Fundamental Problems in Algorithmic Algebra

Chee Keng Yap
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012

September 8, 1993

TO BE PUBLISHED BY PRINCETON UNIVERSITY PRESS

Contents

PREFACE

These notes were first written for a course on *Algebraic Computing: Solving Systems of Polynomial Equations*, given in the Spring Semester of 1989 at the Free University of Berlin. They were thoroughly revised following a similar course at the Courant Institute in the Spring of 1992. Prerequisites are an undergraduate course in algebra and a graduate course in algorithmics.

I regard this course as an introduction to computer algebra. The subject matter ('starting from the Fundamental Theorem of Algebra') is as classical as one gets in theoretical computer science, and yet it is refreshingly contemporary in interest. This is because the complexity viewpoint exposes many classical questions to new light. There is a common misunderstanding that equates computational mathematics with numerical analysis. In fact, it seems to me that the older name of "symbolic manipulation" given to our field arose as a direct contrast to "numerical computation". The preferred name today is "computer algebra", although I feel that "algorithmic algebra" gives a better emphasis to the fundamental nature of the subject. In any case, computer algebra uses quite distinct techniques, and satisfies requirements distinct from that in numerical analysis. In many areas of computer application (robotics, computer-aided design, geometric modeling, etc) computer algebra is now recognized as an essential tool. This is partly driven by the wide-spread availability of powerful computer work-stations, and the rise of a new generation of computer algebra systems to take advantage of this computing power.

The full spectrum of activity in computer algebra today covers many important areas that we do not even give a hint of in these lectures: it ranges from more specialized topics such as algorithmic integration theory, to implementation issues in computer algebra systems, to a highly developed and beautiful complexity theory of algebraic problems, to problems in allied application areas such as robot motion planning. Our material is necessarily selective, although we feel that if one must cut one swath from the elementary into the deeper parts of the subject in an introductory course, this is a choice cut. Historically, what we identified as "Fundamental problems" in these lectures were clearly central to the development of algebra and even of mathematics. There is an enormous amount of relevant classical literature on these fundamental problems, in part a testimony to the strong algorithmic nature of mathematics before the twentieth century. Even when restricted to this corpus of knowledge (classical, supplemented by modern algorithmic development), my colleagues will surely notice important gaps. But I hope they may still find this book useful as a launching point into their own favorite areas.

We have tried to keep the style of the book close to the lecture form in which this material originally existed. Of course, we have considerably expanded on the lecture material. This mainly consisted of the filling in of mathematical background: a well-equipped student may skip this. The teacher could convey the central ideas quickly at the expense of generality, for instance, by assuming that the rings under discussion are the "canonical examples" (Z and $F[X]$). One teaching plan is to choose a subset of the material in each Lecture Section of this book for presentation in a 2-hour class (the typical length of class at Courant), with the rest assigned for further reading.

I thank Frau Schottke from the Free University for her dedicated transcription of my original hand-written notes into the computer.

Chee Yap
Greenwich Village
September 8, 1993

 September 8, 1993

Lecture 0
INTRODUCTION

This lecture is an orientation on the central problems that concern us. Specifically, we identify three families of "Fundamental Problems" in algorithmic algebra (§1 – §3). In the rest of the lecture (§4–§9), we briefly discuss the complexity-theoretic background. §10 collects some common mathematical terminology while §11 introduces computer algebra systems. The reader may prefer to skip §4-11 on a first reading, and only use them as a reference.

> All our rings will contain *unity* which is denoted 1 (and distinct from 0). They are commutative except in the case of matrix rings.

The main algebraic structures of interest are:

$\mathbb{N}$	=	natural numbers $0, 1, 2, \ldots$
$\mathbb{Z}$	=	integers
$\mathbb{Q}$	=	rational numbers
$\mathbb{R}$	=	reals
$\mathbb{C}$	=	complex numbers
$R[\mathbf{X}]$	=	polynomial ring in $d \geq 1$ variables $\mathbf{X} = (X_1, \ldots, X_n)$ with coefficients from a ring R.

Let R be any ring. For a univariate polynomial $P \in R[X]$, we let $\deg(P)$ and $\texttt{lead}(P)$ denote its *degree* and *leading coefficient* (or leading coefficient). If $P = 0$ then by definition, $\deg(P) = -\infty$ and $\texttt{lead}(P) = 0$; otherwise $\deg(P) \geq 0$ and $\texttt{lead}(P) \neq 0$. We say P is a (respectively) integer, rational, real or complex polynomial, depending on whether R is $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ or $\mathbb{C}$.

In the course of this book, we will encounter other rings: (e.g., §I.1). With the exception of matrix rings, all our rings are commutative. The basic algebra we assume can be obtained from classics such as van der Waerden [22] or Zariski-Samuel [27, 28].

§1. Fundamental Problem of Algebra

Consider an integer polynomial

$$P(X) = \sum_{i=0}^{n} a_i X^i \qquad (a_i \in \mathbb{Z}, a_n \neq 0). \tag{1}$$

Many of the oldest problems in mathematics stem from attempts to solve the equation

$$P(X) = 0, \tag{2}$$

i.e., to find numbers α such that $P(\alpha) = 0$. We call such an α a *solution* of equation (2); alternatively, α is a *root* or *zero* of the polynomial $P(X)$. By definition, an *algebraic number* is a zero of some polynomial $P \in \mathbb{Z}[X]$. The *Fundamental Theorem of Algebra* states that every non-constant polynomial $P(X) \in \mathbb{C}[X]$ has a root $\alpha \in \mathbb{C}$. Put another way, $\mathbb{C}$ is algebraically closed. d'Alembert first formulated this theorem in 1746 but Gauss gave the first complete proof in his 1799 doctoral thesis

at Helmstedt. It follows that there are n (not necessarily distinct) complex numbers $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ such that the polynomial in (1) is equal to

$$P(X) \equiv a_n \prod_{i=1}^{n} (X - \alpha_i). \tag{3}$$

To see this, suppose α_1 is a root of $P(X)$ as guaranteed by the Fundamental Theorem. Using the *synthetic division algorithm* to divide $P(X)$ by $X - \alpha_1$, we get

$$P(X) = Q_1(X) \cdot (X - \alpha_1) + \beta_1$$

where $Q_1(X)$ is a polynomial of degree $n-1$ with coefficients in $\mathbb{C}$ and $\beta_1 \in \mathbb{C}$. On substituting $X = \alpha_1$, the left-hand side vanishes and the right-hand side becomes β_1. Hence $\beta_1 = 0$. If $n = 1$, then $Q_1(X) = a_n$ and we are done. Otherwise, this argument can be repeated on $Q_1(X)$ to yield equation (3).

The computational version of the Fundamental Theorem of Algebra is the problem of finding roots of a univariate polynomial. We may dub this the *Fundamental Problem of Computational Algebra* (or *Fundamental Computational Problem of Algebra*). The Fundamental Theorem is about complex numbers. For our purposes, we slightly extend the context as follows. If $R_0 \subseteq R_1$ are rings, the *Fundamental Problem* for the pair (R_0, R_1) is this:

Given $P(X) \in R_0[X]$, solve the equation $P(X) = 0$ in R_1.

We are mainly interested in cases where $\mathbb{Z} \subseteq R_0 \subseteq R_1 \subseteq \mathbb{C}$. The three main versions are where (R_0, R_1) equals $(\mathbb{Z}, \mathbb{Z}), (\mathbb{Z}, \mathbb{R})$ and $(\mathbb{Z}, \mathbb{C})$, respectively. We call them the *Diophantine*, *real* and *complex versions* (respectively) of the Fundamental Problem.

What does it mean "to solve $P(X) = 0$ in R_1"? The most natural interpretation is that we want to enumerate all the roots of P that lie in R_1. Besides this *enumeration interpretation*, we consider two other possibilities: the *existential interpretation* simply wants to know if P has a root in R_1, and the *counting interpretation* wants to know the number of such roots. To enumerate[1] roots, we must address the representation of these roots. For instance, we will study a representation via "isolating intervals".

Recall another classical version of the Fundamental Problem. Let $R_0 = \mathbb{Z}$ and R_1 denote the complex subring comprising all those elements that can be obtained by applying a finite number of field operations (ring operations plus division by non-zero) and taking nth roots ($n \geq 2$), starting from $\mathbb{Z}$. This is the famous *solution by radicals version* of the Fundamental Problem. It is well known that when $\deg P = 2$, there is always a solution in R_1. What if $\deg P > 2$? This was a major question of the 16th century, challenging the best mathematicians of its day. We now know that solution by radicals exists for $\deg P = 3$ (Tartaglia, 1499-1557) and $\deg P = 4$ (variously ascribed to Ferrari (1522-1565) or Bombelli (1579)). These methods were widely discussed, especially after they were published by Cardan (1501-1576) in his classic Ars magna, "The Great Art", (1545). This was *the* algebra book until Descartes' (1637) and Euler's *Algebra* (1770). Abel (1824) (also Wantzel) show that there is no solution by radicals for a general polynomial of degree 5. Ruffini had a prior though incomplete proof. This kills the hope for a single formula which solves *all* quintic polynomials. This still leaves open the possibility that for *each* quintic polynomial, there is a formula to extract its roots. But it is not hard to dismiss this possibility: for example, an explicit quintic polynomial that

[1] There is possible confusion here: the word "enumerate" means to "count" as well as to "list by name". Since we are interested in both meanings here, we have to appropriate the word "enumerate" for only one of these two senses. In this book, we try to use it only in the latter sense.

does not admit solution by radicals is $P(X) = X^5 - 16X + 2$ (see [3, p.574]). Miller and Landau [12] (also [26]) revisits these question from a complexity viewpoint. The above historical comments may be pursued more fully in, for example, Struik's volume [21].

Remarks:. The Fundamental Problem of algebra used to come under the rubric "theory of equations", which nowadays is absorbed into other areas of mathematics. In these lectures, we are interested in general and effective methods, and we are mainly interested in real solutions.

§2. Fundamental Problem of Classical Algebraic Geometry

To generalize the Fundamental Problem of algebra, we continue to fix two rings, $\mathbb{Z} \subseteq R_0 \subseteq R_1 \subseteq \mathbb{C}$. First consider a bivariate polynomial

$$P(X, Y) \in R_0[X, Y]. \tag{4}$$

Let $\text{Zero}(P)$ denote the set of R_1-solutions of the equation $P = 0$, *i.e.*, $(\alpha, \beta) \in R_1^2$ such that $P(\alpha, \beta) = 0$. The *zero set* $\text{Zero}(P)$ of P is generally an infinite set. In case $R_1 = \mathbb{R}$, the set $\text{Zero}(P)$ is a planar curve that can be plotted and visualized. Just as solutions to equation (2) are called algebraic numbers, the zero sets of bivariate integer polynomials are called *algebraic curves*. But there is no reason to stop at two variables. For $d \geq 3$ variables, the zero set of an integer polynomial in d variables is called an *algebraic hypersurface*: we reserve the term *surface* for the special case $d = 3$.

Given two surfaces defined by the equations $P(X, Y, Z) = 0$ and $Q(X, Y, Z) = 0$, their intersection is generally a curvilinear set of triples $(\alpha, \beta, \gamma) \in R_1^3$, consisting of all simultaneous solutions to the pair of simultaneous equations $P = 0,\ Q = 0$. We may extend our previous notation and write $\text{Zero}(P, Q)$ for this intersection. More generally, we want the simultaneous solutions to a *system of* $m \geq 1$ *polynomial equations* in $d \geq 1$ variables:

$$\left.\begin{array}{c} P_1 = 0 \\ P_2 = 0 \\ \vdots \\ P_m = 0 \end{array}\right\} \qquad (\text{where } \ P_i \in R_0[X_1, \ldots, X_d]) \tag{5}$$

A point $(\alpha_1, \ldots, \alpha_d) \in R_1^d$ is called a *solution* of the system of equations (5) or a *zero* of the set $\{P_1, \ldots, P_m\}$ provided $P_i(\alpha_1, \ldots, \alpha_d) = 0$ for $i = 1, \ldots, m$. In general, for any subset $J \subseteq R_0[\mathbf{X}]$, let $\text{Zero}(J) \subseteq R_1^d$ denote the *zero set* of J. To denote the dependence on R_1, we may also write $\text{Zero}_{R_1}(J)$. If R_1 is a field, we also call a zero set an *algebraic set*. Since the primary objects of study in classical algebraic geometry are algebraic sets, we may call the problem of solving the system (5) the *Fundamental (Computational) Problem of classical algebraic geometry*. If each P_i is linear in (5), we are looking at a system of linear equations. One might call this is the *Fundamental (Computational) Problem of linear algebra*. Of course, linear systems are well understood, and their solution technique will form the basis for solving nonlinear systems.

Again, we have three natural meanings to the expression "solving the system of equations (5) in R_1": (i) The existential interpretation asks if $\text{Zero}(P_1, \ldots, P_m)$ is empty. (ii) The counting interpretation asks for the cardinality of the zero set. In case the cardinality is "infinity", we could refine the question by asking for the *dimension* of the zero set. (iii) Finally, the enumeration interpretation poses no problems when there are only finitely many solutions. This is because the coordinates of these solutions turn out to be algebraic numbers and so they could be explicitly enumerated. It becomes problematic when the zero set is infinite. Luckily, when $R_1 = \mathbb{R}$ or $\mathbb{C}$, such zero sets are well-behaved topologically, and each zero set consists of a finite number of connected components.

(For that matter, the counting interpretation can be re-interpreted to mean counting the number of components of each dimension.) A typical interpretation of "enumeration" is "give at least one sample point from each connected component". For real planar curves, this interpretation is useful for plotting the curve since the usual method is to "trace" each component by starting from any point in the component.

Note that we have moved from algebra (numbers) to geometry (curves and surfaces). In recognition of this, we adopt the geometric language of "points and space". The set R_1^d (d-fold Cartesian product of R_1) is called the *d-dimensional affine space* of R_1, denoted $\mathbb{A}^d(R_1)$. Elements of $\mathbb{A}^d(R_1)$ are called *d-points* or simply *points*. Our zero sets are subsets of this affine space $\mathbb{A}^d(R_1)$. In fact, $\mathbb{A}^d(R_1)$ can be given a topology (the Zariski topology) in which zero sets are the closed sets.

There are classical techniques via elimination theory for solving these Fundamental Problems. The recent years has seen a revival of these techniques as well as major advances. In one line of work, Wu Wen-tsun exploited Ritt's idea of characteristic sets to give new methods for solving (5) rather efficiently in the complex case, $R_1 = \mathbb{C}$. These methods turn out to be useful for proving theorems in elementary geometry as well [25]. But many applications are confined to the real case ($R_1 = \mathbb{R}$). Unfortunately, it is a general phenomenon that real algebraic sets do not behave as regularly as the corresponding complex ones. This is already evident in the univariate case: the Fundamental Theorem of Algebra fails for real solutions. In view of this, most mathematical literature treats the complex case. More generally, they apply to any algebraically closed field. There is now a growing body of results for real algebraic sets.

Another step traditionally taken to "regularize" algebraic sets is to consider projective sets, which abolish the distinction between finite and infinite points. A *projective d-dimensional point* is simply an equivalence class of the set $\mathbb{A}^{d+1}(R_1)\backslash\{(0,\ldots,0)\}$, where two non-zero $(d+1)$-points are *equivalent* if one is a constant multiple of the other. We use $\mathbb{P}^d(R_1)$ to denote the *d-dimensional projective space* of R_1.

Semialgebraic sets. The real case admits a generalization of the system (5). We can view (5) as a *conjunction* of basic predicates of the form "$P_i = 0$":

$$(P_1 = 0) \ \wedge \ (P_2 = 0) \wedge \cdots \wedge (P_m = 0).$$

We generalize this to an arbitrary Boolean combination of basic predicates, where a basic predicate now has the form $(P = 0)$ or $(P > 0)$ or $(P \geq 0)$. For instance,

$$((P = 0) \ \wedge \ (Q > 0)) \ \vee \ \neg(R \geq 0)$$

is a Boolean combination of three basic predicates where P, Q, R are polynomials. The set of real solutions to such a predicate is called a *semi-algebraic set* (or a *Tarski set*). We have effective methods of computing semi-algebraic sets, thanks to the pioneering work of Tarski and Collins [7]. Recent work by various researchers have reduced the complexity of these algorithms from double exponential time to single exponential space [15]. This survey also describes to applications of semi-algebraic in algorithmic robotics, solid modeling and geometric theorem proving. Recent books on real algebraic sets include [4, 2, 10].

§3. Fundamental Problem of Ideal Theory

Algebraic sets are basically geometric objects: witness the language of "space, points, curves, surfaces". Now we switch from the geometric viewpoint (back!) to an algebraic one. One of the beauties of this subject is this interplay between geometry and algebra.

Fix $\mathbb{Z} \subseteq R_0 \subseteq R_1 \subseteq \mathbb{C}$ as before. A polynomial $P(\mathbf{X}) \in R_0[\mathbf{X}]$ is said to *vanish* on a subset $U \subseteq \mathbb{A}^d(R_1)$ if for all $\mathbf{a} \in U$, $P(\mathbf{a}) = 0$. Define

$$\textsc{Ideal}(U) \subseteq R_0[\mathbf{X}]$$

to comprise all polynomials $P \in R_0[\mathbf{X}]$ that vanish on U. The set $\textsc{Ideal}(U)$ is an ideal. Recall that a non-empty subset $J \subseteq R$ of a ring R is an *ideal* if it satisfies the properties

1. $a, b \in J \Rightarrow a - b \in J$
2. $c \in R$, $a \in J \Rightarrow ca \in J$.

For any $a_1, \ldots, a_m \in R$ and $R' \supseteq R$, the set $(a_1, \ldots, a_m)_{R'}$ defined by

$$(a_1, \ldots, a_m)_{R'} := \{\sum_{i=1}^{m} a_i b_i : b_1, \ldots, b_m \in R'\}$$

is an ideal, the ideal *generated by* $a_1, \ldots, a_m$ in R'. We usually omit the subscript R' if this is understood.

The Fundamental Problem of classical algebraic geometry (see Equation (5)) can be viewed as computing (some characteristic property of) the zero set defined by the input polynomials $P_1, \ldots, P_m$. But note that

$$\textsc{Zero}(P_1, \ldots, P_m) = \textsc{Zero}(I)$$

where I is the ideal generated by $P_1, \ldots, P_m$. Hence we might as well assume that the input to the Fundamental Problem is the ideal I (represented by a set of generators). *This suggests that we view ideals to be the algebraic analogue of zero sets.* We may then ask for the algebraic analogue of the Fundamental Problem of classical algebraic geometry. A naive answer is that, "given $P_1, \ldots, P_m$, to enumerate the set $(P_1, \ldots, P_m)$". Of course, this is impossible. But we effectively "know" a set S if, for any purported member x, we can decisively say whether or not x is a member of S. Thus we reformulate the enumerative problem as the *Ideal Membership Problem*:

$$\textit{Given } P_0, P_1, \ldots, P_m \in R_0[\mathbf{X}], \textit{ is } P_0 \textit{ in } (P_1, \ldots, P_m)?$$

Where does R_1 come in? Well, the ideal $(P_1, \ldots, P_m)$ is assumed to be generated in $R_1[\mathbf{X}]$. We shall introduce effective methods to solve this problem. The technique of Gröbner bases (as popularized by Buchberger) is notable. There is strong historical basis for our claim that the ideal membership problem is fundamental: van der Waerden [22, vol. 2, p. 159] calls it the "main problem of ideal theory in polynomial rings". Macaulay in the introduction to his 1916 monograph [14] states that the "object of the algebraic theory [of ideals] is to discover those general properties of [an ideal] which will afford a means of answering the question whether a given polynomial is a member of a given [ideal] or not".

How general are the ideals of the form $(P_1, \ldots, P_m)$? The only ideals that might not be of this form are those that cannot be generated by a finite number of polynomials. The answer is provided by what is perhaps the starting point of modern algebraic geometry: the *Hilbert!Basis Theore*. A ring R is called *Noetherian* if all its ideals are finitely generated. For example, if R is a field, then it is Noetherian since its only ideals are (0) and (1). The Hilbert Basis Theorem says that $R[\mathbf{X}]$ is Noetherian if R is Noetherian. This theorem is crucial[2] from a constructive viewpoint: it assures us that although ideals are potentially infinite sets, they are finitely describable.

[2]The paradox is, many view the original proof of this theorem as initiating the modern tendencies toward non-constructive proof methods.

We now have a mapping

$$U \mapsto \textsc{Ideal}(U) \tag{6}$$

from subsets of $\mathbb{A}^d(R_1)$ to the ideals of $R_0[\mathbf{X}]$, and conversely a mapping

$$J \mapsto \textsc{Zero}(J) \tag{7}$$

from subsets of $R_0[\mathbf{X}]$ to algebraic sets of $\mathbb{A}^d(R_1)$. It is not hard to see that

$$J \subseteq \textsc{Ideal}(\textsc{Zero}(J)), \qquad U \subseteq \textsc{Zero}(\textsc{Ideal}(U)) \tag{8}$$

for all subsets $J \subseteq R_0[\mathbf{X}]$ and $U \subseteq \mathbb{A}^d(R_1)$. Two other basic identities are:

$$\begin{aligned} \textsc{Zero}(\textsc{Ideal}(\textsc{Zero}(J))) &= \textsc{Zero}(J), \qquad J \subseteq R_0[\mathbf{X}], \\ \textsc{Ideal}(\textsc{Zero}(\textsc{Ideal}(U))) &= \textsc{Ideal}(U), \qquad U \subseteq \mathbb{A}^d(R_1), \end{aligned} \tag{9}$$

We prove the first equality: If $\mathbf{a} \in \textsc{Zero}(J)$ then for all $P \in \textsc{Ideal}(\textsc{Zero}(J))$, $P(\mathbf{a}) = 0$. Hence $\mathbf{a} \in \textsc{Zero}(\textsc{Ideal}(\textsc{Zero}(J))$. Conversely, if $\mathbf{a} \in \textsc{Zero}(\textsc{Ideal}(\textsc{Zero}(J))$ then $P(\mathbf{a}) = 0$ for all $P \in \textsc{Ideal}(\textsc{Zero}(J))$. But since $J \subseteq \textsc{Ideal}(\textsc{Zero}(J))$, this means that $P(\mathbf{a}) = 0$ for all $P \in J$. Hence $\mathbf{a} \in \textsc{Zero}(J)$. The second equality (9) is left as an exercise.

If we restrict the domain of the map in (6) to algebraic sets and the domain of the map in (7) to ideals, would these two maps be inverses of each other? The answer is no, based on a simple observation: An ideal I is called *radical* if for all integers $n \geq 1$, $P^n \in I$ implies $P \in I$. It is not hard to check that $\textsc{Ideal}(U)$ is radical. On the other hand, the ideal $(X^2) \in \mathbb{Z}[X]$ is clearly non-radical.

It turns out that if we restrict the ideals to radical ideals, then $\textsc{Ideal}(\cdot)$ and $\textsc{Zero}(\cdot)$ would be inverses of each other. This is captured in the *Hilbert Nullstellensatz* (or, Hilbert's Zero Theorem in English). After the Basis Theorem, this is perhaps the next fundamental theorem of algebraic geometry. It states that if P vanishes on the zero set of an ideal I then some power P^n of P belongs to I. As a consequence,

$$I = \textsc{Ideal}(\textsc{Zero}(I)) \Leftrightarrow I \text{ is radical.}$$

In proof: Clearly the left-hand side implies I is radical. Conversely, if I is radical, it suffices to show that $\textsc{Ideal}(\textsc{Zero}(I)) \subseteq I$. Say $P \in \textsc{Ideal}(\textsc{Zero}(I))$. Then the Nullstellensatz implies $P^n \in I$ for some n. Hence $P \in I$ since I is radical, completing our proof.

We now have a bijective correspondence between algebraic sets and radical ideals. This implies that ideals in general carry more information than algebraic sets. For instance, the ideals (X) and (X^2) have the same zero set, *viz.*, $X = 0$. But the unique zero of (X^2) has multiplicity 2.

The ideal-theoretic approach (often attached to the name of E. Noether) characterizes the transition from classical to "modern" algebraic geometry. "Post-modern" algebraic geometry has gone on to more abstract objects such as schemes. Not much constructive questions are raised at this level, perhaps because the abstract questions are hard enough. The reader interested in the profound transformation that algebraic geometry has undergone over the centuries may consult Dieudonné [9] who described the subject in "seven epochs". The current challenge for constructive algebraic geometry appears to be at the levels of classical algebraic geometry and at the ideal-theoretic level. For instance, Brownawell [6]and others have recently given us effective versions of classical results such as the Hilbert Nullstellensatz. Such results yields complexity bounds that are necessary for efficient algorithms (see Exercise).

This concludes our orientation to the central problems that motivates this book. This exercise is pedagogically useful for simplifying the algebraic-geometric landscape for students. However, the richness of this subject and its complex historical development ensures that, in the opinion of some

experts, we have made gross oversimplifications. Perhaps an account similar to what we presented is too much to hope for – we have to leave this to the professional historians to tell us the full story. In any case, having *selected* our core material, the rest of the book will attempt to treat and view it through the lens of computational complexity theory. The remaining sections of this lecture addresses this.

Exercises

Exercise 3.1: Show relation (8), and relation (9). □

Exercise 3.2: Show that the ideal membership problem is polynomial-time equivalent to the problem of checking if two sets of elements generate the same ideal: Is $(a_1, \ldots, a_m) = (b_1, \ldots, b_n)$? [Two problems are polynomial-time equivalent if one can be reduced to the other in polynomial-time and vice-versa.] □

Exercise 3.3*: a) Given $P_0, P_1, \ldots, P_m \in \mathbb{Q}[X_1, \ldots, X_d]$, where these polynomials have degree at most n, there is a known double exponential bound $B(d, n)$ such that if $P_0 \in (P_1, \ldots, P_m)$ there there exists polynomials $Q_1, \ldots, Q_m$ of degree at most $B(d, n)$ such that

$$P_0 = P_1 Q_1 + \cdots + P_m Q_m.$$

Note that $B(d, n)$ does not depend on m. Use this fact to construct a double exponential time algorithm for ideal membership.
b) Does the bound $B(d, n)$ translate into a corresponding bound for $\mathbb{Z}[X_1, \ldots, X_d]$? □

§4. Representation and Size

We switch from mathematics to computer science. To investigate the computational complexity of the Fundamental Problems, we need tools from complexity theory. The complexity of a problem is a function of some size measure on its input instances. The size of a problem instance depends on its representation.

Here we describe the representation of some basic objects that we compute with. For each class of objects, we choose a notion of "size".

Integers: Each integer $n \in \mathbb{Z}$ is given the binary notation and has *(bit-)size*

$$\texttt{size}(n) = 1 + \lceil \log(|n| + 1) \rceil$$

where logarithms are always base 2 unless otherwise stated. The term "$1 + \ldots$" takes care of the sign-bit.

Rationals: Each rational number $p/q \in \mathbb{Q}$ is represented as a pair of integers with $q > 0$. We do not assume the reduced form of a rational number. The *(bit-)size* is given by

$$\texttt{size}\left(\frac{p}{q}\right) = \texttt{size}(p) + \texttt{size}(q) + \log(\texttt{size}(p))$$

where the "$+ \log(\texttt{size}(p))$" term indicates the separation between the two integers.

Matrices: The default is the *dense representation* of matrices so that zero entries must be explicitly represented. An $m \times n$ matrix $M = (a_{ij})$ has *(bit-)size*

$$\texttt{size}(M) = \sum_{i=1}^{m}\sum_{j=1}^{n}(\texttt{size}(a_{ij}) + \log(\texttt{size}(a_{ij})))$$

where the "$+\log(size(a_{ij}))$" term allows each entry of M to indicate its own bits (this is sometimes called the "self-limiting" encoding). Alternatively, a simpler but less efficient encoding is to essentially double the number of bits

$$\texttt{size}(M) = \sum_{i=1}^{m}\sum_{j=1}^{n}(2 + 2\texttt{size}(a_{ij}))\,.$$

This encoding replaces each 0 by "00" and each 1 by "11", and introduces a separator sequence "01" between consecutive entries.

Polynomials: The default is the *dense representation* of polynomials. So a degree-n univariate polynomial is represented as a $(n+1)$-tuple of its coefficients – and the size of the $(n+1)$-tuple is already covered by the above size consideration for matrices. *(bit-)size*

Other representations (especially of multivariate polynomials) can be more involved. In contrast to dense representations, *sparse representations* refer to *sparse representation* those whose sizes grow linearly with the number of non-zero terms of a polynomial. In general, such compact representations greatly increase (not decrease!) the computational complexity of problems. For instance, Plaisted [16, 17] has shown that deciding if two sparse univariate integer polynomials are relatively prime is NP-hard. In contrast, this problem is polynomial-time solvable in in the dense representation (Lecture II).

Ideals: Usually, 'ideals' refer to polynomial ideals. An ideal I is represented by any finite set $\{P_1, \ldots, P_n\}$ of elements that generate it: $I = (P_1, \ldots, P_n)$. The size of this representation just the sum of the sizes of the generators. Clearly, the representation of an ideal is far from unique.

The representations and sizes of other algebraic objects (such as algebraic numbers) will be discussed as they arise.

§5. Computational Models

We briefly review four models of computation: *Turing machines*, *Boolean circuits*, *algebraic programs* and *random access machines*. With each model, we will note some natural complexity measures (time, space, size, etc), including their correspondences across models. We will be quite informal since many of our assertions about these models will be (with some coaching) self-evident. A reference for machine models is Aho, Hopcroft and Ullman [1]. For a more comprehensive treatment of the algebraic model, see Borodin and Munro [5]; for the Boolean model, see Wegener [24].

I. Turing machine model. The Turing (machine) model is embodied in the *multitape Turing machine*, in which inputs are represented by a binary string. Our representation of objects and definition of sizes in the last section are especially appropriate for this model of computation. The machine is essentially a finite state automaton (called its *finite state control*) equipped with a finite set of doubly-infinite tapes, including a distinguished *input tape* Each tape is divided into cells indexed by the integers. Each cell contains a symbol from a finite alphabet. Each tape has a head

which scans some cell at any moment. A Turing machine may operate in a variety of *computational modes* such as *deterministic*, *nondeterministic* or *randomized*; and in addition, the machine can be generalized from sequential to parallel modes in many ways. We mostly assume the deterministic-sequential mode in this book. In this case, a Turing machine operates according to the specification of its finite state control: in each step, depending on the current state and the symbols being scanned under each tape head, the transition table specifies the next state, modifies the symbols under each head and moves each head to a neighboring cell. The main complexity measures in the Turing model are *time* (the number of steps in a computation), *space* (the number of cells used during a computation) and *reversal* (the number of times a tape head reverses its direction).

II. Boolean circuit model. This model is based on *Boolean circuits*. A Boolean circuit is a directed acyclic finite graph whose nodes are classified as either *input nodes* or *gates*. The input nodes have in-degree 0 and are labeled by an input variable; gates are labeled by Boolean functions with in-degree equal to the arity of the label. The set of Boolean functions which can be used as gate labels is called the *basis!of computational models* of the model. In this book, we may take the basis to be the set of Boolean functions of at most two inputs. We also assume no á priori bound on the out-degree of a gate. The three main complexity measures here are *circuit size* (the number of gates), *circuit depth* (the longest path) and *circuit width* (roughly, the largest antichain).

A circuit can only compute a function on a fixed number of Boolean inputs. Hence to compare the Boolean circuit model to the Turing machine model, we need to consider a *circuit family*, which is an infinite sequence $(C_0, C_1, C_2, \ldots)$ of circuits, one for each input size. Because there is no *a priori* connection between the circuits in a circuit family, we call such a family *non-uniform*. *non-uniform*. For this reason, we call Boolean circuits a "non-uniform model" as opposed to Turing machines which is "uniform". Circuit size can be identified with time on the Turing machine. Circuit depth is more subtle, but it can (following Jia-wei Hong be identified with "reversals" on Turing machines.

It turns out that the Boolean complexity of *any* problem is at most $2^n/n$ (see [24]). Clearly this is a severe restriction on the generality of the model. But it is possible to make Boolean circuit families "uniform" in several ways and the actual choice is usually not critical. For instance, we may require that there is a Turing machine using logarithmic space that, on input n in binary, constructs the (encoded) nth circuit of the circuit family. The resulting *uniform Boolean complexity* is now polynomially related to Turing complexity. Still, the non-uniform model suffices for many applications (see §8), and that is what we will use in this book.

Encodings and bit models. The previous two models are called *bit models* because mathematical objects must first be encoded as binary strings before they can be used on these two models. The issue of encoding may be quite significant. But we may get around this by assuming standard conventions such as binary encoding of numbers, list representation of sets, etc. In algorithmic algebra, it is sometimes useful to avoid encodings by incorporating the relevant algebraic structures directly into the computational model. This leads us to our next model.

III. Algebraic program models. In *algebraic programs*, we must fix some algebraic structures (such as $\mathbb{Z}$, polynomials or matrices over a ring R) and specify a set of primitive algebraic operations called the *basis!of computational models* of the model. Usually the basis includes the ring operations $(+, -, \times)$, possibly supplemented by other operations appropriate to the underlying algebraic structure. A common supplement is some form of root finding (e.g., multiplicative inverse, radical extraction or general root extraction), and GCD. The algebraic program model is thus a class of models based on different algebraic structures and different bases.

An *algebraic program* is defined to be a rooted ordered tree T where each node represents either an *assignment step* of the form

$$V \leftarrow F(V_1, \ldots, V_k),$$

or a *branch step* of the form

$$F(V_1, \ldots, V_k) : 0.$$

Here, F is a k-ary operation in the basis and each V_i is either an input variable, a constant or a variable that has been assigned a value further up the tree. The out-degree of an assignment node is 1; the out-degree of a branch node is 2, corresponding to the outcomes $F(V_1, \ldots, V_k) = 0$ and $F(V_1, \ldots, V_k) \neq 0$, respectively. If the underlying algebraic structure is real, the branch steps can be extended to a 3-way branch, corresponding to $F(V_1, \ldots, V_k) < 0, = 0$ or > 0. At the leaves of T, we fix some convention for specifying the output.

The *input size* is just the number of input variables. The main complexity measure studied with this model is *time*, the length of the longest path in T. Note that we charge a unit cost to each basic operation. This could easily be generalized. For instance, a multiplication step in which one of the operands is a constant (*i.e.*, does not depend on the input parameters) may be charged nothing. This originated with Ostrowski who wrote one of the first papers in algebraic complexity.

Like Boolean circuits, this model is non-uniform because each algebraic program solves problems of a fixed size. Again, we introduce the *algebraic program family* which is an infinite set of algebraic programs, one for each input size.

When an algebraic program has no branch steps, it is called a *straight-line program*. To see that in general we need branching, consider algebraic programs to compute the GCD (see Exercise below).

IV. RAM model. Finally, consider the *random access machine* model of computation. Each RAM is defined by a finite set of instructions, rather as in assembly languages. These instructions make reference to operands called *registers* Each register can hold an arbitrarily large integer and is *indexed* by a natural number. If n is a natural number, we can denote its contents by $\langle n \rangle$. Thus $\langle\langle n \rangle\rangle$ refers to the contents of the register whose index is $\langle n \rangle$. In addition to the usual registers, there is an unindexed register called the *accumulator* in which all computations are done (so to speak). The RAM instruction sets can be defined variously and have the simple format

INSTRUCTION OPERAND

where OPERAND is either n or $\langle n \rangle$ and n is the index of a register. We call the operand *direct* or *indirect* depending on whether we have n or $\langle n \rangle$. We have five RAM instructions: a STORE and LOAD instruction (to put the contents of the accumulator to register n and vice-versa), a TEST instruction (to skip the next instruction if $\langle n \rangle$ is zero) and a SUCC operation (to add one to the content of the accumulator). For example, 'LOAD 5' instructs the RAM to put $\langle 5 \rangle$ into the accumulator; but 'LOAD $\langle 5 \rangle$' puts $\langle\langle 5 \rangle\rangle$ into the accumulator; 'TEST 3' causes the next instruction to be skipped if $\langle 3 \rangle = 0$; 'SUCC' will increment the accumulator content by one. There are two main models of time-complexity for RAM models: in the *unit cost model*, each executed instruction is charged 1 unit of time. In contrast, the *logarithmic cost model*, charges $\lceil \lg(|n| + |\langle n \rangle|) \rceil$ whenever a register n is accessed. Note that an instruction accesses one or two registers, depending on whether the operand is direct or indirect. It is known that the logarithmic cost RAM is within a quadratic factor of the Turing time complexity. The above RAM model is called the *successor RAM* to distinguish it from other variants, which we now briefly note. More powerful arithmetic operations (ADDITION, SUBTRACTION and even MULTIPLICATION) are sometimes included in the instruction set. Schönhage describes an even simpler RAM model than the above model,

essentially by making the operand of each of the above instructions implicit. He shows that this simple model is real-time equivalent to the above one.

Exercises

Exercise 5.1:

(a) Describe an algebraic program for computing the GCD of two integers. (Hint: implement the Euclidean algorithm. Note that the input size is 2 and this computation tree must be infinite although it halts for all inputs.)

(b) Show that the integer GCD cannot be computed by a straight-line program.

(c) Describe an algebraic program for computing the GCD of two rational polynomials $P(X) = \sum_{i=0}^{n} a_i X^i$ and $Q(X) = \sum_{i=0}^{m} b_i X^i$. The input variables are $a_0, a_1, \ldots, a_n, b_0, \ldots, b_m$, so the input size is $n + m + 2$. The output is the set of coefficients of $\texttt{GCD}(P, Q)$. □

§6. Asymptotic Notations

Once a computational model is chosen, there are additional decisions to make before we get a "complexity model". This book emphasizes mainly the *worst case time measure* in each of our computational models. To each machine or program A in our computational model, this associates a function $T_A(n)$ that specifies the worst case number of time steps used by A, over all inputs of size n. Call $T_A(n)$ the *complexity of* A. Abstractly, we may define a *complexity model* to comprise a computational model together with an associated complexity function $T_A(n)$ for each A. The complexity models in this book are: *Turing complexity model*, *Boolean complexity model*, *algebraic complexity model*, and *RAM complexity model*. For instance, the Turing complexity model refers to the worst-case time complexity of Turing machines. "Algebraic complexity model" is a generic term that, in any specific instance, must be instantiated by some choice of algebraic structure and basis operations.

We intend to distinguish complexity functions up to constant multiplicative factors and up to their eventual behavior. To facilitate this, we introduce some important concepts.

Definition 1 *A* complexity function *is a real partial function* $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *such that* $f(x)$ *is defined for all sufficiently large natural numbers* $x \in \mathbb{N}$. *Moreover, for sufficiently large* x, $f(x) \geq 0$ *whenever* x *is defined.*

If $f(x)$ is undefined, we write $f(x) \uparrow$, and this is to be distinguished from the case $f(x) = \infty$. *Note that we require that* $f(x)$ *be eventually non-negative.* We often use familiar partial functions such as $\log x$ and 2^x as complexity functions, even though we are mainly interested in their values at $\mathbb{N}$. Note that if f, g are complexity functions then so are

$$f + g, \quad fg, \quad f^g, \quad f \circ g$$

where in the last case, we need to assume that $(f \circ g)(x) = f(g(x))$ is defined for sufficiently large $x \in \mathbb{N}$.

The big-Oh notation. Let f, g be complexity functions. We say f *dominates* g if $f(x) \geq g(x)$ for all sufficiently large x, and provided $f(x), g(x)$ are both defined. By "sufficiently large x" or "large enough x" we mean "for all $x \geq x_0$" where x_0 is some unspecified constant.

The *big-Oh notation*asymptotic notation!big-Oh is the most famous member of a family of asymptotic notations. The prototypical use of this notation goes as follows. We say f *is big-Oh of* g (or, f *is order of* g) and write

$$f = O(g) \tag{10}$$

if there is a constant $C > 0$ such that $C \cdot g(x)$ dominates $f(x)$. As examples of usage, $f(x) = O(1)$ (respectively, $f(x) = x^{O(1)}$) means that $f(x)$ is eventually bounded by some constant (respectively, by some polynomial). Or again, $n \log n = O(n^2)$ and $1/n = O(1)$ are both true.

Our definition in Equation (10) gives a very specific formula for using the big-Oh notation. We now describe an extension. Recursively define *O-expressions* as follows. Basis: If g is a symbol for a complexity function, then g is an O-expression. Induction: If E_i $(i = 1, 2)$ are O-expressions, then so are

$$O(E_1), \quad E_1 \pm E_2, \quad E_1 E_2, \quad E_1^{E_2}, \quad E_1 \circ E_2.$$

Each O-expression *denotes* a set of complexity functions. Basis: The O-expression g denotes the singleton set $\{\overline{g}\}$ where $\overline{g}$ is the function denoted by g. Induction: If E_i denotes the set of complexity functions $\overline{E}_i$ then the O-expression $O(E_1)$ denotes the set of complexity functions $\overline{f}$ such that there is some $\overline{g} \in \overline{E}_1$ and $C > 0$ and $\overline{f}$ is dominated by $C\overline{g}$. The expression $E_1 + E_2$ denotes the set of functions of the form $f_1 + f_2$ where $f_i \in \overline{E}_i$. Similarly for $E_1 E_2$ (product), $E_1^{E_2}$ (exponentiation) and $E_1 \circ E_2$ (function composition). Finally, we use these O-expressions to assert the containment relationship: we write

$$E_1 = E_2,$$

to mean $\overline{E}_1 \subseteq \overline{E}_2$. Clearly, the equality symbol in this context is asymmetric. In actual usage, we take the usual license of confusing a function symbol g with the function $\overline{g}$ that it denotes. Likewise, we confuse the concept of an O-expression with the set of functions it denotes. By convention, the expressions 'c' ($c \in \mathbb{R}$) and 'n' denote (respectively) the constant function c and the identity function. Then 'n^2' and '$\log n$' are O-expressions denoting the (singleton set containing the) square function and logarithm function. Other examples of O-expressions: $2^{n+O(\log n)}$, $O(O(n)^{\log n} + n^{O(n)} \log\log n)$, $f(n) \circ O(n \log n)$. Of course, all these conventions depends on fixing 'n' as the distinguished variable. Note that $1 + O(1/n)$ and $1 - O(1/n)$ are different O-expressions because of our insistence that complexity functions are eventually non-negative.

The subscripting convention. There is another useful way to extend the basic formulation of Equation (10): instead of viewing its right-hand side "$O(g)$" as denoting a set of functions (and hence the equality sign as set membership '$\in$' or set inclusion '$\subseteq$'), we can view it as denoting some *particular* function $C \cdot g$ that dominates f. The big-Oh notation in this view is just a convenient way of hiding the constant 'C' (it saves us the trouble of inventing a symbol for this constant). In this case, the equality sign is interpreted as the "dominated by" relation, which explains the tendency of some to write '$\leq$' instead of the equality sign. Usually, the need for this interpretation arises because we want to obliquely refer to the implicit constant. For instance, we may want to indicate that the implicit constants in two occurrences of the same O-expression are really the same. To achieve this cross reference, we use a subscripting convention: *we can attach a subscript or subscripts to the O, and this particularizes that O-expression to refer to some fixed function.* Two identical O-expressions with identical subscripts refer to the same implicit constants. By choosing the subscripts judiciously, this notation can be quite effective. For instance, instead of inventing a function symbol $T_A(n) = O(n)$ to denote the running time of a linear-time algorithm A, we may simply use the subscripted expression "$O_A(n)$"; subsequent use of this expression will refer to the same function. Another simple illustration is "$O_3(n) = O_1(n) + O_2(n)$": the sum of two linear functions is linear, with different implicit constant for each subscript.

Related asymptotic notations. We say f is *big-Omega of* g and write

$$f(n) = \Omega(g(n))$$

if there exists a real $C > 0$ such that $f(x)$ dominates $C \cdot g(x)$. We say f is *Theta of* g and write

$$f(n) = \Theta(g(n))$$

if $f = O(g)$ and $f = \Omega(g)$. We normally distinguish complexity functions up to Theta-order. We say f is *small-oh of* g and write

$$f(n) = o(g(n))$$

if $f(n)/g(n) \to 0$ as $n \to \infty$. We say f is *small-omega of* g and write

$$f(n) = \omega(g(n))$$

if $f(n)/g(n) \to \infty$ as $n \to \infty$. We write

$$f \sim g$$

if $f = g[1 \pm o(1)]$. For instance, $n + \log n \sim n$ but not $n + \log n \sim 2n$.

These notations can be extended as in the case of the big-Oh notation. The semantics of mixing these notations are less obvious and is, in any case, not needed.

§7. Complexity of Multiplication

We introduce three "intrinsic" complexity functions,

$$\mathrm{M}_B(n), \quad \mathrm{M}_A(n), \quad \mathrm{MM}(n)$$

related to multiplication in various domains under various complexity models. These functions are useful in bounding other complexity functions. This leads to a discussion of intrinsic complexity.

Complexity of multiplication. Let us first fix the model of computation to be the multitape Turing machine. We are interested in the *intrinsic Turing complexity* T_P of a computational problem P, namely the intrinsic (time) cost of solving P on the Turing machine model. Intuitively, we expect $T_P = T_P(n)$ to be a complexity function, corresponding to the "optimal" Turing machine for P. If there is no optimal Turing machine, this is problematic – – see below for a proper treatment of this. If P is the problem of multiplying two binary integers, then the fundamental quantity $T_P(n)$ appears in the complexity bounds of many other problems, and is given the special notation

$$\mathrm{M}_B(n)$$

in this book. For now, we will assume that $\mathrm{M}_B(n)$ is a complexity function. The best upper bound for $\mathrm{M}_B(n)$ is

$$\mathrm{M}_B(n) = O(n \log n \log\log n), \tag{11}$$

from a celebrated result [20] of Schönhage and Strassen (1971). To simplify our display of such bounds (cf. [18, 13]), we write $\mathcal{L}^k(n)$ ($k \geq 1$) to denote some fixed but non-specific function $f(n)$ that satisfies

$$\frac{f(n)}{\log^k n} = o(\log n).$$

If $k = 1$, the superscript in $\mathcal{L}^1(n)$ is omitted. In this notation, equation (11) simplifies to

$$\mathrm{M}_B(n) = n\mathcal{L}(n).$$

Note that we need not explicitly write the big-Oh here since this is implied by the $\mathcal{L}(n)$ notation. Schönhage [19] (cf. [11, p. 295]) has shown that the complexity of integer multiplication takes a simpler form with alternative computational models (see §6): *A successor RAM can multiply two n-bit integers in $O(n)$ time under the unit cost model, and in $O(n \log n)$ time in the logarithmic cost model.*

Next we introduce the *algebraic complexity of multiplying two degree n polynomials*, denoted

$$\mathrm{M}_A(n).$$

The basis (§6) for our algebraic programs is comprised of the ring operations of R, where the polynomials are from $R[X]$. Trivially, $\mathrm{M}_A(n) = O(n^2)$ but Lecture I will show that

$$\mathrm{M}_A(n) = O(n \log n).$$

Finally, we introduce the *algebraic complexity of multiplying two $n \times n$ matrices.* We assume the basis is comprised of the ring operations of a ring R, where the matrix entries come from R. This is another fundamental quantity which will be denoted by

$$\mathrm{MM}(n)$$

in this book. Clearly $\mathrm{MM}(n) = O(n^3)$ but a celebrated result of Strassen (1968) shows that this is suboptimal. The current record (see Lecture I) is

$$\mathrm{MM}(n) = O(n^{2.376}). \tag{12}$$

On Intrinsic Complexity.

The notation "$\mathrm{M}_B(n)$" is not rigorous when naively interpreted as a complexity function. Let us see why. More generally, let us fix a *complexity model M*: this means we fix a computational model (Turing machines, RAM, etc) and associate a complexity function $T_A(n)$ to each program A in M as in §7. But complexity theory really begins when we associate an *intrinsic complexity function $T_P(n)$* with each computational problem P. Thus, $\mathrm{M}_B(n)$ is the intrinsic complexity function for the problem of multiplying two binary integers in the standard (worst-case time) Turing complexity model. But how shall we define $T_P(n)$?

First of all, we need to clarify the concept of a "computational problem". One way is to introduce a logical language for specifying problems. But for our purposes, *we will simply identify a computational problem P with a set of programs in model M*. The set P comprises those programs in M that is said to "solve" the problem. For instance, the integer multiplication problem is identified with the set P_{mult} of all Turing machines that, started with $\overline{m}\#\overline{n}$ on the input tape, eventually halts with the product $\overline{mn}$ on the output tape (where $\overline{n}$ is the binary representation of $n \in \mathbb{N}$). If P is a problem and $A \in P$, we say *A solves P* or *A is an algorithm for P*. A complexity function $f(n)$ is an *upper bound*intrinsic complexity!upper bound on the problem P if there is an algorithm A for P such that $f(n)$ dominates $T_A(n)$. If, for every algorithm A for P, $T_A(n)$ dominates $f(n)$, then we call $f(n)$ a *lower bound*intrinsic complexity!lower bound on the problem P.

Let U_P be the set of upper bounds on P. Notice that there exists a unique complexity function $\ell_P(n)$ such that $\ell_P(n)$ is a lower bound on P and for any other lower bound $f(n)$ on P, $\ell_P(n)$ dominates $f(n)$. To see this, define for each n, $\ell_P(n) := \inf\{f(n) : f \in U_P\}$. On the other hand, there may not exist $T(n)$ in U_P that is dominated by all other functions in U_P; if $T(n)$ exists,

it would (up to co-domination) be equal to $\ell_P(n)$. In this case, we may call $\ell_P(n) = T(n)$ the *intrinsic complexity* $T_P(n)$ *of* P. To resolve the case of the "missing intrinsic complexity", we generalize our concept of a function: An *intrinsic (complexity) function* is *intrinsic (complexity) function* any non-empty family U of complexity functions that is closed under domination, *i.e.*, if $f \in U$ and g dominates f then $g \in U$. The set U_P of upper bounds of P is an intrinsic function: we identify this as the *intrinsic complexity* T_P of P. A subset $V \subseteq U$ is called a *generating set* of U if every $f \in U$ dominates some $g \in V$. We say U is *principal* if U has a generating set consisting of one function f_0; in this case, we call f_0 a *generator* of U. If f is a complexity function, we will identify f with the principal intrinsic function with f as a generator. Note that in non-uniform computational models, the intrinsic complexity of any problem is principal.

Let U, T be intrinsic functions. We extend the standard terminology for ordinary complexity functions to intrinsic functions. Thus

$$U + T, \quad UT, \quad U^T, \quad U \circ T \tag{13}$$

denote intrinsic functions in the natural way. For instance, $U + T$ denotes the intrinsic function generated by the set of functions of the form $u + t$ where $u \in U$ and $t \in T$. We say U *is big-Oh of* T, written

$$U = O(T),$$

if there exists $u \in U$ such that for all $t \in T$, we have $u = O(t)$ in the usual sense. The reader should test these definitions by interpreting $\mathrm{M}_B(n)$, etc, as intrinsic functions (e.g., see (14) in §9). Basically, these definitions allow us to continue to talk about intrinsic functions rather like ordinary complexity functions, provided we know how to interpret them. Similarly, we say U *is big-Omega of* T, written $U = \Omega(T)$, if for all $u \in U$, there exists $t \in T$ such that $u = \Omega(t)$. We say U *is Theta of* T, written $U = \Theta(T)$, if $U = O(T)$ and $U = \Omega(T)$.

Complexity Classes. Corresponding to each computational model, we have complexity classes of problems. Each complexity class is usually characterized by a complexity model (worst-case time, randomized space, etc) and a set of complexity bounds (polynomial, etc). The class of problems that can be solved in polynomial time on a Turing machine is usually denoted P: it is arguably the most important complexity class. This is because we identify this class with the "feasible problems". For instance, the the Fundamental Problem of Algebra (in its various forms) is in P but the Fundamental Problem of Classical Algebraic Geometry is not in P. Complexity theory can be characterized as the study of relationships among complexity classes. Keeping this fact in mind may help motivate much of our activities. Another important class is NC which comprises those problems that can be solved *simultaneously* in depth $\log^{O(1)} n$ and size $n^{O(1)}$, under the Boolean circuit model. Since circuit depth equals parallel time, this is an important class in parallel computation. Although we did not define the circuit analogue of algebraic programs, this is rather straightforward: they are like Boolean circuits except we perform algebraic operations at the nodes. Then we can define NC_A, the algebraic analogue of the class NC. Note that NC_A is defined relative to the underlying algebraic ring.

Exercises

Exercise 7.1: Prove the existence of a problem whose intrinsic complexity is not principal. (In Blum's axiomatic approach to complexity, such problems exist.) □

§8. On Bit versus Algebraic Complexity

We have omitted other important models such as pointer machines that have a minor role in algebraic complexity. But why such a proliferation of models? Researchers use different models depending on the problem at hand. We offer some guidelines for these choices.

1. There is a consensus in complexity theory that the Turing model is the most basic of all general-purpose computational models. To the extent that algebraic complexity seeks to be compatible to the rest of complexity theory, it is preferable to use the Turing model.

2. In practice, the RAM model is invariably used to describe algebraic algorithms because the Turing model is too cumbersome. Upper bounds (*i.e.*, algorithms) are more readily explained in the RAM model and we are happy to take advantage of this in order to make the result more accessible. Sometimes, we could further assert ("left to the reader") that the RAM result extends to the Turing model.

3. Complexity theory proper is regarded to be a theory of "uniform complexity". This means "naturally" uniform models such as Turing machines are preferred over "naturally non-uniform" models such as Boolean circuits. Nevertheless, non-uniform models have the advantage of being combinatorial and conceptually simpler. Historically, this was a key motivation for studying Boolean circuits, since it is hoped that powerful combinatorial arguments may yield super-quadratic lower bounds on the Boolean size of specific problems. Such a result would immediately imply non-linear lower bounds on Turing machine time for the same problem. (Unfortunately, neither kind of result has been realized.) Another advantage of non-uniform models is that the intrinsic complexity of problems is principal. Boolean circuits also seems more natural in the parallel computation domain, with circuit depth corresponding to parallel time.

4. The choice between bit complexity and the algebraic complexity is problem-dependent. For instance, the algebraic complexity of integer GCD would not make much sense (§6, Exercise). But bit complexity is meaningful for any problem (the encoding of the problem must be taken into account). This may suggest that algebraic complexity is a more specialized tool than bit complexity. But even in a situation where bit complexity is of primary interest, it may make sense to investigate the corresponding algebraic complexity. For instance, the algebraic complexity of multiplying integer matrices is $\mathrm{MM}(n) = O(n^{2.376})$ as noted above. Let[3] $\mathrm{MM}(n, N)$ denote the Turing complexity of integer matrix multiplication, where N is an additional bound on the bit size of each entry of the matrix. The best upper bound for $\mathrm{MM}(n, N)$ comes from the trivial remark,

$$\mathrm{MM}(n, N) = O(\mathrm{MM}(n)\mathrm{M}_B(N)). \tag{14}$$

That is, the known upper bound on $\mathrm{MM}(n, N)$ comes from the separate upper bounds on $\mathrm{MM}(n)$ and $\mathrm{M}_B(N)$.

Linear Programming. Equation (14) illustrates a common situation, where the best bit complexity of a problem is obtained as the best algebraic complexity multiplied by the best bit complexity on the underlying operations. We now show an example where this is not the case. Consider the linear programming problem. Let m, n, N be complexity parameters where the linear constraints are represented by $Ax \leq b$, A is an $m \times n$ matrix, and all the numbers in A, b have at most N bits. The linear programming problem can be reduced to checking for the feasibility of the inequality $Ax \leq b$, on input A, b. The Turing complexity $T_B(m, n, N)$ of this problem is known to be polynomial in m, n, N. This result was a breakthrough, due to Khacian in 1979. On the other hand, it is a major open problem whether the corresponding algebraic complexity $T_A(m, n)$ of linear programming is polynomial in m, n.

Euclidean shortest paths. In contrast to linear programming, we now show a problem for which the bit complexity is not known to be polynomial but whose algebraic complexity is polynomial.

[3]The bit complexity bound on any problem is usually formulated to have one more size parameter (N) than the corresponding algebraic complexity bound.

This is the problem of finding the shortest paths between two points on the plane. Let us formulate a version of the *Euclidean shortest path problem*: we are given a planar graph G that is linearly embedded in the plane, *i.e.*, each vertex v of G is mapped to a point $m(v)$ in the plane and each edge (u, v) between two vertices is represented by the corresponding line segment $[m(u), m(v)]$, where two segments may only intersect at their endpoints. We want to find the shortest (under the usual Euclidean metric) path between two specified vertices s, t. Assume that the points $m(v)$ have rational coordinates. Clearly this problem can be solved by Djikstra's algorithm in polynomial time, provided we can (i) take square-roots, (ii) add two sums of square-roots, and (iii) compare two sums of square-roots in constant time. Thus the algebraic complexity is polynomial time (where the basis operations include (i-iii)). However, the current best bound on the bit complexity of this problem is single exponential space. Note that the numbers that arise in this problem are the so-called *constructible reals* (Lecture VI) because they can be finitely constructed by a ruler and a compass.

The lesson of these two examples is that bit complexity and algebraic complexities do not generally have a simple relationship. Indeed, we cannot even expect a polynomial relationship between these two types of complexities: depending on the problem, either one could be exponentially worse than the other.

Exercises

Exercise 8.1*: Obtain an upper bound on the above Euclidean shortest path problem. □

Exercise 8.2: Show that a real number of the form

$$\alpha = n_0 \pm \sqrt{n_1} \pm \sqrt{n_2} \pm \cdots \pm \sqrt{n_k}$$

(where n_i are positive integers) is a zero of a polynomial $P(X)$ of degree at most 2^k, and that all zeros of $P(X)$ are real. □

§9. Miscellany

This section serves as a quick general reference.

Equality symbol. We introduce two new symbols to reduce[4] the semantic overload commonly placed on the equality symbol '='. We use the symbol '$\leftarrow$' for *programming variable assignments* , from right-hand side to the left. Thus, $V \leftarrow V + W$ is an assignment to V (and it could appear on the right-hand side, as in this example). We use the symbol ':=' to denote *definitional equality*, with the term being defined on the left-hand side and the defining terms on the right-hand side. Thus, "$f(n) := n \log n$" is a definition of the function f. Unlike some similar notations in the literature, we refrain from using the mirror images of the definition symbol (we will neither write "$V + W \to V$" nor "$n \log n =: f(n)$").

Sets and functions. The empty set is written $\emptyset$. Let A, B be sets. Subsets and proper subsets are respectively indicated by $A \subseteq B$ and $A \subset B$. Set difference is written $A \setminus B$. Set formation is usually written $\{x : \ldots x \ldots\}$ and sometimes written $\{x | \ldots x \ldots\}$ where $\ldots x \ldots$ specifies some

[4]Perhaps to atone for our introduction of the asymptotic notations.

properties on x. The A is the union of the sets A_i for $i \in I$, we write $A = \cup_{i\in I} A_i$. If the A_i's are pairwise disjoint, we indicate this by writing

$$A = \uplus_{i\in I} A_i.$$

Such a disjoint union is also called a *partition* of A. Sometimes we consider *multisets*. A multiset S can be regarded as sets whose elements can be repeated – the number of times a particular element is repeated is called its *multiplicity*. Alternatively, S can be regarded as a function $S : D \to \mathbb{N}$ where D is an ordinary set and $S(x) \geq 1$ gives the multiplicity of x. We write $f \circ g$ for the composition of functions $g : U \to V$, $f : V \to W$. So $(f \circ g)(x) = f(g(x))$. If a function f is undefined for a certain value x, we write $f(x) \uparrow$.

Numbers. Let $\mathbf{i}$ denote $\sqrt{-1}$, the square-root of -1. For a complex number $z = x + \mathbf{i}y$, let $\mathtt{Re}(z) := x$ and $\mathtt{Im}(z) := y$ denote its real and imaginary part, respectively. Its *modulus* $|z|$ is defined to be the positive square-root of $x^2 + y^2$. If z is real, $|z|$ is also called the *absolute value* . The *(complex) conjugate* of z is defined to be $\overline{z} := \mathtt{Re}(z) - \mathtt{Im}(z)$. Thus $|z|^2 = z\overline{z}$.

But if S is any set, $|S|$ will refer to the *cardinality* , *i.e.*, the number of elements in S. This notation should not cause a confusion with the notion of modulus of z.

For a real number r, we use Iverson's notation (as popularized by Knuth) $\lceil r \rceil$ and $\lfloor r \rfloor$ for the *ceiling* and *floor* functions. We have

$$\lfloor r \rfloor \leq \lceil r \rceil .$$

In this book, we introduce the *symmetric ceiling* and *symmetric floor* functions:

$$\lceil r \rceil_s := \begin{cases} \lceil r \rceil & \text{if } r \geq 0, \\ \lfloor r \rfloor & \text{if } r < 0. \end{cases}$$

$$\lfloor r \rfloor_s := \begin{cases} \lfloor r \rfloor & \text{if } r \geq 0, \\ \lceil r \rceil & \text{if } r < 0. \end{cases}$$

These functions satisfy the following inequalities, valid for all real numbers r:

$$|\, \lfloor r \rfloor_s \,| \leq |r| \leq |\, \lceil r \rceil_s \,|.$$

(The usual floor and ceiling functions fail this inequality when r is negative.) We also use $\lfloor r \rceil$ to denote the *rounding* function, $\lfloor r \rceil := \lceil r - 0.5 \rceil$. So

$$\lfloor r \rfloor \leq \lfloor r \rceil \leq \lceil r \rceil.$$

The base of the *logarithm function* $\log x$, is left unspecified if this is immaterial (as in the notation $O(\log x)$). On the other hand, we shall use

$$\lg x, \qquad \ln x$$

for logarithm to the base 2 and the natural logarithm, respectively.

Let a, b be integers. If $b > 0$, we define the *quotient* and *remainder functions* , $\mathtt{quo}(a, b)$ and $\mathtt{rem}(a, b)$ which satisfy the relation

$$a = \mathtt{quo}(a, b) \cdot b + \mathtt{rem}(a, b)$$

such that $b > \mathtt{rem}(a, b) \geq 0$. We also write these functions using an in-fix notation:

$$(a \,\mathbf{div}\, b) := \mathtt{quo}(a, b); \qquad (a \,\mathbf{mod}\, b) := \mathtt{rem}(a, b).$$

These functions can be generalized to Euclidean domains (lecture II, §2). We continue to use 'mod' in the standard notation "$a \equiv b (\mathrm{mod}\, m)$" for congruence modulo m. We say a *divides* b if $\mathtt{rem}(a, b) = 0$, and denote this by "$a \,|\, b$". If a does not divide b, we denote this by "$a \nmid b$".

Norms. For a complex polynomial $P \in \mathbb{C}[X]$ and for each positive real number k, let $\|P\|_k$ denote[5] the *k-norm* ,

$$\|P\|_k := \left(\sum_{i=0}^{n} |p_i|^k\right)^{1/k}$$

where $p_0, \ldots, p_n$ are the coefficients of P. We extend this definition to $k = \infty$, where

$$\|P\|_\infty := \max\{|p_i| : i = 0, \ldots, n\}. \tag{15}$$

There is a related L_k-norm defined on P where we view P as a complex function (in contrast to L_k-norms, it is usual to refer to our k-norms as "ℓ_k-norms"). The L_k-norms are less important for us. Depending on context, we may prefer to use a particular k-norm: in such cases, we may simply write "$\|P\|$" instead of "$\|P\|_k$". For $0 < r < s$, we have

$$\|P\|_\infty \le \|P\|_s < \|P\|_r \le (n+1)\|P\|_\infty \tag{16}$$

The second inequality (called Jensen's inequality) follows from:

$$\begin{aligned}\frac{(\sum_i |p_i|^s)^{1/s}}{(\sum_j |p_j|^r)^{1/r}} &= \left\{\sum_{i=0}^{n} \frac{|p_i|^s}{(\sum_j |p_j|^r)^{s/r}}\right\}^{\frac{1}{s}} = \left\{\sum_{i=0}^{n} \left(\frac{|p_i|^r}{\sum_j |p_j|^r}\right)^{\frac{s}{r}}\right\}^{\frac{1}{s}} \\ &< \left\{\sum_{i=0}^{n} \left(\frac{|p_i|^r}{\sum_j |p_j|^r}\right)\right\}^{\frac{1}{r}} = 1.\end{aligned}$$

The 1-, 2- and ∞-norms of P are also known as the *weight*, *length*, and *height* of P. If $\mathbf{u}$ is a vector of numbers, we define its *k-norm* $\|\mathbf{u}\|_k$ by viewing $\mathbf{u}$ as the coefficient vector of a polynomial. The following inequality will be useful:

$$\|P\|_1 \le \sqrt{n}\|P\|_2.$$

To see this, note that $n\sum_{i=1}^n a_i^2 \ge (\sum_{i=1}^n a_i)^2$ is equivalent to $(n-1)\sum_{i=1}^n a_i^2 \ge 2\sum_{1\le i<j\le n} a_i a_j$. But this amounts to $\sum_{1\le i<j\le n}(a_i - a_j)^2 \ge 0$.

Inequalities. Let $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$ be real n-vectors. We write $\mathbf{a}\cdot\mathbf{b}$ or $\langle \mathbf{a}, \mathbf{b}\rangle$ for their scalar product $\sum_{i=1}^n a_i b_i$.

Hölder's Inequality: If $\frac{1}{p} + \frac{1}{q} = 1$ then

$$|\langle \mathbf{a}, \mathbf{b}\rangle| \le \|\mathbf{a}\|_p \|\mathbf{b}\|_q,$$

with equality iff there is some k such that $b_i^q = k a_i^p$ for all i. In particular, we have the Cauchy-Schwarz Inequality:

$$|\langle \mathbf{a}, \mathbf{b}\rangle| \le \|\mathbf{a}\|_2 \cdot \|\mathbf{b}\|_2.$$

Minkowski's Inequality: for $k > 1$,

$$\|\mathbf{a} + \mathbf{b}\|_k \le \|\mathbf{a}\|_k + \|\mathbf{b}\|_k.$$

This shows that the k-norms satisfy the triangular inequality.

A real function $f(x)$ defined on an interval $I = [a, b]$ is *convex* on I if for all $x, y \in I$ and $0 \le \alpha \le 1$, $f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y)$. For instance, if $f''(x)$ is defined and $f''(x) \ge 0$ on I implies f is convex on I.

[5] In general, a *norm* of a real vector V is a real function $N : V \to \mathbb{R}$ such that for all $x \in V$, (i) $N(x) \ge 0$ with equality iff $x = \mathbf{0}$, (ii) $N(cx) = |c|N(x)$ for any $c \in \mathbb{R}$, and (iii) $N(x+y) \le N(x) + N(y)$. The k-norms may be verified to be a norm in this sense.

Polynomials. Let $A(X) = \sum_{i=0}^{n} a_i X^i$ be a univariate polynomial. Besides the notation $\deg(A)$ and $\texttt{lead}(A)$ of §1, we are sometimes interested in the largest power $j \geq 0$ such that X^j divides $A(X)$; this j is called the *tail degree* of A. The coefficient a_j is the *tail coefficient* of A, denoted $\texttt{tail}(A)$.

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be $n \geq 1$ (commutative) variables, and consider multivariate polynomials in $R[\mathbf{X}]$. A *power product* over $\mathbf{X}$ is a polynomial of the form $T = \prod_{i=1}^{n} X_i^{e_i}$ where each $e_i \geq 0$ is an integer. In particular, if all the e_i's are 0, then $T = 1$. The *total degree* $\deg(T)$ of T is given by $\sum_{i=1}^{n} e_i$, and the *maximum degree* $\texttt{mdeg}(T)$ is given by $\max_{i=1}^{n} e_i$. Usually, we simply say "degree" for total degree. Let $\texttt{PP}(\mathbf{X}) = \texttt{PP}(X_1, \ldots, X_n)$ denote the set of power products over $\mathbf{X}$.

A *monomial* or *term* is a polynomial of the form cT where T is a power product and $c \in R \setminus \{0\}$. So a polynomial A can be written uniquely as a sum $A = \sum_{i=1}^{k} A_i$ of monomials with distinct power products; each such monomial A_i is said to *belong* to A. The *(term) length* of a polynomial A to be the number of monomials in A, not to be confused with its Euclidean length $\|A\|_2$ defined earlier. The total degree $\deg(A)$ (respectively, maximum degree $\texttt{mdeg}(A)$) of a polynomial A is the largest total (respectively, maximum) degree of a power product in A. Usually, we just say "degree" of A to mean total degree. A polynomial is *homogeneous* if each of its monomials has the same total degree. Again, any polynomial A can be written uniquely as a sum $A = \sum_i H_i$ of homogeneous polynomials H_i of distinct degrees; each H_i is said to be a *homogeneous component* of A.

The degree concepts above can be generalized. If $\mathbf{X}_1 \subseteq \mathbf{X}$ is a set of variables, we may speak of the "$\mathbf{X}_1$-degree" of a polynomial A, or say that a polynomial "homogeneous" in $\mathbf{X}_1$, simply by viewing A as a polynimial in $\mathbf{X}_1$. Or again, if $\mathbf{Y} = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$ is a partition of the variables $\mathbf{X}$, the "$\mathbf{Y}$-maximum degree" of A is the maximum of the $\mathbf{X}_i$-degrees of A $(i = 1, \ldots, k)$.

Matrices. The set of $m \times n$ matrices with entries over a ring R is denoted $R^{m \times n}$. Let $M \in R^{m \times n}$. If the (i,j)th entry of M is x_{ij}, we may write $M = [x_{ij}]_{i,j=1}^{m,n}$ (or simply, $M = [x_{ij}]_{i,j}$). The (i,j)th entry of M is denoted $M(i;j)$. More generally, if $i_1, i_2, \ldots, i_k$ are indices of rows and $j_1, \ldots, j_\ell$ are indices of columns,

$$M(i_1, \ldots, i_k; j_1, \ldots, j_\ell) \tag{17}$$

denotes the *submatrix* obtained by intersecting the indicated rows and columns. In case $k = \ell = 1$, we often prefer to write $(M)_{i,j}$ or $(M)_{ij}$ instead of $M(i;j)$. If we delete the ith row and jth column of M, the resulting matrix is denoted $M[i;j]$. Again, this notation can be generalized to deleting more rows and columns. E.g., $M[i_1, i_2; j_1, j_2, j_3]$ or $[M]_{i_1,i_2;j_1,j_2,j_3}$. The *transpose* of M is the $n \times m$ matrix, denoted M^T, such that $M^T(i;j) = M(j;i)$.

A *minor* of M is the determinant of a square submatrix of M. The submatrix in (17) is *principal* if $k = \ell$ and

$$i_1 = j_1 < i_2 = j_2 < \cdots < i_k = j_k.$$

A minor is *principal* if it is the determinant of a principal submatrix. If the submatrix in (17) is principal with $i_1 = 1, i_2 = 2, \ldots, i_k = k$, then it is called the "kth principal submatrix" and its determinant is the "kth principal minor". (Note: the literature sometimes use the term "minor" to refer to a principal submatrix.)

Ideals. Let R be a ring and I, J be ideals of R. The ideal *generated* by elements $a_1, \ldots, a_m \in R$ is denoted $(a_1, \ldots, a_m)$ and is defined to be the smallest ideal of R containing these elements. Since

this well-known notation for ideals may be ambiguous, we sometimes write[6]

$$\texttt{Ideal}(a_1, \ldots, a_m).$$

Another source of ambiguity is the underlying ring R that generates the ideal; thus we may sometimes write

$$(a_1, \ldots, a_m)_R \quad \text{or} \quad \texttt{Ideal}_R(a_1, \ldots, a_m).$$

An ideal I is *principal* if it is generated by one element, $I = (a)$ for some $a \in R$; it is *finitely generated* if it is generated by some finite set of elements. For instance, the *zero ideal* is $(0) = \{0\}$ and the *unit ideal* is $(1) = R$. Writing $aR := \{ax : x \in R\}$, we have that $(a) = aR$, exploiting the presence of $1 \in R$. A *principal ideal ring* or *domain* is one in which every ideal is principal. An ideal is called *homogeneous* (resp., *monomial*) if it is generated by a set of homogeneous polynomials (resp., monomials).

The following are five basic operations defined on ideals:

Sum: $I + J$ is the ideal consisting of all $a + b$ where $a \in I, b \in J$.

Product: IJ is the ideal generated by all elements of the form ab where $a \in I, b \in J$.

Intersection: $I \cap J$ is just the set theoretic intersection of I and J.

Quotient: $I : J$ is defined to be the set $\{a | aJ \subseteq I\}$. If $J = (a)$, we simply write $I : a$ for $I : J$.

Radical: $\sqrt{I}$ is defined to be set $\{a | (\exists n \geq 1) a^n \in I\}$.

Some simple relationships include $IJ \subseteq I \cap J$, $I(J + J') = IJ + IJ'$, $(a_1, \ldots, a_m) + (b_1, \ldots, b_n) = (a_1, \ldots, a_m, b_1, \ldots, b_n)$. An element b is *nilpotent* if some power of b vanishes, $b^n = 0$. Thus $\sqrt{(0)}$ is the set of nilpotent elements. An ideal I is *maximal* if $I \neq R$ and it is not properly contained in an ideal $J \neq R$. An ideal I is *prime* if $ab \in I$ implies $a \in I$ or $b \in I$. An ideal I is *primary* if $ab \in I, a \notin I$ implies $b^n \in I$ for some positive integer n. A ring with unity is *Noetherian* if every ideal I is finitely generated. It turns out that for Noetherian rings, the basic building blocks are primary ideals (not prime ideals). We assume the reader is familiar with the construction of ideal quotient rings, R/I.

Exercises

Exercise 9.1: (i) Verify the rest of equation (16).
(ii) $\|A \pm B\|_1 \leq \|A\|_1 + \|B\|_1$ and $\|AB\|_1 \leq \|A\|_1 \|B\|_1$.
(iii) (Duncan) $\|A\|_2 \|B\|_2 \leq \|AB\|_2 \sqrt{\binom{2n}{n}\binom{2m}{m}}$ where $\deg(A) = m$, $\deg(B) = n$. □

Exercise 9.2: Show the inequalities of Hölder and Minkowski. □

Exercise 9.3: Let $I \neq R$ be an ideal in a ring R with unity.
a) I is maximal iff R/I is a field.
b) I is prime iff R/I is a domain.
c) I is primary iff every zero-divisor in R/I is nilpotent. □

[6] Cf. the notation $\textsc{Ideal}(U) \subseteq R_0[X_1, \ldots, X_d]$ where $U \in \mathbb{A}^d(R_1)$, introduced in §4. We capitalize the names of maps from an algebraic to a geometric setting or vice-versa. Thus Ideal, Zero.

§10. Computer Algebra Systems

In a book on algorithmic algebra, we would be remiss if we make no mention of *computer algebra systems*. These are computer programs that manipulate and compute on symbolic ("algebraic") quantities as opposed to just numerical ones. Indeed, there is an intimate connection between algorithmic algebra today and the construction of such programs. Such programs range from general purpose systems (e.g., `Maple`, `Mathematica`, `Reduce`, `Scratchpad`, `Macsyma`, etc.) to those that target specific domains (e.g., `Macaulay` (for Gröbner bases), `MatLab` (for numerical matrices), `Cayley` (for groups), `SAC-2` (polynomial algebra), `CM` (celestial mechanics), `QES` (quantum electrodynamics), etc.). It was estimated that about 60 systems exist around 1980 (see [23]). A computer algebra book that discuss systems issues is [8]. In this book, we choose to focus on the mathematical and algorithmic development, *independent of any computer algebra system.* Although it is possible to avoid using a computer algebra system in studying this book, we strongly suggest that the student learn at least one general-purpose computer algebra system and use it to work out examples. If any of our exercises make system-dependent assumptions, it may be assumed that `Maple` is meant.

Exercises

Exercise 10.1: It took J. Bernoulli (1654-1705) less than 1/8 of an hour to compute the sum of the 10th power of the first 1000 numbers: $91, 409, 924, 241, 424, 243, 424, 241, 924, 242, 500$.

(i) Write a procedure `bern(n,e)` in your favorite computer algebra system, so that the above number is computed by calling `bern(1000, 10)`.

(ii) Write a procedure `berns(m,n,e)` that runs `bern(n,e)` m times. Do simple profiling of the functions `bern, berns`, by calling `berns(100, 1000, 10)`. □

References

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[2] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[3] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[4] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[5] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[6] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[7] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[8] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[9] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[10] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[11] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[12] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[13] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[14] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[15] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[16] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[17] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[18] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[19] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[20] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[21] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[22] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[23] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[24] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[25] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[26] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[27] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[28] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

Contents

Lecture I
ARITHMETIC

This lecture considers the *arithmetic operations* (addition, subtraction, multiplication and division) in three basic algebraic structures: polynomials, integers, matrices. These operations are the basic building blocks for other algebraic operations, and hence are absolutely fundamental in algorithmic algebra. Strictly speaking, division is only defined in a field. But there are natural substitutes in general rings: it could be always be replaced by the *divisibility predicate*. In a domain, we can define *exact division*. The the exact division of u by v is defined iff the v divides u; when defined, the result is the unique w such that $vw = u$. In case of Euclidean rings (Lecture II), division could be replaced by the quotient and remainder functions.

Complexity of Multiplication. In most algebraic structures of interest, the obvious algorithms for addition and subtraction take linear time and are easily seen to be optimal. Since we are mainly concerned with asymptotic complexity here, there is nothing more to say about them. As for the division-substitutes, they turn out to be reducible to multiplication. Hence the term "complexity of multiplication" can be regarded a generic term to cover such operations as well. After such considerations, what remains to be addressed is multiplication itself. The pervading influence of Schönhage and Strassen in all these results cannot be overstated.

We use some other algebraic structures in addition to the ones introduced in Lecture 0, §1:

$$\begin{array}{lcl} GF(p^m) & = & \text{Galois field of order } p^m, \ p \text{ prime}, \\ \mathbb{Z}_n & = & \text{integers modulo } n \geq 1, \\ \mathrm{M}_{m,n}(R) & = & m \text{ by } n \text{ matrices over a ring } R, \\ \mathrm{M}_n(R) & = & \mathrm{M}_{n,n}(R). \end{array}$$

Finite structures such as $GF(p^m)$ and $\mathbb{Z}_n$ have independent interest, but they also turn out to be important for algorithms in infinite structures such as $\mathbb{Z}$.

§1. The Discrete Fourier Transform

The key to fast multiplication of integers and polynomials is the discrete Fourier transform.

Roots of unity. In this section, we work with complex numbers. A complex number $\alpha \in \mathbb{C}$ is an nth *root of unity* if $\alpha^n = 1$. It is a *primitive nth root of unity* if, in addition, $\alpha^m \neq 1$ for all $m = 1, \ldots, n-1$. In particular,

$$e^{\frac{2\pi}{n}\mathbf{i}} = \cos\frac{2\pi}{n} + \mathbf{i}\sin\frac{2\pi}{n}$$

($\mathbf{i} = \sqrt{-1}$) is a primitive nth root of unity. There are exactly $\varphi(n)$ primitive nth roots of unity where $\varphi(n)$ is the number of positive integers less than or equal to n that are relatively prime to n. Thus $\varphi(n) = 1, 1, 2, 2, 4, 2, 6$ for $n = 1, 2, \ldots, 7$; $\varphi(n)$ is also known as *Euler's phi-function or totient function*.

Example: A primitive 8th root of unity is $\omega = e^{\frac{2\pi}{8}\mathbf{i}} = \frac{1}{\sqrt{2}} + \mathbf{i}\frac{1}{\sqrt{2}}$. It is easy to check the only other primitive roots are ω^3, ω^5 and ω^7 (so $\varphi(8) = 4$). These roots are easily visualized in the complex plane (see figure 1).

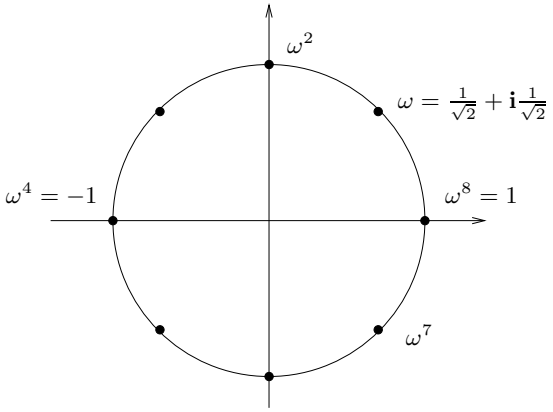


Figure 1: The 8th roots of unity.

Let ω denote any primitive nth root of unity. We note a basic identity.

Lemma 1 (Cancellation Property)

$$\sum_{j=0}^{n-1} \omega^{js} = \begin{cases} 0 & \text{if } s \not\equiv 0 \bmod n \\ n & \text{if } s \equiv 0 \bmod n \end{cases}$$

Proof. The result is clear if $s \equiv 0 \bmod n$. Otherwise, consider the identity $x^n - 1 = (x-1)(\sum_{j=0}^{n-1} x^j)$. Substituting $x = \omega^s$ makes the left-hand side equal to zero. The right-hand side becomes $(\omega^s - 1)(\sum_{j=0}^{n-1} \omega^{js})$. Since $\omega^s \neq 1$ for $s \not\equiv 0 \bmod n$, the result follows. **Q.E.D.**

Let $F(\omega) = F_n(\omega)$ denote the matrix

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ \vdots & & & & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)^2} \end{bmatrix}.$$

Definition 1 (The DFT and its inverse) *Let* $\mathbf{a} = (a_0, \ldots, a_{n-1})^T \in \mathbb{C}^n$. *The* discrete Fourier transform *(abbr. DFT) of* $\mathbf{a}$ *is* $\mathrm{DFT}_n(\mathbf{a}) := \mathbf{A} = (A_0, \ldots, A_{n-1})^T$ *where* $A_i = \sum_{j=0}^{n-1} a_j \omega^{ij}$, *for* $i = 0, \ldots, n-1$. *That is,*

$$\mathrm{DFT}_n(\mathbf{a}) = F(\omega) \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{n-1} \end{bmatrix}.$$

The inverse discrete Fourier transform *of* $\mathbf{A} = (A_0, \ldots, A_{n-1})^T$ *is* $\mathrm{DFT}_n^{-1}(\mathbf{A}) = \frac{1}{n} F(\omega^{-1}) \cdot \mathbf{A}$. *That*

is,

$$\mathrm{DFT}_n^{-1}(\mathbf{A}) := \frac{1}{n}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega^{-1} & \omega^{-2} & \cdots & \omega^{-n+1} \\ \vdots & & & & \vdots \\ 1 & \omega^{-n+1} & \omega^{-2(n-1)} & \cdots & \omega^{-(n-1)^2} \end{bmatrix} \cdot \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{n-1} \end{bmatrix}.$$

Note that $\omega^{-1} = \omega^{n-1}$. We will omit the subscript 'n' in DFT_n when convenient. The following shows that the two transforms are indeed inverses of each other:

Lemma 2 *We have $F(\omega^{-1}) \cdot F(\omega) = F(\omega) \cdot F(\omega^{-1}) = nI_n$ where I_n is the identity matrix.*

Proof. Let $F(\omega^{-1}) \cdot F(\omega) = [c_{j,k}]_{j,k=0}^{n-1}$ where

$$c_{j,k} = \sum_{i=0}^{n-1} \omega^{-ji}\omega^{ik} = \sum_{i=0}^{n-1} \omega^{i(k-j)}.$$

If $j = k$, then $c_{j,k} = \sum_{i=0}^{n-1} \omega^0 = n$. Otherwise, $-n < k - j < n$ and $k - j \neq 0$ implies $c_{j,k} = 0$, using lemma 1. Similarly, $F(\omega) \cdot F(\omega^{-1}) = nI_n$. **Q.E.D.**

Connection to polynomial evaluation and interpolation. Let $\mathbf{a}$ be the coefficient vector of the polynomial $P(X) = \sum_{i=0}^{n-1} a_i X^i$. Then computing $\mathrm{DFT}(\mathbf{a})$ amounts to *evaluating* the polynomial $P(X)$ at all the nth roots of unity, at

$$X = 1, X = \omega, X = \omega^2, \ldots, X = \omega^{n-1}.$$

Similarly, computing $\mathrm{DFT}^{-1}(\mathbf{A})$ amounts to recovering the polynomial $P(X)$ from its values $(A_0, \ldots, A_{n-1})$ at the same n points. In other words, the inverse discrete Fourier transform *interpolates*, or reconstructs, the polynomial $P(X)$ from its values at all the n roots of unity. Here we use the fact (Lecture IV.1) that the interpolation of a degree $n-1$ polynomial from its values at n distinct points is unique. (Of course, we could also have viewed DFT as interpolation and DFT^{-1} as evaluation.)

The Fast Fourier Transform. A naive algorithm to compute DFT and DFT^{-1} would take $\Theta(n^2)$ complex arithmetic operations. In 1965, Cooley and Tukey [47] discovered a method that takes $O(n \log n)$ operations. This has come to be known as the *fast Fourier transform* (FFT). This algorithm is widely used. The basic ideas of the FFT were known prior to 1965. E.g., Runge and König, 1924 (see [105, p. 642]).

Let us now present the FFT algorithm to compute $\mathrm{DFT}(\mathbf{a})$ where $\mathbf{a} = (a_0, \ldots, a_{n-1})$. In fact, it is a fairly straightforward divide-and-conquer algorithm. To simplify discussion, let n be a power of 2. Instead of $\mathbf{a}$, it is convenient to be able to interchangeably talk of the polynomial $P(X)$ whose coefficient vector is $\mathbf{a}$. As noted, computing $\mathrm{DFT}(\mathbf{a})$ amounts to computing the n values

$$P(1), P(\omega), P(\omega^2), \ldots, P(\omega^{n-1}). \tag{1}$$

First, let us express $P(X)$ as the sum of its odd part and its even part:

$$P(X) = P_e(X^2) + X \cdot P_o(X^2)$$

where $P_e(Y), P_o(Y)$ are polynomials of degrees at most $\frac{n}{2}$ and $\frac{n-1}{2}$, respectively. E.g., for $P(X) = 3X^6 - X^4 + 2X^3 + 5X - 1$, we have $P_e(Y) = 3Y^3 - Y^2 - 1$, $P_o(Y) = 2Y + 5$. Thus we have reduced the problem of computing the values in (1) to the following:

FFT Algorithm:
Input: a polynomial $P(X)$ with coefficients given by an n-vector $\mathbf{a}$, and ω, a primitive nth root of unity.
Output: $\mathrm{DFT}_n(\mathbf{a})$.
1. Evaluate $P_e(X^2)$ and $P_o(X^2)$ at $X^2 = 1, \omega^2, \omega^4, \ldots, \omega^n, \omega^{n+2}, \ldots, \omega^{2n-2}$.
2. Multiply $P_o(\omega^{2j})$ by ω^j for $j = 0, \ldots, n-1$.
3. Add $P_e(\omega^{2j})$ to $\omega^j P_o(\omega^{2j})$, for $j = 0, \ldots, n-1$.

Analysis. Note that in step 1, we have $\omega^n = 1$, $\omega^{n+2} = \omega^2, \ldots, \omega^{2n-2} = \omega^{n-2}$. So it suffices to evaluate P_e and P_o at only $n/2$ values, $X = 1, \omega^2, \ldots, \omega^{n-2}$, *i.e.*, at all the $(n/2)$th roots of unity. But this is equivalent to the problem of computing $\mathrm{DFT}_{n/2}(P_e)$ and $\mathrm{DFT}_{n/2}(P_o)$. Hence we view step 1 as two recursive calls. Steps 2 and 3 take n multiplications and n additions respectively. Overall, if $T(n)$ is the number of complex additions and multiplications, we have

$$T(n) = 2T(n/2) + 2n$$

which has the exact solution $T(n) = 2n \log n$ for n a power of 2.

Since the same method can be applied to the inverse discrete Fourier transform, we have shown:

Theorem 3 (Complexity of FFT) *Assuming the availability of a primitive nth root of unity, the discrete Fourier transform* DFT_n *and its inverse can be computed in $O(n \log n)$ complex arithmetic operations.*

Note that this is a result in the algebraic program model of complexity (§0.6). This could be translated into a result about bit complexity (Turing machines or Boolean Circuits) if we make assumptions about how the complex numbers are encoded in the input. However, this exercise would not be very illuminating, and we await a "true" bit complexity result below in §3.

Remark: There are several closely related fast transform methods which have the same framework. For example, [66].

Exercises

Exercise 1.1: Show that the number of multiplications in step 2 can be reduced to $n/2$. HINT: Then half of the additions in step 3 become subtractions. □

§2. Polynomial Multiplication

We consider the multiplication of complex polynomials. To exploit the FFT algorithm, we make a fundamental connection.

Convolution and polynomial multiplication. Assume $n \geq 2$. The *convolution* of two n-vectors $\mathbf{a} = (a_0, \ldots, a_{n-1})^T$ and $\mathbf{b} = (b_0, \ldots, b_{n-1})^T$ is the n-vector

$$\mathbf{c} = \mathbf{a} * \mathbf{b} := (c_0, \ldots, c_{n-1})^T$$

where $c_i = \sum_{j=0}^{i} a_j b_{i-j}$. Let $P(X)$ and $Q(X)$ be polynomials of degrees less than $n/2$. Then $R(X) := P(X)Q(X)$ is a polynomial of degree less than $n-1$. Let $\mathbf{a}$ and $\mathbf{b}$ denote the coefficient vectors of P and Q (padded out with initial zeros to make vectors of length n). Then it is not hard to see that $\mathbf{a} * \mathbf{b}$ gives the coefficient vector of $R(X)$. Thus *convolution is essentially polynomial multiplication*. The following result relates convolution to the usual scalar product, $\mathbf{a} \cdot \mathbf{b}$.

Theorem 4 (Convolution Theorem) *Let $\mathbf{a}, \mathbf{b}$ be n-vectors whose initial $\lfloor n/2 \rfloor$ entries are zeros. Then*

$$\mathrm{DFT}^{-1}(\mathrm{DFT}(\mathbf{a}) \cdot \mathrm{DFT}(\mathbf{b})) = \mathbf{a} * \mathbf{b}. \tag{2}$$

Proof. Suppose $\mathrm{DFT}(\mathbf{a}) = (A_0, \ldots, A_{n-1})^T$ and $\mathrm{DFT}(\mathbf{b}) = (B_0, \ldots, B_{n-1})^T$. Let $\mathbf{C} = (C_0, \ldots, C_{n-1})^T$ where $C_i = A_i B_i$. From the evaluation interpretation of DFT, it follows that C_i is the value of the polynomial $R(X) = P(X)Q(X)$ at $X = \omega^i$. Note that $\deg(R) \leq n-1$. Now, evaluating a polynomial of degree $\leq n-1$ at n distinct points is the inverse of interpolating such a polynomial from its values at these n points (see §IV.1). Since DFT^{-1} and DFT are inverses, we conclude that $\mathrm{DFT}^{-1}(\mathbf{C})$ is the coefficient vector of $R(X)$. We have thus given an interpretion for the left-hand side of (2). But the right-hand side of (2) is also equal to the coefficient vector of $R(X)$, by the polynomial multiplication interpretation of convolution. **Q.E.D.**

This theorem reduces the problem of convolution (equivalently, polynomial multiplication) to two DFT and one DFT^{-1} computations. We immediately conclude from the FFT result (Theorem 3):

Theorem 5 (Algebraic complexity of polynomial multiplication) *Assuming the availability of a primitive nth root of unity, we can compute the product PQ of two polynomials $P, Q \in \mathbb{C}[X]$ of degrees less than n in $O(n \log n)$ complex operations.*

Remark: If the coefficients of our polynomials are not complex numbers but in some other ring, then a similar result holds provided the ring contains an analogue to the roots of unity. Such a situation arises in our next section.

Exercises

Exercise 2.1: Show that polynomial quotient $P \,\mathbf{div}\, Q$ and remainder $P \,\mathbf{mod}\, Q$ can be computed in $O(n \log n)$ complex operations. □

Exercise 2.2: Let $q = p^m$ where $p \in \mathbb{N}$ is prime, $m \geq 1$. Show that in $GF(q)$, we can multiply in $O(m\mathcal{L}(m))$ operations of $\mathbb{Z}_p$ and can compute inverses in $O(m\mathcal{L}^2(m))$ operations. HINT: use the fact that $GF(q)$ is isomorphic to $GF(p)[X]/(F(X))$ where $F(X)$ is any polynomial of degree m that is irreducible over $GF(p)$. □

Exercise 2.3: Let $q = p^m$ as above. Show how to multiply two degree n polynomials over $GF(q)$ in $O(n\mathcal{L}^2(n))$ operations of $GF(q)$. and compute the GCD of two such polynomials in $O(n\mathcal{L}^2(n))$ operations of $GF(q)$. □

§3. Modular FFT

To extend the FFT technique to integer multiplication, a major problem to overcome is how one replaces the complex roots of unity with some discrete analogue. One possibility is to carry out the complex arithmetic to a suitable degree of accuracy. This was done by Strassen in 1968, achieving a time bound that satisfies the recurrence $T(n) = O(nT(\log n))$. For instance, this implies $T(n) = O(n \log n (\log\log n)^{1+\epsilon})$ for any $\epsilon > 0$. In 1971, Schönhage and Strassen managed to improved this to $T(n) = O(n \log n \log\log n)$. While the complexity improvement can be said to be strictly of theoretical interest, their use of modular arithmetic to avoid approximate arithmetic has great interest. They discovered that the discrete Fourier transform can be defined, and the FFT efficiently implemented, in $\mathbb{Z}_M$ where

$$M = 2^L + 1, \tag{3}$$

for suitable values of L. This section describes these elegant techniques.

First, we make some general remarks about $\mathbb{Z}_M$ for an arbitrary modulus $M > 1$. An element $x \in \mathbb{Z}_M$ is a *zero-divisor*ring!zero-divisor if there exists $y \neq 0$ such that $x \cdot y = 0$; a *(multiplicative) inverse*ring!inverse element of x is y such that $xy = 1$. For example, in $\mathbb{Z}_4$, the element 2 has no inverse and $2 \cdot 2 = 0$.

> Claim: an element $x \in \mathbb{Z}_M$ has a multiplicative inverse (denoted x^{-1}) if and only if x is not a zero-divisor.

To see this claim, suppose x^{-1} exists and $x \cdot y = 0$. Then $y = 1 \cdot y = x^{-1}x \cdot y = 0$. Conversely, if x is not a zero-divisor then the elements in the set $\{x \cdot y : y \in \mathbb{Z}_M\}$ are all distinct because if $x \cdot y = x \cdot y'$ then $x(y - y') = 0$ and $y - y' \neq 0$, contradiction. Hence, by pigeon-hole principle, 1 occurs in the set. This proves our claim. We have two basic consequences: (i) If x has an inverse, the inverse is unique. [In proof, if $x \cdot y = 1 = x \cdot y'$ then $x(y - y') = 0$ and so $y = y'$.] (ii) $\mathbb{Z}_M$ is a field iff M is prime. [In proof, if M has the proper factorization xy then x is a zero-divisor. Conversely, if M is prime then every $x \in \mathbb{Z}_M$ has an inverse because the extended Euclidean algorithm (Lecture II§2) implies there exist $s, t \in \mathbb{Z}_M$ such that $sx + tM = 1$, *i.e.*, $s = x^{-1} (\mathrm{mod}\, M)$.]

In the rest of this section and also the next one, we assume M has the form in Equation (3). Then $2^L \equiv -1 (\mathrm{mod}\, M)$ and $2^{2L} = (M-1)^2 \equiv 1 (\mathrm{mod}\, M)$. We also use the fact that every element of the form 2^i $(i \geq 0)$ has an inverse in $\mathbb{Z}_M$, *viz.*, 2^{2L-i}.

Representation and basic operations modulo M. We clarify how numbers in $\mathbb{Z}_M$ are represented. Let $2^L \equiv -1 (\mathrm{mod}\, M)$ be denoted with the special symbol $\overline{1}$. We represent each element of $\mathbb{Z}_M \setminus \{\overline{1}\}$ in the expected way, as a binary string $(b_{L-1}, \ldots, b_0)$ of length L; the element $\overline{1}$ is given a special representation. For example, with $M = 17, L = 4$ then 13 is represented by $(1, 1, 0, 1)$, or simply written as (1101). It is relatively easy to add and subtract in $\mathbb{Z}_M$ under this represention using a linear number of bit operations, *i.e.*, $O(L)$ time. Of course, special considerations apply to $\overline{1}$.

Exercise 3.1: Show that addition and subtraction take $O(L)$ bit operations. □

We will also need to multiply by powers of 2 in linear time. Intuitively, multiplying a number X by 2^j amounts to left-shifting the string X by j positions; a slight complication arises when we get a carry to the left of the most significant bit.

Example: Consider multiplying $13 = (1101)$ by $2 = (0010)$ in $\mathbb{Z}_{17}$. Left-shifting (1101) by 1 position gives (1010), with a carry. This carry represents $16 \equiv -1 = \overline{1}$. So to get the final result, we must add $\overline{1}$ (equivalently, subtract 1) from (1010), yielding (1001). [Check: $13 \times 2 \equiv 9 (\mathrm{mod}\, 17)$ and $9 = (1001)$.]

In general, if the number represented by the string $(b_{L-1}, \ldots, b_0)$ is multiplied by 2^j $(0 < j < L)$, the result is given as a difference:

$$(b_{L-j-1}, b_{L-j-2}, \ldots, b_0, 0, \ldots, 0) - (0, \ldots, 0, b_{L-1}, b_{L-2}, \ldots, b_{L-j}).$$

But we said that subtraction can be done in linear time. So we conclude: *in $\mathbb{Z}_M$, multiplication by 2^j takes $O(L)$ bit operations.*

Primitive roots of unity modulo M. Let $K = 2^k$ and K divides L. We define

$$\omega := 2^{L/K}.$$

For instance, in $\mathbb{Z}_{17}$, and with $K = 2$, we get $\omega^i = 4, 16, 13, 1$ for $i = 1, 2, 3, 4$. So ω is a primitive 4th root of unity.

Lemma 6 *In $\mathbb{Z}_M$, ω is a primitive $(2K)$th root of unity.*

Proof. Note that $\omega^K = 2^L \equiv -1 (\mathrm{mod}\, M)$. Thus $\omega^{2K} \equiv 1 (\mathrm{mod}\, M)$, *i.e.*, it is a $(2K)$th root of unity. To show that it is in fact a primitive root, we must show $\omega^j \not\equiv 1$ for $j = 1, \ldots, (2K-1)$. If $j \leq K$ then $\omega^j = 2^{Lj/K} \leq 2^L < M$ so clearly $\omega^j \not\equiv 1$. If $j > K$ then $\omega^j = -\omega^{j-K}$ where $j - K \in \{1, \ldots, K-1\}$. Again, $\omega^{j-K} < 2^L \equiv -1$ and so $-\omega^{j-K} \not\equiv 1$. **Q.E.D.**

We next need the equivalent of the cancellation property (Lemma 1). The original proof is invalid since $\mathbb{Z}_M$ is not necessarily an integral domain (see remarks at the end of this section).

Lemma 7 *The cancellation property holds:*

$$\sum_{j=0}^{2K-1} \omega^{js} \equiv \begin{cases} 0 (\mathrm{mod}\, M) & \text{if } s \not\equiv 0 \bmod 2K, \\ 2K (\mathrm{mod}\, M) & \text{if } s \equiv 0 \bmod 2K. \end{cases}$$

Proof. The result is true if $s \equiv 0 \bmod 2K$. Assuming otherwise, let $(s \bmod 2K) = 2^p q$ where q is odd, $0 < 2^p < 2K$ and let $r = 2K \cdot 2^{-p} > 1$. Then by breaking up the desired sum into 2^p parts,

$$\begin{aligned} \sum_{j=0}^{2K-1} \omega^{js} &= \sum_{j=0}^{r-1} \omega^{js} + \sum_{j=r}^{2r-1} \omega^{js} + \cdots + \sum_{j=2K-r}^{2K-1} \omega^{js} \\ &= \sum_{j=0}^{r-1} \omega^{js} + \omega^{rs} \sum_{j=0}^{r-1} \omega^{js} + \cdots + \omega^{rs(2^p-1)} \sum_{j=0}^{r-1} \omega^{js} \\ &\equiv 2^p \sum_{j=0}^{r-1} \omega^{js}, \end{aligned}$$

since $\omega^{rs} \equiv 1 \bmod M$. Note that $\omega^{rs/2} = \omega^{Kq} \equiv (-1)^q = -1$. The lemma follows since

$$\sum_{j=0}^{r-1} \omega^{js} = \sum_{j=0}^{\frac{r}{2}-1} \left(\omega^{sj} + \omega^{s(j+\frac{r}{2})} \right) \equiv \sum_{j=0}^{\frac{r}{2}-1} \left(\omega^{sj} - \omega^{sj} \right) = 0.$$

Q.E.D.

Using ω, we define the discrete Fourier transform and its inverse in $\mathbb{Z}_M$ as usual: $\mathrm{DFT}_{2K}(\mathbf{a}) := F(\omega)\cdot\mathbf{a}$ and $\mathrm{DFT}_{2K}^{-1}(\mathbf{A}) := \frac{1}{2K}F(\omega^{-1})\cdot\mathbf{A}$. To see that the inverse transform is well-defined, we should recall that $\frac{1}{2K}$ and ω^{-1} both exist. Our proof that DFT and DFT^{-1} are inverses (Lemma 2) goes through. We obtain the analogue of Theorem 3:

Theorem 8 *The transforms* $\mathrm{DFT}_{2K}(\mathbf{a})$ *and* $\mathrm{DFT}_{2K}^{-1}(\mathbf{A})$ *for* $(2K)$*-vectors* $\mathbf{a}, \mathbf{A} \in (\mathbb{Z}_M)^{2K}$ *can be computed using the Fast Fourier Transform method, taking* $O(KL\log K)$ *bit operations.*

Proof. We use the FFT method as before (refer to the three steps in the FFT display box in §1). View $\mathbf{a}$ as the coefficient vector of the polynomial $P(X)$. Note that ω is easily available in our representation, and ω^2 is a primitive Kth root of unity in $\mathbb{Z}_M$. This allows us to implement step 1 recursively, by calling DFT_K twice, once on the even part $P_e(Y)$ and again on the odd part $P_o(Y)$. In step 2, we need to compute ω^j (which is easy) and multiply it to $P_o(\omega^{2j})$ (also easy), for $j = 0, \ldots, 2K-1$. Step 2 takes $O(KL)$ bit operations. Finally, we need to add $\omega^j P_o(\omega^{2j})$ to $P_e(\omega^{2j})$ in step 3. This also takes $O(KL)$ bit operations. Thus the overall number of bit operations $T(2K)$ satisfies the recurrence

$$T(2K) = 2T(K) + O(KL)$$

which has solution $T(2K) = O(KL\log K)$, as claimed. **Q.E.D.**

Remarks: It is not hard to show (exercise below) that if M is prime then L is a power of 2. Generally, a number of the form $2^{2^n}+1$ is called *Fermat number*. The first 4 Fermat numbers are prime which led Fermat to the rather unfortunate conjecture that they all are. No other primes have been discovered so far and many are known to be composite (Euler discovered in 1732 that the 5th Fermat number $2^{2^5}+1$ is divisible by 641). Fermat numbers are closely related to a more fortunate conjecture of Mersenne, that all numbers of the form $2^p - 1$ are prime (where p is prime): although the conjecture is false, at least there is more hope that there are infinitely many such primes.

Exercises

Exercise 3.2: (i) If $a^L + 1$ is prime where $a \geq 2$, then a is even and L is a power of two.
(ii) If $a^L - 1$ is prime where $L > 1$, then $a = 2$ and L is prime. □

Exercise 3.3: Show that Strassen's recurrence $T(n) = n \cdot T(\log n)$ satisfies

$$T(n) = O\left(\left(\prod_{i=0}^{k-1} \log^{(i)} n\right)(\log^{(k)} n)^{1+\epsilon}\right) \tag{4}$$

for any $k < \log^*(n)$. HINT: use bootstrapping. □

Exercise 3.4: (Karatsuba) The first subquadratic algorithm for integer multiplication uses the fact that if $U = 2^L U_0 + U_1$ and $V = 2^L V_0 + V_1$ where U_i, V_i are L-bit numbers, then $W = UV = 2^{2L}U_0V_0 + 2^L(U_0V_1 + U_1V_0) + U_1V_1$, which we can rewrite as $2^{2L}W_0 + 2^L W_1 + W_2$. But if we compute $(U_0+U_1)(V_0+V_1), W_0, W_2$, we also obtain W_1. Show that this leads to a time bound of $T(n) = O(n^{\lg 3})$. □

§4. Fast Integer Multiplication

The following result of Schönhage and Strassen [185] is perhaps "the fundamental result" of the algorithmic algebra.

Theorem 9 (Complexity of integer multiplication) *Given two integers u, v of sizes at most n bits, we can form their product uv in $O(n \log n \log\log n)$ bit-operations.*

For simplicity, we prove a slightly weaker version of this result, obtaining a bound of $O(n \log^{2.6} n)$ instead.

A simplified Schönhage-Strassen algorithm. Our goal is to compute the product W of the positive integers U, V. Assume U, V are N-bit binary numbers where $N = 2^n$. Choose $K = 2^k, L = 3 \cdot 2^\ell$ where

$$k := \left\lfloor \frac{n}{2} \right\rfloor, \quad \ell := \lceil n - k \rceil .$$

Observe that although k, ℓ are integers, we will not assume that n is integer (*i.e.*, N need not be a power of 2). This is important for the recursive application of the method.

Since $k + \ell \geq n$, we may view U as $2^{k+\ell}$-bit numbers, padding with zeros as necessary. Break up U into K pieces, each of bit-size 2^ℓ. By padding these with K additional zeros, we get the the $(2K)$-vector,

$$\overline{U} = (0, \ldots, 0, U_{K-1}, \ldots, U_0)$$

where U_j are 2^ℓ-bit strings. Similarly, let

$$\overline{V} = (0, \ldots, 0, V_{K-1}, \ldots, V_0)$$

be a $(2K)$-vector where each component has 2^ℓ bits. Now regard $\overline{U}, \overline{V}$ as the coefficient vectors of the polynomials $P(X) = \sum_{j=0}^{K-1} U_j X^j$ and $Q(X) = \sum_{j=0}^{K-1} V_j X^j$. Let

$$\overline{W} = (W_{2K-1}, \ldots, W_0)$$

be the convolution of $\overline{U}$ and $\overline{V}$. Note that each W_i in $\overline{W}$ satisfies the inequality

$$0 \leq W_i \leq K \cdot 2^{2 \cdot 2^\ell} \tag{5}$$

since it is the sum of at most K products of the form $U_j V_{i-j}$. Hence

$$0 \leq W_i < 2^{3 \cdot 2^\ell} < M$$

where $M = 2^L + 1$ as usual. So if arithmetic is carried out in $\mathbb{Z}_M$, $\overline{W}$ will be correctly computed. Recall that $\overline{W}$ is the coefficient vector of the product $R(X) = P(X)Q(X)$. Since $P(2^{2^\ell}) = U$ and $Q(2^{2^\ell}) = V$, it follows that $R(2^{2^\ell}) = UV = W$. Hence

$$W = \sum_{j=0}^{2K-1} 2^{2^\ell j} W_j.$$

We can easily obtain each summand in this sum from $\overline{W}$ by multiplying each W_j with $2^{2^\ell j}$. As each W_j has $k + 2 \cdot 2^\ell < L$ non-zero bits, we illustrate this summation as follows:

From this figure we see that each bit of W is obtained by summing at most 3 bits plus at most 2 carry bits. Since W has at most $2N$ bits, we conclude:

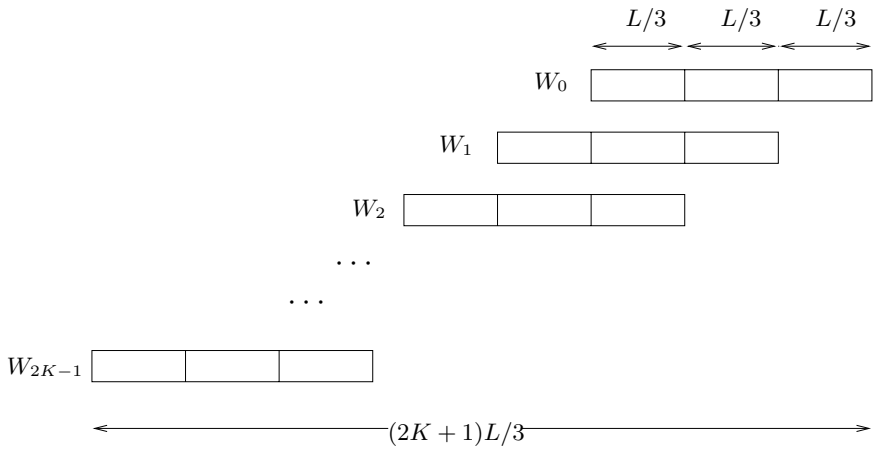


Figure 2: Illustrating forming the product $W = UV$.

Lemma 10 *The product W can be obtained from $\overline{W}$ in $O(N)$ bit operations.*

It remains to show how to compute $\overline{W}$. By the convolution theorem,

$$\overline{W} = \mathrm{DFT}^{-1}(\mathrm{DFT}(\overline{U}) \cdot \mathrm{DFT}(\overline{V})).$$

These three transforms take $O(KL \log K) = O(N \log N)$ bit operations (Theorem 8). The scalar product $\mathrm{DFT}(\overline{U}) \cdot \mathrm{DFT}(\overline{V})$ requires $2K$ multiplications of L-bit numbers, which is accomplished recursively. Thus, if $T(N)$ is the bit-complexity of this algorithm, we obtain the recurrence

$$T(N) = O(N \log N) + 2K \cdot T(L). \tag{6}$$

Write $t(n) := T(N)/N$ where $N = 2^n$. The recurrence becomes

$$\begin{aligned} t(n) &= O(n) + 2\frac{K}{N}T(L) \\ &= O(n) + 2 \cdot \frac{3}{L}T(L) \\ &= O(n) + 6 \cdot t(\frac{n}{2} + c), \end{aligned}$$

for some constant c. Recall that n is not necessarily integer in this notation. To solve this recurrence, we shift the domain of $t(n)$ by defining $s(n) := t(n + 2c)$. Then

$$s(n) = O(n + 2c) + 6t((n/2) + 2c) = O(n) + 6s(n/2).$$

This has solution $s(n) = O(n^{\lg 6})$. Back-substituting, we obtain

$$T(N) = O(N \log^{\alpha} N), \qquad \alpha = \lg 6 < 2.5848. \tag{7}$$

Refinements. Our choice of $L = 3 \cdot 2^{\ell}$ is clearly suboptimal. Indeed, it is not hard to see that our method really implies

$$T(N) = O(N \log^{2+\varepsilon} N)$$

for any $\varepsilon > 0$. A slight improvement (attributed to Karp in his lectures) is to compute each W_i $(i = 0, \ldots, 2K-1)$ in two parts: let $M' := 2^{2 \cdot 2^{\ell}} + 1$ and $M'' := K$. Since M', M'' are relatively prime

and $W_i < M'M''$, it follows that if we have computed $W_i' := W_i \bmod M'$ and $W_i'' := W_i \bmod M''$, then we can recover W_i using the Chinese remainder theorem (Lecture IV). It turns out that computing all the W_i'''s and the reconstruction of W_i from W_i', W_i'' can be accomplished in linear time. The computation of the W_i''s proceeds exactly as the above derivation. The new recurrence we have to solve is

$$t(n) = n + 4t(n/2)$$

which has the solution $t(n) = O(n^2)$ or $T(N) = O(N \log^2 N)$. To obtain the ultimate result, we have to improve the recurrence to $t(n) = n + 2t(n/2)$. In addition to the above ideas (Chinese remainder, etc), we must use a variant convolution called "negative wrapped convolution" and DFT_K instead of DFT_{2K}. Then W_i's can be uniquely recovered.

Exercise 4.1: Carry out the outline proposed by Karp. □

Integer multiplication in other models of computation. In the preceding algorithm, we only counted bit operations and it is not hard to see that this complexity can be achieved on a RAM model. It is tedious but possible to carry out the Schönhage-Strassen algorithm on a Turing machine, in the same time complexity. Thus we conclude

$$\mathrm{M}_B(n) = O(n \log n \log\log n) = n\mathcal{L}(n)$$

where $\mathrm{M}_B(n)$ denotes the Turing complexity of multiplying two n-bit integers (§0.7). This bound on $\mathrm{M}_B(n)$ can be improved for more powerful models of computation. Schönhage [182] has shown that linear time is sufficient on pointer machines. Using general simulation results, this translates to $O(n \log n)$ time on logarithmic-cost successor RAMs (§0.5). In parallel models, $O(\log n)$ time suffices on a parallel RAM.

Extending the notation of $\mathrm{M}_B(n)$, let

$$\mathrm{M}_B(m, n)$$

denote the Turing complexity of multiplying two integers of sizes (respectively) at most m and n bits. Thus, $\mathrm{M}_B(n) = \mathrm{M}_B(n, n)$. It is straightforward to extend the bound on $\mathrm{M}_B(n)$ to $\mathrm{M}_B(m, n)$.

Exercises

Exercise 4.2: Show that $\mathrm{M}_B(m, n) = \max\{m, n\}\mathcal{L}(\min\{m, n\})$. □

Exercise 4.3: Show that we can take remainders $u \bmod v$ and form quotients $u \operatorname{div} v$ of integers in the same bit complexity as multiplication. □

Exercise 4.4: Show how to multiply in $\mathbb{Z}_p$ ($p \in \mathbb{N}$ a prime) in bit complexity $O(\log p\, \mathcal{L}(\log p))$, and form inverses in $\mathbb{Z}_p$ in bit complexity $O(\log p\, \mathcal{L}^2(\log p))$. □

§5. Matrix Multiplication

For arithmetic on matrices over a ring R, it is natural that our computational model is algebraic programs over the base comprising the ring operations of R. Here the fundamental discovery by

Strassen (1968) [195] that the standard algorithm for matrix multiplication is suboptimal started off intense research for over a decade in the subject. Although the final word is not yet in, rather substantial progress had been made. These results are rather deep and we only report the current record, due to Coppersmith and Winograd (1987) [48]:

Proposition 11 (Algebraic complexity of matrix multiplication) *The product of two matrices in* $\mathrm{M}_n(R)$ *can be computed of* $O(n^\alpha)$ *operations in the ring* R*, where* $\alpha = 2.376$*. In other words,*

$$\mathrm{MM}(n) = O(n^\alpha).$$

It is useful to extend this result to non-square matrix multiplication. Let $\mathrm{MM}(m, n, p)$ denote the number of ring operations necessary to compute the product of an $m \times n$ matrix by a $n \times p$ matrix. So $\mathrm{MM}(n) = \mathrm{MM}(n, n, n)$.

Theorem 12 *Let* $\mathrm{MM}(n) = O(n^\alpha)$ *for some* $\alpha \geq 2$*. Then*

$$\mathrm{MM}(m, n, p) = O(mnp \cdot k^{\alpha-3})$$

where $k = \min\{m, n, p\}$.

Proof. Suppose A is a $m \times n$ matrix, B a $n \times p$ matrix. First assume $m = p$ but n is arbitrary. Then the bound in our theorem amounts to:

$$\mathrm{MM}(m, n, m) = \begin{cases} O(nm^{\alpha-1}) & \text{if } m \leq n \\ O(m^2 n^{\alpha-2}) & \text{if } n \leq m. \end{cases}$$

We prove this in two cases. *Case:* $m \leq n$. We partition A into $r = \lceil n/m \rceil$ matrices, $A = [A_1|A_2|\cdots|A_r]$ where each A_i is an m-square matrix except possibly for A_r. Similarly partition B into r m-square matrices, $B^T = [B_1^T|B_2^T|\cdots|B_r^T]$. Then

$$AB = A_1B_1 + A_2B_2 + \cdots + A_rB_r.$$

We can regard A_rB_r as a product of two m-square matrices, simply by padding out A_r and B_r with zeros. Thus each A_iB_i can be computed in m^α operations. To add the products $A_1B_1, \ldots, A_rB_r$ together, we use $O(rm^2) = O(rm^\alpha)$ addition operations. Hence the overall complexity of computing AB is $O(rm^\alpha) = O(nm^{\alpha-1})$ as desired.
Case: $n \leq m$. We similarly break up the product AB into r^2 products of the form A_iB_j, $i, j = 1, \ldots, r$, $r = \lceil m/n \rceil$. This has complexity $O(r^2 n^\alpha) = O(m^2 n^{\alpha-2})$. This completes the proof for the case $m = p$.

Next, since the roles of m and p are symmetric, we may assume $m < p$. Let $r = \lceil p/m \rceil$. We have two cases: (1) If $m \leq n$ then $\mathrm{MM}(m, n, p) \leq r\mathrm{MM}(m, n, m) = O(pnm^{\alpha-2})$. (2) If $n < m$, then $\mathrm{MM}(m, n, p) \leq r\mathrm{MM}(m, n, m) = O(rm^2 n^{\alpha-2}) = O(pmn^{\alpha-2})$. **Q.E.D.**

Notice that this result is independent of any internal details of the $O(n^\alpha)$ matrix multiplication algorithm. Webb Miller [133] has shown that under sufficient conditions for numerical stability, any algorithm for matrix multiplication over a ring requires n^3 multiplications. For a treatment of stability of numerical algorithms (and Strassen's algorithm in particular), we recommend the book of Higham [81].

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture II
The GCD

Next to the four arithmetic operations, the greatest common denominator (GCD) is perhaps the most basic operation in algebraic computing. The proper setting for discussing GCD's is in a unique factorization domain (UFD). For most common UFDs, that venerable algorithm of Euclid is available. In the domains $\mathbb{Z}$ and $F[X]$, an efficient method for implementing Euclid's algorithm is available. It is the so-called half-GCD approach, originating in ideas of Lehmer, Knuth and Schönhage. The presentation here is based on unpublished joint work with Klaus Thull, and gives a unified framework for the half-GCD approach for both integer and polynomial GCD. We also give the first proof for the correctness of the (corrected) polynomial half-GCD algorithm.

> The student will not go far amiss if she interprets all references to rings as either the integers $\mathbb{Z}$ or a polynomial ring $F[X]$ over a field F (even taking $F = \mathbb{Q}$).

§1. Unique Factorization Domain

Let D be a commutative ring. All rings in this book contain unity 1 where $0 \neq 1$. For $a, b \in D$, we say b *divides* a, and write $b \,|\, a$, if there is a $c \in D$ such that $a = bc$. If b does not divide a, we write $b \nmid a$. We also call b a *divisor* of a, and a a *multiple* of b. Thus every element divides 0 but 0 does not divide any non-zero element. A *zero-divisor* is an element b such that $bc = 0$ for some non-zero c. We also call an element *regular* if it is not a zero-divisor. An *integral domain* (or *domain* for short) D is a commutative ring whose only zero-divisor is 0. A *unit* is an element that divides 1. (Alternatively, units are the invertible elements.) Thus the unity element 1 is always a unit and the zero element is never a unit. In a field, all non-zero elements are units. Two elements, a and b, are *associates* if $a = ub$ for some unit u. Clearly the relation of being associates is an equivalence relation. So the elements of D are partitioned into equivalence classes of associates.

Exercise 1.1:
(a) The set of units and the set of zero-divisors are disjoint.
(b) $a \,|\, b$ and $b \,|\, a$ iff a, b are associates. □

Convention. For each equivalence class of associates, we assume that a *distinguished member* is chosen. The following convention captures most cases:
(i) The unity element 1 is always distinguished.
(ii) In $\mathbb{Z}$, the units are $+1$ and -1 and the equivalence classes are $\{-n, +n\}$ for each $n \in \mathbb{N}$. The non-negative elements will be distinguished in $\mathbb{Z}$.
(iii) In the polynomial ring $D[X]$ over a domain D, if we have specified distinguished elements in D then the distinguished elements of $D[X]$ will be those with distinguished leading coefficients. In case D is a field, this means the distinguished elements in $D[X]$ are the *monic polynomials*, *i.e.*, those with leading coefficient 1. Note that the product of distinguished elements are distinguished when $D = \mathbb{Z}$.

A *proper divisor* of b is any divisor that is neither a unit nor an associate of b. An element is *irreducible* if it has no proper divisors; otherwise it is *reducible*ring!reducible element. Since any divisor of a unit is a unit, it follows that units are irreducible. Furthermore, the zero element is irreducible if and only if D is a domain.

A *unique factorization domain* (abbreviated, UFD) is a domain D in which every non-unit b can be written as a product of irreducible non-units,

$$b = b_1\, b_2 \cdots b_n \qquad (n \geq 1).$$

Moreover, these irreducible elements are unique up to reordering and associates. UFD's are also called *factorial domains*.

The importance of UFD's is that its elements are made up of "fundamental building blocks", namely the irreducible elements. Note that $\mathbb{Z}$ is a UFD, by the *Fundamental Theorem of Arithmetic*. In fact, a UFD can be said to be a domain that has an analogue to the Fundamental Theorem of arithmetic! The non-zero irreducible elements of $\mathbb{Z}$ are called primes. But in general, we define the *prime*ring!prime elements elements of a ring R to be those non-units $p \in R$ such that $p \neq 0$ and if p divides any product $a \cdot b$ then p divides either a or b.

One sees that prime elements are irreducible but the converse is not generally true. For example (see [29, page 173]), in $\mathbb{C}[X, Y, Z]/(Z^2 - XY)$, Z is irreducible but not prime because Z divides XY without dividing X or Y. It is easy to see that in a UFD, every irreducible element is also a prime. Hence this is an example of a non-UFD.

Theorem 1 *D is a UFD iff $D[X]$ is a UFD.*

It is clear that $D[X]$ is not a UFD if D is not a UFD. The proof of the other direction is due to Gauss and is deferred to the next lecture. Trivially, a field F is a UFD. Hence, by induction on $d \geq 1$, this theorem shows that $F[X_1, \ldots, X_d]$ is a UFD.

Greatest common divisor. Let D be a UFD and $S \subseteq D$ be a finite non-empty set. We write $a \,|\, S$ (read, a divides S) to mean $a \,|\, b$ for all $b \in S$. An element $d \in D$ is a *greatest common divisor* (abbreviated, GCD) of S if

1) $d \,|\, S$,

2) if $c \,|\, S$ then $c \,|\, d$.

Exercise 1.2: Prove that S has a greatest common divisor, and this is determined up to associates. □

We can therefore define the function $\texttt{GCD}(S)$ by choosing the distinguished greatest common divisor of S. If $S = \{a_1, a_2, \ldots, a_m\}$, we write $\texttt{GCD}(a_1, a_2, \ldots, a_m)$ for $\texttt{GCD}(S)$. Unless otherwise noted, this lecture will assume that S has one or two elements: $S = \{a, b\}$. In this case, the GCD function may be regarded as a two argument function, $\texttt{GCD}(a, b)$. It is called the *simple GCD function*, as opposed to the *multiple GCD function* for general sets. If S has $m \geq 2$ elements, we can compute $\texttt{GCD}(S)$ using $m - 1$ simple GCD computations.

The following is easy.

$\texttt{GCD}(1, b)$	$=$	1	
$\texttt{GCD}(0, b)$	$=$	$\widehat{b}$	where $\widehat{b}$ is the distinguished associate of b
$\texttt{GCD}(a, b)$	$=$	$\texttt{GCD}(b, a)$	
$\texttt{GCD}(a + b, b)$	$=$	$\texttt{GCD}(a, b)$	
$\texttt{GCD}(ua, b)$	$=$	$\texttt{GCD}(a, b)$	where u is a unit

Say a, b are *relatively prime* or *co-prime* if $\texttt{GCD}(a, b) = 1$.

For instance, $\texttt{GCD}(123, 234) = 3$ and $\texttt{GCD}(3X^4-6X^3+13X^2-8X+12, 6X^5+17X^3-3X^2+12X-4) = 3X^2 + 4$.

GCD for ideals. Although we began with UFD's such as $\mathbb{Z}$ and $\mathbb{Q}[X]$, our Fundamental Problems force us to consider more general domains such as number rings (§VI.3). These rings need not be UFD's (exercise below). This led Kummer, Dedekind and Kronecker to develop ideal theory for algebraic numbers[1]. To regain the UFD property, we generalize numbers to ideals and introduce the concept of prime ideals. The ideal theoretic analogue of UFD's is this: a *Dedekind domain* is one in which every ideal is a product of prime ideals. It can be proved that such prime ideal factorizations are unique (e.g., [221, p. 273]). Number rings are Dedekind domains.

We do not define the concept of ideal divisibility via ideal products. Instead, if $I, J \subseteq D$ are ideals, we define I to be a *divisor* of J, and say I *divides* J, to mean $I \supseteq J$.

This definition is a stroke of inspiration from Dedekind (1871). Consider ideals in $\mathbb{Z}$: they have the form (n) where $n \in \mathbb{Z}$ since $\mathbb{Z}$ is a principal ideal domain (§3). Hence we can identify ideals of $\mathbb{Z}$ with numbers. Then $m, n \in \mathbb{Z}$ has the property that $m \,|\, n$ iff $(m) \supseteq (n)$, "agreeing" with our definition. In general, the relationship between ideal quotient and divisor property is only uni-directional: for ideals $I, J \subseteq D$, we have that $I \supseteq IJ$ and so I divides IJ.

The $\texttt{GCD}$ of a set S of ideals is by definition the smallest ideal that divides each $I \in S$, and we easily verify that

$$\texttt{GCD}(S) = \sum_{I \in S} I.$$

For $I = (a_1, \ldots, a_m)$ and $J = (b_1, \ldots, b_n)$, we have

$$\texttt{GCD}(I, J) = I + J = (a_1, \ldots, a_m, b_1, \ldots, b_n). \tag{1}$$

So the GCD problem for ideals is trivial unless we require some other conditions on the ideal generators. For instance, for the ideals of $\mathbb{Z}$, the $\texttt{GCD}$ of (a) and (b) is the ideal (a, b). But since $\mathbb{Z}$ is a principal ideal domain, we know that $(a, b) = (d)$ for some $d \in \mathbb{Z}$. We then interpret the $\texttt{GCD}$ problem in $\mathbb{Z}$ to mean the computation of d from a, b. It is not hard to prove that d is what we have defined to be a greatest common divisor of a, b. Thus, the common notation '(a, b)' for $\texttt{GCD}(a, b)$ is consistent with the ideal theoretic notation! In general, for a, b in a UFD, one should not expect $\texttt{Ideal}(a, b)$ to be generated by the $\texttt{GCD}(a, b)$. For instance, $\mathbb{Z}[X]$ is a UFD, $\texttt{GCD}(2, X) = 1$ but $\texttt{Ideal}(2, X) \neq \texttt{Ideal}(1)$.

Exercises

Exercise 1.3:

(a) Is the set of ideals of a domain D under the ideal sum and ideal product operations a ring? The obvious candidates for the zero and unity elements are (0) and $(1) = D$.

(b) Verify equation (1).

(c) What is the least common multiple, $\texttt{LCM}$, operation for ideal? □

Exercise 1.4: Say a domain D is *factorable* if every non-unit of D is a finite product of irreducible elements. Prove that a factorable domain D is a UFD iff irreducible elements are prime. □

[1] The other historical root for ideal theory is rational function fields in one variable.

Exercise 1.5: We will prove that the number ring $\mathbb{Z}[\sqrt{-5}] = \{x + y\sqrt{-5} : x, y \in \mathbb{Z}\}$ (cf. §VI.3) is not a UFD. The *norm* of an element $x + y\sqrt{-5}$ is $N(x + y\sqrt{-5}) := x^2 + 5y^2$. Show:
(a) $N(ab) = N(a)N(b)$.
(b) $N(a) = 1$ iff a is a unit. Determine the units of $\mathbb{Z}[\sqrt{-5}]$.
(c) If $N(a)$ is a prime integer then a is irreducible in $\mathbb{Z}[\sqrt{-5}]$.
(d) The numbers $2, 3, 1 + \sqrt{-5}, 1 - \sqrt{-5}$ are irreducible and not associates of each other. Since $6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5})$, conclude that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD. □

Exercise 1.6:
(a) In a principal ideal domain, the property "$I \supseteq J$" is equivalent to "there exists an ideal K such that $IK = J$".
(b) In $\mathbb{Z}[X, Y]$, there does not exist an ideal K such that $(X, Y) \cdot K = (X^2, Y^2)$. □

Exercise 1.7: (Lucas 1876) The GCD of two Fibonacci numbers is Fibonacci. □

Exercise 1.8: (Kaplansky) Define a *GCD-domain* to be a domain in which any two elements have a GCD.
a) Show that if D is such a domain, then so is $D[X]$.
b) Show that if for any two elements $u, v \in D$, either $u \mid v$ or $v \mid u$ (D is a *valuation domain*) then D is a GCD-domain. □

§2. Euclid's Algorithm

We describe Euclid's algorithm for computing the GCD of two positive integers

$$m_0 > m_1 > 0.$$

The algorithm amounts to constructing a sequence of remainders,

$$m_0, m_1, m_2, \ldots, m_k, \qquad (k \geq 1) \tag{2}$$

where

$$\begin{aligned} m_{i+1} &= m_{i-1} \bmod m_i \qquad (i = 1, \ldots, k-1) \\ 0 &= m_{k-1} \bmod m_k. \end{aligned}$$

Recall that $a \bmod b$ is the remainder function that returns an integer in the range $[0, b)$. But this is not the only possibility (next section).

Let us prove that m_k equals $\texttt{GCD}(m_0, m_1)$. We use the observation that if any number d divides m_i and m_{i+1}, then d divides m_{i-1} (provided $i \geq 1$) and d divides m_{i+2} (provided $i \leq k-2$). Note that m_k divides m_k and m_{k-1}. So by repeated application of the observation, m_k divides both m_0 and m_1. Next suppose d is any number that divides m_0 and m_1. Then repeated application of the observation implies d divides m_k. Thus we conclude that $m_k = \texttt{GCD}(m_0, m_1)$.

Two pieces of data related to the $\texttt{GCD}(m_0, m_1)$ are often important. Namely, there exist integers s, t such that

$$\texttt{GCD}(m_0, m_1) = sm_0 + tm_1. \tag{3}$$

We call the pair (s, t) a *co-factor* of (m_0, m_1). By the *co-GCD problem*, we mean the problem of computing a co-factor for an input pair of numbers. It is easy to obtain the GCD from a co-factor.

However, most co-GCD algorithms also produces the GCD with no extra effort. By definition, an *extended GCD algorithm* solves both the GCD and co-GCD problems. The existence of co-factors will be proved by our construction of an extended GCD algorithm next.

We proceed as follows. Suppose q_i is the quotient of the ith remaindering step in (2):

$$m_{i+1} = m_{i-1} - q_i m_i \qquad (i = 2, \ldots, k-1) \tag{4}$$

We compute two auxiliary sequences

$$(s_0, s_1, \ldots, s_k) \quad \text{and} \quad (t_0, t_1, \ldots, t_k) \tag{5}$$

so that they satisfy the property

$$m_i = s_i m_0 + t_i m_1, \qquad (i = 0, \ldots, k). \tag{6}$$

Note that when $i = k$, this property is our desired equation (3). The auxiliary sequences are obtained by mirroring the remaindering step (4),

$$\left.\begin{array}{l} s_{i+1} = s_{i-1} - q_i s_i \\ t_{i+1} = t_{i-1} - q_i t_i \end{array}\right\} \qquad (i = 2, \ldots, k-1) \tag{7}$$

To initialize the values of s_0, s_1 and t_0, t_1, observe that

$$m_0 = 1 \cdot m_0 + 0 \cdot m_1$$

and

$$m_1 = 0 \cdot m_0 + 1 \cdot m_1.$$

Thus (6) is satisfied for $i = 0, 1$ if we set

$$(s_0, t_0) := (1, 0), \qquad (s_1, t_1) := (0, 1).$$

Inductively, (6) is satisfied because

$$\begin{aligned} m_{i+1} &= m_{i-1} - q_i m_i \\ &= (s_{i-1} m_0 + t_{i-1} m_1) - q_i (s_i m_0 + t_i m_1) \\ &= (s_{i-1} - q_i s_i) m_0 + (t_{i-1} - q_i t_i) m_1 \\ &= s_{i+1} m_0 + t_{i+1} m_1. \end{aligned}$$

This completes the description and proof of correctness of the extended Euclidean algorithm.

Application. Suppose we want to compute multiplicative inverses modulo an integer m_0. An element m_1 has a multiplicative inverse modulo m_0 if and only if $\texttt{GCD}(m_0, m_1) = 1$. Applying the extended algorithm to m_0, m_1, we obtain s, t such that

$$1 = \texttt{GCD}(m_0, m_1) = s m_0 + t m_1.$$

But this implies

$$1 \equiv t m_1 (\mathrm{mod}\, m_0),$$

i.e., t is the inverse of m_1 modulo m_0. Similarly s is the inverse of m_0 modulo m_1.

Exercises

Exercise 2.1: (i) Show that every two steps of the Euclidean algorithm reduce the (bit) size of the larger integer by at least one. Conclude that the bit complexity of the Euclidean algorithm is $O(nM_B(n))$ where $M_B(n)$ is the bit complexity of integer multiplication.
(ii) Improve this bound to $O(n^2)$. HINT: If the bit length of m_i in the remainder sequence is ℓ_i, then the bit length of q_i is at most $\ell_{i-1} - \ell_i + 1$. The ith step can be implemented in time $O(\ell_i(\ell_{i-1} - \ell_+1))$. □

Exercise 2.2: Consider the extended Euclidean algorithm.
(i) Show that for $i \geq 2$, we have $s_i t_i < 0$ and $s_i > 0$ iff i is even.
(ii) Show that the co-factor (s, t) computed by the algorithm satisfy $\max\{|s|, |t|\} < m_0$. □

Exercise 2.3: (Blankinship) The following is a simple basis for an extended multiple GCD algorithm. Let $N = (n_1, \ldots, n_k)^T$ be a k-column of integers and A the $k \times (k+1)$ matrix whose first column is N, and the remaining columns form an identity matrix. Now perform any sequence of row operations on A of the form "subtract an integer multiple of one row from another". It is clear that we can construct a finite sequence of such operations so that the first column eventually contains only one non-zero entry d where $d = \texttt{GCD}(n_1, \ldots, n_k)$. If the row containing d is $(d, s_1, \ldots, s_k)$, prove that

$$d = \sum_{i=1}^{k} s_i n_i.$$

□

Exercise 2.4:
(i) Let $n_1 > n_2 > \cdots > n_k > 1$ ($k \geq 1$) be integers. Let $S = (s_1, \ldots, s_k) \in \mathbb{Z}^k$ be called a *syzygy* of $N = (n_1, \ldots, n_k)$ if $\sum_{i=1}^{k} s_i n_i = 0$. Prove that the set of syzygies of N forms a $\mathbb{Z}$-module. For instance, let s_{ij} ($1 \leq i < j \leq n$) be the k-vector $(0, \ldots, 0, n_j, 0, \ldots, 0, -n_i, 0, \ldots, 0)$ (where the only non-zero entries are at positions i and j as indicated). Clearly s_{ij} is a syzygy. This module has a finite basis (XI§1). Construct such a basis.
(ii) Two k-vectors S, S' are *equivalent* if $S - S'$ is a syzygy of N. Show that every S is equivalent to some S' where each component c of S' satisfies $|c| < n_1$. □

§3. Euclidean Ring

We define the abstract properties that make Euclid's algorithm work. A ring R is *Euclidean* if there is a function

$$\varphi : R \to \{-\infty\} \cup \mathbb{R}$$

such that

i) $b \neq 0$ and $a|b$ implies $\varphi(a) \leq \varphi(b)$;

ii) for all $r \in \mathbb{R}$, the set $\{\varphi(a) : a \in R, \varphi(a) < r\}$ is finite;

iii) for all $a, b \in R$ ($b \neq 0$), there exists $q, r \in R$ such that

$$a = bq + r \quad \text{and} \quad \varphi(r) < \varphi(b).$$

We say that φ is an *Euclidean value function* for R, and call the q and r in iii) a *quotient* and *remainder* of a, b relative to φ. Property iii) is called the *division property* (relative to φ). We introduce the *remainder* $\texttt{rem}(a, b)$ and *quotient* $\texttt{quo}(a, b)$ functions that pick out some definite pair of remainder and quotient of a, b that simultaneously satisfy property iii). Note that these functions are only defined when $b \neq 0$. Often it is convenient to write these two functions using infix operators **mod** and **div**:

$$\texttt{rem}(a, b) = a \,\mathbf{mod}\, b, \quad \texttt{quo}(a, b) = a \,\mathbf{div}\, b. \tag{8}$$

A *Euclidean domain* is an Euclidean ring that is also a domain.

Exercise 3.1:
(a) $\texttt{rem}(a, b) = 0$ if and only if $b|a$ (in particular, $\texttt{rem}(0, b) = 0$).
(b) $\varphi(a) = \varphi(b)$ when a and b are associates.
(c) $\varphi(0) < \varphi(b)$ for all non-zero b. □

Our two standard domains, $\mathbb{Z}$ and $F[X]$, are Euclidean:
(A) $\mathbb{Z}$ is seen to be an Euclidean domain by letting $\varphi(n) = |n|$, the absolute value of n. There are two choices for $\texttt{rem}(m, n)$ unless $n|m$, one positive and one negative. For instance, $\texttt{rem}(8, 5)$ can be taken to be 3 or -2. There are two standard ways to make $\texttt{rem}(m, n)$ functional. In the present lecture, we choose the non-negative remainder. The corresponding function $\texttt{rem}(m, n) \geq 0$ is called the *non-negative remainder function*. An alternative is to choose the remainder that minimizes the absolute value (choosing the positive one in case of a tie); this corresponds to the *symmetric remainder function*. The function $\texttt{quo}(a, b)$ is uniquely determined once $\texttt{rem}(a, b)$ is fixed. Again, we have the *non-negative quotient function* and the *symmetric quotient function*.
(B) If F is any field, the following *division property for polynomials* holds: for $A, B \in F[X]$ where $B \neq 0$, there exists $Q, R_0 \in F[X]$ such that

$$A = BQ + R_0, \qquad \deg(R_0) < \deg(B).$$

This can be proved by the synthetic division algorithm which one learns in high school. It follows that the polynomial ring $F[X]$ is an Euclidean domain, as witnessed by the choice $\varphi(P) = \deg P$, for $P \in F[X]$. In fact, the synthetic division algorithm shows that $\texttt{rem}(P, Q)$ and $\texttt{quo}(P, Q)$ are uniquely determined. Despite property ii) in the definition of φ, there may be infinitely many $a \in R$ with $\varphi(a) < r$. This is the case if $R = F[X]$ with F infinite.

Lemma 2 *If a is a proper divisor of b then $\varphi(a) < \varphi(b)$.*

Proof. By the division property, $a = bq + r$ where $\varphi(r) < \varphi(b)$. Now $r \neq 0$ since otherwise b divides a, which contradicts the assumption that a properly divides b. Since a properly divides b, let $b = ac$ for some c. Then $r = a - bq = a(1 - cq)$. Then property i) implies $\varphi(a) \leq \varphi(r) < \varphi(b)$. **Q.E.D.**

Theorem 3 *An Euclidean ring is a principal ideal ring. Indeed, if $b \in I \setminus \{0\}$ is such that $\varphi(b)$ is minimum then $I = \texttt{Ideal}(b)$.*

Proof. Let I be any ideal. By property ii), there exists a $b \in I \setminus \{0\}$ such that $\varphi(b)$ is minimum. To show $I = \texttt{Ideal}(b)$, it suffices to show that b divides any $c \in I \setminus \{0\}$. By the division property, $c = bq + r$ where $\varphi(r) < \varphi(b)$. If $r \neq 0$ then we have found an element $r = c - bq \in I \setminus \{0\}$ with $\varphi(r) < \varphi(b)$, contradicting our choice of b. If $r = 0$ then $b|c$. **Q.E.D.**

The converse is not true (see exercise).

Lemma 4 *In a principal ideal ring R, the non-zero irreducible elements are prime.*

Proof. Let $p \in R \setminus \{0\}$ be irreducible. If p divides the product bc, we must prove that p divides b or p divides c. Since R is a principal ideal ring, $\texttt{Ideal}(p, b) = \texttt{Ideal}(u)$ for some $u = \alpha p + \beta b$. So $u|p$. Since p is irreducible, u is a unit or an associate of p. If u is an associate, and since $u|b$, we have $p|b$, which proves the lemma. If u is a unit then $uc = \alpha pc + \beta bc$. Since $p|bc$, this implies $p|uc$, *i.e.*, $p|c$. **Q.E.D.**

Theorem 5 *In a principal ideal ring, the factorization of a non-unit into irreducible non-units is unique, up to reordering and associates.*

Proof. Suppose $b \in R$ is a non-unit with two factorizations into irreducible non-units:

$$b = p_1 p_2 \cdots p_m = q_1 q_2 \cdots q_n, \qquad 1 \le m \le n.$$

We use induction on m. If $m = 1$ then clearly $n = 1$ and $q_1 = p_1$. Assume $m > 1$. Since p_1 is a prime, it must divide some q_i, and we might as well assume $p_1|q_1$. But q_1 is also a prime and so it must be an associate of p_1. Dividing by p_1 on both sides of the expression, it follows that $p_2 \cdots p_m = q_2' q_3 \cdots q_n$ where q_2' is an associate of q_2. By induction, $m = n$ and the two factorizations are unique up to reordering and associates. This implies our theorem. **Q.E.D.**

Corollary 6 *An Euclidean domain is a UFD.*

Remainder sequences. Relative to the remainder and quotient functions, we define a *remainder sequence* for any pair $a, b \in R$ to be a sequence

$$a_0, a_1, \ldots, a_k \qquad (k \ge 1) \tag{9}$$

such that $a_0 = a, a_1 = b$ and for $i = 1, \ldots, k-1$, a_{i+1} is an associate of $\texttt{rem}(a_{i-1}, a_i)$, and $\texttt{rem}(a_{k-1}, a_k) = 0$. Note that termination of this sequence is guaranteed by property ii). The remainder sequence is *strict* if a_{i+1} is any remainder of a_{i-1}, a_i for all i; it is *Euclidean* if $a_{i+1} = \texttt{rem}(a_{i-1}, a_i)$.

Example: In $\mathbb{Z}$, $(13, 8, 5, 3, 2, 1)$, $(13, 8, -3, 2, \pm 1)$ and $(13, 8, 3, -1)$ are all strict remainder sequences for $(13, 8)$. A non-strict remainder sequence for $(13, 8)$ is $(13, 8, -5, 2, 1)$. ■

Associated to each remainder sequence (9) is another sequence

$$q_1, q_2, \ldots, q_k \tag{10}$$

where $a_{i-1} = a_i q_i + u_i a_{i+1}$ ($i = 1, \ldots, k-1$, u_i is a unit) and $a_{k-1} = a_k q_k$. We call (10) the *quotient sequence associated to*remainder sequence!2@
it see also quotient sequence the remainder sequence.

Norms. In some books, the function φ is restricted to the range $\mathbb{N}$. This restriction does not materially affect the concept of Euclidean domains, and has the advantage that property ii) is

automatic. Our formulation makes it more convenient to formulate functions φ that have other desirable properties. For instance, we often find the properties:

$$\varphi(ab) = \varphi(a)\varphi(b)$$

and

$$\varphi(a+b) \leq \varphi(a) + \varphi(b).$$

In this case, we call φ a *multiplicative norm* (or *valuation*), and might as well (why?) assume $\varphi(0) = 0$ and $\varphi(1) = 1$. Similarly, if $\varphi(ab) = \varphi(a) + \varphi(b)$ and $\varphi(a+b) = O(1) + \max\{\varphi(a), \varphi(b)\}$ then we call φ an *additive norm*, and might as well assume $\varphi(0) = -\infty$ and $\varphi(1) = 0$. Clearly φ is multiplicative implies $\log \varphi$ is additive.

Remarks. The number rings $\mathbb{O}_\alpha$ (§VI.3) have properties similar to the integers. In particular, they support the concepts of divisibility and factorization. Gauss pointed out that such rings may not be a UFD (class number 1). Even when $\mathbb{O}_\alpha$ is a UFD, it may not be Euclidean; the "simplest" example is $\mathbb{O}_{\sqrt{-19}}$ (see [174]). An obvious candidate for the Euclidean value function φ is the norm of algebraic numbers, but other functions are conceivable. Turning now to the quadratic number rings (*i.e.*, $\mathbb{O}_\alpha$ where $\alpha = \sqrt{d}$ and d is squarefree), Kurt Heegner [Diophantische Analysis und Modulfunktionen, *Mathematische Zeitschrift*, vol. 56, 1952, pp.227–253] was the first[2] to prove that there are exactly nine such UFD's in which $d < 0$, *viz.*, $d = -1, -2, -3, -7, -11, -19, -43, -67, -163$. In contrast, it is conjectured that there are infinitely many UFD's among the real (*i.e.*, $d > 0$) quadratic number fields. It is known that there are precisely 21 real quadratic domains that support the Euclidean algorithm. Currently, the most general GCD algorithms are from Kaltofen and Rolletschek [98] who presented polynomial-time GCD algorithms for each quadratic number ring $\mathbb{O}_{\sqrt{d}}$ that is a UFD, not necessarily Euclidean.

Exercises

Exercise 3.2: Justify the above remarks about multiplicative and additive norms. □

Exercise 3.3: Verify that the Euclidean algorithm computes the `GCD` in an Euclidean domain, relative to the function $\texttt{rem}(a, b)$. □

Exercise 3.4: Show that the number ring $\mathbb{O}_{\mathbf{i}}$ $(= \mathbb{Z}[\mathbf{i}] = \{a + \mathbf{i}b : a, b \in \mathbb{Z}\})$ of Gaussian integers forms an Euclidean domain with respect to the multiplicative norm $\varphi(a + \mathbf{i}b) = a^2 + b^2$. Use the identity of Fibonacci,

$$\varphi(xy) = \varphi(x)\varphi(y), \qquad x, y \in \mathbb{Z}[\mathbf{i}].$$

What are the possible choices for defining the remainder and quotient functions here? □

Exercise 3.5: Consider the number ring $R = \mathbb{O}_{\sqrt{-19}}$. Note that $\mathbb{O}_{\sqrt{-19}} = \{m + n\omega : m, n \in \mathbb{Z}\}$ where $\omega = \frac{1+\sqrt{-19}}{2}$. The norm of $x + y\sqrt{-19} \in \mathbb{Q}(\sqrt{-19})$ is by $x^2 + 19y^2$.
(a) R is a principal ideal domain.
(b) R is not an Euclidean domain with respect to the standard norm function. HINT: What is the remainder of 5 divided by $\sqrt{-19}$? □

[2]This result is often attributed to Stark and Baker who independently proved this in 1966. See Buell [34].

§4. The Half-GCD Problem

An exercise in §2 shows that that Euclid's algorithm for integers has bit complexity $\Theta(n^2\mathcal{L}(n))$. Knuth [104] is the first to obtain a subquadratic complexity for this problem. In 1971, Schönhage [181] improved it to the current record of $O(\mathrm{M}_B(n)\log n) = n\mathcal{L}^2(n)$. Since $F[X]$ is an Euclidean domain, Euclid's algorithm can be applied to polynomials as well. Given $P_0, P_1 \in F[X]$ with $n = \deg P_0 > \deg P_1 \geq 0$, consider its Euclidean remainder sequence

$$P_0, P_1, P_2, \ldots, P_h \qquad (h \geq 1). \tag{11}$$

Call the sequence *normal* if $\deg P_{i-1} = 1 + \deg P_i$ for $i = 2, \ldots, h$. A random choice for P_0, P_1 gives rise to a normal sequence with high probability (this is because non-normal sequences arise from the vanishing of certain determinants involving the the coefficients of P_0, P_1, Lecture III). The algebraic complexity of this Euclidean algorithm is therefore

$$O(\mathrm{M}_A(n)n) = O(n^2 \log n) \tag{12}$$

where $\mathrm{M}_A(n) = O(n \log n)$ is the algebraic complexity of polynomial multiplication (Lecture 1). Moenck [141] improves (12) to $O(\mathrm{M}_A(n)\log n)$ in case the remainder sequence is normal. Aho-Hopcroft-Ullman [2] incorrectly claimed that the Moenck algorithm works in general. Brent-Gustavson-Yun [26] presented a corrected version without a proof. Independently, Thull and Yap [204] rectified the algorithm with a proof, reproduced below. This lecture follows the unified framework for both the polynomial and integer GCD algorithms, first presented in [204].

Let us motivate the approach of Schönhage and Moenck. These ideas are easiest seen in the case of the polynomials. If the sequence (11) is normal with $n = h$ and $\deg P_i = n - i$ $(i = 0, \ldots, n)$ then

$$\sum_{i=0}^{h} \deg P_i = n(n-1)/2.$$

So any algorithm that explicitly computes each member of the remainder sequence has at least quadratic complexity. On the other hand, if

$$Q_1, Q_2, \ldots, Q_h$$

is the quotient sequence associated to (11), then it is not hard to show that

$$\sum_{i=1}^{h} \deg Q_i = n. \tag{13}$$

Indeed, we can quickly (in $O(n \log^2 n)$ time, see Exercise below) obtain any member of the remainder sequence from the Q_i's. This suggests that we redirect attention to the quotient sequence.

Matrix Terminology. To facilitate description of our algorithms, we resort to the language of matrices and vectors. In this lecture, all matrices will be 2×2 matrices and all vectors will be column 2-vectors. The identity matrix is denoted by E. The Euclidean algorithm as embodied in (11) can be viewed as a sequence of transformations of 2-vectors:

$$\begin{bmatrix} P_0 \\ P_1 \end{bmatrix} \xrightarrow{M_1} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \xrightarrow{M_2} \cdots \xrightarrow{M_{h-1}} \begin{bmatrix} P_{h-1} \\ P_h \end{bmatrix} \xrightarrow{M_h} \begin{bmatrix} P_h \\ 0 \end{bmatrix}. \tag{14}$$

Precisely, if U, V are vectors and M a matrix, we write

$$U \xrightarrow{M} V$$

to mean that $U = MV$. Hence equation (14) can be correctly interpreted if we define

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

In general, an *elementary matrix* refers to a matrix of the form $M = \begin{bmatrix} Q & 1 \\ 1 & 0 \end{bmatrix}$ where Q is a polynomial with positive degree. We call Q the *partial!quotient* in M. A *regular matrix* M is a product of zero or more elementary matrices,

$$M = M_1 M_2 \cdots M_k \qquad (k \geq 0). \tag{15}$$

When $k = 0$, M is interpreted to be E. The sequence $Q_1, \ldots, Q_k$ of partial quotients associated with the elementary matrices $M_1, \ldots, M_k$ in equation (15) is called the *sequence of partial quotients* of M. Also, Q_k is called its *last partial quotient* Note that regular matrices have determinant ± 1 and so are invertible. Regular matrices arise because

$$U \xrightarrow{M} V \text{ and } V \xrightarrow{M'} W \text{ implies } U \xrightarrow{MM'} W.$$

Our terminology here is motivated by the connection to continued fractions (for instance, regular matrices are related to regular continued fractions).

We are ready to define the *half-GCD* (or, HGCD) *problem* for a polynomial ring $F[X]$:

> *Given* $P_0, P_1 \in F[X]$ *where* $n = \deg P_0 > \deg P_1$, *compute a regular matrix*
>
> $$M := \texttt{hGCD}(P_0, P_1)$$
>
> *such that if*
>
> $$\begin{bmatrix} P_0 \\ P_1 \end{bmatrix} \xrightarrow{M} \begin{bmatrix} P_2 \\ P_3 \end{bmatrix}$$
>
> *then*
>
> $$\deg P_2 \geq n/2 > \deg P_3. \tag{16}$$

In general, we say two numbers a, b *straddle* a third number c if $a \geq c > b$. Thus $\deg P_2, \deg P_3$ straddle $n/2$ in equation (16).

Now we show how to compute the GCD using the `hGCD`-subroutine. In fact the algorithm is really a "co-GCD" (§2) algorithm:

POLYNOMIAL CO-GCD ALGORITHM:
Input: A pair of polynomials with $\deg P_0 > \deg P_1$
Output: A regular matrix $M = \texttt{co-GCD}(P_0, P_1)$ such that

$$\begin{bmatrix} P_0 \\ P_1 \end{bmatrix} \xrightarrow{M} \begin{bmatrix} \texttt{GCD}(P_0, P_1) \\ 0 \end{bmatrix}.$$

[1] Compute $M_0 \leftarrow \texttt{hGCD}(P_0, P_1)$.
[2] Recover P_2, P_3 via

$$\begin{bmatrix} P_2 \\ P_3 \end{bmatrix} \leftarrow M_0^{-1} \cdot \begin{bmatrix} P_0 \\ P_1 \end{bmatrix}.$$

[3] if $P_3 = 0$ then return(M_0).
else, perform one step of the Euclidean algorithm,

$$\begin{bmatrix} P_2 \\ P_3 \end{bmatrix} \xrightarrow{M_1} \begin{bmatrix} P_3 \\ P_4 \end{bmatrix}$$

where M_1 is an elementary matrix.
[4] if $P_4 = 0$ then return$(M_0 M_1)$.
else, recursively compute $M_2 \leftarrow \texttt{co-GCD}(P_3, P_4)$
return$(M_0 M_1 M_2)$.

The correctness of this algorithm is clear. The reason for step [3] is to ensure that in our recursive call, the degree of the polynomials is less than $n/2$. The algebraic complexity $T(n)$ of this algorithm satisfies

$$T(n) = T'(n) + O(\mathrm{M}_A(n)) + T(n/2)$$

where $T'(n)$ is the complexity of the HGCD algorithm. Let us assume that

$$\mathrm{M}_A(n) = O(T'(n)), \qquad T'(\alpha n) \leq \alpha T'(n).$$

For instance, the first relation holds if $T'(n) = \Omega(\mathrm{M}_A(n))$; the second relation holds if $T'(n)$ is bounded by a polynomial. In particular, they hold if $T'(n) = \Theta(M(n) \log n)$, which is what we will demonstrate below. Then

$$T(n) = O(T'(n) + T'(n/2) + T'(n/4) + \cdots) = O(T'(n)).$$

In conclusion, the complexity of the GCD problem is the same order of the complexity as the HGCD problem. Henceforth, we focus on the HGCD problem.

Remarks: The above complexity counts ring operations from F. If we count operations in $F[X]$, the complexity is $O(n \log n)$. This counting is more general because it applies also to the case of integer HGCD to be discussed. Strassen [196] has proved that this complexity is optimal: $O(n \log n)$ is both necessary and sufficient.

EXERCISES

Exercise 4.1: Recall the auxiliary sequences $(s_0, s_1, \ldots, s_k)$ and $(t_0, t_1, \ldots, t_k)$ computed in the Extended Euclidean algorithm (§2) for the GCD of a pair of integers $a_0 > a_1 > 1$. Show that the appropriate elementary matrices have the form $\begin{bmatrix} s_i & t_i \\ s_{i+1} & t_{i+1} \end{bmatrix}^{-1}$. □

Exercise 4.2:
(a) Show equation (13).

(b) Show that in $O(n \log^2 n)$ time, we can reconstruct the polynomials S, T from the quotient sequence $Q_1, \ldots, Q_k$ where $SP_0 + TP_1 = \texttt{GCD}(P_0, P_1)$. HINT: note that $\begin{bmatrix} P_i \\ P_{i+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -Q_i \end{bmatrix} \begin{bmatrix} P_{i-1} \\ P_i \end{bmatrix}$, and use a balanced binary tree scheme. □

§5. Properties of the Norm

For the rest of this Lecture, the domain D refers to either $\mathbb{Z}$ or $F[X]$. In this case, we define the *(additive) norm* $\|a\|$ of $a \in D$ thus:

$$\|a\| := \begin{cases} \log_2 |a| & \text{if } a \in \mathbb{Z}, \\ \deg(a) & \text{if } a \in F[X]. \end{cases}$$

The previous section describes the polynomial HGCD problem. A similar, albeit more complicated, development can be carried out for integers. We now describe a common framework for both the integer and polynomial HGCD algorithms.

The following properties are easy to check:

a) $\|a\| \in \{-\infty\} \cup \mathbb{R}^*$ where $\mathbb{R}^*$ is the set of non-negative real numbers.

b) $\|a\| = -\infty \Longleftrightarrow a = 0$

c) $\|a\| = 0 \Longleftrightarrow a$ is a unit.

d) $\|-a\| = \|a\|$

e) $\|ab\| = \|a\| + \|b\|$

f) $\|a + b\| \leq 1 + \max\{\|a\|, \|b\|\}$.

The last two properties imply that the norm is additive (§3). However, polynomials satisfy the stronger *non-Archimedean property* (cf. [111, p.283]):

$$\|a + b\| \leq \max\{\|a\|, \|b\|\}.$$

It is this non-Archimedean property that makes polynomials relatively easier than integers. This property implies

$$\|a + b\| = \max\{\|a\|, \|b\|\} \qquad \text{if } \|a\| \neq \|b\|. \tag{17}$$

Exercise 5.1: Prove this. □

This norm function serves as the Euclidean value function for D. In particular, the *division property* relative to the norm holds: for any $a, b \in D$, $b \neq 0$, there exists $q, r \in D$ such that

$$a = qb + r, \quad \|r\| < \|b\|.$$

The remainder and quotient functions, $\texttt{rem}(a,b)$ and $\texttt{quo}(a,b)$, can be defined as before. Recall that these functions are uniquely defined for polynomials, but for integers, we choose $\texttt{rem}(a,b)$ to be the non-negative remainder function. Note that in the polynomial case,

$$\|a \bmod X^m\| \leq \min\{\|a\|, m-1\} \tag{18}$$

$$\|a \operatorname{div} X^m\| = \begin{cases} \|a\| - m & \text{if } \|a\| \geq m \\ -\infty & \text{else.} \end{cases} \tag{19}$$

Matrices and vectors. The previous section (§4) introduced the matrix concepts we needed. Those definitions extend in the obvious way to our present setting, except for one place, where we need special care: A matrix of the form $M = \begin{bmatrix} q & 1 \\ 1 & 0 \end{bmatrix}$ (where $q \in D$) is denoted $\langle q \rangle$. A matrix is *elementary* if it has the form $\langle q \rangle$ where $\|q\| > 0$ in the case of polynomials (as before), and $q > 0$ in the case of integers. A finite product $\langle q_1 \rangle \langle q_2 \rangle \cdots \langle q_k \rangle$ $(k \geq 0)$ of elementary matrices is again called *regular* and may be denoted $\langle q_1, \ldots, q_k \rangle$. When $k = 2$, the careful reader will note the clash with our notation for scalar products, but this ambiguity should never confuse.

A regular matrix $M = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$ satisfies the following *ordering property*:

$$M \neq E \Rightarrow \quad \|p\| \geq \max\{\|q\|, \|r\|\} \geq \min\{\|q\|, \|r\|\} \geq \|s\|, \quad \|p\| > \|s\|. \tag{20}$$

Exercise 5.2:

a) Prove the ordering property.

b) If all the inequalities in the definition of the ordering property are in fact strict and $\|s\| \geq 0$, we say M satisfies the *strict ordering property*. Show that the product of three or more elementary matrices has the strict ordering property.

c) Bound the norms of the entries of the matrix $\langle q_1, \ldots, q_k \rangle$ in terms of the individual norms $\|q_i\|$. □

For vectors U, V and matrix M, we write

$$U \xrightarrow{M} V$$

(or simply, $U \longrightarrow V$) if $U = MV$. We say M *reduces*matrix!regular!reducing a vector U to V if M is a regular matrix. If, in addition, $U = \begin{bmatrix} a \\ b \end{bmatrix}, V = \begin{bmatrix} a' \\ b' \end{bmatrix}$ such that $\|a\| > \|b\|$ and $\|a'\| > \|b'\|$, then we say this is an *Euclidean reduction*.

A matrix is *unimodular* if[3] it has determinant ± 1. Clearly regular matrices are unimodular. Thus their inverses are easily obtained: if $M = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$ is regular with determinant $\det M = \delta$ then

$$M^{-1} = \delta \begin{bmatrix} s & -q \\ -r & p \end{bmatrix} = \pm \begin{bmatrix} s & -q \\ -r & p \end{bmatrix}. \tag{21}$$

If $U = \begin{bmatrix} a \\ b \end{bmatrix}$ then we write $\texttt{GCD}(U)$ for the GCD of a and b. We say U, V are *equivalent* if $U = MV$ for some unimodular matrix M.

[3] In some literature, "unimodular" refers to determinant $+1$.

Lemma 7 *U and V are equivalent if and only if* $\texttt{GCD}(U) = \texttt{GCD}(V)$.

Proof. It is easy to check that if the two vectors are equivalent then they must have the same GCD. Conversely, by Euclid's algorithm, they are both equivalent to the vector $\begin{bmatrix} g \\ 0 \end{bmatrix}$ where g is their common GCD. **Q.E.D.**

It follows that this binary relation between vectors is indeed a mathematical equivalence relation. The following is a key property of Euclidean remainder sequences (§3):

Lemma 8 *Given a, b, a', b' such that $\|a\| > \|b\| \geq 0$. The following are equivalent:*
(i) *a', b' are consecutive elements in the Euclidean remainder sequence of a, b.*
(ii) *There is a regular matrix M such that*

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M} \begin{bmatrix} a' \\ b' \end{bmatrix} \tag{22}$$

and either $\|a'\| > \|b'\| \geq 0$ (polynomial case) or $a' > b' > 0$ (integer case).

Proof. If (i) holds then we can (by Euclid's algorithm) find some regular matrix M satisfying (ii). Conversely assume (ii). We show (i) by induction on the number of elementary matrices in the product M. The result is immediate if $M = E$. If M is elementary, then (i) follows from the division property for our particular choices for D. Otherwise let $M = M''M'$ where M' is elementary and M'' is regular. Then for some a'', b'',

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M''} \begin{bmatrix} a'' \\ b'' \end{bmatrix} \xrightarrow{M'} \begin{bmatrix} a' \\ b' \end{bmatrix}.$$

But $a'' = a'q' + b'$ and $b'' = a'$ where q' is the partial quotient of M'. We verify that this means $\|a''\| > \|b''\|$. By induction, a'', b'' are consecutive elements in a strict remainder sequence of a, b. Then (i) follows. **Q.E.D.**

Exercises

Exercise 5.3: In Exercise 2.1, we upper bound the length of the integer Euclidean remainder sequence of $a > b > 0$ by $2\log_2 a$. We now give a slight improvement.

(a) Prove that for $k \geq 1$,

$$\underbrace{\langle 1, \ldots, 1 \rangle}_{k} = \langle 1 \rangle^k = \begin{bmatrix} F_{k+1} & F_k \\ F_k & F_{k-1} \end{bmatrix}$$

where $\{F_i\}_{i \geq 0}$ is the Fibonacci sequence defined by: $F_0 = 0, F_1 = 1$ and $F_{i+1} = F_i + F_{i-1}$ $(i \geq 1)$.

(b) Let ϕ be the positive root of the equation $X^2 - X - 1 = 0$ (so $\phi = (1 + \sqrt{5})/2 = 1.618...$). Prove inductively that

$$(1.5)^{k-1} \leq F_{k+1} \leq \phi^k.$$

(c) Say $(q_1, q_2, \ldots, q_k)$ is the quotient sequence associated to the remainder sequence of $a > b > 0$. If $\langle q_1, \ldots, q_k \rangle = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$ prove that

$$\|p\| \leq \|a\| - \|b\|.$$

(d) Conclude that $k < 1 + \log_{1.5} a$.
(e) (Lamé) Give an exact worst case bound on k in terms of ϕ and its conjugate $\widehat{\phi}$. □

§6. Polynomial HGCD

We describe the polynomial HGCD algorithm and prove its correctness.

Parallel reduction. The idea we exploit in the HGCD algorithm might be called "parallel reduction". Suppose we want to compute the HGCD of the pair of polynomials $A, B \in F[X]$ where $\deg A = 2m$. First we truncate these polynomials to define

$$\begin{bmatrix} A_0 \\ B_0 \end{bmatrix} := \begin{bmatrix} A \,\mathbf{div}\, X^m \\ B \,\mathbf{div}\, X^m \end{bmatrix}. \tag{23}$$

Suppose that R is the matrix returned by $HGCD(A_0, B_0)$; so

$$\begin{bmatrix} A_0 \\ B_0 \end{bmatrix} \xrightarrow{R} \begin{bmatrix} A_0' \\ B_0' \end{bmatrix} \tag{24}$$

for some A_0', B_0'. Then we can define A', B' via

$$\begin{bmatrix} A \\ B \end{bmatrix} \xrightarrow{R} \begin{bmatrix} A' \\ B' \end{bmatrix}. \tag{25}$$

Two reductions by the same matrix are said to be *parallel*. Thus (24) and (25) are parallel reductions. If A', B' turn out to be two consecutive terms in the remainder sequence of A, B, then we may have gained something! This is because we had computed R without looking at the lower order coefficients of A, B. But we need another property for R to be useful. We want the degrees of A', B' to straddle a sufficiently small value below $2m$. By definition of HGCD, the degrees of A_0', B_0' straddle $m/2$. A reasonable expectation is that the degrees of A', B' straddle $3m/2$. This would be the case if we could, for instance, prove that

$$\deg(A') = m + \deg(A_0'), \qquad \deg(B') = m + \deg(B_0').$$

This is not quite correct, as we will see. But it will serve to motivate the following outline of the HGCD algorithm.

Outline. Given A, B with $\deg(A) = 2m > \deg(B) > 0$, we recursively compute $R \leftarrow HGCD(A_0, B_0)$ as above. Now use R to carry out the reduction of (A, B) to (A', B'). Note that although the degrees of A', B' straddle $3m/2$, we have no upper bound on the degree of A'. Hence we perform one step of the Euclidean reduction:

$$\begin{bmatrix} A' \\ B' \end{bmatrix} \xrightarrow{\langle Q \rangle} \begin{bmatrix} C \\ D \end{bmatrix}.$$

Now the degree of $C = B'$ is less than $3m/2$. We can again truncate the polynomials C, D via

$$\begin{bmatrix} C_0 \\ D_0 \end{bmatrix} := \begin{bmatrix} C \,\mathbf{div}\, X^k \\ D \,\mathbf{div}\, X^k \end{bmatrix}.$$

for a certain $k \geq 0$. Intuitively, we would like to pick $k = m/2$. Then we make our second recursive call to compute $S \leftarrow HGCD(C_0, D_0)$. We use S to reduce (C, D) to (C', D'). Hopefully, the degrees of C' and D' straddle m, which would imply that our output matrix is

$$R \cdot \langle Q \rangle \cdot S$$

The tricky part is that k cannot be simply taken to be $m/2$. This choice is correct only if the remainder sequence is normal, as Moenck assumed. Subject to a suitable choice of k, we have described the HGCD algorithm.

We are ready to present the actual algorithm. We now switch back to the norm notation, $\|A\|$ instead of $\deg(A)$, to conform to the general framework.

Algorithm Polynomial HGCD(A, B):
Input: A, B are univariate polynomials with $\|A\| > \|B\| \geq 0$.
Output: a regular matrix M which reduces (A, B) to (C', D') where $\|C'\|, \|D'\|$ straddle $\|A\|/2$.

[1] $m \leftarrow \left\lceil \frac{\|A\|}{2} \right\rceil$; {This is the magic threshold}
if $\|B\| < m$ then return(E);
[2] $\begin{bmatrix} A_0 \\ B_0 \end{bmatrix} \leftarrow \begin{bmatrix} A \,\mathbf{div}\, X^m \\ B \,\mathbf{div}\, X^m \end{bmatrix}$.
{now $\|A_0\| = m'$ where $m + m' = \|A\|$}
$R \leftarrow \texttt{hGCD}(A_0, B_0)$;
{$\left\lceil \frac{m'}{2} \right\rceil$ is the magic threshold for this recursive call}
$\begin{bmatrix} A' \\ B' \end{bmatrix} \leftarrow R^{-1} \begin{bmatrix} A \\ B \end{bmatrix}$;
[3] if $\|B'\| < m$ then return(R);
[4] $Q \leftarrow A' \,\mathbf{div}\, B'$; $\begin{bmatrix} C \\ D \end{bmatrix} \leftarrow \begin{bmatrix} B' \\ A' \,\mathbf{mod}\, B' \end{bmatrix}$;
[5] $l \leftarrow \|C\|$; $k \leftarrow 2m - l$; {now $l - m < \left\lceil \frac{m'}{2} \right\rceil$}
[6] $C_0 \leftarrow C \,\mathbf{div}\, X^k$; $D_0 \leftarrow D \,\mathbf{div}\, X^k$; {now $\|C_0\| = 2(l - m)$}
$S \leftarrow \texttt{hGCD}(C_0, D_0)$;
{$l - m$ is magic threshold for this recursive call.}
[7] $M \leftarrow R \cdot \langle Q \rangle \cdot S$; return$(M)$;

The programming variables in this algorithm are illustrated in the following figure.

To prove the correctness of this lemma, we must show that the output matrix M satisfies

$$\begin{bmatrix} A \\ B \end{bmatrix} \xrightarrow{M} \begin{bmatrix} C' \\ D' \end{bmatrix}, \qquad \|C'\| \geq \left\lceil \frac{\|A\|}{2} \right\rceil > \|D'\|. \tag{26}$$

The Basic Setup. Let $A, B \in F[X]$, $\|A\| > \|B\| \geq 0$ and $m \geq 1$ be given. Define A_0, B_0 as in equation (23). This determines A_1, B_1 via the equation

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} A_0 X^m + A_1 \\ B_0 X^m + B_1 \end{bmatrix} = \begin{bmatrix} A_0 & A_1 \\ B_0 & B_1 \end{bmatrix} \begin{bmatrix} X^m \\ 1 \end{bmatrix}. \tag{27}$$

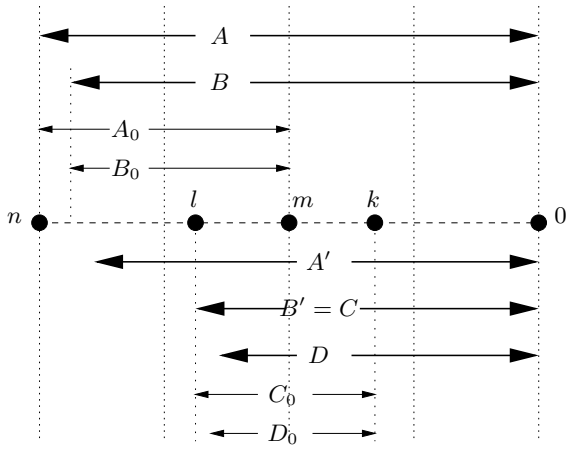


Figure 1: Variables in the polynomial HGCD algorithm.

Now let M be any given regular matrix. This determines A'_0, B'_0, A'_1, B'_1 via

$$\begin{bmatrix} A'_0 & A'_1 \\ B'_0 & B'_1 \end{bmatrix} := M^{-1} \begin{bmatrix} A_0 & A_1 \\ B_0 & B_1 \end{bmatrix}. \tag{28}$$

Finally, define A', B' via

$$\begin{bmatrix} A' \\ B' \end{bmatrix} := \begin{bmatrix} A'_0 & A'_1 \\ B'_0 & B'_1 \end{bmatrix} \begin{bmatrix} X^m \\ 1 \end{bmatrix}. \tag{29}$$

Hence we have the "parallel" reductions,

$$\begin{bmatrix} A_0 \\ B_0 \end{bmatrix} \xrightarrow{M} \begin{bmatrix} A'_0 \\ B'_0 \end{bmatrix}, \quad \begin{bmatrix} A \\ B \end{bmatrix} \xrightarrow{M} \begin{bmatrix} A' \\ B' \end{bmatrix}.$$

Lemma 9 (Correctness Criteria) *Let A, B, m, M be given, as in the Basic Setup, and define the remaining notations $A_i, B_i, A'_i, B'_i, A', B'$ ($i = 0, 1$) as indicated. If*

$$\|A'_0\| > \|B'_0\|, \tag{30}$$
$$\|A_0\| \leq 2\|A'_0\| \tag{31}$$

then

$$\|A'\| = m + \|A'_0\|, \qquad \|B'\| \leq m + \max\{\|B'_0\|, \|A_0\| - \|A'_0\| - 1\}.$$

In particular,

$$\|A'\| > \|B'\|.$$

Proof. Let $M = \begin{bmatrix} P & Q \\ R & S \end{bmatrix}$. First observe that $\|A'_0\| > \|B'_0\|$ and $A_0 = A'_0P + B'_0Q$ implies $\|A_0\| = \|A'_0\| + \|P\|$. Hence (31) is equivalent to

$$\|P\| \leq \|A'_0\|. \tag{32}$$

Since $M^{-1} = \pm \begin{bmatrix} S & -Q \\ -R & P \end{bmatrix}$ and $A'_1 = \pm(A_1S - B_1Q)$,

$$\|A'_1\| \leq \max\{\|A_1S\|, \|B_1Q\|\} < m + \|P\| \leq m + \|A'_0\|$$

Hence $A' = A'_0 X^m + A'_1$ implies $\|A'\| = m + \|A'_0\|$, as desired.

From $B'_1 = \pm(-A_1 R + B_1 P)$ we get $\|B'_1\| \leq m - 1 + \|P\| = m - 1 + \|A_0\| - \|A'_0\|$. From $B' = B'_0 X^m + B'_1$ we get the desired inequality $\|B'\| \leq m + \max\{\|B'_0\|, \|A_0\| - \|A'_0\| - 1\}$. **Q.E.D.**

We call the requirement (31) the (lower) "threshold" for $\|A'_0\|$. This threshold is the reason for the lower bound on $\|C'\|$ in the HGCD output specification (26).

Finally we prove the correctness of the HGCD algorithm.

Lemma 10 (HGCD Correctness) *Algorithm HGCD is correct: with input polynomials A, B where $\|A\| > \|B\| \geq 0$, it returns a regular matrix M satisfying (26).*

Proof. To keep track of the proof, the following sequence of reductions recalls the notations of the algorithm:

$$\begin{bmatrix} A \\ B \end{bmatrix} \xrightarrow{R} \begin{bmatrix} A' \\ B' \end{bmatrix} \xrightarrow{\langle Q \rangle} \begin{bmatrix} C \\ D \end{bmatrix} \xrightarrow{S} \begin{bmatrix} C' \\ D' \end{bmatrix}. \tag{33}$$

The algorithm returns a matrix in steps [1], [3] or [7]. Only when the algorithm reaches step [7] does the full sequence (33) take effect. It is clear that the returned matrix is always regular. So it remains to check the straddling condition of equation (26). In step [1], the result is clearly correct.

Consider the matrix R returned in step [3]: the notations m, A_0, B_0, A', B' in the algorithm conform to those in Correctness Criteria (lemma 9), after substituting R for M. By induction hypothesis, the matrix R returned by the first recursive call (step [2]) satisfies

$$\|A'_0\| \geq \lceil m'/2 \rceil > \|B'_0\|, \qquad (m' = \|A_0\|)$$

where $\begin{bmatrix} A_0 \\ B_0 \end{bmatrix} \xrightarrow{R} \begin{bmatrix} A'_0 \\ B'_0 \end{bmatrix}$. Then lemma 9 implies $\|A'\| = m + \|A'_0\| \geq m$. Since $m > \|B'\|$ is a condition for exit at step [3], it follows that the straddling condition (26) is satisfied at this exit.

Finally consider the matrix M returned in step [7]. Since we did not exit in step [3], we have $m \leq \|B'\|$. In step [4] we form the quotient Q and remainder D of A' divided by B'. Also we renamed B' to C. Hence $m \leq l$ where $l = \|C\|$. To see that C_0 is properly computed, let us verify

$$l \geq k \geq 0. \tag{34}$$

The first inequality in (34) follows from $l \geq m \geq m + (m - l) = k$. To see the second, $l = \|B'\| \leq m + \max\{\|B'_0\|, \|A_0\| - \|A'_0\| + 1\}$ (Correctness Criteria) and so $l \leq m + \max\{\lceil m'/2 \rceil - 1, \lfloor m'/2 \rfloor + 1\} \leq m + \lfloor m'/2 \rfloor + 1$. Thus $l - m \leq \lfloor m'/2 \rfloor + 1 \leq m$. Hence $k = m - (l - m) \geq 0$, proving (34). In the second recursive call, HGCD(C_0, D_0) returns S. By induction,

$$\|C'_0\| \geq \lceil \|C_0\|/2 \rceil > \|D'_0\|, \qquad \text{where } \begin{bmatrix} C_0 \\ D_0 \end{bmatrix} \xrightarrow{S} \begin{bmatrix} C'_0 \\ D'_0 \end{bmatrix}. \tag{35}$$

But $\|C_0\| = l - k = 2(l - m)$ so (35) becomes

$$\|C'_0\| \geq l - m > \|D'_0\|.$$

Now let $\begin{bmatrix} C \\ D \end{bmatrix} \xrightarrow{S} \begin{bmatrix} C' \\ D' \end{bmatrix}$. Another application of lemma 9 shows that

$$\|C'\| = k + \|C'_0\| \geq k + l - m = m$$

and

$$\begin{aligned} \|D'\| &\leq k + \max\{\|D'_0\|, \|C_0\| - \|C'_0\| - 1\} \\ &\leq k + \max\{l - m - 1, l - m - 1\} \\ &= k + l - m - 1 = m - 1. \end{aligned}$$

This shows $\|C'\| \geq m > \|D'\|$ and hence (26). **Q.E.D.**

Remark: The proof shows we could have used $k \leftarrow 2m - l - 1$ as well. Furthermore, we could modify our algorithm so that after step [4], we return $R \cdot \langle Q \rangle$ in case $\|D\| < m$. This may be slightly more efficient.

Complexity analysis. The HGCD algorithm makes two recursive calls to itself, $\texttt{hGCD}(A_0, B_0)$ and $\texttt{hGCD}(C_0, D_0)$. We check that $\|A_0\|$ and $\|C_0\|$ are both bounded by $n/2$. The work in each call to the algorithm, exclusive of recursion, is $O(M_B(n)) = O(n \log n)$. Hence the algebraic complexity $T(n)$ of this HGCD algorithm satisfies

$$T(n) = 2T(n/2) + O(n \log n).$$

This yields $T(n) = O(n \log^2 n)$.

Exercises

Exercise 6.1: Generalize the HGCD problem to the following: the function $\text{FGCD}(A, B, f)$ whose arguments are polynomials A, B as in the HGCD problem, and f is a rational number between 0 and 1. $\text{FGCD}(A, B, f)$ returns a matrix M that reduces the pair (A, B) to (A', B') such that $\|A'\|, \|B'\|$ straddle $f\|A\|$. Thus $\text{FGCD}(A, B, 1/2) = \texttt{hGCD}(A, B)$. Show that FGCD can be computed in the same complexity as `hGCD` by using `hGCD` as a subroutine. □

Exercise 6.2: Modify the polynomial HGCD algorithm so that in step [5], the variable k is set to $\lceil m/2 \rceil$. This is essentially the algorithm of Moenck-Aho-Hopcroft-Ullman [2]. We want to construct inputs to make the algorithm return wrong answers. Note that since the modified algorithm works for inputs with normal remainder sequence (see §3), we are unlikely to find such inputs by generating random polynomials. Suppose the output M reduces the input (A, B) to (A', B'). The matrix M may be wrong for several reasons:
(i) $\|B'\| \geq \lceil \|A\|/2 \rceil$.
(ii) $\|A'\| < \lceil \|A\|/2 \rceil$.
(iii) A', B' are not consecutive entries of the Euclidean remainder sequence of A, B.
Construct inputs to induce each of these possibilities. (The possibilities (i) and (ii) are known to occur.) □

§A. APPENDIX: Integer HGCD

For the sake of completeness, we present an integer version of the HGCD algorithm. We initially use two simple tricks. The first is to recover the non-Archimedean property thus: for $a, b \in \mathbb{Z}$,

$$ab \leq 0 \qquad \Longrightarrow \qquad \|a+b\| \leq \max\{\|a\|, \|b\|\}.$$

One consequence of the non-Archimedean property we exploited was that if $\|a\| \neq \|b\|$ then $\|a+b\| = \max\{\|a\|, \|b\|\}$. Here is an integer analogue:

$$\|a\| - \|b\| \geq 1 \qquad \Longrightarrow \qquad \|a+b\| = \|a\| \pm \epsilon, \quad (0 \leq \epsilon \leq 1).$$

To carry out a parallel reduction, the integer analogue would perhaps be to call HGCD on $a \,\mathbf{div}\, 2^m, b \,\mathbf{div}\, 2^m$ for suitable m. Instead, the second trick will call HGCD on

$$a_0 := 1 + (a \,\mathbf{div}\, 2^m), \quad b_0 := b \,\mathbf{div}\, 2^m. \tag{36}$$

The Basic Setup. We begin by proving the analogue of the Correctness Criteria (lemma 9). The following notations will be fixed for the next two lemmas.

Assume that we are given $a > b > 0$ and $m \geq 1$ where $a \geq 2^m$. This determines the non-negative values a_0, a_1, b_0, b_1 via

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a_0 & -a_1 \\ b_0 & b_1 \end{bmatrix} \begin{bmatrix} 2^m \\ 1 \end{bmatrix}, \quad 0 < a_1 \leq 2^m, \quad 0 \leq b_1 < 2^m. \tag{37}$$

Note that both tricks are incorporated in (37). Defining a_0, b_0 as in (36) is the same as choosing $a_1 := 2^m - (a \,\mathbf{mod}\, 2^m)$ and $b_1 := b \,\mathbf{mod}\, 2^m$. This choice ensures $a_0 > b_0$, as we assume in the recursive call to the algorithm on a_0, b_0.

We are also given a regular matrix M. This determines the values $a'_0, b'_0, a'_1, b'_1, a', b'$ via

$$\begin{bmatrix} a'_0 & a'_1 \\ b'_0 & b'_1 \end{bmatrix} := M^{-1} \begin{bmatrix} a_0 & -a_1 \\ b_0 & b_1 \end{bmatrix} \tag{38}$$

and

$$\begin{bmatrix} a' \\ b' \end{bmatrix} := \begin{bmatrix} a'_0 & a'_1 \\ b'_0 & b'_1 \end{bmatrix} \begin{bmatrix} 2^m \\ 1 \end{bmatrix}. \tag{39}$$

Hence we have the parallel reductions

$$\begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \xrightarrow{M} \begin{bmatrix} a'_0 \\ b'_0 \end{bmatrix}, \quad \begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M} \begin{bmatrix} a' \\ b' \end{bmatrix}.$$

Finally, we assume two key inequalities:

$$a'_0 > b'_0 \geq 0 \tag{40}$$

$$2\|a'_0\| - 2 \geq \|a_0\| \tag{41}$$

Now write M as

$$M = \begin{bmatrix} p & q \\ r & s \end{bmatrix}, \qquad M^{-1} = \delta \begin{bmatrix} s & -q \\ -r & p \end{bmatrix} \tag{42}$$

where $\delta = \det M = \pm 1$. From (38) we obtain

$$a'_1 = -\delta(sa_1 + qb_1), \quad b'_1 = \delta(ra_1 + pb_1). \tag{43}$$

The proof below uses (43) to predict the signs of a'_1, b'_1, assuming the sign of δ. This is possible thanks to the second trick.

The following is the integer analogue of lemma 9:

Lemma 11 (Partial Correctness Criteria)
With the Basic Setup:

(–) *Suppose* $\det M = -1$.
(–a) $\|a'\| = m + \|a'_0\| + \epsilon_1$, $(0 \le \epsilon_1 < 1)$.
(–b) $\|b'\| \le m + \max\{\|b'_0\|, \|a_0\| - \|a'_0\| + 1\}$.
Moreover, $\|a'\| > \|b'\|$.

(+) *Suppose* $\det M = +1$.
(+a) $\|a'\| = m + \|a'_0\| - \epsilon_2$, $(0 \le \epsilon_2 < 1)$.
(+b) $\|b'\| \le 1 + m + \max\{\|b'_0\|, \|a_0\| - \|a'_0\| + 1\}$.
Furthermore $b' \ge 0$.

In both cases (–) *and* (+), $a' > 0$.

Proof. Since $a_0 = pa'_0 + qb'_0$, the ordering property (20) and (40) yields

$$\|a_0\| = \|p\| + \|a'_0\| + \epsilon_3 \quad (0 \le \epsilon_3 < 1).$$

Hence (41) is equivalent to

$$\|p\| + \epsilon_3 \le \|a'_0\| - 2.$$

We now prove cases (–) and (+) in parallel.

Part (a). From equation (43),

$$\begin{aligned} \|a'_1\| &\le \max\{\|sa_1\|, \|qb_1\|\} + 1 \\ &< \|p\| + m + 1 \qquad (\text{by (20)}, \|a_1\| \le m, \|b_1\| < m) \\ &\le \|a'_0\| + m - 1. \end{aligned}$$

Hence $\|a'_0 2^m\| > 1 + \|a'_1\|$ and so $a' = a'_0 2^m + a'_1 > 0$. This proves the desired $a' > 0$. If $\delta = -1$ then $a'_1 \ge 0$ (by equation (43)) and hence $\|a'\| = m + \|a'_0\| + \epsilon_1$ as required by subcase (–a). On the other hand, if $\delta = +1$ then $a'_1 \le 0$ and $a' = a'_0 2^m + a'_1 > a'_0 2^{m-1}$ and hence subcase (+a) follows.

Part (b). Again from (43),

$$\begin{aligned} \|b'_1\| &\le \max\{\|ra_1\|, \|pb_1\|\} + 1 \\ &\le \|p\| + m + 1 \\ &\le \|a_0\| - \|a'_0\| + m + 1. \end{aligned}$$

In case $\delta = +1$, $b'_1 \ge 0$ and hence $b' = b'_0 2^m + b'_1 \ge 0$, as desired. Also subcase (+b) easily follows. In case $\delta = -1$, $b'_1 \le 0$ and $b'_0 b'_1 \le 0$. This gives the non-Archimedean inequality:

$$\|b'\| \le \max\{\|b'_0 2^m\|, \|b'_1\|\},$$

which proves subcase (–b).

Finally, we must show that $\delta = -1$ implies $\|a'\| > \|b'\|$: this follows immediately from (40), (41) and subcases (–a) and (–b). **Q.E.D.**

To see inadequacies in the Partial Correctness Criteria, we state our precise algorithmic goal.

Integer HGCD Criteria: *On input $a > b \geq 0$, the integer HGCD algorithm outputs a regular matrix M such that*

$$a \leq 3 \Rightarrow \quad M = E \tag{44}$$

$$a \geq 4 \Rightarrow \|a'\| \geq 1 + \left\lceil \frac{\|a\|}{2} \right\rceil > \|b'\|, \quad a' > b' \geq 0 \tag{45}$$

where $\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M} \begin{bmatrix} a' \\ b' \end{bmatrix}$.

Note that if $a \geq 4$ then $\|a\| \geq 1 + \lceil \|a\|/2 \rceil$ and so the desired matrix M exists.

Discussion. We see that the pair a', b' obtained in the Partial Correctness Criteria lemma may fail two properties needed for our HGCD algorithm:
Case $\det M = -1$: b' can be negative.
Case $\det M = +1$: the inversion $b' \geq a'$ may occur.
Clearly these two failures are mutually exclusive. On deeper analysis, it turns out that we only have to modify M slightly to obtain some regular matrix M^* such that

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M^*} \begin{bmatrix} a^* \\ b^* \end{bmatrix}$$

and a^*, b^* satisfy the correctness criteria, $\|a^*\| \geq m > \|b^*\|$, $a^* > b^* \geq 0$. The "Fixing Up lemma" below shows how to do this. The fixing up is based on three simple transformations of regular matrices: advancing, backing up and toggling.

In the following, let $a > b > 0$ and $M = \langle q_1, \ldots, q_k \rangle$ be a regular matrix such that

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M} \begin{bmatrix} a' \\ b' \end{bmatrix}.$$

(I) Advancing: If $q' = a' \,\mathbf{div}\, b'$, then we say that M has *advanced by one step* to the matrix $\langle q_1, \ldots, q_k, q' \rangle$. Note that this operation defines a regular matrix iff $q \geq 1$, *i.e.*, $\|a'\| \geq \|b'\|$. In general, we may speak of advancing M by more than one step.

(II) Backing Up: We call the matrix $\langle q_1, \ldots, q_{k-i} \rangle$ the *backing up* of M by i steps ($0 \leq i \leq k$); in case $i = 1$, we simply call it the *backup* of M. To do backing up, we need to recover the last partial quotient x from a regular matrix $M = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$. Note that $M = E$ if and only if $q = 0$, but in this case x is undefined. Hence assume $M \neq E$. Then M is elementary if and only if $s = 0$, and in this case $x = p$. So we next assume that M is not elementary. Write

$$M' = \begin{bmatrix} p' & q' \\ r' & s' \end{bmatrix}, \qquad M = M' \cdot \begin{bmatrix} x & 1 \\ 1 & 0 \end{bmatrix}$$

where $M' \neq E$ and $p = xp' + q', q = p'$. There are two cases. **Case of $q = 1$:** Clearly $p' = 1$. Since $p' \geq q' \geq 1$ (ordering property), we must have $q' = 1$. Hence x equals $p - 1$. **Case**

of $q > 1$: Then $p' > q'$ (there are two possibilities to check, depending on whether M' is elementary or not). This implies $x = p \,\mathbf{div}\, q$. In summary, the last partial quotient of M is given by

$$x = \begin{cases} \text{undefined} & \text{if } q = 0 \\ p & \text{if } s = 0 \\ p - 1 & \text{if } q = 1 \\ p \,\mathbf{div}\, q & \text{otherwise} \end{cases}$$

(III) Toggling: We call $T = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}$ the *toggle matrix*, so-called because T is idempotent ($T^2 = E$). The matrix MT is the *toggle of* M. We observe that MT is equal to $\langle q_1, \ldots, q_{k-1}, q_k - 1, 1\rangle$ in case $q_k > 1$, and $MT = \langle q_1, \ldots, q_{k-2}, q_{k-1} + 1\rangle$ in case $q_k = 1$ and $k > 1$. However, if $q_k = 1$ and $k = 1$, MT is not a regular matrix. In any case, we have

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{MT} \begin{bmatrix} a' + b' \\ -b' \end{bmatrix}.$$

Exercise A.1: Verify the remarks on the toggling matrix T. □

Lemma 12 (Fixing Up)
With the notations from the Basic Setup, let t be any number (the "fixup threshold") such that

$$\|a'_0\| \geq t > \max\{\|b'_0\|, \|a_0\| - \|a'_0\| + 1\}. \tag{46}$$

Moreover, if we write M as $\langle q_1, \ldots, q_k\rangle$ and M^* is as specified below, then

$$\|a^*\| \geq m + t > \|b^*\| \tag{47}$$

and

$$b^* \geq 0, \tag{48}$$

where $\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M^*} \begin{bmatrix} a^* \\ b^* \end{bmatrix}$. Here M^* is the regular matrix specified as follows:

(–) *Suppose* $\det M = -1$.
(–A) *If $b' \geq 0$ then $M^* := M$.*
(–B) *Else if $\|a' + b'\| \geq m + t$ then M^* is the toggle of M.*
(–C) *Else if $q_k \geq 2$ then $M^* := \langle q_1, \ldots, q_{k-1}, q_k - 1\rangle$ is the backup of the toggle of M.*
(–D) *Else M^* is the backing up of M by two steps.*

(+) *Suppose* $\det M = +1$.
(+A) *If $\|a'\| \leq \|b'\|$ then M^* is the advancement of $\langle q_1, \ldots, q_{k-1}\rangle$ by at most two steps.*
(+B) *Else if $\|a'\| < m + t$ then M^* is the backing up of M by one or two steps.*
(+C) *Else M^* is the advancement of M by at most two steps.*

Proof. The Partial Correctness Criteria lemma will be repeatedly exploited. First assume $\det M = -1$.

Subcase (–A). In this subcase, (48) is automatic, and (47) follows from case (–) of the Partial Correctness Criteria lemma.

Subcase (–B). In this subcase, $M^* = MT$ reduces $\begin{bmatrix} a \\ b \end{bmatrix}$ to $\begin{bmatrix} a^* \\ b^* \end{bmatrix} = \begin{bmatrix} a'+b' \\ -b' \end{bmatrix}$, as noted earlier. Recall that MT is not regular in case $k = 1$ and $q_1 = 1$. But if this were the case then

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

This implies $a' + b' = a > b = a'$ and so $b' > 0$, contradicting the fact that subcase (–A) has been excluded. Again, (48) is immediate, and (47) follows from case (–) of the Partial Correctness Criteria Lemma ($\|a^*\| = \|a' + b'\| \geq m + t$ by assumption).

Subcase (–C). In this subcase, M^* can be written as $M \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$, and so $\begin{bmatrix} a^* \\ b^* \end{bmatrix} = \begin{bmatrix} a' \\ a'+b' \end{bmatrix}$. We see that (48) holds by virtue of $a' + b' > 0$ (since $\|a'\| > \|b'\|, a' > 0$). Also (47) holds because the Partial Correctness Criteria lemma implies $\|a'\| \geq m + t$ and, since subcase (–B) fails, $\|a' + b'\| < m + t$.

Subcase (–D). Now $q_k = 1$ and M^* omits the last two partial quotients $(q_{k-1}, q_k) = (x, 1)$ where we write x for q_{k-1}. We ought to show $k \geq 2$, but this is the same argument as in subcase (–B). Hence $M^* = M \begin{bmatrix} 1 & -x \\ -1 & x+1 \end{bmatrix}$ and $\begin{bmatrix} a^* \\ b^* \end{bmatrix} = \begin{bmatrix} a'(x+1)+b'x \\ a'+b' \end{bmatrix}$. Hence $\|a^*\| = \|xb^* + a'\| > \|b^*\|$ and so a^*, b^* are consecutive elements of the remainder sequence of a, b. Then (48) holds because $a' + b' > 0$. To see (47), it is clear that $m + t > \|b^*\|$ (otherwise subcase (–B) applies) and it remains to show $\|a^*\| \geq m + t$. But this follows from $\|a^*\| = \|a' + x(a' + b')\| > \|a'\| \geq m + t$.

Now consider the case $\det M = +1$.

Subcase (+A). So there is inversion, $a' \leq b'$. Let us back up M to $\langle q_1, \ldots, q_{k-1} \rangle$:

$$\begin{bmatrix} a \\ b \end{bmatrix} \overset{\langle q_1,\ldots,q_{k-1}\rangle}{\longrightarrow} \begin{bmatrix} a'' \\ a' \end{bmatrix} \overset{\langle q_k \rangle}{\longrightarrow} \begin{bmatrix} a' \\ b' \end{bmatrix}.$$

Hence $a'' = a'q_k + b' > a'$. Thus a'', a' are consecutive members of the remainder sequence of a, b. Now $\|a''\| \geq \|2a'\| = \|a'\| + 1 > m + \|a'_0\| \geq m + t$. Also $\|a'\| \leq \|b'\| < 1 + m + t$ (by Partial Correctness Criteria). Therefore, if we advance $\langle q_1, \ldots, q_{k-1} \rangle$ by at most two steps, we would reduce a'', a' to a^*, b^* where $\|b^*\| < m + t$.

Subcase (+B). Now $a' > b' \geq 0$ and $\|a'\| < m + t$. So a', b' are consecutive members of the remainder sequence of a, b. Consider the entry $a'' = a'q_k + b'$ preceding a' in the remainder sequence of a, b. If $\|a''\| \geq m + t$, we are done since $\|a''\|, \|a'\|$ straddle $m + t$. Otherwise, consider the entry $a''' = a''q_{k-1} + a'$ preceding a''. We have

$$\|a'''\| = \|(a'q_k + b')q_{k-1} + a'\| \geq \|a'\| + 1 \geq m + t.$$

Thus $\|a'''\|, \|a''\|$ straddle $m + t$.

Subcase (+C). Again a', b' are consecutive members of the remainder sequence with $\|a'\| \geq m + t$. But $\|b'\| - 1 < m + t$ implies that if we advance M by at most two steps, the pair a', b' would be reduced to a^*, b^* where $\|b^*\| < m + t$.

Q.E.D.

It is interesting to note that in tests on randomly generated numbers of about 45 digits, subcase (–D) never arose.

The Fixup Procedure. The Fixing Up lemma and its proof provides a specific procedure to convert the tentative output matrix M of the HGCD algorithm into a valid one. To be specific, let

$$\texttt{Fixup}(M, a, b, m, t)$$

denote the subroutine that returns M^*, as specified in the Fixing Up lemma:

$$\begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{M^*} \begin{bmatrix} a^* \\ b^* \end{bmatrix}, \qquad \|a^*\| \geq m + t > \|b^*\|.$$

The correct behavior of the Fixup procedure depends on its input parameters fulfilling the conditions of the Fixing Up lemma. In particular, it must fulfil the conditions of the Basic Setup (mainly the inequalities (40) and (41)) and also the inequality (46).

In a typical invocation of $\texttt{Fixup}(M, a, b, m, t)$, the values M, a, b, m are available as in the Basic Setup. To pick a value of t, we use the fact that the following typically holds:

$$\|a'_0\| \geq 1 + \lceil \|a_0\|/2 \rceil > \|b'_0\| \tag{49}$$

(cf. (45)). In this case, it is easily verified that the choice $t = 1 + \lceil \|a_0\|/2 \rceil$ will satisfy (46). Of course inequality (49) also implies inequality (41).

However, our Fixup procedure may also be called in a situation when the Fixing Up lemma does not hold, namely, when $a_0 = 1 + (a \,\mathbf{div}\, 2^m) \leq 3$. In this case, no choice of t satisfying inequality (46) is possible. Note that $b_0 < a_0 \leq 3$ implies $b = b_0 2^m + b_1 < 3 \cdot 2^m$. It is easy to check that if we take at most three of the usual Euclidean remaindering steps, we reduce a, b to a^*, b^* where $\|a^*\| \geq m > \|b^*\|$. In such a situation, if we assume that the Fixup procedure is called with $M = E$ and $t = 0$, the returned matrix M^* is the advancement of E by at most three steps. More generally, if $\|a\|, \|b\|$ straddle $m + i$ where $i = 0, 1, 2$, and we call Fixup with the arguments

$$\texttt{Fixup}(E, a, b, m, 0),$$

we say this is the "easy fixup" case, because M^* is the advancement of E by at most 4 steps.

We present the integer HGCD algorithm.

ALGORITHM INTEGER HGCD(a, b):
Input: integers a, b with $a > b \geq 0$.
Output: a regular matrix M satisfying the integer HGCD criteria (44) or (45).

[1] $m \leftarrow 1 + \left\lceil \frac{\|a\|}{2} \right\rceil$; {this is the magic threshold}
if $a \leq 3$ or $\|b\| < m$ then return(E);
[2] $a_0 \leftarrow 1 + (a \operatorname{\mathbf{div}} 2^m)$; $b_0 \leftarrow b \operatorname{\mathbf{div}} 2^m$;
if $a_0 \leq 3$ then $t \leftarrow 0$ else $t \leftarrow 1 + \lceil \frac{\|a_0\|}{2} \rceil$;
$R \leftarrow$ `Fixup`(`hGCD`$(a_0, b_0), a, b, m, t)$;
$\begin{bmatrix} a' \\ b' \end{bmatrix} \leftarrow R^{-1} \begin{bmatrix} a \\ b \end{bmatrix}$;
[3] if $\|b'\| < m$ then return(R);
[4] $q \leftarrow a' \operatorname{\mathbf{div}} b'$; $\begin{bmatrix} c \\ d \end{bmatrix} \leftarrow \begin{bmatrix} b' \\ a' \operatorname{\mathbf{mod}} b' \end{bmatrix}$;
if $1 + (c \operatorname{\mathbf{div}} 2^m) \leq 3$ then return$(R \cdot$ `Fixup`$(E, c, d, m, 0))$;
[5] $l \leftarrow \lceil \|c\| \rceil$; $k \leftarrow 2m - l - 1$;
{Now $\|c\| \geq m + 1 \geq 4$. We claim $\|c\| - 1 \geq k \geq 0$}
[6] $c_0 \leftarrow 1 + (c \operatorname{\mathbf{div}} 2^k)$; $d_0 \leftarrow d \operatorname{\mathbf{div}} 2^k$;
if $c_0 \leq 3$ then $t' \leftarrow 0$ else $t' \leftarrow 1 + \left\lceil \frac{\|c_0\|}{2} \right\rceil$;
$S \leftarrow$ `Fixup`(`hGCD`$(c_0, d_0), c, d, k, t')$; {We claim $k + t' = m + 1$.}
[7] $\begin{bmatrix} c' \\ d' \end{bmatrix} \leftarrow S^{-1} \begin{bmatrix} c \\ d \end{bmatrix}$; {So $\|c'\|, \|d'\|$ straddle $k + t'$}
$T \leftarrow$ `Fixup`$(E, c', d', m, 0)$;
$M \leftarrow R \cdot \langle q \rangle \cdot S \cdot T$; return$(M)$;

Correctness. Procedure `hGCD` returns in four places (steps [1], [3], [4] and [7]) in the algorithm. We show that the matrix returned at each of these places is correct. Since these matrices are regular, we basically have to check the straddling property (45) when $a \geq 4$. We will also need to check that each call to the Fixup procedure is proper.

a) In case the algorithm returns the identity matrix E in step [1], the correctness is trivial.

b) In step [2], we must check that the proper conditions are fulfilled for calling `Fixup`. When $a_0 \leq 3$ we have the "easy fixup" case. Otherwise $a_0 \geq 4$ and the first recursive call in `hGCD` returns some regular matrix R' which is fixed up as R by `Fixup`. The conditions of the Basic Setup are fulfilled with a, b, m as usual and $M = R'$. If $\begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \xrightarrow{R'} \begin{bmatrix} a_0' \\ b_0' \end{bmatrix}$, then inductively, the correctness of the HGCD procedure implies equation (49) (and hence (41)) holds. As discussed following equation (49), the choice $t = 1 + \lceil \|a_0\|/2 \rceil$ then satisfies (46),

c) Suppose the matrix R is returned at step [3]. This is correct since $\|a'\|, \|b'\|$ straddle $m + t$ (by correctness of the Fixup procedure) and the condition for exit is $\|b'\| < m$.

d) In step [4], the call to `Fixup` is the "easy fixup" case since $\|c\| \geq m$ and $\|c\| \leq m + 2$. The returned matrix is clearly correct.

e) Suppose we reach step [5]. We show the claim $\|c\| - 1 \geq k \geq 0$. We have $\|c\| - 1 \geq m - 1 \geq (m - 1) + (m - l) = k$. Next, $k \geq 0$ is equivalent to $2m - 1 \geq l$. This follows from:

$$l = \lceil \|b'\| \rceil \quad \leq \quad m + t \qquad \text{(from the first FIXUP)}$$

$$\begin{aligned}
&= m+1+\left\lceil \frac{\|1+(a \,\mathbf{div}\, 2^m)\|}{2} \right\rceil \\
&\le m+1+\left\lceil \frac{1+\lfloor \|a \,\mathbf{div}\, 2^m\| \rfloor}{2} \right\rceil \quad (\text{since } \|1+x\| \le 1+\lfloor \|x\| \rfloor) \\
&= m+1+\left\lceil \frac{1+\lfloor \|a/2^m\| \rfloor}{2} \right\rceil \quad (\text{since } \lfloor \|x \,\mathbf{div}\, y\| \rfloor = \lfloor \|x/y\| \rfloor) \\
&= m+1+\left\lceil \frac{1+\lfloor \|a\| \rfloor - m}{2} \right\rceil \\
&\le m+\lceil (m-1)/2 \rceil \quad (\text{since } m = 1+\lceil \|a\|/2 \rceil) \\
&\le 2m-1 \qquad (m \ge 2)
\end{aligned}$$

f) The call to `Fixup` in step [6] fulfills the appropriate conditions. [Reasoning as before: note that $\|c\|-1 \ge k$ implies that $c_0 \ge 3$. Hence, the "easy fixup" case occurs iff $c_0 = 3$. Otherwise, the Basic Setup conditions prevail with a, b, m, M in the Basic Setup replaced by $c, d, k, \texttt{hGCD}(c_0, d_0)$.] Next we prove the claim $k + t' = m + 1$:

$$\begin{aligned}
t' &= 1+\lceil \|c_0\|/2 \rceil \\
&= 1+\left\lceil \frac{\lceil \|c_0\| \rceil}{2} \right\rceil \\
&= 1+\left\lceil \frac{\lceil \|\epsilon + (c/2^k)\| \rceil}{2} \right\rceil \quad (0 < \epsilon \le 1, \quad c_0 = 1+\lfloor c/2^k \rfloor) \\
&= 1+\left\lceil \frac{l-k+\delta}{2} \right\rceil \qquad (\delta = 0 \text{or } 1, \quad l = \lceil \|c\| \rceil) \\
&= 1+\left\lceil \frac{2l-2m+1+\delta}{2} \right\rceil \quad (k = 2m-l-1) \\
&= 2+l-m.
\end{aligned}$$

Thus $k + t' = k + (2 + l - m) = m + 1$, as desired.

g) We reach step [7]. By the correctness of the Fixup procedure, $\|c'\|, \|d'\|$ straddle $k + t' = m + 1$. Hence we have the right conditions for the "easy fixup" case. The final output is clearly correct.

This concludes our correctness proof.

Computational details and analysis. The algorithm has to perform comparisons of the kind

$$\|a\| : m,$$

and compute ceiling functions in the special forms

$$\lceil \|a\| \rceil, \qquad \lceil \|a\|/2 \rceil,$$

where a, m are positive integers. (In the algorithm a may be zero, but we can treat those as special cases.) Since $\|a\|$ is generally not rational, we do not want to explicitly compute it. Instead we reduce these operations to integer comparisons, checking if a is a power of two, and to computing the function $\texttt{bit}(a)$, which is defined to be the number of bits in the binary representation of a positive integer a. So $\texttt{bit}(a) = 1 + \lfloor \log_2 a \rfloor$ and clearly this function is easily computed in linear time in the usual Turing machine model. Then we have

$$\|a\| \ge m \Leftrightarrow \texttt{bit}(a) - 1 \ge m$$

and

$$\|a\| > m \Leftrightarrow \begin{cases} \texttt{bit}(a) - 1 > m & \text{if } a \text{ is a power of 2} \\ \texttt{bit}(a) > m & \text{else.} \end{cases}$$

and finally,

$$\left\lceil \frac{\|a\|}{2} \right\rceil = \begin{cases} \left\lceil \frac{\texttt{bit}(a)-1}{2} \right\rceil & \text{if } a \text{ is a power of 2} \\ \left\lceil \frac{\texttt{bit}(a)}{2} \right\rceil & \text{else} \end{cases}$$

The global structure of the complexity analysis is similar to the polynomial HGCD case: with $M_B(n) = n\mathcal{L}(n)$ denoting as usual the bit complexity of integer multiplication, it is not hard to see that `Fixup` takes time $O(M_B(n))$, under the conditions stipulated for its invocation. In the two recursive calls to `hGCD`, it is easy to check that the integers have bit size $\frac{n}{2} + O(1)$. Hence, if $T(n)$ is the bit complexity of our HGCD algorithm on inputs of size at most n, then

$$T(n) = O(M_B(n)) + 2T(\frac{n}{2} + O(1)).$$

This has solution $T(n) = O(M_B(n) \log n) = n\mathcal{L}^2(n)$.

Exercises

Exercise A.2:
(a) Verify the remarks on reducing operations involving $\|a\|$ to integer operations, the function $\texttt{bit}(a)$ and testing if a is a power of 2.
(b) Derive the time bound $T(n) = n\mathcal{L}^2(n)$ for the HGCD algorithm. □

Exercise A.3: Try to simplify the integer HGCD algorithm by separating the truncation value t (as in $a_0 := a \,\mathbf{div}\, 2^t$) from the straddling value s (as in $\|a'\| \geq s > \|b'\|$). Currently $t = s = 1 + \lceil \|a\|/2 \rceil$. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture III
Subresultants

We extend the Euclidean algorithm to the polynomial ring $D[X]$ where D is a unique factorization domain. The success of this enterprise depends on the theory of subresultants. Subresultant sequences are special remainder sequences which have many applications including Diophantine equations, Sturm theory, elimination theory, discriminants, and algebraic cell decompositions. Our approach to subresultants follows Ho and Yap [84, 83], who introduced the pseudo-subresultants to carry out Loos' program [119] of studying subresultants via specialization from the case of indeterminate coefficients. This approach goes back to Habicht [76].

One of the most well-studied problems in the early days of computer algebra (circa 1970) is the problem of computing the GCD in the polynomial ring $D[X]$ where D is a UFD. See the surveys of Loos [33] and Collins [45]. This led to a development of efficient algorithms whose approach is quite distinct from the HGCD approach of the previous lecture. The reader may be surprised that any new ideas are needed: why not use the previous techniques to compute the GCD in $Q_D[X]$ (where Q_D is the quotient field of D) and then "clear denominators"? One problem is that computing remainders in $Q_D[X]$ can be quite non-trivial for some D (say, $D = F[X_1, \ldots, X_d]$). Another problem is that clearing denominators is really a multiple GCD computation (in its dual form of a multiple LCM computation). Multiple GCD is expensive in practical terms, even when it is polynomial-time as in the case $D = \mathbb{Z}$. Worst, in case $D = F[X_1, \ldots, X_d]$, we are involved in a recursive situation of exponential complexity. Hence the challenge is to develop a direct method avoiding the above problems.

In this lecture, D refers to a unique factorization domain with quotient field Q_D. The reader may safely take $D = \mathbb{Z}$ and so $Q_D = \mathbb{Q}$.

§1. Primitive Factorization

The goal of this section is to extend the arithmetic structure of a unique factorization domain D to its quotient field Q_D and to $Q_D[X]$. It becomes meaningful to speak of irreducible factorizations in Q_D and $Q_D[X]$.

Content of a polynomial. Let $q \in D$ be an irreducible element. For any non-zero element $a/b \in Q_D$ where $a, b \in D$ are relatively prime, define the *q-order* of a/b to be

$$\mathtt{ord}_q(a/b) := \begin{cases} n & \text{if } q^n|a \text{but not } q^{n+1}|a, \\ -n & \text{if } q^n|b \text{but not } q^{n+1}|b. \end{cases} \tag{1}$$

Exactly one of the two conditions in this equation hold unless $n = 0$. In this case, $\mathtt{ord}_q(a/b) = 0$ or equivalently, q does not divide ab. For example, in $D = \mathbb{Z}$ we have $\mathtt{ord}_2(4/3) = 2$, $\mathtt{ord}_2(3/7) = 0$ and $\mathtt{ord}_2(7/2) = -1$.

We extend this definition to polynomials. If $P \in Q_D[X] \setminus \{0\}$, define the q-order of P to be

$$\mathtt{ord}_q(P) := \min_i \{\mathtt{ord}_q(c_i)\}$$

where c_i ranges over all the non-zero coefficients of P. For example, $\mathtt{ord}_2(X^3 - \frac{5}{4}X + 2) = -2$ in $\mathbb{Z}[X]$. Any associate of $q^{\mathtt{ord}_q(P)}$ is called a *q-content* of P. Finally, we define a *content* of P to be

$$u \prod_q q^{\mathtt{ord}_q(P)}$$

where q ranges over all distinguished irreducible elements of D, u is any unit. This product is well-defined since all but finitely many $\mathtt{ord}_q(P)$ are zero. For the zero element, we define $\mathtt{ord}_q(0) = -\infty$ and the q-content and content of 0 are both 0.

Primitive polynomials. A polynomial of $Q_D[X]$ is *primitive* if it has content 1; such polynomials are elements of $D[X]$. Thus every non-zero polynomial P has a factorization of the form

$$P = cQ$$

where c is a content of P and Q is primitive. We may always choose Q so that its leading coefficient is distinguished. In this case, c is called *the* content of P, and Q the *primitive part* of P. These are denoted $\mathtt{cont}(P)$ and $\mathtt{prim}(P)$, respectively. We call the product expression "$\mathtt{cont}(P)\mathtt{prim}(P)$" the *primitive factorization* of P.

If $\mathtt{prim}(P) = \mathtt{prim}(Q)$, we say that P, Q are *similar* and denote this by

$$P \sim Q.$$

Hence $P \sim Q$ iff there exist $\alpha, \beta \in D$ such that $\alpha P = \beta Q$. In particular, if P, Q are associates then they are similar.

For instance, the following are primitive factorizations:

$$\begin{aligned} -4X^3 - 2X + 6 &= (-2) \cdot (2X^3 + X - 3) \\ (15/2)X^2 - (10/3)X + 5 &= (5/6) \cdot (9X^2 - 4X + 6). \end{aligned}$$

Also, $-4X^3 - 2X + 6 \sim 6X^3 + 3X - 9$.

The following is one form of a famous little lemma[1]:

Lemma 1 (Gauss' Lemma) *If D is a UFD and $P, Q \in D[X]$ are primitive, then so is their product PQ.*

Proof. We must show that for all irreducible $q \in D$, $\mathtt{ord}_q(PQ) = 0$. We can uniquely write any polynomial $P \in D[X]$ as

$$P = qP_0 + P_1, \qquad (P_0, P_1 \in D[X])$$

where $\deg(P_0)$ is less than the tail degree of P_1 and the tail coefficient $\mathtt{tail}(P_1)$ is not divisible by q. [If $\mathtt{tail}(P)$ is not divisible by q, then $P_0 = 0$ and $P_1 = P$.] Moreover,

$$\mathtt{ord}_q(P) = 0 \quad \text{iff} \quad P_1 \neq 0.$$

Thus $P_1 \neq 0$. Let $Q = qQ_0 + Q_1$ be the similar expression for Q and again $Q_1 \neq 0$. Multiplying the expressions for P and Q, we get an expression of the form

$$PQ = qR_0 + R_1, \qquad R_1 = P_1 Q_1 \neq 0.$$

[1] We refer to Edwards [63] for a deeper investigation of this innocuous lemma.

By the uniqueness of such expressions, we conclude that $\mathtt{ord}_q(PQ) = 0$. **Q.E.D.**

If $P_i = a_iQ_i$ $(i = 1, 2)$ is a primitive factorization, we have $\mathtt{cont}(P_1P_2) = \mathtt{cont}(a_1Q_1a_2Q_2) = a_1a_2\mathtt{cont}(Q_1Q_2)$ and $\mathtt{prim}(P_1P_2) = \mathtt{prim}(Q_1Q_2)$. By Gauss' lemma, $\mathtt{cont}(Q_1Q_1) = \epsilon$ and $\mathtt{prim}(Q_1Q_2) = \epsilon' Q_1Q_2$, where ϵ, ϵ' are units. Hence we have shown

Corollary 2 *For $P_1, P_2 \in Q_D[X]$,*

$$\mathtt{cont}(P_1P_2) = \epsilon \cdot \mathtt{cont}(P_1)\mathtt{cont}(P_2), \qquad \mathtt{prim}(P_1P_2) = \epsilon' \cdot \mathtt{prim}(P_1)\mathtt{prim}(P_2).$$

Another corollary to Gauss' lemma is this:

Corollary 3 *If $P(X) \in D[X]$ is primitive and $P(X)$ is reducible in $Q_D[X]$ then $P(X)$ is reducible in $D[X]$.*

To see this, suppose $P = QR$ with $Q, R \in Q_D[X]$. By the above corollary, $\mathtt{cont}(P) = \epsilon \cdot \mathtt{cont}(Q)\mathtt{cont}(R) = \epsilon''$ for some unit ϵ''. Then $P = \epsilon'' \cdot \mathtt{prim}(P)$. By the same corollary again,

$$P = \epsilon'' \cdot \epsilon' \cdot \mathtt{prim}(Q)\mathtt{prim}(R).$$

Since $\mathtt{prim}(Q), \mathtt{prim}(R)$ belongs to $D[X]$, this shows P is reducible.

We are ready to prove the non-trivial direction of the theorem in §II.1: *if D is a UFD, then $D[X]$ is a UFD.*

Proof. Suppose $P \in D[X]$ and without loss of generality, assume P is not an element of D. Let its primitive factorization be $P = aP'$. Clearly P' is a non-unit. We proceed to give a unique factorization of P' (as usual, unique up to reordering and associates). In the last lecture, we proved that a ring of the form $Q_D[X]$ (being Euclidean) is a UFD. So if we view P' as an element of $Q_D[X]$, we get a unique factorization, $P' = P'_1P'_2 \cdots P'_\ell$ where each P'_i is an irreducible element of $Q_D[X]$. Letting the primitive factorization of each P'_i be c_iP_i, we get

$$P' = c_1 \cdots c_\ell P_1 \cdots P_\ell.$$

But $c_1 \cdots c_\ell = \epsilon$ (some unit). Thus

$$P' = (\epsilon \cdot P_1)P_2 \cdots P_\ell$$

is a factorization of P' into irreducible elements of $D[X]$. The uniqueness of this factorization follows from the fact that $Q_D[X]$ is a UFD. If a is a unit, then

$$P = (a \cdot \epsilon \cdot P_1)P_2 \cdots P_\ell$$

gives a unique factorization of P. Otherwise, since D is a UFD, a has a unique factorization, say $a = a_1 \cdots a_k$. Then

$$P = a_1 \cdots a_k(\epsilon \cdot P_1)P_2 \cdots P_\ell$$

gives a factorization of P. It is easy to show uniqueness. **Q.E.D.**

The divide relation in a quotient field. We may extend the relation 'b divides c' to the quotient field of a UFD. For $b, c \in Q_D$, we say b *divides* c, written

$$b|c,$$

if for all irreducible q, either $0 \leq \mathtt{ord}_q(b) \leq \mathtt{ord}_q(c)$ or $\mathtt{ord}_q(c) \leq \mathtt{ord}_q(b) \leq 0$. Clearly Q_D is also a "unique factorization domain" whose irreducible elements are q, q^{-1} where $q \in D$ is irreducible. Hence the concept of `GCD` is again applicable and we extend our previous definition to Q_D in a natural way. We call b a *partial content* of P if b divides $\mathtt{cont}(P)$.

Exercises

Exercise 1.1: Assume that elements in Q_D are represented as a pair (a, b) of relatively prime elements of D. Reduce the problem of computing GCD in Q_D to the problem of GCD in D. □

Exercise 1.2: (Eisenstein's criterion) Let D be a UFD and $f(X) = \sum_{i=0}^{n} a_i X^i$ be a primitive polynomial in $D[X]$.
(i) If there exists an irreducible element $p \in D$ such that

$$\begin{aligned} a_n &\not\equiv 0 (\mathrm{mod}\, p), \\ a_i &\equiv 0 (\mathrm{mod}\, p) \qquad (i = 0, \ldots, n-1), \\ a_0 &\not\equiv 0 (\mathrm{mod}\, p^2), \end{aligned}$$

then $f(X)$ is irreducible in $D[X]$.
(ii) Under the same conditions as (i), conclude that the polynomial $g(x) = \sum_{i=0}^{n} a_i X^{n-i}$ is irreducible. □

Exercise 1.3: (i) $X^n - p$ is irreducible over $\mathbb{Q}[X]$ for all prime $p \in \mathbb{Z}$.
(ii) $f(X) = X^{p-1} + X^{p-2} + \cdots + X + 1$ $(= \frac{X^p - 1}{X - 1})$ is irreducible in $\mathbb{Q}[X]$ for all prime $p \in \mathbb{Z}$. HINT: apply Eisenstein's criterion to $f(X + 1)$.
(iii) Let ζ be a primitive 5-th root of unity. Then $\sqrt{5} \in \mathbb{Q}(\zeta)$.
(iv) The polynomial $g(X) = X^{10} - 5$ is irreducible over $\mathbb{Q}[X]$ but factors as $(X^5 - \sqrt{5})(X^5 + \sqrt{5})$ over $\mathbb{Q}(\zeta)[X]$. □

§2. Pseudo-remainders and PRS

Since $D[X]$ is a UFD, the concept of GCD is meaningful. It easily follows from the definitions that for $P, Q \in D[X]$,

$$\mathtt{cont}(\mathtt{GCD}(P, Q)) = \epsilon \cdot \mathtt{GCD}(\mathtt{cont}(P), \mathtt{cont}(Q)), \qquad (\epsilon = \text{unit}) \tag{2}$$

$$\mathtt{prim}(\mathtt{GCD}(P, Q)) = \mathtt{GCD}(\mathtt{prim}(P), \mathtt{prim}(Q)). \tag{3}$$

Thus the GCD problem in $D[X]$ can be separated into the problem of multiple GCD's in D (to extract contents and primitive parts) and the GCD of primitive polynomials in $D[X]$.

To emulate the Euclidean remaindering process for GCD's in $D[X]$, we want a notion of remainders. We use a basic observation, valid in any domain D, not just in UFD's. If $A, B \in D[X]$, we may still define $\mathtt{rem}(A, B)$ by treating A, B as elements of the Euclidean domain $Q_D[X]$. In general, $\mathtt{rem}(A, B) \in Q_D[X]$.

Lemma 4 (Pseudo-division Property) *In any domain D, if $A, B \in D[X]$ where $d = \deg A - \deg B \geq 0$ and $\beta = \texttt{lead}(B)$. Then $\texttt{rem}(\beta^{d+1}A, B)$ is an element of $D[X]$.*

Proof. By the division property in $Q_D[X]$, there exists $S, R \in Q_D[X]$ such that

$$A = BS + R, \qquad \deg R < \deg B. \tag{4}$$

Write $A = \sum_{i=0}^{m} a_i X^i$, $B = \sum_{i=0}^{n} b_i X^i$ and $S = \sum_{i=0}^{d} c_i X^i$. Then we see that

$$\begin{aligned} a_m &= b_n c_d, \\ a_{m-1} &= b_n c_{d-1} + b_{n-1} c_d, \\ a_{m-2} &= \cdots. \end{aligned}$$

From the first equation, we conclude that c_d can be written as $a_n/\beta = \alpha_0 \beta^{-1}$ ($\alpha_0 = a_n$). From the next equation, we further deduce that c_{d-1} can be written in the form $\alpha_1 \beta^{-2}$ for some $\alpha_1 \in D$. By induction, we deduce $c_{d-i} = \alpha_i \beta^{-(i+1)}$ for some $\alpha_i \in D$. Hence $\beta^{d+1} S \in D[X]$. Multiplying equation (4) by β^{d+1}, we conclude that $\texttt{rem}(\beta^{d+1}A, B) = \beta^{d+1}R$. The lemma follows since $\beta^{d+1}R = \beta^{d+1}A - B(\beta^{d+1}S)$ is an element of $D[X]$. **Q.E.D.**

So it is natural to define the *pseudo-remainder* of $P, Q \in D[X]$ as follows:

$$\texttt{prem}(P, Q) := \begin{cases} P & \text{if} \quad \deg P < \deg Q \\ \texttt{rem}(\beta^{d+1}P, Q) & \text{if} \quad d = \deg P - \deg Q \geq 0, \beta = \texttt{lead}(Q). \end{cases}$$

Pseudo-remainders are elements of $D[X]$ but they are not guaranteed to be primitive. We now generalize the concept of remainder sequences. A sequence of non-zero polynomials

$$(P_0, P_1, \ldots, P_k) \quad (k \geq 1)$$

is called a *polynomial remainder sequence* (abbreviated, PRS) of P, Q if $P_0 = P, P_1 = Q$ and

$$\begin{aligned} P_{i+1} &\sim \texttt{prem}(P_{i-1}, P_i) \quad (i = 2, \ldots, k-1) \\ 0 &= \texttt{prem}(P_{k-1}, P_k). \end{aligned}$$

If $d_i = \deg P_i$ for $i = 0, \ldots, k$, we call

$$(d_0, d_1, \ldots, d_k)$$

the *degree sequence* of the PRS. The degree sequence of a PRS is determined by the first two elements of the PRS. The PRS is *regular* if $d_i = 1 + d_{i+1}$ for $i = 1, \ldots, k-1$.

Discussion: We are usually happy to compute GCD's up to similarity. The concept of a PRS captures this indifference: the last term of a PRS is similar to the GCD of the first two terms. Consider how we might compute a PRS. Assuming we avoid computing in $Q_D[X]$, we are presented with several strategies. Here are two obvious ones:
(a) *Always maintain primitive polynomials.* Each step of the PRS algorithm is implemented by a pseudo-remainder computation followed by primitive factorization of the result.
(b) *Avoid all primitive factorizations until the last step.* Repeatedly compute pseudo-remainders, and at the end, extract the content with one primitive factorization.
Both strategies have problems. In case (a), we are computing multiple GCD at each step, which we said is too expensive. In case (b), the final polynomial can have exponentially large coefficients (this will be demonstrated below). In this lecture, we present a solution of G. E. Collins involving an interesting middle ground between strategies (a) and (b), *which is sufficient to avoid exponential growth of the coefficients without repeated multiple GCD computation.*

The PRS sequences corresponding to strategies (a) and (b) above are:

a) Primitive PRS This is a PRS $(P_0, \ldots, P_k)$ where each member (except possibly for the first two) is primitive:

$$P_{i+1} = \texttt{prim}(\texttt{prem}(P_{i-1}, P_i)) \qquad (i = 1, \ldots, k-1).$$

b) Pseudo-Euclidean PRS This is a PRS $(P_0, \ldots, P_k)$ where

$$P_{i+1} = \texttt{prem}(P_{i-1}, P_i) \qquad (i = 1, \ldots, k-1).$$

The following illustrates the explosive coefficient growth in the Pseudo-Euclidean PRS.

Example: (Knuth's example) Displaying only coefficients, the following is an Pseudo-Euclidean PRS in $\mathbb{Z}[X]$ where each polynomial is represented by its list of coefficients.

X^8	X^7	X^6	X^5	X^4	X^3	X^2	X	1
1	0	1	0	−3	−3	8	2	−5
		3	0	5	0	−4	−9	21
				−15	0	3	0	−9
						15795	30375	−59535
							1254542875143750	−1654608338437500
								12593338795500743100931141992187500

§3. Determinantal Polynomials

In this section, we introduce the connection between PRS and determinants. The concept of "determinantal polynomials" [119] is key to understanding the connection between elimination and remainders.

Let M be an $m \times n$ matrix, $m \leq n$. The *determinantal polynomial*polynomial!determinantal of M is

$$\texttt{dpol}(M) := \det(M_m) X^{n-m} + \det(M_{m+1})\, X^{n-m-1} + \cdots + \det(M_n)$$

where M_i is the square submatrix of M consisting of the first $m-1$ columns and the ith column of M $(i = m, \ldots, n)$. Call

$$\det(M_m)$$

the *nominal leading coefficient* and $n - m$ the *nominal degree nominal degree* of $\texttt{dpol}(M)$. Of course the degree of $\texttt{dpol}(M)$ could be less than its nominal degree.

Notation: If $P_1, \ldots, P_m$ are polynomials and $n \geq 1 + \max_i\{\deg P_i\}$ then

$$\texttt{mat}_n(P_1, \ldots, P_m)$$

is the $m \times n$ matrix whose ith row contains the coefficients of P_i listed in order of decreasing degree, treating P_i as having nominal degree $n-1$. Write

$$\texttt{dpol}_n(P_1, \ldots, P_m)$$

for $\texttt{dpol}(\texttt{mat}_n(P_1, \ldots, P_m))$. The subscript n is normally omitted when understood or equal to $1 + \max_i\{\deg P_i\}$.

Sylvester's matrix. Let us illustrate this notation. We often apply this notation to "shifted polynomials" (where we call X^iP a "shifted version" of the polynomial P). If P and Q are polynomials of degree m and n respectively then the following $m+n$ by $m+n$ square matrix is called *Sylvester matrix* of P and Q:

$$\texttt{mat}(\underbrace{X^{n-1}P, X^{n-2}P, \ldots, X^1P, X^0P}_{n}, \underbrace{X^{m-1}Q, X^{m-2}Q, \ldots, X^0Q}_{m})$$

$$= \begin{bmatrix} a_m & a_{m-1} & \cdots & & a_0 & & & \\ & a_m & a_{m-1} & \cdots & & a_0 & & \\ & & \ddots & & & & \ddots & \\ & & & a_m & a_{m-1} & \cdots & & a_0 \\ b_n & b_{n-1} & \cdots & b_1 & b_0 & & & \\ & b_n & b_{n-1} & \cdots & b_1 & b_0 & & \\ & & \ddots & & & & \ddots & \\ & & & b_n & b_{n-1} & \cdots & & b_0 \end{bmatrix}$$

where $P = \sum_{i=0}^{m} a_i X^i$ and $Q = \sum_{i=0}^{n} b_i X^i$. The above matrix may also be written as

$$\texttt{mat}(X^{n-1}P, X^{n-2}P, \ldots, X^1P, X^0P; X^{m-1}Q, X^{m-2}Q, \ldots, X^0Q),$$

with a semicolon to separate the P's from the Q's. [In general, we may replace commas with semicolons, purely as a visual aid to indicate groupings.] Since this matrix is square, its determinantal polynomial is a constant called the *resultant* of P and Q, and denoted $\texttt{res}(P,Q)$. We shall return to resultants in Lecture VI.

The basic connection between determinants and polynomials is revealed in the following:

Lemma 5 *Let $P, Q \in D[X]$, $\deg P = m \geq n = \deg Q$. If*

$$M = \texttt{mat}(\underbrace{X^{m-n}Q, X^{m-n-1}Q, \ldots, X^1Q, X^0Q}_{m-n+1}, P)$$

then

$$\texttt{dpol}(M) = \texttt{prem}(P,Q).$$

Proof. Let

$$M' = \texttt{mat}(\underbrace{X^{m-n}Q, X^{m-n-1}Q, \ldots, XQ, Q}_{m-n+1}, b^{m-n+1}P),$$

where $b = \texttt{lead}(Q) = b_n$.

1. Since M' is obtained from the matrix M in the lemma by multiplying the last row by b^{m-n+1}, it follows that $\texttt{dpol}(M') = b^{m-n+1}\texttt{dpol}(M)$.

2. If we do Gaussian elimination on M', by repeated elimination of leading coefficients of the last row we finally get a matrix

$$M'' = \begin{bmatrix} b_n & b_{n-1} & & \cdots & b_0 & & \\ & b_n & b_{n-1} & & \cdots & b_0 & \\ & & \ddots & & & & \\ & & & b_n & b_{n-1} & \cdots & b_0 \\ & & & & c_{n-1} & c_{n-2} & \cdots & c_0 \end{bmatrix}$$

where the polynomial represented by the last row is $R = \sum_{i=0}^{n-1} c_i X^i$ with nominal degree $n-1$. It is seen that $R = \texttt{prem}(P, Q)$.

3. From the definition of determinantal polynomials, $\texttt{dpol}(M'') = b_n^{m-n+1}\texttt{prem}(P, Q)$.

4. Gaussian row operations on a matrix do not change the determinantal polynomial of a matrix:

$$\texttt{dpol}(M') = \texttt{dpol}(M''). \tag{5}$$

The lemma follows from these remarks. **Q.E.D.**

Thus if $Q(X)$ is monic, then the remainder (which is equal to the pseudo-remainder) of $P(X)$ divided by $Q(X)$ is a determinantal polynomial. Another consequence is this:

Corollary 6 *Let $P, Q \in D[X], \deg P = m \geq n = \deg Q$ and $a, b \in D$. Then*

$$\texttt{prem}(aP, bQ) = ab^{m-n+1}\texttt{prem}(P, Q).$$

From equation (5) we further conclude

Corollary 7 *With $b = \texttt{lead}(Q)$,*

$$\texttt{dpol}(\underbrace{X^{m-n}Q, \ldots, Q}_{m-n+1}, P) = \texttt{dpol}(\underbrace{X^{m-n}Q, \ldots, Q}_{m-n+1}, b^{-(m-n+1)}\texttt{prem}(P, Q)). \tag{6}$$

Application. We show that coefficients in the Pseudo-Euclidean PRS can have sizes exponentially larger that those in the corresponding Primitive PRS. Suppose

$$(P_0, P_1, \ldots, P_k)$$

is the Pseudo-Euclidean PRS and $(d_0, d_1, \ldots, d_k)$ associated degree sequence. Write

$$(\delta_1, \ldots, \delta_k)$$

where $\delta_i = d_{i-1} - d_i$. Let $\alpha = \texttt{cont}(P_2), Q_2 = \texttt{prim}(P_2)$:

$$P_2 = \alpha Q_2.$$

Then corollary 6 shows that

$$\alpha^{\delta_2+1} | \texttt{prem}(P_1, \alpha Q_2) = P_3.$$

Writing $P_3 = \alpha^{\delta_2+1} Q_3$ for some Q_3, we get next

$$\alpha^{(\delta_2+1)(\delta_3+1)} | \texttt{prem}(P_2, \alpha^{\delta_2+1} Q_3) = \texttt{prem}(P_2, P_3) = P_4.$$

Continuing in this way, we eventually obtain

$$\alpha^N | P_k \qquad (\text{where } \; N = \prod_{i=2}^{k-1} (\delta_i + 1)).$$

Since $\delta_i \geq 1$, we get $\alpha^{2^k} | P_k$. Assuming that the size of an element $\alpha \in D$ is doubled by squaring, this yields the desired conclusion. Note that this exponential behavior arises even in a regular PRS (all δ_i equal 1).

Exercises

Exercise 3.1: What is the main diagonal of the Sylvester matrix? Show that $a_m^n b_0^m$ and $b_n^m a_0^n$ are terms in the resultant polynomial. What is the general form of such terms? □

Exercise 3.2:
a) The content of P_k is larger than the α^N indicated. [For instance, the content of P_4 is strictly larger than the $\alpha^{(\delta_2+1)(\delta_3+1)}$ indicated.] What is the correct bound for N? (Note that we are only accounting for the content arising from α.)
b) Give a general construction of Pseudo-Euclidean PRS's with coefficient sizes growing at this exponential rate. □

§4. Polynomial Pseudo-Quotient

As a counterpart to lemma 5, we show that the coefficients of the pseudo-quotient can also be characterized as determinants of a suitable matrix M. This fact is not used in this lecture.

Let $P(X) = \sum_{i=0}^m a_i X^i, Q(X) = \sum_{i=0}^n b_i X^i \in D[X]$. We define the *pseudo-quotient* of $P(X)$ divided by $Q(X)$ to be the (usual) quotient of $b^{m-n+1}P(X)$ divided by $Q(X)$, where $b = b_n$ and $m \geq n$. If $m < n$, the pseudo-quotient is just $P(X)$ itself. In the following, we assume $m \geq n$.

The desired matrix is

$$M := \mathtt{mat}(P, X^{m-n}Q, X^{m-n-1}Q, \ldots, XQ, Q) = \begin{bmatrix} a_m & a_{m-1} & a_{m-2} & \cdots & a_{m-n} & \cdots & a_1 & a_0 \\ b_n & b_{n-1} & b_{n-2} & \cdots & b_0 & \cdots & 0 & 0 \\ & b_n & b_{n-1} & \cdots & b_1 & \cdots & 0 & 0 \\ & & \ddots & & & & & \vdots \\ & & & b_n & b_{n-1} & \cdots & b_1 & b_0 \end{bmatrix}.$$

Let M_i denote the $(i+1) \times (i+1)$ principal submatrix of M.

Lemma 8 *Let $C(X) = \sum_{i=0}^{m-n} c_i X^{m-n-i}$ be the pseudo-quotient of $P(X)$ divided by $Q(X)$. Then for each $i = 0, \ldots, m-n$,*

$$c_i = (-1)^i b^{m-n-i} \det M_i, \qquad b = \mathtt{lead}(Q).$$

Proof. Observe that the indexing of the coefficients of $C(X)$ is reversed. The result may be directly verified for $i = 0$. For $i = 1, 2, \ldots, m-n+1$, observe that

$$b^{m-n+1}P(X) - \left(\sum_{j=0}^{i-1} c_j X^{m-n-j} \right) \cdot Q(X) = c_i X^{m-i} + O(X^{m-i-1}) \tag{7}$$

where $O(X^\ell)$ refers to terms of degree at most ℓ. Equation (7) amounts to multiplying the $(j+2)$nd row of M by c_j and subtracting this from the first row, for $j = 0, \ldots, i-1$. Since the determinant of a matrix is preserved by this operation, we deduce that

$$\det \begin{bmatrix} a'_m & a'_{m-1} & \cdots & a'_{m-i+1} & a'_{m-i} \\ b_n & b_{n-1} & \cdots & b_{n-i+1} & b_{n-i} \\ & b_n & & \cdots & b_{n-i+1} \\ & & \ddots & & \vdots \\ & & & b_n & b_{n-1} \end{bmatrix} = \det \begin{bmatrix} 0 & 0 & \cdots & 0 & c_i \\ b_n & b_{n-1} & \cdots & b_{n-i+1} & b_{n-i} \\ & b_n & & \cdots & b_{n-i+1} \\ & & \ddots & & \vdots \\ & & & b_n & b_{n-1} \end{bmatrix}$$

where $a'_j := a_j b^{m-n+1}$. But the LHS equals $b^{m-n+1} \det M_i$ and the RHS equals $(-b)^i c_i$. **Q.E.D.**

§5. The Subresultant PRS

We now present Collins's PRS algorithm.

A PRS $(P_0, P_1, \ldots, P_k)$ is said to be *based* on a sequence

$$(\beta_1, \beta_2, \ldots, \beta_{k-1}) \quad (\beta_i \in D) \tag{8}$$

if

$$P_{i+1} = \frac{\texttt{prem}(P_{i-1}, P_i)}{\beta_i} \quad (i = 1, \ldots, k-1). \tag{9}$$

Note that the Pseudo-Euclidean PRS and Primitive PRS are based on the appropriate sequences. We said the Primitive PRS is based on a sequence whose entries β_i are relatively expensive to compute. We now describe one sequence that is easy to obtain (even in parallel). Define for $i = 0, \ldots, k-1$,

$$\left.\begin{array}{l} \delta_i := \deg(P_i) - \deg(P_{i+1}), \\ a_i := \texttt{lead}(P_i). \end{array}\right\} \tag{10}$$

Then let

$$\beta_{i+1} := \begin{cases} (-1)^{\delta_0+1} & \text{if} \quad i = 0, \\ (-1)^{\delta_i+1} (\psi_i)^{\delta_i} a_i & \text{if} \quad i = 1, \ldots, k-2, \end{cases} \tag{11}$$

where $(\psi_0, \ldots, \psi_{k-1})$ is an auxiliary sequence given by

$$\left.\begin{array}{lcl} \psi_0 & := & 1, \\ \psi_{i+1} & := & \psi_i (\frac{a_{i+1}}{\psi_i})^{\delta_i} = \frac{(a_{i+1})^{\delta_i}}{(\psi_i)^{\delta_i - 1}}, \end{array}\right\} \tag{12}$$

for $i = 0, \ldots, k-2$.

By definition, the *subresultant PRS* is based on the sequence $(\beta_1, \ldots, \beta_{k-1})$ just defined. The *subresultant PRS algorithm* computes this sequence. It is easy to implement the algorithm in the style of the usual Euclidean algorithm: the values $P_0, P_1, a_0, a_1, \delta_0, \psi_0, \psi_1$ and β_1 are initially available. Proceeding in *stages*, in the ith stage, $i \geq 1$, we compute the quintuple (in this order)

$$P_{i+1}, a_{i+1}, \delta_i, \psi_{i+1}, \beta_{i+1} \tag{13}$$

according to (9),(10),(12) and (11), respectively.

This algorithm was discovered by Collins in 1967 [44] and subsequently simplified by Brown [30]. It is the best algorithm in the family of algorithms based on sequences of β's.

It is not easy to see why this sequence of β_i works: Superficially, equation (9) implies that P_{i+1} lies in $Q_D[x]$ rather than $D[x]$. Moreover, it is not clear from (12) that ψ_i (and hence β_{i+1}) belongs to D rather than Q_D. In fact the ψ_i's turn out to be determinants of coefficients of P_0 and P_1, a fact not known in the early papers on the subresultant PRS algorithm. This fact implies that the β_i's have sizes that are polynomial in the input size. In other words, this algorithm succeeded in curbing the exponential growth of coefficients (unlike the Pseudo-Euclidean PRS) without incurring expensive multiple GCD computations (which was the bane of the primitive PRS). *The theory of subresultants will explain all this, and more. This is ostensibly the goal of the rest of this lecture, although subresultants have other uses as well.*

Complexity. It is easy to see that implementation (13) takes $O(n^2 \log n)$ operations of D. Schwartz [188] applied the Half-GCD idea (Lecture II) in this setting to get an $O(n \log^2 n)$ bound, provided we only compute the sequence of partial quotients and coefficients of similarities

$$(Q_1, \alpha_1, \beta_1), \ldots, (Q_{k-1}, \alpha_{k-1}, \beta_{k-1})$$

where $\alpha_i P_{i+1} = \beta_i P_{i-1} + P_i Q_i$. This amounts to an extended GCD computation.

EXERCISES

Exercise 5.1: Modify the HGCD algorithm (see Lecture VIII) to compute the subresultants. □

§6. Subresultants

We introduce subresultants.

Definition: Let $P, Q \in D[X]$ with

$$\deg(P) = m > n = \deg(Q) \geq 0.$$

For $i = 0, 1, \ldots, n$, the ith *subresultant!ith@ith subresultant* of P and Q is defined as

$$\texttt{sres}_i(P,Q) := \texttt{dpol}(\underbrace{X^{n-i-1}P, X^{n-i-2}P, \ldots, P}_{n-i}, \underbrace{X^{m-i-1}Q, X^{m-i-2}Q, \ldots, Q}_{m-i}). \tag{14}$$

Observe that the defining matrix

$$\texttt{mat}(X^{n-i-1}P, \ldots, P; X^{m-i-1}Q, \ldots, Q)$$

has $m+n-2i$ rows and $m+n-i$ columns. If $n = 0$, then $i = 0$ and P does not appear in the matrix and the matrix is $m \times m$. The *nominal degree* of $\texttt{sres}_i(P,Q)$ is i. The nominal leading coefficient of $\texttt{sres}_i(P,Q)$ is called the ith *principal subresultant coefficient* of P and Q, denoted $\texttt{psc}_i(P,Q)$. .

Note that the zeroth subresultant is in fact the resultant,

$$\texttt{sres}_0(P,Q) = \texttt{res}(P,Q),$$

and thus subresultants are a generalization of resultants. Furthermore,

$$\texttt{sres}_n(P,Q) = \texttt{lead}(Q)^{m-n-1}Q \sim Q.$$

It is convenient to extend the above definitions to cover the cases $i = n+1, \ldots, m$:

$$\mathtt{sres}_i(P,Q) := \begin{cases} 0 & \text{if } i = n+1, n+2, \ldots, m-2 \\ Q & \text{if } i = m-1 \\ P & \text{if } i = m \end{cases} \tag{15}$$

Note that this extension is consistent with the definition (14) because in case $n = m-1$, the two definitions of $\mathtt{sres}_n(P,Q)$ agree. Although this extension may appear contrived, it will eventually prove to be the correct one. Again, the subscript in $\mathtt{sres}_i(P,Q)$ indicates its nominal degree. The sequence

$$(S_m, S_{m-1}, \ldots, S_1, S_0), \qquad \text{where } S_i = \mathtt{sres}_i(P,Q),$$

is called the *subresultant chain* of P and Q. A member $\mathtt{sres}_i(P,Q)$ in the chain is *regular* if its degree is equal to the nominal degree i; otherwise it is *irregular*. We say the chain is *regular* if $\mathtt{sres}_i(P,Q)$ is regular for all $i = 0, \ldots, n$ (we ignore $i = n+1, \ldots, m$).

Likewise, we extend the definition of principal subresultant coefficient $\mathtt{psc}_i(P,Q)$ to the cases $i = n+1, \ldots, m$:

$$\mathtt{psc}_i(P,Q) := \begin{cases} \text{nominal leading coefficient of } \mathtt{sres}_i(P,Q) & \text{for } i = n+1, \ldots, m-1 \\ 1 & \text{for } i = m. \end{cases} \tag{16}$$

Note that $\mathtt{psc}_m(P,Q)$ is not defined as $\mathtt{lead}(P) = \mathtt{lead}(S_m(P,Q))$ as one might have expected.

We will see that the subresultant PRS is just a subsequence of the corresponding subresultant chain.

Remark: This concept of "regular" polynomials is quite generic: if a polynomial has a 'nominal degree' (which is invariably an upper bound on the actual degree), then its "regularity" simply means that the nominal degree equals the actual degree.

Exercises

Exercise 6.1: Let the coefficients of $P(X) = \sum_{i=0}^{m} a_i X^i$ and $Q(X) = \sum_{j=0}^{n} b_j X^j$ be indeterminates. Let the *weights* of a_i and b_j be i and j, respectively. If M is a monomial in the a_i's and b_j's, the *weight* of M is the sum of the weights of each indeterminate in M. E.g., $M = (a_m)^n (b_0)^m$ has weight mn. Let c_k be the leading coefficient of $\mathtt{sres}_k(P,Q)$, viewed as a polynomial in the a_i's and b_j's.

(i) Show that the weight of each term in the polynomial c_k is

$$m(n-k) + (m-k)k = mn - k^2.$$

HINT: note that the principal diagonal of the matrix defining $\mathtt{sres}_k(P,Q)$ produces a term with this weight. Use the fact that if π, π' are two permutations of $m+n-2k$ that differ by a transposition, the terms in c_k arising from π, π' have the same weight, provided they are not zero. What if one of terms is zero?

(ii) Generalize this to the remaining coefficients of $\mathtt{sres}_k(P,Q)$. □

§7. Pseudo-subresultants

The key to understanding polynomial remainder sequences lies in the prediction of *unavoidable* contents of polynomials in the PRS. This prediction is simpler for regular subresultant chains.

Regular chains can be studied using indeterminate coefficients. To be precise, suppose the given polynomials

$$P = \sum_{i=0}^{m} a_i X^i, \quad Q = \sum_{i=0}^{n} b_i X^i, \qquad (n = m-1) \tag{17}$$

come from the ring

$$\mathbb{Z}[X,\ a_m, \ldots, a_0, b_{m-1}, \ldots, b_0] = \mathbb{Z}[X][a_m, \ldots, a_0, b_{m-1}, \ldots, b_0]$$

where a_i, b_i are indeterminates. Assuming $\deg P = 1 + \deg Q$ is without loss of generality for indeterminate coefficients. After obtaining the properties of subresultants in this setting, we can "specialize" the indeterminates a_i, b_j to values $\overline{a}_i, \overline{b}_i$ in D. This induces a ring homomorphism Φ from $\mathbb{Z}[X;\ a_m, \ldots, a_0, b_{m-1}, \ldots, b_0]$ to $D[X]$. We indicate the Φ-image of an element $e \in \mathbb{Z}[X;\ a_m, \ldots, a_0, b_{m-1}, \ldots, b_0]$ by $\overline{e}$, called the *specialization of* e. Thus if $(S_m, \ldots, S_0)$ is the subresultant chain of P, Q, we can observe the behavior of the specialized chain[2]

$$(\overline{S}_m, \ldots, \overline{S}_0) \tag{18}$$

in $D[X]$. This approach was first used by Loos [33] who also noted that this has the advantage of separating out the two causes of irregularity in chains: (a) the irregularity effects caused by the specialization, and (b) the similarity relations among subresultants that are *independent* of the specialization. The similarity relations of (b) are captured in Habicht's theorem (see exercise). The proper execution of this program is slightly complicated by the fact that in general,

$$\overline{S}_i = \overline{\mathtt{sres}_i(P, Q)} \neq \mathtt{sres}_i(\overline{P}, \overline{Q}). \tag{19}$$

To overcome this difficulty, the concept of "pseudo-subresultant chains" was introduced in [84]. It turns out that (18) is precisely the pseudo-subresultant chain of $\overline{P}, \overline{Q}$, provided $\deg \overline{P} = m$. In this way, Loos' program is recaptured via pseudo-subresultants without an explicit use of specialization.

Definition 1 *Let $P, Q \in D[X]$ and $m = \deg P > \deg Q \geq -\infty$. For $i = 0, 1, \ldots, m-1$, define the ith* pseudo-subresultant *of P and Q to be*

$$\mathtt{psres}_i(P, Q) := \mathtt{dpol}_{2m-i-1}(\underbrace{X^{m-i-2}P, X^{m-i-3}P, \ldots, P}_{m-i-1}, \underbrace{X^{m-i-1}Q, X^{m-i-2}Q, \ldots, Q}_{m-i}).$$

Note that

$$\mathtt{psres}_{m-1}(P, Q) = Q.$$

Extending these definitions as before,

$$\mathtt{psres}_m(P, Q) := P.$$

The sequence

$$(S_m, S_{m-1}, \ldots, S_1, S_0), \qquad \text{where } S_i = \mathtt{psres}_i(P, Q)$$

is called the pseudo-subresultant chain *of P and Q. The ith* pseudo-principal subresultant coefficient *of P and Q, denoted $\mathtt{ppsc}_i(P, Q)$ is defined to be the nominal leading coefficient of $\mathtt{psres}_i(P, Q)$ for $i = 0, \ldots, m-1$ but (again) $\mathtt{ppsc}_m(P, Q) := 1$.*

Pseudo-subresultants of P, Q are basically their subresultants except that we give Q a nominal degree of $\deg(P) - 1$. The *defining matrix* for $\mathtt{psres}_i(P, Q)$ has shape $(2m - 2i - 1) \times (2m - i - 1)$. This definition, unlike the definition of subresultants, allows $\deg Q = -\infty$ $(Q = 0)$, in which case $\mathtt{psres}_i(P, Q) = 0$ for all $i < m$. It is not hard to see that

$$\mathtt{psres}_i(aP, bQ) = a^{m-i-1} b^{m-i} \mathtt{psres}_i(P, Q).$$

[2] We prefer to write '$\overline{S}_j$' instead of the more accurate '$\overline{S_j}$'.

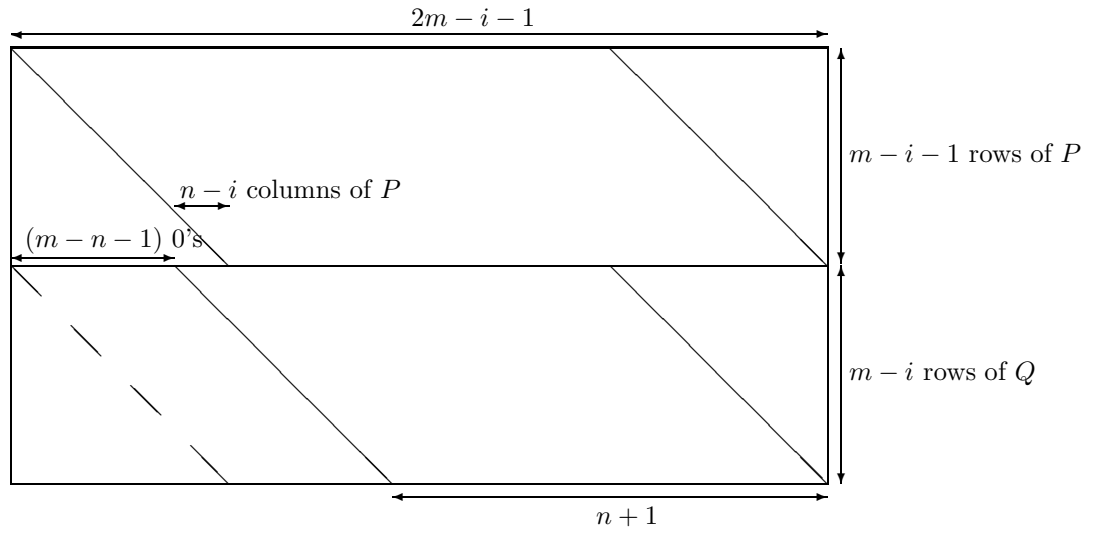


Figure 1: The matrix associated to $\texttt{psres}_i(P,Q)$

Furthermore, pseudo-subresultants are similar to subresultants:

$$\texttt{psres}_i(P,Q) = \begin{cases} \texttt{lead}(P)^{m-n-1}\texttt{sres}_i(P,Q) & \text{for } i = 0,1,\ldots,m-2 \\ \texttt{sres}_i(P,Q) & \text{for } i = m-1, m. \end{cases}$$

Our initial goal is to prove a weak version of the Subresultant Theorem. We first give a preparatory lemma.

Lemma 9 (Basic Lemma) *Let $P, Q \in D[X]$ with $\deg P = m > n = \deg Q \geq -\infty$. If $a = \texttt{lead}(P), b = \texttt{lead}(Q)$ then for $i = 0,\ldots,m-2$:*

$$\texttt{psres}_i(P,Q) = 0 \quad \text{if } i \geq n+1 \tag{20}$$

$$\texttt{psres}_n(P,Q) = (ab)^{m-n-1}Q \tag{21}$$

$$\texttt{psres}_i(P,Q) = a^{m-n-1}b^{-(m-n+1)(n-i-1)}(-1)^{(n-i)(m-i)}\texttt{psres}_i(Q,\texttt{prem}(P,Q)), \quad \text{if } i \leq n-1. \tag{22}$$

Proof. The result is clearly true if $Q = 0$, so assume $\deg Q \geq 0$. We use the aid of Figure 1. Let column 1 refer to the rightmost column of the matrix

$$\texttt{mat}(X^{m-i-2}P,\ldots,P;X^{m-i-1}Q,\ldots,Q)$$

in the figure. Thus column $m+1$ contains the leading coefficient of the row corresponding to P. The column containing the leading coefficient of the row corresponding to $X^{m-i-1}Q$ is $(m-i-1)+(n+1) = m+n-i$. But P and $X^{m-i-1}Q$ correspond to consecutive rows. Hence if $i = n$, the leftmost $2m-2i-1$ columns form an upper triangular square matrix with determinant $a^{m-n-1}b^{m-n}$. It is not hard to see that this proves equation (21). If $i > n$ then the last two rows of the leftmost $2m-2i-2$ columns are identically zero. This means that any square matrix obtained by adding a column to these $2m-2i-2$ columns will have zero determinant. This proves equation (20). Finally, to prove equation (22), suppose $i \leq n-1$. We get

$\texttt{psres}_i(P,Q)$

$$
\begin{aligned}
&= \texttt{dpol}(\underbrace{X^{m-i-2}P,\ldots,P}_{m-i-1},\underbrace{X^{m-i-1}Q,\ldots,Q}_{m-i}) \\
&= \texttt{dpol}(\underbrace{X^{m-i-2}P,\ldots,X^{n-i}P}_{m-n-1},\underbrace{X^{n-i-1}P,\ldots,P}_{n-i},\underbrace{X^{m-i-1}Q,\ldots,Q}_{m-i}) \\
&= \texttt{dpol}(\underbrace{X^{n-i-1}P,\ldots,P}_{n-i},\underbrace{X^{m-i-1}Q,\ldots,Q}_{m-i})\cdot a^{m-n-1} \\
&\qquad \text{(expanding the leftmost } m-n-1 \text{ columns)} \\
&= \texttt{dpol}(\underbrace{X^{n-i-1}b^{m-n+1}P,\ldots,b^{m-n+1}P}_{n-i},\underbrace{X^{m-i-1}Q,\ldots,Q}_{m-i})\cdot a^{m-n-1}\cdot b^{-(m-n+1)(n-i)} \\
&= \texttt{dpol}(\underbrace{X^{n-i-1}\texttt{prem}(P,Q),\ldots,\texttt{prem}(P,Q)}_{n-i},\underbrace{X^{m-i-1}Q,\ldots,Q}_{m-i})\cdot a^{m-n-1}\cdot b^{-(m-n+1)(n-i)} \\
&\qquad \text{(by corollary 7)} \\
&= \texttt{dpol}(\underbrace{X^{m-i-1}Q,\ldots,X^{n-i-1}Q}_{m-n+1},\underbrace{X^{n-i-2}Q,\ldots,Q}_{n-i-1},\underbrace{X^{n-i-1}\texttt{prem}(P,Q),\ldots,\texttt{prem}(P,Q)}_{n-i}) \\
&\quad \cdot a^{m-n-1}\cdot b^{-(m-n+1)(n-i)}\cdot(-1)^{(n-i)(m-i)} \\
&\qquad \text{(transposing columns)} \\
&= \texttt{dpol}(\underbrace{X^{n-i-2}Q,\ldots,Q}_{n-i-1},\underbrace{X^{n-i-1}\texttt{prem}(P,Q),\ldots,\texttt{prem}(P,Q)}_{n-i}) \\
&\quad \cdot a^{m-n-1}\cdot b^{-(m-n+1)(n-i)}\cdot(-1)^{(n-i)(m-i)}\cdot b^{m-n+1} \\
&\qquad \text{(expanding the leftmost } m-n+1 \text{ columns)} \\
&= \texttt{psres}_i(Q,\texttt{prem}(P,Q))\cdot a^{m-n-1}\cdot b^{-(m-n+1)(n-i-1)}\cdot(-1)^{(n-i)(m-i)}.
\end{aligned}
$$

Q.E.D.

The case $i = n-1$ in equation (22) is noteworthy:

$$\texttt{psres}_{n-1}(P,Q) = (-a)^{m-n-1}\texttt{prem}(P,Q). \tag{23}$$

We define a *block* to be a sequence

$$B = (P_1, P_2, \ldots, P_k), \quad k \geq 1$$

of polynomials where $P_1 \sim P_k$ and $0 = P_2 = P_3 = \cdots = P_{k-1}$. We call P_1 and P_k (respectively) the *top* and *base* of the block. Two special cases arise: In case $k = 1$, we call B a *regular block*; in case $P_1 = 0$, we call B a *zero block*. Thus the top and the base of a regular block coincide.

Using the Basic Lemma, we deduce the general structure of a subresultant chain.

Theorem 10 (Block Structure Theorem) *A subresultant or pseudo-subresultant chain*

$$(S_m, S_{m-1}, \ldots, S_0)$$

is uniquely partitioned into a sequence

$$B_0, B_1, \ldots, B_k, \qquad (k \geq 1)$$

of blocks such that
a) B_0 is a regular block.
b) If U_i is the base polynomial of block B_i then U_i is regular and $U_{i+1} \sim \texttt{prem}(U_{i-1}, U_i)$ $(0 < i < k)$.
c) There is at most one zero block; if there is one, it must be B_k.

Proof. Since pseudo-subresultants are similar to their subresultant counterparts, it is sufficient to prove the theorem assuming $(S_m, \ldots, S_0)$ is a pseudo-subresultant chain.

Assertion a) is immediate since $B_0 = (S_m)$. We verify assertion b) by induction on i: If $\deg S_{m-1} = n$, the Basic Lemma (21) implies that $(S_{m-1}, S_{m-2}, \ldots, S_n)$ forms the next block B_1. Moreover, S_n is regular and $S_{n-1} \sim \texttt{prem}(S_m, S_{m-1})$ (23). Thus $U_2 \sim \texttt{prem}(U_0, U_1)$. Inductively, assuming that block B_i has been defined and the polynomial following the base of B_i is similar to $\texttt{prem}(U_{i-1}, U_i)$, we can repeat this argument to define the next block B_{i+1} and show that U_{i+1} is regular and $U_{i+1} \sim \texttt{prem}(U_{i-1}, U_i)$. This argument terminates when $\texttt{prem}(U_{i-1}, U_i) = 0$. Then the rest of the pseudo-subresultants are zero, forming the final zero block, which is assertion c). **Q.E.D.**

By definition, a sequence of polynomials that satisfies this Block Structure theorem is called *block-structured*. This structure is graphically illustrated in figure 2. Here $m = 12$ and there are 5 blocks in this particular chain. Each non-zero polynomial in the chain is represented by a horizontal line segment and their constant terms are vertically aligned. The leading coefficient of regular polynomials lies on the main diagonal. The top and base polynomials in the ith block are denoted by T_i and U_i, respectively.

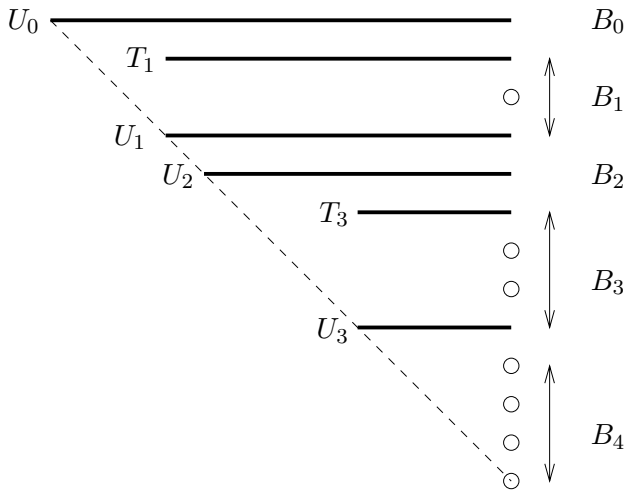


Figure 2: Block structure of a chain with $m = 12$

Exercises

Exercise 7.1: Construct an example illustrating (19). □

Exercise 7.2: Deduce the following from the Block Structure Theorem. Suppose $P, Q \in D[X]$ has the remainder sequence $(P_0, P_1, \ldots, P_\ell)$ in $\mathbf{Q}_D[X]$. Let the blocks of their subresultant sequence be $B_0, B_1, \ldots$, where U_i is the base of block B_i.
(i) $U_i \sim P_i$ for $i \geq 0$. If the last non-zero block is B_ℓ, then $P_\ell \sim \texttt{GCD}(P_0, P_1)$.
(ii) The number of non-zero blocks in the subresultant chain of P, Q is equal to the length of any remainder sequence of P, Q. Moreover, the base of each block is similar to the corresponding member of the remainder sequence.
(iii) The last non-zero element in the subresultant chain is similar to $\texttt{GCD}(P, Q)$.
(iv) The smallest index i such that the principal subresultant coefficient $\texttt{psc}_i(P, Q)$ is non-zero

is equal to $\deg(\texttt{GCD}(P,Q))$.
(v) Two polynomials P, Q are relatively prime if and only if their resultant does not vanish, $\texttt{res}(P,Q) \neq 0$. □

§8. Subresultant Theorem

The Block Structure Theorem does not tell us the coefficients of similarity implied by the relation $b_{i+1} \sim \texttt{prem}(b_{i-1}, b_i)$. It is a tedious exercise to track down these coefficients in *some* form; but the challenge is to present them in a *useful* form. It is non-obvious that these coefficients bear simple relations to the principal pseudo-subresultant coefficients; the insight for such a relation comes from the case of indeterminate coefficients (Habicht's theorem, see Exercise). These relations, combined with the Block Structure Theorem, constitute the Subresultant Theorem which we will prove. We begin with an analogue to Habicht's theorem.

Theorem 11 (Pseudo Habicht's theorem) *Let $(S_m, \ldots, S_0)$ be a pseudo-subresultant chain, and let $(c_m, \ldots, c_0)$ be the corresponding sequence of principal pseudo-subresultant coefficients. If S_k is regular $(1 \leq k \leq m)$ then*

$$S_i = c_k^{-2(k-i-1)} \texttt{psres}_i(S_k, S_{k-1}), \qquad i = 0, \ldots, k-1.$$

Proof. We use induction on k. If $k = m$ then the result is true by definition (recall $c_m = 1$). Let $P = S_m$, $Q = S_{m-1}$, $n = \deg Q$, $a = \texttt{lead}(P)$ and $b = \texttt{lead}(Q)$. So S_n is the next regular pseudo-subresultant. Unfortunately, the argument is slightly different for $k = n$ and for $k < n$.

CASE $k = n$: The Basic Lemma implies

$$S_n = (ab)^{(m-n-1)}Q, \qquad S_{n-1} = (-a)^{(m-n-1)}\texttt{prem}(P,Q).$$

Taking coefficients of S_n, we get $c_n = a^{m-n-1}b^{m-n}$. From the Basic Lemma (22), for $i = 0, \ldots, n-1$,

$$\begin{aligned}
& a^{-(m-n-1)}b^{(m-n+1)(n-i-1)}(-1)^{-(n-i)(m-i)}S_i \\
&= \texttt{psres}_i(Q, \texttt{prem}(P,Q)) \\
&= \texttt{psres}_i((ab)^{-(m-n-1)}S_n, (-a)^{-(m-n-1)}S_{n-1}) \\
&\qquad \text{(substituting for } Q, \texttt{prem}(P,Q)) \\
&= (ab)^{-(m-n-1)(n-i-1)}(-a)^{-(m-n-1)(n-i)}\texttt{psres}_i(S_n, S_{n-1}). \\
S_i &= a^{-2(m-n-1)(n-i-1)}b^{-2(m-n)(n-i-1)}\texttt{psres}_i(S_n, S_{n-1}) \\
&= c_n^{-2(n-i-1)}\texttt{psres}_i(S_n, S_{n-1}).
\end{aligned}$$

CASE $1 \leq k < n$: By the Block Structure Theorem, there is some regular S_ℓ $(\ell \leq n)$ such that $k = \deg(S_{\ell-1})$. By induction hypothesis, the lemma is true for ℓ. Let $a_i = \texttt{lead}S_i$ (so $a_i \neq 0$ unless $S_i = 0$). We have

$$\begin{aligned}
c_\ell^{2(\ell-k-1)}S_k &= \texttt{psres}_k(S_\ell, S_{\ell-1}) \qquad \text{(by induction)} \\
&= (c_\ell a_{\ell-1})^{\ell-k-1}S_{\ell-1} \qquad \text{(Basic Lemma).} \\
S_{\ell-1} &= (c_\ell a_{\ell-1}^{-1})^{\ell-k-1}S_k. \qquad (24)
\end{aligned}$$

Taking coefficients,

$$c_k = c_\ell^{-(\ell-k-1)}a_{\ell-1}^{\ell-k}. \qquad (25)$$

Again,

$$\begin{aligned} c_\ell^{2(\ell-k)} S_{k-1} &= \mathtt{psres}_{k-1}(S_\ell, S_{\ell-1}) \qquad \text{(by induction)} \\ &= (-c_\ell)^{\ell-k-1}\mathtt{prem}(S_\ell, S_{\ell-1}) \qquad \text{(by equation (23)).} \\ \mathtt{prem}(S_\ell, S_{\ell-1}) &= (-c_\ell)^{\ell-k+1} S_{k-1}. \end{aligned} \tag{26}$$

Hence

$$\begin{aligned} & c_\ell^{2(\ell-i-1)} S_i \\ &= \mathtt{psres}_i(S_\ell, S_{\ell-1}) \qquad \text{(by induction)} \\ &= c_\ell^{\ell-k-1} a_{\ell-1}^{-(\ell-k+1)(k-i-1)} (-1)^{(\ell-i)(k-i)} \mathtt{psres}_i(S_{\ell-1}, \mathtt{prem}(S_\ell, S_{\ell-1})) \qquad \text{(Basic Lemma)} \\ &= c_\ell^{\ell-k-1} a_{\ell-1}^{-(\ell-k+1)(k-i-1)} (-1)^{(\ell-i)(k-i)} \mathtt{psres}_i((c_\ell a_{\ell-1}^{-1})^{\ell-k-1} S_k, (-c_\ell)^{\ell-k+1} S_{k-1}) \\ & \qquad\qquad \text{(by (24), (26))} \\ &= c_\ell^{2(k-i)(\ell-k)} a_{\ell-1}^{-2(\ell-k)(k-i-1)} \mathtt{psres}_i(S_k, S_{k-1}) \qquad \text{(more manipulations).} \\ S_i &= (c_\ell)^{2(\ell-k-1)(k-i-1)} (a_{\ell-1})^{-2(\ell-k)(k-i-1)} \mathtt{psres}_i(S_k, S_{k-1}) \\ &= (c_k)^{-2(k-i-1)} \mathtt{psres}_i(S_k, S_{k-1}) \qquad \text{(by (25)).} \end{aligned}$$

Q.E.D.

Combined with the Basic Lemma, it is straightforward to infer:

Theorem 12 (Pseudo-Subresultant Theorem) *Let $(S_m, \ldots, S_0)$ be a pseudo-subresultant chain, and let $(a_m, \ldots, a_0)$ be the corresponding sequence of leading coefficients. This chain is block-structured such that if S_ℓ, S_k $(m \geq \ell > k \geq 1)$ are two consecutive regular pseudo-subresultants in this sequence then:*

$$S_k = \begin{cases} (a_\ell a_{\ell-1})^{\ell-k-1} S_{\ell-1} & \text{if } \ell = m, \\ (a_\ell^{-1} a_{\ell-1})^{\ell-k-1} S_{\ell-1} & \text{if } \ell < m. \end{cases} \tag{27}$$

$$S_{k-1} = \begin{cases} (-a_\ell)^{\ell-k-1} \mathtt{prem}(S_\ell, S_{\ell-1}) & \text{if } \ell = m, \\ (-a_\ell)^{-(\ell-k+1)} \mathtt{prem}(S_\ell, S_{\ell-1}) & \text{if } \ell < m. \end{cases} \tag{28}$$

Finally, we transfer the result from pseudo-subresultants to subresultants:

Theorem 13 (Subresultant Theorem) *Let $(R_m, \ldots, R_0)$ be a subresultant chain, and let $(c_m, \ldots, c_0)$ be the corresponding sequence of principal subresultant coefficients. This chain is block-structured such that if R_ℓ, R_k $(m \geq \ell > k \geq 1)$ are two consecutive regular subresultants in this sequence then:*

$$R_k = (c_\ell^{-1} \mathtt{lead}(R_{\ell-1}))^{\ell-k-1} R_{\ell-1}, \tag{29}$$

$$R_{k-1} = (-c_\ell)^{-\ell+k-1} \mathtt{prem}(R_\ell, R_{\ell-1}). \tag{30}$$

Proof. Let $(S_m, \ldots, S_0)$ be the corresponding pseudo-subresultant chain with leading coefficients $(a_m, \ldots, a_0)$. Write a instead of a_m and let $n = \deg S_{m-1}$. We exploit the relation

$$R_i = \begin{cases} S_i & \text{if } i = m-1, m, \\ a^{-(m-n-1)} S_i & \text{if } i = 0, \ldots, m-2. \end{cases}$$

Hence, if R_i is regular and $i < m$, we have

$$c_i = a^{-(m-n-1)} a_i.$$

We show the derivation of R_{k-1}, leaving the derivation of R_k to the reader:

$$\begin{aligned}
R_{k-1} &= a^{-(m-n-1)} S_{k-1} \\
&= \begin{cases} a^{-(m-n-1)}(-a_\ell)^{\ell-k-1}\mathtt{prem}(S_\ell, S_{\ell-1}) & \text{if } \ell = m \\ a^{-(m-n-1)}(-a_\ell)^{-(\ell-k+1)}\mathtt{prem}(S_\ell, S_{\ell-1}) & \text{if } \ell < m \end{cases} \\
&= \begin{cases} (-1)^{\ell-k-1}\mathtt{prem}(S_\ell, S_{\ell-1}) & \text{if } \ell = m \\ a^{-(m-n-1)}(-a_\ell)^{-(\ell-k+1)}\mathtt{prem}(S_\ell, S_{\ell-1}) & \text{if } \ell < m \end{cases} \\
&= \begin{cases} (-1)^{\ell-k-1}\mathtt{prem}(R_\ell, R_{\ell-1}) & \text{if } \ell = m \\ a^{-(m-n-1)}(-a_\ell)^{-(\ell-k+1)}\mathtt{prem}(R_\ell, a^{m-n-1}R_{\ell-1}) & \text{if } \ell = m-1 \\ a^{-(m-n-1)}(-a_\ell)^{-(\ell-k+1)}\mathtt{prem}(a^{m-n-1}R_\ell, a^{m-n-1}R_{\ell-1}) & \text{if } \ell < m \end{cases} \\
&= \begin{cases} (-1)^{\ell-k-1}\mathtt{prem}(R_\ell, R_{\ell-1}) & \text{if } \ell = m \\ (-a_\ell)^{-(\ell-k+1)}a^{(m-n-1)(\ell-k+1)}\mathtt{prem}(R_\ell, R_{\ell-1}) & \text{if } \ell = m-1 (= n) \\ (-a_\ell)^{-(\ell-k+1)}a^{(m-n-1)(\ell-k+1)}\mathtt{prem}(R_\ell, R_{\ell-1}) & \text{if } \ell < m \end{cases} \\
&= (-c_\ell)^{-(\ell-k+1)}\mathtt{prem}(R_\ell, R_{\ell-1}).
\end{aligned}$$

The last equality is justified since:
(i) $\ell = m$: this is because $c_\ell = 1$.
(ii) and (iii): $\ell < m$: this is because $c_\ell = a^{-(m-n-1)}a_\ell$. **Q.E.D.**

So equation (29) gives the coefficients of similarity between the top and base polynomials in each block.

Exercises

Exercise 8.1:
(i) Verify the Pseudo-Subresultant Theorem.
(ii) Complete the proof for the Subresultant Theorem. □

Exercise 8.2: Show that the problem of computing the GCD of two integer polynomials is in the complexity class $NC = NC_B$. □

Exercise 8.3: Prove that if P, Q have indeterminate coefficients as in (17), then
(i) $\mathtt{sres}_{m-2}(P, Q) = \mathtt{prem}(P, Q)$.
(ii) For $k = 0, \ldots, m-3$,

$$b_{m-1}^{2(m-k-2)}\mathtt{sres}_k(P, Q) = \mathtt{sres}_k(Q, \mathtt{prem}(P, Q)).$$

(iii) [Habicht's theorem] If $S_i = \mathtt{sres}_i(P, Q)$ and $c_i = \mathtt{psc}_i(P, Q)$, for $j = 1, \ldots, m-1$,

$$c_{j+1}^{2(j-k)} S_k = \mathtt{sres}_k(S_{j+1}, S_j), \qquad (k = 0, \ldots, j-1) \tag{31}$$

$$c_{j+1}^2 S_{j-1} = \mathtt{prem}(S_{j+1}, S_j). \tag{32}$$

□

§9. Correctness of the Subresultant PRS Algorithm

We relate the subresultant PRS

$$(P_0, P_1, \ldots, P_k) \tag{33}$$

described in §5 (equations (11) and (12)) to the subresultant chain

$$(R_m, R_{m-1}, \ldots, R_0) \tag{34}$$

where $R_m = P_0$ and $R_{m-1} = P_1$. (Note that the convention for subscripting PRS's in increasing order is opposite to that for subresultant chains.) The basic connection, up to similarity, is already established by the Block Structure Theorem. Th real task is to determine the coefficients of similarity between the top of B_i and P_i. As a matter of fact, we have not even established that members of the Subresultant PRS are in $D[X]$. This is captured in the next theorem. Recall the computation of P_i involves the following two auxiliary sequences

$$(\beta_1, \ldots, \beta_{k-1}), \qquad (\psi_0, \ldots, \psi_{k-1})$$

as given in (11) and (12), where

$$\delta_i = \deg P_i - \deg P_{i+1}, \qquad a_i = \texttt{lead}(P_i).$$

Theorem 14 (Subresultant PRS Correctness) *Let T_i, U_i be the top and base polynomials of block B_i, where $(B_0, \ldots, B_k)$ are the non-zero blocks of our subresultant chain.*

a) $\psi_i = \texttt{lead}(U_i)$, $i = 1, \ldots, k$. *(Note that $\psi_0 = 1$.)*

b) *The sequence $(T_0, \ldots, T_k)$ is precisely $(P_0, \ldots, P_k)$, the subresultant PRS.*

Proof. We use induction on i.
BASIS: Part a): from (12), we have $\psi_1 = (a_1)^{\delta_0}$. We verify from equation (29) that $\texttt{lead}(U_1) = \psi_1$.
Part b): By definition, $T_i = P_i$ for $i = 0, 1$. Using the Subresultant Theorem,

$$\begin{aligned}
P_2 &= \frac{\texttt{prem}(P_0, P_1)}{\beta_1} \quad \text{(by definition)} \\
&= \frac{\texttt{prem}(T_0, T_1)}{(-1)^{\delta_0+1}} \quad (\beta_1 = (-1)^{\delta_0+1}) \\
&= \frac{(-1)^{\delta_0+1} T_2}{(-1)^{\delta_0+1}} \qquad \text{(from (30))} \\
&= T_2. \\
P_3 &= \frac{\texttt{prem}(P_1, P_2)}{\beta_2} \\
&= \frac{\texttt{prem}(T_1, T_2)}{(-1)^{\delta_1+1} \psi_1^{\delta_1} a_1} \quad \text{(by definition of } \beta_2) \\
&= \frac{\texttt{prem}(U_1, T_2)}{(-1)^{\delta_1+1} \psi_1^{\delta_1} a_1^{\delta_0}} \qquad (\text{since } \; U_1 = a_1^{\delta_0 - 1} T_1) \\
&= \frac{(-\psi_1)^{\delta_1+1} T_3}{(-1)^{\delta_1+1} \psi_1^{\delta_1} a_1^{\delta_0}} \qquad (\text{by (30)}, \texttt{prem}(U_1, T_2) = (-\psi_1)^{1+\delta_1} T_3) \\
&= T_3 \qquad (\text{since } \; \psi_1 = a_1^{\delta_0}).
\end{aligned}$$

INDUCTION: Let $i \geq 2$ and assume that part a) is true for $i-1$ and part b) is true for i and $i+1$. Rewriting equation (29) from the Subresultant Theorem in the present terminology:

$$\mathtt{lead}(U_{i-1})^{\delta_{i-1}-1}U_i = \mathtt{lead}(T_i)^{\delta_{i-1}-1}T_i. \tag{35}$$

By inductive hypothesis, $(\psi_{i-1})^{\delta_{i-1}-1}U_i = \mathtt{lead}(P_i)^{\delta_{i-1}-1}P_i$. Comparing leading coefficients, $(\psi_{i-1})^{\delta_{i-1}-1}\mathtt{lead}(U_i) = a_i^{\delta_{i-1}}$. Hence,

$$\mathtt{lead}(U_i) = \frac{a_i^{\delta_{i-1}}}{\psi_{i-1}^{\delta_{i-1}-1}}.$$

But the latter is defined to be ψ_i, hence we have shown part a) for i. For part b), again rewrite equation (30) from the Subresultant Theorem:

$$(-\mathtt{lead}(U_i))^{\delta_i+1}T_{i+2} = \mathtt{prem}(U_i, T_{i+1}). \tag{36}$$

Then

$$\begin{aligned}
\beta_{i+1}P_{i+2} &= \mathtt{prem}(P_i, P_{i+1}) \\
&= \mathtt{prem}(T_i, T_{i+1}) \quad \text{(by inductive hypothesis.)} \\
&= \mathtt{prem}(\frac{\mathtt{lead}(U_{i-1})^{\delta_{i-1}-1}}{\mathtt{lead}(T_i)^{\delta_{i-1}-1}}U_i, T_{i+1}) \quad \text{(by (35))} \\
&= \frac{\psi_{i-1}^{\delta_{i-1}-1}}{a_i^{\delta_{i-1}-1}}\mathtt{prem}(U_i, T_{i+1}) \quad \text{(by inductive hypothesis)} \\
&= \frac{\psi_{i-1}^{\delta_{i-1}-1}}{a_i^{\delta_{i-1}-1}}(-\psi_i)^{\delta_i+1}T_{i+2} \quad \text{(by (36) and part a))} \\
&= \beta_{i+1}T_{i+2}.
\end{aligned}$$

So $T_{i+2} = P_{i+2}$, extending the induction for part b). **Q.E.D.**

Part a) may be reexpressed:

Corollary 15 *The sequence of the ψ_i's in the Subresultant PRS Algorithm on input P_0, P_1 are the principal subresultant coefficients of the subresultant chain of P_0, P_1.*

This confirms the original claim that $\psi_i \in D$ and that (being determinants) their sizes are polynomially bounded when $D = \mathbb{Z}$.

Exercises

Exercise 9.1: (C.-J. Ho) Berkowitz has shown that the determinant of an $m \times m$ matrix has parallel complexity $O(\log^2 m, m^{3.5})$, *i.e.*, can be computed in parallel time $O(\log^2 m)$ using $O(m^{3.5})$ processors. Use this to conclude that the parallel complexity of computing the Subresultant PRS of P_0, P_1 is

$$O(\log^2 m, rnm^{3.5})$$

where $m = \deg(P_0) > \deg(P_1) = n > 0$ and r is the length of the Subresultant PRS. HINT: first compute the principal subresultant coefficients. Then use the parallel-prefix of Ladner-Fisher to obtain a sequence of the r indices of the non-zero principal subresultant coefficients. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture IV
Modular Techniques

We introduce modular techniques based on the Chinese Remainder Theorem, and efficient techniques for modular evaluation and interpolation. This leads to an efficient GCD algorithm in $\mathbb{Z}[X]$.

The rings in this lecture need not be domains.

§1. Chinese Remainder Theorem

The Chinese Remainder Theorem stems from observations about integers such as these: assume we are interested in computing with non-negative integers that are no larger than $3 \cdot 4 \cdot 5 = 60$. Then any integer of interest, say 18, can be represented by its vector of residues modulo $3, 4, 5$ (respectively),

$$(18 \bmod 3, 18 \bmod 4, 18 \bmod 5) = (0, 2, 3).$$

Two numbers in this representation can be added and multiplied, in the obvious componentwise manner. It turns out this sum or product can be faithfully recovered provided it does not exceed 60. For instance $36 = 18 \times 2$ is represented by $(0, 2, 3) \times (2, 2, 2) = (0, 0, 1)$, which represents 36. A limited form of the theorem was stated by Sun Tsŭ (thought to be between 280 and 473 A.D.); the general statement, and proof, by Chhin Chin Shao came somewhat later in 1247 (see [10, p. 271]).

We proceed to put these ideas in its proper algebraic setting (following Lauer [3]). ?? FULL NAME? The above illustration uses an implicit ring, $\mathbb{Z}_3 \otimes \mathbb{Z}_4 \otimes \mathbb{Z}_5$. In general, if $R_1, \ldots, R_n$ are rings, we write

$$R_1 \otimes R_2 \otimes \cdots \otimes R_n \quad \text{or} \quad \bigotimes_{i=1}^{n} R_i$$

for the Cartesian product of $R_1, \ldots, R_n$, endowed with a ring structure by componentwise extension of the individual ring operations. The zero and unity elements of this Cartesian product are $(0, 0, \ldots, 0)$ and $(1, 1, \ldots, 1)$, respectively. For instance, $(u_1, \ldots, u_n) + (v_1, \ldots, v_n) = (u_1 + v_1, \ldots, u_n + v_n)$ where the ith component arithmetic is done in R_i. The reader should always keep in sight where the ring operations are taking place because our notations (to avoid clutter) will not show this explicitly.

Let R be a ring. Two ideals $I, J \subseteq R$ are *relatively prime* if $I + J = R$. This is equivalent to the existence of $a \in I, b \in J$ such that $a + b = 1$. For an ideal I, let R/I denote the quotient ring whose elements are denoted $u + I$ ($u \in R$) under the canonical map. For example, if $R = \mathbb{Z}$ and $I = (n)$ then R/I is isomorphic to $\mathbb{Z}_n$. Two ideals (n) and (m) are relatively prime iff n, m are relatively prime integers. So the ideals $(3), (4), (5)$ implicit in our introductory example are relatively prime. We now present the ideal-theoretic version of the Chinese Remainder Theorem.

Theorem 1 (Chinese Remainder Theorem) *Let $(I_1, \ldots, I_n)$ be a sequence of pairwise relatively prime ideals of R. Then the map*

$$\Phi : u \in R \longmapsto (u + I_1, \ldots, u + I_n) \in \bigotimes_{i=1}^{n} (R/I_i)$$

is an onto homomorphism with kernel

$$\ker \Phi = \bigcap_{i=1}^{n} I_i.$$

In short, this is the content of the Chinese Remainder Theorem:

$$R/(\ker \Phi) \cong (R/I_1) \otimes \cdots \otimes (R/I_n).$$

Proof. It is easy to see that Φ is a homomorphism with kernel $\bigcap_{i=1}^{n} I_i$. The nontrivial part is to show that Φ is onto. Let

$$\overline{u} = (u_1 + I_1, \ldots, u_n + I_n) \in \bigotimes_{i=1}^{n} (R/I_i).$$

We must show the existence of $u \in R$ such that $\Phi(u) = \overline{u}$, *i.e.*,

$$u \equiv u_i (\mathrm{mod}\, I_i) \quad \text{for all} \quad i = 1, \ldots, n. \tag{1}$$

Suppose for each $i = 1, \ldots, n$ we can find b_i such that for all $j = 1, \ldots, n$,

$$b_i \equiv \delta_{i,j} (\mathrm{mod}\, I_j) \tag{2}$$

where $\delta_{i,j}$ is Kronecker's delta-function. Then the desired u is given by

$$u := \sum_{i=1}^{n} u_i b_i.$$

To find the b_i's, we use the fact that for all $i \neq j$, I_i and I_j are relatively prime implies there exist elements

$$a_i^{(j)} \in I_i$$

such that

$$a_i^{(j)} + a_j^{(i)} = 1.$$

We then let

$$b_i := \prod_{\substack{j=1 \\ j \neq i}}^{n} a_j^{(i)}.$$

To see that b_i satisfies equation (2), note that if $j \neq i$ then $a_j^{(i)} \mid b_i$ and $a_j^{(i)} \in I_j$ imply $b_i \equiv 0 (\mathrm{mod}\, I_j)$. On the other hand,

$$b_i = \prod_{\substack{j=1 \\ j \neq i}}^{n} (1 - a_i^{(j)}) \equiv 1 (\mathrm{mod}\, I_i).$$

Q.E.D.

In our introductory example we have $R = \mathbb{Z}$ and the ideals are $I_i = (q_i)$ where $q_1, \ldots, q_n$ are pairwise relatively prime numbers. The numbers $a_i^{(j)}$ can be computed via the extended Euclidean algorithm, applied to each pair q_i, q_j. The kernel of this homomorphism is the ideal $(q_1 q_2 \cdots q_m)$.

The use of the map Φ gives the name "homomorphism method" to this approach. The hope offered by this theorem is that computation in the quotient rings R/I_i may be easier than in R. It is important to notice the part of the theorem stating that the kernel of the homomorphism is $\cap_{i=1}^{n} I_n$. Translated, the price we pay is that elements that are equivalent modulo $\cap_{i=1}^{n} I_n$ are indistinguishable.

Lagrange and Newton interpolations. The proof of the Chinese Remainder Theorem involves constructing an element u that satisfies the system of modular equivalences (1). This is the *modular interpolation problem*. Conversely, constructing the u_i's from u is the *modular evaluation problem*. The procedure used in our proof is called *Lagrange interpolation*. An alternative to Lagrange interpolation is *Newton interpolation*. The basic idea is to build up partial solutions. Suppose we want to solve the system of equivalences $u \equiv u_i (\mathrm{mod}\, I_i)$ for $i = 1, \ldots, n$. We construct a sequence $u^{(1)}, \ldots, u^{(n)}$ where $u^{(i)}$ is a solution to the first i equivalences. We construct $u^{(i)}$ from $u^{(i-1)}$ using

$$u^{(i)} = u^{(i-1)} + (u_i - u^{(i-1)}) \prod_{j=1}^{i-1} a_j^{(i)}$$

where the $a_j^{(i)} \in I_j$ are as in the Lagrange method. Thus $u^{(1)} = u_1$ and $u^{(2)} = u_1 + (u_2 - u_1)a_1^{(2)}$. Lagrange interpolation is easily parallelizable, while Newton interpolation seems inherently sequential in nature. On the other hand, Newton interpolation allows one to build up partial solutions in an on-line manner.

Exercises

Exercise 1.1: Carry out the details for Newton interpolation. □

Exercise 1.2: Give an efficient parallel implementation of the Lagrange interpolation. □

§2. Evaluation and Interpolation

Polynomial evaluation and interpolation. An important special case of solving modular equivalences is when $R = F[X]$ where F is a field. For any set of n distinct elements, $a_1, \ldots, a_n \in F$, the set of ideals $\mathtt{Ideal}(X - a_1), \ldots, \mathtt{Ideal}(X - a_n)$ are pairwise relatively prime. It is not hard to verify by induction that

$$\mathtt{Ideal}(X - a_1) \cap \mathtt{Ideal}(X - a_2) \cap \cdots \cap \mathtt{Ideal}(X - a_n) = \prod_{i=1}^{n} \mathtt{Ideal}(X - a_i) = \mathtt{Ideal}(\prod_{i=1}^{n} (X - a_i)).$$

It is easy to see that $P(X)\,\mathbf{mod}(X - a_i)$ is equal to $P(a_i)$ for any $P(X) \in F[X]$ (Lecture VI.1). Hence the quotient ring $F[X]/(X - a_i)$ is isomorphic to F. Applying the Chinese Remainder Theorem with $I_i := (X - a_i)$, we obtain the homomorphism

$$\Phi : P(X) \in F[X] \mapsto (P(a_1), \ldots, P(a_n)) \in F^n.$$

Computing this map Φ is the *polynomial evaluation problem*; reconstructing a degree $n-1$ polynomial $P(X)$ from the pairs $(a_1, A_1), \ldots, (a_n, A_n)$ such that $P(a_i) = A_i$ for all $i = 1, \ldots, n$ is the *polynomial interpolation problem*. A straight forward implementation of polynomial evaluation has algebraic complexity $O(n^2)$. In Lecture I, we saw that evaluation and interpolation at the n roots of unity has complexity $O(n \log n)$. We now show the general case can be solved almost as efficiently.

The simple observation exploited for polynomial evaluation is this: If $M'(X)|M(X)$ then

$$P(X)\,\mathbf{mod}\, M'(X) = (P(X)\,\mathbf{mod}\, M(X))\,\mathbf{mod}\, M'(X). \tag{3}$$

Theorem 2 *The evaluation of degree $n-1$ polynomials at n arbitrary points has algebraic complexity $O(n \log^2 n)$.*

Proof. We may assume the polynomial $P(X)$ is monic and its degree is $\leq n-1$, where $n = 2^k$ is a power of 2. Suppose we want to evaluate $P(X)$ at $a_0, \ldots, a_{n-1}$. We construct a balanced binary tree T with n leaves. The n leaves are associated with the polynomials $X - a_j$ for $j = 0, \ldots, n-1$. If an internal node u has children v, w with associated polynomials $M_v(X), M_w(X)$ then u is associated with the polynomial $M_u(X) = M_v(X)M_w(X)$. There are 2^i polynomials at level i, each of degree 2^{k-i} (the root has level 0). As the algebraic complexity of multiplying polynomials is $O(n \log n)$, computing all the polynomials associated with level i takes $O(2^i(k-i)2^{k-i}) = O(n \log n)$ operations. Hence we can proceed in a bottom-up manner to compute the set $\{M_u : u \in T\}$ of polynomials in $O(n \log^2 n)$ operations.

We call T the *moduli tree* for $a_0, \ldots, a_{n-1}$. This terminology is justified by our intended use of T: given the polynomial $P(X)$, we want to compute

$$P_u(X) := P(X) \bmod M_u(X)$$

at each node u in T. This is easy to do in a top-down manner. If node u is the child of v and $P(X) \bmod M_v(X)$ has been computed, then we can compute $P_u(X)$ via

$$P_u(X) \leftarrow P_v(X) \bmod M_u(X),$$

by exploiting equation (3). If u is at level $i \geq 1$ then this computation takes $O((k-i)2^{k-i})$ operations since the polynomials involved have degree at most 2^{k-i+1}. Again, the computation at each level takes $O(n \log n)$ operations for a total of $O(n \log^2 n)$ operations. Finally, note that if u is a leaf node with $M_u(X) = X - a_j$, then $P_u(X) = P(X) \bmod M_u(X) = P(a_j)$, which is what we wanted. **Q.E.D.**

To achieve a similar result for interpolation, we use Lagrange interpolation. In the polynomial case, the formula to interpolate $(a_1, A_1), \ldots, (a_n, A_n)$ has the simple form

$$P(X) := \sum_{k=1}^{n} \Delta_k(X) A_k$$

provided $\Delta_k(X)$ evaluated at a_i is equal to $\delta_{k,i}$ (Kronecker's delta). The polynomial $\Delta_k(X)$ can be defined as follows:

$$\begin{aligned} \Delta_k(X) &:= D_k(X)/d_k; \\ D_k(X) &:= \prod_{\substack{i=1 \\ i \neq k}}^{n} (X - a_i); \\ d_k &:= \prod_{\substack{i=1 \\ i \neq k}}^{n} (a_k - a_i). \end{aligned}$$

Note that $\Delta_k(X)$ has degree $n-1$. First consider the problem of computing the d_k's. Let

$$\begin{aligned} M(X) &:= \prod_{i=1}^{n} (X - a_i), \\ M'(X) &= \frac{dM(X)}{dX} \\ &= \sum_{k=1}^{n} \left(\prod_{\substack{i=1 \\ i \neq k}}^{n} (X - a_i) \right). \end{aligned}$$

Then it is easy to see that $M'(a_k) = d_k$. It follows that computing $d_1, \ldots, d_n$ is reduced to evaluating $M'(X)$ at the n points $X = a_1, \ldots, X = a_n$. By the previous theorem, this can be accomplished with $O(n \log^2 n)$ ring operations, assuming we have $M'(X)$. Now $M'(X)$ can be obtained from $M(X)$ in $O(n)$ operations. Since $M(X)$ is the label of the root of the moduli tree T for $a_1, \ldots, a_n$, we can construct $M(X)$ in $O(n \log^2 n)$ operations.

We now seek to split $P(X)$ into two subproblems. First, write $M(X) = M_0(X)M_1(X)$ where

$$M_0(X) = \prod_{i=1}^{n/2}(X - a_i), \qquad M_1(X) = \prod_{i=1+n/2}^{n}(X - a_i).$$

Note that $M_0(X), M_1(X)$ are polynomials in the moduli tree T, which we may assume has been precomputed. Then

$$\begin{aligned} P(X) &= \sum_{k=1}^{n/2} D_k(X)\frac{A_k}{d_k} + \sum_{k=1+n/2}^{n} D_k(X)\frac{A_k}{d_k} \\ &= M_1(X)\sum_{k=1}^{n/2} D_k^*(X)\frac{A_k}{d_k} + M_0(X)\sum_{k=1+n/2}^{n} D_k^*(X)\frac{A_k}{d_k} \\ &= M_1(X)P_0(X) + M_0(X)P_1(X), \end{aligned}$$

where $D_k^*(X) = D_k(X)/M_1(X)$ for $k \leq n/2$ and $D_k^*(X) = D_k(X)/M_0(X)$ for $k > n/2$, and $P_0(X), P_1(X)$ have the same form as $P(X)$ except they have degree $n/2$. By recursively solving for $P_0(X), P_1(X)$, we can reconstruct $P(X)$ in two multiplications and one addition. The multiplications take $O(n \log n)$ time, and so we see that the time $T(n)$ to compute $P(X)$ (given the moduli tree and the d_k's) satisfies the recurrence

$$T(n) = 2T(n/2) + \Theta(n \log n)$$

which has solution $T(n) = \Theta(n \log^2 n)$. It follows that the overall problem has this same complexity. This proves:

Theorem 3 *The interpolation of a degree $n-1$ polynomial from its values at n distinct points has algebraic complexity $O(n \log^2 n)$.*

Solving integer modular equations. There are similar results for the integer case, which we only sketch since they are similar in outline to the polynomial case.

Lemma 4 *Given $s+1$ integers u and $q_1, \ldots, q_s$ where $u\prod_{i=1}^{s} q_i$ has bit size of d, we can form $u_1, \ldots, u_s$ where $u_i = (u \bmod q_i)$ in $O(d\mathcal{L}^2(d))$ bit operations.*

Proof. We proceed in two phases:
1. Bottom-up Phase: Construct a balanced binary tree T with s leaves. With each leaf, we associate a q_i and with each internal node v, we associate the product m_v of the values at its two children. The product of all the m_v's associated with nodes v in any single level has at most d bits. Proceeding in a bottom-up manner, the values at a single level can be computed in time $O(d\mathcal{L}(d))$. Summing over all $\log s$ levels, the time is at most $O(d \log s \cdot \mathcal{L}(d))$.

2. Top-down Phase: Our goal now is to compute at every node v the value $u \bmod m_v$. Proceeding "top-down", assuming that $u \bmod m_v$ at a node v has been computed, we then compute the value $u \bmod m_w$ at each child w of v. This takes time $O(M_B(\log m_v))$. This work is charged to the node v. Summing the work at each level, we get $O(d\mathcal{L}(d))$. Summing over all levels, we get $O(d \log s\mathcal{L}(d)) = O(d\mathcal{L}^2(d))$ time. **Q.E.D.**

Lemma 5 *Given non-negative u_i, q_i $(i = 1, \ldots, s)$ where $u_i < q_i$, the q_i's being pairwise co-prime and each q_i having bit size d_i, we can compute u satisfying $u \equiv u_i (\bmod q_i)$ in bit complexity*

$$O(d^2\mathcal{L}^2(d))$$

where $d = \sum_{i=1}^{s} d_i$.

Proof. We use the terminology in the proof of the Chinese Remainder Theorem in §1.
1. Each pair $a_i^{(j)}, a_j^{(i)}$ is obtained by computing the extended Euclidean algorithm on q_i, q_j. We may assume $a_i^{(i)} = 1$. Note that $a_i^{(j)}$ has bit size $\leq d_j + d_i$. This takes time $O((d_i + d_j)\mathcal{L}^2(d))$. Summing over all i, j, we get a bound of $O(d^2\mathcal{L}^2(d))$.
2. For each $i = 1, \ldots, s$, we compute $b_i = \prod_{j=1, j\neq i}^{s} a_j^{(i)}$ as follows. First note that the bit size of b_i is $\leq sd_i + d$. As in the previous lemma, compute the product b_i using the pattern of a balanced tree T_i with leaves labeled by $a_j^{(i)}$ $(j = 1, \ldots, s)$. This takes time $O((sd_i + d)\mathcal{L}(d))$ per level or $O((sd_i + d) \log s\mathcal{L}(d))$ for the entire tree. Summed over all i, the cost is $O(sd \log s\mathcal{L}(d))$.
3. Finally, we compute the answer $u = \sum_{i=1}^{s} u_i b_i$. Each term $u_i b_i$ can be computed in $O((sd_i + d)\mathcal{L}(d))$. Thus the sum can be computed in $O(sd\mathcal{L}(d))$ time.
4. Summing over all the above, we get a bound of $O(d^2\mathcal{L}^2(d))$. **Q.E.D.**

Exercises

Exercise 2.1: Solve the following modular interpolation problems:
i) $u \equiv 1 (\bmod 2), u \equiv 1 (\bmod 3), u \equiv 1 (\bmod 5)$. (Of course, we know the answer, but you should go through the general procedure.)
ii) $u \equiv 1 (\bmod 2), u \equiv 1 (\bmod 3), u \equiv 1 (\bmod 5), u \equiv 3 (\bmod 7)$.
iii) $P(0) = 2, P(1) = -1, P(2) = -4, P(3) = -1$ where $\deg P = 3$. □

Exercise 2.2:
i) Verify the assertions about polynomial evaluation and interpolation, in particular, equation (3).
ii) Let $J = \texttt{Ideal}(X - a) \cap \texttt{Ideal}(X - b)$ where $a, b \in F$ are distinct. Prove that $F[X]/J$ and $F[X]/\texttt{Ideal}(X^2)$ are not isomorphic as rings. □

Exercise 2.3: It should be possible to improve the complexity of modular integer interpolation above. □

§3. Finding Prime Moduli

To apply the Chinese Remainder Theorem for GCD in $\mathbb{Z}[X]$, we need to find a set of relatively prime numbers whose product is sufficiently large. In this section, we show how to find a set of prime numbers whose product is larger than some prescribed bound.

The following function is useful: for a positive integer n, $\theta(n)$ is defined to be the natural logarithm of the product of all primes $\leq n$. The following estimate from Langemyr [11] (see also Rosser-Schoenfeld [15]) is useful.

Proposition 6 *For $n \geq 2$,*

$$0.31n < \theta(n) < 1.02n.$$

Consider the following problem: given a number $N > 0$, list all the primes $\leq N$. We can do this quite simply using the *sieve of Eratosthenes* (276-194 B.C.). Let L be a Boolean array of length N, initialized to 1. We want $L[i] = 1$ to indicate that i is a candidate for a prime. We can immediately set $L[1] = 0$. In the general step, let p be the smallest index such that $L[p] = 1$, and p is prime. Then we do the following "step" repeatedly, until the entire array is set to 0:

$$\text{Output } p \text{ and set } L[ip] = 0 \text{ for } i = 1, 2, \ldots, \lfloor \frac{N}{p} \rfloor. \tag{4}$$

The correctness of this procedure is easy. Note that (4) costs $\frac{N}{p}$ array accesses. The total number of array accesses over all steps is

$$\sum_{p<N} \frac{N}{p} = N \cdot \sum_{p<N} \frac{1}{p}$$

where the summation is over all primes less thatn N. Clearly $\sum_{p<N} \frac{1}{p} \leq \sum_{i=1}^{N} \frac{1}{i} = O(\log N)$. But it is well-known [6, p.351] that, in fact,

$$\sum_{p<N} \frac{1}{p} = \ln \ln N + O(1).$$

So the total number of array accesses is $O(N \log \log N)$. In the RAM complexity model, this procedure has a complexity of $O(N\mathcal{L}(N))$.

Lemma 7 *We can find a list of primes whose product is at least n in time $O(\log n\, \mathcal{L}(\log n))$ in the RAM model.*

Proof. Choose $N = \lceil \frac{\ln n}{0.31} \rceil$. Then $\theta(N) > 0.31N \geq \ln n$. So the product of all primes at most N is at least n. The above algorithm of Erathosthenes has the desired complexity bound. **Q.E.D.**

§4. Lucky homomorphisms for the GCD

Let p be a fixed prime. The key homomorphism we consider is the map

$$(\cdot)_p : \mathbb{Z}[X] \to \mathbb{Z}_p[X] \tag{5}$$

where $(A)_p$ denotes the polynomial obtained by the modulo p reduction of each coefficient of $A \in \mathbb{Z}[X]$. Where there is no ambiguity, we write A_p for $(A)_p$. We will also write

$$A \equiv A_p (\mathrm{mod}\, p).$$

Note that the GCD is meaningful in $\mathbb{Z}_p[X]$ since it is a UFD. This section investigates the connection between $\mathtt{GCD}(A, B)_p$ and $\mathtt{GCD}(A_p, B_p)$.

Begin with the observation

$$(AB)_p = (A_p \cdot B_p), \qquad A, B \in \mathbb{Z}[X]$$

where the second product occurs in $\mathbb{Z}_p[X]$. It follows that

$$A|B \text{implies} \;\; A_p|B_p.$$

Similarly, $\texttt{GCD}(A,B)|A$ implies $\texttt{GCD}(A,B)_p|A_p$. By symmetry, $\texttt{GCD}(A,B)_p|B_p$. Hence

$$\texttt{GCD}(A,B)_p|\texttt{GCD}(A_p,B_p).$$

However, it is not generally true that

$$\texttt{GCD}(A,B)_p = \texttt{GCD}(A_p,B_p). \tag{6}$$

A simple example is

$$A = X - 1, \quad B = X + 1$$

and $p = 2$. Here $A_p = B_p$ and hence $\texttt{GCD}(A_p,B_p) = A_p$. But A, B are relatively prime so that $\texttt{GCD}(A,B)_p = 1$.

We study the conditions under which equation (6) holds. Roughly speaking, we call such a choice of p "lucky". The basic strategy for the modular GCD algorithm is this: we pick a set $p_1, \ldots, p_n$ of lucky primes and compute $\texttt{GCD}(A_{p_i}, B_{p_i})$ for $i = 1, \ldots, n$. Since $\texttt{GCD}(A_{p_i}, B_{p_i}) = \texttt{GCD}(A,B)_{p_i}$, we can reconstruct $G = \texttt{GCD}(A,B)$ by solving the system of modular equivalences

$$G \equiv \texttt{GCD}(A,B)_{p_i} \pmod{p_i}.$$

Definition: A prime $p \in \mathbb{N}$ is *lucky* for $A, B \in \mathbb{Z}[X]$ if p does not divide $\texttt{lead}(A) \cdot \texttt{lead}(B)$ and

$$\deg(\texttt{GCD}(A_p,B_p)) = \deg(\texttt{GCD}(A,B)).$$

To be sure, this definition may appear odd because we are trying to compute $\texttt{GCD}(A,B)$ via mod p computation where p is lucky. But to know if p is lucky, the definition requires us to know the degree of $\texttt{GCD}(A,B)$.

Lemma 8 *If p does not divide at least one of $\texttt{lead}(A)$ and $\texttt{lead}(B)$ then $\texttt{GCD}(A_p,B_p)$ has degree at least as large as $\texttt{GCD}(A,B)$.*

Proof. Let $G = \texttt{GCD}(A,B)$, $g = \texttt{lead}G$, $a = \texttt{lead}A$ and $b = \texttt{lead}B$. Note that $g|\texttt{GCD}(a,b)$. If p does not divide a, then p does not divide $\texttt{GCD}(a,b)$ and hence p does not divide g. Then $\deg(G) = \deg(G_p)$. But $\deg(G_p) \leq \deg \texttt{GCD}(A_p,B_p)$. **Q.E.D.**

To generalize this lemma, suppose we have a homomorphism between domains,

$$\Phi : D \to D'$$

extended to

$$\Phi : D[X] \to D'[X]$$

in the coefficient-wise fashion (but still denoted by the same symbol Φ). Let

$$(A_m, A_{m-1}, \ldots, A_0)$$

be a subresultant chain in $D[X]$ and

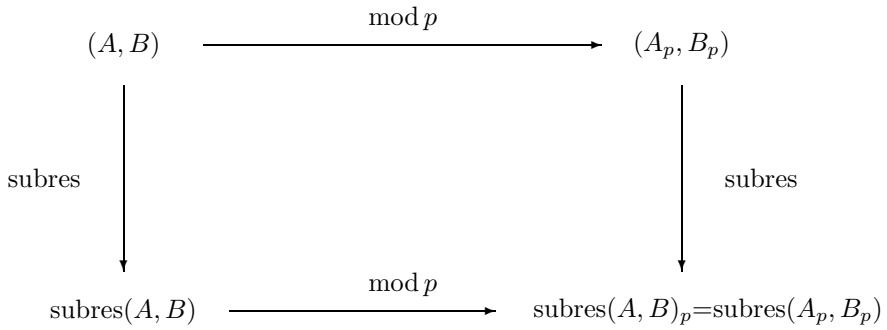
$$(\overline{A}_m, \overline{A}_{m-1}, \ldots, \overline{A}_0)$$

be the Φ-image of the chain, $\overline{A}_i = \Phi(A_i)$.

Lemma 9 *If* $\deg(A_m) = \deg(\overline{A}_m)$ *and* $\deg(A_{m-1}) = \deg(\overline{A}_{m-1})$ *then* $(\overline{A}_m, \ldots, \overline{A}_0)$ *is a subresultant chain in* $D'[X]$.

Proof. The hypothesis of this lemma is simply that the leading coefficients of A_{m-1} and A_m must not be in the kernel of Φ. The result follows since subresultants are determinants of matrices whose shape is solely a function of the degrees of the first 2 polynomials in the chain. **Q.E.D.**

We conclude that the following diagram commutes if p does not divide $\texttt{lead}(A)\texttt{lead}(B)$:



Here, $\text{subres}(P, Q)$ denotes the subresultant chain of P, Q in $\mathbb{Z}[X]$ or in $\mathbb{Z}_p[X]$. The following generalizes lemma 8.

Lemma 10 *Under the same assumption as the previous lemma, if* $\overline{A}_i$ $(i = 0, \ldots, m)$ *is nonzero then* $\deg(\overline{A}_i)$ *also occurs as* $\deg(A_j)$ *for some* $j \leq i$. *In particular,* $\texttt{GCD}(\overline{A}_m, \overline{A}_{m-1})$ *has degree at least as large as that of* $\texttt{GCD}(A_m, A_{m-1})$.

Proof. By the Block Structure Theorem (§III.7), if $\deg(\overline{A}_i) = j$ then $\overline{A}_j$ is regular and hence A_j is regular. The conclusion about the GCD uses the fact that the non-zero subresultant of smallest degree is similar to the GCD. **Q.E.D.**

We justify our definition of luckiness:

Lemma 11 (Luckiness Lemma) *If* p *is lucky for* A *and* B *then* $\texttt{GCD}(A_p, B_p) \sim \texttt{GCD}(A, B)_p$.

Proof. Let $(A_m, \ldots, A_0)$ be the subresultant chain for A, B and $\overline{A}_i = (A_i)_p$ for $i = 0, \ldots, m$. Lemma 9 implies that $(\overline{A}_m, \ldots, \overline{A}_0)$ is a subresultant sequence for A_p, B_p. By definition, p is lucky

means that if $\texttt{GCD}(A, B)$ has degree d then $\overline{A}_d$ and A_d are the last nonzero polynomials in their respective subresultant sequence. The lemma follows from

$$\texttt{GCD}(A, B)_p \sim (A_d)_p = \overline{A}_d \sim \texttt{GCD}(A_p, B_p).$$

Q.E.D.

Lemma 12 *Let $A, B \in \mathbb{Z}[X]$ with $n = \max\{\deg A, \deg B\}$ and $N = \max\{\|A\|_2, \|B\|_2\}$. If P is the product of all the unlucky primes of A, B, then*

$$P \leq N^{2n+2}.$$

Proof. If $\texttt{GCD}(A, B)$ has degree d then the dth principal subresultant coefficient C_d is non-zero. If p is unlucky and does not divide ab then by lemma 8, $\deg \texttt{GCD}(A, B) < \deg \texttt{GCD}(A_p, B_p)$. This means $p|C_d$. Hence all unlucky primes for A, B are among the divisors of $a \cdot b \cdot C_d$, where $a = \texttt{lead}(A)$ and $b = \texttt{lead}(B)$. The product of all prime divisors of $a \cdot b \cdot C_d$ is at most $|a \cdot b \cdot C_d|$. Since $|a| \leq N, |b| \leq N$, the lemma follows if we show

$$|C_d| \leq N^{2n}.$$

To see this, C_d is the determinant of a submatrix M of the Sylvester matrix of A, B. Each row r_i of M has non-zero entries coming from coefficients of A or of B. Thus $\|r_i\|_2 \leq N$. Since M has at most $2n$ rows, the bound on $|C_d|$ follows immediately from Hadamard's determinant bound (§IX.1).

Q.E.D.

§5. Coefficient Bounds for Factors

Assume that

$$A(X), B(X) \in \mathbb{C}[X]$$

where $B \,|\, A$. We derive an upper bound on $\|B\|_2$ in terms of $\|A\|_2$. Such bounds are needed in our analysis of the modular GCD algorithm and useful in other contexts (e.g., factorization algorithms). Begin with the following equality:

Lemma 13 *Let $A(X) \in \mathbb{C}[X], c \in \mathbb{C}$. Then $\|(X - c) \cdot A(X)\|_2 = \|(\overline{c}X - 1) \cdot A(X)\|_2$, where $\overline{c}$ is the complex conjugate of c.*

Proof. If $A(X) = \sum_{i=0}^{m} a_i X^i$ then

$$\begin{aligned}
(X - c) \cdot A(X) &= \sum_{i=0}^{m+1} (a_{i-1} - ca_i)X^i, \qquad (a_{-1} = a_{m+1} = 0), \\
\|(X - c) \cdot A(X)\|_2 &= \sum_{i=0}^{m+1} (a_{i-1} - ca_i)(\overline{a_{i-1}} - \overline{ca_i}) \\
&= \sum_{i=0}^{m+1} ((|a_{i-1}|^2 + |c|^2 \cdot |a_i|^2) - (ca_i\overline{a_{i-1}} + \overline{ca}_i a_{i-1})) \\
&= (1 + |c|^2) \sum_{i=0}^{m} |a_i|^2 - \sum_{i=1}^{m} (ca_i\overline{a_{i-1}} + \overline{ca}_i \cdot a_{i-1}).
\end{aligned}$$

Similarly, $\|(\overline{c}X - 1) \cdot A(X)\|_2$ can be expanded to give the same expression. **Q.E.D.**

For any $A(X) \in \mathbb{C}[X]$ whose complex roots are $\alpha_1, \ldots, \alpha_m$ (not necessarily distinct), define the *measure,*
it see under polynomial of A to be

$$M(A) = |a| \cdot \prod_{i=1}^{m} \max\{1, |\alpha_i|\}$$

where a is the leading coefficient of $A(X)$. The effect of the max function is simply to discard from the product any root within the unit disc. Measures have the nice property that

$$A|B \text{implies} \quad M(A)/|a| \leq M(B)/|b|$$

where $a = \texttt{lead}(A)$ and $b = \texttt{lead}(B)$. The following proof is from Mignotte [13]:

Theorem 14 *Let $A(X) \in \mathbb{C}[X]$ has lead coefficient a and tail coefficient a'.*
(i) Then $M(A) \leq \|A\|_2$.
(ii) If A is not a monomial then

$$M(A)^2 + \left(\frac{aa'}{M(A)}\right)^2 \leq \|A\|_2^2.$$

Proof. Let $\alpha_1, \ldots, \alpha_m \in \mathbb{C}$ be the not-necessarily distinct roots of A, arranged so that

$$|\alpha_1| \geq \cdots \geq |\alpha_k| \geq 1 > |\alpha_{k+1}| \geq \cdots \geq |\alpha_m|$$

for some $k = 0, \ldots, m$. By repeated applications of the previous lemma,

$$\begin{aligned}
\|A\|_2 &= \|a \prod_{i=1}^{m} (X - \alpha_i)\|_2 \\
&= \|a(\overline{\alpha}_1 X - 1) \prod_{i=2}^{m} (X - \alpha_i)\|_2 \\
&= \cdots \\
&= \|a \prod_{j=1}^{k} (\overline{\alpha}_j X - 1) \prod_{i=k+1}^{m} (X - \alpha_i)\|_2.
\end{aligned}$$

Let B denote the last polynomial,

$$B = a \prod_{j=1}^{k} (\overline{\alpha}_j X - 1) \prod_{i=k+1}^{m} (X - \alpha_i).$$

Then

$$\texttt{lead}B = a \prod_{j=1}^{k} \overline{\alpha}_j, \qquad \texttt{tail}B = a \prod_{i=k+1}^{n} \alpha_i = \frac{aa'}{M(A)}.$$

Clearly $\|A\|_s \geq |\texttt{lead}B| = a \prod_{j=1}^{k} |\alpha_j| = M(A)$, proving (i). Part (ii) is also immediate since $\|A\|_2 \geq |\texttt{lead}B|^2 + |\texttt{tail}B|^2$ when A is not a monomial. **Q.E.D.**

Part (i) is often called the bound of Landau (1905); the improvement in (ii) is attributed to Vicente Gonçalves (1956) (cf. [16, p. 162]).

Corollary 15 *If α is a root of $A(X) \in \mathbb{C}[X]$ then $|\alpha| \leq \|A\|_2/|a|$ where $a = \texttt{lead}A$.*

Lemma 16 *If $B(X) = \sum_{i=0}^{n} bX^i$ then $|b_{n-i}| \leq \binom{n}{i} M(B)$.*

Proof. Let $B(X) = \sum_{i=0}^{n} b_i X^i = b\prod_{i=1}^{n}(X - \beta_i)$. Then for $i = 0, \ldots, n$:

$$\begin{aligned} |b_{n-i}| &\leq |b| \sum_{1\leq j_1<\cdots<j_i\leq n} |\beta_{j_1}\beta_{j_2}\cdots\beta_{j_i}| \\ &\leq \sum_{1\leq j_1<\cdots<j_i\leq n} M(B) \\ &= \binom{n}{i} M(B). \end{aligned}$$

Q.E.D.

Theorem 17 (Mignotte) *Let $A, B \in \mathbb{C}[X]$, $b = \texttt{lead}(B)$, $a = \texttt{lead}(A)$ and $n = \deg(B)$. If $B|A$ then*

$$\begin{aligned} \|B\|_\infty &\leq \left|\frac{b}{a}\right| \cdot \binom{n}{\lfloor n/2 \rfloor} \cdot \|A\|_2 \\ \|B\|_1 &\leq \left|\frac{b}{a}\right| \cdot 2^n \cdot \|A\|_2 \end{aligned}$$

Proof. The first inequality is an immediate consequence of the previous lemma, using the fact $\binom{n}{i} \leq \binom{n}{\lfloor n/2 \rfloor}$ and $M(B) \leq |b/a|M(A) \leq |b/a| \cdot \|A\|_2$. For the second inequality, we bound $\|B\|_1$ by suming up the upper bounds (again from previous lemma) for $|b_0|, \ldots, |b_n|$, giving

$$\|B\|_1 \leq 2^n M(B).$$

Q.E.D.

Since $\|B\|_2 \leq \|B\|_1$ (§0.10), we get an upper bound on $\|B\|_2$ as well. If C, B are integer polynomials and $C|B$ then $|\texttt{lead}(C)/\texttt{lead}(B)| \leq 1$. Therefore:

Corollary 18 *Let $A, B, C \in Z[X]$ with $C = \texttt{GCD}(A, B)$. Then*

$$\|C\|_1 \leq 2^n \min\{\|A\|_2, \|B\|_2\}$$

where $\deg(A) \geq \deg(B) = n$.

We refer to Mignotte [13] for more information about measures. The above bounds have been sharpened by Beauzamy [1] by using a weighted L_2-norm of polynomials: for a polynomial A of degree m with coefficients a_i, define

$$[A]_2 := \left(\sum_{i=0}^{m} \frac{|a_i|^2}{\binom{m}{i}}\right)^{1/2}.$$

Then if $B|A$, it is shown that

$$\|B\|_\infty \leq \frac{3^{3/4}}{2\sqrt{\pi}} \frac{3^{m/2}}{\sqrt{m}} [A]_2.$$

In general, for any norm $N(A)$ on polynomials, we can define two constants $\beta > 1$ and $\delta > 1$ which are the smallest values such that

$$N(B)N(C) \leq \delta^n N(A), \quad N(B) \leq \beta^n N(A)$$

holds for all monic polynommials A, B, C such that $A = B \cdot C$ and $\deg(A) = n$. For any two standard norms, $N(A)$ and $N'(A)$, we have the basic inequality $N(A) \leq (n+1)N'(A)$ (§0.9). Hence the constants δ, β are the same for all such norms. It is also easy to see that $\delta \geq \beta$, since $N(C) \geq 1$. The above result of Mignotte implies $\beta \leq 2$. Boyd [2] has determined $\delta = M(P_1) = 1.79162\ldots$ and $\beta = M(P_0) = 1.38135\ldots$, where $P_1 = 1 + X + Y - XY$ and $P_0 = 1 + X + Y$.

Exercises

Exercise 5.1: Conclude from the bounds for the coefficient sizes of factors that the problem of factorizing integer polynomials is finite. □

Exercise 5.2: (Davenport-Trager) Construct examples in which the coefficients of the GCD of $A, B \in \mathbb{Z}[X]$ grow much larger than the coefficients of A, B. HINT: $A = (X+1)^{2k}(X-1)$, $B = (X+1)^{2k}(X^2 - X + 1)$. □

Exercise 5.3: (Cassels) Let z, β be real or complex, $|z| \leq 1$. Then

$$\begin{array}{rcll} |z - \overline{\beta}| & \leq & |1 - \beta z| & \text{if } |\beta| < 1, \\ |z - \overline{\beta}| & \geq & |1 - \beta z| & \text{if } |\beta| > 1. \end{array}$$

Equality holds in both cases iff $|z| = 1$. □

Exercise 5.4: (i) Show that the measure of a polynomial $A(Z)$ can also be defined as

$$M(A) = \exp\left(\int_0^1 \log|A(e(\theta))| d\theta\right)$$

where $e(\theta) = \exp(2\pi \mathbf{i}\theta)$. If $A = A(X_1, \ldots, X_n)$ is a multivariate polynomial, this definition generalizes to the multiple integral:

$$M(A) = \exp\left(\int_0^1 \cdots \int_0^1 \log|A(e(\theta_1), \ldots, e(\theta_n))| d\theta_1 \cdots d\theta_n\right).$$

(We can view $M(A)$ is the geometric mean of $|A|$ on the torus T^n.) (ii) (Mahler) If $n = \deg A$ then

$$\binom{n}{\lfloor n/2 \rfloor}^{-1} \|A\|_\infty \leq M(A) \leq \|A\|_\infty \sqrt{n+1}.$$

$$2^{-n}\|A\|_1 \leq M(A) \leq \|A\|_1.$$

(iii) $M(A \pm B) \leq \|A\|_1 + \|B\|_1 \leq 2^n(M(A) + M(B))$.
(iv) (Duncan [4])

$$\binom{2n}{n}^{-1/2} \|A\|_2 \leq M(A) \leq \|A\|_1.$$

(v) $M(A) \leq \|A\|_2$. HINT: Use induction on degree, Jensen's inequality $\int_0^1 \log|F(t)|dt \leq \log\int_0^1 |F(t)|dt$ and Parseval's formula for a univariate polynomial $F(X)$: $\int_0^1 (\sum_{i=1}^n |F(e^{\mathbf{i}2\pi t})|^2 dt = \|F\|_2^2$. □

Exercise 5.5: Referring to the weighted L_2-norm $[A]_2$ for a univariate A:
(i) $[A]_2 \leq \|A\|_2$.
(ii) $[AB]_2 \leq [A]_2[B]_2$.
(iii) $\binom{d}{\lfloor d/2\rfloor}^{-1/2} M(A) \leq [A]_2 \leq 2^{d/2} M(A)$.
(iv) Compare the bounds of Beauzamy to that of Mignotte.

□

Exercise 5.6: If $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, let $|\alpha| := \sum_{i=1}^n \alpha_i$ and $\alpha! := \alpha_1! \cdots \alpha_n!$. If $A(X_1, \ldots, X_n) = \sum_\alpha a_\alpha \mathbf{X}^\alpha$ is homogeneous, then define

$$[A]_2 := \left(\sum_{|\alpha|=m} \frac{\alpha!}{m!} |a_\alpha|^2 \right)^{1/2}$$

Clearly $[A]_2 \leq \|A\|_2$.
(i) (Beauzamy,Bombieri,Enflo,Montgomery) $[AB]_2 \geq \sqrt{\frac{m!n!}{(m+n)!}}[A]_2[B]_2$.
(ii) (Beauzamy) In case A is not homogeneous, define $[A]_2$ to be $[A^\wedge]_2$ where $A^\wedge$ is the homogenization of A with respect to a new variable. If A, B are univariate and $B|A$ and $A(0) \neq 0$ then $\|B\|_\infty \leq \frac{3^{3/4} 3^{d/2}}{2\sqrt{\pi d}} [A]_2$. □

§6. A Modular GCD algorithm

We present the modular algorithm of Brown and Collins for computing the `GCD` of $A, B \in \mathbb{Z}[X]$ where

$$\begin{aligned} n_0 &= \max\{\deg(A), \deg(B)\}, \\ N_0 &= \max\{\|A\|_2^2, \|B\|_2^2\}. \end{aligned}$$

We have shown that the product of all unlucky primes is $\leq N_0^{n_0+2}$, and that each coefficient of $\texttt{GCD}(A, B)$ has absolute value $\leq 2^{n_0} N_0$. Let

$$K_0 = 2 \cdot N_0^{n_0+2} \cdot 2^{n_0} N_0 = 2^{n_0+1} N_0^{n_0+3}.$$

First compute the list of all initial primes until their product is just $\geq K_0$. The lucky primes in this list have product at least

$$K_0 \cdot N_0^{-(n_0+2)} = 2^{n_0+1} N_0.$$

To identify these lucky primes, we first omit from our list all primes that divide the leading coefficients of A or of B. Among the remaining primes p, we compute A_p, B_p and then $\texttt{GCD}(A_p, B_p)$. Let $\delta(p)$ be the degree of $\texttt{GCD}(A_p, B_p)$ and

$$\delta^* = \min_p \delta(p).$$

Clearly δ^* is the degree of $\texttt{GCD}(A, B)$ since there is a lucky prime p_0 that remains and this $\delta(p_0)$ would attain the minimum δ^*. We discard all p where $\delta(p) > \delta^*$. We have now identified the set L^* of all lucky primes in our original list.

Let

$$C(X) := \sum_{i=0}^{\delta^*} c_i X^i \sim \texttt{GCD}(A, B). \tag{7}$$

Our goal is to compute some such $C(X)$ using the Chinese Remainder Theorem. We must be careful as $C(X)$ is determined by $\texttt{GCD}(A, B)$ only up to similarity. To see what is needed, assume that for each lucky p, we have computed

$$C_p(X) := \sum_{i=0}^{\delta^*} c_{i,p} X^i \sim \texttt{GCD}(A_p, B_p). \tag{8}$$

How shall we ensure that these $C_p(X)$'s are "consistent"? That is, is there one polynomial $C(X) \in \mathbb{Z}[X]$ such that each $C_p(X)$ is the image of $C(X)$ under the canonical map (5)? We shall *pick* equation (7) such that

$$\texttt{lead}(C) = c_{\delta^*} = \texttt{GCD}(\texttt{lead}(A), \texttt{lead}(B)).$$

Such a choice exists because if $C|A$ and $C|B$ then $\texttt{lead}(C)|\texttt{GCD}(\texttt{lead}(A), \texttt{lead}(B))$. "Consistency" then amounts to the requirement

$$c_{\delta^*,p} = (c_{\delta^*} \bmod p)$$

for each p. Now we can reconstruct the c_i's in equation (7) as the solution to the system of congruences.

$$c_i \equiv c_{i,p} \pmod{p}, \quad p \in L^*.$$

The correctness of this solution depends on the Luckiness Lemma, and the fact that the product of lucky primes (being at least $2^{n_0+1}N_0$) is at least twice as large as $|c_i|$.

To recapitulate:

1. Compute the list of initial primes whose product is $\geq K_0$.

2. Omit all those primes that divide $\texttt{lead}(A)$ or $\texttt{lead}(B)$.

3. For each remaining prime p, compute A_p, B_p, $C_p(X) \sim \texttt{GCD}(A_p, B_p)$ and $\delta(p)$.

4. Find δ^* as the minimum of the $\delta(p)$'s. Omit the remaining unlucky primes.

5. Use Chinese Remainder to reconstruct $C(X) \sim \texttt{GCD}$.

Timing Analysis. We bound the time of the above steps. Let $k_0 = \log K_0 = O(n_0 \log N_0)$.

1. This step takes $O(k_0 \mathcal{L}(k_0))$.

2. This is negligible compared to other steps: for each prime p, to check if $p|\texttt{lead}(A)$ takes $O(\log p + \log N_0)\mathcal{L}(\log p + \log N_0))$. Summed over all p, this is $k_0 \mathcal{L}(N_0)\mathcal{L}(k_0)$.

3(a). To compute A_p (similarly for B_p) for all p, we exploit the integer evaluation result (lemma 4). Hence all A_p can be computed in time $O(n_0 k_0 \mathcal{L}^2(k_0))$, the n_0 coming from the coefficients of A.

3(b). To compute $\texttt{GCD}(A_p, B_p)$ (for any p) requires (Lecture III)

$$n_0\mathcal{L}^2(n_0)$$

operations in $\mathbb{Z}_p$, and each $\mathbb{Z}_p$ operation costs $O(\mathcal{L}^2(\log p)) = O(\log p)$. Summing over all p's, we get order of:

$$n_0\mathcal{L}^2(n_0)\sum_p \log p = n_0\mathcal{L}^2(n_0)\cdot k_0.$$

4. Negligible.

5. Applying the integer interpolation result (lemma 5), we get a time of

$$O(k_0^2\mathcal{L}^2(k_0)).$$

Summing up these costs, we conclude:

Theorem 19 *The above algorithm computes the* $\texttt{GCD}$ *of* $A, B \in \mathbb{Z}[X]$ *in time*

$$O(k_0^2\mathcal{L}^2(k_0)).$$

If the input A, B *has size* n*, the complexity bound becomes*

$$O(n^2\mathcal{L}^2(n)).$$

In practice, one could try to rely on luck for lucky primes and this can be the basis of fast probabilistic algorithms.

§7. What else in GCD computation?

There are several further directions in the study of $\texttt{GCD}$ computation:

1. extend to multivariate polynomials
2. extend to multiple GCD
3. algorithms that are efficient for sparse polynomials
4. extend to other number fields
5. use randomization techniques to speed up computation

1. In principle, we know how to compute the multiple GCD of a set $S \subseteq D[X_1, \ldots, X_n]$ where D is a UFD: treating S_n as polynomials in X_n, then $G = \texttt{GCD}(S)$ can be factored (§III.2) into its content and primitive part: $G = \texttt{cont}(G)\texttt{prim}(G)$. But $\texttt{cont}(G) = \texttt{GCD}(\texttt{cont}(S))$ and $\texttt{prim}(G) = \texttt{GCD}(\texttt{prim}(S))$ where $\texttt{cont}(S) = \{\texttt{cont}(A) : A \in S\}$ and $\texttt{prim}(S) = \{\texttt{prim}(A) : A \in S\}$. Now $\texttt{cont}(A)$ amounts to computing the multiple GCD of the coefficients of A, and this can be achieved by using induction on n. The basis case amounts to computing GCD in D, which we assume is known. For $n > 1$, the computations of $\texttt{cont}(S)$ and $\texttt{prim}(S)$ are reduced to operations in $D[X_1, \ldots, X_{n-1}]$. Finally, $\texttt{cont}(G)$ is also reduced to GCD in $D[X_1, \ldots, X_{n-1}]$ and $\texttt{prim}(G)$ is done using, say, the methods of the previous lecture.

2. As the preceding procedure shows, even if we start with computing a simple GCD of two polynomials, we may recursively have to deal with multiple GCD. Although multiple GCD can be reduced to simple GCD, the efficient computation of the multiple GCD is not well-understood. One such algorithm is the Jacobi-Perron algorithm for multiple GCD for integers. Chung-jen Ho [7, 8] has generalized the concept of subresultants to several univariate polynomials, and presented a multiple GCD algorithm for $F[X]$.

3. The use of sparse polynomial representation is important (especially in multivariable case) but it makes apparently "simple" problems such as univariate GCD inherently intractable (see §0.5).

4. Another direction is to consider factorization and GCD problems in non-commutative rings. For instance, see [9, chapter 14] for the ring of integer matrices that has a form of unique factorization and GCD. The study of GCD algorithms for quadratic integer rings that are UFD's is related to the problem of finding shortest vectors in a 2-dimensional lattice. We return to the last topic in Lecture IX.

§8. Hensel Lifting

[NOTE: the introduction to this chapter needs to be change to reflect this insertion. Some general remarks – including the terminology "homomorphism techniques" for modular techniques.]

Consider the problem of computing the `GCD` of two multivariate polynomials. The approach of the preceding sections can be generalized for this problem. Unfortunately, the number of homomorpic subproblems we need to solve grows exponentially with the number of variables. This section investigates an alternative approach called "Hensel lifing". Instead of a growing number of homomorphic subproblems, we solve one homomorphic subproblem, and then "lift" the solution back to the original domain. The emphasis here is on the "lifting", which turns out to be computationally more expensive than in Chinese Remaindering methods. Fortunately, for many multivariate computations such as GCD and factorization, this approach turns out to be more efficient.

One of the first papers to exploit Hensel's lifting is Musser's thesis [14] on polynomial factorization. Yun [17, 19] observed that the lifing process in Hensel's method is the algebraic analogue of Newton's iteration for finding roots (see Chapter 6, Section 10). The book [5] gives an excellent treatment of modular techniques. We follow the general formulation of Lauer [12].

Motivating example: polynomial pactorization. Before considering the general framework, consider the special case of integer polynomials. Let $A, B, C \in \mathbb{Z}[X]$. Fix a prime number p and for any $n \geq 1$, consider the homomorphism

$$(\cdot)_{p^n} : \mathbb{Z}[X] \to \mathbb{Z}_p[X]$$

(cf. Section 4). As usual, we write $A \equiv B (\mathrm{mod}\, p^n)$ if $(A)_{p^n} = (B)_{p^n}$.

Lemma 20 (Hensel) *uppose*

$$AB \equiv C (\mathrm{mod}\, p^n)$$

and A, B are relatively prime, modulo p^m for some $1 \leq m \leq n$. Then there exists $A^*, B^* \in \mathbb{Z}[X]$ such that

$$A^* B^* = C \bmod p^{n+m}$$

and $A^* \equiv A (\mathrm{mod}\, p^n)$, $B^* \equiv B (\mathrm{mod}\, p^n)$.

Proof. Suppose $AB = C + p^n\widehat{C}$, $A^* = A + p^n\widetilde{A}$ and $B^* = B + p^n\widetilde{B}$. Here $\widetilde{C}$ is determined by the given data, but we will choose $\widetilde{A}$ and $\widetilde{B}$ to verify the lemma. Then we have

$$\begin{aligned} A^*B^* &= (A + p^n\widetilde{A})(B + p^n\widetilde{B}) \\ &\equiv AB + p^n(A\widetilde{B} + \widetilde{A}B)(\mathrm{mod}\, p^{n+m}) \\ &\equiv C + p^n(\widetilde{C} + A\widetilde{B} + \widetilde{A}B)(\mathrm{mod}\, p^{n+m}) \\ &\equiv C(\mathrm{mod}\, p^{n+m}). \end{aligned}$$

The last equivalence is true provided we choose $\widetilde{A}, \widetilde{B}$ such that

$$\widetilde{C} + A\widetilde{B} + \widetilde{A}B \equiv 0(\mathrm{mod}\, p^m).$$

But since A, B are relatively prime modulo p^m, there are polynomials A', B', D such that $AB' + A'B \equiv 1(\mathrm{mod}\, p^m)$. Thus $\widetilde{C} - AB'\widetilde{C} - A'B\widetilde{C} \equiv 0(\mathrm{mod}\, p^m)$. We therefore choose $\widetilde{B} = -B'\widetilde{C}$ and $\widetilde{A} = -A'\widetilde{C}$. **Q.E.D.**

This lemma says that, given a factorization A, B of C modulo p^n, and under a suitable "non-degeneracy" condition on this factorization, we can "lift" the factorization up to another factorization (A^*, B^*) modulo p^{n+m}. the lemma can be applied again. Zassenhaus [20] generalized this to obtain a quadratically convergent factorization method.

References

[1] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[2] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[3] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[4] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[5] K. O. Geddes, S. R. Czapor, and G. Labahn. *Algorithms for Computer Algebra*. Kluwer Academic Publishers, Boston, 1992.

[6] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[7] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[8] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[9] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[10] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[11] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[12] M. Lauer. Generalized p-adic constructions. *SIAM J. Computing*, 12(2):395–410, 1983.

[13] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[14] D. R. Musser. *Algorithms for Polynomial Factorization*. PhD thesis, University of Wisconsin, 1971. Technical Report 134, Department of Computer Science.

[15] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[16] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[17] D. Y. Y. Yun. *The Hensel Lemma in Algebraic Manipulation*. PhD thesis, Massachusetts Institute of Technology, Cambridge MA, 1974. Project MAC Report TR-138.

[18] D. Y. Y. Yun. Algebraic algorithms using p-adic construction. In *Proc. ACM Symposium on Symbolic and Algebraic Computation*, pages 248–258. ACM, 1976.

[19] H. Zassenhaus. On Hensel factorization, I. *Journal of Number Theory*, 1:291–311, 1969.

Contents

Lecture V
Fundamental Theorem of Algebra

This lecture has primarily mathematical rather than computational goals. Our main objective is the Fundamental Theorem of Algebra. We choose a slightly circuitous route, via an investigation of the underlying real field $\mathbb{R}$. It is said that the Fundamental Theorem of Algebra depends on two distinct sets of properties: the algebraic properties of a field and the analytic properties of real numbers. But we will see that these "analytic properties" could be formulated purely in algebraic terms. Of course, there is no avoiding some standard construction (such as Dedekind cuts or Cauchy sequences) of the real numbers $\mathbb{R}$, and verifying that they satisfy our algebraic axioms. But even such constructions can made in a purely algebraic setting. This development originated from Artin's solution to Hilbert's 17th Problem[1]. The solution is based on the theory of real closed fields (see [111, 209]). Much of this theory has been incorporated into the algebraic theory of quadratic forms [176, 164], as well as into real semi-algebraic topology [3, 20, 23].

§1. Elements of Field Theory

We briefly review the some basic algebraic properties of field. For a proper treatment, there are many excellent textbooks (including van der Waerden's classic [209]).

Fields. A field is a commutative ring in which each non-zero element is invertible. This implies that a field is a domain. Often, F arises as the quotient field of a domain D. This underlying domain gives F its "arithmetical structure" which is important for other considerations. For instance, in Lecture III.1, we showed that the concepts of divisibility and unique factorization in a domain extend naturally to its quotient field. If there is a positive integer p such that $\underbrace{1+1+\cdots+1}_{p} = 0$. If p is chosen as small as possible, we say the field has *characteristic* p; if no such p exists, it has characteristic 0. One verifies that p must be prime.

Extension Fields. If $F \subseteq G$ where G is a field and F is a field under the induced operations of G, then F is a *subfield* of G, and G an *extension field* of F. An element $\theta \in G$ is *algebraic* over F if $p(\theta) = 0$ for some $p(X) \in F[X]$; otherwise θ is *transcendental* G is an *algebraic extension* of F if every element of G is algebraic over F. If $S \subseteq G$ then $F(S)$, the *adjunction of F by S*, denotes the smallest subfield of G that contains $F \cup S$. In case $S = \{\theta_1, \ldots, \theta_k\}$ is a finite set, we write $F(\theta_1, \ldots, \theta_k)$ for $F(S)$ and call this a *finite extension*. If $k = 1$, $F(\theta_1)$ is called a *simple extension*. It is easy to see G can be viewed as a vector space over F. Let $[G : F]$ denote the dimension of this vector space. We call $[G : F]$ the *degree* of G over F.

Simple extensions. To study the simple extension $F(\theta)$, consider the natural map $\phi : F[X] \to G$ that takes X to θ and which fixes F (this just means $\phi(x) = x$ for $x \in F$). It is clear that ϕ is a homomorphism. If I is the kernel of ϕ then the image of ϕ is isomorphic to $F[X]/I$. Furthermore, we have $I = (p)$ for some $p \in F[X]$, since I is an ideal and $F[X]$ is a principal ideal domain. Note that p must be irreducible. [Otherwise, $p = p_1 p_2$ for some non-trivial factor p_1. Then $0 = \phi(p) = \phi(p_1)\phi(p_2)$ implies $\phi(p_1) = 0$ or $\phi(p_2) = 0$ (since G is a domain). This proves p_1 or p_2 is in the kernel I,

[1] Let K be the field of rational numbers. Hilbert asks if a rational function $f \in K(X_1, \ldots, X_n)$ that is non-negatives at every point $(a_1, \ldots, a_n) \in K^n$ for which $f(a_1, \ldots, a_n)$ is defined, is necessarily a sum of squares of rational functions. Artin answered affirmatively in the more general case of any real closed field K.

contradiction.] There are now two possibilities: either $p = 0$ or $p \neq 0$. In the former case, the image is isomorphic to $F[X]/(0) = F[X]$ and θ is a transcendental element; in the latter case, we find that $\phi(p) = p(\theta) = 0$ so that θ is algebraic. In case $p \neq 0$, it is also easy to see that $F[\theta] = F(\theta)$: every non-zero element of $F[\theta]$ has the form $q(\theta)$ for some polynomial $q(X) \in F[X]$. We show that $q(\theta)$ has a multiplicative inverse. By the extended Euclidean algorithm, there exists $a(X), b(X) \in F[X]$ such that $q(X)a(X) + p(X)b(X) = 1$. Then $p(\theta) = 0$ implies $q(\theta)a(\theta) = 1$, *i.e.*, $a(\theta)$ is the inverse of $q(\theta)$.

Splitting fields. Above we started out with a given extension field G of F and ask how we find simple extensions of F into G. There is a converse problem: given a field F, we want to construct an extension with prescribed properties. In case we want a simple transcendental extension, this is easy: such a G is isomorphic to $F(X)$. If G is to be an algebraic extension, assume we are given a polynomial $p(X) \in F[X]$ and G is to be the smallest extension such that $p(X)$ splits into linear factors in $G[X]$. Then G is called the *splitting field* of $p(X)$, and is unique up to isomorphism. We now show such a splitting field may be constructed, proceeding in stages. First let us split off all linear factors $X-\alpha$ ($\alpha \in F$) of $p(X)$. If a non-linear polynomial $p_1(X)$ remains after removing the linear factors, let $q_1(X)$ be any irreducible non-linear factor of $p_1(X)$. Then the quotient ring $F[X]/(q_1)$ is a domain. But it is in fact a field because (q_1) is a maximal ideal. [For, if $q \notin (q_1)$ then the irreducibility of q_1 implies $\texttt{GCD}(q, q_1) = 1$, and by the extended Euclidean algorithm $F[X] = (1) = (q, q_1)$.] This extension field can be written as $F(\theta_1)$ where θ_1 is the equivalent class of X in $F[X]/(q_1)$. Now in $F(\theta_1)$, the polynomial $p_1/(X - \theta_1)$ may split off additional linear factors. If a non-linear polynomial p_2 remains after removing these linear factors, we again pick any irreducible factor q_2 of p_2, and extend $F(\theta_1)$ to $F(\theta_1)[X]/(q_2)$, which we write as $F(\theta_1, \theta_2)$, etc. This process must eventually stop. The splitting field G has the form $F(\theta_1, \ldots, \theta_k)$ and can be shown to be unique up to isomorphism. We have shown: *for any polynomial $p(X) \in F[X]$ there exists an extension field G of F in which $p(X)$ has* $\deg(p)$ *roots.*

Normal extensions. A field G is said to be a *normal extension* of F (or, *normal over* F) if G is an algebraic extension and for every irreducible polynomial $p(X) \in F[X]$, either G has no roots of $p(X)$ or G contains the splitting field of $p(X)$. We can equivalently characterize normal extensions as follows: two elements of G are *conjugates* of each other *over* F if they have the same minimal polynomial in $F[X]$. Then G is a normal extension of F iff G is closed under conjugates over F, *i.e.*, if $a \in G$ then G contains all the conjugates of a over F. If G is also a finite extension of F, it can be shown that G must be a splitting field of some polynomial over F. For instance, a quadratic extension $F(\sqrt{a})$ is normal over F. On the other hand, $\mathbb{Q}(a^{1/3}) \subseteq \mathbb{R}$ is not normal over $\mathbb{Q}$ for any positive integer a that is square-free. To see this, note that by Eisenstein's criterion (§III.1, Exercise), $X^3 - a$ is irreducible over $\mathbb{Z}$ and hence over $\mathbb{Q}$. But

$$X^3 - a = (X - a^{1/3})(X - \rho a^{1/3})(X - \rho^2 a^{1/3})$$

where $\rho, \rho^2 = (-1 \pm \sqrt{-3})/2$ are the two primitive cube-roots of unity. If $\mathbb{Q}(a^{1/3})$ were normal over $\mathbb{Q}$ then $\mathbb{Q}(a^{1/3})$ would contain a non-real element $\rho a^{1/3}$, which is impossible. It is not hard to show that splitting fields of F are normal extensions. A normal extension of a normal extension of F need not be a normal extension of F (Exercise).

Separable extensions. An irreducible polynomial $f(X) \in F[X]$ may well have multiple roots α in its splitting field. Such an α is said to be *inseparable* over F. But if α is a multiple root of $f(X)$, then it is a common root of $f(X)$ and $df(X)/dX = f'(X)$. Since $f(X)$ is irreducible, this implies $f'(X)$ is identically zero. Clearly this is impossible if F has characteristic zero (in general, such fields are called *perfect*). In characteristic $p > 0$, it is easy to verify that $f'(X) \equiv 0$ implies $f(X) = \phi(X^{p^e})$ for some $e \geq 1$ and $\phi(Y)$ is irreducible in $F[Y]$. If α is a simple root of an irreducible polynomial,

then it is *separable*. An extension G of F is *separable* over F if if all its elements are separable over F. An extension is *Galois* if it is normal and separable.

Galois theory. If E is an extension field of F, let $\Gamma(E/F)$ denote the group of automorphisms of E that fixes F. We call $g \in \Gamma(E/F)$ an *automorphism* of E *over* F. We *claim*: g must map each $\theta \in E$ to a conjugate element $g(\theta)$ over F. [In proof, note that if $p(X) \in F[X]$ then

$$g(p(\theta)) = p(g(\theta)).$$

The claim follows if we let $p(X)$ be the minimal polynomial of θ, whereby $p(\theta) = 0$ and $g(p(\theta)) = g(0) = 0$, so that $p(g(\theta)) = 0$.] In consequence, the group $\Gamma(E/F)$ is finite when E is a splitting field of some polynomial $p(X)$ over F. To see this, note that our claim implies that each $g \in \Gamma(E/F)$ determines a permutation π of $\alpha_1, \ldots, \alpha_n$. Conversely, each permutation π can extend to at most one $g \in \Gamma(E/F)$ since g is completely determined by its action on the roots of $p(X)$ because E is generated by the roots of $p(X)$ over F.

If G' is any subgroup of $\Gamma(E/F)$, then the *fixed field of* G' is the set of elements $x \in E$ such that $g(x) = x$ for all $g \in G$. Galois theory relates subgroups of $\Gamma(E/F)$ to the subfields of E over F. Two subfields K, K' of E over F are *conjugate* if there is an automorphism σ of E over F such that $\sigma(K) = K'$.

The Fundamental theorem of Galois theory says this. Suppose $p(X) \in F[X]$ is separable over F and E is the splitting field of $p(X)$.

(i) There is a one-one correspondence between subfields of E over F and the subgroups of $\Gamma(E/F)$: a subfield K corresponds to a subgroup H iff the fixed field of H is equal to K.

(ii) If K' is another subfield that corresponds to $H' \subseteq \Gamma(E/F)$ then $K \subseteq K'$ iff $H' \subseteq H$.

(iii) If K and K' are conjugate subfields then H and H' are conjugate subgroups.

Primitive element. Suppose $G = F(\theta_1, \ldots, \theta_k)$ is a finite separable extension of an infinite field F. Then it can be shown that $G = F(\theta)$ for some θ. Such an element θ is called a *primitive element*field!primitive element of G over F. The existence of such elements is easy to show provided we accept the fact[2] that are only finitely many fields that are intermediate between F and G: it is enough to show this when $k = 2$. Consider $F(\theta_1 + c\theta_2)$ for all $c \in F$. Since there are only finitely many such fields (being intermediate between F and G), suppose $F(\theta_1 + c\theta_2) = F(\theta_1 + c'\theta_2)$ for some $c \neq c'$. Letting $\theta = \theta_1 + c\theta_2$, it is clear that $F(\theta) \subseteq F(\theta_1, \theta_2)$. To see the converse inclusion, note that $(c - c')\theta_2 = \theta - (\theta_1 + c'\theta_2)$. Hence $\theta_2 \in F(\theta)$ and also $\theta_1 \in F(\theta)$.

Zorn's Lemma. A powerful principle in mathematical arguments is the Axiom of Choice. This usually appears in algebraic settings as Zorn's lemma (following Kneser): *if P is a partially ordered set such that every chain C in P has an upper bound in P, then P contains a maximal element.* A set $C \subseteq P$ is a chain if for every $x, y \in C$, either $x < y$ or $x > y$ or $x = y$. A typical application is this: let P be a collection of fields, partially ordered by set inclusion. If C is a chain in P, we note that its *union* $\cup C$ is also a field defined in the natural way: if $x, y \in \cup C$ then there is a field $F \in C$ that contains x, y and we define $x + y$ and xy as if they are elements in F. Assume that P is closed under unions of chains. Then Zorn's lemma implies that P contains a maximal field.

Algebraic Closure. If every non-linear polynomial in $F[X]$ is reducible then we say that F is *algebraically closed*. The *algebraic closure* of F, denoted $\overline{F}$, is a smallest algebraically closed field

[2] This is a result of Artin (see Jacobson [90]).

containing F. A theorem of Steinitz says that every field F has an algebraic closure, and this closure is unique up to isomorphism. The proof uses Zorn's lemma. But the existence of algebraic closures is intuitively clear: we simply iterate the splitting field construction for each polynomial, using transfinite induction. The Fundamental Theorem of Algebra is the assertion that $\mathbb{C}$ is algebraically closed.

Exercises

Exercise 1.1:
(i) A quadratic extension is a normal extension.
(ii) Let a be a positive square-free integer. If α is any root of $X^3 - a$ then $\mathbb{Q}(\alpha)$ is not normal. (This is more general than stated in the text.)
(iii) $\mathbb{Q}(\sqrt[4]{2})$ is not a normal extension of $\mathbb{Q}$. Thus, a normal extension of a normal extension need not be a normal extension. HINT: $X^4 - 2 = (X^2 - \sqrt{2})(X^2 + \sqrt{2})$. □

Exercise 1.2: The splitting field E of $f(X) \in F[X]$ over F has index $[E : F] \leq n!$ where $n = \deg(f)$.
HINT: use induction on n. □

Exercise 1.3:
(i) Compute a basis of E over $F = \mathbb{Q}$ in the following cases of E: $E = \mathbb{Q}(\sqrt{2}, \sqrt{3})$, $\mathbb{Q}(\sqrt{2}, \sqrt{-2})$, $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2})$, $\mathbb{Q}(\sqrt{2}, \omega)$ where $\omega = (1 + \sqrt{-3})/2$, $\mathbb{Q}(\sqrt[3]{2}, \omega)$, $\mathbb{Q}(\sqrt{2}, \sqrt[3]{2}, \sqrt[3]{5})$.
(ii) Compute the group $\Gamma(E/F)$, represented as a subgroup of the permutations on the previously computed basis. Which of these extensions are normal? □

§2. Ordered Rings

To study the real field $\mathbb{R}$ algebraically, we axiomatize one of its distinguishing properties, namely, that it can[3] be ordered.

Let R be a commutative ring (as always, with unity). A subset $P \subseteq R$ is called a *positive set* if it satisfies these properties:
(I) For all $x \in R$, either $x = 0$ or $x \in P$ or $-x \in P$, and these are mutually exclusive cases.
(II) If $x, y \in P$ then $x + y$ and $xy \in P$.
We say R is *ordered* (by P) if R contains a positive set P, and call (R, P) an *ordered ring*.

As examples, $\mathbb{Z}$ is naturally ordered by the set of positive integers. If R is an ordered ring, we can extend this ordering to the polynomial ring $R[X]$, by defining the positive set P to comprise all polynomials whose leading coefficients are positive (in R).

Let $P \subseteq R$ be a fixed positive set. We call a non-zero element x *positive* or *negative* depending on whether it belongs to P or not. For $x, y \in R$, we say x is *less than* y, written "$x < y$", if $y - x$ is positive. Similarly, x is *greater than* y if $x - y \in P$, written "$x > y$". In particular, positive and negative elements are denoted $x > 0$ and $x < 0$, respectively. We extend in the usual way the terminology to *non-negative*, *non-positive*, *greater or equal to* and *less than or equal to*, written $x \geq 0$, $x \leq 0$, $x \geq y$ and $x \leq y$. Define the *absolute value* $|x|$ of x to be x if $x \geq 0$ and $-x$ if $x < 0$.

[3]It is conventional to define "ordered fields". But the usual concept applies to rings directly. Moreover, we are interested in the order in rings such as $\mathbb{Z}$ and $\mathbb{Q}[X]$.

We now show that notations are consistent with some familiar properties of these inequality symbols.

Lemma 1 *Let x, y, z be elements of an ordered ring R.*
(i) $x > 0$ and $xy > 0$ implies $y > 0$.
(ii) $x \neq 0$ implies $x^2 > 0$. In particular, $1 > 0$.
(iii) $x > y$ implies $x + z > y + z$.
(iv) $x > y$ and $z > 0$ implies $xz > yz$.
(v) $x > 0$ implies $x^{-1} > 0$.
(vi) $x > y > 0$ implies $y^{-1} > x^{-1}$ (provided these are defined).
(vii) (transitivity) $x > y$ and $y > z$ implies $x > z$.
(viii) $x \neq 0, y \neq 0$ implies $xy \neq 0$.
(ix) $|xy| = |x| \cdot |y|$.
(x) $|x + y| \leq |x| + |y|$.
(xi) $x^2 > y^2$ implies $|x| > |y|$.

The proof is left as an exercise.

From property (II) in the definition of an ordered ring R, we see that R has characteristic 0 (otherwise if $p > 0$ is the characteristic of R then $\underbrace{1 + 1 + \cdots + 1}_{p} = 0$ is positive, contradiction). Parts (ii) and (iii) of the lemma implies that $0 < 1 < 2 < \cdots$. Part (vii) of this lemma says that R is totally ordered by the '$>$' relation. Part (viii) implies R is a domain.

An *ordered domain* (or *field*) is an ordered ring that happens to be a domain (or field). If D is an ordered domain, then its quotient field Q_D is also ordered: define an element $u/v \in Q_D$ to be *positive* if uv is positive in D. It is easy to verify that this defines an ordering on Q_D that extends the ordering on D.

Exercises

Exercise 2.1: Verify lemma 1. □

Exercise 2.2: In an ordered field F, the polynomial $X^n - c$ has at most one positive root, denoted $\sqrt[n]{c}$. If n is odd, it cannot have more than one root; if n is even, it has at most two roots (one is the negative of the other). □

Exercise 2.3: If the ordering of Q_D preserves the ordering of D, then this ordering of Q_D is unique. □

§3. Formally Real Rings

Sum of Squares. In the study of ordered rings, those elements that can be written as sums of squares have a special role. At least for the real fields, these are necessarily positive elements. Are they necessarily positive in an ordered ring R? To investigate this question, let us define

$$R^{(2)}$$

to denote the set of elements of the form $\sum_{i=1}^{m} x_i^2$, $m \geq 1$, where the x_i's are non-zero elements of R. The x_i's here are not necessarily distinct. But since the x_i's are non-zero, it is not automatic that 0 belongs to $R^{(2)}$. Indeed, whether 0 belongs to $R^{(2)}$ is critical in our investigations.

Lemma 2
(i) $1 \in R^{(2)}$ *and* $R^{(2)}$ *is closed under addition and multiplication.*
(ii) If $x, y \in R^{(2)}$ *and* y *is invertible then* $x/y \in R^{(2)}$.
(iii) If $P \subseteq R$ *is a positive set, then* $R^{(2)} \subseteq P$.

Proof. (i) is easy. To see (ii), note that if y^{-1} exists then $x/y = (xy)(y^{-1})^2$ is a product of elements in $R^{(2)}$, so $x/y \in R^{(2)}$. Finally, (iii) follows from (i) because squares are positive. **Q.E.D.**

This lemma shows that $R^{(2)}$ has some attributes of a positive set. Under what conditions can $R^{(2)}$ be extended into a positive set? From (iii), we see that $0 \notin R^{(2)}$ is a necessary condition. This further implies that R has characteristic $p = 0$ (otherwise $\underbrace{1 + 1 + \cdots + 1}_{p} = 0 \in R^{(2)}$).

A ring R is *formally real* if $0 \notin R^{(2)}$. This notion of "real" is only formal because R need not be a subset of the real numbers $\mathbb{R}$ (Exercise). The following is immediately from lemma 2(iii):

Corollary 3 *If* R *is ordered then* R *is formally real.*

To what extent is the converse true? If R is formally real, then $0 \notin R^{(2)}$, and $x \in R^{(2)}$ implies $-x \notin R^{(2)}$. So, $R^{(2)}$ has some of the attributes of a positive set. In the next section, we show that if R is a formally real domain then $R^{(2)}$ can be extended to a positive set of some extension of R.

Exercises

Exercise 3.1: (a) If the characteristic of R is not equal to 2 and R is a field then $0 \in R^{(2)}$ implies $R = R^{(2)}$.
(b) If $R^{(2)}$ does not contain 0 then R has no nilpotent elements.
(c) Every element in $GF(q)$ is a sum of two squares. □

Exercise 3.2:
(a) Let $\alpha \in \mathbb{C}$ be any root of $X^3 - 2$. Then $\mathbb{Q}(\alpha)$ is formally real (but not necessarily real).
(b) Let $\mathbb{Q}(\alpha)$ be an algebraic number field and $f(X)$ is the minimal polynomial of α. Then $\mathbb{Q}(\alpha)$ is formally real iff $f(X)$ has a root in $\mathbb{R}$. □

Exercise 3.3: Let K be a field.
(a) Let $G_2(K)$ denote the set $\{x \in K \setminus \{0\} : x = a^2 + b^2, a, b \in K\}$. Show that $G_2(K)$ is a group under multiplication. HINT: consider the identity $|zz'| = |z| \cdot |z'|$ where z, z' are complex numbers.
(b) Let $G_4(K)$ denote the set $\{x \in K \setminus \{0\} : x = a^2 + b^2 + c^2 + d^2, a, b, c, d \in K\}$. Show that $G_4(K)$ is a group under multiplication. HINT: consider the identity $|qq'| = |q| \cdot |q'|$ where q, q' are quarternions. □

§4. Constructible Extensions

Let F be a formally real field. For instance, if D be a domain, then its quotient field $F = Q_D$ is formally real iff D is formally real. It is immediate that

$$F \text{ is formally real iff } \quad -1 \notin F^{(2)}.$$

We call a field extension of the form

$$G = F(\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_n}), \qquad a_i \in F,$$

a *finite constructible extension* of F provided G is formally real. If $n = 1$, we call G a *simple constructible extension*. Note that "$F(\sqrt{a})$" is just a convenient notation for the splitting field of the polynomial $X^2 - a$ over F (that is, we do not know if $\sqrt{a}$ can be uniquely specified as the "positive square root of a" at this point).

Ruler and compass constructions. Our "constructible" terminology comes from the classical problem of ruler-and-compass constructions. More precisely, a number is *(ruler-and-compass) constructible* if it is equal to the distance between two *constructed points*. By definition, constructible numbers are positive. Initially, we are given two points (regarded as constructed) that are unit distance apart. Subsequent points can be constructed as an intersection point of two constructed curves where a *constructed curve* is either a line through two constructed points or a circle centered at a constructed point with radius equal to a constructed number. [Thus, our ruler is only used as a "straight-edge" and our compass is used to transfer the distance between two constructed points as well as to draw circles.] The following exercise shows that "constructible numbers" in this sense coincides with our abstract notion of constructible real numbers over $\mathbb{Q}$.

Exercise 4.1: In this exercise, constructible means "ruler-and-compass constructible".

i) Show that if $S \subseteq \mathbb{R}$ is a set of constructible numbers, so are the positive elements in the smallest field $F \subseteq \mathbb{R}$ containing S. [In particular, the positive elements in $\mathbb{Q}$ are constructible.]

ii) Show that if the positive elements in a field $F \subseteq \mathbb{R}$ are constructible, so are the positive elements in $F(\sqrt{a})$, for any positive $a \in F$. [In view of i), it suffices to construct $\sqrt{a}$.]

iii) Show that if x is any number constructible from elements of $F \subseteq \mathbb{R}$ then x is in $F(\sqrt{a_1}, \ldots, \sqrt{a_k})$ for some positive numbers $a_i \in F$, $k \geq 0$. □

Lemma 4 *If F is a formally real field, $a \in F$ and $F(\sqrt{a})$ is not formally real then $a \notin F^{(2)}$ and $-a \in F^{(2)}$.*

Proof. $F(\sqrt{a})$ is not formally real is equivalent to $0 \in F(\sqrt{a})^{(2)}$. Hence

$$\begin{aligned} 0 &= \sum_i (b_i + c_i\sqrt{a})^2, \qquad (b_i, c_i \in F) \\ &= \sum_i (b_i^2 + c_i^2 a) + 2\sqrt{a}\sum_i b_i c_i \\ &= u + v\sqrt{a}, \end{aligned}$$

where the last equation defines u and v. If $u \neq 0$ then $v \neq 0$; hence $\sqrt{a} = -u/v \in F$ and $F(\sqrt{a}) = F$ is formally real, contradicting our assumption. Hence we may assume $u = 0$. If $a \in F^{(2)}$ then u

which is defined as $\sum_i (b_i^2 + c_i^2 a)$ also belongs to $F^{(2)}$. This give the contradiction $0 = u \in F^{(2)}$. This proves $a \notin F^{(2)}$, as required. We also see from the definition of u that

$$-a = (\sum_i b_i^2)/(\sum_i c_i^2) \in F^{(2)},$$

as required. **Q.E.D.**

Corollary 5
(i) (Real Square-root Extension) $a \in F^{(2)}$ implies $F(\sqrt{a})$ is formally real.
(ii) $a \in F$ implies $F(\sqrt{a})$ or $F(\sqrt{-a})$ is formally real.

Constructible closure. Let H be any extension of F. Call $x \in H$ a *constructible element* of H over F provided $x \in G \subseteq H$ where G is any finite constructible extension of F. Call H a *constructible extension* of F if every element in H is constructible. A field F is *constructible closed* if for any $a \in F$, $F(\sqrt{a}) = F$. We call F *(formally) real constructible closed* if F is formally real and for any $a \in F^{(2)}$, $F(\sqrt{a}) = F$. Beware that if F is constructible closed then it cannot be formally real because $\sqrt{-1} \in F$. We define a *(formally) real constructible closure* $\widehat{F}$ of a formally real field F to be a real constructible closed extension of F that is minimal, *i.e.*, for any field G, if $F \subseteq G \subset \widehat{F}$ then G is not real constructible closed.

Let U be a set of formally real extensions of F, closed under two operations, (a) real square-root extension (á la corollary 5(i)), and (b) forming unions of chains (§1). Zorn's lemma implies that U has a maximal element F' that is an extension of F. Clearly F' is real constructible closed. To obtain a real constructible closure of F, the set V of all real constructible closed fields between F and F'. If C is a chain in V, the intersection $\cap C$ can be made into a field that contains F in a natural way. We see that V is closed under intersection of chains. By another application of Zorn's lemma to V, we see that there is minimal element $\widehat{F}$ in V. This shows:

Theorem 6 *Every formally real field F has a real constructible closure $G = \widehat{F}$.*

For instance, suppose $F = \mathbb{Q}(X)$. A real constructive closure $G = \widehat{F}$ contains either $\sqrt{X}$ or $\sqrt{-X}$, but not both. Let x_1 be the element in G such that $x_1^2 = X$ or $x_1^2 = -X$. The choice of x_1 will determine the sign of X. In general, G contains elements x_n ($n \geq 1$) such that $x_n^2 = x_{n-1}$ or $x_n^2 = -x_{n-1}$ (by definition, $x_0 = X$). Thus G is far from being uniquely determined by F. On the other hand, the next result shows that each choice of G induces a unique ordering of F.

Lemma 7 *If G is real constructible closed then it has a unique positive set, and this set is $G^{(2)}$.*

Proof. We know that $G^{(2)}$ is contained in any positive set of G. Hence it suffices to show that $G^{(2)}$ is a positive set. We already know $G^{(2)}$ is closed under addition and multiplication, and $0 \notin G^{(2)}$. We must show that for every non-zero element x, either x or $-x$ belongs to $G^{(2)}$. But this is a consequence of corollary 5 since either $\sqrt{x}$ or $\sqrt{-x}$ is in G. **Q.E.D.**

We now investigate the consequence of adding $\mathbf{i} = \sqrt{-1}$ to a real constructible closed field. This is analogous to the extension from $\mathbb{R}$ to $\mathbb{C} = \mathbb{R}(\mathbf{i})$.

Theorem 8 *If G is real constructible closed then $G(\mathbf{i})$ is constructible closed.*

Proof. Let $a + b\mathbf{i}$ where $a, b \in G$, not both 0. We must show that there exist $c, d \in G$ such that $(c + d\mathbf{i})^2 = a + b\mathbf{i}$. Begin by defining a positive element $e := \sqrt{a^2 + b^2}$. Clearly e belongs to G. We have $e \geq |a|$ since $e^2 = a^2 + b^2 \geq |a|^2$. Hence both $(e - a)/2$ and $(e + a)/2$ are non-negative. So there exists $c, d \in G$ satisfying

$$c^2 = \frac{e + a}{2}, \qquad d^2 = \frac{e - a}{2}. \tag{1}$$

This determines c, d only up to sign, so we further require that $cd \geq 0$ iff $b \geq 0$. Hence we have

$$c^2 - d^2 = a, \qquad 2cd = b. \tag{2}$$

It follows that

$$(c + d\mathbf{i})^2 = (c^2 - d^2) + 2cd\mathbf{i} = a + b\mathbf{i},$$

as desired. **Q.E.D.**

Exercises

Exercise 4.2: (See [4]) For $a, b \in \mathbb{R}$, we say that the complex number $a + b\mathbf{i}$ is *constructible* if a and b are both constructible real numbers.

(i) The sum, difference, product and quotient of constructible complex numbers are constructible.

(ii) The square-root of a constructible complex number is constructible. □

§5. Real Closed Fields

A field is *real closed* if it is formally real and any algebraic proper extension is formally non-real.

Lemma 9 *Let F be formally real. Let $p(X) \in F[X]$ be irreducible of odd degree n and α be any root of $p(X)$ in the algebraic closure of F, Then $F(\alpha)$ is formally real.*

Proof. The result is true for $n = 1$. Suppose inductively that the result holds for all smaller odd values of n. If $F(\alpha)$ is not formally real then

$$-1 = \sum_{i \in I} q_i(\alpha)^2$$

for some finite index set I and $q_i(X) \in F[X]$, $\deg q_i \leq n - 1$. Since $p(X)$ is irreducible, $F(\alpha)$ is isomorphic to $F[X]/(p(X))$. Thus we get

$$-1 = \sum_{i \in I} q_i(X)^2 + r(X)p(X)$$

for some $r(X) \in F[X]$. But $\sum_{i \in I} q_i(X)^2$ has a positive even degree of at most $2n - 2$. Hence, in order for this equation to hold, the degree of $r(X)p(X)$ must equal that of $\sum_{i \in I} q_i(X)^2$. Then $r(X)$ has odd degree at most $n - 2$. If $r'(X)$ is any irreducible factor of $r(X)$ of odd degree and β is a root of $r'(X)$ then we get

$$-1 = \sum_{i \in I} q_i(\beta)^2.$$

This proves $F(\beta)$ is not formally real, contradicting our inductive assumption. **Q.E.D.**

For instance, $X^3 - 2$ is irreducible in $\mathbb{Q}[X]$ (Eisenstein's criterion). Thus $\mathbb{Q}(\alpha)$ is formally formally real for α any cube-root of 2. If we choose $\alpha \notin \mathbb{R}$, then $\mathbb{Q}(\alpha)$ is not a subset of $\mathbb{R}$.

Corollary 10 *If F is real closed, then every irreducible polynomial of $F[X]$ has even degree.*

Proof. Let $p(X) \in F[X]$ have odd degree. If α is any root of $p(X)$ in the algebraic closure of F then $F(\alpha)$ is formally real. Since F is real closed, this means $\alpha \in F$. As $X - \alpha$ divides $p(X)$, we conclude that $p(X)$ is reducible in $F[X]$. **Q.E.D.**

Theorem 11 (Characterization of real closed fields)
The following statements are equivalent.
(i) F is real closed.
(ii) F is real constructible closed and every polynomial in $F[X]$ of odd degree has a root in F.
(iii) F is not algebraically closed but $F(\mathbf{i})$ is.

Proof.

(i) implies (ii): this follows from the above corollary and the following observation: a real closed field F is real constructible closed. To see this, if $a \in F^{(2)}$ then $F(\sqrt{a})$ is formally real and hence $\sqrt{a} \in F$.

(ii) implies (iii): clearly F is formally real implies it is not algebraically closed since $X^2 + 1$ has no solution in F. To see that $F(\mathbf{i})$ is algebraically closed, it suffices to prove that any non-constant polynomial $f(X) \in F(\mathbf{i})[X]$ has a root in $F(\mathbf{i})$. Write $\overline{f(X)}$ for the *conjugate* of $f(X)$, obtained by conjugating each coefficient (the conjugate of a coefficient $x + y\mathbf{i} \in F(\mathbf{i})$ is $x - y\mathbf{i}$). It is not hard to verify that

$$g(X) = f(X)\overline{f}(X)$$

is an element of $F[X]$. Moreover, if $g(X)$ has a root $\alpha \in F(\mathbf{i})$, this implies $f(\alpha) = 0$ or $\overline{f}(\alpha) = 0$. But the latter is equivalent to $f(\overline{\alpha}) = 0$. So $g(X)$ has a root in $F(\mathbf{i})$ iff $f(X)$ has a root in $F(\mathbf{i})$.

We now focus on $g(X)$. Let $\deg(g) = n = 2^i q$ where q is odd and $i \geq 0$. We use induction on n to show that g has a root in $F(\mathbf{i})$. If $i = 0$, then by assumption $g(X)$ has a root in F. So assume $i \geq 1$. Let $\alpha_1, \ldots, \alpha_n$ be the roots of g in an algebraic extension of F. We may assume these roots are distinct since otherwise $\texttt{GCD}(g, dg/dX)$ has a root α in $F(\mathbf{i})$, by induction on n. Consider the set of values

$$B = \{\alpha_j\alpha_k + c(\alpha_j - \alpha_k) : 1 \leq j < k \leq n\}$$

where $1 \leq j < k \leq n$, for a suitable choice of $c \in F$. Let $N := \binom{n}{2}$. Clearly $|B| \leq N$ and there are $O(n^4)$ values of c for which this inequality is strict. This is because each coincidence of values uniquely determines a c. Since F is infinite, we may pick c so that $|B| = N$. Let s_i denote ith elementary symmetric function (§VI.5) of the elements in B (thus $s_1 = \sum_{x \in B} x$). But s_i is also symmetric in $\alpha_1, \ldots, \alpha_n$. Hence the s_i are rational integral polynomials in the elementary symmetric functions $\sigma_0, \ldots, \sigma_n$ on $\alpha_1, \ldots, \alpha_n$. But these σ_i's are precisely the coefficients of $g(X)$. Thus the polynomial $G(X) = \sum_{i=0}^{N} s_i X^i$ belongs to $F[X]$ and its roots are precisely the elements of B.

Notice the degree of G is $N = 2^{i-1}q'$ for some odd q'. By induction hypothesis, G has a root in $F(\mathbf{i})$. Without loss of generality, let this root be

$$\phi = \alpha_1\alpha_2 + c(\alpha_1 - \alpha_2) \in F(\mathbf{i}).$$

Similar to the construction of $G(X)$, let $u(X) \in F[X]$ be the polynomial whose roots are precisely the set $\{\alpha_j\alpha_k : 1 \le j < k \le n\}$, and likewise $v(X) \in F[X]$ be the polynomial whose roots are $\{\alpha_j - \alpha_k : 1 \le j < k \le n\}$. We note that

$$u(\alpha_1\alpha_2) = 0, \qquad v(\frac{\phi - \alpha_1\alpha_2}{c}) = v(\alpha_1 - \alpha_2) = 0.$$

Moreover, for any $\alpha_j\alpha_k$ where $(j,k) \neq (1,2)$, $v(\frac{\phi-\alpha_j\alpha_k}{c}) \neq 0$ because our choice of c implies $\frac{\phi-\alpha_j\alpha_k}{c} \neq \alpha_\ell - \alpha_m$ for any $1 \le \ell < m \le n$. This means that the polynomials

$$u(X), \quad v(\frac{\phi - X}{c}) \in F(\phi)[X] \subseteq F(\mathbf{i})[X]$$

have $\alpha_1\alpha_2$ as their only common root. Their GCD is thus $X - \alpha_1\alpha_2$, which must be an element of $F(\mathbf{i})[X]$. This proves that $\alpha_1\alpha_2 \in F(\mathbf{i})$ and therefore $\alpha_1 - \alpha_2 \in F(\mathbf{i})$. We can determine α_1 and α_2 by solving a quadratic equation in $F(\mathbf{i})$. This proves $g(X)$ has solutions α_1, α_2 in $F(\mathbf{i})$.

(iii) implies (i): we must show that F is formally real. We first observe that an irreducible polynomial $f(X)$ in $F[X]$ must have degree 1 or 2 because of the inequality

$$2 = [F(\mathbf{i}) : F] \ge [E : F] = \deg f$$

where $E \subseteq F(\mathbf{i})$ is the splitting field of f over F. Next we see that it is sufficient to show that every sum $a^2 + b^2$ of squares ($a, b \in F \setminus \{0\}$) is a square in F. For, by induction, this would prove that every element in $F^{(2)}$ is a square, and the formal reality of F then follows because -1 is not a square in F. To show $a^2 + b^2$ is a square, consider the polynomial $f(X) = (X^2 - a)^2 + b^2 \in F[X]$. It factors as

$$(X^2 - a - b\mathbf{i})(X^2 - a + b\mathbf{i})$$

over $F(\mathbf{i})$. Since $F(\mathbf{i})$ is algebraically closed, there are $c, d \in F(\mathbf{i})$ such that

$$c^2 = a + b\mathbf{i}, \qquad d^2 = a - b\mathbf{i}.$$

This gives $f(X) = (X - c)(X + c)(X - d)(X + d)$. Note that $\pm a \pm b\mathbf{i}$ are not elements in F. Thus $f(X)$ has no linear factors in $F[X]$. It must therefore split into quadratic factors. Consider the factor that contains $X - c$. This cannot be $(X - c)(X + c) = X^2 - c^2$. Hence it must be $(X - c)(X \pm d)$. In either case, notice that $\pm cd = \sqrt{a^2 + b^2}$ is the constant term of $(X - c)(X \pm d)$. Hence $\sqrt{a^2 + b^2} \in F$, as we wanted to show. **Q.E.D.**

Exercises

Exercise 5.1: Show that $f(X)\overline{f}(X) \in F[X]$ if $f(X) \in F(\mathbf{i})[X]$. □

Exercise 5.2: Let $a, b \in F$ and $f(X) \in F[X]$ where F is real closed. If $f(a) < 0$ and $f(b) > 0$ then there exists c between a and b such that $f(c) = 0$. HINT: it suffices to show this for $\deg f = 2$. □

Exercise 5.3: (See [164])
(a) The *Pythagorean number* of R is the least number $h = h(R)$ such that if $x \in R^{(2)}$, then x is a sum of at most h squares. Thus $h(R) = 1$ for $R = \mathbb{R}$. Show Lagrange's theorem that $h(\mathbb{Z}) = 4$.
(b) If K is real closed field then $h(K) = 2$.
(c) We call a field K *Pythagorean* if $h(K) = 2$; alternatively, any sum of 2 squares is a square in such a field. Show that the field of constructible real numbers is Pythagorean.
(d) Let $P \subseteq \mathbb{R}$ be the smallest Pythagorean field containing $\mathbb{Q}$. Then P is properly contained in the field of constructible real numbers. HINT: $\sqrt{\sqrt{2}-1} \notin P$ is constructible. □

§6. Fundamental Theorem of Algebra

In this section, we are again interested in the standard "reals" $\mathbb{R}$, not just "formal reals". In 1746, d'Alembert (1717–1783) published the first formulation and proof of the Fundamental Theorem of Algebra. Gauss (1777–1855) is credited with the first proof[4] that is acceptable by modern standards in 1799. We note two analytic properties of real numbers:

1. The reals are ordered (and hence formally real).
2. (Weierstraß's Nullstellensatz) If $f(X)$ is a real function continuous in an interval $[a, b]$, and $f(a)f(b) < 0$ then $f(c) = 0$ for some c between a and b.

Recall that $f(X)$ is continuous at a point $X = a$ if for all $\epsilon > 0$ there is a $\delta > 0$ such that $|f(a + d) - f(a)| < \epsilon$ whenever $|d| < \delta$. It is not hard to show that the constant functions, the identity function, the sums and products of continuous functions are all continuous. In particular, polynomials are continuous. One then concludes from Weierstraß's Nullstellensatz:
(i) A positive real number c has a positive real square root, $\sqrt{c}$.
(ii) A real polynomial $f(X)$ of odd degree has a real root.
It follows from theorem 11 that $\mathbb{R}$ is real closed. Since $\mathbb{C}$ is defined to be $\mathbb{R}(\sqrt{-1})$, we obtain:

Theorem 12 (Fundamental Theorem of Algebra)
$\mathbb{C}$ *is algebraically closed.*

Exercise 6.1: Use Weierstraß's Nullstellensatz to verify the above assertions, in particular, properties (i) and (ii). □

Cantor's construction of the reals. Since the Fundamental Theorem of Algebra is about a very specific structure, $\mathbb{R}$, it is worthwhile recalling one construction (albeit, a highly non-constructive one!) of this set. If F is a field and $\widehat{F}$ is a minimal real closed field containing F, then we call $\widehat{F}$ a *real closure* of F. As in our proof of theorem 6, the real closure of any real field F exists, by Zorn's lemma. For instance, $\widehat{\mathbb{Q}}$ is the set of real algebraic numbers and hence forms a countable set. But $\mathbb{R}$, being uncountable, must necessarily include many elements not in $\widehat{Q}$.

[4] One of Gauss' proofs was in turn found wanting, presumably also by modern standards only! Apropos of a footnote in (§VI.1) concerning the second-class status of complex numbers, Gauss personally marked the transition to the modern view of complex numbers: in his 1799 dissertation on the Fundamental Theorem of Algebra, he deliberately avoided imaginary numbers (by factoring polynomials only up to linear or quadratic factors). In 1849, he returned to give his fourth and last proof of the theorem, this time using imaginaries. The symbol **i** for $\sqrt{-1}$ is due to Euler (1707–1783). See [197, p. 116,122].

We now outline the method to obtain a real closed field containing an ordered field F. This just mirrors the construction of $\mathbb{R}$ from $\mathbb{Q}$ using Cauchy sequences. A *Cauchy or fundamental sequence* of F is a infinite sequence $a = (a_1, a_2, \ldots)$ such that for all positive ϵ there exists $n = n(\epsilon)$ such that $|a_i - a_j| < \epsilon$ for all $i, j > n$. We may add and multiply such sequences in a componentwise manner. In particular, if $b = (b_1, b_2, \ldots)$ then $ab = (a_1 b_1, a_2 b_2, \ldots)$. It is easy to check that Cauchy sequences form a commutative ring. We define a sequence $(a_1, a_2, \ldots)$ to be *null* if for all $\epsilon > 0$ there exists $n = n(\epsilon)$ such that $|a_i| < \epsilon$ for $i > n$. Similarly, we define the sequence to be *positive* if there is an $\epsilon > 0$ and n such that $a_i > \epsilon$ for $i > n$. The set of null sequences form a maximal ideal in the ring of fundamental sequences; hence the ring of Cauchy sequences modulo this null ideal is a field $\widetilde{F}$ that extends F in a canonical way. If $F = \mathbb{Q}$ then $\widetilde{F}$ is, by definition, equal to $\mathbb{R}$. Since $\mathbb{R}$ is uncountable, it is not equal to the $\widehat{Q}$. This shows that $\widetilde{F}$ is, in general, not equal to the real closure $\widehat{F}$ of F. Now $\widetilde{F}$ is an ordered field because the set of positive Cauchy sequences correspond to the positive elements of $\widetilde{F}$. This construction, if repeated on $\widetilde{F}$ yields nothing new: $\widetilde{\widetilde{F}}$ is isomorphic to $\widetilde{F}$. An ordering of R is *Archimedean* if for all $a \in R$, there is an $n \in \mathbb{Z}$ such that $a < n$. If F is Archimedean ordered, we have a canonical isomorphism between $\widetilde{F}$ and $\mathbb{R}$.

Exercises

Exercise 6.2: Show that $\widehat{\mathbb{Q}}$ is the set of real algebraic numbers, and hence a countable set. □

Exercise 6.3: Verify the assertions made of the Cantor construction. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture VI
Roots of Polynomials

From a historical viewpoint, it seems appropriate to call root finding for polynomials the *Fundamental Computational Problem of algebra*. It occupied mathematicians continuously from the earliest days, and was until the 19th century one of their major preoccupations. Descartes, Newton, Euler, Lagrange and Gauss all wrote on this subject. This interest has always been intensely computational in nature, which strikes a chord with modern computer science. The various extensions of natural numbers (negative[1], rational, algebraic and complex numbers) were but attempts to furnish entities as roots. Group theory, for instance, originated in this study of roots.

There are two distinct lines of investigation. The first is the algebraic approach which began with with the Italian algebraists (see §0.2) after the introduction of algebra through the Arab mathematicians in the 13th century. The algebraic approach reached a new level of sophistication with the impossibility demonstrations of Abel and Wantzel. The numerical approximation of roots represent the other approach to the Fundamental Problem. Here, Viète (1600) published the first solution. These were improved by others, culminating in the well known method of Newton (1669). Horner's contribution was to organize Newton's method for polynomials in a very efficiently hand-calculable form. One ought not minimize such a contribution: contemporary research in algorithms follows the same spirit. Horner's method resembles a method that was perfected by Chin Kiu-Shao about 1250 [194, p.232]. Simpson, the Bernoullis, Lagrange and others continued this line of research. Goldstine's history of numerical analysis [73] treats numerical root finding; Nový [146] focuses on the algebraic side, 1770–1870.

Modern treatments of the fundamental problem may be found in Henrici [79], Obreschkoff [147], Ostrowski [151] and Marden [127]. Our treatment here is slanted towards finding real roots. In principle, finding complex roots can be reduced to the real case. We are interested in "infallible methods". Collins [46], an early advocate of this approach, noted that we prefer to make infallible algorithms faster, whereas others have sought to make fast algorithms less fallible (cf. [93]). Along this tradition, recent work of Schönhage [184], Pan [152] and Renegar [166] show how to approximate all complex roots of a polynomial to any prescribed accuracy $\epsilon > 0$, in time $O(n^2 \log n(n \log n + \log \frac{1}{\epsilon}))$. Neff [144] shows that this problem is "parallelizable" (in NC, cf. §0.8). However, this does not imply that the problem of root isolation is in NC. There is a growing body of literature related to Smale's approach [192, 193]. Pan [153] gives a recent history of the bit complexity of the problem. This (and the next) lecture is necessarily a selective tour of this vast topic.

§1. Elementary Properties of Polynomial Roots

There is a wealth of material on roots of polynomials (e.g. [127, 147, 135]). Here we review some basic properties under two categories: complex roots and real roots. But these two categories could also be taken as any algebraically closed field and any real closed field, respectively.

Complex Polynomials. Consider a complex polynomial,

$$A(X) = \sum_{i=0}^{n} a_i X^i, \quad a_n \neq 0, n \geq 1.$$

[1]The term "imaginary numbers" shows the well-known bias in favor of real numbers. Today, the term "negative numbers" has hardly any lingering negative (!) connotation but a bias was evident in the time of Descartes: his terms for positive and negative roots are (respectively) "true" and "false" roots [197, p. 90].

C1. Let $c \in \mathbb{C}$. By the Division Property for polynomials (§II.3), there exists a $B(X) \in \mathbb{C}[X]$,

$$A(X) = B(X) \cdot (X - c) + A(c).$$

Thus c is a root of $A(X)$ iff $A(c) = 0$, iff $A(X) = B(X) \cdot (X - c)$. Since $\deg B = \deg(A) - 1$, we conclude by induction that $A(X)$ has at most $\deg A$ roots. This is the easy half of the fundamental theorem of algebra.

C2. *Taylor's expansion* of $A(X)$ at $c \in \mathbb{C}$:

$$A(X) = A(c) + \frac{A'(c)}{1!}(X - c) + \frac{A''(c)}{2!}(X - c)^2 + \cdots + \frac{A^{(n)}(c)}{n!}(X - c)^n.$$

C3. *$A(X)$ is determined by its value at $n + 1$ distinct values of X.* This can be seen as a consequence of the Chinese Remainder Theorem (§IV.1).

C4. **(The Fundamental Theorem of Algebra)** *$A(X)$ has exactly n (not necessarily distinct) complex roots, $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$.* This was proved in the last lecture. By repeated application of Property C1, we can write

$$A(X) = a_n \prod_{i=1}^{n} (X - \alpha_i). \tag{1}$$

Definition: If α occurs $m \geq 0$ times among the roots $\alpha_1, \ldots, \alpha_n$, we say α is a *root of multiplicity m* of $A(X)$. Alternatively, we say α is an *m-fold root* of $A(X)$. However, when we say "α is a root of A" without qualification about its multiplicity, we presuppose the multiplicity is positive, $m \geq 1$. A root is *simple* or *multiple* according as $m = 1$ or $m \geq 2$. We say A is *square-free* if it has no multiple roots.

C5. *If α is a root of $A(X)$ of multiplicity $m \geq 1$ then α is a root of its derivative $A'(X)$ of multiplicity $m - 1$.* *Proof.* Write $A(X) = (X - \alpha)^m B(X)$ where $B(\alpha) \neq 0$. Then $A'(X) = m(X - \alpha)^{m-1} B(X) + (X - \alpha)^m B'(X)$. Clearly α has multiplicity $\geq m - 1$ as a root of $A'(X)$. Writing

$$C(X) = \frac{A'(X)}{(X - \alpha)^{m-1}} = mB(X) + (X - \alpha)B'(X),$$

we conclude $C(\alpha) = mB(\alpha) \neq 0$. Hence α has multiplicity exactly $m - 1$ as a root of $A'(X)$. **Q.E.D.**

Corollary 1 *The polynomial*

$$\frac{A(X)}{\texttt{GCD}(A(X), A'(X))} \tag{2}$$

is square-free and contains exactly the distinct roots of $A(X)$.

C6. The derivative $A'(X)$ can be expressed in the form

$$\frac{A'(X)}{A(X)} = \frac{1}{X - \alpha_1} + \frac{1}{X - \alpha_2} + \cdots + \frac{1}{X - \alpha_n}.$$

This follows by taking derivatives on both sides of equation (1) and dividing by $A(X)$. The rational function $A'(X)/A(X)$ is also called the *logarithmic derivative* since it is equal to $\frac{d \log A(X)}{dX}$. There are very interesting physical interpretations of A'/A. See [127, p. 6].

C7. *$A(X)$ is a continuous function of X.* This follows from the continuity of the multiplication and addition functions.

C8. *The roots $A(X)$ are continuous functions of the coefficients of $A(X)$.* We need to state this precisely: suppose $\alpha_1, \ldots, \alpha_k$ are the distinct roots of $A(X)$ where α_i has multiplicity $m_i \geq 1$, and let $D_1, \ldots, D_k$ be any set of discs such that each D_i contains α_i but not α_j if $j \neq i$. Then there exists an $\epsilon > 0$ such that for all $\epsilon_0, \ldots, \epsilon_n$ with $|\epsilon_i| < \epsilon$, the polynomial

$$B(X) = \sum_{i=0}^{n} (a_i + \epsilon_i) X^i$$

has exactly m_i roots (counted with multiplicity) inside D_i for $i = 1, \ldots, k$. For a proof, see [127, p. 3].

C9. For any $c \in \mathbb{C}$, there is a root $\alpha^* \in \mathbb{C}$ of $A(X)$ such that

$$|c - \alpha^*| \leq (|A(c)|/|a_n|)^{1/n}.$$

In proof, observe that $|A(c)| = |a_n| \prod_{i=1}^{n} |c - \alpha_i|$ in the notation of equation (1). We just choose α^* to minimize $|c - \alpha_i|$. As a corollary, the root α^* of smallest modulus satisfies

$$|\alpha^*| \leq \left(\frac{|a_0|}{|a_n|} \right)^{1/n}.$$

Real Polynomials. *The remaining properties assume that $A(X) \in \mathbb{R}[X]$.*

R1. *The non-real roots of $A(X)$ appear in conjugate pairs.*

Proof. For a real polynomial $A(X)$, we may easily verify that $\overline{A(\alpha)} = A(\overline{\alpha})$ for any complex number α, *i.e.*, complex conjugation and polynomial evaluation commute. Thus $A(\alpha) = 0$ implies $A(\overline{\alpha}) = 0$. **Q.E.D.**

As $(X - \alpha)(X - \overline{\alpha})$ is a real polynomial, we conclude:

Corollary 2
1) *$A(X)$ can be written as a product of real factors that are linear or quadratic.*
2) *If $n = \deg A$ is odd, then $A(X)$ has at least one real root.*

R2. Let X range over the reals. The sign of $A(X)$ as $|X| \to \infty$ is the sign of $a_n X^n$. The sign of $A(X)$ as $|X| \to 0$ is the sign of $a_i X^i$ where i is the smallest index such that $a_i \neq 0$.

R3. If $\alpha < \beta$ are two real numbers such that $A(\alpha)A(\beta) < 0$ then there exists γ ($\alpha < \gamma < \beta$) such that $A(\gamma) = 0$.

R4. Let $\epsilon > 0$ approach 0. Then for any real root α,

$$\frac{A'(\alpha - \epsilon)}{A(\alpha - \epsilon)} \to -\infty, \qquad \frac{A'(\alpha + \epsilon)}{A(\alpha + \epsilon)} \to +\infty.$$

In other words, when X is just slightly smaller than α, $A'(X)$ and $A(X)$ have different signs and when X is just slightly greater than α, they have the same signs. See Figure 1. *Proof.* From Property C6. we see that of $\frac{A'(\alpha-\epsilon)}{A(\alpha-\epsilon)}$ approaches $\frac{1}{-\epsilon}$ as ϵ approaches 0+. But $\frac{1}{-\epsilon} \to -\infty$. Similarly for the other case. **Q.E.D.**

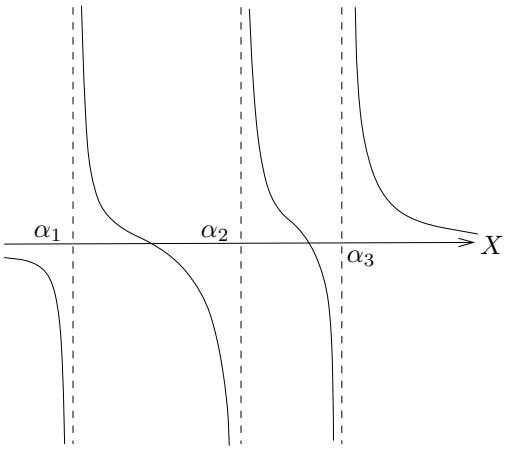


Figure 1: Illustrating the function $A'(X)/A(X)$.

R5. **Theorem 3 (Rolle's theorem)** *Between two consecutive real roots of $A(X)$ there is an odd number of real roots of $A'(X)$.*

Proof. Apply the previous property: if $\alpha < \beta$ are two consecutive real roots then $A(X)$ has constant sign in the interval (α, β). However $A'(X)$ has the same sign as $A(X)$ near α^+ and has a different sign from $A(X)$ near β^-. By continuity of $A'(X)$, it must be zero an odd number of times. **Q.E.D.**

Corollary 4 *Between any two consecutive real roots of $A'(X)$ there is at most one real root of $A(X)$.*

Exercises

Exercise 1.1: (Lucas, 1874) Any convex region K of the complex plane containing all the complex roots of $A(X)$ also contains all the complex roots of $A'(X)$. NOTE: This is the complex analogue of Rolle's theorem for real roots. Much is known about the location of the roots of the derivative of a polynomial; see [127]. □

Exercise 1.2: (Jensen, 1912) Let $A(X)$ be a real polynomial, so its non-real roots occur in conjugate pairs. A *Jensen circle* of $A(X)$ is a circle with diameter determined by one of these conjugate pair of roots. Then all the non-real zeros of $A'(X)$ lie on or inside the union of the Jensen circles of $A(X)$. □

Exercise 1.3: (Rouché, 1862) Suppose $P(Z), Q(Z)$ are analytic inside a Jordan curve C, are continuous on C, and satisfy $|P(Z)| < |Q(Z)|$ on C. Then $F(Z) = P(Z) + Q(Z)$ has the same number of zeros inside C as $Q(Z)$. □

Exercise 1.4: (Champagne) We improve the root bound in Property C9. Suppose the roots $\alpha_1, \ldots, \alpha_k$ $(k = 0, \ldots, n-1)$ have been "isolated": this means that there are discs D_i

$(i = 1, \ldots, k)$ centered in c_i with radii $r_i > 0$ such that each D_i contains α_i and no other roots and the D_i's are pairwise disjoint. Then for any c chosen outside the union $\cup_{i=1}^{k} D_i$ of these discs, there is a root $\alpha^* \in \{\alpha_{k+1}, \ldots, \alpha_n\}$ such that

$$|c - \alpha^*| \leq \frac{|A(c)|}{|a_n|} \prod_{i=2}^{n} (|c - c_i| - r_i)^{-1}.$$

□

Exercise 1.5: Give a lower bound on $|\alpha_i|$ using the technique in Property C9. □

Exercise 1.6: [Newton] If the polynomial $A(X) = a_n X^n + \binom{n}{1} a_{n-1} X^{n-1} + \cdots + \binom{n}{n-1} a_1 X + a_0$, $a_n \neq 0$ has real coefficients and n real roots then $a_i^2 \geq a_{i-1} a_{i+1}$ for $i = 1, \ldots, n-1$. HINT: Obvious for $n = 2$ and inductively use Rolle's theorem. □

§2. Root Bounds

Let

$$A(X) = \sum_{i=0}^{n} a_i X^i, \quad a_n \neq 0$$

where $a_i \in \mathbb{C}$, and let $\alpha \in \mathbb{C}$ denote any root of $A(X)$. To avoid exceptions below, we will also assume $a_0 \neq 0$ so that $\alpha \neq 0$. Our goal here is to give upper and lower bounds on $|\alpha|$. One such bound (§IV.5) is the Landau bound,

$$|\alpha| \leq \|A\|_2 / |a_n|. \tag{3}$$

And since $1/\alpha$ is a root of $X^n A(1/X^n)$, we also get $1/|\alpha| \leq \|A\|_2/|a_i|$ where i is the largest subscript such that $a_i \neq 0$. Thus $|\alpha| \geq |a_i|/\|A\|_2$. We next obtain a number of similar bounds.

Knuth attributes the following to Zassenhaus but Ostrowski [151, p.125] says it is well-known, and notes an improvement (Exercise) going back to Lagrange.

Lemma 5 *We have $|\alpha| < 2\beta$ where*

$$\beta := \max\left\{ \frac{|a_{n-1}|}{|a_n|}, \sqrt{\frac{|a_{n-2}|}{|a_n|}}, \sqrt[3]{\frac{|a_{n-3}|}{|a_n|}}, \ldots, \sqrt[n]{\frac{|a_0|}{|a_n|}} \right\}.$$

Proof. The lemma is trivial if $|\alpha| \leq \beta$; so assume otherwise. Since $A(\alpha) = 0$, $a_n \alpha^n = -(a_{n-1}\alpha^{n-1} + \cdots + a_0)$. Hence

$$\begin{aligned}
|a_n| \cdot |\alpha|^n &\leq |a_{n-1}| \cdot |\alpha|^{n-1} + |a_{n-2}| \cdot |\alpha|^{n-2} + \cdots + |a_0|. \\
1 &\leq \frac{|a_{n-1}|}{|a_n|} \cdot \frac{1}{|\alpha|} + \frac{|a_{n-2}|}{|a_n|} \cdot \frac{1}{|\alpha|^2} + \cdots + \frac{|a_0|}{|a_n|} \cdot \frac{1}{|\alpha|^n} \\
&\leq \frac{\beta}{|\alpha|} + \frac{\beta^2}{|\alpha|^2} + \cdots + \frac{\beta^n}{|\alpha|^n} \\
&< \frac{\beta/|\alpha|}{1 - (\beta/|\alpha|)},
\end{aligned}$$

where the last step uses our assumption $|\alpha| > \beta$. This yields the bound $|\alpha| < 2\beta$. **Q.E.D.**

Assuming $|a_n| \geq 1$ (as when $A(X) \in \mathbb{Z}[X]$), we get

$$|\alpha| < 2\|A\|_\infty. \tag{4}$$

Similarly, by applying this result to the polynomial $X^n A(1/X)$, assuming $|a_0| \geq 1$, we get

$$|\alpha| > \frac{1}{2\|A\|_\infty}. \tag{5}$$

Now define

$$\gamma := \max\left\{ \frac{|a_{n-1}|}{\binom{n}{1}|a_n|}, \sqrt{\frac{|a_{n-2}|}{\binom{n}{2}|a_n|}}, \sqrt[3]{\frac{|a_{n-3}|}{\binom{n}{3}|a_n|}}, \ldots, \sqrt[n]{\frac{|a_0|}{\binom{n}{n}|a_n|}} \right\}.$$

We have not seen this bound in the literature:

Lemma 6

$$|\alpha| \leq \frac{\gamma}{\sqrt[n]{2} - 1}.$$

Proof. As before, we obtain

$$\begin{aligned} 1 &\leq \frac{|a_{n-1}|}{|a_n|} \cdot \frac{1}{|\alpha|} + \frac{|a_{n-2}|}{|a_n|} \cdot \frac{1}{|\alpha|^2} + \cdots + \frac{|a_0|}{|a_n|} \cdot \frac{1}{|\alpha|^n} \\ &\leq \binom{n}{1}\frac{\gamma}{|\alpha|} + \binom{n}{2}\frac{\gamma^2}{|\alpha|^2} + \cdots + \binom{n}{n}\frac{\gamma^n}{|\alpha|^n}, \\ 2 &\leq \left(1 + \frac{\gamma}{|\alpha|}\right)^n, \end{aligned}$$

from which the desired bound follows. **Q.E.D.**

Both β and γ are invariant under "scaling", *i.e.*, multiplying $A(X)$ by any non-zero constant.

Lemma 7 (Cauchy)

$$\frac{|a_0|}{|a_0| + \max\{|a_1|, \ldots, |a_n|\}} < |\alpha| < 1 + \frac{\max\{|a_0|, \ldots, |a_{n-1}|\}}{|a_n|}$$

Proof. We first show the upper bound for $|\alpha|$. If $|\alpha| \leq 1$ then the desired upper bound is immediate. So assume $|\alpha| > 1$. As in the previous proof

$$\begin{aligned} |a_n||\alpha|^n &\leq |a_{n-1}| \cdot |\alpha|^{n-1} + |a_{n-2}| \cdot |\alpha|^{n-2} + \cdots + |a_0| \\ &\leq \max\{|a_0|, \ldots, |a_{n-1}|\} \sum_{i=0}^{n-1} |\alpha|^i \\ &\leq \max\{|a_0|, \ldots, |a_{n-1}|\} \cdot \frac{(|\alpha|^n - 1)}{|\alpha| - 1}, \\ |\alpha| - 1 &< \frac{\max\{|a_0|, \ldots, |a_{n-1}|\}}{|a_n|}. \end{aligned}$$

This shows the upper bound. The lower bound is obtained by applying the upper bound to the polynomial $X^n A\left(\frac{1}{X}\right)$ which has $\frac{1}{\alpha}$ as root (recall we assume $\alpha \neq 0$). We get

$$\left|\frac{1}{\alpha}\right| < 1 + \frac{\max\{|a_1|, \ldots, |a_n|\}}{|a_0|}$$

from which the lower bound follows. **Q.E.D.**

Corollary 8

$$\frac{|a_0|}{|a_0| + \|A\|_\infty} < |\alpha| < 1 + \frac{\|A\|_\infty}{|a_n|}.$$

If $A(X)$ is an integer polynomial and $a_0 \neq 0$, we may conclude

$$\frac{1}{1 + \|A\|_\infty} < |\alpha| < 1 + \|A\|_\infty.$$

Lemma 9 (Cauchy)

$$|\alpha| \leq \max\left\{\frac{n|a_{n-1}|}{|a_n|}, \sqrt{\frac{n|a_{n-2}|}{|a_n|}}, \sqrt[3]{\frac{n|a_{n-3}|}{|a_n|}}, \ldots, \sqrt[n]{\frac{n|a_0|}{|a_n|}}\right\}$$

Proof. If k is the index such that

$$|a_k| \cdot |\alpha|^k$$

is maximum among all $|a_i| \cdot |\alpha|^i \quad (i = 0, \ldots, n-1)$ then the initial inequality of the previous proof yields

$$\begin{aligned} |a_n| \cdot |\alpha|^n &\leq n|a_k| \cdot |\alpha|^k \\ |\alpha| &\leq \sqrt[n-k]{\tfrac{n \cdot |a_k|}{|a_n|}}. \end{aligned}$$

The lemma follows. **Q.E.D.**

Many other types of root bounds are known. For instance, for polynomials with real coefficients, we can exploit the signs of these coefficients (see exercises). One can also bound the product of the absolute values of the roots of a polynomial [132]. See also [147, 131].

Exercises

Exercise 2.1: i) Apply the β and γ root bounds to the roots of the characteristic polynomial of an $n \times n$ matrix A whose entries are integers of magnitude at most c.
ii) Compare the β and γ bounds generally. □

Exercise 2.2: (Cauchy) Let R be the unique positive root of the polynomial

$$B(Z) = |a_n|Z^n - (|a_{n-1}|Z^{n-1} + \cdots + |a_0|) = 0, \quad a_n \neq 0.$$

Then all the zeros of $A(Z) = \sum_{i=0}^n a_i Z^i$ lie in the circle $|Z| \leq R$. □

Exercise 2.3: (Real root bounds for real polynomials) Consider the polynomial $A(X) = \sum_{i=0}^{n} a_i x^i$ ($a_n \neq 0$) with real coefficients a_i. The following describes L which is understood to be a bound on the real roots of $A(X)$, not a bound on the absolute values of these roots.

i) [Rule of Lagrange and MacLaurin] Let $a = \max\{|a_i/a_n| : a_i < 0\}$, and let $n - m$ be the largest index i of a negative coefficient a_i. Then $L = 1 + a^{1/m}$.

ii) [Newton's rule] Let L be[2] any real number such that each of the first n derivatives of $A(x)$ evaluates to a positive number at L.

iii) [Laguerre's rule] Let $L > 0$ have the property that when $A(X)$ is divided by $X - L$, the quotient $B(X)$ is a polynomial with positive coefficients and the remainder r is a positive constant.

iv) [Cauchy's rule] Let the negative coefficients in $A(X)$ have indices $i_1, i_2, \ldots, i_k$ for some k. Then let L be the maximum of

$$(k \cdot |a_{i_1}|)^{1/(n-i_1)}, (k \cdot |a_{i_2}|)^{1/(n-i_2)}, \ldots, (k \cdot |a_{i_k}|)^{1/(n-i_k)}.$$

v) [Grouping method] Let $f(X)$ and $g(X)$ be polynomials with non-negative coefficients such that the biggest exponent of X in $g(X)$ is not bigger than the smallest exponent of X in $f(X)$. If $L > 0$ is such that $F(L) := f(L) - g(L) > 0$, then $F(X) > 0$ for all $X > L$. □

Exercise 2.4: Let $A(X) = X^5 - 10X^4 + 15X^3 + 4X^2 - 16X + 400$. Apply the various root bounds to $A(X)$. □

Exercise 2.5: (Ostrowski, Lagrange) Improve the root bound above attributed to Zassenhaus: if α is a root of $A(X) = \sum_{i=0}^{n} a_i X_i$, ($a_0 = 1$), and we arrange the terms $a_i^{1/(n-i)}$ ($i = 1, \ldots, n$) in non-decreasing order, then the sum of the last two terms in this sequence is a bound on $|\alpha|$. □

Exercise 2.6: (J. Davenport) There is a root of the polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$ whose absolute value is at least $\beta/2n$ where β is the bound of Zassenhaus. HINT: Assume $A(X)$ is monic and say $\beta^k = |a_{n-k}|$ for some k. Then there are k roots whose product has absolute value at least $|a_{n-k}|\binom{n}{k}^{-1}$. □

Exercise 2.7: The following bounds are from Mahler [125]. Let $F(X)$ be a complex function, $F(0) \neq 0$, $\xi_1, \ldots, \xi_N$ are all the zeros of $F(X)$ (multiplicities taken into account) satisfying $|\xi_i| \leq r$ for all i, $F(X)$ is regular inside the closed disc $\{x : |x| \leq r\}$. Then Jensen's formula in analytic function theory says

$$\frac{1}{2\pi}\int_0^{2\pi} \log|F(re^{\mathbf{i}\theta})|d\theta = \log|F(0)| + \sum_{i=1}^{N} \log\frac{r}{|\xi_i|}.$$

(a) Let $f(X) = \sum_{i=0}^{n} a_i X^i \in \mathbb{C}[X]$, $a_0 a_n \neq 0$. Let the roots of $f(X)$ be $\xi_1, \ldots, \xi_n$ such that

$$|\xi_1| \leq \cdots \leq |\xi_N| \leq 1 < |\xi_{N+1}| \leq \cdots \leq |\xi_n|.$$

Apply Jensen's formula with $r = 1$ to show

$$\frac{1}{2\pi}\int_0^{2\pi} \log|f(e^{\mathbf{i}\theta})|d\theta = \log|a_0\xi_{N+1}\cdots\xi_n| = \log|M(f)|.$$

[2]To apply this rule, first try to find a number L_1 such that the derivative $A^{(n-1)}(X)$ (which is linear in X) is positive or vanishes at L_1. Next consider $A^{(n-2)}(X)$. If this is negative, choose a number $L_2 > L_1$, etc.

(b) Show

$$\log(2^{-n}\|f\|_1) \leq \frac{1}{2\pi}\int_0^{2\pi} \log|f(e^{\mathbf{i}\theta})|d\theta \leq \log\|f\|_1.$$

HINT: The second inequality is immediate, and gives a bound on $M(f)$ weaker than Landau's.

(c) [Feldman-Mahler] Show for any subset $\{\xi_{i_1}, \ldots, \xi_{i_m}\}$ of $\{\xi_1, \ldots, \xi_n\}$,

$$|a_0\xi_{i_1}\cdots\xi_{i_m}| \leq \|f\|_1.$$

(d) [Gelfond] If $f(X) = \prod_{i=1}^{s} f_i(X)$, $n = \deg f$, then

$$2^n\|f\|_1 \geq \prod_{i=1}^{s}\|f_i\|_1.$$

□

§3. Algebraic Numbers

Our original interest was finding roots of polynomials in $\mathbb{Z}[X]$. By extending from $\mathbb{Z}$ to $\mathbb{C}$, every integer polynomial has a solution in $\mathbb{C}$. But it is not necessary to go so far: every integer polynomial has a root in the algebraic closure[3] $\overline{\mathbb{Z}}$ of $\mathbb{Z}$. In general, we let

$$\overline{D}$$

denote the algebraic closure of a domain D. Just as the concept of a UFD is characterized as a domain in which the Fundamental Theorem of Arithmetic holds, *an algebraically closed domain can be characterized as one in which the Fundamental Theorem of Algebra holds.*

By definition, an *algebraic number* α is an element in $\mathbb{C}$ that is the zero of some polynomial $P(X) \in \mathbb{Z}[X]$. For instance $\sqrt{2}$ and $\mathbf{i} = \sqrt{-1}$ are algebraic numbers. We call $P(X)$ a *minimal polynomial* of α if the degree of $P(X)$ is minimum. To make this polynomial unique (see below) we further insist that $P(X)$ be primitive and its leading coefficient is a distinguished element of $\mathbb{Z}$; then we call $P(X)$ *the* minimal polynomial of α. Note that minimal polynomials must be irreducible. The *degree of* α is the degree of its minimal polynomial. By definition, if $\alpha = 0$, its unique minimal polynomial is 0 with degree $-\infty$. Clearly, every algebraic number belongs to $\overline{\mathbb{Z}}$. In §5, we show that $\overline{\mathbb{Z}}$ is equal to the set of algebraic numbers; this justifies calling $\overline{\mathbb{Z}}$ the *field of algebraic numbers*.

A non-algebraic number in $\mathbb{C}$ is called a *transcendental number*. By Cantor's diagonalization argument, it is easy to see that transcendental numbers exist, and in abundance. Unfortunately, proofs that special numbers such π (circumference of the circle with unit diameter) and e (base of the natural logarithm) are transcendental are invariably non-trivial. Nevertheless, it is not hard to show explicit transcendental numbers using a simple argument from Liouville (1844). It is based on the fact that algebraic numbers cannot be approximated too closely. Here is the precise statement: if α is the irrational zero of an integral polynomial $A(X)$ of degree n, then there exists a constant $c = c(\alpha) > 0$ such that for all $p, q \in \mathbb{Z}$, $q > 0$,

$$|\alpha - \frac{p}{q}| \geq \frac{c}{q^n}. \tag{6}$$

[3]Concepts that are defined for fields can be applied to domains in the natural way: we simply apply the concept to the quotient field of the said domain. Thus the "algebraic closure" (which was defined for fields in §V.1) of a domain D refers to the algebraic closure of the quotient field of D.

Without loss of generality, assume $A'(\alpha) \neq 0$ (otherwise replace $A(X)$ by $A(X)/\texttt{GCD}(A(X), A'(X))$). Pick $\varepsilon > 0$ such that for all $\beta \in [\alpha \pm \varepsilon]$, $|A'(\beta)| \geq |A'(\alpha)|/2$. Now the one-line proof goes as follows:

$$\frac{1}{q^n} \leq \left|A\left(\frac{p}{q}\right)\right| = \left|A\left(\frac{p}{q}\right) - A(\alpha)\right| = \left|\alpha - \frac{p}{q}\right| \cdot |A'(\beta)| \tag{7}$$

for some $\beta = t\alpha + (1-t)p/q$, $0 \leq t \leq 1$, and the last equality uses the Mean Value Theorem or Taylor's expansion (§10). This proves

$$\left|\alpha - \frac{p}{q}\right| \geq \min\{\varepsilon, \frac{2}{q^n}|A'(\alpha)\} \geq \frac{c}{q^n}$$

($c = \min\{\varepsilon, 2/|A'(\alpha)|\}$. With this, we can deduce that certain explicitly constructed numbers are transcendental; for instance,

$$\alpha = \frac{1}{2} + \frac{1}{2^{2!}} + \frac{1}{2^{3!}} + \frac{1}{2^{4!}} + \cdots + \frac{1}{2^{n!}} + \cdots. \tag{8}$$

There touches on the deep subject of Diophantine approximation. Let us define $\kappa(n)$ to be the least number such that for all $\nu > \kappa(n)$ and all algebraic numbers α of degree $n \geq 2$, there exists a constant $c = c(\alpha, \nu)$ such that

$$|\alpha - \frac{p}{q}| \geq \frac{c}{q^\nu}. \tag{9}$$

Thus (6) shows $\kappa(n) \leq n$. This bound on $\kappa(n)$ steadily decreased starting with Thue (1909), Siegel (1921), independently by Dyson and Gelfond (1947), until Roth (1955) finally showed that $\kappa(n) \leq 2$. This is the best possible since for every irrational α, there are infinitely many solutions to $|\alpha - p/q| < 1/q^2$ (see §XIV.6). It should be noted that the constant $c = c(\alpha, \nu)$ here is ineffective in all the cited development (but it is clearly effective in the original Liouville argument).

Number Fields. All the computations in our lectures take place exclusively in $\overline{\mathbb{Z}}$. In fact, in any particular instance of a problem, all computations occur in a subfield of the form

$$\mathbb{Q}(\alpha) \subseteq \overline{\mathbb{Z}}$$

for some $\alpha \in \overline{\mathbb{Z}}$. This is because all computations involve only finitely many algebraic numbers, and by the primitive element theorem, these belong to one field $\mathbb{Q}(\alpha)$ for some α. Subfields of the form $\mathbb{Q}(\alpha)$ are called *number fields*.

Basic arithmetic in a number field $\mathbb{Q}(\alpha)$ is quite easy to describe. Let $P(X)$ be the minimal polynomial of α. If α is of degree d, then it follows from $P(\alpha) = 0$ that α^d can be expressed as a polynomial of degree at most $d-1$ in α, with coefficients in $\mathbb{Q}$. It follows that every element of $\mathbb{Q}(\alpha)$ can be written as a polynomial of degree at most $d-1$ in α. Viewed as a vector space over $\mathbb{Q}$, $\mathbb{Q}(\alpha)$ has dimension d; in particular,

$$1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$$

is a basis of this vector space. An element $\beta = \sum_{i=0}^{d-1} b_i \alpha^i$ in $\mathbb{Q}(\alpha)$ is uniquely determined by its coefficients $(b_0, b_1, \ldots, b_{d-1})$ (otherwise, we would have a vanishing polynomial in α of degree less than d). Addition and subtraction in $\mathbb{Q}(\alpha)$ are performed component-wise in these coefficients. Multiplication can be done as in polynomial multiplication, followed by a reduction to a polynomial of degree at most $d-1$. What about the inverse of an element? Suppose $Q(X) = \sum_{i=0}^{d-1} b_i X^i$ and we want the inverse of $\beta = Q(\alpha)$. Since $P(X)$ is irreducible, $Q(X), P(X)$ are relatively prime in $\mathbb{Q}[X]$. So by the extended Euclidean algorithm (§II.4), there exist $A(X), B(X) \in \mathbb{Q}[X]$ such that $A(X)P(X) + B(X)Q(X) = 1$. Hence $B(\alpha)Q(\alpha) = 1$, or $B(\alpha) = \beta^{-1}$.

Arithmetical Structure of Number Fields. An algebraic number α is *integral* (or, an *algebraic integer*) if α is the root of a monic integer polynomial. The set of algebraic integers is denoted $\mathbb{O}$. As expected, ordinary integers are algebraic integers. In the next section, we show that $\mathbb{O}$ is a subring of $\overline{\mathbb{Z}}$. This subring plays a role analogous to that of $\mathbb{Z}$ inside $\mathbb{Q}$ (or, as we say, gives the algebraic numbers its "arithmetical structure"). Denote the set of algebraic integers in a number field $\mathbb{Q}(\alpha)$ by

$$\mathbb{O}_\alpha := \mathbb{Q}(\alpha) \cap \mathbb{O},$$

which we call *number rings*. The simplest example of a number ring is $\mathbb{O}_{\mathbf{i}}$, called the ring of *Gaussian integer*. It turns out that $\mathbb{O}_{\mathbf{i}} = \mathbb{Z}[\mathbf{i}]$. On the other hand, $\frac{1}{2}(1-\sqrt{5})$ is an algebraic integer since it is the root of $X^2 - X - 1$. This shows that $\mathbb{O}(\alpha)$ is not always of the form $\mathbb{Z}[\alpha]$.

Let the minimal polynomial of α be $P(X)$; if α is also an algebraic integer then it is the root of a monic polynomial $Q(X)$ of minimal degree. The following shows that $P(X) = Q(X)$.

Lemma 10
(i) Let $P(X), Q(X) \in \mathbb{Z}[X]$ such that $P(X)$ is the minimal polynomial of α and $Q(\alpha) = 0$. Then $P(X)$ divides $Q(X)$.
(ii) The minimal polynomial of an algebraic number is unique.
(iii) The minimal polynomial of an algebraic integer is monic.

Proof. (i) By the Division Property for polynomials (§II.3), $Q(X) = b(X)P(X) + r(X)$ for some rational polynomials $b(X), r(X) \in \mathbb{Q}[X]$ where $\deg(r) < \deg(P)$. Hence $Q(\alpha) = P(\alpha) = 0$ implies $r(\alpha) = 0$. Since P is the minimal polynomial, this means $r(X)$ is the zero polynomial, *i.e.*, $Q(X) = b(X)P(X)$. We may choose $\xi \in \mathbb{Q}$ such that $\xi \cdot b(X)$ is a primitive integral polynomial. By Gauss' Lemma (§III.1), $\xi Q(X)$ is a primitive polynomial since it is the product of two primitive polynomials, $\xi Q(X) = \xi b(X) \cdot P(X)$. Thus $P(X)$ divides $\xi Q(X)$. By primitive factorization in $\mathbb{Z}[X]$ (§III.1), $\xi Q(X)$ equals $\texttt{prim}(Q(X))$. Hence $P(X)$ divides $\texttt{prim}(Q(X))$ which divides $Q(X)$.
(ii) If $\deg Q = \deg P$ this means P, Q are associates. But the distinguished element of a set of associates is unique. Uniqueness of the minimal polynomial for α follows.
(iii) If in (i) we also have that $Q(X)$ is primitive then unique factorization implies $\xi = 1$. If α is an algebraic integer, let $Q(X)$ be a monic polynomial such that $Q(\alpha) = 0$. Then in part (i), $\xi = 1$ implies $Q(X) = b(X)P(X)$ and hence $P(X)$ must be monic. **Q.E.D.**

From this lemma (iii), we see that the only algebraic integers in $\mathbb{Q}$ is $\mathbb{Z}$:

$$\mathbb{O} \cap \mathbb{Q} = \mathbb{Z}.$$

The justifies calling the elements of $\mathbb{Z}$ the *rational integers* (but colloquially we just say *ordinary integers*).

Lemma 11 *Every algebraic number has the form $\alpha = \beta/n$ where β is an algebraic integer and $n \in \mathbb{Z}$.*

Proof. Say α is a root of $P(X) \in \mathbb{Z}[X]$. If $P(X) = \sum_{i=0}^n a_i X^i$ where $a = a_n$ then

$$\begin{aligned} a^{n-1}P(X) &= \sum_{i=0}^{n} a_i a^{n-1-i}(aX)^i \\ &= Q(aX) \end{aligned}$$

where $Q(Y) = \sum_{i=0}^{n}(a_i a^{n-1-i})Y^i$ is a monic polynomial. So $a\alpha$ is an algebraic integer since it is a root of $Q(Y)$. **Q.E.D.**

We extend in a routine way the basic arithmetical concepts to number rings. Let us fix a number ring $\mathbb{O}_\alpha$ and simply say "integer" for an element of this ring. For integers a, b, we say *a divides b* (denoted $a|b$) if there exists an integer c such that $ac = b$. A *unit* u is an integer that divides 1 (and hence every integer in the number ring). Alternatively, u and u^{-1} are both integers iff they are both units. It is easy to see that u is the root of a monic polynomial $P(X)$ whose constant term is unity, 1. Two integers a, b are *associates* if $a = ub$ for some unit u. An integer is *irreducible* if it is only divisible by units and its associates.

Just as a number field is a vector space, its underlying number ring is a lattice. We will investigate the geometric properties of lattices in Lecture VIII. A basic property about $\mathbb{O}_\alpha$ is that it has a integral basis, $\omega_1, \ldots, \omega_n$, meaning that ω_i are integers and every integer is a rational integral combination of these ω_i's.

Remarks. The recent book of Cohen [43] is a source on algebraic number computations. See also Zimmer [223].

Exercises

Exercise 3.1:

a) Complete the one-line argument in Liouville's result: choose the constant c in equation (6) from (7).

b) Show that α in equation (8) is transcendental. HINT: take $q = 2$.

c) Extend Liouville's argument to show that $|\alpha - \sqrt{p/q}| \geq Cq^{-(n+1)/2}$.

□

Exercise 3.2: Let $R \subseteq R'$ be rings. An element $\alpha \in R'$ that satisfies a monic polynomial in $R[X]$ is said to be *integral over* R. The set R^* of elements in R' that are integral over R' is called the *integral closure of* R in R'. Show that the integral closure R^* is a ring that contains R. NOTE: $\mathbb{O}_\alpha$ can thus be defined to be the integral closure of $\mathbb{Z}$ in $\mathbb{Q}(\alpha)$. □

Exercise 3.3:

a) Show that $\mathbb{O}_{\sqrt{-1}} = \mathbb{Z}[\mathbf{i}]$ (the Gaussian integers).

b) Show that $\mathbb{O}_{\sqrt{-3}} = \mathbb{Z}[\omega] = \{m + n\omega : m, n \in \mathbb{Z}\}$ where $\omega = \frac{1+\sqrt{-3}}{2}$. NOTE: $\omega^2 = \omega - 1$.

c) Determine the quadratic integers. More precisely, determine $\mathbb{O}_{\sqrt{d}}$ for all square-free $d \in \mathbb{Z}$, $d \neq 1$. HINT: $\mathbb{O}_{\sqrt{d}} = \mathbb{Z}[\sqrt{d}]$ if $d \equiv 2$ or $d \equiv 3 \pmod 4$ and $\mathbb{O}_{\sqrt{d}} = \mathbb{Z}[\frac{\sqrt{d}-1}{2}]$ if $d \equiv 1 \pmod 4$.

d) Prove that $\mathbb{O}_\alpha$ is a subring of $\mathbb{Q}(\alpha)$. □

Exercise 3.4:

a) Show that $\alpha \in \mathbb{Q}(\sqrt{d})$ is integer iff the trace $Tr(\alpha) = \alpha + \overline{\alpha}$ and the norm $N(\alpha) = \alpha\overline{\alpha}$ are ordinary integers.

b) Every ideal $I \subseteq \mathbb{O}_{\sqrt{d}}$ is a *module*, *i.e.*, has the form $\mathbb{Z}[\alpha, \beta] := \{m\alpha + n\beta : m, n \in \mathbb{Z}\}$, for some $\alpha, \beta \in \mathbb{Q}(\sqrt{d})$.

c) A module $M = \mathbb{Z}[\alpha, \beta]$ is an ideal iff its *coefficient ring* $\{x \in \mathbb{Q}(\sqrt{d}) : xM \subseteq M\}$ is precisely $\mathbb{O}_{\sqrt{d}}$. □

Exercise 3.5: (H. Mann) Let θ be a root of $X^3 + 4X + 7$. Show that $1, \theta, \theta^2$ is an integral basis for $\mathbb{Q}(\theta)$. □

§4. Resultants

There are two classical methods for obtaining basic properties of algebraic numbers, namely the theory of symmetric functions and the theory of resultants (§III.3). Here we take the latter approach. We first develop some properties of the resultant.

The results in this section are valid in any algebraically closed field. We fix such a field $D = \overline{D}$.

Lemma 12 *Let $A, B \in D[X]$ with $\deg A = m, \deg B = n$ and let $\alpha, \beta \in D$.*

(i) $\texttt{res}(\alpha, B) = \alpha^n$. *By definition,* $\texttt{res}(\alpha, \beta) = 1$.

(ii) $\texttt{res}(X - \alpha, B) = B(\alpha)$.

(iii) $\texttt{res}(A, B) = (-1)^{mn}\texttt{res}(B, A)$.

(iv) $\texttt{res}(\alpha A, B) = \alpha^n \texttt{res}(A, B)$.

We leave the simple proof of this lemma as an exercise (the proof of (ii) is instructive).

Lemma 13 $\texttt{res}(A, B) = 0$ *if and only if* $\texttt{GCD}(A, B)$ *is non-constant.*

Proof. Although this is a special case of the Fundamental theorem of subresultants, it is instructive to give a direct proof. Suppose $\texttt{res}(A, B) = 0$. Consider the matrix equation

$$w \cdot S = \mathbf{0}, \quad w = (u_{n-1}, u_{n-2}, \ldots, u_0, v_{m-1}, \ldots, v_0), \tag{10}$$

where S is the $(n+m)$–square Sylvester matrix of A, B and w is a row $(m+n)$-vector of unknowns. There is a non-trivial solution w since $\det(S) = \texttt{res}(A, B) = 0$. Now define $U = \sum_{j=0}^{n-1} u_j X^j$ and $V = \sum_{i=0}^{m-1} v_i X^i$. Then (10), or rather $w \cdot S \cdot x = 0$ where $x = (X^{m+n-1}, X^{m+n-2}, \ldots, X, 1)^T$, amounts to the polynomial equation

$$UA + VB = 0.$$

But by the unique factorization theorem for $D[X]$ (recall D is a field), A has a factor of degree at most $m-1$ in G and hence a factor of degree at least 1 in B. This show $\texttt{GCD}(A, B)$ is non-constant, as desired. Conversely, if $\texttt{GCD}(A, B)$ has positive degree, then the equation

$$\widehat{B}A - \widehat{A}B = 0$$

holds where $\widehat{B} = B/\texttt{GCD}(A, B), \widehat{A} = A/\texttt{GCD}(A, B)$. This can be written in the matrix form (10) as before. Thus $0 = \det(S) = \texttt{res}(A, B)$. **Q.E.D.**

Proposition 12(ii) is a special case of the following (see [33, p. 177] for a different proof):

Lemma 14 *Let $A, B \in D[X]$, $\alpha \in D$, $\deg B > 0$. Then*

$$\texttt{res}((X - \alpha) \cdot A, B) = B(\alpha)\texttt{res}(A, B).$$

Proof. Let $A^* = (X-\alpha)\cdot A$, $m = \deg A^*, n = \deg B$ and M the Sylvester matrix of A^*, B. Writing $A(X) = \sum_{i=0}^{m-1} a_i X^i$ and $B(X) = \sum_{i=0}^{n} b_i X^i$, then M is given by

$$\left[\begin{array}{ccccccccc}
a_{m-1} & a_{m-2}-\alpha a_{m-1} & a_{m-3}-\alpha a_{m-2} & \cdots & a_0-\alpha a_1 & -\alpha a_0 & & & \\
 & a_{m-1} & a_{m-2}-\alpha a_{m-1} & \cdots & a_1-\alpha a_2 & a_0-\alpha a_1 & -\alpha a_0 & & \\
 & & \ddots & & & & & \ddots & \\
 & & & a_{m-1} & \cdots & & & a_0-\alpha a_1 & -\alpha a_0 \\
\hline
b_n & b_{n-1} & \cdots & b_1 & b_0 & & & & \\
 & b_n & \cdots & b_2 & b_1 & b_0 & & & \\
 & & \ddots & & & & \ddots & & \\
 & & & & b_n & b_{n-1} & \cdots & b_1 & b_0
\end{array}\right]$$

We now apply the following operations to M, in the indicated order: add α times column 1 to column 2, then add α times column 2 to column 3, etc. In general, we add α times column i to column $i+1$, for $i = 1, \ldots, m+n-1$. The resulting matrix M' can be succinctly described by introducing the notation

$$B/X^i, \quad (i \in \mathbb{Z})$$

to denote the integral part of $B(X)$ divided by X^i. For instance, $B/X^n = b_n$, $B/X^{n-1} = b_n X + b_{n-1}$ and $B/X = b_n X^{n-1} + b_{n-1}X^{n-2} + \cdots + b_2 X + b_1$. Note that if $i \leq 0$, then we are just multiplying $B(X)$ by X^{-i}, as in $B/X^0 = B(X)$ and $B/X^{-2} = X^2 B(X)$. Finally, we write B/α^i to denote the substitution of α for X in B/X^i. The matrix M' is therefore

$$M' = \left[\begin{array}{ccccccccc}
a_{m-1} & a_{m-2} & a_{m-3} & \cdots & a_0 & 0 & & & \\
 & a_{m-1} & a_{m-2} & \cdots & a_1 & a_0 & 0 & & \\
 & & \ddots & & & & \ddots & & \\
 & & & a_{m-1} & \cdots & & & a_0 & 0 \\
\hline
B/\alpha^n & B/\alpha^{n-1} & \cdots & B/\alpha^1 & B(\alpha) & B/\alpha^{-1} & \cdots & B/\alpha^{-m+2} & B/\alpha^{-m+1} \\
 & B/\alpha^n & \cdots & & B/\alpha^1 & B(\alpha) & \cdots & B/\alpha^{-m+3} & B/\alpha^{-m+2} \\
 & & \ddots & & & & \ddots & & \vdots \\
 & & & B/\alpha^n & B/\alpha^{n-1} & \cdots & B/\alpha^1 & B(\alpha) & B/\alpha^{-1} \\
 & & & & B/\alpha^n & B/\alpha^{n-1} & \cdots & B/\alpha^1 & B(\alpha)
\end{array}\right]$$

Note that if we subtract α times the last row from the last-but-one row, we transform that row into

$$(0, \ldots, 0, b_n, b_{n-1}, \ldots, b_1, b_0, 0).$$

In general, we subtract α times row $m+n-i+1$ from the $(m+n-i)$th row (for $i = 1, \ldots, m-1$), we obtain the matrix

$$M'' = \left[\begin{array}{ccccccccc}
a_{m-1} & a_{m-2} & a_{m-3} & \cdots & a_0 & 0 & & & \\
 & a_{m-1} & a_{m-2} & \cdots & a_1 & a_0 & 0 & & \\
 & & \ddots & & & & \ddots & \ddots & \\
 & & & a_{m-1} & \cdots & & a_1 & a_0 & 0 \\
\hline
b_n & b_{n-1} & \cdots & b_1 & b_0 & 0 & \cdots & 0 & 0 \\
 & b_n & \cdots & & b_1 & b_0 & & 0 & 0 \\
 & & \ddots & & & & \ddots & & \vdots \\
 & & & b_n & b_{n-1} & \cdots & & b_0 & 0 \\
 & & & & B/\alpha^n & B/\alpha^{n-1} & \cdots & B/\alpha^1 & B(\alpha)
\end{array}\right]$$

But the last column of M'' contains only one non-zero entry $B(\alpha)$ at the bottom right corner, and the co-factor of this entry is the determinant of the Sylvester matrix of A, B. Hence $\det M'' = B(\alpha)\texttt{res}(A,B)$. **Q.E.D.**

Theorem 15 *Let* $A, B \in D[X], a = \texttt{lead}(A), b = \texttt{lead}(B), \deg A = m, \deg B = n$ *with roots*

$$\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n \in \overline{D}.$$

Then $\texttt{res}(A, B)$ *is equal to each of the following expressions:*

(i) $a^n \prod_{i=1}^{m} B(\alpha_i)$

(ii) $(-1)^{mn} b^m \prod_{j=1}^{n} A(\beta_j)$

(iii) $a^n b^m \prod_{i=1}^{m} \prod_{j=1}^{n} (\alpha_i - \beta_j)$

Proof. Writing $A = a \prod_{i=1}^{m} (X - \alpha_i)$, we get from the previous lemma,

$$\begin{aligned} \texttt{res}(A, B) &= a^n \texttt{res}\left(\prod_{i=1}^{m} (X - \alpha_i), B \right) \\ &= a^n B(\alpha_1) \texttt{res}\left(\prod_{i=2}^{m} (X - \alpha_i), B \right) \\ &= \cdots \\ &= a^n B(\alpha_1) \cdots B(\alpha_m). \end{aligned}$$

This shows (i), and (ii) is similar. We deduce (iii) from (i) since

$$B(\alpha_i) = b \prod_{j=1}^{n} (\alpha_i - \beta_j).$$

Q.E.D.

The expression in part (i) of the theorem is also known as Poisson's definition of the resultant.

If A, B are multivariate polynomials, we can take their resultant by viewing them as univariate polynomials in any one of the variables Y. To indicate this, we write $\texttt{res}_Y(A, B)$.

Lemma 16 *Let* $A, B \in D[X]$ *and* $\alpha, \beta \in D$ *such that* $A(\alpha) = B(\beta) = 0$ *and* $\deg A = m, \deg B = n$.

(i) $1/\alpha$ *is the root of* $X^m A(1/X)$ *provided* $\alpha \neq 0$.

(ii) $\beta \pm \alpha$ *is a root of* $C(X) = \texttt{res}_Y(A(Y), B(X \mp Y))$.

(iii) $\alpha\beta$ *is a root of* $C(X) = \texttt{res}_Y(A(Y), Y^n B(\frac{X}{Y}))$.

Proof.

(i) This is immediate.

(ii) $$\begin{aligned}\texttt{res}_Y(A(Y), B(X \mp Y)) &= a^n \prod_{i=1}^{m} B(X \mp \alpha_i) \\ &= a^n b^m \prod_{i=1}^{m}\prod_{j=1}^{n}(X \mp \alpha_i - \beta_j).\end{aligned}$$

(iii) $$\begin{aligned}\texttt{res}_Y(A(Y), Y^n B(\tfrac{X}{Y})) &= a^n \prod_{i=1}^{m}(\alpha_i^n B(\frac{X}{\alpha_i})) \\ &= a^n \prod_{i=1}^{m}(b\alpha_i^n \prod_{j=1}^{n}(\frac{X}{\alpha_i} - \beta_j)) \\ &= a^n b^m \prod_{i=1}^{m}\prod_{j=1}^{n}(X - \alpha_i\beta_j).\end{aligned}$$

Q.E.D.

The proof of (ii) and (iii) shows that if $A(X), B(X)$ are monic then $C(X)$ is monic. Thus:

Corollary 17 *The algebraic integers form a ring: if α, β are algebraic integers, so are $\alpha \pm \beta$ and $\alpha\beta$.*

Corollary 18 *The set of algebraic numbers forms a field extension of $\mathbb{Q}$. Furthermore, if α, β are algebraic numbers of degrees m and n respectively then both $\alpha + \beta$ and $\alpha\beta$ have degrees $\leq mn$.*

Proof. We only have to verify the degree bounds. For $\alpha \pm \beta$, we must show that $\texttt{res}_Y(A(Y), B(X \mp Y))$ has X-degree at most mn. Let $M = (a_{i,j})$ be the $(m+n)$-square Sylvester matrix whose determinant equals $\texttt{res}_Y(A(Y), B(X \mp Y))$. Then the first n rows of M have constant entries (corresponding to coefficients of $A(Y)$) while the last m rows have entries that are polynomials in X (corresponding to coefficients of $B(X \mp Y)$, viewed as a polynomial in Y). Moreover, each entry in the last m rows has X-degree at most n. Thus each of $(m+n)!$ terms in the determinant of M is of X-degree at most mn. A similar argument holds for $\texttt{res}_Y(A(Y), Y^n B(X/Y))$. **Q.E.D.**

Computation of Resultants. The computation of resultants can be performed more efficiently than using the obvious determinant computation. This is based on the following simple observation: let $A(X) = B(X)Q(X) + R(X)$ where $m = \deg A, n = \deg B, r = \deg R$ and $m \geq n > r$. Then

$$\begin{aligned}\texttt{res}(A, B) &= (-1)^{n(m-r)} b^{m-r} \texttt{res}(R, B), \qquad (b = \texttt{lead}(B)) &(11)\\ &= (-1)^{mn} b^{m-r} \texttt{res}(B, R). &(12)\end{aligned}$$

Thus, the resultant of A, B can be expressed in terms of the resultant of B, R. Since R is the remainder of A divided by B, we can apply an Euclidean-like algorithm in case the coefficients come from a field F: given $A, B \in F[X]$, we construct the Euclidean remainder sequence (Lecture III):

$$A_0 = A, A_1 = B, A_2, \ldots, A_h$$

where $A_{i+1} = A_{i-1} \bmod A_i$ and $A_{h+1} = 0$. If $\deg A_h$ is non-constant, then $\texttt{res}(A, B) = 0$. Otherwise, we can repeatedly apply the formula of equation (11) until the basis case given by $\texttt{res}(A_{h-1}, A_h) = A_h^{\deg(A_{h-1})}$. This computation can further be sped up: J. Schwartz [186] has shown that the Half-GCD technique (Lecture II) can be applied to this problem.

Exercises

Exercise 4.1: Compute the minimal polynomial of $\sqrt{3} - \sqrt[3]{3} + 1$. □

Exercise 4.2: $\texttt{res}(AB, C) = \texttt{res}(A, C) \cdot \texttt{res}(B, C)$. □

Exercise 4.3: Let α, β be real algebraic numbers. Construct a polynomial $C(X) \in \mathbb{Z}[X]$ such that $C(\alpha + \mathbf{i}\beta) = 0$. Hence $\alpha + \mathbf{i}\beta$ is algebraic. Writing $C(\alpha + \mathbf{i}\beta) = C_0(\alpha, \beta) + \mathbf{i}C_1(\alpha, \beta)$, can you give direct constructions of C_0, C_1? □

Exercise 4.4: An algebraic integer that divides 1 is called a *unit*.
(i) An algebraic integer is a unit iff its minimal polynomial has trailing coefficient ± 1.
(ii) The inverse of a unit is a unit; the product of two units is a unit. □

Exercise 4.5: A root of a monic polynomial with coefficients that are algebraic integers is an algebraic integer. □

Exercise 4.6: (Projection) Let $R(Y)$ be the resultant, with respect to the variable X, of $F(X, Y)$ and $G(X, Y)$.
(i) Justify the interpretation of the roots of $R(Y)$ to be the projection of the set $F(X, Y) = G(X, Y) = 0$.
(ii) Suppose $G(X, Y)$ is derivative of $F(X, Y)$ with respect to X. Give the interpretation of the roots of $R(Y)$. □

Exercise 4.7: [Bezout-Dixon Resultant] With $A(X), B(X)$ as above, consider the bivariate polynomial,

$$D(X, Y) := \det \begin{bmatrix} A(X) & B(X) \\ A(Y) & B(Y) \end{bmatrix}.$$

(i) Show that $\Delta(X, Y) := \frac{D(X,Y)}{X-Y}$ is a polynomial.
(ii) The polynomial $\Delta(X, Y)$, regarded as a polynomial in Y is therefore of degree $m - 1$. Show that for every common root α of $A(X)$ and $B(X)$, $\Delta(\alpha, Y) = 0$. Conversely, show that if $\deg(A) = \deg(B)$ and $\Delta(\alpha, Y) = 0$ then α is a common root of A, B.
(iii) Construct a determinant $R(A, B)$ in the coefficients of $A(X)$ and $B(X)$ whose vanishing corresponds to the existence of a common root of $A(X)$ and $B(X)$¿
(iv) Construct $R(A, B)$ where $A(X)$ and $B(X)$ are polynomials of degree 2 with indeterminate coefficients. Confirm that this is (up to sign) equal to the Sylvester resultant of A, B.
(v) Construct $R(A, B)$ where A, B again has indeterminate coefficients and $\deg(A) = 2$ and $\deg(B) = 3$. What is the relation between $R(A, B)$ and the Sylvester resultant of A, B? In general, what can you say if $\deg(A) \neq \deg(B)$?
(vi) Design and analyze an efficient algorithm to compute $R(A, B)$.
(vii) [Dixon] Generalize this resultant construction to three bivariate polynomials, $A(X, Y), B(X, Y)$ and $C(X, Y)$. That is, construct a polynomial $R(A, B, C)$ in the coefficients of A, B, C such that A, B, C have a common solution iff $R(A, B, C) = 0$. □

§5. Symmetric Functions

The other approach to the basic properties of algebraic numbers is via the theory of symmetric functions. This approach is often the simplest way to show existence results (cf. theorem 23 below). But computationally, the use of symmetric functions seems inferior to resultants.

Consider polynomials in $D[\mathbf{X}] = D[X_1, \ldots, X_n]$ where D is a domain. Let S_n denote the set of all permutations on the set $\{1, 2, \ldots, n\}$. This is often called the *symmetric!group* on n-symbols. A polynomial $A(X_1, \ldots, X_n)$ is *symmetric in* $X_1, \ldots, X_n$ if for all permutations $\pi \in S_n$, we have $A(X_1, \ldots, X_n) = A(X_{\pi(1)}, \ldots, X_{\pi(n)})$. For example, the following set of functions are symmetric:

$$\begin{aligned}
\sigma_1(X_1, \ldots, X_n) &= \sum_{i=1}^{n} X_i, \\
\sigma_2(X_1, \ldots, X_n) &= \sum_{1 \le i < j \le n} X_i X_j, \\
&\vdots \\
\sigma_k(X_1, \ldots, X_n) &= \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} X_{i_1} X_{i_2} \cdots X_{i_k}, \\
&\vdots \\
\sigma_n(X_1, \ldots, X_n) &= X_1 X_2 \cdots X_n.
\end{aligned}$$

We call σ_i the *ith elementary symmetric function* (on $X_1, \ldots, X_n$). We could also define $\sigma_0 = 1$.

Let $e = (e_1, \ldots, e_n)$ where $e_1 \ge e_2 \ge \cdots \ge e_n \ge 0$. If $\pi \in S_n$, let $\mathbf{X}_\pi^e$ denote the power product

$$X_{\pi(1)}^{e_1} X_{\pi(2)}^{e_2} \cdots X_{\pi(n)}^{e_n}.$$

In case π is the identity, we write $\mathbf{X}^e$ instead of $\mathbf{X}_\pi^e$. In our inductive proofs, we need the *lexicographic!ordering* on n-tuples of numbers:

$$(d_1, \ldots, d_n) \underset{\texttt{LEX}}{\ge} (e_1, \ldots, e_n)$$

is defined to mean that the first non-zero entry of $(d_1 - e_1, \ldots, d_n - e_n)$, if any exists, is positive. If we identify a power product $\mathbf{X}^e$ with the n-tuple e, then the set $\texttt{PP} = \texttt{PP}(\mathbf{X})$ of power products can be identified with $\mathbb{N}^n$ and hence given the lexicographical ordering. There is a unique minimal element in PP, namely 1. In our proof below, we use the fact that PP is well-ordered[4] by the lexicographic ordering. This will be proved in a more general context in §XII.1.

We now introduce two classes of symmetric polynomials: first, define $G_e = G_{e_1,\ldots,e_n}$ to be the sum over all distinct terms in the multiset

$$\{\mathbf{X}_\pi^e : \pi \in S_n\}.$$

For example, $\sigma_1 = G_{1,0,\ldots,0}$, $\sigma_2 = G_{1,1,0,\ldots,0}$ and $\sigma_n = G_{1,1,\ldots,1}$.

Lemma 19 *G_e is symmetric.*

Proof. Clearly the expression

$$G'_e = \sum_{\pi \in S_n} \mathbf{X}_\pi^e$$

[4] A linearly ordered set S is *well-ordered* if every non-empty subset has a least element.

is symmetric. We only have to show that there is a constant c_e such that $G'_e = c_e G_e$. Let

$$Aut(e) := \{\pi \in S_n \,:\, \mathbf{X}^e_\pi = \mathbf{X}^e\}.$$

It is easy to see that $Aut(e)$ is a subgroup of S_n. For any $\rho \in S_n$, we have

$$Aut(e) \cdot \rho = \{\pi \in S_n \,:\, \mathbf{X}^e_{\pi\rho^{-1}} = \mathbf{X}^e\} = \{\pi \in S_n \,:\, \mathbf{X}^e_\pi = \mathbf{X}^e_\rho\}.$$

This shows that $\mathbf{X}^e_\rho$ occurs exactly $|Aut(e)|$ times in G'_e. As this number does not depend on ρ, we conclude the desired constant is $c_e = |Aut(e)|$. **Q.E.D.**

A *basic polynomial* is one of the form $a \cdot G_e$ for some $a \in D \setminus \{0\}$ and $e = (e_1, \ldots, e_n)$ where $e_1 \geq \cdots e_n \geq 0$. Clearly a symmetric polynomial $A(X_1, \ldots, X_n)$ can be written as a sum E of basic polynomials:

$$E = \sum_{i=1}^{m} a_i G_{e(i)}$$

where each $a_i G_{e(i)}$ is a basic polynomial. If the $e(i)$'s in this expression are distinct then the expression is unique (up to a permutation of the terms). Call this unique expression E the *basic decomposition* of $A(X_1, \ldots, X_n)$.

With $e = (e_1, \ldots, e_n)$ as before, the second class of polynomials has the form

$$\Gamma_e := \sigma_1^{e_1 - e_2} \sigma_2^{e_2 - e_3} \cdots \sigma_{n-1}^{e_{n-1} - e_n} \sigma_n^{e_n}. \tag{13}$$

Γ_e is symmetric since it is a product of symmetric polynomials. A *σ-basic polynomial* is one of the form $a \cdot \Gamma_e$, $a \in D \setminus \{0\}$. If a symmetric polynomial can be written as a sum E' of σ-basic polynomials, then E' is unique in the same sense as in a basic decomposition (Exercise). We call E' a *σ-basic decomposition*. The *σ-degree* of E' is the total degree of E' when viewed as a polynomial in the σ_i's. Likewise, say E' is *σ-homogeneous* if E' is homogeneous as a polynomial in the σ_i's. The next result shows that every symmetric polynomial has a σ-basic decomposition.

Examples. Let $n = 2$. The symmetric polynomial $A_1(X, Y) = X^2 + Y^2$ can be expressed as $A_1(X, Y) = (X+Y)^2 - 2XY = \sigma_1^2 - 2\sigma_2$. In fact, $A_1(X, Y)$ is the basic polynomial $G_{2,0}$, and it has the σ-basic decomposition $\Gamma_{2,0} - 2\Gamma_{2,2}$. Now let $n = 3$. $A_2(X, Y, Z) = (XY)^2 + (YZ)^2 + (ZX)^2$ can be written $\sigma_2^2 - 2\sigma_1\sigma_3$. Then $A_2 = G_{2,2,0}$ and has the σ-basic decomposition $\Gamma_{2,2,0} - 2\Gamma_{2,1,1}$.

The maximum degree (§0.10) of $A(X_1, \ldots, X_n)$ is the maximum X_i-degree of A, $i = 1, \ldots, n$. If A is symmetric, the maximum degree of A is equal to the X_i-degree for any i. Thus the maximum degrees of both A_1 and A_2 in these examples are 2. The σ-degree of their σ-basic decompositions are also 2.

We are ready to prove:

Theorem 20 (σ-Basic Decomposition of Symmetric Polynomials)
(i) Every symmetric polynomial $A(X_1, \ldots, X_n) \in D[X_1, \ldots, X_n]$ has a σ-basic decomposition E'.
(ii) If A has maximum degree d then E' has σ-degree d.
(iii) If A is homogeneous then E' is σ-homogeneous.

Proof. (i) The result is trivial if A is a constant polynomial. So assume otherwise. For some $e = (e_1, \ldots, e_n)$, the basic decomposition of $A(X_1, \ldots, X_n)$ has the form

$$A(X_1, \ldots, X_n) = a \cdot G_e + A', \qquad (0 \neq a \in D) \tag{14}$$

where A' involves power products that are lexicographically less than $\mathbf{X}^e$. Now consider Γ_e in equation (13): expanding each σ_i term into sums of monomials,

$$\begin{aligned}\Gamma_e &= (X_1 + \cdots + X_n)^{e_1-e_2}(X_1X_2 + \cdots + X_{n-1}X_n)^{e_2-e_3}\cdots(X_1X_2\cdots X_n)^{e_n} \\ &= \left\{(X_1)^{e_1-e_2}(X_1X_2)^{e_2-e_3}\cdots(X_1\cdots X_n)^{e_n}\right\} + \cdots \\ &\qquad \cdots + \left\{(X_n)^{e_1-e_2}(X_{n-1}X_n)^{e_2-e_3}\cdots(X_1\cdots X_n)^{e_n}\right\} \\ &= \left\{X_1^{e_1}X_2^{e_2}\cdots X_n^{e_n}\right\} + \cdots + \left\{X_n^{e_1}X_{n-1}^{e_2}\cdots X_1^{e_n}\right\}. \qquad (15)\end{aligned}$$

The basic decomposition of Γ_e contains the basic polynomial G_e: this follows from the presence of the term $X_1^{e_1}X_2^{e_2}\cdots X_n^{e_n}$ in equation (15) and the fact that Γ_e is symmetric. Thus

$$\Gamma_e = G_e + G' \qquad (16)$$

for some symmetric polynomial G'. Moreover, the basic decomposition of G' involves only power products that are lexicographically less than $\mathbf{X}^e$: this is clear since $\mathbf{X}^e$ is obtained by multiplying together the lexicographically largest power product in each σ-term. From equations (14) and (16), we conclude that

$$A(X_1,\ldots,X_n) = a\cdot\Gamma_e - a\cdot G' + A' \qquad (17)$$

where $-a\cdot G' + A'$ involves power products that are lexicographically less than $\mathbf{X}^e$. By the principal of induction for well-ordered sets (see Exercise), we easily conclude that A has a σ-basic decomposition.

(ii) Note that if A has maximum degree d then in equation (17), the σ-degree of $a\cdot\Gamma_e$ is d, and the maximum degree of $-a\cdot G' + A'$ is at most d. The result follows by induction.

(iii) Immediate. **Q.E.D.**

Since a σ-basic polynomial is a polynomial in the elementary symmetric functions, we conclude:

Theorem 21 (Fundamental Theorem of Symmetric Functions)
If $A(X_1,\ldots,X_n) \in D[\mathbf{X}]$ is a symmetric polynomial of maximum degree d, then there is a polynomial $B(X_1,\ldots,X_n) \in D[\mathbf{X}]$ of total degree d such that

$$A(X_1,\ldots,X_n) = B(\sigma_1,\ldots,\sigma_n)$$

and σ_i is the ith elementary symmetric function on $X_1,\ldots,X_n$. If A is homogeneous, so is B.

Let

$$\gamma^{(1)},\gamma^{(2)},\ldots,\gamma^{(n)}$$

be all the roots (not necessarily distinct) of a polynomial

$$G(X) = \sum_{i=0}^{n} g_iX^i, \qquad (g_i \in D).$$

Hence $G(X) = g_n\prod_{i=1}^n(X-\gamma^{(i)})$. Equating coefficients in the two expressions for $G(X)$,

$$(-1)^i g_{n-i} = g_n\cdot\sigma_i(\gamma^{(1)},\ldots,\gamma^{(n)})$$

for $i = 0,1,\ldots,n$. So the coefficient of X^{n-i} in $G(X)$ is, up to signs, equal to the product of g_n with the ith elementary symmetric function on the roots of $G(X)$. If $A \in D[X_1,\ldots,X_n]$ is symmetric and homogeneous of degree d, then by the Fundamental Theorem there exists $B \in D[X_1,\ldots,X_n]$ such that $g_n^dA(\gamma^{(1)},\ldots,\gamma^{(n)}) = B(-g_{n-1},\ldots,(-1)^ng_0)$, which is an element of D. If A is not homogeneous, the same argument can be applied to each homogeneous component of A, thus showing:

Theorem 22 *Let $G(X) \in D[X]$ be as above. If $A \in D[X_1, \ldots, X_n]$ is symmetric of degree d then $(g_n)^d A(\gamma^{(1)}, \ldots, \gamma^{(n)})$ is a polynomial in the coefficients of $G(X)$. In particular,*

$$(g_n)^d A(\gamma^{(1)}, \ldots, \gamma^{(n)}) \in D.$$

We give another application of the Fundamental Theorem:

Theorem 23 *If α is algebraic over E, and E is an algebraic extension of a domain D, then α is algebraic over D.*

Proof. Since α is algebraic over D iff it is algebraic over the quotient field of D, we may assume that D is a field in the following proof. Let α be the root of the polynomial $B(X) = \sum_{i=0}^n \beta_i X^i$ where $\beta_i \in E$. Let

$$R_i := \{\beta_i^{(j)} : j = 1, \ldots, d_i\} \tag{18}$$

be the set of conjugates of β_i over D. For each choice of $j_0, j_1, \ldots, j_n$, consider the 'conjugate' of $B(X)$,

$$B_{j_0,\ldots,j_n}(X) := \sum_{i=0}^n \beta_i^{(j_i)} X^i.$$

Form the polynomial

$$A(X) := \prod_{j_0, j_1, \ldots, j_n} B_{j_0,\ldots,j_n}(X).$$

Note that $B(X)|A(X)$ and hence α is a root of $A(X)$. The theorem follows if we show that $A(X) \in D[X]$. Fix any coefficient a_k in $A(X) = \sum_k a_k X^k$. Let $D_i := D[R_i, R_{i+1}, \ldots, R_n]$, $i = 0, 1, \ldots, n$. Note that D_i is also a field, since D is a field (see §V.1). View a_k as a polynomial in the variables R_0, with coefficients in D_1, *i.e.*, $a_k \in D_1[R_0]$. But a_k is symmetric in R_0 and so $a_k \in D_1$ by theorem 22 above. Next we view a_k as a polynomial in $D_2[R_1]$. Again, a_k is symmetric in R_1 and so $a_k \in D_2$. Repeating this argument, we finally obtain $a_k \in D$. **Q.E.D.**

We have shown that the set of algebraic numbers forms a field extension of $\mathbb{Z}$. By the preceding, this set is algebraically closed. Clearly, it is the smallest such extension of $\mathbb{Z}$. This proves:

Theorem 24 *The algebraic closure $\overline{\mathbb{Z}}$ of $\mathbb{Z}$ is the set of algebraic numbers.*

Exercises

Exercise 5.1: Show the uniqueness of σ-basic decompositions. □

Exercise 5.2: Let $s_k(X_1, \ldots, X_n) = \sum_{i=1}^n X_i^k$ for $k = 1, \ldots, n$. Express s_k in terms of $\sigma_1, \ldots, \sigma_k$. Conversely, express σ_k in terms of $s_1, \ldots, s_k$. □

Exercise 5.3: The *principle of induction* for well-ordered sets S is this: Suppose a statement $P(x)$ is true for all minimal elements x of S, and for any $y \in S$, whenever $P(x)$ is true for all $x < y$, then $P(y)$ also holds. Then $P(x)$ is true for all $x \in S$. □

Exercise 5.4: Generalize the above proof that G_e is symmetric: fix any subgroup $\Delta \leq S_n$ and we need not assume that the e_i's are non-decreasing. As above, define G_e to be the sum over the distinct terms in the multiset $\{\mathbf{X}_\pi^e : \pi \in \Delta\}$. Let $Aut_\Delta(e) = \{\rho \in \Delta : \mathbf{X}_\rho^e = \mathbf{X}^e\}$ be the group of Δ-automorphisms of $\mathbf{X}^e$.
(a) Show that there is a constant c_e such that $\sum_{\pi\in\Delta} \mathbf{X}_\pi^e$ is equal to $c_e G_e$.
(b) The number of terms in G_e is equal to the number of cosets of $Aut_\Delta(\mathbf{X}^e)$ in Δ. □

Exercise 5.5: Fix $e = (e_1, \ldots, e_n)$, $e_i \geq 0$, and subgroup $\Delta \leq S_n$, as in the last exercise. Let $U = \{\mathbf{X}_\pi^e : \pi \in S_n\}$. For $t \in U$, the Δ-*orbit* of t is $t_\Delta = \{t_\pi : \pi \in \Delta\}$. Let $\Lambda \leq \Delta$ be a subgroup. Let Σ be the *normalizer* of Λ in Δ, defined as $\Sigma = \{\pi \in \Delta : \pi^{-1}\Lambda\pi = \Lambda\}$. (Note: in the permutation, $\pi^{-1}\rho\pi$ we first apply π^{-1}, then ρ, finally π.) Define (cf. [148, 149]):

$$\begin{aligned} N_\Delta(\Lambda) &:= \{t \in U : Aut_\Delta(t) = \Lambda\}, \\ \widehat{N}_\Delta(\Lambda) &:= \{Q : Q \text{ is an } \Delta\text{-orbit }, Q \cap N_\Delta(\Lambda) \neq \emptyset\}. \end{aligned}$$

Let $t \in N_\Delta(\Lambda)$.
(a) Show that $Aut_\Delta(t_\pi) = \pi^{-1}\Lambda\pi$.
(b) $t_\Delta \cap N_\Delta(\Lambda) = t_\Sigma$.
(c) $|t_\Sigma|$ equals the number of cosets of Λ. □

Exercise 5.6: Suppose $A(X_1, \ldots, X_n)$ is a symmetric polynomial with integer coefficients.
(a) If we write A as a polynomial $P(\sigma_1, \ldots, \sigma_n)$ in the elementary symmetric functions, what is a bound on the $\|P\|_\infty$ as a function of $\|A\|_\infty$, n and $\deg A$?
(b) Give an algorithm to convert a symmetric polynomial $A(X_1, \ldots, X_n)$ (given as a sum of monomials) into a polynomial $B(\sigma_1, \ldots, \sigma_n)$ in the elementary symmetric functions σ_i. Analyze its complexity. In particular, bound $\|B\|_\infty$ in terms of $\|A\|_\infty$.
(c) We want to make the proof of theorem 23 a constructive result: suppose that coefficients of the minimal polynomials of the β_i's are given. Give an algorithm to construct the polynomial $A(X)$. What is its complexity? □

Exercise 5.7: A polynomial $A = A(X_1, \ldots, X_n)$ is **alternating** if for all transpositions of a pair of variables, $X_i \leftrightarrow X_j$ $(i \neq j)$, the sign of A changes. Show that A can be expressed (up to sign) as

$$A = B \prod_{i<j} (X_i - X_j)$$

for some symmetric polynomial B. □

§6. Discriminant

For any non-constant polynomial $A(X) \in \mathbb{C}[X]$, we define its *minimum root separation* to be

$$\mathtt{sep}(A) := \min_{1\leq i<j\leq k} |\alpha_i - \alpha_j|$$

where the distinct roots of $A(X)$ are $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$. Clearly $k \geq 1$ and if $k = 1$, then $\mathtt{sep}(A) = \infty$ by definition. In order to get a bound on $\mathtt{sep}(A)$, we introduce a classical tool in the study of polynomials.

Let D be any domain. The *discriminant* of $A \in D[X]$ is

$$\mathtt{disc}(A) = a^{2m-2} \prod_{1\leq i<j\leq m} (\alpha_i - \alpha_j)^2$$

where $\alpha_1, \ldots, \alpha_m \in \overline{D}$ are the roots (not necessarily distinct) of A, $\deg(A) = m \geq 2$, $a = \texttt{lead}(A)$. If $A(X) = aX^2 + bx + c$, then its discriminant is the familiar $\texttt{disc}(A) = b^2 - 4ac$. It is clear from this definition that $\texttt{disc}(A) = 0$ iff A has repeated roots. To see that $\texttt{disc}(A) \in D$, note that the function $\prod_{i \leq i < j \leq m}(\alpha_i - \alpha_j)^2$ is a symmetric function of the roots of $A(X)$. Since this function has maximum degree $2(m-1)$, theorem 22 implies that our expression for $\texttt{disc}(A)$ is an element of D. But this gives no indication on how to compute it. The following result gives the remedy. With A' denoting the derivative of A, we have:

Lemma 25 $a \cdot \texttt{disc}(A) = (-1)^{\binom{m}{2}} \texttt{res}(A, A')$.

Proof.

$$\begin{aligned}
\texttt{res}(A, A') &= a^{m-1} \prod_{i=1}^{m} A'(\alpha_i) \quad \text{(by theorem 15)} \\
&= a^{m-1} \prod_{i=1}^{m} \left(a \prod_{j=1}^{m} (X - \alpha_j) \right)' \Bigg|_{X=\alpha_i} \\
&= a^{2m-1} \prod_{i=1}^{m} \left(\sum_{k=1}^{m} \prod_{\substack{j=1 \\ j \neq k}}^{m} (X - \alpha_j) \right) \Bigg|_{X=\alpha_i} \\
&= a^{2m-1} \prod_{i=1}^{m} \left(\prod_{\substack{j=1 \\ j \neq i}}^{m} (\alpha_i - \alpha_j) \right) \\
&= a^{2m-1} \prod_{1 \leq i < j \leq m} (-1)(\alpha_i - \alpha_j)^2 \\
&= (-1)^{\binom{m}{2}} a \cdot \texttt{disc}(A).
\end{aligned}$$

Q.E.D.

The following matrix

$$V_m = V_m(\alpha_1, \alpha_2, \ldots, \alpha_m) := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_m \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_m^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^{m-1} & \alpha_2^{m-1} & \cdots & \alpha_m^{m-1} \end{bmatrix}$$

is called a *Vandermonde matrix*, and its determinant is a *Vandermonde determinant*.

Lemma 26

$$\prod_{1 \leq i < j \leq m} (\alpha_i - \alpha_j) = (-1)^{\binom{m}{2}} \det V_m(\alpha_1, \alpha_2, \ldots, \alpha_m).$$

Proof. One can evaluate $\det V_m$ recursively as follows. View $\det V_m$ as a polynomial $P_m(\alpha_m)$ in the variable α_m. The degree of α_m in P_m is $m-1$, as seen by expanding the determinant by the last

column. If we replace α_m in $P_m(\alpha_m)$ by α_i $(i = 1, \ldots, m-1)$, we get the value $P_m(\alpha_i) = 0$. Hence α_i is a root and by basic properties of polynomials (§2),

$$P_m(\alpha_m) = U \cdot (\alpha_m - \alpha_1)(\alpha_m - \alpha_2) \cdots (\alpha_m - \alpha_{m-1})$$

where U is the coefficient of α_m^{m-1} in P_m. But U is another Vandermonde determinant P_{m-1}:

$$P_{m-1} = \det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_{m-1} \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_{m-1}^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^{m-2} & \alpha_2^{m-2} & \cdots & \alpha_{m-1}^{m-2} \end{bmatrix} = \det V_{m-1}(\alpha_1, \ldots, \alpha_{m-1}).$$

Inductively, let

$$(-1)^{\binom{m-1}{2}} P_{m-1} = \prod_{1 \le i < j \le m-1} (\alpha_i - \alpha_j).$$

Hence

$$\begin{aligned} P_m &= U \cdot \prod_{i=1}^{m-1} (\alpha_m - \alpha_i) \\ &= (-1)^{\binom{m-1}{2}} P_{m-1} \cdot (-1)^{m-1} \prod_{i=1}^{m-1} (\alpha_i - \alpha_m) \\ &= (-1)^{\binom{m}{2}} \prod_{1 \le i < j \le m-1} (\alpha_i - \alpha_j) \cdot \prod_{i=1}^{m-1} (\alpha_i - \alpha_m) \\ &= (-1)^{\binom{m}{2}} \prod_{1 \le i < j \le m} (\alpha_i - \alpha_j). \end{aligned}$$

Q.E.D.

For a monic polynomial A, $\sqrt{\texttt{disc}(A)}$ is equal to a Vandermonde determinant, up to sign. Note that $\sqrt{\texttt{disc}(A)}$ is however not a symmetric function (although, up to sign, it is symmetric).

Exercises

Exercise 6.1: The discriminant of $A(X) = \sum_{i=0}^{3} a_i X^i$ is $a_2^2 a_1^2 + 18 a_3 a_2 a_1 a_0 - 4 a_3 a_1^3 - 4 a_2^3 a_0 - 27 a_3^2 a_0^2$. □

Exercise 6.2: $\texttt{disc}(AB) = \texttt{disc}(A)\texttt{disc}(B)(\texttt{res}(A, B))^2$. □

Exercise 6.3: Show that

$$\texttt{disc}(A) = \det \begin{bmatrix} s_0 & s_1 & \cdots & s_{m-1} \\ s_1 & s_2 & \cdots & s_m \\ \vdots & & & \\ s_{m-1} & s_m & \cdots & s_{2m-2} \end{bmatrix}$$

where $s_i = \sum_{j=1}^{m} \alpha_j^i \quad (i = 0, \ldots, 2m-2)$. □

Exercise 6.4: Let $A(X)$ be a monic polynomial of degree d. Show that the sign of its discriminant is $(-1)^{(d^2-r)/2}$ where r is the number of its real roots. □

§7. Root Separation

We now prove a root separation bound. But first we quote Hadamard's bound whose proof is delayed to a later lecture (§VIII.2).

Lemma 27 (Hadamard's determinantal inequality) *Let $M \in \mathbb{C}^{n\times n}$. Then $|\det(M)| \leq \prod_{i=1}^{n} \|R_i\|_2$ where R_i is the ith row of M. Equality holds iff $\langle R_i, \overline{R_j}\rangle = 0$ for all $i \neq j$. Here $\overline{R_j}$ denotes the complex conjugate of each component in R_j and $\langle , \rangle$ is scalar product.*

Recall (§IV.5) that the measure $M(A)$ of a complex polynomial A is equal to the product $|\mathtt{lead}(A)|\cdot \prod_i |\alpha_i|$ where i ranges over all complex roots α_i of A with absolute value $|\alpha_i| \geq 1$. (If i ranges over an empty set, then $M(A) = |\mathtt{lead}(A)|$.)

Theorem 28 (Davenport-Mahler) *Assume $A(X) \in \mathbb{C}[X]$ has roots $\alpha_1, \ldots \alpha_m \in \mathbb{C}$. For any $k+1$ of these roots, say $\alpha_1, \ldots, \alpha_{k+1}$ $(k = 1, \ldots, m-1)$, we reorder them so that*

$$|\alpha_1| \geq |\alpha_2| \geq \cdots \geq |\alpha_{k+1}|.$$

Then

$$\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}| > \sqrt{|\mathtt{disc}(A)|} \cdot M(A)^{-m+1} \cdot m^{-m/2} \cdot \left(\frac{\sqrt{3}}{m}\right)^k.$$

Proof. First let us assume A is monic. Let us give an upper bound on $\sqrt{|\mathtt{disc}(A)|}$ which by the previous lemma is, up to sign, given by the Vandermonde determinant

$$\det V_m = \det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_m \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_m^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^{m-1} & \alpha_2^{m-1} & \cdots & \alpha_m^{m-1} \end{bmatrix}.$$

We modify Vandermonde's matrix by subtracting the 2nd from the 1st column, the 3rd column from the 2nd, and so on, and finally the $(k+1)$st column from the kth column. Hence the first k column becomes (for $i = 1, \ldots, k$):

$$(\alpha_i - \alpha_{i+1}) \begin{bmatrix} 0 \\ 1 \\ \beta_i^{(2)} \\ \beta_i^{(3)} \\ \vdots \\ \beta_i^{(m-1)} \end{bmatrix}$$

where $\beta_i^{(j)} = \frac{\alpha_i^j - \alpha_{i+1}^j}{\alpha_i - \alpha_{i+1}} = \sum_{\ell=0}^{j-1} \alpha_i^\ell \alpha_{i+1}^{j-1-\ell}$. Hence

$$\det V_m = \prod_{i=1}^{k}(\alpha_i - \alpha_{i+1}) \cdot \det \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 & \alpha_{k+1} & \cdots & \alpha_m \\ \beta_1^{(2)} & \beta_2^{(2)} & \cdots & \beta_k^{(2)} & \alpha_{k+1}^2 & \cdots & \alpha_m^2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \beta_1^{(m-1)} & \beta_2^{(m-1)} & \cdots & \beta_k^{(m-1)} & \alpha_{k+1}^{m-1} & \cdots & \alpha_m^{m-1} \end{bmatrix}. \tag{19}$$

Let us upper bound the 2-norm of each column in this last matrix. There are two cases to consider:

Case 1: The column is $(0, 1, \beta_i^{(2)}, \ldots, \beta_i^{(m-1)})^T$. Each $\beta_i^{(j)}$ satisfies

$$\begin{aligned} |\beta_i^{(j)}|^2 &\leq \left(\sum_{\ell=0}^{j-1} |\alpha_i|^\ell |\alpha_{i+1}|^{j-1-\ell}\right)^2 \\ &\leq \left(j|\alpha_i|^{j-1}\right)^2 \\ &\leq j^2(\max\{1, |\alpha_i|\})^{2(m-1)}. \end{aligned}$$

So the 2-norm of the column is

$$\left(\sum_{j=1}^{m-1} |\beta_i^{(j)}|^2\right)^{\frac{1}{2}} < \sqrt{\frac{m^3}{3}} \cdot (\max\{1, |\alpha_i|\})^{m-1},$$

using the fact that $\sum_{j=1}^{m} j^2 = \frac{m^3}{3} + \frac{m^2}{2} + \frac{m}{6} < \frac{(m+1)^3}{3}$.

Case 2: The column is $(1, \alpha_i, \alpha_i^2, \ldots, \alpha_i^{m-1})^T$. Its 2-norm is

$$\left(\sum_{\ell=0}^{m-1} |\alpha_i|^{2\ell}\right)^{1/2} < \sqrt{m} \cdot (\max\{1, |\alpha_i|\})^{m-1}.$$

The product of the 2-norms of all the m columns is therefore less than

$$\left(\sqrt{\frac{m^3}{3}}\right)^k (\sqrt{m})^{m-k} \prod_{i=1}^{m} \max\{1, |\alpha_i|\}^{m-1} = \left(\frac{m}{\sqrt{3}}\right)^k \cdot m^{m/2} \cdot (M(A))^{m-1}$$

where $M(A)$ is the measure of A.

By Hadamard's inequality, this product is an upper bound on the determinant in (19). Hence

$$\sqrt{|\mathtt{disc}(A)|} = |\det V_m| < \left(\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|\right) \cdot \left(\frac{m}{\sqrt{3}}\right)^k m^{m/2} (M(A))^{m-1}.$$

This proves the theorem for monic A. It remains to remove the assumption that A is monic. Suppose $\mathtt{lead}(A) = a \neq 1$. Then clearly we have proved that

$$\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}| > \sqrt{|\mathtt{disc}(A/a)|} \cdot M(A/a)^{-m+1} \cdot m^{-m/2} \cdot \left(\frac{\sqrt{3}}{m}\right)^k.$$

But $\texttt{disc}(A/a) = \frac{\texttt{disc}(A)}{a^{2m-2}}$ and $M(A/a)^{-m+1} = a^{m-1}M(A)^{-m+1}$. Hence the extraneous factors involving a cancel out, as desired.

Q.E.D.

The preceding proof for $k = 1$ is from Mahler (1964), generalized here to $k > 1$ by Davenport (1985). Since $M(A) \leq \|A\|_2$ (§IV.5), we obtain:

Corollary 29
(i) $\texttt{sep}(A) > \sqrt{3|\texttt{disc}(A)|} \cdot \|A\|_2^{-m+1} \cdot m^{-(m+2)/2}$ *where* $m = \deg A$.
(ii) $|\texttt{disc}(A)| \leq m^m (M(A))^{2m-2} \leq m^m \|A\|_2^{2m-2}$.

Proof. Part (i) comes from the theorem with $k = 1$. Part (ii) is a corollary of the proof of the main theorem (essentially with $k = 0$). **Q.E.D.**

For integer polynomials, let us express part (i) in simpler, if cruder, terms. The bit-size of an integer polynomial (§0.8) is simply the sum of the bit-sizes of its coefficients in the dense representation.

Lemma 30 *If* $A \in \mathbb{Z}[X]$ *is square-free of degree* m *and has bit-size* $s \geq 4$ *then* $\texttt{sep}(A) > \|A\|_2^{-m+1} m^{-(m+2)/2} \geq 2^{-2s^2}$.

Proof. Note that $\sqrt{|\texttt{disc}(A)|} \geq 1$, $\|A\|_2 \leq 2^s$, and $m \leq s$. **Q.E.D.**

In other words, with $O(s^2)$ bits of accuracy we can separate the roots of A. Instead of the trivial bound $|\texttt{disc}(A)| \geq 1$ above, Siegel [191, p. 27] shows that a situation where this can be improved: if $A(X) \in \mathbb{Z}[X]$ is irreducible and monic with only real zeros then

$$\texttt{disc}(A) \geq \left(\frac{m^m}{m!}\right)^2, \qquad m = \deg A.$$

Remark: Our root separation bound is useless when A has multiple roots, since the discriminant is then zero. Of course, we can still obtain a bound indirectly, by computing the root separation of the square-free part $A^* := A/\texttt{GCD}(A, A')$ of A, as follows. Since $A^*|A$, we have $\|A^*\|_1 \leq 2^m \|A\|_2$ (§IV.5), assuming integer coefficients. Then

$$\texttt{sep}(A) = \texttt{sep}(A^*) > (2^m \|A\|_2)^{-m} m^{-(m+2)/2}. \tag{20}$$

This bound is inferior to what can be obtained by direct arguments. Rump [173] (as rectified by Schwartz [187]) has shown

$$\texttt{sep}(A) > \left(2 \cdot m^{(m/2)+2} (\|A\|_\infty + 1)^m\right)^{-1}. \tag{21}$$

Exercises

Exercise 7.1: [Mahler]
Show that the root separation bound in corollary 29(i) is tight up to some constant. HINT: The polynomial $A(X) = X^m - 1$ has $|\mathtt{disc}(A)| = m^m$ and $\mathtt{sep}(A) = 2\sin(\pi/m)$. □

Exercise 7.2: (Sellen-Yap) Let $\varepsilon = a + \sqrt{b} - \sqrt{c}$ where a, b, c are all L-bit integers.
(i) Show that $\log_2(1/|\varepsilon|) = 3L/2 + O(1)$. HINT: a is at most $1 + (L/2)$ bits long.
(ii) Show that this is the best possible by an infinite family of examples with $L \to \infty$. HINT: $a = 2^{L/2-1}$, $b = 2^{L-2}-1$ and $c = 2^{L-1}-2$. These numbers are the best possible for $L = 6, 8, 10$, as verified by exhaustive computation. □

Exercise 7.3: (Mignotte) Consider the following polynomial $A(X) = X^n - 2(aX - 1)^2$ where $n \geq 3, a \geq 3$ are integers.
i) Show that $A(X)$ is irreducible (using Eisenstein's criterion).
ii) Show that $A(X)$ has two real roots close to $1/a$ and their separation is at most $2a^{-(n+2)/2}$.
iii) Compute bounds for the absolute value of the roots and root separation, using the above formulas. □

§8. A Generalized Hadamard Bound

Let $A(X), B(X) \in \mathbb{Z}[X]$ where $A(X)B(X)$ has only simple roots, and $n = \max\{\deg A, \deg B\}$. We want a lower bound on $|\alpha - \beta|$ where α, β are roots of (respectively) $A(X), B(X)$. Using the fact that $|\alpha - \beta| \geq \mathtt{sep}(AB)$, we derive a bound:

$$\begin{aligned} |\alpha - \beta| &\geq \mathtt{sep}(A(X)B(X)) \\ &\geq \sqrt{3 \cdot \mathtt{disc}(A(X)B(X))} \cdot \|AB\|_2^{-(2n-1)} (2n)^{-(2n+2)/2} \\ &\geq (\|A\|_2 \|B\|_2 (1+n))^{(-2n+1)} (2n)^{-n-1}, \end{aligned}$$

using the fact $\|AB\|_2 \leq \|A\|_\infty \|B\|_\infty (1+n)$. This section gives a slightly sharper bound, based on a generalization of the Hadamard bound [72]. The proof further applies to complex polynomials $A, B \in \mathbb{C}[X]$ that need not be square-free. Let $W = [w_{ij}]_{i,j} \in \mathbb{C}^{n \times n}$. Define

$$H(W) := (\prod_{i=1}^{n} \sum_{j=1}^{n} |w_{ij}|^2)^{\frac{1}{2}}.$$

Then the Hadamard's determinantal bound (Lemma 27) gives $|\det(W)| \leq H(W)$. The following generalizes this.

Theorem 31 (Goldstein-Graham) *Let $M(X) = (M_{ij}(X))$ be an n-square matrix whose entries $M_{ij}(X)$ are polynomials in $\mathbb{C}[X]$. Let $W = [w_{ij}]_{i,j}$ be the matrix where $w_{ij} = \|M_{ij}(X)\|_1$. Then* $\det(M(X)) \in \mathbb{C}[X]$ *satisfies*

$$\|\det(M(X))\|_2 \leq H(W).$$

Proof. For any real t,

$$|M_{ij}(e^{\mathbf{i}\,t})| \leq w_{ij}$$

where $\mathbf{i} = \sqrt{-1}$. Hence Hadamard's inequality implies

$$\begin{aligned} |\det(M(e^{\mathbf{i}\,t}))|^2 &\leq \prod_{k=1}^{n}\sum_{\ell=1}^{n}|M_{k\ell}(e^{\mathbf{i}\,t})|^2 \\ &\leq \prod_{k=1}^{n}\sum_{\ell=1}^{n} w_{k\ell}^2 \\ &= (H(W))^2. \end{aligned}$$

But, if the polynomial $\det(M(X))$ is $a_0 + a_1X + a_2X^2 + \cdots$, then

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} |\det(M(e^{\mathbf{i}\,t}))|^2 dt &= \frac{1}{2\pi}\int_0^{2\pi} ((\sum_k a_k e^{\mathbf{i}\,kt})(\sum_\ell \overline{a}_\ell \overline{e}^{\mathbf{i}\,\ell t}))dt \\ &= \sum_k |a_k|^2 \end{aligned}$$

since $\frac{1}{2\pi}\int_0^{2\pi} e^{-\mathbf{i}\,kt}dt = \delta_{k,0}$ (Kronecker's delta). (This is also known as Parseval's identity.) Hence

$$\begin{aligned} \|\det(M(X))\|_2^2 &= \sum_k |a_k|^2 \qquad \text{(by definition)} \\ &= \frac{1}{2\pi}\int_0^{2\pi} |\det(M(e^{\mathbf{i}\,t}))|^2 dt \\ &\leq \frac{1}{2\pi}\int_0^{2\pi} (H(W))^2 dt \\ &= (H(W))^2. \end{aligned}$$

Q.E.D.

Applications. (I) Consider our original problem of bounding the minimum separation between distinct roots of $A(X)$ and $B(X)$ in $\mathbb{Z}[X]$, where $A(X)B(X)$ need not be square-free. Let

$$C(X) := \mathtt{res}_Y(A(Y), B(X+Y))$$

where Y is a new variable. Note that $\alpha - \beta$ is a root of $C(X)$. Assume $m = \deg A$ and $n = \deg B$. Writing $B(X) = \sum_{i=0}^n b_i X^i$, we have

$$\begin{aligned} B(X+Y) &= \sum_{i=0}^{n} b_i (X+Y)^i \\ &= \sum_{i=0}^{n} b_i \sum_{j=0}^{i} \binom{i}{j} X^{i-j}Y^j \\ &= \sum_{j=0}^{n} Y^j \left(\sum_{i=j}^{n} b_i \binom{i}{j} X^{i-j} \right). \end{aligned}$$

Let $S(X)$ be the Sylvester matrix corresponding to $\mathtt{res}_Y(A(Y), B(X+Y))$. Consider a row of $S(X)$ corresponding to $B(X+Y)$: each non-zero entry is a polynomial of the form

$$B_j(X) := \sum_{i=j}^{n} b_i \binom{i}{j} X^{i-j}.$$

Its 1-norm is bounded as $\|B_j(X)\|_1 \leq \|B\|_\infty \sum_{i=j}^n \binom{i}{j}$. Thus the 2-norm of such a row is at most

$$\begin{aligned}
\|B\|_\infty \left(\sum_{j=0}^n \left(\sum_{i=j}^n \binom{i}{j} \right)^2 \right)^{\frac{1}{2}} &\leq \|B\|_\infty \sum_{j=0}^n \sum_{i=j}^n \binom{i}{j} \\
&= \|B\|_\infty \sum_{i=0}^n \sum_{j=0}^i \binom{i}{j} \\
&= \|B\|_\infty \sum_{i=0}^n 2^i \\
&= \|B\|_\infty (2^{n+1} - 1).
\end{aligned}$$

Since there are m such rows, the product of all these 2-norms is at most

$$\|B\|_\infty^m 2^{m(n+1)}.$$

The remaining rows of $S(X)$ have as non-zero entries the coefficients of $A(X)$. Their 2-norms are clearly $\|A\|_2$. Again, there are n such rows, so their product is $\|A\|_2^n$. The generalized Hadamard bound yields

$$\begin{aligned}
\|C(X)\|_2 &\leq \|B\|_\infty^m 2^{m(n+1)} \cdot \|A\|_2^n \\
&\leq (2^{n+1}\|B\|_2)^m \|A\|_2^n. \qquad (22)
\end{aligned}$$

Of course, if $n < m$, we could interchange the roles of m and n in this bound. Applying Landau's bound (3), we conclude:

Lemma 32 *If $\alpha \neq \beta$ then*

$$|\alpha - \beta| > \frac{1}{\|C\|_2} \geq \frac{1}{2^{nm+\min\{m,n\}} \|B\|_2^m \|A\|_2^n}.$$

If s is the sum of the bit sizes of $A(X)$ and $B(X)$ then

$$|\alpha - \beta| > \frac{1}{(2^s \cdot 2^s)^{s-1} 2^{s-1}} \geq 2^{-2s^2}.$$

Letting $A = B$, we further obtain (cf. equation (20)):

Corollary 33 *If A is an integer polynomial, not necessarily square-free,*

$$\mathtt{sep}(A) > (2^{m+1}\|A\|_2^2)^{-m}.$$

(II) The next application is useful when computing in a number field $\mathbb{Q}(\alpha)$ (cf. [173]):

Lemma 34 *Let $A, B \in \mathbb{Z}[X]$, $\deg A = m > 0$ and $\deg B = n > 0$. For any root α of A, if $B(\alpha) \neq 0$ then*

$$|B(\alpha)| > \frac{1}{\|\widehat{B}\|_2^m \cdot \|A\|_2^n + 1}$$

where $\widehat{B(X)}$ is the same polynomial as $B(X)$ except that its constant term b_0 has been replaced by $1 + |b_0|$.

Proof. Let Y be a new variable and consider the resultant of $A(Y)$ and $X - B(Y)$ with respect to Y:

$$C(X) = \mathtt{res}_Y(A(Y), X - B(Y)) = a_m^n \prod_{i=1}^{m}(X - B(\alpha_i))$$

where α_i's are the roots of A and $a_m = \mathtt{lead}(A)$. From the determinantal bound of Goldstein and Graham,

$$\|C(X)\|_2 \leq \|A\|_2^n \cdot \|\widehat{B}\|_2^m.$$

Again applying Landau's bound, any root γ of $C(Y)$ satisfies

$$|\gamma| > \frac{1}{\|C\|_2}.$$

Since $B(\alpha)$ is such a root, the lemma follows. **Q.E.D.**

(III) Our final application arises in an implementation of a *real algebraic expression package* [57]. In particular, we are interested in real expressions E that are recursively built-up from the rational constants, using the operations of

$$+, -, \times, \div, \sqrt{}. \tag{23}$$

Thus E denotes a constructible real number (§V.4). With each expression E, the user can associate a *precision bound* of the form $[a, r]$ where $a, r \in \mathbb{Z}$. If the value of E is α, this means the system will find an approximate value $\widehat{\alpha}$ satisfying

$$|\alpha - \widehat{\alpha}| \leq \max\{|\alpha|2^{-r}, 2^{-a}\}.$$

Thus, $\widehat{\alpha}$ has "absolute precision a" or "relative precision r". Note that by changing the precision bound, we may force the approximate value to be recomputed. The most important case is $[a, r] = [\infty, 1]$, which guarantees one relative bit of α. This ensures that the sign of α is correctly determined. The system is the first one that could automatically determine the sign of arbitrary real constructible expression. To achieve this, we need an easily computable lower bound on $|\alpha|$ when $\alpha \neq 0$. There are several ways to do this, but we maintain with each node of the expression E an upper bound on the degree and length of the algebraic number represented at that node. If α is an algebraic number, we call the pair (d, ℓ) a *degree-length* bound on α if there exists a polynomial $A(X) \in \mathbb{Z}[X]$ such that $A(\alpha) = 0$, $\deg(A) \leq d$ and $\|A\|_2 \leq \ell$. Note that this implies that $|\alpha| \geq 1/\ell$ (Landau's bound) and so we only need to compute α to about $\lg \ell$ bits in absolute precision in order to determine its sign. We now derive the recursive rules for maintaining this bound.

Suppose the algebraic number β is obtained from α_1 and α_2 by one of the 5 operations in (23). Inductively, assume a degree-length bound of (d_i, ℓ_i) on α_i, $(i = 1, 2)$, and let $A_i(X)$ be a polynomial that achieves this bound. We will describe a polynomial $B(X)$ such that $B(\beta) = 0$, and a corresponding degree-length bound (d, ℓ) on β.

- (BASIS) $\beta = p/q$ is a rational number, where $p, q \in \mathbb{Z}$. Choose $B(X) = qX - p$, $d = 1$ and $\ell = \sqrt{p^2 + q^2}$.
- (INVERSE) $\beta = 1/\alpha_1$: choose $B(X) = X^{d_1}A_1(1/X)$, $d = d_1$ and $\ell = \ell_1$.
- (SQUARE-ROOT) $\beta = \sqrt{\alpha_1}$: choose $B(X) = A_1(X^2)$, $d = 2d_1$ and $\ell = \ell_1$.
- (PRODUCT) $\beta = \alpha_1\alpha_2$: choose $B(X) = \mathtt{res}_Y(A_1(Y), Y^{d_2}A_2(X/Y))$, $d = d_1d_2$ and

$$\ell = \ell_1^{d_2}\ell_2^{d_1}.$$

- (SUM/DIFFERENCE) $\beta = \alpha_2 \pm \alpha_1$: choose $B(X) = \mathtt{res}_Y(A_1(Y), A_2(X \mp Y))$, $d = d_1 d_2$ and

$$\ell = \ell_1^{d_2} \ell_2^{d_1} 2^{d_1 d_2 + \min\{d_1, d_2\}}.$$

The BASIS, INVERSE and SQUARE-ROOT cases are obvious. The choices of $B(X)$ and d the remaining cases are justified by the theory of resultants (§4). It remains to justify the choices of ℓ. For PRODUCT, the choice ℓ is an easy application of the generalized Hadamard bound. In case of SUM/DIFFERENCE, the choice ℓ is derived in application (I). Finally, these bounds can easily be extended to the class of general algebraic expressions (see Exercise 3).

Exercises

Exercise 8.1: Let $A = a_0 + a_1 X$ and $B = b_0 + b_1 X$. Then $\|AB\|_2 \leq \|A\|_2 \|B\|_2$ iff $a_0 a_1 b_0 b_1 \leq 0$. □

Exercise 8.2: Suppose that we wish to maintain a "degree-height" bound (d, h) instead of the degree-length bound (d, ℓ). Recall that the height of a polynomial $A(X)$ is $\|A(X)\|_\infty$. Derive the corresponding recursive rules for maintaining such bounds. □

Exercise 8.3: Extend the class of expressions in Application (III) as follows. We treat only real algebraic numbers (extension to the complex case is similar). If $E_0, E_1, \ldots, E_n$ are expressions and i is a number between 1 and n then $\mathrm{POLY}(E_0, \ldots, E_n, i)$ is a new expression that denotes the ith largest real root of the polynomial $P(X) = \sum_{j=0}^{n} \alpha_j X^j$, where α_j is the value of E_j. This POLY expression is considered ill-formed if $P(X)$ has less than j real roots. Show how to maintain the (degree-length) bound for such expressions. □

§9. Isolating Intervals

The existence of a root separation bound justifies the following representation of real algebraic numbers.

(i) Let $I = [a, b]$ be an interval, $(a, b \in \mathbb{R}, a \leq b)$. For any polynomial $A(X) \in \mathbb{R}[X]$, we call I an *isolating interval* of A if I contains exactly one distinct real root α of A. The *width* of I is $b - a$.

(ii) Let $\alpha \in \overline{\mathbb{Z}} \cap \mathbb{R}$ be a real algebraic number. An *isolating interval representation* of α is a pair

$$(A(X), I)$$

where $A \in \mathbb{Z}[X]$ is square-free and primitive, $A(\alpha) = 0$, and I is an isolating interval of A that contains α and has rational endpoints: $I = [a, b], a < \alpha < b$, $(a, b \in \mathbb{Q})$. As a special case, we allow $a = b = \alpha$. We write

$$\alpha \cong (A, I)$$

to denote this relationship.

This isolating interval representation motivates the *root isolation problem*: given a real polynomial $P(X)$, determine an isolating interval for each real root of $P(X)$. This problem is easily solved in principle: we know a lower bound L and an upper bound U on all real roots of $P(X)$. We partition the interval $[L, U]$ into subintervals of width at most 4^{-s^2} where s is the bit-size of $P(X)$. We evaluate $P(X)$ at the end points of these subintervals. By our root separation bounds, such a subinterval is isolating for $P(X)$ iff $P(X)$ does not vanish at both end points but has opposite signs there. Of course, this procedure is grossly inefficient. The next lecture uses Sturm sequences to perform more this efficiently.

Clearly, an isolating interval representation is far from unique. We do not insist that $A(X)$ be the minimal polynomial of α because this is too expensive computationally. We also note that the rational endpoints of I are usually *binary rationals i.e.*, they have finite binary expansions. Note that once A is fixed, then minimum root separation tells us I need not have more than $O(s^2)$ bits to isolate any root of interest (s is the bit-size A). The interval I serves to distinguish the root of $A(X)$ from the others. This is not the main function of I, however – otherwise we could as well represent α as $(A(X), i)$ if α is the ith smallest real root of $A(X)$. The advantage of isolating interval is that it facilitates numerical computations. In the following, let α, β be two real algebraic numbers represented in this way:

$$\alpha \cong (A(X), I),\ \beta \cong (B(X), J)$$

(A) We can compute $\alpha \cong (A, I)$ to any desired degree of accuracy using repeated bisections of I: if $I = [a, b]$ is not degenerate then clearly $A(a) \cdot A(b) < 0$. We begin by evaluating $A(\frac{a+b}{2})$. If $A(\frac{a+b}{2}) = 0$ then we have found α exactly. Otherwise, either $[a, \frac{a+b}{2}]$ or $[\frac{a+b}{2}\, b]$ contains α. It is easy to determine exactly which half-interval contains α: $a < \alpha < \frac{a+b}{2}$ iff $A(a)A((a+b)/2) < 0$. Note that if a, b are binary rationals, then $\frac{a+b}{2}$ is still a binary rational.

(B) We can compare α and β to see which is bigger: this comparison is immediate if $I \cap J = \emptyset$. Otherwise, we could repeatedly refine I and J using bisections until they are disjoint. But what if $\alpha = \beta$? In other words, when do we stop bisecting in case $\alpha = \beta$? If s is a bound on the sum of the bit sizes of A and B, then the previous section says we can stop when I and J have widths $\leq 4^{-s^2}$, concluding $\alpha = \beta$ iff $I \cap J \neq \emptyset$.

(C) We can perform any of the four arithmetic operations on α and β. We just illustrate the case of multiplication. We can (§4) compute a polynomial $C(X)$ that contains the product $\alpha\beta$ as root. It remains to compute an isolating interval K of $C(X)$ for $\alpha\beta$. Can we choose $K = I \times J = \{xy : x \in I, y \in J\}$? The answer is yes, provided the width of K is smaller than the root-separation bound for $C(X)$. For instance, suppose both $I = [a, a']$ and $J = [b, b']$ have width at most some $w > 0$. Then the width of K is $W := a'b' - ab$ (assuming $a > 0, b > 0$). But $W = a'b' - ab \leq (a + b + w)w$, so it is easy to determine a value w small enough so that W is less than the root separation bound for C. Then we just have to refine I and J until their widths are at most w. It is similarly easy to work out the other cases.

The above methods are simple to implement.

Exercises

Exercise 9.1: Show that a number is algebraic iff it is of the form $\alpha + \mathbf{i}\beta$ where α, β are real algebraic numbers. Hence any representation for real algebraic numbers implies a representation of all algebraic numbers. □

Exercise 9.2: Give complete algorithms for the four arithmetic operations on algebraic numbers, using the isolating interval representation. □

§10. On Newton's Method

Most books on Numerical Analysis inform us that when one has a "sufficiently good" initial approximation to a root, Newton's method rapidly converges to the root. Newton's method is much more efficient than the bisection method in the previous section for refining isolating intervals: in each iteration, the bisection method increases the root approximation by one bit while Newton's method doubles the number of bits. Hence in practice, we should begin by applying some bisection method until our isolating interval is "sufficiently good" in the above sense, whereupon we switch to Newton's method. In fact, we may then replace the isolating interval by any point within the interval.

In this section, we give an á priori bound on how close an initial approximation must be to be "sufficiently good". Throughout this section, we assume $f(X)$ is a real function whose zeros we want to approximate.

We view Newton's method as giving a suitable transformation of $f(X)$ into another function $F(X)$, such that a *fixed point* X^* of $F(X)$ is a root of $f(X)$:

$$X^* = F(X^*) \Rightarrow f(X^*) = 0.$$

As a simple example of a transformation of $f(X)$, we can let $F(X) = X - f(X)$. More generally, let $F(X) = X - g(X) \cdot f(X)$ for a suitable function $g(X)$. In the following, we assume the standard Newton method where

$$F(X) = X - \frac{f(X)}{f'(X)}. \tag{24}$$

The rest of the method amounts to finding a fixed point of $F(X)$ via an *iterative process*: begin with an *initial value* $X_0 \in \mathbb{R}$ and generate the sequence $X_1, X_2, \ldots$ where

$$X_{i+1} = F(X_i). \tag{25}$$

We say the iterative process *converges from* X_0 if the sequence of X_i's converges to some value X^*. Assuming that $F(X)$ is continuous at X^*, we may conclude that $F(X^*) = X^*$. Hence X^* is a root of $f(X)$.

To study the convergence of this iteration, let us assume that F is n-fold differentiable and the process converges to X^* starting at X_0. Using *Taylor's expansion of F at X^* with error term*,

$$\begin{aligned} F(X) &= F(X^*) + (X - X^*) \cdot F'(X^*) + (X - X^*)^2 \cdot \frac{F''(X^*)}{2!} + \cdots + \\ &\quad (X - X^*)^{n-1} \cdot \frac{F^{(n-1)}(X^*)}{(n-1)!} + (X - X^*)^n \cdot \frac{F^{(n)}(\xi)}{n!}, \end{aligned} \tag{26}$$

where $F^{(i)}$ denotes the i-fold differentiation of $F(X)$ and ξ denotes some value between X and X^*:

$$\xi = X + \theta(X^* - X), \qquad 0 \le \theta \le 1.$$

We say $F(X)$ gives an *n-th order iteration* at X^* if $F^{(i)}(X^*) = 0$ for $i = 1, \ldots, n-1$. Then, since $F(X^*) = X^*$, we have

$$F(X) - X^* = (X - X^*)^n \cdot \frac{F^{(n)}(\xi)}{n!}. \tag{27}$$

Let us suppose that for some real $k_0 > 0$,

$$\left|\frac{F^{(n)}(\xi)}{n!}\right| < k_0 \tag{28}$$

for all ξ where $|X^* - \xi| \leq 1/k_0$. Repeated application of equation (27) yields

$$\begin{aligned}
|X_1 - X^*| &< k_0|X_0 - X^*|^n, \\
|X_2 - X^*| &< k_0|X_1 - X^*|^n < k_0^{n+1}|X_0 - X^*|^{n^2}, \\
&\vdots \\
|X_i - X^*| &< \begin{cases} k_0^{\frac{n^i-1}{n-1}}|X_0 - X^*|^{n^i} & \text{if} \quad n > 1, \\ k_0^i|X_0 - X^*| & \text{if} \quad n = 1. \end{cases}
\end{aligned}$$

If $n = 1$ then convergence is assured if $k_0 < 1$. Let us assume $n > 1$. Then

$$|X_i - X^*| < \left(k_0^{\frac{1}{n-1}}|X_0 - X^*|\right)^{n^i} \cdot k_0^{\frac{1}{1-n}}$$

and a sufficient condition for convergence is

$$k_0^{\frac{1}{n-1}}|X_0 - X^*| < 1. \tag{29}$$

Remark: Newton's method works in very general settings. In particular, it applies when $f(X)$ is a complex function. But if $f(X)$ is a real polynomial and X^* is a complex root, it is clear that the initial value X_0 must be complex if convergence to X^* is to happen. More generally, Newton's method can be used to solve a system of equations, *i.e.*, when $f(X)$ is a vector-valued multivariate function, $f : \mathbb{C}^n \to \mathbb{C}^m$, $X = (X_1, \ldots, X_n)$.

Exercises

Exercise 10.1: Apply Newton's method to finding the square-root of an integer n. Illustrate it for $n = 9, 10$. □

Exercise 10.2: (Schroeppel, MIT AI Memo 239, 1972) Let $f(X) \in \mathbb{C}[X]$ be a quadratic polynomial with distinct roots α, β. Viewing $\mathbb{C}$ as the Euclidean plane, let L be the perpendicular bisector of the segment connecting α and β.
(a) Show that Newton's method converges to the closest root if the initial guess z_0 lies in $\mathbb{C} \setminus L$.
(b) If $z_0 \in L$, the method does not converge.
(c) There is a (relatively) dense set of points which involve division by zero.
(d) There is a dense set of points that loop, but all loops are unstable. □

Exercise 10.3: (Smale) Show that $f(X) = X^3/2 - X + 1$ has two neighborhoods centered about $X = 0$ and $X = 1$ such that Newton's method does not converge when initialized in these neighborhoods. What are the complex roots of f? □

§11. Guaranteed Convergence of Newton Iteration

Most books on Numerical Analysis inform us that when X_0 is sufficiently close to a root X^*, Newton iteration gives a second order rate of convergence. What we seek now is an explicit *á priori* upper bound on $|X_0 - X^*|$ which guarantees the said rate of convergence to X^*. Smale (e.g., [192]) describes implicit conditions for convergence. Following Smale, one may call such an X_0 is called an *approximate root* (and the set of approximate roots that converges to X^* the *Newton basin* of X^*). See also Friedman [68].

We now carry out the estimates for $F(X) = X - \frac{f(X)}{f'(X)}$ where $f(X) \in \mathbb{Z}[X]$ is square-free of degree m with real root X^*, and X_0 is a real satisfying $|X_0 - X^*| \leq 1$. Then

$$F'(X) = \frac{f_0 f_2}{f_1^2} \tag{30}$$

$$F''(X) = \frac{f_0 f_1 f_3 + f_1^2 f_2 - 2 f_0 f_2^2}{f_1^3} \tag{31}$$

where we write f_i for $f^{(i)}(X)$. For any root X^*, equation (30) shows that $F'(X^*) = 0$ since X^* is a simple root. Hence Newton's method gives rise to a second order iteration. Our goal is to find a real bound $\delta_0 > 0$ such that for any real number ξ satisfying $|X^* - \xi| \leq \delta_0$,

$$\left| \frac{F''(\xi)}{2!} \delta_0 \right| < 1. \tag{32}$$

Note that this implies (28) and (29) with the choice $k_0 = \frac{1}{\delta_0}$.

Lemma 35 *If $|\xi - X^*| \leq 1$ and $f(X^*) = 0$ then for all $i = 0, 1, \ldots, m$,*

$$|f^{(i)}(\xi)| \leq \frac{m!}{(m-i)!}(1 + M)^{1+m},$$

where $M = 1 + \|f\|_\infty$.

Proof. By Cauchy's bound (§2, lemma 7), $|X^*| < M$. Then

$$\begin{aligned} |f^{(i)}(\xi)| &\leq \frac{m!}{(m-i)!}\|f\|_\infty \sum_{j=0}^{m-i} |\xi|^j \\ &< \frac{m!}{(m-i)!}\|f\|_\infty \sum_{j=0}^{m} (1+M)^j \\ &< \frac{m!}{(m-i)!}(1+M)^{1+m}. \end{aligned}$$

Q.E.D.

Lemma 36 *Let $f(X) \in \mathbb{Z}[X]$ be square-free, and X^* a root of $f(X)$. If $m = \deg f$ then*

$$|f'(X^*)| \geq \frac{1}{m^{m-3/2}\|f\|_\infty^{m-2}}.$$

Proof. Let $g(X) = \frac{f(X)}{X - X^*}$. Then (by property C6. in §1) we have $f'(X^*) = g(X^*)$. We claim that

$$\mathtt{disc}(f) = \mathtt{disc}(g) \cdot f'(X^*)^2.$$

To see this, if $f(X) = a\prod_{i=1}^{m}(X-\alpha_i)$ then

$$\begin{aligned}\texttt{disc}(f) &= a^{2m-2}\prod_{1\le i<j\le m}(\alpha_i-\alpha_j)^2 \\ &= a^{2m-4}\prod_{1\le i<j<m}(\alpha_i-\alpha_j)^2\cdot\left[a\prod_{i=1}^{m}(\alpha_i-\alpha_m)\right]^2.\end{aligned}$$

Choosing $X^* = \alpha_m$ then $a\prod_{i=1}^{m}(\alpha_i-\alpha_m) = \pm g(X^*)$, which verifies our claim. By Corollary 29(ii) of Mahler (§7) we see that

$$|\texttt{disc}(g)| \le (m-1)^{m-1}M(g)^{2m-4}$$

where $M(g)$ is the measure of g. Also $M(g) \le M(f) \le \|f\|_2$. This implies

$$\begin{aligned}|\texttt{disc}(g)| &\le (m-1)^{m-1}\|f\|_2^{2m-4} \\ &< m^{m-1}\left(m^{\frac{1}{2}}\|f\|_\infty\right)^{2m-4} \\ &= m^{2m-3}\|f\|_\infty^{2m-4}.\end{aligned}$$

Hence $|f'(X^*)| \ge \sqrt{|\texttt{disc}(f)|}\cdot m^{-m+3/2}\|f\|_\infty^{-m+2}$ and the lemma follows. **Q.E.D.**

Let us now pick

$$\delta_0 := \frac{1}{m^{3m+9}(1+M)^{6m}}.$$

We first derive a lower bound on

$$|f'(\xi)|, \qquad \text{where } |X^*-\xi| \le \delta_0.$$

We have

$$f'(\xi) = f'(X^*) + (X^*-\xi)f''(\eta)$$

for some η between X^* and ξ, so $|X^*-\eta| \le \delta_0$. Using the preceding two lemmas,

$$\begin{aligned}|f'(\xi)| &\ge |f'(X^*)| - |X^*-\xi|\cdot|f''(\eta)| \\ &\ge \frac{1}{m^{m-3/2}(M-1)^{m-2}} - |\delta_0|\cdot m^2(1+M)^{1+m} \\ &\ge \frac{1}{m^{m-3/2}(M-1)^{m-2}} - \frac{m^2(1+M)^{1+m}}{m^{3m+9}(1+M)^{6m}} \\ &\ge \frac{1}{m^m(M-1)^{m-2}}\end{aligned}$$

From (31) we see that

$$|F''(\xi)| \le \frac{4\cdot K^3}{|f'(\xi)|^3}$$

where $K \ge \max_{i=0,\ldots,3}\{|f^{(i)}(\xi)|\}$. It suffices to choose $K = m^3(1+M)^{1+m}$ by lemma 35. Thus

$$\begin{aligned}|F''(\xi)| &\le 4(m^3(1+M)^{1+m})^3\cdot(m^m(M-1)^{m-2})^3 \\ &< 4m^{3m+9}(1+M)^{6m-3} \\ &< \delta_0^{-1}.\end{aligned}$$

Our goal, inequality (32) is thus achieved, and we have proved:

Theorem 37 *Let $f(X) \in \mathbb{Z}[X]$ be square-free with $m = \deg f$ and $M = 1 + \|f\|_\infty$. Then Newton iteration for $f(X)$ is guaranteed to converge to a root X^* provided the initial approximation is at most*

$$\delta_0 = (m^{3m+9}(1+M)^{6m})^{-1}$$

from X^*.

If s is the bit-size of $f(X)$ then $m \leq s$, $M \leq 2^s$ and δ_0 has about $6s^2$ bits of accuracy. In practice, it would be of interest to "dynamically" check when an approximate root is close enough for Newton iteration, since in practice one expects the effective δ_0 is much larger than the one guaranteed by this theorem. In Collin's computer algebra system SAC-II, such checks are apparently used.

Exercises

Exercise 11.1: Suppose that $f(X) \in \mathbb{Z}[X]$ is not square-free. Show that Newton's iteration works with $g(x) = f(X)/f'(X)$ instead of $f(X)$. Derive a similar guaranteed convergence bound for $g(X)$. □

Exercise 11.2: (a) Let $f(X) \in \mathbb{C}[X]$ be square-free with roots $\alpha_1, \ldots, \alpha_m$. Show that

$$\prod_{i=1}^{m} f'(\alpha_i) = (-1)^{\binom{m}{2}} a^{-m+2} \mathtt{disc}(f).$$

Deduce a lower bound on $|f'(\alpha_i)|$ from this.
(b) By modifying the proof of the Davenport-Mahler root separation bound show

$$\pm\sqrt{\frac{|\mathtt{disc}(f)|}{a^{2m-2}}} = \prod_{j=2}^{n}(\alpha_j - \alpha_1) \cdot \det \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \alpha_1 & 1 & & 1 \\ \alpha_1^2 & \beta_2^{(2)} & & \beta_m^{(2)} \\ \vdots & \vdots & & \vdots \\ \alpha_1^{m-1} & \beta_2^{(m-1)} & & \beta_m^{(m-1)} \end{bmatrix}$$

where $\beta_j^{(k)} = (\alpha_j^k - \alpha_1^k)/(\alpha_j - \alpha_1)$. Note that $\left| a \prod_{j=2}^{n}(\alpha_j - \alpha_1) \right| = |f'(\alpha_1)|$. If $f(X)$ is monic, deduce a lower bound for $|f'(\alpha)|$.
(c) Consider the resultant $R(f,g)$ of $f(X), g(X)$. Show that

$$R(f,g) = \begin{bmatrix} f_m & f_{m-1} & \cdots & & f_0 & & & \\ & f_m & f_{m-1} & \cdots & & f_0 & & \\ & & \ddots & & & & \ddots & \\ & & & f_m & f_{m-1} & \cdots & & f_0 \\ g_n & g_{n-1} & \cdots & & g_0 & & & \\ & g_n & g_{n-1} & \cdots & & g_0 & & \\ & & \ddots & & & & \ddots & \\ & & & g_n & & \cdots & & g_0 \end{bmatrix}$$

is equal to

$$\begin{bmatrix}
f_m & f_{m-1} & \cdots & & f_0 & & & \alpha^{n-1}f(\alpha) \\
 & f_m & f_{m-1} & \cdots & & f_0 & & \alpha^{n-2}f(\alpha) \\
 & & \ddots & & & & \ddots & \vdots \\
 & & & f_m & f_{m-1} & \cdots & f_1 & f(\alpha) \\
g_n & g_{n-1} & \cdots & & g_0 & & & \alpha^{m-1}g(\alpha) \\
 & g_n & g_{n-1} & \cdots & & g_0 & & \alpha^{m-2}g(\alpha) \\
 & & \ddots & & & & \ddots & \vdots \\
 & & & g_n & & \cdots & g_1 & g(\alpha)
\end{bmatrix}$$

for any value α. Derive from this a lower bound for $|f'(\alpha)|$ where α is a root of $f(X)$.
(d) (Lang) Show that if $f(X)$ has only simple roots α_i $(i = 1, \ldots, n)$, and ω is any complex number then

$$\min_i |\omega - \alpha_i| \leq |f(\omega)|e^{5n(\log n + h + 4)}$$

where $h = \|f\|_\infty$. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture VII
Sturm Theory

We owe to Descartes the problem of counting the number of real roots of a polynomial, and to Waring (1762) and Lagrange (1773) the problem of separating these roots. Lagrange gave the first complete algorithm for separating roots, which Burnside and Panton [2] declared "practically useless", a testimony to some implicit efficiency criteria. The decisive technique was found by Sturm in 1829. It superseded the research of his contemporaries, Budan (1807) and Fourier (1831) who independently improved on Descartes and Lagrange. In one sense, Sturm's work culminated a line of research that began with Descartes' rule of sign. According to Burnside and Panton, the combination of Horner and Sturm gives the best root separation algorithm of their day. Hurwitz, Hermite and Routh all made major contributions to the subject. Sylvester was especially interested in Sturm's work, as part of his interest in elimination theory and theory of equations [16]. In [14], he alludes to a general theory encompassing Sturm theory. This is apparently the tome of an article [15]. Uspensky [17] rated highly a method of root separation based on a theorem of Vincent (1836). Of course, all these evaluations of computational methods are based on some implicit model of the human-hand-computer. With the advent of complexity theory we have more objective methods of evaluating algorithms.

One profound generalization of Sturm's theorem is obtained by Tarski, in his famous result showing the decidability of elementary algebra and geometry (see [7]). Hermite had interest in generalizing Sturm's theory to higher dimensions, and considered some special cases; the general case has recently been achieved in the theses of Pedersen [11] and Milne [9].

§1. Sturm Sequences from PRS

We introduce Sturm's remarkable computational tool for counting the real zeros of a real function. We also show a systematic construction of such sequences from a PRS (§III.2). Our next definition is slightly more general than the usual.

Let $A(X), B(X) \in \mathbb{R}[X]$ be non-zero polynomials. By a *(generalized) Sturm sequence* for $A(X)$, $B(X)$ we mean a PRS

$$\overline{A} = (A_0, A_1, \ldots, A_h), \quad h \geq 1,$$

for $A(X)$, $B(X)$ such that for all $i = 1, \ldots, h$, we have

$$\beta_i A_{i+1} = \alpha_i A_{i-1} + Q_i A_i \tag{1}$$

$(\alpha_i, \beta_i \in \mathbb{R}, Q_i \in \mathbb{R}[X])$ such that $A_{h+1} = 0$ and $\alpha_i \beta_i < 0$.

We call $\overline{A}$ a *Sturm sequence for* A if it is a Sturm sequence for A, A' where A' denotes the derivative of A.

Note that we do not assume $\deg A \geq \deg B$ in this definition. However, if $\deg A < \deg B$ then it is clear that $A = A_0$ and A_2 are equal up to a negative constant factor. In any case, the degrees of all subsequent polynomials are strictly decreasing, $\deg A_1 > \deg A_2 > \cdots > \deg A_h \geq 0$. Note that the relation (1) exists by the definition of PRS.

Connection between a PRS and a Sturm sequence. Essentially, a Sturm sequence differs from a PRS only by virtue of the special sign requirements on the coefficients of similarity α_i, β_i.

Although this connection is well-known, the actual form of this connection has not been clearly elucidated. Our goal here is to do this, and in a way that the transformation of a PRS algorithm into a Sturm sequence algorithm can be routine.

Assume that we are given a PRS $\overline{A} = (A_0, \ldots, A_h)$. We need not know the values α_i, β_i or Q_i in equation (1), but we do require knowledge of the product

$$s_i := -\mathtt{sign}(\alpha_i \beta_i) \tag{2}$$

of signs, for $i = 1, \ldots, h-1$. Here $\mathtt{sign}(x)$ is a real function defined as expected,

$$\mathtt{sign}(x) := \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases} . \tag{3}$$

In the known PRS algorithms, these signs can be obtained as a byproduct of computing the PRS. We will now construct a sequence

$$(\sigma_0, \sigma_1, \ldots, \sigma_h),$$

of signs where $\sigma_0 = \sigma_1 = +1$ and $\sigma_i \in \{-1, 0, +1\}$ such that

$$(\sigma_0 A_0, \sigma_1 A_1, \ldots, \sigma_h A_h) \tag{4}$$

is a Sturm sequence. From (1) we see that

$$(\beta_i \sigma_{i+1})(\sigma_{i+1} A_{i+1}) = (\alpha_i \sigma_{i-1})(\sigma_{i-1} A_{i-1}) + Q_i A_i.$$

Hence (4) is a Sturm sequence provided that $\mathtt{sign}(\alpha_i \sigma_{i+1} \beta_i \sigma_{i-1}) = -1$ or, using equation (2),

$$\mathtt{sign}(s_i \sigma_{i+1} \sigma_{i-1}) = 1.$$

Multiplying together j $(2 \le 2j \le h)$ of these equations,

$$(\sigma_0 s_1 \sigma_2)(\sigma_2 s_3 \sigma_4)(\sigma_4 \sigma_5 \sigma_6) \cdots (\sigma_{2j-2} s_{2j-1} \sigma_{2j}) = 1.$$

Telescoping, we obtain the desired formula for σ_{2j}:

$$\sigma_{2j} = \prod_{i=1}^{j} s_{2i-1}. \tag{5}$$

Similarly, we have the formula for σ_{2j+1} $(2 \le 2j+1 \le h)$:

$$\sigma_{2j+1} = \prod_{i=1}^{j} s_{2i}. \tag{6}$$

Thus the sequence $(\sigma_1, \ldots, \sigma_h)$ of signs splits into two alternating subsequences whose computation depends on two disjoint subsets of $\{s_1, \ldots, s_{h-1}\}$. Also (5) and (6) can be rapidly computed in parallel, using the so-called parallel prefix algorithm.

Descartes' Rule of Sign. As noted in the introduction, the theory of Sturm sequences basically supersedes Descartes' Rule of Sign (or its generalizations) as a tool for root counting. The rule says:

> *The sign variation in the sequence $(a_n, a_{n-1}, \ldots, a_1, a_0)$ of coefficients of the polynomial $P(X) = \sum_{i=0}^{n} a_i X^i$ is more than the number of positive real roots of $P(X)$ by some non-negative even number.*

The proof of this and its generalization is left to an exercise.

Exercises

Exercise 1.1: Modify the subresultant algorithm (§III.5) of Collins to produce a Sturm Sequence. NOTE: in §III.5, we assume that the input polynomials P, Q satisfy $\deg P > \deg Q$. A small modification must now be made to handle the possibility that $\deg P \leq \deg Q$. □

Exercise 1.2: Prove Descartes' Rule of Sign. HINT: let $Q(X)$ be a real polynomial and α a positive real number. The number of sign variations in the coefficient sequence of $(X-\alpha)Q(X)$ is more than that of the coefficient sequence of $Q(X)$ by a positive odd number. □

Exercise 1.3: (i) Give the analogue of Descartes' rule of sign for negative real roots.
(ii) Prove that if $P(X)$ has only real roots, then the number of sign variations in $P(X)$ and $P(-X)$ is exactly n.
(iii) Let $(a_n, \ldots, a_1, a_0)$ be the sequence of coefficients of $P(X)$. If $a_n a_0 \neq 0$ and $P(X)$ has only real roots, then the sequence has the property that $a_i = 0$ implies $a_{i-1}a_{i+1} < 0$. □

Exercise 1.4: Newton's rule for counting the number of imaginary roots (see quotation preceding this lecture) is modified in case a polynomial has a block of two or more consecutive terms that are missing. Newton specifies the following rule for such terms:

> *If two or more terms are simultaneously lacking, beneath the first of the deficient terms, the sign − must be placed, beneath the second, +, etc., except that beneath the last of the terms simultaneously lacking, you must always place the sign + when the terms next on either sides of the deficient ones have contrary signs.*

He gives the following examples:

$$\begin{array}{ccccccccccc} & & \frac{2}{5} & & \frac{1}{2} & & \frac{1}{2} & & \frac{2}{5} & & \\ X^5 & + & aX^4 & + & 0 & + & 0 & + & 0 & + & a^5 \quad (4 \text{ imaginary roots}) \\ + & & + & & - & & + & & - & & + \end{array}$$

$$\begin{array}{ccccccccccc} & & \frac{2}{5} & & \frac{1}{2} & & \frac{1}{2} & & \frac{2}{5} & & \\ X^5 & + & aX^4 & + & 0 & + & 0 & + & 0 & - & a^5 \quad (2 \text{ imaginary roots}) \\ + & & + & & - & & + & & + & & + \end{array}$$

(i) Restate Newton's rule in modern terminology.
(ii) Count the number of imaginary roots of the polynomials $X^7 - 2X^6 + 3X^5 - 2X^4 + X^3 - 3 = 0$, and $X^4 + 14X^2 - 8X + 49$. □

§2. A Generalized Sturm Theorem

Let $\overline{\alpha} = (\alpha_0, \ldots, \alpha_h)$ be a sequence of real numbers. We say there is a *sign variation* in $\overline{\alpha}$ *at position* i $(i = 1, \ldots, h)$ if for some $j = 0, \ldots, i-1$ we have

(i) $\alpha_j \, \alpha_i < 0$

(ii) $\alpha_{j+1} = \alpha_{j+2} = \cdots = \alpha_{i-1} = 0$.

The *sign variation* of $\overline{\alpha}$ is the number of positions in $\overline{\alpha}$ where there is a sign variation.

For instance, the sequence $(0, -1, 0, 3, 8, -7, 9, 0, 0, 8)$ has sign variations at positions $3, 5$ and 6. Hence its sign variation is 3.

For any sequence $\overline{A} = (A_0, \ldots, A_h)$ of polynomials and $\alpha \in \mathbb{R}$, let $\overline{A}(\alpha)$ denote the sequence $(A_0(\alpha), \ldots, A_h(\alpha))$. Then the sign variation of $\overline{A}(\alpha)$ is denoted

$$\mathtt{Var}_{\overline{A}}(\alpha),$$

where we may omit the subscript when $\overline{A}$ is understood. If $\overline{A}$ is the Sturm sequence for A, B, we may write $\mathtt{Var}_{A,B}(\alpha)$ instead of $\mathtt{Var}_{\overline{A}}(\alpha)$. If $\alpha < \beta$, we define the *sign variation difference* over the interval $[\alpha, \beta]$ to be

$$\mathtt{Var}_{\overline{A}}[\alpha, \beta] := \mathtt{Var}_{\overline{A}}(\alpha) - \mathtt{Var}_{\overline{A}}(\beta). \tag{7}$$

There are different forms of "Sturm theory". Each form of Sturm theory amounts to giving an interpretation to the sign variation difference (7), for a suitable notion of the "Sturm sequence" $\overline{A}$. In this section, we prove a general (apparently new) theorem to encompass several known Sturm theories.

In terms of counting sign variations, Exercise 7.2.1 indicates that all Sturm sequences for A, B are equivalent. Hence, we may loosely refer to *the* Sturm sequence of A, B.

Let $r \geq 0$ be a non-negative integer. Recall that α is a root of *multiplicity* r (equivalently, α is an *r-fold root*) of an r-fold differentiable function $f(X)$ if

$$f^{(0)}(\alpha) = f^{(1)}(\alpha) = \cdots = f^{(r-1)}(\alpha) = 0, \qquad f^{(r)}(\alpha) \neq 0.$$

So we refer (awkwardly) to a non-root of f as a 0-fold root. However, if we simply say 'α is a root of f' then it is understood that the multiplicity r is positive. If h is sufficiently small and α is an r-fold root, then Taylor's theorem with remainder gives us

$$f(\alpha + h) = \frac{h^r}{r!} \cdot f^{(r)}(\alpha + \theta h)$$

for some θ, $0 \leq \theta \leq 1$. So for $h > 0$, $f(\alpha + h)$ has the sign of $f^{(r)}(\alpha)$; for $h < 0$, $f(\alpha + h)$ has the sign of $(-1)^r f^{(r)}(\alpha)$. Hence:

> *If r is odd, $f(X)$ changes sign in the neighborhood of α;*
> *If r is even, $f(X)$ maintains its sign in the neighborhood of α.*

Let $\overline{A} = (A_0, \ldots, A_h)$ be a sequence of non-zero polynomials and α a real number.
i) We say α is *regular for* $\overline{A}$ if each $A_i(X) \in \overline{A}$ is non-vanishing at $X = \alpha$; otherwise, α is *irregular*.
ii) We say α is *degenerate for* $\overline{A}$ if each $A_i(X) \in \overline{A}$ vanishes at $X = \alpha$; otherwise α is *nondegenerate*.
iii) A closed interval $[\alpha, \beta]$ where $\alpha < \beta$ is called a *fundamental interval* (at γ_0) for $\overline{A}$ if α, β are non-roots of A_0 and there exists $\gamma_0 \in [\alpha, \beta]$ such that for all $\gamma \in [\alpha, \beta]$, if $\gamma \neq \gamma_0$ then γ is regular for $\overline{A}$. Note that γ_0 can be equal to α or β.

Hence α may be neither regular nor degenerate for $\overline{A}$, *i.e.*, it is both irregular and nondegenerate for $\overline{A}$. The following characterizes nondegeneracy.

Lemma 1 *Let $\overline{A} = (A_0, \ldots, A_h)$ be a Sturm sequence.*
a) The following are equivalent:

(i) *α is degenerate for $\overline{A}$.*
(ii) *Two consecutive polynomials in $\overline{A}$ vanish at α.*
(iii) *A_h vanishes at α.*

b) If α is nondegenerate and $A_i(\alpha) = 0$ ($i = 1, \ldots, h-1$) then $A_{i-1}(\alpha)A_{i+1}(\alpha) < 0$.

Proof.
a) If α is degenerate for $\overline{A}$ then clearly any two consecutive polynomials would vanish at α. Conversely, if $A_{i-1}(\alpha) = A_i(\alpha) = 0$, then from equation (1), we see that $A_{i+1}(\alpha) = 0$ ($i + 1 \leq h$) and $A_{i-2}(\alpha) = 0$ ($i - 2 \geq 0$). Repeating this argument, we see that every A_j vanishes at α. Thus α is degenerate for $\overline{A}$. This proves the equivalence of (i) and (ii). The equivalence of (ii) and (iii) is easy once we recall that A_h divides A_{h-1}, by definition of a PRS. Hence A_h vanishes at α implies A_{h-1} vanishes at α.
b) This follows from the fact that $\alpha_i\beta_i < 0$ in equation (1). **Q.E.D.**

The importance of fundamental intervals arises as follows. Suppose we want to evaluate $\mathtt{Var}_{A,B}[\alpha, \beta]$ where α, β are non-roots of A. Clearly, there are only a finite number of irregular values in the interval $[\alpha, \beta]$. If there are no irregular values in the interval, then trivially $\mathtt{Var}_{A,B}[\alpha, \beta] = 0$. Otherwise, we can find values

$$\alpha = \alpha_0 < \alpha_1 < \cdots < \alpha_k = \beta$$

such that each $[\alpha_{i-1}, \alpha_i]$ is a fundamental interval. Clearly

$$\mathtt{Var}_{A,B}[\alpha, \beta] = \sum_{i=1}^{k} \mathtt{Var}_{A,B}[\alpha_{i-1}, \alpha_i].$$

So we have reduced our problem to sign variation difference on fundamental intervals.

Given real polynomials $A(X), B(X)$, we say $A(X)$ *dominates* $B(X)$ if for each root α of $A(X)$, we have

$$r \geq s \geq 0$$

where α is an r-fold root of $A(X)$ and an s-fold root of $B(X)$.

Note that $r \geq 1$ here since α is a root of $A(X)$. Despite the terminology, "domination" is neither transitive nor asymmetric as a binary relation on real polynomials. We use the concept of domination in the following four situations, where in each case $A(X)$ dominates $B(X)$:

- $B(X)$ is the derivative of $A(X)$.
- $A(X)$ and $B(X)$ are relatively prime.
- $A(X)$ and $B(X)$ are both square-free.
- $B(X)$ divides $A(X)$.

We have invented the concept of domination to unify these We come to our key lemma.

Lemma 2 *Let $\overline{A} = (A_0, \ldots, A_h)$ be a Sturm sequence for A, B where A dominates B. If $[\alpha, \beta]$ is a fundamental interval at γ_0 for $\overline{A}$ then*

$$\mathtt{Var}_{\overline{A}}[\alpha, \beta] = \begin{cases} 0 & \text{if } r = 0 \text{ or } r + s \text{ is even} \\ \\ \mathrm{sign}(A^{(r)}(\gamma_0)B^{(s)}(\gamma_0)) & \text{if } r \geq 1 \text{ and } r + s \text{ is odd,} \end{cases}$$

where γ_0 is an r-fold root of $A(X)$ and also an s-fold root of $B(X)$.

Proof. We break the proof into two parts, depending on whether γ_0 is degenerate for $\overline{A}$.

Part I. Suppose γ_0 is nondegenerate for $\overline{A}$. Then $A_h(\gamma_0) \neq 0$. We may define the unique sequence

$$0 = \pi(0) < \pi(1) < \cdots < \pi(k) = h, \qquad (k \geq 1)$$

such that for all $i > 0$, $A_i(\gamma_0) \neq 0$ iff $i \in \{\pi(1), \pi(2), \ldots, \pi(k)\}$. Note that $\pi(0) = 0$ has special treatment in this definition. Define for each $j = 1, \ldots, k$, the subsequence $\overline{B}_j$ of $\overline{A}$:

$$\overline{B}_j := (A_{\pi(j-1)}, A_{\pi(j-1)+1}, \ldots, A_{\pi(j)}).$$

Since two consecutive polynomials of $\overline{A}$ cannot vanish at a nondegenerate γ_0, it follows that since $\pi(j) - \pi(j-1)$ equals 1 or 2 (*i.e.*, each $\overline{B}_j$ has 2 or 3 members). Indeed, $\overline{B}_j$ has 3 members iff its middle member vanishes at γ_0. Then the sign variation difference can be expressed as

$$\mathtt{Var}_{A,B}[\alpha, \beta] = \sum_{i=1}^{k} \mathtt{Var}_{\overline{B}_i}[\alpha, \beta]. \tag{8}$$

Let us evaluate $\mathtt{Var}_{\overline{B}_i}[\alpha, \beta]$ in two cases:

CASE 1: $\mathtt{Var}_{\overline{B}_i}[\alpha, \beta]$ has three members. The signs of the first and third member do not vary in the entire interval $[\alpha, \beta]$. In fact, the signs of the first and third member must be opposite. On the other hand, the signs of the middle member at α and at β are different (one of them can be the zero sign). But regardless, it is now easy to conclude $\mathtt{Var}_{\overline{B}_i}[\alpha, \beta] = 1 - 1 = 0$.

CASE 2: $\mathtt{Var}_{\overline{B}_i}[\alpha, \beta]$ has two members. There are two possibilities, depending on whether the first member of the sequence $\overline{B}_i$ vanishes at γ_0 or not. In fact, the first member vanishes iff $i = 1$ (so $\overline{B}_1 = (A, B)$ and $A(\gamma_0) = 0$). If $A(\gamma_0) \neq 0$, then the signs of both members in $\overline{B}_i$ do not vary in the entire interval $[\alpha, \beta]$. This proves $\mathtt{Var}_{\overline{B}_i}[\alpha, \beta] = 0$, as required by the lemma when $A(\gamma_0) \neq 0$.

Before we consider the remaining possibility where $A(\gamma_0) = 0$, we may simplify equation (8), using the fact that all the cases we have considered until now yield $\mathtt{Var}_{\overline{B}_i}[\alpha, \beta] = 0$:

$$\mathtt{Var}_{A,B}[\alpha, \beta] = \begin{cases} \mathtt{Var}_{\overline{B}_1}[\alpha, \beta] & \text{if } A(\gamma_0) = 0, \\ \\ 0 & \text{else.} \end{cases} \tag{9}$$

Note that if $A(\gamma_0) \neq 0$ then $r = 0$. Thus equation (9) verifies our lemma for the case $r = 0$.

Hence assume $A(\gamma_0) = 0$, *i.e.*, $r \geq 1$. We have $s = 0$ because γ_0 is assumed to be nondegenerate for $\overline{A}$. Also $\alpha < \gamma_0 < \beta$ since $A(X)$ does not vanish at α or β (definition of fundamental interval). There are two subcases.

SUBCASE: r is even. Then $A(X)$ and $B(X)$ both maintain their signs in the neighborhood of γ_0 (except temporarily vanishing at γ_0). Then we see that

$$\mathtt{Var}_{\overline{B}_1}(\alpha) = \mathtt{Var}_{\overline{B}_1}(\beta),$$

proving the lemma in this subcase.

SUBCASE: r is odd. Then $A(X)$ changes sign at γ_0 while $B(X)$ maintains its sign in $[\alpha, \beta]$. Hence $\mathtt{Var}_{\overline{B}_1}[\alpha, \beta] = \pm 1$. In fact, the following holds:

$$\mathtt{Var}_{\overline{B}_1}[\alpha, \beta] = \mathtt{sign}(A^{(r)}(\gamma_0) B^{(s)}(\gamma_0)), \tag{10}$$

proving the lemma when $s = 0$ and $r \geq 1$ is odd. [Let us verify equation (10) in case $B(X) > 0$ throughout the interval. There are two possibilities: if $A^{(r)}(\gamma_0) < 0$ then we get $\mathtt{Var}_{\overline{B}_1}(\alpha) = 0$ and $\mathtt{Var}_{\overline{B}_1}(\beta) = 1$ so that $\mathtt{Var}_{\overline{B}_1}[\alpha,\beta] = \mathtt{sign}(A^{(r)}(\gamma_0))$. If $A^{(r)}(\gamma_0) > 0$ then $\mathtt{Var}_{\overline{B}_1}(\alpha) = 1$ and $\mathtt{Var}_{\overline{B}_1}(\beta) = 0$, and again $\mathtt{Var}_{\overline{B}_1}[\alpha,\beta] = \mathtt{sign}(A^{(r)}(\gamma_0))$.]

Part II. Now assume γ_0 is degenerate. This means $\alpha < \gamma_0 < \beta$. Let

$$\overline{C} = (A_0/A_h, A_1/A_h, \ldots, A_h/A_h)$$

be the *depressed sequence* derived from $\overline{A}$. This is a Sturm sequence for $C_0 = A_0/A_h$, $C_1 = A_1/A_h$. Moreover, γ_0 is no longer degenerate for $\overline{C}$, and we have

$$\mathtt{Var}_{\overline{A}}(\gamma) = \mathtt{Var}_{\overline{C}}(\gamma),$$

for all $\gamma \in [\alpha,\beta]$, $\gamma \neq \gamma_0$. Since $[\alpha,\beta]$ remains a fundamental interval at γ_0 for $\overline{C}$, the result of part I in this proof can now be applied to $\overline{C}$, showing

$$\mathtt{Var}_{\overline{C}}[\alpha,\beta] = \begin{cases} 0 & \text{if } r^* = 0 \text{ or } r^* + s^* \text{ is even,} \\ \\ \mathtt{sign}(C_0^{(r^*)}(\gamma_0)C_1^{(s^*)}(\gamma_0)) & \text{if } r^* \geq 1 \text{ and } r^* + s^* \text{ is odd.} \end{cases} \tag{11}$$

Here r^*, s^* are the multiplicities of γ_0 as roots of C_0, C_1 (respectively). Clearly, if γ_0 is an m-fold root of $A_h(X)$, then $r = r^* + m$, $s = s^* + m$. Hence $r^* + s^* =$ even iff $r + s =$ even. This shows

$$\mathtt{Var}_{\overline{A}}[\alpha,\beta] = \mathtt{Var}_{\overline{C}}[\alpha,\beta] = 0$$

when $r + s =$ even, as desired. If $r^* + s^* =$ odd and $r^* \geq 1$, we must show

$$\mathtt{sign}(C_0^{(r^*)}(\gamma_0)C_1^{(s^*)}(\gamma_0)) = \mathtt{sign}(A^{(r)}(\gamma_0)B^{(s)}(\gamma_0)). \tag{12}$$

For clarity, let $A_h(X)$ be rewritten as $D(X)$ so that

$$\begin{aligned} A(X) &= C_0(X) \cdot D(X) \\ A^{(r)}(X) &= \sum_{i=0}^{r} \binom{r}{i} C_0^{(i)}(X) D^{(r-i)}(X) \\ A^{(r)}(\gamma_0) &= \binom{r}{r^*} C_0^{(r^*)}(\gamma_0) D^{(m)}(\gamma_0) \end{aligned}$$

since $C_0^{(i)}(\gamma_0) = 0$ for $i < r^*$, and $D^{(r-i)}(\gamma_0) = 0$ for $i > r^*$. Similarly,

$$B^{(s)}(\gamma_0) = \binom{s}{s^*} C_1^{(s^*)}(\gamma_0) D^{(m)}(\gamma_0).$$

This proves (12).

Finally suppose $r^* = 0$. But the assumption that A dominates B implies $s^* = 0$. [This is the only place where domination is used.] Hence $s^* + r^*$ is even and $\mathtt{Var}_{\overline{C}}[\alpha,\beta] = 0$. Hence $s + r$ is also even and $\mathtt{Var}_{\overline{A}}[\alpha,\beta] = 0$. This completes the proof. **Q.E.D.**

This lemma immediately yields the following:

Theorem 3 (Generalized Sturm) *Let A dominate B and let $\alpha < \beta$ so that $A(\alpha)A(\beta) \neq 0$. Then*

$$\mathtt{Var}_{A,B}[\alpha,\beta] = \sum_{\gamma,r,s} \mathtt{sign}(A^{(r)}(\gamma)B^{(s)}(\gamma)) \tag{13}$$

where γ ranges over all roots of A in $[\alpha, \beta]$ of multiplicity $r \geq 1$, and B has multiplicity s at γ, and $r + s =$ odd.

The statement of this theorem can be generalized in two ways without modifying the proof:
(a) We only need to assume that A dominates B within the interval $[\alpha, \beta]$, *i.e.*, at the roots of A in the interval, the multiplicity of A is at least that of the multiplicity of B.
(b) The concept of domination can be extended to mean that at each root γ of A (restricted to $[\alpha, \beta]$ as in (a) if we wish), if A, B have multiplicities r, s (respectively) at γ, then $\max\{0, s - r\}$ is even.

Exercises

Exercise 2.1: Suppose $\overline{A}$ and $\overline{B}$ are both Sturm sequences for $A, B \in \mathbb{R}[X]$. Then they have the same length and corresponding elements of $\overline{A}$ and $\overline{B}$ are related by positive factors: $A_i = \alpha_i B_i$ where α_i is a positive real number. □

Exercise 2.2: The text preceding Lemma 7.2 specified four situations were $A(X)$ dominates $B(X)$. Verify domination in each case. □

Exercise 2.3: (Budan-Fourier) Let $A_0(X)$ be a polynomial, $\alpha < \beta$ and $A_0(\alpha)A_0(\beta) \neq 0$. Let $\overline{A} = (A_0, A_1, \ldots, A_h)$ be the sequence of non-zero derivatives of A_0, *viz.*, A_i is the ith derivative of A_0. Then the number of real zeros of $A_0(X)$ in $[\alpha, \beta]$ is less than the $\mathtt{Var}_{\overline{A}}[\alpha, \beta]$ by an even number. HINT: Relate the location of zeros of $A(X)$ and its derivative $A'(X)$. Use induction on $\deg A_0$. □

Exercise 2.4: a) Deduce Descartes' Rule of Sign (§1) from the Budan-Fourier Rule (see previous exercise).
b) (Barbeau) Show that Descartes' Rule gives a sharper estimate for the number of negative zeros than Budan-Fourier for the polynomial $X^4 + X^2 + 4X - 3$. □

§3. Corollaries and Applications

We obtain four useful corollaries to the generalized Sturm theorem. The first is the classic theorem of Sturm.

Corollary 4 (Sturm) *Let $A(X) \in \mathbb{R}[X]$ and suppose $\alpha < \beta$ are both non-roots of A. Then the number of distinct real roots of $A(X)$ in the interval $[\alpha, \beta]$ is given by $\mathtt{Var}_{A,A'}[\alpha, \beta]$.*

Proof. With $B(X) = A'(X)$, we see that $A(X)$ dominates $B(X)$ so that the generalized Sturm theorem gives:

$$\mathtt{Var}_{A,B}[\alpha, \beta] = \sum_{\gamma, r, s} \mathtt{sign}(A^{(r)}(\gamma)B^{(s)}(\gamma)),$$

where γ is an r-fold root of A in (α, β), γ is an s-fold root of B and $r \geq 1$ with $r + s$ being odd. But at every root of A, these conditions are satisfied since $r = s + 1$. Hence the summation applies to

every root γ of A. Furthermore, we see that $A^{(r)}(\gamma) = B^{(s)}(\gamma)$ so that $\mathtt{sign}\,(A^{(r)}(\gamma)B^{(s)}(\gamma)) = 1$. So the summation yields the number of roots of A in $[\alpha, \beta]$. **Q.E.D.**

Note that it is computationally convenient that our version of Sturm's theorem does not assume $A(X)$ is square-free (which is often imposed).

Corollary 5 (Schwartz-Scharir) *Let $A(X), B(X) \in \mathbb{R}[X]$ be square-free polynomials. If $\alpha < \beta$ are both non-roots of A then*

$$\mathtt{Var}_{A,B}[\alpha, \beta] = \sum_{\gamma} \mathtt{sign}(A'(\gamma)B(\gamma))$$

where γ ranges over all roots of $A(X)$ in $[\alpha, \beta]$.

Proof. We may apply the generalized Sturm theorem to evaluate $\mathtt{Var}_{A,B}[\alpha, \beta]$ in this corollary. In the sum of (13), consider the term indexed by the triple (γ, r, s) with $r \geq 1$ and $r + s$ is odd. By square-freeness of A and B, we have $r \leq 1$ and $s \leq 1$. Thus $r = 1$, $s = 0$ and equation (13) reduces to

$$\mathtt{Var}_{A,B}[\alpha, \beta] = \sum_{\gamma} \mathtt{sign}(A'(\gamma)B(\gamma)),$$

where the summation is over roots γ of A in $[\alpha, \beta]$ which are not roots of B. But if γ is both a root of A and of B then $\mathtt{sign}(A'(\gamma)B(\gamma)) = 0$ and we may add these terms to the summation without any effect. This is the summation sought by the corollary. **Q.E.D.**

The next corollary will be useful in §7:

Corollary 6 (Sylvester, revisited by Ben-Or, Kozen, Reif) *Let $\overline{A}$ be a Sturm sequence for $A, A'B$ where $A(X)$ is square-free and $A(X), B(X)$ are relatively prime. Then for all $\alpha < \beta$ which are non-roots of A,*

$$\mathtt{Var}_{\overline{A}}[\alpha, \beta] = \sum_{\gamma} \mathtt{sign}(B(\gamma))$$

where γ ranges over the roots of $A(X)$ in $[\alpha, \beta]$.

Proof. Again note that A dominates $A'B$ and we can proceed as in the proof of the previous corollary. But now, we get

$$\begin{aligned} \mathtt{Var}_{\overline{A}}[\alpha, \beta] &= \sum_{\gamma} \mathtt{sign}(A'(\gamma) \cdot A'(\gamma)B(\gamma)) \\ &= \sum_{\gamma} \mathtt{sign}(B(\gamma)), \end{aligned}$$

as desired. **Q.E.D.**

In this corollary, the degree of $A_0 = A$ is generally less than the degree of $A_1 = A'B$ so that the remainder sequence typically looks like this: $\overline{A} = (A, A'B, -A, \ldots)$.

Our final corollary concerns the concept of the Cauchy index of a rational function. Let $f(X)$ be a real continuous function defined in an open interval (α, β) where $-\infty \leq \alpha < \beta \leq +\infty$. We

allow $f(X)$ to have isolated poles in the interval (α, β). Recall that $\gamma \in (\alpha, \beta)$ is a *pole* of $f(X)$ if $1/f(X) \to 0$ as $X \to \gamma$. The *Cauchy index* of f at a pole γ is defined[1] to be

$$\frac{\mathtt{sign}(f(\gamma^-)) - \mathtt{sign}(f(\gamma^+))}{2}.$$

For instance, the index is -1 if $f(X)$ changes from $-\infty$ to $+\infty$ as X increases through γ, and the index is 0 if the sign of $f(X)$ does not change in passing through γ. The *Cauchy index* of f over an interval (α, β) is then

$$I_\alpha^\beta f(X) := \sum_\gamma \frac{\mathtt{sign}(f(\gamma^-)) - \mathtt{sign}(f(\gamma^+))}{2}$$

where the sum is taken over all poles $\gamma \in (\alpha, \beta)$. Typically, $f(X)$ is a rational function $A(X)/B(X)$ where $A(X), B(X)$ are relatively prime polynomials.

Corollary 7 (Cauchy Index) *Let $A(X), B(X) \in \mathbb{R}[X]$ be relatively prime and $f(X) = A(X)/B(X)$. Then*

$$I_\alpha^\beta f(X) = -\mathtt{Var}_{A,B}[\alpha, \beta].$$

Proof. Let (γ, r, s) index a summation term in (13). We have $s = 0$ since A, B are relatively prime. This means that r is odd, and

$$\begin{aligned}
\mathtt{sign}(A^{(r)}(\gamma)) &= \frac{\mathtt{sign}(A(\gamma^+)) - \mathtt{sign}(A(\gamma^-))}{2}, \\
\mathtt{sign}(A^{(r)}(\gamma)B^{(0)}(\gamma)) &= \frac{\mathtt{sign}(A(\gamma^+)B(\gamma^+)) - \mathtt{sign}(A(\gamma^-)B(\gamma^-))}{2} \\
&= \frac{\mathtt{sign}(f(\gamma^+)) - \mathtt{sign}(f(\gamma^-))}{2}.
\end{aligned}$$

Summing the last equation over each (γ, r, s), the left-hand side equals $\mathtt{Var}_{A,B}[\alpha, \beta]$, by the generalized Sturm theorem. But the right-hand side equals $I_\alpha^\beta f$. **Q.E.D.**

This result is used in §5. For now, we give two applications of the corollary of Schwartz-Scharir (cf. [13]).

A. The sign of a real algebraic number. The first problem is to determine the sign of a number β in a real number field $\mathbb{Q}(\alpha)$. We assume that β is represented by a rational polynomial $B(X) \in \mathbb{Q}[X]$: $\beta = B(\alpha)$. Assume α is represented by the isolating interval representation (§VI.9)

$$\alpha \cong (A, [a, b])$$

where $A \in \mathbb{Z}[X]$ is a square-free polynomial. First let us assume $B(X)$ is square-free. To determine the sign of β, first observe that

$$\mathtt{sign}(A'(\alpha)) = \mathtt{sign}(A(b) - A(a)). \tag{14}$$

Using the corollary of Schwartz-Sharir,

$$\mathtt{Var}_{A,B}[a, b] = \mathtt{sign}(A'(\alpha) \cdot B(\alpha)).$$

[1] Here, $\mathtt{sign}(f(\gamma^-))$ denotes the sign of $f(X)$ for when $\gamma - X$ is positive but arbitrarily small. When $f(X)$ is a rational function, this sign is well-defined. Similarly $\mathtt{sign}(f(\gamma^+))$ is the sign of $f(X)$ when $X - \gamma$ is positive but arbitrarily small.

Hence,

$$\begin{aligned} \mathtt{sign}(B(\alpha)) &= \mathtt{sign}((\mathtt{Var}_{A,B}[a,b]) \cdot A'(\alpha)) \\ &= \mathtt{sign}((\mathtt{Var}_{A,B}[a,b]) \cdot (A(b) - A(a))). \end{aligned}$$

If $B(X)$ is not square-free, we can first decompose it into a product of square-free polynomials. That is, B has a *square-free decomposition* $B_1 \cdot B_2 \cdot \ldots \cdot B_k$ where B_1 is the square-free part of B and $B_2 \cdot \ldots \cdot B_k$ is recursively the square-free decomposition of B/B_1. Then $\mathtt{sign}(B(\alpha)) = \prod_{i=1}^{k} \mathtt{sign}(B_i(\alpha))$.

Exercise 3.1: Alternatively, use the Sylvester corollary to obtain the sign of $B(\alpha)$. □

B. Comparing two real algebraic numbers. Given two real algebraic numbers

$$\alpha \cong (A, I), \qquad \beta \cong (B, J)$$

represented as indicated by isolating intervals, we wish to compare them. Of course, one method is to determine the sign of $\alpha - \beta$, by a suitable reduction to the problem in Section 7.3.1. But we give a more direct reduction. If $I \cap J = \emptyset$ then the comparison is trivially done. Otherwise, if either $\alpha \notin I \cap J$ or $\beta \notin I \cap J$ then again we can easily determine which of α or β is bigger. Hence assume α and β are both in a common isolating interval $I \cap J = [a, b]$.

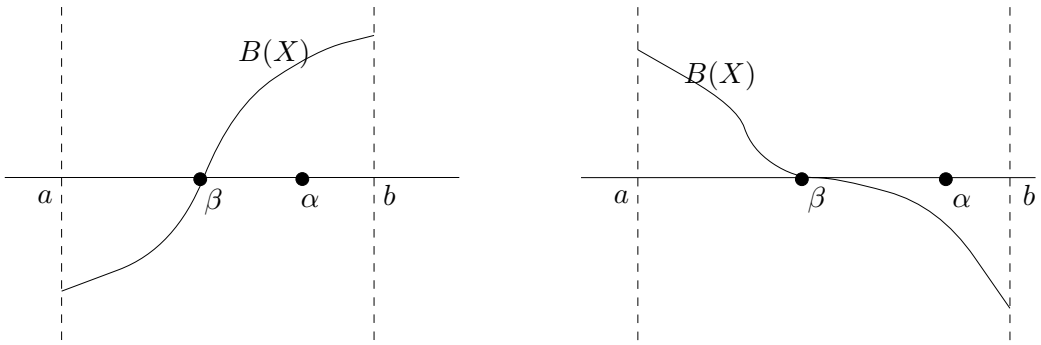


Figure 1: Two cases for $\alpha > \beta$ in isolating interval $[a, b]$.

It is not hard to verify (see Figure 1) that

$$\alpha \geq \beta \quad \Leftrightarrow \quad B(\alpha) \cdot B'(\beta) \geq 0,$$

with equality on the left-hand side if and only if equality is attained on the right-hand side (note that $B'(\beta) \neq 0$ by square-freeness of B). Since we already know how to obtain the signs of $B(\alpha)$ and of $B'(\beta)$ (Section 7.3.1) we are done:

$$B(\alpha) \cdot B'(\beta) \geq 0 \quad \Leftrightarrow \quad (\mathtt{Var}_{A,B}[a,b]) \cdot (A(b) - A(a)) \cdot (B(b) - B(a)) \geq 0.$$

Complexity of one incremental-bit of an algebraic number. Let α be an algebraic number, given as the ith real root of a square-free polynomial $A(X) \in \mathbb{Z}[X]$. Consider the following question: what is the complexity of finding out one *incremental-bit* of α? More precisely, suppose we already know that α lies within an interval I. How much work does it

take to halve the interval? There are three stages. *Sturm stage:* Initially, I can be taken to be $[-M, M]$ where $M = 1 + \|A\|_\infty$ is Cauchy's bound. We can halve I by counting the number of real roots of A in the interval $[-M, 0]$ and $[0, M]$. This takes two "Sturm queries" as given by corollary 4. Subsequently, assuming we already know the number of real roots inside I, each incremental-bit of α costs only one Sturm query. This continues until I is an isolating interval. *Bisection stage:* Now we may assume that we know the sign of $A(X)$ at the end-points of I. Henceforth, each incremental-bit costs only one polynomial evaluation, *viz.*, evaluating the sign of $A(X)$ at the mid-point of I. We continue this until the size Δ of I is within the range of guaranteed Newton convergence. *Newton stage:* According to §VI.11, it suffices to have $\Delta \leq m^{-3m-9}M^{-6m}$ where $m = \deg A$ and $M = 2 + \|A\|_\infty$. Let X_0 be the midpoint of I when Δ first reaches this bound. If Newton iteration transforms X_i to X_{i+1}, then the point X_i is within distance 2^{-2^i} of α (§VI.10). The corresponding interval I_i may be taken to have size $2^{1-2^i}\Delta$, centered at X_i. That is, we obtain about 2^i incremental-bits for i Newton steps. Each Newton step is essentially two polynomial evaluations. In an amortized sense, the cost is about 2^{-i+1} polynomial evaluations per incremental-bit for the ith Newton iteration.

Exercises

Exercise 3.2: Isolate the roots of:

(a) $(X^2 + 7)^2 - 8X = X^4 + 14X^2 - 8X + 49$.

(b) $X^{16} - 8X^{14} + 8X^{12} + 64X^{10} - 98X^8 - 184X^6 + 200X^4 + 224X^2 - 113$.

These are the minimal polynomials of $\sqrt{2} + \sqrt{5}$ and $\sqrt{1 + \sqrt{5 - 3\sqrt{1 + \sqrt{2}}}}$, respectively. □

Exercise 3.3: Isolate the roots of the following polynomials:

$$\begin{aligned} P_2(X) &= \frac{3}{2}X^2 - \frac{1}{2}, \\ P_3(X) &= \frac{5}{2}X^3 - \frac{3}{2}X, \\ P_4(X) &= \frac{35}{8}X^4 - \frac{15}{4}X^2 + \frac{3}{8}. \end{aligned}$$

These are the Legendre polynomials, which have all real and distinct roots lying in the interval $[-1, 1]$. □

Exercise 3.4: Give an algorithm for the square-free decomposition of a polynomial $B(X) \in \mathbb{Z}[X]$: $B(X) = B_1 B_2 \cdots B_k$ as described in the text. Analyze the complexity of your algorithm. □

.

Exercise 3.5: What does $\mathtt{Var}_{A,A''}[\alpha, \beta]$ count, assuming $\alpha < \beta$ and $A(\alpha)A(\beta) \neq 0$? □

.

Exercise 3.6: (a) Let $Q(Y) \in \mathbb{Q}(\alpha)[Y]$, where α is a real root of $P(X) \in \mathbb{Q}[X]$. Assume that we have an isolating interval representation for α (relative to $P(X)$) and the coefficients of

$Q(Y)$ are represented by rational polynomials in α. Show how to carry out a Sturm sequence computation to isolate the real roots of $Q(Y)$. Analyze the complexity of your algorithm.
(b) This gives us a method of representing elements of the double extension $\mathbb{Q}(\alpha)(\beta)$. Extend the method to multiple (real) extensions: $\mathbb{Q}(\alpha_1)\cdots(\alpha_k)$. Explain how arithmetic in such representations might be carried out. □

Exercise 3.7: (Schwartz-Sharir) Given an integer polynomial $P(X)$ (not necessarily square-free) and an isolating interval I of $P(X)$ for one of its real roots α, determine the multiplicity of $P(X)$ at α. □

Exercise 3.8: In order for all the roots of $P(X)$ to be real, it is necessary that the leading coefficients of a Sturm sequence of $P(X)$ be all positive. □

Exercise 3.9: Give a version of the generalized Sturm's theorem where we replace the condition that α, β are non-roots of A by the condition that these are nondegenerate. □

Exercise 3.10: Let $\alpha_1, \ldots, \alpha_k$ be real algebraic numbers with isolating interval representations. Preprocess this set of numbers so that, for any subsequently given integers $n_1, \ldots, n_k \in \mathbb{Z}$, you can efficiently test if $\sum_{i=1}^{k} n_i\alpha_i$ is zero. □

Exercise 3.11: (Sederberg and Chang)
(a) Let $P(X), B(X)$ and $C(X)$ be non-zero real polynomials and define

$$A(X) := B(X)P'(X) + C(X)P(X).$$

Then between any two adjacent real roots of $P(X)$ there is at least one real root of $A(X)$ or $B(X)$. (This statement can be interpreted in the natural way in case the two adjacent roots coincide.) In general, any pair $A(X), B(X)$ of polynomials with this property is called an *isolator pair* for $P(X)$.
(b) Let $P(X) = X^3 + aX^2 + bX + c$. Construct two linear polynomials $A(X)$ and $B(X)$ which form an isolator pair for $P(X)$. What are the roots $A(X)$ and $B(X)$? HINT: choose $B(X) = \frac{1}{3}(X + \frac{a}{3})$ and $C(X) = -1$.
(c) Relate the concept of isolator pairs to the polynomial remainder sequence of $P(X)$. □

Exercise 3.12*: Is there a simple method to decide if an integer polynomial has only real roots?
□

§4. Integer and Complex Roots

We discuss the special cases of integer and rational roots, and the more general case of complex roots.

Integer and Rational Roots. Let $A(X) = \sum_{i=0}^{n} a_i X^i$ be an integer polynomial of degree n. We observe that if u is an integer root of $A(X)$ then

$$a_0 = -\sum_{i=1}^{n} a_i u^i = -u\left(\sum_{i=1}^{n} a_i u^{i-1}\right)$$

and hence u divides a_0. Hence, checking if $A(X)$ has any integer roots it can be reduced to factorization of integers: we factor a_0 and for each integer factor u, we check if $A(u) = 0$. Similarly, if u/v is a rational root of $A(X)$ with $\texttt{GCD}(u, v) = 1$ it is easily checked that u divides a_0 and v divides a_n. [Thus, if u/v is a rational root of a monic integer polynomial then $v = 1$, *i.e.*, the set of algebraic integers that are rational is precisely $\mathbb{Z}$.] We can thus reduce the search for rational roots to the factorization of a_0 and a_n.

Hilbert's 10th problem asks for an algorithm to decide if an input integer polynomial has any integer roots. Matiyasevich (1970), building on the work of Davis, Putnam and Robinson [5], proved that no such algorithm exists, by showing that this is (many-one) equivalent to the Halting Problem. For an exposition of this result, see the book of Davis [4, Appendix 2] or [8]. It is an open problem whether there is an algorithm to decide if an input integer polynomial has any rational roots. This can be shown to be equivalent to restricting the inputs to Hilbert's 10th problem to homogeneous polynomials.

Complex Roots. We reduce the extraction of complex roots to the real case. The real and complex component of a complex algebraic number may be separately represented using isolating intervals. Suppose $P(X) \in \mathbb{C}[X]$ and $\overline{P}(X)$ is obtained by complex conjugation of each coefficient of $P(X)$. Then for $\alpha \in \mathbb{C}$, $\overline{P(\alpha)} = \overline{P}(\overline{\alpha})$. So $P(\alpha) = 0$ iff $\overline{P}(\overline{\alpha}) = 0$. It follows that if $P(X) = \prod_{i=1}^{n}(X - \alpha_i)$ then

$$P(X) \cdot \overline{P}(X) = (\prod_{i=1}^{n} X - \alpha_i)(\prod_{i=1}^{n} X - \overline{\alpha_i}).$$

Hence $P(X) \cdot \overline{P}(X)$ is a real polynomial, as $(X - \alpha_i)(X - \overline{\alpha_i}) \in \mathbb{R}[X]$. This shows that even when we are interested in complex roots, we may only work with real polynomials. But it may be more efficient to allow polynomials with complex coefficients (cf. next section). In practice, we assume that $P(X)$ has Gaussian integers $\mathbb{Z}[\mathbf{i}]$ as coefficients.

If $F(X) \in \mathbb{C}[X]$ and $\alpha + \mathbf{i}\beta \in \mathbb{C}$ $(\alpha, \beta \in \mathbb{R})$ is a root of $F(X)$ then we may write

$$F(\alpha + \mathbf{i}\beta) = P(\alpha, \beta) + \mathbf{i}Q(\alpha, \beta)$$

where $P(X, Y), Q(X, Y)$ are bivariate real polynomials determined by F. This reduces the problem of finding α, β to solving the simultaneous system

$$\begin{aligned} P(\alpha, \beta) &= 0, \\ Q(\alpha, \beta) &= 0. \end{aligned}$$

We solve for α using resultants:

$$R(X) := \texttt{res}_Y(P(X, Y), Q(X, Y)).$$

For each real root α of $R(X)$, we can plug α into $P(\alpha, Y)$ to solve for $Y = \beta$. (We have not explicitly described how to handle polynomials with algebraic coefficients but in principle we know how to perform arithmetic operations for algebraic numbers.) Alternatively, we can find β among the real roots of $\texttt{res}_X(P, Q)$ and check for each pair α, β that may serve as a root $\alpha + \mathbf{i}\beta$ of $F(X)$. This will be taken up again in the next section.

It is instructive to examine the above polynomials P, Q in greater detail. To this end, let us write $F(X)$ as

$$F(X) = A(X) + \mathbf{i}B(X), \quad A(X), B(X) \in \mathbb{R}[X].$$

Then by Taylor's expansion,

$$A(\alpha + \mathbf{i}\beta) = A(\alpha) + \frac{A'(\alpha)}{1!} \cdot (\mathbf{i}\beta) + \frac{A''(\alpha)}{2!} \cdot (\mathbf{i}\beta)^2 + \cdots + \frac{A^{(n)}(\alpha)}{n!}(\mathbf{i}\beta)^n$$

where $n = \max\{\deg A, \deg B\}$. Similarly,

$$B(\alpha + \mathbf{i}\beta) = B(\alpha) + \frac{B'(\alpha)}{1!}(\mathbf{i}\beta) + \cdots + \frac{B^{(n)}}{n!}(\mathbf{i}\beta)^n.$$

Hence the real and imaginary parts of $F(\alpha + \mathbf{i}\beta)$ are, respectively,

$$\begin{aligned} P(\alpha,\beta) &= A(\alpha) + \frac{B'(\alpha)}{1!}(-\beta) + \frac{A^{(2)}(\alpha)}{2!}(-\beta^2) + \cdots, \\ Q(\alpha,\beta) &= B(\alpha) + \frac{A'(\alpha)}{1!}(\beta) + \frac{B^{(2)}(\alpha)}{2!}(-\beta^2) + \cdots. \end{aligned}$$

So $P(\alpha,\beta)$ and $Q(\alpha,\beta)$ are polynomials of degree $\leq n$ in β with coefficients that are polynomials in α of degree $\leq n$. Hence $R(\alpha)$ is a polynomial of degree n^2 in α. Moreover, the bit-size of $R(X)$ remains polynomially bounded in the bit-size of $A(X)$, $B(X)$. Hence, any polynomial-time solution to real root isolation would lead to a polynomial-time solution to complex root isolation.

Remarks: See Householder [6] for more details on this approach.

Exercises

Exercise 4.1: Work out the algorithmic details of the two methods for finding complex roots as outlined above. Determine their complexity. □

Exercise 4.2: Express $P(\alpha,\beta)$ and $Q(\alpha,\beta)$ directly in terms of $F^{(i)}(\alpha)$ and β^i by a different Taylor expansion, $F(\alpha + \mathbf{i}\beta) = F(\alpha) + F'(\alpha)(\mathbf{i}\beta) + \cdots$. □

Exercise 4.3: A **Diophantine polynomial** is a polynomial $D(X_1, \ldots, X_n)$ with (rational) integer coefficients and whether the X_i's are integer variables. Hilbert's 10th Problem asks whether a given Diophantine polynomial $D(X_1, \ldots, X_n)$ is solvable. Show that the decidability of Hilbert's 10th Problem is equivalent to the decidability of each of the following problems:
(i) The problem of deciding if a system of Diophantine equations is solvable.
(ii) The problem of deciding if a Diophantine equation of total degree 4 is solvable. **Remark:** It is an unknown problem whether '4' here can be replaced by '3'. HINT: First convert the single Diophantine polynomial to an equivalent system of polynomials of total degree at most 2.
(iii) The problem of deciding if a Diophantine equation of degree 4 has solution in non-negative integers. HINT: In one direction, use the fact that every non-negative integer is the sum of four squares of integers. □

Exercise 4.4: A **Diophantine set of dimension** n is one of the form

$$\{(a_1, \ldots, a_n) \in \mathbb{Z}_n : (\exists b_1, \ldots, b_m \in \mathbb{Z}) D(a_1, \ldots, a_n, b_1, \ldots, b_m) = 0\}$$

where $D(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is a Diophantine polynomial. A Diophantine set $S \subseteq \mathbb{Z}^n$ can be viewed as **Diophantine relation** $R(X_1, \ldots, X_n)$ where $R(a_1, \ldots, a_n)$ holds iff $(a_1, \ldots, a_n) \in S$.
(i) Show that the following relations are Diophantine: $X_1 \neq X_2$, $X_1 = (X_2 \bmod X_3)$, $X_1 = \texttt{GCD}(X_2, X_3)$
(ii) A set $S \subseteq \mathbb{Z}$ is Diophantine iff

$$S = \{D(a_1, \ldots, a_m) : (\exists a_1, \ldots, a_n \in \mathbb{Z}\}$$

for some Diophantine polynomial $D(Y_1, \ldots, Y_m)$.
(iii) Show that Diophantine sets are closed under union and intersection.
(iv) (M.Davis) Diophantine sets are not closed under complement. The complementation is with respect to $\mathbb{Z}^n$ if the dimension is n.
(v) (Y.Matijasevich) The exponentiation relation $X = Y^Z$, where X, Y, Z are restricted to natural numbers, is Diophantine. This is a critical step in the solution of Hilbert's 10th Problem. □

§5. The Routh-Hurwitz Theorem

We now present an alternative method for isolating complex zeros using Sturm's theory. First we consider a special subproblem: to count the number of complex roots in the upper complex plane. This problem has independent interest in the theory of stability of dynamical systems, and was first solved by Routh in 1877, using Sturm sequences. Independently, Hurwitz in 1895 gave a solution based on the theory of residues and quadratic forms. Pinkert [12] exploited this theory to give an algorithm for isolating complex roots. Here, we present a variant of Pinkert's solution.

In this section we consider complex polynomials as well as real polynomials.

We begin with an elementary result, a variant of the so-called *principle!of argument*. Let $F(Z) \in \mathbb{C}[Z]$ and L be an oriented line in the complex plane. Consider the *increase* in the argument of $F(Z)$ as Z moves along the entire length of L, denoted

$$\Delta_L \arg F(Z).$$

Note that if $F = G \cdot H$ then

$$\Delta_L \arg F = (\Delta_L \arg G) + (\Delta_L \arg H). \tag{15}$$

Lemma 8 *Suppose no root of $F(Z)$ lies on L, $p \geq 0$ of the complex roots of $F(Z)$ lie to the left-hand side of L, and $q \geq 0$ of the roots lie to the right-hand side, multiplicity counted. Then $\Delta_L \arg F(Z) = \pi(p - q)$.*

Proof. Without loss of generality, let $F(Z) = \prod_{i=1}^{p+q}(Z - \alpha_i)$, $\alpha_i \in \mathbb{C}$. Then $\arg F(Z) = \sum_{i=1}^{p+q} \arg(Z - \alpha_i)$. Suppose α_i lies to the left of L. Then as Z moves along the entire length of L, $\arg(Z - \alpha_i)$ increases by π *i.e.*, $\Delta_L \arg(Z - \alpha_i) = \pi$. Similarly, if α_i lies to the right of L, $\Delta_L \arg(Z - \alpha_i) = -\pi$. The lemma follows by summing over each root. **Q.E.D.**

Since $p + q = \deg F(Z)$, we conclude:

Corollary 9

$$\begin{aligned} p &= \frac{1}{2}\left[\deg F + \frac{1}{\pi}\Delta_L \arg F(Z)\right], \\ q &= \frac{1}{2}\left[\deg F - \frac{1}{\pi}\Delta_L \arg F(Z)\right]. \end{aligned}$$

Number of roots in the upper half-plane. Our immediate goal is to count the number of roots above the real axis. Hence we now let L be the real axis. By the foregoing, the problem amounts to deriving a suitable expression for $\Delta_L \arg F(Z)$. Since Z is going to vary over the reals, we prefer to use 'X' to denote a real variable. Let

$$F(X) = F_0(X) + \mathbf{i}F_1(X)$$

where $F_0(X), F_1(X) \in \mathbb{R}[X]$. Observe that α is a real root of $F(X)$ iff α is a real root of $G = \texttt{GCD}(F_0, F_1)$. Before proceeding, we make three simplifications:

- We may assume $F_0(X)F_1(X) \neq 0$. If $F_1 = 0$ then the complex roots of $F(X)$ come in conjugate pairs and their number can be determined from the number of real roots. Similarly if $F_0 = 0$ then the same argument holds if we replace F by $\mathbf{i}F$.
- We may assume F_0, F_1 are relatively prime, since we can factor out any common factor $G = \texttt{GCD}(F_0, F_1)$ from F, and and apply equation (15) to F/G and G separately.
- We may assume $\deg F_0 \geq \deg F_1$. Otherwise, we may replace F by $\mathbf{i}F$ which has the same set of roots. This amounts to replacing (F_0, F_1) by $(-F_1, F_0)$ throughout the following.

We define

$$\rho(X) := \frac{F_0(X)}{F_1(X)}.$$

Thus $\rho(X)$ is well-defined for all X (we never encounter $0/0$). Clearly $\arg F(X) = \cot^{-1}\rho(X)$. Let

$$\alpha_1 < \alpha_2 < \cdots < \alpha_k$$

be the real roots of $F_0(X)$. They divide the real axis L into $k+1$ segments,

$$L = L_0 \cup L_1 \cup \cdots \cup L_k, \qquad (L_i = [\alpha_i, \alpha_{i+1}])$$

where $\alpha_0 = -\infty$ and $\alpha_{k+1} = +\infty$. Thus,

$$\Delta_L \arg F(X) = \sum_{i=0}^{k} \Delta_{\alpha_i}^{\alpha_{i+1}} \cot^{-1}\rho(X).$$

Here the notation

$$\Delta_\alpha^\beta f(Z)$$

denotes the increase in the argument of $f(Z)$ as Z moves along the line segment from α to β. Since $F(X)$ has no real roots, $\rho(X)$ is defined for all X (we do not get $0/0$) and $\rho(X) = 0$ iff $X \in \{\alpha_i : i = 1, \ldots, k\}$. We will be examining the signs of $\rho(\alpha_i^-)$ and $\rho(\alpha_i^+)$, and the following graph of the cotangent function is helpful:

Note that $\cot^{-1}\rho(\alpha_i) = \cot^{-1}0 = \pm\pi/2$ (taking values in the range $[-\pi, +\pi]$), and

$$\Delta_{\alpha_i}^{\alpha_{i+1}} \cot^{-1}\rho(X) = \lim_{\epsilon \to 0} \Delta_{\alpha_i+\epsilon}^{\alpha_{i+1}-\epsilon} \cot^{-1}\rho(X).$$

But $\rho(X)$ does not vanish in the interval $[\alpha_i + \epsilon, \alpha_{i+1} - \epsilon]$. Hence for $i = 1, \ldots, k-1$,

$$\begin{aligned} \Delta_{\alpha_i}^{\alpha_{i+1}} \cot^{-1}\rho(X) &= \begin{cases} 0 & \text{if } \rho(\alpha_i^+)\rho(\alpha_{i+1}^-) > 0 \\ \pi & \text{if } \rho(\alpha_i^+) < 0, \quad \rho(\alpha_{i+1}^-) > 0 \\ -\pi & \text{if } \rho(\alpha_i^+) > 0, \quad \rho(\alpha_{i+1}^-) < 0 \end{cases} \\ &= \pi \left[\frac{\texttt{sign}(\rho(\alpha_{i+1}^-)) - \texttt{sign}(\rho(\alpha_i^+))}{2} \right]. \end{aligned} \tag{16}$$

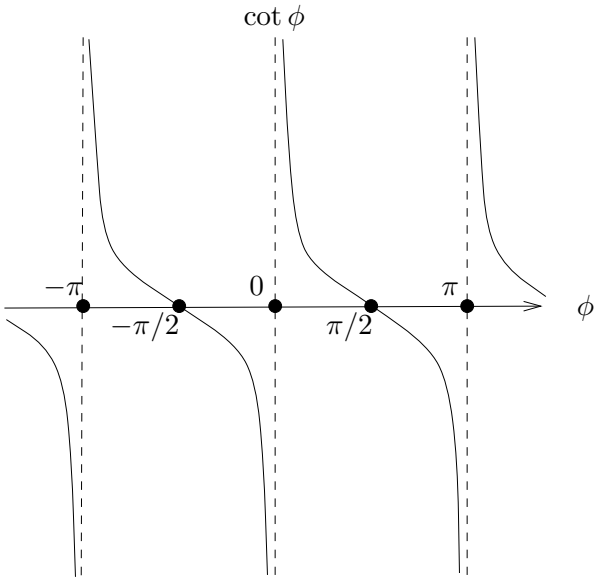


Figure 2: The cotangent function.

This is seen by an examination of the graph of $\cot\phi$. For $i = 0, k$, we first note that if $\deg F_0 > \deg F_1$ then $\rho(-\infty) = \pm\infty$ and $\rho(+\infty) = \pm\infty$. It follows that

$$\begin{aligned}\Delta_{-\infty}^{\alpha_1}\cot^{-1}\rho(X) &= \frac{\pi}{2}\mathtt{sign}(\rho(\alpha_1^-)),\\ \Delta_{\alpha_k}^{+\infty}\cot^{-1}\rho(X) &= -\frac{\pi}{2}\mathtt{sign}(\rho(\alpha_k^+)),\end{aligned}$$

and so

$$\Delta_{-\infty}^{\alpha_1}\cot^{-1}\rho(X) + \Delta_{\alpha_k}^{+\infty}\cot^{-1}\rho(X) = \frac{\pi}{2}\mathtt{sign}(\rho(\alpha_1^-)) - \frac{\pi}{2}\mathtt{sign}(\rho(\alpha_k^+)). \tag{17}$$

If $\deg F_0 = \deg F_1$ then $\rho(-\infty) = \rho(+\infty) = (\mathtt{lead}(F_0))/(\mathtt{lead}(F_1))$ and again (17) holds. Combining equations (16) and (17), we deduce:

Lemma 10

$$\Delta_L \arg F(X) = \pi\sum_{i=1}^{k}\frac{\mathtt{sign}(\rho(\alpha_i^-)) - \mathtt{sign}(\rho(\alpha_i^+))}{2}.$$

But α_i is a pole of $\rho^{-1} = F_1/F_0$. Hence the expression $\frac{\mathtt{sign}(\rho(\alpha_i^-))-\mathtt{sign}(\rho(\alpha_i^+))}{2}$ is the Cauchy index of ρ^{-1} at α_i. By Corollary 7 (§3), this means $-\mathtt{Var}_{F_1,F_0}[-\infty,+\infty]$ gives the Cauchy index of ρ^{-1} over the real line L. Thus $\Delta_L \arg F(X) = -\mathtt{Var}_{F_1,F_0}[-\infty,+\infty]$. Combined with corollary 9, we obtain:

Theorem 11 (Routh-Hurwitz) *Let $F(X) = F_0(X) + \mathbf{i}F_1(X)$ be monic with $\deg F_0 \geq \deg F_1 \geq 0$ and F_0, F_1 relatively prime. The number of roots of $F(X)$ lying above the real axis L is given by*

$$\frac{1}{2}\left(\deg F - \mathtt{Var}_{F_1,F_0}[-\infty,+\infty]\right).$$

To exploit this result for a complex root isolation method, we proceed as follows.

1. Counting Roots to one side of the imaginary axis. Suppose we want to count the number p of roots of $F(Z)$ to the right of the imaginary axis, assuming $F(Z)$ does not have any purely imaginary roots. Note that α is a root of $F(Z)$ to the right of the imaginary axis iff $\mathbf{i}\alpha$ is a root of $F(Z/\mathbf{i}) = F(-\mathbf{i}Z)$ lying above the real axis. It is easy (previous section) to construct the polynomial $G(Z) := F(-\mathbf{i}Z)$ from $F(Z)$.

2. Roots in two opposite quadrants. We can count the number of roots in the first and third quadrant as follows: from $F(Z)$ construct a polynomial $F^*(Z)$ whose roots are precisely the squares of roots of $F(Z)$. This means that α is a root of $F(Z)$ in the first (I) or third (III) quadrant iff α^2 is a root of $F^*(Z)$ in the upper half-plane (which we know how to count). Similarly, the roots of $F(Z)$ in (II) and (IV) quadrants are sent into the lower half-plane. It remains to construct $F^*(Z)$. This is easily done as follows: Let $F(Z) = F_o(Z) + F_e(Z)$ where $F_o(Z)$ consists of those monomials of odd degree and $F_e(Z)$ consisting of those monomials of even degree. This means $F_o(Z)$ is an odd function (*i.e.*, $F_o(-Z) = -F_o(Z)$), and $F_e(Z)$ is an even function (*i.e.*, $F_e(-Z) = F_e(Z)$). Consider

$$\begin{aligned} G(Z) &= F_e(Z)^2 - F_o(Z)^2 \\ &= (F_e(Z) + F_o(Z))(F_e(Z) - F_o(Z)) \\ &= F(Z)(F_e(-Z) + F_o(-Z)) \\ &= F(Z)F(-Z). \end{aligned}$$

If $F(Z) = c\prod_{i=1}^{n}(Z - \beta_i)$ where β_i are the roots of $F(Z)$ then

$$F(Z)F(-Z) = c^2 \prod_{i=1}^{n}(Z - \beta_i)(-Z - \beta_i) = (-1)^n c^2 \prod_{i=1}^{n}(Z^2 - \beta_i^2).$$

Hence, we may define our desired polynomial $F^*(Y)$ by the relation $F^*(Z^2) = G(Z)$. In fact, $F^*(Y)$ is trivially obtained from the coefficients of $G(Z)$.

3. Roots inside a quadrant. We can count the number $\#(I)$ of roots in the first quadrant, since

$$\#(I) = \frac{1}{2}\left[(\#(I) + \#(II)) + (\#(I) + \#(IV)) - (\#(II) + \#(IV))\right]$$

where $\#(I) + \#(II)$ and $\#(I) + \#(IV)$ are half-plane counting queries, and $\#(II) + \#(IV)$ is a counting query for an opposite pair of quadrants. But we have shown how to answer such queries.

4. Roots in a translated quadrant. If the origin is translated to a point $\alpha \in \mathbb{C}$, we can count the number of roots of $F(Z)$ in any of the four quadrants whose origin is at α, by counting the number of roots of $F(Z + \alpha)$ in the corresponding quadrant.

5. Putting these together. In the last section, we have shown how to isolate a sequence $x_1 < x_2 < \cdots < x_k$ of real numbers that contain among them all the real parts of complex roots of $F(Z)$. Similarly, we can isolate a sequence $y_1 < y_2 < \cdots < y_\ell$ of real numbers that contains among them all the imaginary parts of complex roots of $F(Z)$. So finding all roots of $F(Z)$ is reduced to testing if each $x_i + \mathbf{i}y_j$ is a root. We may assume from the root isolation that we know (rational) numbers a_i, b_j such that

$$x_1 < a_1 < x_2 < a_2 < \cdots < a_{k-1} < x_k < a_k, \quad y_1 < b_1 < y_2 < b_2 < \cdots < b_{\ell-1} < y_\ell < b_\ell.$$

Then for $j = 1, \ldots, \ell$ and for $i = 1, \ldots, k$, we determine the number $n(i,j)$ of roots of $F(Z)$ in the quadrant (III) based at $a_i + \mathbf{i}b_j$. Note that $n(1,1) = 1$ or 0 depending on whether $x_1 + \mathbf{i}y_1$ is a root or not. It is easy to work out a simple scheme to similarly determine whether each $x_i + \mathbf{i}y_j$ is a root or not.

Exercises

Exercise 5.1: Determine the complexity of this procedure. Exploit the fact that the testings of the various $x_i + \mathbf{i}y_j$'s are related. □

Exercise 5.2: Isolate the roots of $F(Z) = (Z^2-1)(Z^2+0.16)$ using this procedure. [This polynomial has two real and two non-real roots. Newton iteration will fail in certain open neighborhoods (attractor regions).] □

Exercise 5.3: Derive an algorithm to determine if a complex polynomial has all its roots inside any given circle of the complex plane.
HINT: the transformation $w \mapsto z = r\frac{1+w}{1-w}$ (for any real $r > 0$) maps the half-plane $\mathtt{Re}(w) < 0$ into the open disc $|z| < r$. □

Exercise 5.4: If $F(X)$ is a real polynomial whose roots have no positive real parts then the coefficients of $F(X)$ have no sign variation.
HINT: write $F(X) = \prod_{i=1}^{n}(X - \alpha_i)$ and divide the n roots into the k real roots and 2ℓ complex roots ($n = k + 2\ell$).

□

Exercise 5.5: Let $F_n(X), F_{n-1}(X), \ldots, F_0(X)$ be a sequence of real polynomials where each $F_i(X)$ has degree i and positive leading coefficient. Moreover, $F_i(x) = 0$ implies $F_{i-1}(x)F_{i+1}(x) < 0$ (for $i = 1, 2, \ldots, n-1$, and $x \in \mathbb{R}$). Then each $F_i(X)$ $(i = 1, \ldots, n)$ has i simple real roots and between any two consecutive roots is a root of F_{i-1}. □

Exercise 5.6: (Hermite, Biehler) If all the roots of $F(X) = A(X) + \mathbf{i}B(X)$ $(A(X), B(X) \in \mathbb{R}[X])$ lie on one side of the real axis of the complex plane, then $A(X)$ and $B(X)$ have only simple real roots, and conversely. □

§6. Sign Encoding of Algebraic Numbers: Thom's Lemma

We present an alternative representation of real algebraic numbers as suggested by Coste and Roy [3]. If $\overline{A} = [A_1(X), A_2(X), \ldots, A_m(X)]$ is a sequence[2] of real polynomials, then a *sign condition* of $\overline{A}$ is any sequence of signs,

$$[s_1, s_2, \ldots, s_m], \qquad s_i \in \{-1, 0, +1\}.$$

We say $[s_1, s_2, \ldots, s_m]$ is the *sign condition* of $\overline{A}$ at $\alpha \in \mathbb{R}$ if $s_i = \mathtt{sign}(A_i(\alpha))$ for $i = 1, \ldots, m$. This will be denoted

$$\mathtt{sign}_\alpha(\overline{A}) = [s_1, \ldots, s_m].$$

[2] In this section, we use square brackets '[...]' as a stylistic variant of the usual parentheses '(...)' for writing certain sequences.

A sign condition of $\overline{A}$ is *consistent* if there exists such an α. Define the sequence

$$\mathrm{Der}[A] := [A(X), A'(X), A^{(2)}(X), \ldots, A^{(n)}(X)], \qquad \deg A = n,$$

of derivatives of $A(X) \in \mathbb{R}[X]$. The representation of algebraic numbers is based on the following "little lemma" of Thom. Let us call a subset of $\mathbb{R}$ *simple* if it is empty, a singleton or an open interval.

Lemma 12 (Thom) *Let $A(X) \in \mathbb{R}[X]$ have degree $n \geq 0$ and let $s = [s_0, s_1, \ldots, s_n] \in \{-1, 0, +1\}^{n+1}$ be any sign condition. Then the set*

$$S := \{x \in \mathbb{R} : \mathtt{sign}(A^{(i)}(x)) = s_i, \text{for all } \ i = 0, \ldots, n\}$$

is simple.

Proof. We may use induction on n. If $n = 0$ then $A(X)$ is a non-zero constant and S is either empty or equal to $\mathbb{R}$. So let $n \geq 1$ and let $s' = [s_1, \ldots, s_n]$. Then the set

$$S' := \{x \in \mathbb{R} : \mathtt{sign}(A^{(i)}(x)) = s_i, i = 1, \ldots, n\}$$

is simple, by the inductive hypothesis for $A'(X)$. Note that $S = S' \cap S_0$ where $S_0 := \{x \in \mathbb{R} : \mathtt{sign}(A(x)) = s_0\}$. Now the set S_0 is a disjoint union of simple sets. In fact, viewing $A(X)$ as a continuous real function, S_0 is equal to $A^{-1}(0)$, $A^{-1}(\mathbb{R}_{>0})$ or $A^{-1}(\mathbb{R}_{<0})$, depending on whether $s_0 = 0, +1$ or -1. In any case, we see that if $S' \cap S_0$ is a connected set, then it is simple. So assume it is disconnected. Then S' contains two distinct roots of $A(X)$. By Rolle's theorem (§VI.1), $A'(X)$ must have a root in S'. This implies S' is contained in the set $\{x \in \mathbb{R} : \mathtt{sign}(A'(x)) = 0\}$, which is a finite set. Since S' is connected, it follows that S' is empty or a singleton. This contradicts the assumption that $S' \cap S_0$ is disconnected. **Q.E.D.**

Lemma 13 *Let α, β be distinct real roots of $A(X)$, $\deg A(X) = n \geq 2$. Let $s = [s_0, \ldots, s_n]$ and $s' = [s'_0, \ldots, s'_n]$ be the sign conditions of Der$[A]$ at α and at β (respectively).*
(i) s and s' are distinct.
(ii) Let i be the largest index such that $s_i \neq s'_i$. Then $0 < i < n$ and $s_{i+1} = s'_{i+1} \neq 0$. Furthermore, $\alpha < \beta$ iff one of the following conditions holds:

$$(a) \qquad s_{i+1} = +1 \text{ and } s_i < s'_i;$$
$$(b) \qquad s_{i+1} = -1 \text{ and } s_i > s'_i.$$

Proof. Let I be the open interval bounded by α, β.
(i) If $s = s'$ then by Thom's lemma, every $\gamma \in I$ also achieves the sign condition s. In particular, this means $A(\gamma) = 0$. Since there are infinitely many such γ, $A(X)$ must be identically zero, contradiction.
(ii) It is clear that $0 < i < n$ since $s_0 = s'_0 = 0$ and $s_n = s'_n$. Thom's lemma applied to the polynomial $A^{(i+1)}(X)$ implies that $A^{(i+1)}(\gamma)$ has constant sign throughout the interval I. If $s_{i+1} = s'_{i+1} = 0$ then we obtain the contradiction that $A^{(i+1)}(X)$ is identically zero in I. So suppose $s_{i+1} = s'_{i+1} = +1$ (the other case being symmetrical). Again by Thom's lemma, we conclude that $A^{(i+1)}(\gamma) > 0$ for all $\gamma \in I$, *i.e.*, $A^{(i)}(X)$ is strictly increasing in I. Thus $\alpha < \beta$ iff

$$A^{(i)}(\alpha) < A^{(i)}(\beta). \tag{18}$$

Since the signs of $A^{(i)}(\alpha)$ and $A^{(i)}(\beta)$ are distinct, the inequality (18) amounts to $s_i < s'_i$. **Q.E.D.**

This result suggests that we code a real algebraic number α by specifying a polynomial $A(X)$ at which α vanishes, and by specifying its sign condition at $\mathrm{Der}[A']$, written

$$\alpha \cong (A(X), \mathtt{sign}(\mathrm{Der}[A'])).$$

This is the same notation ($\cong$) used when α is represented by an isolating interval (§VI.9), but it should not lead to any confusion. We call $(A(X), \mathtt{sign}(\mathrm{Der}[A']))$ a *sign encoding* of α. For example, $\sqrt{2} \cong (X^2 - 2, [+1, +1])$ and $-\sqrt{2} \cong (X^2 - 2, [-1, +1])$.

This encoding has some advantages over the isolating interval representation in that, once A is fixed, the representation is unique (and we can make A unique by choosing the distinguished minimal polynomial of α). It's discrete nature is also desirable. On the other hand, the isolating intervals representation gives an explicit numerical approximation, which is useful. Coste and Roy [3] also generalized the sign encoding to the multivariate situation.

Exercises

Exercise 6.1: Let $s = [s_0, \ldots, s_n]$ be a sequence of *generalized sign condition* that is, s_i belongs to the set $\{< 0, \leq 0, 0, \geq 0, > 0\}$ of generalized signs (rather than $s_i \in \{-1, 0, +1\}$). If $A(X)$ has degree $n \geq 0$, show that the set $\{x \in \mathbb{R} : s = \mathtt{sign}_x(\mathrm{Der}[A])\}$ is connected (possibly empty). □

Exercise 6.2: Give an algorithm to compare two arbitrary real algebraic numbers in this representation. □

§7. Problem of Relative Sign Conditions

Uses of the sign encoding of real algebraic numbers depend on a key algorithm from Ben-Or, Kozen and Reif [1]. This algorithm has come to be known as the "BKR algorithm". We first describe the problem solved by this algorithm.

Let $\overline{B} = [B_1, B_2, \ldots, B_m]$ be a sequence of real polynomials, and A another real polynomial. A sign condition $s = [s_1, \ldots, s_m]$ of $\overline{B}$ is *consistent relative to* A (or, A-*consistent*) if $[0, s_1, \ldots, s_m]$ is consistent for the sequence $[A, B_1, \ldots, B_m]$. In other words, s is A-consistent if $s = \mathtt{sign}_\alpha[\overline{B}]$ for some root α of A. The *weight of* s *relative to* A is the number of roots of A at which $\overline{B}$ achieves the sign condition s. Thus s is relatively consistent iff $[0, s_1, \ldots, s_m]$ has positive weight. If A is understood, we may simply call s a *relatively consistent sign condition* of $\overline{B}$.

The *problem of relative sign consistency*, on input $A, \overline{B}$, asks for the set of all A-consistent sign conditions of $\overline{B}$; a stronger version of this problem is to further ask for the weight of each A-consistent sign condition.

There are numerous other applications of this problem, but we can see immediately its applications to the sign encoding representation:

- To determine the sign encoding of all roots of $A(X)$, it suffices to call the BKR algorithm on $A, \overline{B}$ where $\overline{B} = \mathrm{Der}[A']$.

- To determine the sign of a polynomial $P(X)$ at the roots of A, we call BKR on $A, \overline{B}$ where $\overline{B} = [P, A', A^{(2)}, \ldots, A^{(m-1)}]$.

The original BKR algorithm is described only for the case where $A, B_1, \ldots, B_m$ are relatively prime, as the general case can be reduced to this special case. Still, it is convenient to give a direct algorithm. Mishra and Pedersen [10] observed that corollary 6 used in the original BKR algorithm in fact holds without any conditions on the polynomials A, B:

Lemma 14 *Let $A, B \in \mathbb{R}[X]$ such that $A(\alpha)A(\beta) \neq 0$, $\alpha < \beta$. Then*

$$\mathtt{Var}_{A,A'B}[\alpha, \beta] = \sum_{\gamma} \mathtt{sign}(B(\gamma))$$

where γ ranges over the distinct real roots of A.

Proof. Again, it suffices to prove this for a fundamental interval $[\alpha, \beta]$ at some $\gamma_0 \in [\alpha, \beta]$. Let γ_0 be an r-fold root of A and an s-fold root of $A'B$. If $r \geq s$, then this has been proved in corollary 6. So assume $s > r$. The sign variation difference over $[\alpha, \beta]$ in the Sturm sequence $[A_0, A_1, \ldots, A_h]$ for $A, A'B$ is evidently equal to that in the depressed sequence $[A_0/A_h, A_1/A_h, \ldots, 1]$. But the sign variation difference in the depressed sequence is 0 since γ_0 is a non-root of A_0/A_h (here we use the fact that γ_0 is an r-fold root of A_h). Since $B(\gamma_0) = 0$ (as $s > r$), we have verified

$$\mathtt{Var}_{A,A'B}[\alpha, \beta] = 0 = \mathtt{sign}(B(\gamma_0)).$$

Q.E.D.

In the following, we fix A and $\overline{B} = [B_1, \ldots, B_m]$. If ε is a sign condition of $\overline{B}$, write

$$W^{\varepsilon} := \{\alpha : A(\alpha) = 0, \mathtt{sign}_{\alpha}[\overline{B}] = \varepsilon\} \tag{19}$$

for the set of real roots α of A at which $\overline{B}$ achieves the condition ε. So the weight of ε is given by

$$w^{\varepsilon} := |W^{\varepsilon}|.$$

For instance, when $m = 1$, the roots of A are partitioned into W^0, W^+, W^-. When $m = 3$, w^{+-0} is the number of roots of A at which B_1 is positive, B_2 is negative and B_3 vanishes.

So the BKR algorithm amounts to determining these weights. First consider some initial cases of the BKR algorithm (for small m).

CASE $m = 0$: In this case, the A-consistent sign condition is [ ] (the empty sequence) and its weight is (by definition) just the number of real roots of A. By the original Sturm theorem (§3), this is given by

$$v_A(1) := \mathtt{Var}_{A,A'}[-\infty, +\infty].$$

In general, we shall abbreviate $\mathtt{Var}_{A,A'B}[-\infty, +\infty]$ by $v_A(B)$, or simply, $v(B)$ if A is understood. In this context, computing $v(B)$ is sometimes called "making a Sturm query on B".

CASE $m = 1$: By the preceding lemma,

$$v_A(B_1) = w^+ - w^-, \qquad v_A(B_1^2) = w^+ + w^-.$$

Case $m = 0$ shows that

$$v_A(1) = w^0 + w^+ + w^-.$$

We put these together in the matrix format,

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} w^0 \\ w^+ \\ w^- \end{bmatrix} = \begin{bmatrix} v(1) \\ v(B_1) \\ v(B_1^2) \end{bmatrix}. \tag{20}$$

Thus we can solve for w^0, w^+, w^- since we know the right hand side after making the three Sturm queries $v(1), v(B_1), v(B_1^2)$.

CASE $m = 2$: If we let M_1 be the matrix in equation (20), it is not hard to verify

$$\begin{bmatrix} M_1 & M_1 & M_1 \\ \mathbf{0} & M_1 & -M_1 \\ \mathbf{0} & M_1 & M_1 \end{bmatrix} \cdot \begin{bmatrix} w^{00} \\ w^{0+} \\ w^{0-} \\ w^{+0} \\ w^{++} \\ w^{+-} \\ w^{-0} \\ w^{-+} \\ w^{--} \end{bmatrix} = \begin{bmatrix} v(1) \\ v(B_1) \\ v(B_1^2) \\ v(B_2) \\ v(B_1B_2) \\ v(B_1^2B_2) \\ v(B_2^2) \\ v(B_1B_2^2) \\ v(B_1^2B_2^2) \end{bmatrix}. \tag{21}$$

Again, we can solve for the weights after making some Sturm queries. The case $m = 2$ will illustrate the general development of the BKR algorithm below. If the square matrix in (21) is denoted M_2 then M_2 can be viewed as the "Kronecker product" of M_1 with itself.

Exercises

Exercise 7.1: Let α have the sign encoding $E = (A(X), [s_1, \ldots, s_m])$.
(i) What is the sign encoding of $-\alpha$ in terms of E?
(ii) Give a method to compute the sign encoding E' of $1/\alpha$. Assume that the polynomial in E' is $X^m A(1/X)$. HINT: consider $\mathrm{Der}[A](1/X)$ instead of $\mathrm{Der}[X^m A(1/X)]$. □

§8. The BKR algorithm

We now develop the BKR algorithm.

Let $M \in R^{m\times n}$ and $M' \in R^{m'\times n'}$ where R is any ring. The *Kronecker product* $M \otimes M'$ of M and M' is the $mm' \times nn'$ matrix partitioned into $m \times n$ blocks, with the (i, j)th block equal to

$$(M)_{ij} \cdot M'.$$

In other words, $M \otimes M'$ is defined by

$$\begin{aligned}(M \otimes M')_{(i-1)m'+i',(j-1)m'+j'} &= M_{ij}M_{i'j'},\\ i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}, &i' \in \{1,\ldots,m'\}, j' \in \{1,\ldots,n'\}.\end{aligned}$$

For instance, the matrix M_2 in (21) can be expressed as $M_1 \otimes M_1$. Again, if u, u' are m-vectors and m'-vectors, respectively, then $u \otimes u'$ is a (mm')-vector.

Lemma 15 *Let $M \in R^{m\times m}$ and $M' \in R^{m'\times m'}$ and u, u' be m-vectors and m'-vectors, respectively.*
(i) $(M \otimes M')(u \otimes u') = (Mu) \otimes (M'u')$.
(ii) If M, M' are invertible, so is $M \otimes M'$, with inverse $M^{-1} \otimes M'^{-1}$.

Proof. (i) This is a straightforward exercise.
(ii) Consider the action of the matrix product $(M^{-1} \otimes M'^{-1}) \cdot (M \otimes M')$ on $u \otimes u'$:

$$\begin{aligned}(M^{-1} \otimes M'^{-1}) \cdot (M \otimes M') \cdot u \otimes u' &= (M^{-1} \otimes M'^{-1}) \cdot (M \cdot u \otimes M' \cdot u') \\ &= (M^{-1} \cdot M \cdot u) \otimes (M'^{-1} \cdot M' \cdot u'). \\ &= u \otimes u'.\end{aligned}$$

As u, u' are arbitrary, this proves that $(M^{-1} \otimes M'^{-1}) \cdot (M \otimes M')$ is the identity matrix. **Q.E.D.**

The real algebra of vectors. We describe the BKR algorithm by "shadowing" its action in the ring $R = \mathbb{R}^k$ of k-vectors over $\mathbb{R}$. This notion of shadowing will be clarified below; but it basically makes the correctness of the algorithm transparent.

Note that $R = \mathbb{R}^k$ is a ring under component-wise addition and multiplication. The real numbers $\mathbb{R}$ are embedded in R under the correspondence $\alpha \in \mathbb{R} \mapsto (\alpha, \alpha, \ldots, \alpha) \in R$. Thus R is a real algebra[3].

To describe the BKR algorithm on inputs $A(X)$ and $\overline{B} = [B_1, \ldots, B_m]$, we first choose the k in the definition of R to be the number of distinct real roots of the polynomial $A(X)$; let these roots be

$$\overline{\alpha} = (\alpha_1, \ldots, \alpha_k). \tag{22}$$

We shall use R in two distinct ways:

- A vector in R with entries from $-1, 0, +1$ will be called a *root sign vector*. Such vectors[4] represent the signs of a polynomial $Q(X)$ at the k real roots of $A(X)$ in the natural way:

  $$\mathtt{sign}_{A(X)}(Q(X)))$$

  denotes the sign vector $[s_1, \ldots, s_k]$ where $s_i = \mathtt{sign}(Q(\alpha_i))$. If $s_i = \mathtt{sign}_A(Q_i)$ $(i = 0, 1)$ then notice that $s_0 \cdot s_1 = \mathtt{sign}_A(Q_0 Q_1)$.

  In the BKR algorithm, Q will be a power product of $B_1, \ldots, B_m$.

- A 0/1 vector in R will be called a *Boolean vector*. Such a vector u represents a subset U of the roots of $A(X)$ in the natural way: the i-th component of u is 1 iff $\alpha_i \in U$. If the Boolean vectors $u_0, u_1 \in R$ represent the subsets U_0, U_1 (respectively) then observe that $U_0 \cap U_1$ is represented by the vector product $u_0 \cdot u_1$.

  In the BKR algorithm, the subsets U are determined by sign conditions of $\overline{B}$: such subsets have the form W^ε (see equation (19)) where $\varepsilon = [s_1, \ldots, s_\ell]$ is a sign condition of $\overline{C} = [C_1, \ldots, C_\ell]$ and $\overline{C}$ is a subsequence of $\overline{B}$. Note that ε is not to be confused with the root sign vectors in R. In fact, we define a rather different product operation on such sign conditions: let $\varepsilon = [s_1, \ldots, s_\ell]$ be a sign condition of $\overline{C} = [C_1, \ldots, C_\ell]$ and $\varepsilon' = [s_{\ell+1}, \ldots, s_{\ell'}]$ be a sign condition of $\overline{C}' = [C_{\ell+1}, \ldots, C_{\ell'}]$, $\ell < \ell'$. Assuming that $\overline{C}$ and $\overline{C}'$ are disjoint, we define

  $$\varepsilon \cdot \varepsilon' := [s_1, \ldots, s_\ell, s_{\ell+1}, \ldots, s_{\ell'}],$$

  i.e., the concatenation of ε with ε'. This definition of product is consistent with the product in R in the following sense: if $u_0, u_1 \in R$ represent $W^\varepsilon, W^{\varepsilon'}$ (respectively) then $u_0 \cdot u_1$ (multiplication in R) represents

  $$W^{\varepsilon \cdot \varepsilon}.$$

[3] In general, a ring R containing a subfield K is called a K-algebra.

[4] Although root sign vectors are formally sign conditions, notice that root sign vectors arise quite differently, and hence the new terminology. By the same token, Boolean vectors are formally a special type of sign condition, but they are interpreted very differently.

We come to a key definition: let $\overline{C} = [C_1, \ldots, C_\ell]$ be a subsequence of $\overline{B}$. Let $M \in \mathbb{R}^{\ell \times \ell}$, $\overline{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_\ell]$ where each ε_i is a sign condition for $\overline{C} = [C_1, \ldots, C_\ell]$, and $\overline{Q} = [Q_1, \ldots, Q_\ell]$ be a sequence of real polynomials. We say that

$$(M, \overline{\varepsilon}, \overline{Q})$$

is a *valid triple* for $\overline{C}$ if the following conditions hold:

- M is invertible.
- Every A-consistent sign condition for $\overline{C}$ occurs in $\overline{\varepsilon}$ (so $\overline{\varepsilon}$ may contain relatively inconsistent sign conditions).
- The equation
$$M \cdot u = s \tag{23}$$
holds in R where $u = (u_1, \ldots, u_\ell)^T$ with each u_i a Boolean vector representing W^{ε_i}, and $s = (s_1, \ldots, s_\ell)^T$ with s_i equal to the root sign vector $\mathtt{sign}_A(Q_i) \in R$. Equation (23) is called the *underlying equation* of the triple.

We can view the goal of the BKR algorithm to be the computation of valid triples for $\overline{B}$ (note that A is implicit in our definition of valid triples).

Example: $(M_1, ([0], [+], [-]), [1, B_1, B_1^2])$ is a valid triple for $\overline{B} = [B_1]$. The underlying equation is

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} u^0 \\ u^+ \\ u^- \end{bmatrix} = \begin{bmatrix} \mathtt{sign}_A(1) \\ \mathtt{sign}_A(B_1) \\ \mathtt{sign}_A(B_1^2) \end{bmatrix}. \tag{24}$$

where we write u^0, u^+, u^- for the Boolean vectors representing the sets W^0, W^+, W^-. Compare this equation to equation (20).

We define the "Kronecker product" of two triples $(M, \overline{\varepsilon}, \overline{Q})$ and $(M', \overline{\varepsilon}', \overline{Q}')$ as

$$(M \otimes M', \overline{\varepsilon} \otimes \overline{\varepsilon}', \overline{Q} \otimes \overline{Q}')$$

where the underlying "multiplication" in $\overline{\varepsilon} \otimes \overline{\varepsilon}'$ and $\overline{Q} \otimes \overline{Q}'$ are (respectively) concatenation of sign conditions and multiplication of polynomials. For example,

$$(0, +, -) \otimes (+-, -0) = (0+-, 0-0, ++-, +-0, -+-, --0)$$

and

$$[Q_1, Q_2] \otimes [Q_3, Q_4] = [Q_1Q_3, Q_1Q_4, Q_2Q_3, Q_3Q_4].$$

Lemma 16 *Suppose $(M, \overline{\varepsilon}, \overline{Q})$ is valid for $[B_1, \ldots, B_\ell]$ and $(M', \overline{\varepsilon}', \overline{Q}')$ is valid for $[B_{\ell+1}, \ldots, B_{\ell+\ell'}]$. Then*

$$(M \otimes M', \overline{\varepsilon} \otimes \overline{\varepsilon}', \overline{Q} \otimes \overline{Q}') \tag{25}$$

is valid for $[B_1, \ldots, B_\ell, B_{\ell+1}, \ldots, B_{\ell+\ell'}]$.

Proof. (i) First we note that $M \otimes M'$ is invertible.
(ii) Next note that every A-consistent sign condition for $[B_1, \ldots, B_{\ell+\ell'}]$ is listed in $\overline{\varepsilon} \otimes \overline{\varepsilon}'$.
(iii) Let the underlying equations of $(M, \overline{\varepsilon}, \overline{Q})$ and $(M', \overline{\varepsilon}', \overline{Q}')$ be $M \cdot u = s$ and $M' \cdot u' = s'$, respectively. By lemma 15(i),

$$(M \otimes M')(u \otimes u') = s \otimes s'. \tag{26}$$

Then it remains to see that equation (26) is the underlying equation for equation (25). This follows since for each i, $(u \otimes u')_i$ represents the set $W^{(\overline{\varepsilon} \otimes \overline{\varepsilon}')_i}$, and $(s \otimes s')_i = \mathtt{sign}_A((\overline{Q} \otimes \overline{Q}')_i)$. **Q.E.D.**

Pruning. It follows from this lemma that

$$(M_2, ([0],[+],[-]) \otimes ([0],[+],[-]), [1, B_1, B_1^2] \otimes [1, B_2, B_2^2])$$

is a valid triple for $[B_1, B_2]$. We can repeat this formation of Kronecker product m times to obtain a valid triple $(M, \overline{\varepsilon}, \overline{Q})$ for $[B_1, \ldots, B_m]$. But the size of the matrix M would be $3^m \times 3^m$, which is too large for practical computation. This motivates the idea of "pruning". Observe that the number of A-consistent sign conditions cannot be more than k. This means that in the underlying equation $Mu = s$, all but k of the Boolean vectors $(u)_i$ must be the zero vector $\mathbf{0}$ (representing the empty set). The following steps reduces the matrix M to size at most $k \times k$:

Pruning Procedure for the equation $Mu = s$:

1. Detect and eliminate the zero vectors in u.
   Call the resulting vector u'.
   So the length of u' is ℓ where $\ell \leq k$.
2. Omit the columns in M corresponding to eliminated entries of u.
   We get a new matrix M'' satisfying $M''u' = s$.
3. Since M is invertible, find ℓ rows in M'' that form
   an invertible $\ell \times \ell$ matrix M'.
4. If s' are the entries corresponding to these rows,
   we finally obtain the "pruned equation" $M'u' = s'$.

After we have pruned the underlying equation of the valid triple $(M, \overline{\varepsilon}, \overline{Q})$, we can likewise "prune" the valid triple to a new triple $(M', \overline{\varepsilon}', \overline{Q}')$ whose underlying equation is $M'u' = s'$. It is not hard to verify that that this new triple is valid. The resulting matrix M' has size at most $k \times k$.

Shadowing. The Pruning Procedure above is not intended to be effective because we have no intention of computing over R. Instead, we apply the linear map

$$\lambda : R \to \mathbb{R}$$

defined by $\lambda(x) = \sum_{i=1}^{k} x_i$ for $x = (x_1, \ldots, x_k)$. Notice

- If x is a Boolean vector representing W^{ε} then $\lambda(x) = w^{\varepsilon}$.
- If x is a root sign condition for a polynomial Q then $\lambda(x) = v_A(Q)$, a Sturm query on Q.

If $u \in R^{\ell}$, then $\lambda(u) \in \mathbb{R}^{\ell}$ is defined by applying λ component-wise to u. The underlying equation is transformed by λ into the real matrix equation,

$$M \cdot \lambda(u) = \lambda(s).$$

This equation is only a "shadow" of the underlying equation, but we can effectively compute with this equation. More precisely, we can compute $\lambda(s)$ since it is just a sequence of Sturm queries:

$$\lambda(s) = (v_A(Q_1), \ldots, v_A(Q_{\ell}))^T$$

where $\overline{Q} = (Q_1, \ldots, Q_{\ell})$. From this, we can next compute $\lambda(u)$ as $M^{-1} \cdot \lambda(s)$. The A-inconsistent sign conditions in $\overline{\varepsilon}$ correspond precisely to the 0 entries in $\lambda(u)$. Thus step 1 in the Pruning

Procedure can be effectively carried out. The remaining steps of the Pruning Procedure can now be carried out since we have direct access to the matrix M (we do not need u or s). Finally we can compute the pruned valid triple.

All the ingredients for the BKR algorithm are now present:

BKR Algorithm

Input: $A(X)$ and $\overline{B} = [B_1, \ldots, B_m]$.

Output: a valid triple $(M, \overline{\varepsilon}, \overline{Q})$ for $\overline{B}$.

1. If $m = 1$, we output $(M_1, ([0], [+], [-]), (1, B_1, B_1^2))$ as described above.
2. If $m \geq 2$, recursively compute $(M', \overline{\varepsilon}', \overline{Q}')$ valid for $[B_1, \ldots, B_\ell]$ $(\ell = \lfloor m/2 \rfloor)$, and also $(M'', \overline{\varepsilon}'', \overline{Q}'')$ valid for $[B_{\ell+1}, \ldots, B_m]$.
3. Compute the Kronecker product of $(M', \overline{\varepsilon}', \overline{Q}')$ and $(M'', \overline{\varepsilon}'', \overline{Q}'')$.
4. Compute and output the pruned Kronecker product.

The correctness of this algorithm follows from the preceding development. The algorithm can actually be implemented efficiently using circuits.

Mishra and Pedersen [10] describe extensions of this algorithm useful for various operations on sign encoded numbers.

References

[1] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[2] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[3] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[4] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[5] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[6] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[7] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[8] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[9] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[10] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[11] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[12] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[13] S. M. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[14] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[15] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[16] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[17] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

Contents

Lecture VIII
Gaussian Lattice Reduction

The subject known as the geometry of numbers was initiated by Minkowski. Its power and elegance comes from converting algebraic problems into a geometric setting (which, we might say, is an inversion of the program of Descartes to algebraize geometry). The central object of study in the geometry of numbers is lattices. Cassels [40] (see also [191]) gives a classical treatment of this subject; recent development may be found in the book of Grötschel, Lovász and Schrijver [75]. H. W. Lenstra (1983) first introduced these methods to complexity theory, leading to a polynomial-time algorithm for integer programming in fixed dimensions. Current polynomial-time algorithms for factoring integer polynomials also depend on lattice-theoretic techniques. A key ingredient in these major results are efficient algorithms for lattice reduction. General lattice reduction and factoring of integer polynomials will be treated in the next lecture. In this lecture, we introduce lattice reduction by focusing on 2-dimensional lattices. Here, an algorithm of Gauss lays claim to being the natural extension of Euclid's algorithm to 2-dimensions. The algorithm originally arises in Gauss's theory of reduction of integral binary quadratic forms [41, 177, 108]. See [206, 207, 218] for some recent work on the Gaussian algorithm. Note that we use "Gaussian algorithm" to refer to the 2-dimensional case only, although there are some higher dimensional analogues.

§1. Lattices

This section gives a general introduction to lattices.

Fix $d \geq 1$. Let $S \subseteq \mathbb{R}^d$ be a non-empty finite set. The *lattice generated by* S is the set of integer linear combinations of the elements in S,

$$\Lambda = \Lambda(S) := \{m_1 u_1 + m_2 u_2 + \cdots + m_k u_k \ : \ k \geq 1, u_i \in S, m_i \in \mathbb{Z}\}.$$

The set S is called a *generating set* for the lattice Λ. If S has the minimum cardinality among generating sets for Λ, we call S a *basis* of Λ. The cardinality of a basis of Λ is the *dimension*, $\dim \Lambda$, of Λ. Instead of $\Lambda(S)$, we also write $\mathbb{Z}u$ if $S = \{u\}$; or

$$\Lambda(u_1, \ldots, u_k) = \mathbb{Z}u_1 + \mathbb{Z}u_2 + \cdots + \mathbb{Z}u_k$$

if $S = \{u_1, \ldots, u_k\}$.

Even for $d = 1$, the dimension of a lattice can be arbitrarily large or even infinite. But in our applications, it is sufficient and customary to restrict Λ to the case where $u_1, \ldots, u_k$ are linearly independent as real vectors. In this case, $1 \leq k \leq d$. Viewing S as an ordered sequence $(u_1, \ldots, u_k)$ of vectors, we let

$$A = [u_1, \ldots, u_k] \in \mathbb{R}^{d \times k}$$

denote a $d \times k$ real matrix, and write $\Lambda(A)$ instead of $\Lambda(S)$. Under the said customary convention, A has matrix rank k. We say $\Lambda(A)$ is *full-dimensional* iff $k = d$. Our applications require lattices that are not full-dimensional.

A lattice Λ with only integer coordinates, $\Lambda \subseteq \mathbb{Z}^d$, is called an *integer lattice*. The simplest example of a lattice is the *unit integer lattice* $\Lambda = \mathbb{Z}^d$. A basis for this lattice is the set $S = \{e_1, \ldots, e_d\}$ of elementary vectors in $\mathbb{R}^d$ (equivalently, the identity matrix $E = [e_1, \ldots, e_d]$ is a basis). If we replace any e_i by the vector consisting of all 1's, we get another basis for $\mathbb{Z}^d$.

We examine the conditions for two bases A, B to generate the same lattice. If U is a $k \times k$ real non-singular matrix, we can transform a basis A to AU.

Definition: A square matrix $U \in \mathbb{C}^{k\times k}$ is *unimodular* if $\det U = \pm 1$. A *real*, *integer*, etc, unimodular matrix is one whose entries are all real, all integer, etc.

A unimodular[1] matrix U represents a unimodular transformation of lattice bases, $A \mapsto AU$. Note that the inverse of a (real or integer, respectively) unimodular matrix is still (real or integer) unimodular. The next theorem shows why we are interested in integer unimodular matrices.

Theorem 1 *Let $A, B \in R^{d\times k}$ be two bases. Then $\Lambda(A) = \Lambda(B)$ iff there exists an integer unimodular matrix U such that $A = BU$.*

Proof. ($\Rightarrow$) Since each column of A is in $\Lambda(B)$, there is an integer matrix U_A such that

$$A = BU_A.$$

Similarly, there is an integer matrix U_B such that

$$B = AU_B.$$

Hence $A = AU_BU_A$. If A' is a $k \times k$ submatrix of A such that $\det A' \neq 0$, then $A' = A'U_BU_A$ shows that $\det(U_BU_A) = 1$. Since U_A, U_B are integer matrices this implies $|\det U_A| = |\det U_B| = 1$.

($\Leftarrow$) If $A = BU$ then $\Lambda(A) \subseteq \Lambda(B)$. But since $B = AU^{-1}$, $\Lambda(B) \subseteq \Lambda(A)$. **Q.E.D.**

Definition: The *determinant* of a lattice Λ is given by

$$\det \Lambda := \sqrt{\det A^T A}$$

where A is any basis with $\Lambda(A) = \Lambda$.

By definition, the determinant of a lattice is always positive. Using the previous theorem, it is easy to show that $\det \Lambda$ is well-defined: if $A = BU$ for some unimodular matrix U (this demonstration does not depend on U being integer) then

$$\begin{aligned} \det A^T A &= \det U^T B^T BU = \det(U^T)\det(B^T B)\det(U) \\ &= \det B^T B. \end{aligned}$$

Geometrically, $\det \Lambda$ is the smallest volume of a parallelepiped formed by k independent vectors of Λ ($k = \dim \Lambda$). For instance, the unit integer lattice has determinant 1. The reader may also verify that $\Lambda(u, v) = \mathbb{Z}^2$ where $u = (2, 1)^T, v = (3, 2)^T$. Note that $\det[u, v] = 1$.

It is easy to check that given any basis $A = [a_1, \ldots, a_n]$, the following transformations of A are unimodular transformations:

(i) Multiplying a column of A by -1:

$$A' = [a_1, \ldots, a_{i-1}, -a_i, a_{i+1}, \ldots, a_n].$$

(ii) Adding a constant multiple c of one column to a different column:

$$A' = [a_1, \ldots, a_j, \ldots, a_i + ca_j, \ldots, a_n].$$

[1] Unimodular literally means "of modulus one". The terminology is also used, for instance, to refer to complex numbers $z = x + y\mathbf{i}$ where $|z| = \sqrt{x^2 + y^2} = 1$.

(iii) Permuting two columns of A:

$$A' = [a_1, \ldots, a_{j-1}, a_i, a_{j+1}, \ldots, a_{i-1}, a_j, a_{i+1}, , \ldots, a_n].$$

It is important that $i \neq j$ in (ii). We call these the *elementary column operations*. There is clearly an analogous set of *elementary row operations*. Together, they are called the *elementary unimodular transformations*. If c in (ii) is an integer, then (i), (ii) and (iii) constitute the elementary *integer* row operations.

The unimodular matrices corresponding to the elementary transformations are called *elementary unimodular matrices*. We leave it as an exercise to describe these elementary unimodular matrices explicitly.

A fundamental result which we will not prove here (but see [86, p. 382]) is that the group of unimodular matrices in $\mathbb{Z}^{n\times n}$ can be generated by the following three matrices:

$$U_0 = \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \; U_1 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & (-1)^{n-1} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \; U_2 = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

It is easy to see that U_0, U_1, U_2 are each a product of elementary unimodular transformations. We conclude: *a matrix is unimodular iff it is a product of the elementary unimodular transformations.*

Short vectors. Let $|u|$ denote the *(Euclidean) length* of $u \in \mathbb{R}^d$. So $|u| := \|u\|_2$, in our general notation. When $d = 2$, this notation conveniently coincides with the absolute value of u as a complex number. The *unit vector* along direction u is $\widehat{u} := u/|u|$. Scalar product of u, v is denoted by $\langle u, v\rangle$. We have the basic inequality

$$|\langle u, v\rangle| \leq |u| \cdot |v|. \tag{1}$$

Note that the zero vector $\mathbf{0}$ is always an element of a lattice. We define $u \in \Lambda$ to be a *shortest vector* in Λ if it has the shortest length among the non-zero vectors of Λ. More generally, we call a sequence

$$(u_1, u_2, \ldots, u_k), \qquad k \geq 1$$

of vectors a *shortest k-sequence* of Λ if for each $i = 1, \ldots, k$, u_i is a shortest vector in the set $\Lambda \setminus \Lambda(u_1, u_2, \ldots, u_{i-1})$. We call[2] a vector a kth *shortest vector* if it appears as the kth entry in some shortest k-sequence. Clearly $k \leq \dim \Lambda$. For instance, if u, v are both shortest vectors and are independent, then $(\pm u, \pm v)$ and $(\pm v, \pm u)$ are shortest sequences and so both u, v are 2nd shortest vectors. We will not distinguish u from $-u$ when discussing shortest vectors. So we may say u is the *unique* ith shortest vector if u and $-u$ are the only ith shortest vectors. In a 2-dimensional lattice Λ, we will see that the shortest 2-sequence forms a basis for Λ. Hence we may speak of a *shortest basis* for Λ. But in higher dimensions, a shortest k-sequence (where $k > 2$ is the dimension of the lattice) need not form a basis of the lattice (Exercise).

A fundamental computational problem in lattices is to compute another basis B for a given lattice $\Lambda(A)$ consisting of "short" vectors. The present lecture constructs an efficient algorithm in the two dimensional case. The general case will be treated in the subsequent lecture.

EXERCISES

[2]Evidently, this terminology can be somewhat confusing. For instance, the 2nd shortest vector is not always what you expect.

Exercise 1.1: Show that there exist lattices $\Lambda \subseteq \mathbb{R}$ of arbitrarily large dimension. □

Exercise 1.2: Determine all bases for the unit integer lattice $\mathbb{Z}^2$ where the components of each basis vector are between -4 and 4. (Distinguish a vector up to sign, as usual.) □

Exercise 1.3:
(i) For $A \in \mathbb{Z}^{d\times d}$, we have $\Lambda(A) = \mathbb{Z}^d$ iff $\det A = \pm 1$.
(ii) Give $A, B \in \mathbb{Z}^2$ such that $\det A = \det B$ but $\Lambda(A) \neq \Lambda(B)$. □

Exercise 1.4: The unimodular matrices in $\mathbb{Z}^{n\times n}$ with determinant 1 are called *positive unimodular matrices*. Clearly this is a subgroup of the unimodular matrices. For $n = 2$, show that this subgroup is generated by

$$S = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad T = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

HINT: What is T^2? S^m (the mth power of S)? Use S and T to transform a positive unimodular matrix $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ so that it satisfies $0 \leq b < a$. Now use induction on a to show that M is generated by S and T. □

Exercise 1.5: Show that the set of 2×2 integer unimodular matrices is generated by the following two elementary unimodular matrices:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Note that, in contrast, the general case seems to need three generators. HINT: you may reduce this to the previous problem. □

Exercise 1.6: Show that every lattice Λ has a basis that includes a shortest vector. □

Exercise 1.7: (Dubé) Consider $e_1, e_2, \ldots, e_{n-1}, h$ where e_i is the elementary n-vector whose ith component equals 1 and all other components equal zero, and $h = (\underbrace{\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}}_{n})$. Show that this set of vectors form a basis for the lattice $\Lambda = \mathbb{Z} \cup (\frac{1}{2} + \mathbb{Z})$. What is the shortest n-sequence of Λ? Show that for $n \geq 5$, this shortest n-sequence is not a basis for Λ. □

§2. Shortest vectors in planar lattices

In the rest of this lecture, we focus on lattices in $\mathbb{R}^2$. We identify $\mathbb{R}^2$ with $\mathbb{C}$ via the correspondence $(a, b) \in \mathbb{R}^2 \mapsto a + \mathbf{i}b \in \mathbb{C}$, and speak of complex numbers and 2-vectors interchangeably.

Thus we may write an expression such as '$\langle u/v, w\rangle$' where 'u/v' only makes sense if u, v are treated as complex numbers but the scalar product treats the result u/v as a vector. No ambiguity arises in such mixed notations. Let

$$\angle(u, v) = \angle(v, u)$$

denote the non-reflex angle between the vectors u and v. The *unit normal* $u^\perp$ to u is defined as

$$u^\perp := \widehat{u}\mathbf{i}.$$

Note that multiplying a complex number by $\mathbf{i}$ amounts to rotating the corresponding vector counter-clockwise by 90°.

First, let us relate the shortest vector problem to the GCD problem. Consider the 1-dimensional lattice generated by a set $u_1, \ldots, u_k$ of integers: $\Lambda = \Lambda(u_1, \ldots, u_k)$. It is easy to see that $\Lambda = \Lambda(g)$ where $g = \texttt{GCD}(u_1, \ldots, u_k)$. So g is the shortest vector in Λ. Hence computing shortest vectors is a generalization of the GCD problem. Hence it is not surprising that the GCD problem can be reduced to the shortest vector problem (Exercise). The following definition is key to a characterization of shortest vectors.

Fundamental Region. The *fundamental region* of $u \in \mathbb{C}\setminus\{0\}$ is the set $\mathcal{F}(u)$ of complex numbers $v \in \mathbb{C}$ such that

1. $|v| \geq |u|$;
2. $-\frac{|u|^2}{2} < \langle u, v\rangle \leq \frac{|u|^2}{2}$.

Figure 1 illustrates the fundamental region of u.

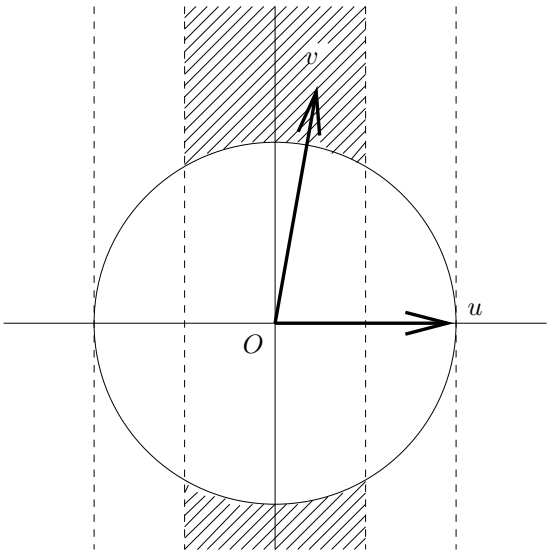


Figure 1: Fundamental Region of u is shaded.

This figure is typical in that, when displaying the fundamental region of $u \in \mathbb{C}$, we usually rotate the axes so that u appears horizontal. Note that $v \in \mathcal{F}(u)$ implies that $\angle(u, v) \geq 60°$.

Lemma 2 *If v is in the fundamental region of u then the sequence (u, v) is a shortest 2-sequence in $\Lambda(u, v)$. Moreover, this shortest 2-sequence is unique (up to sign) unless $|u| = |v|$.*

Proof. We first show that u is a shortest vector of $\Lambda = \Lambda(u, v)$. Let $w = mu + nv \in \Lambda$ be a shortest vector. Projecting w onto $u^\perp$, and noting that $|\langle u^\perp, v\rangle| \geq \sqrt{3}|v|/2$, we have

$$|\langle w, u^\perp\rangle| \geq \frac{|n|\sqrt{3}|v|}{2}. \tag{2}$$

Thus if $|n| > 1$, we have $|w| > |\langle w, u^\perp\rangle| > |v| \geq |u|$, contradicting the choice of w as a shortest vector. Hence we must have $|n| \leq 1$. Next we have

$$|\langle w, \widehat{u}\rangle| \geq \frac{(|m| - 1/2)|u|}{2}.$$

Thus if $|m| > 1$, we have $|w| > |\langle w, \widehat{u}\rangle| > |u|$, again a contradiction. Hence we must have $|m| \leq 1$. If $|m| = 1$ and $|n| = 1$, we have

$$|w|^2 = \langle w, u^\perp\rangle^2 + \langle w, \widehat{u}\rangle^2 > \frac{3}{4}|u|^2 + \frac{1}{4}|u|^2 = |u|^2,$$

contradiction. Hence we must have $|m| + |n| = 1$. There remain two possibilities: (I) If $n = 0$ then $|m| = 1$ and so $|w| = |u|$ and hence u is a shortest vector. (II) If $m = 0$ then $|n| = 1$ and so $|w| = |v|$. Since $|u| \leq |v|$, we conclude u and v are both shortest vectors.

Summarizing, we say that either (I) u is the unique shortest vector (as always, up to sign), or else (II) both u and v are shortest vectors.

We proceed to show that (u, v) is a shortest 2-sequence in $\Lambda(u, v)$. It suffices to show that v is a shortest vector in $\Lambda \setminus \Lambda(u)$. If $w = mu + nv$ is a second shortest vector, then $n \neq 0$. This implies $|n| = 1$ (otherwise, $|w| > |v|$ as shown above). Clearly $|\langle w, u^\perp\rangle| = |\langle v, u^\perp\rangle|$. Also $|\langle w, \widehat{u}\rangle| = m|u| \pm \langle v, \widehat{u}\rangle \geq |\langle v, \widehat{u}\rangle|$, with equality iff $m = 0$. Hence

$$|w|^2 \geq |\langle v, \widehat{u}\rangle|^2 + |\langle v, u^\perp\rangle|^2 = |v|^2,$$

with equality iff $m = 0$. Hence $|w| = |v|$. This proves that (u, v) is a shortest 2-sequence. Moreover, this is unique up to sign unless case (II) occurs. **Q.E.D.**

In the exceptional case of this lemma, we have at least two shortest 2-sequences: $(\pm u, \pm v)$ and $(\pm v, \pm u)$. There are no other possibilities unless we also have $\angle(u, v) = 60^\circ$ or 120°. Then let $w := u + v$ if $\angle(u, v) = 120^\circ$, and $w := u - v$ otherwise. There are now 4 other shortest 2-sequences: $(\pm w, \pm v)$ or $(\pm v, \pm w)$, $(\pm u, \pm w)$ or $(\pm w, \pm u)$.

Coherence and Order. Let $0 \leq \alpha \leq 1$. We say a pair (u, v) of complex numbers is *α-coherent* if

$$\langle \widehat{u}, \widehat{v}\rangle \geq \alpha, \quad u \neq 0, v \neq 0.$$

If $\alpha = 0$ then we simply say *coherent*; if $\alpha = 1/2$ then we say *strongly coherent*. If (u, v) is not coherent, we say it is *incoherent*. We say a pair (u, v) is *ordered* if $|u| > |v|$, otherwise it is *inverted*. We say (u, v) is *admissible* if it is ordered and coherent; otherwise it is *inadmissible*. So α-coherence of u, v amounts to

$$\angle(u, v) \leq \cos^{-1}(\alpha).$$

Thus (u, v) is strongly coherent means $\angle(u, v) \leq 60^\circ$.

Let us investigate the transformation of pairs $(u, v) \in \mathbb{C}^2$ to (v, w) where $w = u - q \cdot v$ for some positive integer q:

$$(u, v) \xrightarrow{q} (v, w).$$

The critical restriction here is that q must be positive. A pair (u, v) can be in one of four different *states*, taken from the set $\{coherent, incoherent\} \times \{ordered, inverted\}$, and defined in the natural way. Focusing only on the states of the pairs involved, it is not hard to verify that the only possible transitions are given in figure 2.

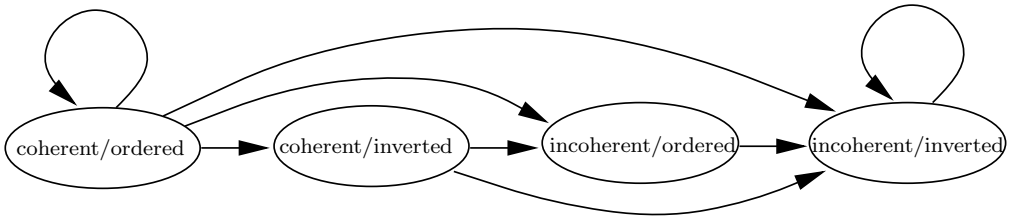


Figure 2: Possible state transitions.

Referring to figure 2, we may call the coherent/inverted and incoherent/ordered states *transitory*, since these states immediately transform to other states.

We record these observations:

Lemma 3 *Let $q_1, q_2, \ldots, q_{k-1}$ $(k \geq 2)$ be arbitrary positive integers. Assuming $u_0, u_1 \in \mathbb{C}$ are not collinear, consider the sequence $(u_0, u_1, u_2, \ldots, u_k)$ where $u_{i+1} = u_{i-1} - q_i u_i$. Also define*

$$p_i := (u_{i-1}, u_i), \qquad \theta_i := \angle(u_{i-1}, u_i), \qquad s_i := |u_{i-1}| + |u_i|, \qquad (i = 1, \ldots, k).$$

(i) *The sequence of angles $\theta_1, \theta_2, \ldots, \theta_k$ is strictly increasing.*

(ii) *The sequence of pairs $p_1, \ldots, p_k$ comprises a prefix of admissible pairs, followed by a suffix of inadmissible pairs. The prefix or the suffix may be empty. The suffix may begin with up to two transitory pairs.*

(iii) *The sequence of sizes $s_1, \ldots, s_k$ comprises a decreasing prefix, followed by an increasing suffix. In case both prefix and suffix are non-empty, say*

$$\cdots s_{i-2} > s_{i-1} > s_i > s_{i+1} > s_{i+2} \cdots$$

then either p_i or p_{i+1} is the first inadmissible pair.

We let the reader verify these remarks. Since we are interested in short lattice basis, lemma 3(iii) suggests that we study sequences whose pairs are admissible. This is taken up next.

EXERCISES

Exercise 2.1: Find the shortest 2-sequence for the lattice $\Lambda(u, v)$ where $u = \begin{bmatrix} 9 \\ 5 \end{bmatrix}$ and $v = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$. □

Exercise 2.2: Show how to compute the second shortest vector in $\Lambda(u, v)$, given that you have the shortest vector w. You may assume that you know $m, n \in \mathbb{Z}$ such that $w = mu + nv$. □

Exercise 2.3: (Zeugmann, v. z. Gathen) Let $a \geq b$ be positive integers. Show that the shortest vector in the lattice $\Lambda \subseteq \mathbb{Z}^2$ generated by $(a(a+1), 0)$ and $(b(a+1), 1)$ is $(0, a')$ where $a' = a/\texttt{GCD}(a, b)$. Conclude that integer GCD computation can be reduced to shortest vectors in $\mathbb{Z}^2$. □

Exercise 2.4: Show that if v, v' are distinct members of $\mathcal{F}(u)$ then $\Lambda(u, v) \neq \Lambda(u, v')$. □

§3. Coherent Remainder Sequences

If v is non-zero, we define the *coherent quotient* of u divided by v as follows:

$$\texttt{quo}_+(u, v) := \left\lfloor \frac{\langle u, v \rangle}{|v|^2} \right\rfloor .$$

Note that (u, v) is coherent iff $\texttt{quo}_+(u, v) \geq 0$. In this case, $\texttt{quo}_+(u, v)$ is the largest $j_0 \in \mathbb{Z}$ such that $(u - j_0 v, v)$ remains coherent. The *coherent remainder* of u, v is defined to be

$$\texttt{rem}_+(u, v) := u - \texttt{quo}_+(u, v) \cdot v.$$

Figure 3 illustrates geometrically the taking of coherent remainders. We are only interested in this definition when (u, v) is coherent.

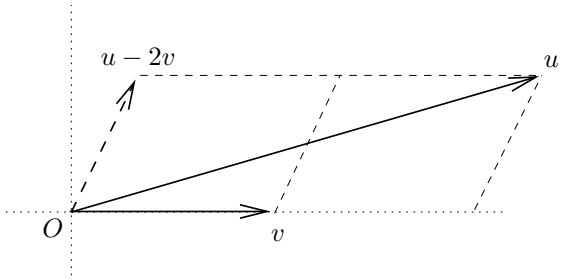


Figure 3: Coherent remainder of u, v where $\texttt{quo}_+(u, v) = 2$.

Note that the pair $(v, \texttt{rem}_+(u, v))$ is coherent unless $\texttt{rem}_+(u, v) = \mathbf{0}$.

If (u_0, u_1) is admissible, define the *coherent remainder sequence* (abbreviated, CRS) of u_0, u_1 to be the maximal length sequence

$$\texttt{CRS}(u_0, u_1) := (u_0, u_1, \ldots, u_{i-1}, u_i, \ldots)$$

such that for each $i \geq 1$,

1. Each pair (u_{i-1}, u_i) is admissible.
2. $u_{i+1} = \texttt{rem}_+(u_{i-1}, u_i)$.

This definition leaves open the possibility that $\texttt{CRS}(u_0, u_1)$ does not terminate – we prove below that this possibility does not arise. A *pair* in the CRS is just any two consecutive members, (u_i, u_{i+1}). The *initial pair* of the CRS is (u_0, u_1), and assuming the CRS has a last term u_k, the *terminal pair* of the CRS is (u_{k-1}, u_k). Clearly a pair (u, v) is a terminal pair in some CRS iff (u, v) is a CRS. So we may call (u, v) a "terminal pair" without reference to any larger CRS. The "maximal length" requirement in the definition of a coherent remainder sequence means that if u_{k+1} is the coherent remainder of a terminal pair (u_{k-1}, u_k), then either $u_{k+1} = 0$ or $|u_{k+1}| \geq |u_k|$.

The next lemma shows that every terminal pair can be easily transformed into a shortest 2-sequence. The proof refers to three regions on the plane illustrated in figure 4. These regions are defined as follows.

- $(I) = \{w \in \mathbb{C} : |w| \geq |v|,\ 0 \leq \langle w, v\rangle \leq |v|^2/2\}$,
- $(II) = \{w \in \mathbb{C} : |v|^2/2 < \langle w, v\rangle \leq |v|^2,\ |w - v| \geq |v|\}$,
- $(III) = \{w \in \mathbb{C} : \langle w, v\rangle \leq |v|^2,\ |w| \geq |v|,\ |w - v| < |v|\}$.

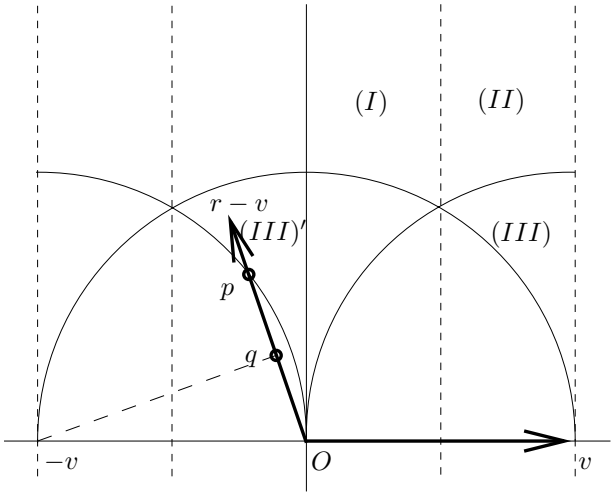


Figure 4: A terminal pair (u, v) where $r = \texttt{rem}_+(u, v) \neq 0$.

Lemma 4 *Let (u, v) be a terminal pair and $r = \texttt{rem}_+(u, v)$. If $|r| \geq |v|$ then one of the following holds. With the notations of Figure 4:*

(i) $r \in \mathcal{F}(v)$ *if* $r \in (I)$.

(ii) $r - v \in \mathcal{F}(v)$ *if* $r \in (II)$.

(iii) $-v \in \mathcal{F}(r - v)$ *if* $r \in (III)$.

Proof. Without loss of generality, assume $\langle r, v^\perp\rangle > 0$ ($v^\perp$ points upwards in the figure). Clearly, r belongs to one of the three regions (I), (II) or (III). If r is in (I) or (II), it clearly satisfies the lemma. In case (III), $r - v$ lies in region $(III)'$, simply defined as $\{z : z + v \in (III)\}$. The line segment from 0 to $r - v$ intersects the circle centered at $-v$ of radius $|v|$ at some point p. Dropping the perpendicular from $-v$ to the point q on the segment Op, we see that

$$\langle (-v), p\rangle = \frac{|p|^2}{2}$$

and hence $\langle(-v),(r-v)\rangle \leq \frac{|r-v|^2}{2}$, *i.e.*, $-v \in \mathcal{F}(r-v)$. **Q.E.D.**

Combined with lemma 2, we conclude:

Corollary 5 *Let (u,v) be a terminal pair and $r = \mathtt{rem}_+(u,v)$.*

(i) *Either v or $r-v$ is a shortest vector in the lattice $\Lambda(u,v)$.*

(ii) *If $r = 0$ then the lattice is one dimensional. Otherwise, a simple unimodular transformation of (u,v) leads to a shortest 2-sequence of $\Lambda(u,v)$. Namely, one of the following is a shortest 2-sequence of $\Lambda(u,v)$:*

$$(v,u),(v,r-v),(r-v,v).$$

We had noted that the angles θ_i defined by pairs (u_i,u_{i+1}) of a CRS is increasing with i (assuming $\theta_i > 0$). The following implies that if $\theta_i \geq 60°$, equivalently, (u_i,u_{i+1}) is not strongly coherent, then (u_i,u_{i+1}) is terminal.

Lemma 6 (60-degree Lemma) *Let (u,v) be admissible. If $\angle(u,v) \geq 60°$ then v is a shortest vector in $\Lambda(u,v)$.*

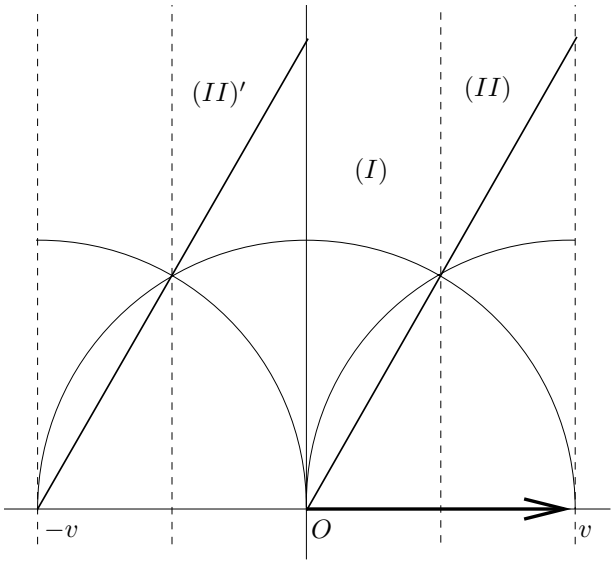


Figure 5: Sixty Degrees Region of u.

Proof. Referring to Figure 5, define the following regions:

- region (I) comprises those $w \in \mathcal{F}(v)$ inside the vertical strip $0 \leq \mathtt{Re}(w/v) \leq 1/2$;
- region (II) comprises those w inside $1/2 < \mathtt{Re}(w/v) \leq 1$ and $\angle(w,v) \geq 60°$.

Clearly, $r = \texttt{rem}_+(u, v)$ belongs to one of these two regions. If $r \in (I)$, and since $(I) \subseteq \mathcal{F}(v)$, we conclude that v is a shortest vector of $\Lambda(r, v) = \Lambda(u, v)$. If $r \in (II)$ then $r - v \in (II)'$ where $(II)'$ is defined to be $(II) - v$. But $(II)' \subseteq \mathcal{F}(v)$. Again v is a shortest vector of $\Lambda(r - v, v) = \Lambda(u, v)$. **Q.E.D.**

One more lemma is needed to prove termination of a remainder sequence: we show that the angles increase at some discrete pace.

Lemma 7 *Let (u_0, u_1, u_2) be the first 3 terms of a CRS and let $\theta_0 = \angle(u_0, u_1)$ and $\theta_1 = \angle(u_1, u_2)$. If $\theta_1 \leq 60°$ then*

$$\sin\theta_1 \geq \frac{2}{\sqrt{3}} \sin\theta_0.$$

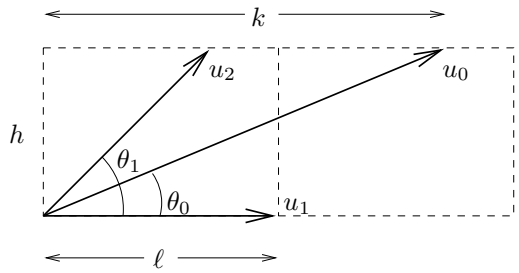


Figure 6: Three consecutive terms in a CRS.

Proof. Let $u_2 = u_0 - qu_1$. It is sufficient to prove the lemma for $q = 1$. Let h, k, ℓ be the lengths shown in figure 6. Since $\theta_1 \leq 60°$, we have $h \leq \sqrt{3}\ell$. Also $\ell \leq k \leq 2\ell$. Thus

$$\left(\frac{\sin\theta_0}{\sin\theta_1}\right)^2 = \frac{h^2 + (k-\ell)^2}{h^2 + k^2} = 1 - \left(\frac{2k\ell - \ell^2}{h^2 + k^2}\right).$$

It is easy to see that

$$\left(\frac{2k\ell - \ell^2}{h^2 + k^2}\right) \geq \frac{\ell^2}{3\ell^2 + 4\ell^2} = \frac{1}{7}.$$

This implies $\sin\theta_1 \geq \sqrt{7/6}\sin\theta_0$. To get the improved bound of the lemma, define the function

$$f(k) := \frac{2k\ell - \ell^2}{3\ell^2 + k^2}.$$

Then $df/dk = 0$ implies $k^2 - k\ell - 3\ell^2 = 0$. This has solution $k = \ell(-1 \pm \sqrt{13})/2$. Hence $f(k)$ has no minimum within the range $[\ell, 2\ell]$, and in this range, the minimum is attained at an end-point. We check that $f(k) \geq f(\ell) = 1/4$ for all $k \in [\ell, 2\ell]$. Hence $\frac{\sin\theta_0}{\sin\theta_1} \leq \sqrt{1 - f(\ell)} = \sqrt{3}/2$. **Q.E.D.**

Theorem 8 *For every admissible pair (u_0, u_1), the number of terms in $\texttt{CRS}(u_0, u_1)$ is at most*

$$3 - 2\log_{4/3}(2\sin\theta_0)$$

where $\theta_0 = \angle(u_0, u_1)$.

Proof. Let

$$\texttt{CRS}(u_0, u_1) = (u_0, u_1, u_2, \ldots, u_i, u_{i+1}, \ldots).$$

To show that the sequence terminates, consider the angles $\theta_i := \angle(u_i, u_{i+1})$. We have

$$\sin\theta_0 \leq \sqrt{3/4}\sin\theta_1 \leq \cdots \leq (3/4)^{i/2}\sin\theta_i$$

provided $\theta_i \leq 60°$. Since $\sin\theta_i \leq \sin 60° = 1/2$, we get $(4/3)^{i/2} \leq 1/(2\sin\theta_0)$ or

$$i \leq -2\log_{4/3}(2\sin\theta_0).$$

If θ_{i+1} is defined and $i+1 > -2\log_{4/3}(2\sin\theta_0)$, then $\theta_{i+1} > 60°$. By the 60-degree Lemma, u_{i+2} is the shortest vector. Hence u_{i+2} must be the last term in the CRS, and the CRS has $i+3$ terms. **Q.E.D.**

Our final result shows the existence of "shortest bases":

Theorem 9 *Every lattice Λ has a basis that is a shortest 2-sequence.*

Proof. Let $\Lambda = \Lambda(u, v)$. It is easy to transform u, v to an admissible pair (u', v') such that $\Lambda(u, v) = \Lambda(u', v')$. By the previous theorem, the CRS of (u', v') has a terminal pair, say, (u'', v''). Since each consecutive pair of a CRS is produced by unimodular transformations, these pairs are bases for Λ. In particular, (u'', v'') is a basis. By corollary 5, a unimodular transformation of (u'', v'') creates a shortest 2-sequence which is therefore a basis for Λ. **Q.E.D.**

The preceding development reduces the shortest vector and shortest basis problem to computing coherent remainder sequences.

Theorem 10 *Given $u, v \in \mathbb{Z}[\mathbf{i}]$ where each component number is n-bits, the shortest basis for $\Lambda(u, v)$ can be computed in $O(n\mathrm{M}_B(n))$ time.*

Proof. By replacing u with $-u$ if necessary, we assume (u, v) is admissible, possibly after reordering. Let $\theta = \angle(u, v)$. We claim that $-\log\sin\theta = O(n)$. To see this, consider the triangle $(0, u, v)$. By the cosine formula,

$$\sin\theta = \frac{\sqrt{(2|u|\cdot|v|)^2 - (|u|^2 + |v|^2 - |u-v|^2)^2}}{2|u|\cdot|v|} \geq \frac{1}{2|u|\cdot|v|}.$$

Since both $|u|$ and $|v|$ are $O(2^n)$, so $(\sin\theta)^{-1} \leq 2|u|\cdot|v| = O(2^n)$, and our claim follows. By theorem 8, the number of steps in $\texttt{CRS}(u, v)$ is $-\log\sin\theta = O(n)$. The proof is complete now because each step of the CRS can be computed in $O(\mathrm{M}_B(n))$ time. **Q.E.D.**

Remarks:
The study of unimodular transformations is a deep topic. Our definition of "fundamental regions" is adapted from the classical literature. For basic properties of the fundamental region in the classical setting, see for example, [86]. See also §XIV.5 for the connection to Möbius transformations. The process of successive reductions of 2-vectors by subtracting a multiple of the last vector from the last-but-one vector may be called the "generic Gaussian algorithm". The "coherent version" of this generic Gaussian algorithm was described in [218]: it is, of course, analogous to non-negative

remainder sequences for rational integers (§II.3). The more commonly studied version of the Gaussian algorithm is analogous to the symmetric remainder sequences for integers. See [206] for the description of other variants of the Gaussian algorithm. A half-Gaussian algorithm (in analogy of half-GCD in Lecture II) was described in [218]. This leads to an improved complexity bound of $O(\log n \mathrm{M}_B(n))$ for computing shortest basis for $\Lambda(u, v)$.

Exercises

Exercise 3.1: Compute the sequence $\texttt{CRS}(33 + 4\mathbf{i}, 20 + \mathbf{i})$. □

Exercise 3.2: Suppose $u_0, u_1 \in \mathbb{Z}[\mathbf{i}]$ are Gaussian integers where each component has at most n bits. Bound the length of $\texttt{CRS}(u_0, u_1)$ as a function of n. □

Exercise 3.3: Let $(u_0, u_1, \ldots, u_k)$ be a CRS.

i) For any complex θ, the sequence $(u_0\theta, u_1\theta, \ldots, u_k\theta)$ is also a CRS.

ii) Assume u_1 is real and u_0 lies in the first quadrant. The angle between consecutive entries always contains the real axis and the subsequent u_i's alternately lie in the first and fourth quadrants. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture IX
Lattice Reduction and Applications

In the previous lecture, we studied lattice reduction in 2 dimensions. Now we present an algorithm that is applicable in all dimensions. This is essentially the algorithm in [117], popularly known as the "LLL algorithm". The complexity of the LLL-algorithm has been improved by Schönhage [183] and Schnorr [180].

This algorithm is very powerful, and we will describe some of its applications. Its most striking success was in solving a major open problem, that of factoring polynomials efficiently. The problem of *(univariate) integer polynomial factorization* can be thus formulated: given $P(X) \in \mathbb{Z}[X]$, find all the irreducible integer polynomial factors of $P(X)$, together with their multiplicities. E.g., with $P(X) = X^4 + X^3 + X + 1$, we want to return the factors $X + 1$ and $X^2 - X + 1$ with multiplicities (respectively) of 2 and 1. This answer is conventionally expressed as

$$P(X) = (X + 1)^2(X^2 - X + 1).$$

The polynomial factorization problem depends on the underlying polynomial ring (which should be a UFD for the problem to have a unique solution). For instance, if we regard $P(X)$ as a polynomial over $\overline{\mathbb{Z}}$ (the algebraic closure of $\mathbb{Z}$, §VI.3), then the answer becomes

$$P(X) = (X + 1)^2(X - \frac{1 - \sqrt{-3}}{2})(X - \frac{1 + \sqrt{-3}}{2}).$$

Since the factors are all linear, we have also found the roots of $P(X)$ in this case. Indeed, factoring integer polynomials over $\overline{\mathbb{Z}}$ amounts to root finding.

This connection goes in the other direction as well: this lecture shows that if we can approximate the roots of integer polynomials with sufficient accuracy then this can be used to factor integer polynomials over $\mathbb{Z}[X]$ in polynomial time. The original polynomial-time algorithm for factoring integer polynomials was a major result of A. K. Lenstra, H. W. Lenstra and Lovász [117].

Kronecker was the first to give an algorithm for factoring multivariate integer polynomials. Known methods for factoring multivariate polynomials are obtained by a reduction to univariate polynomial factorization. Using such a reduction, Kaltofen has shown that factorization of integer polynomials over a fixed number of variables is polynomial-time in the total degree and size of coefficients [96]. One can also extend these techniques to factor polynomials with coefficients that are algebraic numbers. See [99, 109, 115, 116, 83]. A closely related problem is testing if a polynomial is irreducible. This can clearly be reduced to factorization. For integer polynomials $P(X, Y)$, a theorem of Hilbert is useful: $P(X, Y)$ is irreducible implies $P(a, Y)$ is irreducible for some integer a. This can be generalized to multivariate polynomials and made effective in the sense that we show that random substitutions from a suitable set will preserve irreducibility with some positive probability [95]. Testing irreducibility of polynomials over arbitrary fields is, in general, undecidable (Frölich and Shepherdson, 1955). An polynomial is *absolutely irreducible* if it is irreducible even when viewed as a polynomial over the algebraic closure of its coefficient ring. Thus, X^2+Y^2 is irreducible over integers but it is not absolutely irreducible (since the complex polynomials $X \pm \mathbf{i}Y$ are factors). E. Noether (1922) has shown absolute irreducibility is decidable by a reduction to field operations. Again, absolute irreducibility for integer polynomials can be made efficient. For a history of polynomial factorization up to 1986, we refer to Kaltofen's surveys [33, 97].

In this lecture, the 2-norm $\|a\|_2$ of a vector a is simply written $\|a\|$.

§1. Gram-Schmidt Orthogonalization

We use the lattice concepts introduced in §VIII.1. Let $A = [a_1, \ldots, a_m] \in \mathbb{R}^{n \times m}$ be a lattice basis. Note that $1 \leq m \leq n$. The matrix A is *orthogonal* if for all $1 \leq i < j \leq m$, $\langle a_i, a_j \rangle = 0$. The following is a well-known procedure to convert A into an orthogonal basis $A^* = [a_1^*, \ldots, a_m^*]$:

Gram-Schmidt Procedure
Input: $A = [a_1, \ldots, a_m]$.
Output: $A^* = [a_1^*, \ldots, a_m^*]$, the Gram-Schmidt version of A.

1. $a_1^* \leftarrow a_1$.
2. for $i = 2, \ldots, m$ do

$$\mu_{ij} \leftarrow \frac{\langle a_i, a_j^* \rangle}{\langle a_j^*, a_j^* \rangle}, \qquad (\text{for } j = 1, \ldots, i-1) \tag{1}$$

$$a_i^* \leftarrow a_i - \sum_{j=1}^{i-1} \mu_{ij} \cdot a_j^*. \tag{2}$$

This is a very natural algorithm: for $m = 2, 3$, we ask the reader to visualize the operation $a_i \mapsto a_i^*$ as a projection. Let us verify that A^* is orthogonal by induction. As basis of induction,

$$\langle a_2^*, a_1^* \rangle = \langle a_2 - \mu_{21} a_1^*, a_1^* \rangle = \langle a_2, a_1^* \rangle - \mu_{21} \langle a_1^*, a_1^* \rangle = 0.$$

Proceeding inductively, if $i > j$ then

$$\langle a_i^*, a_j^* \rangle = \langle a_i - \sum_{k=1}^{i-1} \mu_{ik} a_k^*, a_j^* \rangle = \langle a_i, a_j^* \rangle - \mu_{ij} \langle a_j^*, a_j^* \rangle = 0,$$

as desired. We shall call A^* the *Gram-Schmidt version* of A. We say that two bases are *Gram-Schmidt equivalent* if they have a common Gram-Schmidt version.

Exercise 1.1: In §VIII.1, we described three elementary unimodular operations. Show that two of them (multiplying a column by -1, and adding a multiple of one column to another) preserve Gram-Schmidt equivalence. The first operation (exchanging two columns) does not. □

Let us rewrite (2) as

$$a_i = a_i^* + \sum_{j=1}^{i-1} \mu_{ij} a_j^*. \tag{3}$$

Then

$$\begin{aligned} \langle a_i, a_i^* \rangle &= \langle a_i^* + \sum_{j=1}^{i-1} \mu_{ij} a_j^*, a_i^* \rangle \\ &= \langle a_i^*, a_i^* \rangle. \end{aligned}$$

Hence (1) may be extended to

$$\mu_{ii} := \frac{\langle a_i, a_i^* \rangle}{\langle a_i^*, a_i^* \rangle} = 1,$$

whence (3) simplifies to

$$a_i = \sum_{j=1}^{i} \mu_{ij} a_j^*. \tag{4}$$

In matrix form,

$$A = A^* M^T \tag{5}$$

where $A^* = [a_1^*, \ldots, a_m^*]$ and M^T is the transpose of a lower diagonal matrix

$$M = \begin{bmatrix} \mu_{11} & 0 & 0 & \cdots & 0 \\ \mu_{21} & \mu_{22} & 0 & & 0 \\ \vdots & & & & \\ \mu_{m1} & \mu_{m2} & \mu_{m3} & \cdots & \mu_{mm} \end{bmatrix}. \tag{6}$$

Since $\mu_{ii} = 1$, it follows that $\det M = 1$ and so the Gram-Schmidt version of A is a unimodular transformation of A. However, M need not be an integer matrix.

Lemma 1

(i) $\det(A^T A) = \prod_{i=1}^{m} \|a_i^*\|^2$.

(ii) $\|a_i\| \geq \|a_i^*\|$ *for* $i = 1, \ldots, m$ *with equality iff* a_i *is orthogonal to all* a_j^* $(j = 1, \ldots, i-1)$.

Proof. (i)

$$\begin{aligned} \det(A^T A) &= \det(M A^{*T} \cdot A^* M^T) \\ &= \det(M) \det(A^{*T} A^*) \det(M^T) \\ &= \det(A^{*T} A) \\ &= \prod_{i=1}^{m} \|a_i^*\|^2. \end{aligned}$$

(ii) From (3), we get

$$\begin{aligned} \|a_i\|^2 &= \|a_i^*\|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 \|a_j^*\|^2 \\ &\geq \|a_i^*\|^2, \end{aligned}$$

with equality iff $\mu_{ij} = 0$ for all j. **Q.E.D.**

From this lemma, we deduce immediately

$$\sqrt{\det A^T A} \leq \prod_{i=1}^{m} \|a_i\|.$$

By part(ii), equality is attained iff each a_i is orthogonal to $a_1^*, \ldots, a_{i-1}^*$. But the latter condition is seen to be equivalent to saying that the a_i's are mutually orthogonal. In particular, when $m = n$, we get Hadamard's determinantal bound

$$|\det A| \leq \prod_{i=1}^{n} \|a_i\|.$$

In Lecture VI.7, for the proof of the Goldstein-Graham bound, we needed the complex version of Hadamard's bound. The preceding proof requires two simple modifications:

1. Instead of the transpose A^T, we now use the *Hermitian transpose* A^H, defined as $A^H := \overline{A}^T$, where $\overline{A}$ is obtained by taking the complex conjugate of each entry of A.

2. The 2–norm of a complex vector $u = (v_1, \ldots, v_n)$ is defined as $\|v\| = \left(\sum_{i=1}^n |v_i|^2\right)^{\frac{1}{2}} = (\sum_{i=1}^n v_i \overline{v}_i)^{\frac{1}{2}}$.

It is easy to verify that the preceding argument goes through. Thus we have the following generalization of Hadamard's bound:

Theorem 2 *Let $A = [a_1, \ldots, a_m] \in \mathbb{C}^{n\times m}$, $1 \le m \le n$. Then*

$$\sqrt{\det(A^H A)} \le \prod_{i=1}^m \|a_i\|.$$

Equality in this bound is achieved iff for all $1 \le i < j \le m$, $\langle a_i, \overline{a}_j\rangle = 0$ where $\overline{a}_j$ is the conjugation of each entry in a_j.

We revert to the setting in which A is a real $n \times m$ matrix. The quantity

$$\delta(A) := \frac{\|a_1\| \cdot \|a_2\| \cdots \|a_m\|}{\sqrt{\det(A^T A)}}$$

is called the *(orthogonality) defect* of A. Note that $\delta(A) \ge 1$. Intuitively, it measures the amount of A's distortion from its Gram-Schmidt version.

This suggests the following *minimum defect basis problem*: given a basis A, find another basis B with $\Lambda(A) = \Lambda(B)$ such that $\delta(B)$ is minimized. Lovász [75, p. 140] has shown this problem to be NP-complete. For many applications, it is sufficient to find a B such that $\delta(B)$ is at most some constant K that depends only on m and n. Call this the *K-defect basis problem*. In case $m = n$, Hermite has shown that there exists such a basis B with

$$\delta(B) \le K_n$$

where K_n depends only on n. The current bound for K_n is $O(n^{1/4}(0.97n)^n)$. We will show a polynomial-time algorithm in case $K = 2^{\binom{m}{2}}$.

Exercises

Exercise 1.2:
(i) If L is any linear subspace of $\mathbb{R}^n$ and $u \in \mathbb{R}^n$ then u can be decomposed as $u = u_L + u_N$ where $u_L \in L$ and u_N is normal to L (*i.e.*, $\langle u_N, a\rangle = 0$ for all $a \in L$). HINT: use the Gram-Schmidt algorithm.
(ii) This decomposition is unique. □

Exercise 1.3: Suppose we are given $B = [b_1, \ldots, b_m] \in \mathbb{Q}^{n\times m}$ and also its Gram-Schmidt version B^*. Note that since B is rational, so is B^*. Suppose $u \in \mathbb{Q}^n$ such that $[b_1, \ldots, b_m, u]$ has linearly dependent columns. Show how to find integers $s, t_1, \ldots, t_m$ such that $su = \sum_{i=1}^m t_i b_i$. HINT: project u to each b_i^*. □

§2. Minkowski's Convex Body Theorem

In this section, we prove a fundamental theorem of Minkowski. We assume full-dimensional lattices here.

Given any lattice basis $A = [a_1, \ldots, a_n] \in \mathbb{R}^{n\times n}$ we call the set

$$\mathcal{F}(A) := \{\alpha_1 a_1 + \ldots + \alpha_n a_n : \ 0 \le \alpha_i < 1, \quad i = 1, \ldots, n\}$$

the *fundamental parallelopiped* of A. The (n-dimensional) volume of $\mathcal{F}(A)$ is given by

$$\texttt{Vol}(\mathcal{F}(A)) = |\det(A)|.$$

It is not hard to see that $\mathbb{R}^n$ is partitioned by the family of sets

$$u + \mathcal{F}(A), \quad u \in \Lambda(A).$$

Any bounded convex set $B \subseteq \mathbb{R}^n$ with volume $\texttt{Vol}(B) > 0$ is called a *body*. The body is 0-*symmetric* if for all $x \in B$, we have also $-x \in B$.

Theorem 3 (Blichfeldt 1914) *Let $m \ge 1$ be an integer, Λ a lattice, and B any body with volume*

$$\texttt{Vol}(B) > m \cdot \det \Lambda.$$

Then there exist $(m+1)$ distinct points $p_1, \ldots, p_{m+1} \in B$ such that for all i, j,

$$p_i - p_j \in \Lambda.$$

Proof. Let $A = [a_1, \ldots, a_n]$ be a basis for Λ and $F = \mathcal{F}(A)$ be the fundamental parallelopiped. For $u \in \Lambda$, define

$$F_u = \{x \in F : x + u \in B\}.$$

Hence $(u + F) \cap B = u + F_u$. It follows that

$$\sum_{u\in\Lambda} \texttt{Vol}(F_u) = \texttt{Vol}(B) > m \cdot \texttt{Vol}(F).$$

We claim that there is a point $p_0 \in F$ that belongs to $m+1$ distinct set $F_{u_1}, \ldots, F_{u_{m+1}}$. If not, we may partition F into

$$F = F^{(0)} \cup F^{(1)} \cup \cdots \cup F^{(m)}$$

where $F^{(i)}$ consists of all those points $X \in F$ that belong to exactly i sets of the form F_u, $(u \in \Lambda)$. Then

$$\sum_{u} \texttt{Vol}(F_u) = \sum_{i=0}^{m} i\texttt{Vol}(F^{(i)}) \le m\texttt{Vol}(F)$$

which is a contradiction. Hence p_0 exists. Since p_0 belongs to F_{u_i} $(i = 1, \ldots, m+1)$, we see that each of the points

$$p_i := p_0 + u_i$$

belong to B. It is clear that the points $p_1, \ldots, p_{m+1}$ fulfill the theorem. **Q.E.D.**

Note that in the proof we use the fact that $\texttt{Vol}(F^{(i)})$ is well defined. We now deduce Minkowski's Convex Body theorem (as generalized by van der Corput).

Theorem 4 (Minkowski) *Let $B \subseteq \mathbb{R}^n$ be an O-symmetric body. For any integer $m \geq 1$ and lattice $\Lambda \subseteq \mathbb{R}^n$, if*

$$\texttt{Vol}(B) > m2^n \det(\Lambda) \tag{7}$$

then $B \cap \Lambda$ contains at least m pairs of points

$$\pm q_1, \ldots, \pm q_m$$

which are distinct from each other and from the origin O.

Proof. Let

$$\frac{1}{2}B = \{p \in \mathbb{R}^n : 2p \in B\}.$$

Then $\texttt{Vol}(\frac{1}{2}B) = 2^{-n}\texttt{Vol}(B) > m \cdot \det(\Lambda)$. By Blichfeldt's theorem, there are $m+1$ distinct points $\frac{p_1}{2}, \ldots, \frac{p_m}{2}, \frac{p_{m+1}}{2} \in \frac{1}{2}B$ such that $\frac{1}{2}p_i - \frac{1}{2}p_j \in \Lambda$ for all i, j. We may assume

$$p_1 \underset{\texttt{LEX}}{\geq} p_2 \underset{\texttt{LEX}}{\geq} \cdots \underset{\texttt{LEX}}{\geq} p_{m+1}$$

where $\underset{\texttt{LEX}}{\geq}$ denotes the lexicographical ordering: $p_i \underset{\texttt{LEX}}{\geq} p_j$ iff $p_i \neq p_j$ and the first non-zero component of $p_i - p_j$ is positive. Then let

$$q_i := \frac{1}{2}p_i - \frac{1}{2}p_{m+1}$$

for $i = 1, \ldots, m$. We see that

$$0, \pm q_1, \pm q_2, \ldots, \pm q_m$$

are all distinct ($q_i - q_j \neq 0$ since $p_i \neq p_j$ and $q_i + q_j \neq 0$ since it has a positive component). Finally, we see that

$$q_i \in B \quad (i = 1, \ldots, m)$$

because $p_i \in B$ and $-p_{m+1} \in B$ (by the O-symmetry of B) implies $\frac{1}{2}(p_i - p_{m+1}) \in B$ (since B is convex). So $\pm q_1, \ldots, \pm q_m$ satisfy the theorem. **Q.E.D.**

We remark that premise (7) of this theorem can be replaced by $\texttt{Vol}(B) \geq m2^n \det(\Lambda)$ provided B is compact. As an application, we now give an upper bound on the length of the shortest vector in a lattice.

Theorem 5 *In any lattice $\Lambda \subseteq \mathbb{R}^n$, there is a lattice point $\xi \in \Lambda$ such that*

$$\|\xi\| \leq \sqrt{\frac{2n}{\pi}} \cdot \det(\Lambda)^{\frac{1}{n}}.$$

Proof. Let B be the n-dimensional ball of radius r centered at the origin. It is well-known [145] that

$$\texttt{Vol}(B) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} \cdot r^n.$$

We choose r large enough so that Minkowski's Convex Body theorem implies B contains a lattice point $\xi \in \Lambda$:

$$\texttt{Vol}(B) \geq 2^n \det \Lambda$$

or

$$r \geq \frac{2}{\sqrt{\pi}} \left(\Gamma(\frac{n}{2}+1) \cdot \det(\Lambda)\right)^{\frac{1}{n}}.$$

Since $\Gamma(x+1) \leq x^x$, it suffices to choose r to be

$$r = \sqrt{\frac{2n}{\pi}}(\det(\Lambda))^{\frac{1}{n}}.$$

Then $\xi \in \Lambda \cap B$ satisfies $\|\xi\| \leq r$. **Q.E.D.**

If n is large enough, it is known that the constant $\sqrt{2/\pi}$ can be replaced by 0.32.

Exercises

Exercise 2.1: Give an upper bound on the length of the shortest vector ξ in a lattice $\Lambda(B)$ that is not necessarily full-dimensional. □

Exercise 2.2: (cf. [145])
i) Show that $\texttt{Vol}(B_n) = \pi^{n/2}/\Gamma(\frac{n}{2}+1)$ where B_n is the unit n-ball.
ii) If B is an $n \times n$ positive definite symmetric matrix, the set of n-vectors $x \in \mathbb{R}^n$ such that $x^T Bx \leq c$ ($c \in \mathbb{R}$) is an ellipsoid E. Determine $\texttt{Vol}(E)$ via a deformation of E into B_n. □

§3. Weakly Reduced Bases

As an intermediate step towards constructing bases with small defects, we introduce the concept of a weakly reduced basis. The motivation here is very natural. Given a basis $B = [b_1, \ldots, b_m]$, we see that its Gram-Schmidt version $B^* = [b_1^*, \ldots, b_m^*]$ has no defect: $\delta(B^*) = 1$. Although B and B^* are related by a unimodular transformation M, unfortunately M is not necessarily integer. So we aim to transform B via an integer unimodular matrix into some $\overline{B} = [\overline{b}_1, \ldots, \overline{b}_m]$ that is as close as possible to the ideal Gram-Schmidt version. To make this precise, recall that for $i = 1, \ldots, m$,

$$b_i = \sum_{j=1}^{i} \mu_{ij} b_j^* \tag{8}$$

where $\mu_{ij} = \frac{\langle b_i, b_j^* \rangle}{\langle b_j^*, b_j^* \rangle}$, and $\mu_{ii} = 1$ (see equation (4) §1).

We say that B is *weakly reduced* if in the relation (8), the μ_{ij}'s satisfy the constraint

$$|\mu_{ij}| \leq \frac{1}{2}, \qquad (1 \leq j < i \leq m).$$

Weakly reduced bases are as close to its Gram-Schmidt version as one can hope for, using only the elementary unimodular transformations but without permuting the columns. Let us consider how to construct such bases. If B is not weakly reduced, there is a pair of indices (i_0, j_0), $1 \leq j_0 < i_0 \leq m$, such that

$$|\mu_{i_0 j_0}| > \frac{1}{2}.$$

Pick (i_0, j_0) to be the *lexicographically largest* such pair: if $|\mu_{ij}| > 1/2$ then $(i_0, j_0) \underset{\texttt{LEX}}{\geq} (i, j)$, *i.e.*, either $i_0 > i$ or $i_0 = i, j_0 \geq j$. Let

$$c_0 = \lfloor \mu_{i_0 j_0} \rceil$$

be the integer closest to $\mu_{i_0 j_0}$. Note that $c_0 \neq 0$. Consider the following unimodular transformation

$$B = [b_1, \ldots, b_{i_0}, \ldots, b_m] \longrightarrow \overline{B} = [\overline{b}_1, \ldots, \overline{b}_{i_0}, \ldots, \overline{b}_m]$$

where

$$\overline{b}_i = \begin{cases} b_i & \text{if} \quad i \neq i_0 \\ b_{i_0} - c_0 b_{j_0} & \text{if} \quad i = i_0 \end{cases}$$

We call the $B \to \overline{B}$ transformation a *weak reduction step*. We observe that $\overline{B}$ and B are Gram-Schmidt equivalent. So we may express $\overline{B}$ in terms of its Gram-Schmidt version (which is still $B^* = [b_1^*, \ldots, b_m^*]$) thus:

$$\overline{b}_i = \sum_{j=1}^{i} \overline{\mu}_{ij} b_j^*$$

where it is easy to check that

$$\overline{\mu}_{ij} = \frac{\langle \overline{b}_i, b_j^* \rangle}{\langle b_j^*, b_j^* \rangle} = \begin{cases} \mu_{ij} & \text{if} \quad i \neq i_0, \\ \mu_{ij} - c_0 \mu_{j_0 j} & \text{if} \quad i = i_0 \end{cases}$$

In particular,

$$|\overline{\mu}_{i_0 j_0}| \quad = \quad |\mu_{i_0 j_0} - c_0| \leq \frac{1}{2}.$$

As usual, $\mu_{j_0 j} = 0$ if $j > j_0$. Hence, if (i, j) is any index such that $(i, j) \underset{\text{LEX}}{\geq} (i_0, j_0)$ then $\overline{\mu}_{ij} = \mu_{ij}$ so $|\overline{\mu}_{ij}| \leq \frac{1}{2}$. This immediately gives us the following.

Lemma 6 (Weak Reduction) *Given any basis $B \in \mathbb{R}^{n \times m}$, we can obtain a weakly reduced basis $\overline{B}$ where $\Lambda(B) = \Lambda(\overline{B})$ by applying at most $\binom{m}{2}$ weak reduction steps to B.*

§4. Reduced Bases and the LLL algorithm

Let us impose a restriction on weakly reduced bases B.

A weakly reduced basis B is *reduced* if in addition it satisfies

$$\|b_i^*\|^2 \leq 2\|b_{i+1}^*\|^2 \tag{9}$$

for $i = 1, \ldots, m-1$, where $B^* = [b_1^*, \ldots, b_m^*]$ is the Gram-Schmidt version of B. We first show that reduced bases have bounded defect.

Lemma 7 *If $B = [b_1, \ldots, b_m]$ is a reduced basis then its defect is bounded: $\delta(B) \leq 2^{\frac{1}{2}\binom{m}{2}}$.*

Proof. If $B^* = [b_1^*, \ldots, b_m^*]$ is the Gram-Schmidt version of B then, by induction using (9), we have

$$\|b_{i-j}^*\|^2 \leq 2^j \|b_i^*\|^2$$

for $0 \leq j \leq i$. But from the usual relation

$$b_i = b_i^* + \sum_{j=1}^{i-1} \mu_{ij} b_j^*$$

and $|\mu_{ij}| \leq \frac{1}{2}$, we get

$$\begin{aligned}
\|b_i\|^2 &\leq \|b_i^*\|^2 + \frac{1}{4}\sum_{j=1}^{i-1}\|b_j^*\|^2 \\
&\leq \|b_i^*\|^2 + \frac{1}{4}\sum_{j=1}^{i-1}2^{i-j}\|b_i^*\|^2 \\
&\leq \|b_i^*\|^2\left(1 + \sum_{j=1}^{i-1}2^{i-j-2}\right) \\
&\leq 2^{i-1}\|b_i^*\|^2 \qquad (i \geq 1). \\
\prod_{i=1}^{m}\|b_i\|^2 &\leq 2^{\binom{m}{2}}\prod_{i=1}^{m}\|b_i^*\|^2.
\end{aligned}$$

Q.E.D.

To measure how close a basis B is to being reduced, we introduce a real function $V(B)$ defined as follows:

$$V(B) := \prod_{i=1}^{m} V_i(B)$$

where

$$V_i(B) := \prod_{j=1}^{i}\|b_j^*\| = \sqrt{\det(B_i^T B_i)}$$

and B_i consists of the first i columns of B. Observe that $V_i(B)$ depends only on the Gram-Schmidt version of B_i. In particular, if B' is obtained by applying the weak reduction step to B, then

$$V(B') = V(B)$$

since B' and B are Gram-Schmidt equivalent. Since $\|b_i\| \geq \|b_i^*\|$ for all i, we deduce that

$$V(B) = \prod_{i=1}^{n}\|b_i^*\|^{n-i+1} \leq \{\max_i \|b_i\|\}^{\binom{n}{2}}.$$

Now suppose $B = [b_1, \ldots, b_m]$ is not reduced by virtue of the inequality

$$\|b_i^*\|^2 > 2\|b_{i+1}^*\|^2$$

for some $i = 1, \ldots, m$. It is natural to perform the following *reduction step* which exchanges the ith and $(i+1)$st columns of B. Let the new basis be

$$C = [c_1, \ldots, c_m] \leftarrow [b_1, \ldots, b_{i-1}, b_{i+1}, b_i, b_{i+2}, \ldots, b_m].$$

Thus $c_j = b_j$ whenever $j \neq i$ or $i+1$. The choice of i for this reduction step is not unique. Nevertheless we now show that $V(B)$ is decreased.

Lemma 8 *If C is obtained from B by a reduction step and B is weakly reduced then*

$$V(C) < \frac{\sqrt{3}}{2}V(B).$$

Proof. Let $C = [c_1, \ldots, c_m]$ be obtained from $B = [b_1, \ldots, b_m]$ by exchanging columns b_i and b_{i+1}. As usual, let $B^* = [b_1^*, \ldots, b_m^*]$ be the Gram-Schmidt version of B with the matrix $(\mu_{jk})_{j,k=1}^m$ connecting them (see (8)). Similarly, let $C^* = [c_1^*, \ldots, c_m^*]$ be the Gram-Schmidt version of C with the corresponding matrix $(\nu_{jk})_{j,k=1}^m$. We have

$$V_j(C) = V_j(B), \qquad (j = 1, \ldots, i-1, i+1, \ldots, m).$$

This is because for $j \neq i$, $C_j = B_j U_j$ where C_j, B_j denotes the matrix comprising the first j columns of C, B (respectively) and U_j is a suitable $j \times j$ unimodular matrix. Hence $|\det(C_j^T C_j)| = |\det(B_j^T B_j)|$. It follows that

$$\frac{V(C)}{V(B)} = \frac{V_i(C)}{V_i(B)} = \frac{\|c_i^*\|}{\|b_i^*\|}. \tag{10}$$

It remains to relate $\|c_i^*\|$ to $\|b_i^*\|$. By equation (8) for μ_{jk}, and a similar one for ν_{jk}, we have

$$c_i = b_{i+1} = b_{i+1}^* + \sum_{j=1}^{i} \mu_{i+1,j} b_j^* = b_{i+1}^* + \mu_{i+1,i} b_i^* + \sum_{j=1}^{i-1} \nu_{ij} c_j^*. \tag{11}$$

The last identity is easily seen if we remember that c_j^* is the component of c_j normal to the subspace spanned by $\{c_1, \ldots, c_{j-1}\}$. Hence $b^* j = c^* j$ and $\mu_{i+1,j} = \nu_{ij}$ for $j = 1, \ldots, i = 1$. Hence

$$\begin{aligned} c_i^* &= c_i - \sum_{j=1}^{i-1} \nu_{ij} c_j^* \\ &= b_{i+1}^* + \mu_{i+1,i} b_i^*. \end{aligned}$$

Since we switched b_i and b_{i+1} in the reduction step, we must have $\|b_i^*\|^2 > 2\|b_{i+1}^*\|^2$. Thus

$$\begin{aligned} \|c_i^*\|^2 &= \|b_{i+1}^*\|^2 + \mu_{i+1,i}^2 \|b_i^*\|^2 \\ &\leq \|b_{i+1}^*\|^2 + \frac{1}{4}\|b_i^*\|^2 \\ &< \frac{1}{2}\|b_i^*\|^2 + \frac{1}{4}\|b_i^*\|^2 = \frac{3}{4}\|b_i^*\|^2. \end{aligned}$$

This, with equation (10), proves the lemma. **Q.E.D.**

We now describe a version of the LLL algorithm (cf. Mishra [136]). In the following, *weak-reduce*(B) denotes a function call that returns a weakly reduced basis obtained by repeated application of the weak reduction step to B. Similarly, *reduce-step*(B) denotes a function that applies a single reduction step to a weakly-reduced B.

LLL Algorithm
- Input: $B \in \mathbb{Q}^{n \times m}$, a basis.
- Output: A reduced basis $\overline{B}$ with $\Lambda(\overline{B}) = \Lambda(B)$.
1. Initialize $\overline{B} \leftarrow$ *weak-reduce*(B).
2. while $\overline{B}$ is not reduced do
   - 2.1. $\overline{B} \leftarrow$ *reduce-step*$(\overline{B})$
   - 2.2. $\overline{B} \leftarrow$ *weak-reduce*$(\overline{B})$.

Correctness: It is clear that if the algorithm halts, then the output $\overline{B}$ is correct. It remains to prove halting. Write

$$B = \frac{1}{d} C \tag{12}$$

for some $C \in \mathbb{Z}^{n\times m}$ and $d \in \mathbb{Z}$. We may assume that in any elementary integer unimodular transform of the matrix in (12), the common denominator d is preserved. Hence it is sufficient to focus on the integer part C. If $C = [c_1, \ldots, c_m]$ then $\|c_i\| = d\|b_i\|$, so

$$\begin{aligned} V_i(C) &= d^i V_i(B) \\ V(C) &= d^{\binom{n}{2}} V(B) \\ &\leq \{d \max_{i=1,\ldots,m} \|b_i\|\}^{\binom{n}{2}}. \end{aligned}$$

If s is the maximum bit-size of entries of B then

$$\log \|b_i\| = O(s + \log n).$$

Each weak reduction step preserves $V(C)$ but a reduction step reduces it by a factor of $\sqrt{3}/2$. Since $|V(C)| \geq 1$, we conclude that the algorithm stops after

$$\log_{\sqrt{3}/2} V(C) = O(n^2 \log \left\{ d \max_i \|b_i\| \right\}) = O(n^2(s + \log n))$$

reduction steps. Each weak reduction of B involves one call to the Gram-Schmidt procedure and $O(n^2)$ vector operations of the form $b_i \leftarrow b_i - cb_j$, $(c \in \mathbb{Z})$. These take $O(n^3)$ arithmetic operations. We conclude with:

Theorem 9 *Given a basis $A \in \mathbb{Q}^{n\times m}$, we can compute a reduced basis B with $\Lambda(A) = \Lambda(B)$ using $O(n^5(s + \log n))$ arithmetic operations, where s is the maximum bit-size of entries in A.*

Exercises

Exercise 4.1: The reduction factor of $\sqrt{3}/2$ in this lemma is tight in the planar case ($n = 2$) (cf. §VIII.3). □

Exercise 4.2: *Bit Complexity.* For simplicity, assume $s \geq \log n$. Show that all intermediate numbers in the LLL algorithm have bit-size $O(ns)$. Conclude that if we use the classical algorithms for rational arithmetic operations, the bit-complexity of the algorithm is $O(n^7 s^3)$. □

Exercise 4.3: By keeping track of the updates to the basis in the weak reduction step we can save a factor of n. Using fast integer arithmetic algorithms, we finally get $O(n^5 s^2 \mathcal{L}(ns))$. □

Exercise 4.4: The LLL algorithm above assumes the columns of the input matrix B forms a basis. In some applications this assumption is somewhat inconvenient. Show how to modify LLL algorithm to accept B whose columns need not be linearly independent. □

§5. Short Vectors

Let $B = [b_1, \ldots, b_m] \in \mathbb{R}^{n\times m}$ be a basis and let $\xi_1 \in \Lambda = \Lambda(B)$ denote the shortest lattice vector, $\xi_1 \neq 0$. We do not know if computing the shortest vector from B is NP-complete. This assumes that the length of a vector is its Euclidean norm. If we use the ∞-norm instead, van Emde Boas (1981)

has shown that the problem becomes NP-complete. In §2, we show that if Λ is a full-dimensional lattice, $\|\xi_1\|$ is bounded by $\sqrt{\frac{2n}{\pi}}\det(\Lambda)^{\frac{1}{n}}$. We do not even have an efficient algorithm to compute *any* lattice vector with length within this bound. But this lecture shows that we can efficiently construct a vector ξ whose length is bounded by a slightly larger constant. Moreover, $\|\xi\|$ is also bounded relative to the length of the shortest vector: $\|\xi\|/\|\xi_1\| \leq 2^{(m-1)/2}$. Indeed, finding such a ξ is trivially reduced to the LLL-algorithm by showing that ξ can be chosen from a reduced base.

Lemma 10 *Let $B^* = [b_1^*, \ldots, b_m^*]$ be the Gram-Schmidt version of B. Then the shortest vector ξ_1 satisfies*

$$\|\xi_1\| \geq \min_{i=1,\ldots,m} \|b_i^*\|.$$

Proof. Suppose

$$\xi_1 = \sum_{i=1}^{k} \lambda_i b_i \qquad (\lambda_i \in \mathbb{Z}, \lambda_k \neq 0)$$

for some $1 \leq k \leq m$. Then

$$\begin{aligned} \xi_1 &= \sum_{i=1}^{k} \lambda_i \sum_{j=1}^{i} \mu_{ij} b_j^* \qquad \text{, by } \textit{equation } (8) \\ &= \lambda_k b_k^* + \sum_{i=1}^{k-1} \mu_i' b_i^* \end{aligned}$$

for some suitable $\mu_i' \in \mathbb{Q}$. Hence

$$\|\xi_1\| \geq |\lambda_k| \cdot \|b_k^*\| \geq \|b_k^*\|.$$

Q.E.D.

We deduce from the above:

Lemma 11 *Let $B = [b_1, \ldots, b_m]$ be a reduced basis and ξ_1 be a shortest vector in $\Lambda(B)$.*

(i) $\|b_1\| \leq 2^{(m-1)/2}\|\xi_1\|$,

(ii) $\|b_1\| \leq 2^{(m-1)/4}(\det \Lambda(B))^{1/m}$.

Proof. (i) Let b_i be the shortest vector in B. Since B is reduced,

$$\|b_1\|^2 = \|b_1^*\|^2 \leq 2^{i-1}\|b_i^*\|^2 \leq 2^{m-1}\|\xi_1\|^2.$$

(ii) $\|b_1\|^{2m} \leq \prod_{i=1}^{m} 2^{i-1}\|b_i^*\|^2 = 2^{\binom{m}{2}} \det(B^T B)$. **Q.E.D.**

Thus we can use the LLL-algorithm to construct a short vector ξ satisfying both

$$\|\xi\|/\|\xi_1\| \leq 2^{(m-1)/2} \quad \text{and} \quad \|\xi\| \leq 2^{(m-1)/4}(\det(B^T B))^{1/2m}. \tag{13}$$

Simultaneous Approximation. Let us give an application to the problem of simultaneous approximation: *given rational numbers $\alpha_1, \ldots, \alpha_n$ and positive integer bounds N, s, find integers $p_1, \ldots, p_n, q$ such that*

$$|q| \leq N \quad \text{and} \quad |q\alpha_i - p_i| \leq 2^{-s}, \qquad (i = 1, \ldots, n). \tag{14}$$

In other words, we want to simultaneously approximate the numbers $\alpha_1, \ldots, \alpha_n$ by rationals numbers $p_1/q, \ldots, p_n/q$ with a common denominator. It is not hard to see that if N is large enough relative to s, there will be a solution; conversely there may be no solutions if N is small relative to s.

Lemma 12 (Dirichlet) *If $N = 2^{s(n-1)}$ then there is a solution to the simultaneous approximation problem.*

By way of motivation, note that the system of inequalities (14) translates into

$$\|p_1 e_1 + p_2 e_2 + \cdots + p_n e_n - q\alpha\|_\infty \leq 2^{-s}.$$

where e_i is the ith elementary n-vector $(0, \ldots, 0, 1, 0, \ldots, 0)$ with a "1" in the ith position and $\alpha = (\alpha_1, \ldots, \alpha_n)$. So this becomes the problem of computing a short vector

$$\xi = p_1 e_2 + p_2 e_2 + \cdots + p_n e_n - q\alpha \tag{15}$$

in the lattice generated by $B = [\alpha, e_1, \ldots, e_n]$.

Let us now prove Dirichlet's theorem: it turns out to be an immediate application of Minkowski's convex body theorem (§2). But we cannot directly apply Minkowski's theorem with the formulation of (15): the columns of B are not linearly independent. To circumvent this, we append an extra coordinate to each vectors in B. In particular, the $n+1$st coordinate of α can be given a non-zero value c, and each e_i is now an elementary $(n+1)$-vector. The modified B is

$$B = [\alpha, e_1, \ldots, e_n] = \begin{bmatrix} \alpha_1 & 1 & 0 & \cdots & 0 \\ \alpha_2 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ \alpha_n & 0 & 0 & \cdots & 1 \\ c & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{16}$$

Note that $\det(B) = c$, where we are free to choose c. Let C be the cube

$$C = \{(x_0, \ldots, x_n) \in \mathbb{R}^{n+1} : |x_i| \leq 2^{-s}\}.$$

The volume of C is $2^{(1-s)n}$. If we choose $c = 2^{-sn}$, then $\texttt{Vol}(C) = 2^n \det(B)$. Since C is compact, using a remark after Minkowski's theorem in §2, we conclude that C contains a point ξ of $\Lambda(B)$:

$$\xi = -q\alpha + \sum_{i=1}^{n} p_i e_i, \quad (q, p_i \in \mathbb{Z}).$$

Since $\xi \in C$, we have $|cq| \leq 2^{-s}$ (or $|q| \leq 2^{s(n-1)}$) and $|q\alpha_i - p_i| \leq 2^{-s}$, proving Dirichlet's theorem.

Unfortunately, there is no known polynomial-time algorithm to find Dirichlet's solution. In contrast, we have:

Theorem 13 *If $N = 2^{ns+n(n+1)/4}$ then the simultaneous approximation problem has a solution that can be found in polynomial time.*

The only algorithmic technique we have for short vectors involve the LLL-algorithm, and somehow we must reduce our problem computing a "short" vector in the sense of (13). The basic setup of the proof of Dirichlet's theorem can be used, with only a different choice of c. With B as in (16), we see from (13) that it suffices to choose c sufficiently small:

$$\|\xi\| \leq 2^{n/4} \det(B)^{1/(n+1)} = 2^{n/4} c^{1/(n+1)}.$$

Hence $\|\xi\|_\infty \leq \|\xi\| \leq 2^{-s}$ provided we choose

$$c = 2^{-(n/4+s)(n+1)}.$$

Now the $(n+1)$st coordinate of ξ is equal to $-qc$ hence $|qc| \leq 2^{-s}$ or

$$|q| \leq 2^{(n/4+s)(n+1)-s} = N.$$

This proves theorem 13.

Exercises

Exercise 5.1: Assuming that m is fixed, show that the shortest vector can be found in polynomial time. □

Exercise 5.2: Show by a general example that Dirichlet's result is tight. □

Exercise 5.3: (Babai 1986) A generalization of the short vector problem is the problem of *near vector*: given a basis $B = [b_1, \ldots, b_m] \in \mathbb{Q}^{n \times m}$ and a vector $u \in \mathbb{Q}^n$, find a lattice point $\xi \in \Lambda(B)$ that is "near to u" in this sense:

$$\|u - \xi\| \leq 2^{(m/2)-1} \|u - \xi_1\|$$

where ξ_1 is the nearest lattice point to u. Show that this problem can be solved in polynomial time. HINT: choose ξ such that

$$u - \xi = \sum_{i=1}^{m} \lambda_i b_i^*, \quad |\lambda_i| \leq 1/2,$$

where $[b_1^*, \ldots, b_m^*]$ is the Gram-Schmidt version of B □

§6. Factorization via Reconstruction of Minimal Polynomials

Approximate roots. Suppose we are given a pair

$$(\overline{\alpha}, s), \qquad \overline{\alpha} \in \mathbb{C}, s \in \mathbb{Z}$$

where $s \geq 4$. In our algorithmic applications, the real $\mathtt{Re}(\overline{\alpha})$ and imaginary $\mathtt{Im}(\overline{\alpha})$ parts of α will be rational numbers, but the mathematical results do not depend on this assumption. Let us call the pair $(\overline{\alpha}, s)$ an *approximate root* if there exists an algebraic number $\alpha \in \mathbb{C}$ with minimal polynomial $F(X)$ such that

$$\text{(i)} \quad |\alpha - \overline{\alpha}| \leq 2^{-5s^3}, \tag{17}$$

$$\text{(ii)} \quad F(X) \text{ has bit-size less than } s. \tag{18}$$

We also say the approximate root $(\overline{\alpha}, s)$ *belongs to* α. Note that equations (17) and (18) imply

$$\|F(X)\|_\infty \le 2^{s-1}, \quad \deg(F) \le s-1, \quad |\alpha| \le 2^s. \tag{19}$$

Our main task is to show that $F(X)$ is uniquely determined under assumptions (17) and (18), and to efficiently reconstruct $F(X)$ from $(\overline{\alpha}, s)$. Incidentally, this suggests that approximate roots can serve as yet another representation of algebraic numbers. Although the representation is very simple, it is not clear whether it is useful for general algebraic manipulations.

Application to Polynomial Factorization. Before addressing the main task, let us show how this leads to an efficient procedure for factoring integer polynomials. Suppose we wish to factor the integer polynomial $G(X)$ of degree $n \ge 2$. According to §IV.5, if $F(X)$ divides $G(X)$ then

$$\begin{aligned} \|F\|_\infty &\le |\mathtt{lead}(F)| \cdot \binom{n}{\lfloor n/2 \rfloor} \|G\| \\ &\le |\mathtt{lead}(G)| \cdot \binom{n}{\lfloor n/2 \rfloor} \|G\|. \end{aligned}$$

If t denotes the last expression, then the bit size of F is bounded by $s := (n+1)t$. We may, following §VII.5, isolate a complex root α of $G(X)$ to the accuracy required by (17). This gives us an approximation $\overline{\alpha}$ such that $|\alpha - \overline{\alpha}| \le 2^{-5s^3}$. Applying the minimal polynomial reconstruction algorithm to the approximate root $(\overline{\alpha}, s)$, we get a minimal polynomial F for α. Note that if $\deg F = n$ then F is just equal to the primitive part of $G(X)$ and we have verified that $G(X)$ is irreducible. Otherwise, $F(X)$ is a nontrivial irreducible factor of $G(X)$. We can now continue with the factorization of $G(X)/F(X)$.

The rest of this section considers the problem of reconstructing minimal polynomials, following Kannan, Lenstra and Lovász [99].

Lattice points as polynomials. Fix an arbitrary integer m, $1 \le m \le s$, and let

$$c := 2^{-4s^3}. \tag{20}$$

Consider the following matrix

$$B_m := \begin{bmatrix} \mathtt{Re}(\overline{\alpha}^0) & \mathtt{Re}(\overline{\alpha}^1) & \cdots & \mathtt{Re}(\overline{\alpha}^m) \\ \mathtt{Im}(\overline{\alpha}^0) & \mathtt{Im}(\overline{\alpha}^1) & \cdots & \mathtt{Im}(\overline{\alpha}^m) \\ c & 0 & \cdots & 0 \\ 0 & c & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & c \end{bmatrix}. \tag{21}$$

Clearly B_m is a basis for a $(m+1)$-dimensional lattice in $\mathbb{R}^{m+3}$. Let the columns of B_m be $b_0, \ldots, b_m$,

$$B_m = [b_0, \ldots, b_m],$$

and consider the following correspondence between polynomials $G(X) \in \mathbb{Z}[X]$ of degree at most m and lattice points $\overline{G} \in \Lambda(B_m)$:

$$G(X) = \sum_{i=0}^{m} g_i X^i \quad \Longleftrightarrow \quad \overline{G} = \sum_{i=0}^{m} g_i b_i. \tag{22}$$

This is seen to be a bijection once we notice

$$\overline{G} = (\mathtt{Re}(G(\overline{\alpha})), \mathtt{Im}(G(\overline{\alpha})), cg_0, cg_1, \ldots, cg_m). \tag{23}$$

It follows that

$$\begin{aligned} \|\overline{G}\|^2 &= \mathtt{Re}(G(\overline{\alpha}))^2 + \mathtt{Im}(G(\overline{\alpha}))^2 + c^2 \sum_{i=0}^{m} g_i^2 \\ \|\overline{G}\|^2 &= |G(\overline{\alpha})|^2 + c^2 \|G(X)\|^2. \end{aligned} \tag{24}$$

This important identity suggests a connection between the minimal polynomial $F(X)$ and short vectors in the lattice $\Lambda(B_m)$: *Assume m is the degree of the minimal polynomial $F(X)$.* Then the lattice point $\overline{F}$ corresponding to $F(X)$ has small length since $F(\overline{\alpha})$ is close to $F(\alpha) = 0$ and c is small implies $\|\overline{F}\| = |F(\overline{\alpha})|^2 + c^2\|F(X)\|$ is small. In fact for any $\overline{G} \neq \overline{F}$ we show that $\|\overline{G}\| \geq 2^m \|\overline{F}\|$. Intuitively, this means that the LLL-algorithm can distinguish $\overline{F}$ from all other lattice points. One complication is that we do not know in advance the degree m of the minimal polynomial. In any case, the idea is to reduce the search for a minimal polynomial to the search for a short vector in a suitable lattice. The following theorem answers the basic questions of this approach.

Theorem 14 (Correctness)
Let $(\overline{\alpha}, s)$ be an approximate root belonging to α. With $1 \leq m \leq s$ and $c = 2^{-4s^3}$, construct the basis B_m as above. Let $F(X)$ be the minimal polynomial of α and

$$C := 2^{2s} c = 2^{2s-4s^3}.$$

(i) *If $\overline{G} \in \Lambda(B_m)$ satisfies $\|\overline{G}\| \leq C$ then $G(\alpha) = 0$.*

(ii) *If $m = \deg F(X)$ then $\|\overline{F}\| \leq C$.*

(iii) *If $m = \deg F(X)$ then for any lattice point $\overline{G} \in \Lambda(B_m)$, $G(\alpha) \neq 0$ implies $\|\overline{G}\| > 2^s C$.*

It is easiest to appreciate the theorem by seeing how it justifies the following algorithm for reconstructing minimal polynomials:

Minimal Polynomial Algorithm:
Input: Approximate root $(\overline{\alpha}, s)$ belonging to α, *i.e.*, satisfying (17) and (18).
Output: $F(X)$, the minimal polynomial of α.
Method:
for $m = 1, 2, \ldots,$ do forever
1. Construct the basis B_m.
2. Use the LLL-algorithm to find a reduced basis B for B_m.
3. Let $\overline{G}$ be the first vector in B, and $G(X)$ the corresponding polynomial.
4. Let $H(X)$ be the primitive part of $G(X)$.
5. If $\|\overline{H}\| \leq C$ then output $H(X)$ and halt.

Justification of Algorithm: Let $F(X)$ be the minimal polynomial of α. Part (i) of the theorem shows that the algorithm cannot halt at any stage m where $m < \deg F \leq s - 1$. So suppose we have reached stage $m = \deg F$. From part (ii), the shortest vector in $\Lambda(B_m)$ has length at most C. Let $\overline{G}$ be first vector in the reduced basis. By a property of reduced bases (lemma 11(i)), $\|\overline{G}\| \leq 2^s \cdot C$.

Then part (iii) of the theorem implies $G(\alpha) = 0$. Hence the primitive part of $G(X)$ must be the minimal polynomial $F(X)$. Therefore in stage $m = \deg F$, the polynomial $H(X)$ is indeed equal to $F(X)$. Hence $H(X)$ will satisfy the condition $\|\overline{H}\| \leq C$, necessary for halting. So our algorithm will surely halt at the correct stage, and when it halts, the output is correct. This concludes the justification.

Proof of the Correctness Theorem. The proof will occupy the rest of this section. We first state without proof a simple estimate.

Lemma 15 *Let $A(X) \in \mathbb{Z}[X]$ and $\alpha, \overline{\alpha} \in \mathbb{C}$. If $m \geq \deg A(X)$ and $M \geq \max\{1, |\alpha|, |\overline{\alpha}|\}$ then*

$$|A(\alpha) - A(\overline{\alpha})| \leq |\alpha - \overline{\alpha}| \cdot \|A\|_\infty m^2 M^m.$$

Proof of part (i) of Theorem. From (24) and the assumption $\|\overline{G}\| \leq C$, we get

$$|G(\overline{\alpha})| \leq C \quad \text{and} \quad c\|G(X)\| \leq C. \tag{25}$$

The latter inequality implies

$$\|G(X)\| \leq c^{-1}C = 2^{2s}.$$

By lemma 15, and since $|\alpha| \leq 2^s$ and $m \leq s-1$,

$$\begin{aligned} |G(\alpha) - G(\overline{\alpha})| &\leq |\alpha - \overline{\alpha}| \cdot \|G\|_\infty \cdot m^2 \cdot 2^{sm} \\ &\leq |\alpha - \overline{\alpha}| \cdot \|G\| \cdot m^2 \cdot 2^{sm} \\ &\leq 2^{-5s^3} \cdot 2^{2s} \cdot 2^s \cdot 2^{s^2} \\ &\leq 2^{-4s^3} \qquad (\text{provided } \ s \geq 4). \\ |G(\alpha)| &\leq |G(\alpha) - G(\overline{\alpha})| + |G(\overline{\alpha})| \\ &\leq 2^{-4s^3} + C \\ &\leq 2C. \end{aligned} \tag{26}$$

Let

$$F(X) = \sum_{i=0}^{n} f_i X^i \tag{27}$$

be the minimal polynomial of α and $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$ be the conjugates of α. Consider the expression

$$f_n^m \prod_{i=1}^{n} G(\alpha_i). \tag{28}$$

Since $\prod_{i=1}^n G(\alpha_i)$ is symmetric in the α_i's and of degree $\leq m$ in α_i, the fundamental theorem on symmetric functions (§VI.5) implies that the expression (28) yields an integer. Since $|f_n| \leq 2^s, |\alpha_i| \leq 2^s$,

$$\begin{aligned} |G(\alpha_i)| &\leq \sum_{j=0}^{m} |g_j| |\alpha_i|^j \\ &\leq \|G\|_\infty \sum_{j=0}^{m} 2^{sj} \\ &\leq 2^{2s} \cdot 2^{sm+1} \leq 2^{2s^2}. \end{aligned}$$

Hence

$$\begin{aligned} |f_n^m \prod_{i=1}^{n} G(\alpha_i)| &\leq |f_n|^m \cdot |G(\alpha)| \cdot \prod_{i=2}^{n} |G(\alpha_i)| \\ &\leq 2^{sm} \cdot 2C \cdot (2^{2s^2})^{n-1} \\ &\leq C \cdot 2^{2s^3} \\ &< 1. \end{aligned}$$

It follows that the expression (28) is equal to 0. Hence $G(\alpha_i) = 0$ for some i. But the α_i's are conjugates, so $G(\alpha) = 0$. This concludes the proof of part(i).

Proof of part (ii). We now want to upper bound $\|\overline{F}\|$. Since $\|F(X)\|_\infty \leq 2^s$ and $|\alpha| \leq 2^s$, lemma 15 shows

$$\begin{aligned} |F(\overline{\alpha})| &= |F(\overline{\alpha}) - F(\alpha)| \\ &\leq |\alpha - \overline{\alpha}| \cdot \|F\|_\infty \cdot s^2 \cdot 2^{s^2} \\ &\leq 2^{-5s^3} \cdot 2^s \cdot 2^s \cdot 2^{s^2} \\ &\leq 2^{-4s^3}. \end{aligned}$$

Hence (cf. (24))

$$\begin{aligned} \|\overline{F}\| &= (|F(\overline{\alpha})|^2 + c^2 \|F(X)\|^2)^{\frac{1}{2}} \\ &\leq ((2^{-4s^3})^2 + (2^{-4s^3} \cdot 2^s)^2)^{\frac{1}{2}} \\ &\leq 2^{2s-4s^3} = C. \end{aligned}$$

Proof of part (iii). Let $\overline{G} \in \Lambda(B_m)$ and $G(\alpha) \neq 0$. We need a lower bound on $\|\overline{G}\|$. We will use a lower bound on $G(\alpha)$ provided by the following lemma:

Lemma 16 *Let $A(X), B(X) \in \mathbb{Z}[X]$ be non-constant and relatively prime. Suppose*

$$\begin{aligned} m &\geq \max\{\deg A, \deg B\}, \\ M &\geq 1 + \max\{\|A\|_\infty, \|B\|_\infty\}. \end{aligned}$$

Then at any root α of $A(X)$,

$$|B(\alpha)| > \frac{1}{m^m M^{3m}}.$$

Proof. Let $\deg A = k, \deg B = \ell$. So $k \geq 1, \ell \geq 1$ and there exist $U(X), V(X) \in \mathbb{Q}[X]$ such that $U = \sum_{i=0}^{\ell-1} u_i X^i$, $V = \sum_{i=0}^{k-1} v_i X^i$ and

$$U(X)A(X) + V(X)B(X) = 1. \tag{29}$$

We rewrite equation (29) using Sylvester's matrix $S = S(A, B)$ (§III.3):

$$(u_{\ell-1}, \ldots, u_0, v_{k-1}, \ldots, v_0) \cdot S = (0, \ldots, 0, 1)$$

Since $\det S$ is the resultant of the relatively prime A and B, we have $\det S \neq 0$. By Cramer's rule, $v_i = \frac{\det S_i}{\det S}$ where S_i is obtained by replacing a suitable row of S by the $(k+\ell)$-vector $(0, \ldots, 0, 1)$. By Hadamard's bound,

$$|v_i| \leq |\det S_i| \leq m^m M^{2m-1}$$

since each row of S has 2-norm at most $\sqrt{m}M$. Using Cauchy's bound that $|\alpha| \leq M$, we obtain

$$\begin{aligned} |V(\alpha)| &\leq |\sum_{i=0}^{k-1} v_i \alpha^i| \\ &\leq \sum_{i=0}^{k-1} (m^m M^{2m-1}) M^i \\ &\leq m^m M^{3m-1}. \end{aligned}$$

From (29) we get $V(\alpha) \cdot B(\alpha) = 1$ and hence $|B(\alpha)| \geq m^{-m} M^{1-3m}$. **Q.E.D.**

To conclude the proof of part(iii), let assume

$$\|G(X)\| \leq 2^{3s}$$

since otherwise

$$\|\overline{G}\| \geq c\|G(X)\| > c \cdot 2^{3s} = 2^s \cdot C,$$

and we are done. Note that $G(X)$ and $F(X)$ are relatively prime because $F(X)$ is a minimal polynomial. Applying lemma 16, we see that (with $M = 2^{3s}, m = s$),

$$|G(\alpha)| \geq s^{-s} 2^{-9s^2} = 2^{-s \log s - 9s^2} > 2^{s+1} C.$$

Applying lemma 15, we get $|G(\alpha) - G(\overline{\alpha})| \leq 2^{-4s^3} < C$ (just as (26) above). Finally,

$$\begin{aligned} |G(\overline{\alpha})| &\geq |G(\alpha)| - |G(\alpha) - G(\overline{\alpha})| \\ &\geq 2^{s+1} C - C > 2^s C. \end{aligned}$$

This completes the proof of the Correctness Theorem.

Remark: The constants c and C have $\Theta(s^3)$ bits each. In [99], similar constants use only $\Theta(s^2)$ bits but it seems that they only proved the equivalent of part (iii). It is our proof of part (i) that seems to require $\Omega(s^3)$ bits. Still, there is not much difference in complexity between computing roots to $\Theta(s^3)$ bits of accuracy as opposed to $\Theta(s^2)$ bits. This is because Newton iteration (§VI.11) can be used once we have about s^2 bits of accuracy.

Exercises

Exercise 6.1:

(i) Estimate the complexity of the Minimal Polynomial Algorithm.

(ii) Estimate the complexity of the factorization based on this minimal polynomial algorithm.

□

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture X
Linear Systems

Determinants and solving linear systems of equations are fundamental in algebraic computation.

The basic facts of this topic are well-known when the underlying algebraic structure R is a field. Gaussian elimination is the main method in this case. But our main interest is when R is a domain D. Of course, we can still embed a determinant computation in the quotient field Q_D of D. But this method turns out to be inefficient, for reasons similar to those that argue against computing the GCD over $D[X]$ via the Euclidean algorithm for GCD over $Q_D[X]$. In this lecture, special techniques will be described for three problems:

(i) Computing determinants. The method to be described has similarities to the subresultant PRS algorithm, and in the modern form, is due to Bareiss [10, 11]. Bareiss noted that the method is known to Jordan. This method seems quite effective when D is the integers or univariate polynomials [51] (see also Sasaki and Murao [175]).

(ii)&(iii) Computing Hermite and Smith normal forms of integer matrices. These have applications to lattice-theoretic questions, solving linear Diophantine equations, and finitely generated Abelian groups.

The results for computing determinants apply to any domain D. For the Hermite and Smith normal forms, we describe the results for $D = \mathbb{Z}$ although the basic method could be extended to any UFD.

The set of $m \times n$ matrices with entries in D is denoted $D^{m\times n}$. The (i, j)th entry of a matrix M is denoted $(M)_{i,j}$ or $(M)_{ij}$.

§1. Sylvester's Identity

Bareiss' algorithm for computing determinants of matrices over D is based on a determinantal identity which we derive in this section. When applied to $D = \mathbb{Z}$, Bareiss' algorithm is polynomial-time.

> It is instructive to understand why the usual Gaussian elimination is inadequate. Using fairly standard notations, suppose we initially have a matrix $M^{(1)}$ with L-bit integers. In the kth stage, we transform $M^{(k)}$ to $M^{(k+1)}$. The transformation applies to all entries of $M^{(k)}$ with index (i, j) where $i, j \geq k$. If the (i, j) entry of $M^{(k)}$ is $x_{ij}^{(k)}$, we have
>
> $$x_{ij}^{(k+1)} \leftarrow x_{ij}^{(k)} - x_{kj}^{(k)} \frac{x_{i1}^{(k)}}{x_{k1}^{(k)}}.$$
>
> The entries in $M^{(k+1)}$ are rational with bit size up to 4 times the bit sizes of entries in $M^{(k)}$. Thus after m steps, the entries can have bit size up to $4^m b$. It is an open problem if this exponential upper bound can be attained. Pivoting does not seem to help (see Exercise). Hence new methods are needed. Edmonds [62] appears to be the first to give a polynomial time solution for this problem.

Let $M \in D^{n\times n}$ with $(M)_{i,j} = x_{i,j}$. Note that x_{ij} may also be an indeterminate if D is suitably defined. The (i, j)-*cofactor* of M is denoted $[M]_{ij}$ and defined thus:

$$[M]_{ij} := (-1)^{i+j} \det M[i; j] \tag{1}$$

where $M[i;j]$ is the matrix obtained by deleting the ith row and jth column of M. The *adjoint* of M is the matrix $\mathtt{adj}(M)$ whose (i,j)th entry is the (j,i)-cofactor of M (note the transposed subscripts). For any $i,j = 1,\ldots,n$, it is easy to see that the sum

$$x_{i1}[M]_{j1} + x_{i2}[M]_{j2} + \cdots + x_{in}[M]_{jn}$$

is equal to $\det M$ if $i = j$, and equal to 0 if $i \neq j$. This immediately yields the following fundamental identity:

$$M \cdot \mathtt{adj}(M) = \det(M) \cdot I. \tag{2}$$

Taking determinants on both sides, it follows that $\det(\mathtt{adj}M) = \det(M)^{n-1}$. If $\det(M)$ is an invertible element of D, we infer that the inverse of M exists and is given by

$$M^{-1} = (\det(M))^{-1}\mathtt{adj}(M). \tag{3}$$

Let ABC denote a triple matrix product of shape $m \times n \times n \times p$ (so B is an n-square matrix). For any b from domain D, we have the identity

$$bABC = A(bB)C. \tag{4}$$

This follows by looking at the typical element of $bABC$:

$$\begin{aligned}
(bABC)_{rs} &= b\sum_{i=1}^{n} a_{ri}(BC)_{is} \\
&= b\sum_{i=1}^{n} a_{ri}\sum_{j=1}^{n}(B)_{ij}c_{js} \\
&= \sum_{i=1}^{n} a_{ri}\sum_{j=1}^{n}(bB)_{ij}c_{js} \\
&= \sum_{i=1}^{n} a_{ri}(bBC)_{is} \\
&= (A(bB)C)_{rs}.
\end{aligned}$$

Next we express the matrix M in the form

$$M = \begin{bmatrix} A & B \\ C & X \end{bmatrix}$$

where A is $(k-1)$-square and X is $(n-k+1)$-square. For the following derivation, assume

$$\delta = \det A \neq 0.$$

Lemma 1

$$\delta^{n-k}\det M = \det(\delta X - C \cdot \mathtt{adj}(A)B).$$

Proof. If we express

$$M = \begin{bmatrix} A & 0 \\ C & I \end{bmatrix}\begin{bmatrix} I & A^{-1}B \\ 0 & X - CA^{-1}B \end{bmatrix} \tag{5}$$

then we see that

$$\begin{aligned}
\det M &= \delta\det(X - CA^{-1}B) \\
\delta^{n-k}\det M &= \det(\delta X - \delta CA^{-1}B) \\
&= \det(\delta X - C \cdot \mathtt{adj}(A)B)
\end{aligned}$$

where the last step exploits equations (3) and (4). **Q.E.D.**

But what is $\delta X - C \cdot \mathtt{adj}(A)B$? To see this, introduce the "(r,s)-bordered matrix of order k" defined to be

$$M_{r,s}^{(k)} := \left[\begin{array}{cccc|c} x_{1,1} & x_{1,2} & \cdots & x_{1,k-1} & x_{1,s} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k-1} & x_{2,s} \\ \cdots & & \cdots & & \cdots \\ x_{k-1,1} & x_{k-1,2} & \cdots & x_{k-1,k-1} & x_{k-1,s} \\ \hline x_{r,1} & x_{r,2} & \cdots & x_{r,k-1} & x_{r,s} \end{array}\right]$$

for $k \leq \min\{r,s\}$. By definition, $M_{r,s}^{(1)}$ is the 1×1 matrix $[x_{r,s}]$. Also, define

$$x_{rs}^{(k)} := \det M_{rs}^{(k)}. \tag{6}$$

For instance, $M = M_{nn}^{(n)}$ and $\det M = x_{nn}^{(n)}$. Now let us look at a typical element of $\delta X - C \cdot \mathtt{adj}(A)B$: for $r \geq k$ and $s \geq k$, we have

$$\begin{aligned} (\delta X - C \cdot \mathtt{adj}(A)B)_{r-k+1,s-k+1} &= \delta x_{r,s} - (C \cdot \mathtt{adj}(A)B)_{r-k+1,s-k+1} \\ &= \delta x_{r,s} - \sum_{i=1}^{k-1} C_{r-k+1,i} \sum_{j=1}^{k-1} (\mathtt{adj}(A))_{ij} B_{j,s-k+1} \\ &= \delta x_{r,s} - \sum_{i=1}^{k-1} x_{r,i} \sum_{j=1}^{k-1} [A]_{ji} x_{j,s} \end{aligned}$$

where $[A]_{ji}$ is the (j,i)-cofactor of A. But the last expression can be seen to be equal to the determinant of the bordered matrix $M_{rs}^{(k)}$ (cf. exercise below giving the cofactors of $x_{r,i}x_{j,s}$). This proves:

Lemma 2

$$(\delta X - C \cdot \mathtt{adj}(A)B)_{rs} = x_{rs}^{(k)}.$$

Note that $\delta = x_{k-1,k-1}^{(k-1)}$. Combining the last two lemmas:

Lemma 3 (Sylvester's identity)

$$(x_{k-1,k-1}^{(k-1)})^{n-k} \det M = \det \left[\begin{array}{cccc} x_{k,k}^{(k)} & x_{k,k+1}^{(k)} & \cdots & x_{k,n}^{(k)} \\ x_{k+1,k}^{(k)} & x_{k+1,k+1}^{(k)} & \cdots & x_{k+1,n}^{(k)} \\ \vdots & & \ddots & \vdots \\ x_{n,k}^{(k)} & x_{n,k+1}^{(k)} & \cdots & x_{n,n}^{(k)} \end{array}\right].$$

Exercises

Exercise 1.1: (i) (Wilkinson 1961) With the notations for Gaussian elimination in the introduction, let us now assume total pivoting. Show that

$$|a_{ij}^{(k)}| \leq k^{1/2} \left(2 \cdot 3^{1/2} \cdot 4^{1/3} \cdot \cdots \cdot k^{1/(k-1)}\right)^{1/2} \|A\|_\infty.$$

NOTE: a rough estimate is that $\log\left(2 \cdot 3^{1/2} \cdot 4^{1/3} \cdot \cdots \cdot k^{1/(k-1)}\right)$ is $O(\log^2 k)$. Thus the magnitude of the entries are polynomial.
(ii) Why is it not obvious that this leads to a polynomial time solution for integer matrices?
(iii) (Open) Construct examples with exponential bit sizes in the intermediate entries. (You may assume that no pivoting is needed.) □

Exercise 1.2: (Bareiss) Show that Sylvester's identity holds even when $x^{(k-1)}_{k-1,k-1} = 0$. HINT: perturb the singular submatrix. □

§2. Fraction-free Determinant Computation

Lemma 3 with $n - k = 1$ is called the "first order" Sylvester identity. With the notations of the previous section, this identity for the matrix $M^{(k+1)}_{i,j}$ amounts to

$$x^{(k-1)}_{k-1,k-1} \det M^{(k+1)}_{ij} \quad = \quad \det \begin{bmatrix} x^{(k)}_{k,k} & x^{(k)}_{k,j} \\ x^{(k)}_{i,k} & x^{(k)}_{i,j} \end{bmatrix}.$$

Hence,

$$x^{(k+1)}_{ij} = \frac{x^{(k)}_{k,k} x^{(k)}_{i,j} - x^{(k)}_{i,k} x^{(k)}_{k,j}}{x^{(k-1)}_{k-1,k-1}}. \tag{7}$$

The important point is that the division by $x^{(k-1)}_{k-1,k-1}$ in this equation is exact (*i.e.*, with no remainder). Equation (7) is the defining step of the fraction-free Gaussian elimination algorithm of Bareiss (and Jordan):

Bareiss Algorithm
Input: M an n-square matrix,
assuming its principal minors $x^{(k)}_{kk}$ are all non-zero.
Output: The matrix entry $(M)_{n,n}$ contains the determinant of M.
In general, for $i \geq k$, we have
$(M)_{ik} = x^{(k)}_{ik}$, $(M)_{ki} = x^{(k)}_{ki}$.

1. $(M)_{0,0} \leftarrow 1$; {Note: $(M)_{0,0}$ is a special variable.}
2. for $k = 1, \ldots, n-1$ do
3. for $i = k+1, \ldots, n$ do
4. for $j = k+1, \ldots, n$ do
5. $(M)_{ij} \leftarrow \frac{(M)_{ij}(M)_{kk} - (M)_{ik}(M)_{kj}}{(M)_{k-1,k-1}}$

The program structure of this algorithm amounts to a simple triple-loop, as in the standard Gaussian elimination. Its correctness is easily shown by induction on k, and by appeal to equation (7) (we leave this as an exercise).

In case the assumption about principal minors turns out to be false, this is easily detected and the algorithm may be aborted. Alternatively, it is not hard to add the code (between lines 2 and 3) to perform some kind of pivoting: say, if $(M)_{k-1,k-1} = 0$ and some $(M)_{k-1,i} \neq 0$ $(i = k, \ldots, n)$ then

we can exchange the ith column with the $k-1$st column. But we defer a more complete discussion of this to an extension of Bareiss' algorithm below.

The division in line 5 is exact since $(M)_{k-1,k-1} = x^{(k-1)}_{k-1,k-1}$. Hence all computed values remain inside the domain D.

This is an "in-place" algorithm that destroys the contents of the original matrix. But we can easily preserve the original matrix is desired. The output M has the following shape:

$$M = \begin{bmatrix} x^{(1)}_{1,1} & \cdots & & & & x^{(1)}_{1,n} \\ \vdots & \ddots & & & & \\ & & x^{(k)}_{kk} & x^{(k)}_{k,k+1} & \cdots & x^{(k)}_{k,n} \\ & & x^{(k)}_{k+1,k} & \ddots & & \\ & & \vdots & \ddots & & \\ x^{(1)}_{n,1} & & x^{(k)}_{n,k} & & & x^{(n)}_{n,n} \end{bmatrix}$$

In other words, for each $k = 1, \ldots, n$, there is an (rotated) L-shaped band in M that contains determinants of order k bordered matrices, as indicated.

In view of the definition of $x^{(k)}_{ij}$ (equation (6)) as subdeterminants of M, we obtain at once:

Lemma 4 *The intermediate values encountered in the algorithm have absolute values at most $n^n 2^{Ln}$ where 2^L bounds the absolute values of the entries of M.*

Since the algorithm takes $O(n^3)$ arithmetic steps, and each entry has at most $n(\log n + L)$ bits, we conclude that the bit-complexity of this method is

$$O(n^3 \mathrm{M}_B(n(\log n + L)))$$

where $\mathrm{M}_B(s)$ is the bit-complexity of multiplying two s-bit integers.

One can exploit the higher order Sylvester identities to obtain analogous algorithms. We will not explore this but see Bareiss [10] for an analysis of an algorithm exploiting the second order identity. We will instead describe two other extensions.

Extended Bareiss Algorithm. We extend the algorithm to matrices of arbitrary $m \times n$ shape and of general rank ρ. This can be formulated as follows:

> (*) *Given a matrix $A \in \mathbb{Z}^{m \times n}$, we want to compute its rank ρ, a permutation P of its rows, a permutation Q of its columns, and non-zero values $d_1, \ldots, d_\rho$, such that each d_i is the ith principal minor of PAQ.*

This will be needed for the Hermite normal form algorithm in §7.

Partially Processed Matrices. To describe the solution, it is useful to introduce some terminology. Let $A, M \in \mathbb{Z}^{m \times n}$. We say that M is a *partially processed version* of A if each (i,j)th entry

$(M)_{i,j}$ is an (i,j)-bordered determinant of A of some order $k = k(i,j)$. Clearly we have

$$1 \leq k \leq \min\{i,j\}.$$

If $k(i,j) = \min\{i,j\}$ (respectively, $k(i,j) = 1$), we say the (i,j)th entry is *fixed* (resp., *original*). We call $k(i,j)$ the *fixing order* of the (i,j)th entry. For instance, A is a partially processed version of itself, with every entry original. If every entry of M is fixed (for A), then M is said to be *completely processed* (for A). In this terminology, we can view Bareiss' algorithm on input A as trying to fix every entry of A. An operation of the form

$$(M)_{i,j} \leftarrow \frac{(M)_{k,k}(M)_{i,j} - (M)_{i,k}(M)_{k,j}}{(M)_{k-1,k-1}} \tag{8}$$

is called a *fixing step* for the (i,j)th entry. The fixing step is *valid* provided the fixing order of $(M)_{k,k}$, $(M)_{i,m}$, $(M)_{i,k}$ and $(M)_{k,m}$ is k and the fixing order of $(M)_{k-1,k-1}$ is $k-1$. If M is partially processed before a valid step for the (i,j)th entry, it remains partially processed after the step, with the fixing order of the (i,j) entry equal to $k+1$ (cf. equation (10)).

Let us return to problem (*). Suppose M is initialized to A and the row and column permutations P, Q are initially identities. Inductively, assume M is a partially processed version of PAQ. We proceed in stages. In stage s ($s = 1, \ldots, \rho$) our goal is to ensure the following properties:
(i) Every entry in the sth principal submatrix of M is fixed.
(ii) The first s diagonal entries of M are non-zero.
(iii) To maintain an index $m_0 \geq s$ such that each row $i > m_0$ is known to be dependent on the first $s-1$ rows. Moreover, the entries in rows $s+1$ to m_0 are original. Initially, $m_0 = m$.

For $s \geq 1$, suppose that stage $s-1$ is completed. [For $s = 1$, the properties (i)-(iii) are vacuously true.] Observe that we can fix the first $s-1$ entries of the sth row of C in $O(s^2)$ fixing steps. Applying this observation once more, we can fix the first s entries in the sth column in $O(s^2)$ fixing steps. Thus goal (i) is attained. The problem is that the (s,s)th entry may be zero. We may go on to fix the first s entries of column j for $j = s+1, s+2, \ldots, n$. There are two cases:
(A) We find a column j which can be exchanged with column s to satisfy goal (ii). Then we exchange column j and column s, and update permutation Q. If $s = m_0$, we halt, else go to stage $s+1$.
(B) No such column exists. In this case we conclude that row s is dependent on the first $s-1$ rows. If $m_0 = s$, we stop, since $\rho = s-1$. Otherwise, we exchange row m_0 with row s, decrement m_0, update permutation P and continue with stage s.

This completes the description of our solution to problem (*). For reference, call this the *Extended Bareiss Algorithm*. It's correctness is clear. For its complexity, first note that each matrix entry is a minor of A and hence has bit-size at most $L' := \rho(\lg \rho + L)$ where $L = \lg \|A\|_\infty$. [Recall (§0.9) that $\lg = \log_2$.] Second, we use $O(mn\rho)$ arithmetic operations on integers of bit-size $O(\rho(\lg \rho + L))$. This is because the fixing step for each entry is applied at most $\rho + 1$ times. To see this, we verify two facts: (a) In case (A), when we exchange columns s and j, note that the first s entries in the columns between s and j have been fixed. This information is still valid after the column exchange. So we just need a Boolean flag at each entry to indicate whether it is fixed or not. (b) In case (B), note that after the exchange between rows s and row m_0, the entire row m_0 is already fixed and row s is original. Thus we have:

Theorem 5 *On input $A \in \mathbb{Z}^{m\times n}$ with rank $\rho \geq 1$, the Extended Bareiss Algorithm has bit complexity $O(mn\rho \mathrm{M}_B(L'))$ where $L' = \rho(\lg \rho + L)$ and $L = \lg \|A\|$.*

A Preprocessing Problem. We note another useful extension to Bareiss' algorithm. If $N \in \mathbb{Z}^{m\times(m-1)}$, let N_i ($i = 1, \ldots, m$) be the submatrix of N with the ith row deleted. Consider this problem:

(**) *Given $N \in \mathbb{Z}^{m\times(m-1)}$, compute the m subdeterminants*

$$\det N_1, \ldots, \det N_m. \tag{9}$$

This is a "preprocessing problem" in the sense that once the $\det N_i$'s are available, for any given column $a = (a_1, \ldots, a_m)^T \in \mathbb{Z}^m$, if A is obtained by appending a to N, we can quickly compute $\det A$ as

$$\det A = \sum_{i=1}^{m} a_i (-1)^{i+m+1} \det N_i.$$

The solution is simple: let M be the $m \times m$ matrix obtained from N by appending a column

$$Z = (Z_1, \ldots, Z_m)^T$$

where the Z_i's are indeterminates. Let $X = (x_{i,j})_{i,j=1}^{m,m}$ be the output matrix when M is input to Bareiss' algorithm. The entries $x_{i,j}$ are computed as expected. But the entry $x_{i,m}$ in the last column is a linear function in $Z_1, \ldots, Z_i$, which we denote by $f_i(Z_1, \ldots, Z_i)$.

In the notations of §1, let $x_{i,j}^{(k)}$ denote the (i,j)-bordered determinant of M of order k $(i, j \geq k)$. Thus each output entry $x_{i,j}$ is equal $x_{i,j}^{(k)}$ where $k = \min\{i, j\}$. Let $\delta_k = x_{k,k}^{(k)}$. For example, the reader may verify that

$$\begin{aligned} f_1 &= Z_1, \\ f_2 &= \delta_1 Z_2 - x_{2,1} Z_1, \\ f_3 &= \delta_2 Z_3 - x_{3,2} Z_2 - \left(\frac{\delta_2 x_{3,1} - x_{3,2} x_{2,1}}{\delta_1} \right) Z_1. \end{aligned}$$

For any $(a_1, \ldots, a_m) \in \mathbb{Z}^m$, the value of $f_m(a_1, \ldots, a_m)$ yields the determinant of the matrix M where each Z_i is replaced by a_i. Thus, up to signs, the desired subdeterminants (9) may be read off from the coefficients of f_m.

What is the complexity of this procedure? The bit-sizes of entries in the first $m-1$ columns of M and the time to compute them are exactly as in Bareiss' algorithm. Consider the mth column. In stage $k+1$ $(k = 1, \ldots, m-1)$ of the outermost for-loop in Bareiss' algorithm, the entries of the column m that are computed are $x_{i,m}^{(k+1)}$ $(i = k+1, \ldots, m)$. We have

$$x_{i,m}^{(k+1)} = \frac{x_{k,k}^{(k)} x_{i,m}^{(k)} - x_{i,k}^{(k)} x_{k,m}^{(k)}}{x_{k-1,k-1}^{(k-1)}} \tag{10}$$

$$= \frac{\delta_k x_{i,m}^{(k)}(Z_1, \ldots, Z_{k-1}, Z_i) - x_{i,k}^{(k)} f_k(Z_1, \ldots, Z_k)}{\delta_{k-1}}. \tag{11}$$

Each of the linear functions $x_{i,m}^{(k)}(Z_1, \ldots, Z_{k-1}, Z_i)$ and $f_k(Z_1, \ldots, Z_k)$ has k coefficients that are minors of N of order $k-1$, and hence has bit-size at most $m(L + \lg m)$. Hence it takes takes $O(k)$ arithmetic operations to compute $x_{i,m}^{(k+1)}$. Summing over the cost for computing the entries of column m, we again have $O(m^3)$ arithmetic operations on integers of bit size $O(m(L + \lg m))$. So the overall complexity is exactly as in the original Bareiss' algorithm.

Exact Division. Exact division turns out to be slightly more efficient than division-with-remainder (by a small constant factor). We briefly describe the method (see Jebelean [92] for more details). Suppose $C = A \cdot B$ is an integer product. Consider the problem of computing B given

C, A where integers are in base β. Let $0 \leq \ell(A) < \beta$ denote the least significant digit (LSD) of A. Then it is easy to see that $\ell(C) \equiv \ell(A)\ell(B) (\mathrm{mod}\, \beta)$. Thus

$$\ell(B) = (\ell(C) \cdot (\ell(A)^{-1}) \,\mathbf{mod}\, \beta)$$

provided $\ell(A)$ is invertible mod β. For simplicity, assume β is prime so that this is always possible. This immediately leads to an exact division algorithm that produces the digits of B sequentially, beginning with the LSD of B. Clearly, this is the opposite of the classical division algorithm [105], and avoids the "guess-and-correct" step of the classical method:

```
Exact Division
    Input: A, C ∈ ℤ in a prime base β, A|C.
    Output: B such that AB = C.
NORMALIZATION:
1.      while β|A do
2.          A ← A/β; C ← C/β.
MAIN LOOP:
3.      while C > 0 do
4.          b ← ℓ(C)/ℓ(A) mod β;
5.          Output b;
6.          C ← (C − b · A)/β.
```

Let $\mathtt{len}(A)$ denote the number of digits in A. Step 6 is considered the inner loop. To speed up this step, observe that only the lowest $\mathtt{len}(B)$ digits of C are involved in the inner loop. Hence the main loop can be improved as follows:

```
        ...
MAIN LOOP:
3.      for k ← (len(C) − len(A) + 1) downto 1
4.          b ← ℓ(C)/ℓ(A) mod β;
5.          Output b;
6.          C ← ((C − b · A) mod β^k)/β.
```

If β is a power of 2, then $\ell(A)$ would be invertible if A is odd. We achieve this by a simple modification to the normalization stage, namely, by inserting steps 2.1 and 2.2 below:

```
NORMALIZATION:
1.      while β|A do
2.          A ← A/β; C ← C/β.
2.1     while 2|A do
2.2         A ← A/2; C ← C/2.
MAIN LOOP:
            ...
```

At the end of normalization, A is odd, and hence $\ell(A)$ is odd. This guarantees that $\ell(A)$ is invertible. The bit analysis of this algorithm is left to an exercise.

Exercises

Exercise 2.1: M is a square matrix with $(M)_{ij} = x_{ij}$ for all i, j. The "cofactor" of $x_{ij}x_{i'j'}$ is defined to be the expression E that is multiplied by $x_{ij}x_{i'j'}$ when we collect terms in the determinant of M that involve both x_{ij} and $x_{i'j'}$. E.g., if M is 2×2, the cofactor of $x_{12}x_{21}$ is -1 and the cofactor of $x_{11}x_{22}$ is 1. If $i = i'$ or $j = j'$ then $E = 0$; otherwise, show that

$$E = (-1)^{i+i'+j+j'+\delta} \det M[i, i'; j, j']$$

where

$$\delta = \begin{cases} 1 & \text{if } (i > i') \oplus (j > j'), \qquad (\oplus \text{ is exclusive-or}) \\ 0 & \text{else,} \end{cases}$$

and $M[i, i'; j, j']$ is the submatrix of M obtained by deleting rows i and i' and deleting columns j and j'. □

Exercise 2.2: Show that $\texttt{adj}(\texttt{adj}A) = \det(A)^n A$. □

Exercise 2.3: What is the 3×3 matrix analogue of equation (5)? □

Exercise 2.4: Carry out the dth order version of Bareiss algorithm, by exploiting the order d Sylvester identity. For instance, for $d = 2$, we must construct $x_{ij}^{(k)}$ for even values of k, evaluating 3×3 determinants. □

Exercise 2.5: Modify Bareiss' algorithm in order to compute the determinant of a matrix with rational entries. Carry out comparison experiments in this setting of rational inputs. HINT: make each row have a common denominator first. □

Exercise 2.6: Suppose M is $n \times m$ where $n \leq m$. In Bareiss' algorithm, we replace line 4 with "for $j = k+1, \ldots, m$ do" (*i.e.*, we extend change the limit from n to m). Describe the contents of the entries $(M)_{n,j}$ for $j = n, n+1, \ldots, m$. What is the complexity of this modification? How is this related to the determinantal polynomials (§III.3)? □

Exercise 2.7: In the exact division algorithm, show that when the length m of C is less than twice the length n of A, this method uses no more than half the number of bit-operations of the classical method. Quantify this bit-operation count more generally in terms of m, n. □

§3. Matrix Inversion

Matrix inverse is easily computed using the standard Gaussian elimination procedure. Following [69], let us compute the more general product

$$CA^{-1}B$$

where the triple product CAB has shape $m \times n \times n \times p$. We proceed as follows: apply Gaussian elimination to the $(m+n) \times (n+p)$ matrix

$$M = \begin{bmatrix} A & B \\ -C & \mathbf{0} \end{bmatrix}, \qquad (A \text{ is nonsingular}) \tag{12}$$

and obtain (after operations that zero the first n columns below A)

$$M' = \begin{bmatrix} A' & B' \\ \mathbf{0} & D' \end{bmatrix}. \tag{13}$$

Here $\mathbf{0}$ denotes a matrix of zeros. Note that if A is singular, we would have discovered this fact during Gaussian elimination. Henceforth, assume A is non-singular. We claim that D' is precisely what we wanted:

Lemma 6

$$D' = CA^{-1}B.$$

Block Gaussian elimination. This lemma is a slight generalization of lemma 1 to the non-square case. It is instructive to briefly consider Gaussian elimination for block-size elements. Let

$$M = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}$$

where A_{ij} is a $k_i \times \ell_j$ matrix ("block"). We may say that M has shape $(k_1, \ldots, k_m) \times (\ell_1, \ldots, \ell_n)$. Consider the following transformation of M: for $1 \le r, s \le m$, $r \ne s$, and any $k_r \times k_s$ matrix H, we replace the rth row of M with the sum of the rth row and H right multiplied by the sth row. That is,

$$A_{rt} \leftarrow A_{rt} + HA_{st}, \qquad (\text{for } t = 1, \ldots, n).$$

This is equivalent to left multiplying M by the matrix

$$T_{r,s}(H) := \begin{bmatrix} I_1 & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ \mathbf{0} & \cdots & I_r & \cdots & H & \cdots & \mathbf{0} \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & I_s & \cdots & \mathbf{0} \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \cdots & I_m \end{bmatrix}$$

where each I_j is the k_j-square identity matrix, the I_j's occurring along the main diagonal, $\mathbf{0}$-blocks occurring everywhere else except at the (r, s)th position which is occupied by H. Clearly, for any minor Δ of M (now viewed as an ordinary matrix of shape $(\sum_{i=1}^m k_i) \times (\sum_{j=1}^n \ell_j)$) that contains rows in the ith and jth block rows of M, this operation preserves Δ. If $m = n$ and for all i, $k_i = \ell_i$ and $A_{i,i}$ is non-singular, then we can effect Gaussian elimination at this block level: for instance, $T_{i1}(A_{i1}A_{11}^{-1})$ will make the $(i, 1)$th entry zero. In particular, if A is square and non-singular, we may transform the matrix

$$N = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{14}$$

to

$$N' = \begin{bmatrix} A' & B' \\ \mathbf{0} & D - CA^{-1}B \end{bmatrix}.$$

We may conclude:

Lemma 7 *The rank of N is equal to the rank of A iff $D = CA^{-1}B$.*

We complete the proof of lemma 6. Suppose, instead of M in equation (12), we had

$$P = \begin{bmatrix} A & B \\ -C & -CA^{-1}B \end{bmatrix}.$$

Then by lemma 7, the rank of P is equal to the rank of M. Applying Gaussian elimination to P to zero the entries below A, we obtain

$$P' = \begin{bmatrix} A' & B' \\ \mathbf{0} & D' - CA^{-1}B \end{bmatrix}.$$

The matrices A', B', D' here are the same as those in M' (equation (13)) since the same multipliers were used to do the elimination. But the rank of P' equals the rank of P and hence of M. This can only mean that $D' - CA^{-1}B = \mathbf{0}$, as we wanted shown.

Exercises

Exercise 3.1: Let N be square in equation (14).
(i) Show that $\det N = \det(AD - ACA^{-1}B)$ if A^{-1} exists.
(ii) Under what conditions is $\det N = \det(AD - CB)$? $\det N = \det(AD - BC)$?
(iii) Consider the four cases of these identities depending on which of the 4 blocks of N are non-singular.
(iv) Obtain the corresponding formulas using column operations (multiplication from the right). □

Exercise 3.2: Use the method to compute $CA^{-1}B$ to solve the system $Ax = b$ of linear equations. □

Exercise 3.3: For $n > 1$, let U be the n-square matrix all of whose entries are 1 and let $S = U - I$. So S has zeros along its main diagonal. Compute S^{-1} using the above algorithm for small n. NOTE: $S^{-1} = \frac{1}{n-1}U - I$. □

Exercise 3.4: Let $T(n)$ be the algebraic complexity of computing the determinant of an n-square matrix. Show that $T(n) = O(\mathrm{MM}(n))$ where $\mathrm{MM}(n)$ is the complexity of multiplying two n-square matrices. □

§4. Hermite Normal Form

Let $A, B \in \mathbb{Z}^{m\times n}$. In §VIII.1, we viewed the columns of A as a generating set of the lattice $\Lambda(A) \subseteq \mathbb{Z}^m$. Alternatively, we may regard $\Lambda(A)$ as a subgroup of the Abelian group $\mathbb{Z}^m$. We have shown (§VIII.1)

$$\Lambda(A) = \Lambda(B) \text{iff } \ A = B \cdot U \tag{15}$$

for some integer unimodular matrix U. The original proof requires A and B to be bases but this assumption can be dropped (see below). This raises the question: how can we decide if two matrices A, B generate the same subgroup (or lattice)? The result (15) does not appear to be helpful for this purpose – there is no obvious way to find U even if one is known to exist. In this lecture, and unlike §VIII.1, we will no longer assume that A has rank n, so the columns of A form a generating set but

not necessarily a basis of $\Lambda(A)$. The tool for answering such questions about $\Lambda(A)$ is a normal form which we now describe.

By applying the elementary *integer* column operations (§VIII.1) to A, we can transform A to a matrix $H = H(A)$ of the following shape:

1. For some $r = 1, \ldots, n$, the first r columns of H are non-zero and the remaining columns (if any) are zero.
2. For $i = 1, \ldots, r$, let $(H)_{c(i),i}$ be the first non-zero entry of column i. Then
$$1 \leq c(1) < c(2) < \cdots < c(r) \leq n. \tag{16}$$
3. For $i = 1, \ldots, r$, $(H)_{c(i),i} > 0$.
4. If $1 \leq j < i \leq r$ then $0 \leq (H)_{c(i),j} < (H)_{c(i),i}$.

Such a matrix $H = H(A)$ is said to be in *Hermite normal norm* (HNF), or H is *the HNF of A*. For instance, the following 6×4 matrix H_1 is in HNF:

$$H_1 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 2 & 10 & -4 & 0 \\ 9 & -3 & 0 & 0 \end{bmatrix}. \tag{17}$$

If only the first two conditions in the definition of HNF hold, H is said to be in *column echelon form* Column echelon form is just a generalization of "lower triangular matrices". If H is a column echelon form of A, then rows $c(1), \ldots, c(r)$ are called the *critical rows* of H. Each of the entries $(H)_{c(i),i}$ is called a *critical entry* of H. The number r of non-zero columns in H is called the *rank* of A. For the matrix H_1 in (17), the critical rows are the 1st, 3rd and 4th, the critical entries are $(1,1), (3,2)$ and $(4,3)$.

Since the elementary column operations preserve the underlying lattice, $\Lambda(A) = \Lambda(H(A))$. It is then easy to see: *row i is critical in $H(A)$ iff there is a vector $\xi \in \Lambda(H(A))$ whose first non-zero entry is in the ith position.* This latter characterization of critical rows depends only on the lattice. This shows that the set of critical rows of $H(A)$ depends only on $\Lambda(A)$.

Theorem 8 *Let A, B have the same shape. $\Lambda(A) = \Lambda(B)$ if and only if A and B have the same Hermite normal form: $H(A) = H(B)$.*

Proof. If $H(A) = H(B)$ then clearly $\Lambda(A) = \Lambda(B)$. In the other direction, let us assume $\Lambda(A) = \Lambda(B)$. We wish to prove $H(A) = H(B)$. First observe that A and B have the same rank, d. This is because A, B have the same set of critical rows, by a preceding observation. But d is the number of critical rows.

We use induction on d. We assume that $d \geq 1$ since $d = 0$ is trivial. Let a, b be the first columns of $H(A)$ and $H(B)$, respectively. Let $c(1)$ and $c'(1)$ denote the first non-zero entry of a and b, respectively. It is easy to see that $c(1) = c'(1)$. Indeed, these two entries must be identical since they each generate all the $c(1)$th entries of vectors in $\Lambda(A) = \Lambda(B)$.

Let $H_1(A)$ and $H_1(B)$ be obtained by deleting a and b, respectively. Notice that $\Lambda(H_1(A))$ generates the subgroup of $\Lambda(A)$ whose first $c(1)$ entries are zero. Similarly $\Lambda(H_1(B))$ generates the subgroup of $\Lambda(B)$ whose first $c(1)$ entries are zero. Since $\Lambda(A) = \Lambda(B)$, we conclude that $\Lambda(H_1(A)) = \Lambda(H_1(B))$. Since $H_1(A)$ and $H_1(B)$ are in HNF, we deduce by induction that $H_1(A) = H_1(B)$.

It remains to prove that $a = b$. Suppose $a - b$ is not identically zero. It follows from the preceding that the first $c(1)$ entries of $a-b$ are zero. If k is the first non-zero column of $a-b$, then $k > c(1)$ and it follows that there exists a column c in $H_1(A)$ whose first non-zero entry is in position k. If $(a)_k$ denotes the kth component of a, then by the definition of HNF, $(c)_k > (a)_k \geq 0$ and $(c)_k > (b)_k \geq 0$. Hence $(a-b)_k$ has absolute value less than $(c)_k$. This is impossible since $a - b \in \Lambda(H_1(A))$ means that $(a-b)_k$ must be a multiple of $(c)_k$. **Q.E.D.**

As corollary, we also see that (15) holds without the restriction that the columns of A, B (respectively) are linearly independent. It suffices to show that if $\Lambda(A) = \Lambda(B)$ then there is a unimodular matrix U such that $AU = B$. But there are unimodular matrices U_A, U_B such that $AU_A = H(A) = H(B) = BU_B$. Take $U = U_A U_B^{-1}$.

It follows that our basic question of deciding if $\Lambda(A) = \Lambda(B)$ is reduced to computing and comparing their HNF's. Other questions such as checking whether a given vector ξ belongs to $\Lambda(A)$ can similarly be answered (Exercise). We next address the computation of HNF.

Generic HNF algorithm. Assume the input is the $m \times n$ matrix A. By the "subrow" at an entry $(A)_{i,j}$ we mean the set of entries of the form $(A)_{i,k}$ for $k = j, j+1, \ldots, n$. The best way to understand this algorithm is to imagine that our task (the main loop below) is to determine the critical rows of A.

Generic HNF Algorithm:
Input: an $m \times n$ integer matrix A.
Output: the Hermite normal norm of A.
MAIN LOOP:
1. Initialize $i \leftarrow 1$ and $j \leftarrow 1$.
2. while $i \leq m$ and $j \leq n$ do:
3. While the subrow at $(A)_{ij}$ has only zero entries, increment i.
4. Now assume the subrow at $(A)_{ij}$ has non-zero entries.
5. By adding multiples of one column to another, eliminate all but one non-zero element in the subrow at $(A)_{ij}$.
6. By a column-exchange, bring this sole non-zero entry to position (i, j) and increment j.
end{while}
CLEAN UP:
7. At this point, the matrix is in column-echelon form.
8. By adding multiples of one column to another, achieve the remaining two conditions for the definition of HNF.

Exercise 4.1: (i) There are several ways to fill in details in the generic algorithm. Describe a reasonable choice.
(ii) For your version of the generic algorithm, analyze the number of arithmetic operations for a $m \times n$ input matrix whose entries are L-bit integers.
(iii) Bound the bit complexity of your algorithm. □

Although part (i) of the exercise should give a polynomial bound in m, n, L, the bit complexity in part (ii) should be exponential in $\min\{m, n\}$. Note that these exponential bounds arise from potential sizes of intermediate matrix entries; the final entries in the HNF can be shown to be polynomially bounded (below). Nevertheless, it is an open problem to construct examples that actually exhibit exponential behavior. In random examples, huge intermediate entries routinely appear. For instance, Hafner and McCurley [77] reported that for random 20×20 matrices with entries between 0 and 10, most gave rise to an entry exceeding 10^{500}. One example has an entry exceeding 10^{5011}. We will develop a modular technique to achieve polynomial bit-complexity bounds.

Invariants of the Hermite normal form. Let $A \in \mathbb{Z}^{m\times n}$ and $1 \le k \le \min\{m, n\}$. For $1 \le i_1 < i_2 < \cdots < i_k \le m$, let $A(i_1, i_2, \ldots, i_k)$ denote the submatrix of A formed by rows $i_1, \ldots, i_k$. Let $\gamma(A; i_1, \ldots, i_k)$ denote the GCD of all the $k \times k$ subdeterminants (*i.e.*, order k minors) of $A(i_1, i_2, \ldots, i_k)$. Note that $\gamma(A; i_1, \ldots, i_k) = 0$ iff every order k minor of $A(i_1, i_2, \ldots, i_k)$ is zero.

Lemma 9 *The elementary integer column operations on A preserve $\gamma(A; i_1, i_2, \ldots, i_k)$.*

Proof. The column operations that interchange two columns or multiply a column by -1 do not change the GCD (since GCD is defined up to associates and we always pick the positive member.) Suppose $c \in \mathbb{Z}$ times the ith column of A is added to the jth column. Certain of the $k \times k$ subdeterminants of $A(i_1, i_2, \ldots, i_k)$ change. If a subdeterminant value D is changed, say to D', it is easy to see that $D - D' = \pm c \cdot D''$ where D'' is another subdeterminant. Moreover, D'' is among the subdeterminants of $A(i_1, \ldots, i_k)$ that have not changed. To see this, observe that a subdeterminant D is changed iff D involves column j but not column i. Schematically, if the old GCD is $\texttt{GCD}(\ldots, D, \ldots, D'', \ldots)$ then the new one is

$$\texttt{GCD}(\ldots, D', \ldots, D'', \ldots) = \texttt{GCD}(\ldots, D \pm c \cdot D'', \ldots, D'', \ldots).$$

But the later expression is equal to the old GCD. **Q.E.D.**

Let r be the rank of A. For $k = 1, \ldots, r$, we use the shorthand

$$\gamma_k(A) := \gamma(A; c(1), c(2), \ldots, c(k)).$$

We define $\gamma_i(A) = 0$ for $i = r + 1, \ldots, n$.

Corollary 10 *The value $\gamma_k(A)$ is invariant under the elementary column operations on A. If H is the Hermite normal form of A, then the product of the first k critical values of H is equal to $\gamma_k(A)$. In particular, $\gamma_k(A)$ divides $\gamma_{k+1}(A)$ for $k = 1, \ldots, n-1$.*

Exercises

Exercise 4.2: Describe the HNF of an $m \times 1$ matrix. □

Exercise 4.3: Discuss other strategies for implementing the generic HNF algorithm. □

Exercise 4.4: Solve the following problem efficiently:
(i) Given $x \in \mathbb{Z}^m$ and $A \in \mathbb{Z}^{m\times n}$, decide if $x \in \Lambda(A)$. If $x \in \Lambda(A)$, find the n-vector ξ such that $A\xi = x$.
(ii) Given $A, B \in \mathbb{Z}^{m\times n}$, check whether $\Lambda(A) = \Lambda(B)$. In case they are equal, construct matrices U, V such that $A = BU$ and $B = AV$. [These matrices need not be unimodular unless A, B are bases for $\Lambda(A)$.] □

Exercise 4.5: Suppose A, B are both in column-echelon form, and $\Lambda(A) \subseteq \Lambda(B)$. If the critical entries in A and the critical entries in B are in the same position and corresponding critical entries have the same values, then $\Lambda(A) = \Lambda(B)$ □

Exercise 4.6: (i) Every subgroup of $\mathbb{Z}^n$ is finitely generated.
(ii) Every finitely generated Abelian group G is isomorphic to a subgroup of $\mathbb{Z}^n$ for some n. □

Exercise 4.7: Call H a *generalized HNF* for A if H is in HNF and obtained from A by the usual elementary column operations, but now we allow the permutation of rows as well. How do the various generalized HNF's of A relate to each other? □

Exercise 4.8: (Open) Given a matrix $A \in \mathbb{Z}^{m\times n}$ and its Hermite normal form $H(A)$. Let L bound the bit sizes of entries in A and $H(A)$. What is best upper bound $B(m, n, L)$ such that there exists a sequence of elementary integer column operations from A to $H(A)$ where all intermediate entries have bit size at most $B(m, n, L)$? □

§5. A Multiple GCD Bound and Algorithm

Computing multiple GCD's is a key subproblem in Hermite normal forms. For example, in the generic HNF algorithm (§4), the process of zeroing out all but one entry of a subrow amounts to computing the multiple GCD of the entries in the subrow. Of course, multiple GCD can be reduced to simple GCD, *i.e.*, GCD for two elements. But this is not the most efficient method. In this section, we present an efficient multiple GCD algorithm over integers. This is based on a co-factor bound that we first derive.

In the following, fix

$$(a_1, a_2, \ldots, a_k) \qquad (k \geq 2)$$

such that the a_i's are distinct and positive. Let $d = \texttt{GCD}(a_1, \ldots, a_k)$. By definition, an integer sequence $(s_1, \ldots, s_k)$ is called a *co-factor* of $(a_1, \ldots, a_k)$ if

$$d = s_1a_1 + s_2a_2 + \cdots + s_ka_k.$$

The "co-GCD problem" refers to the problem of computing a co-factor of $(a_1, \ldots, a_k)$. Note that once a co-factor is available, we can easily compute the GCD. Our first goal is to prove the existence of a co-factor with each $|s_i|$ upper bounded by a_1. We use an argument of Hongwei Cheng:

Lemma 11 *If $d = \texttt{GCD}(a_1, \ldots, a_k)$ then there exists a co-factor $(s_1, \ldots, s_k)$ for $(a_1, \ldots, a_k)$ such that*

$$|s_1| \quad \leq \quad \frac{a_k}{2},$$

$$\begin{aligned} |s_i| &< \frac{a_{i-1}+a_k}{2} \qquad i = 2, \ldots, k-1, \\ |s_k| &\leq \frac{d}{a_k} + \frac{a_{k-1}}{2}. \end{aligned}$$

Proof. Suppose $(t_1, \ldots, t_k)$ is any co-factor for $(a_1, \ldots, a_k)$. Define

$$s_i := \begin{cases} t_i - q_i a_k & i = 1, \ldots, k-1 \\ t_k + \sum_{j=1}^{k-1} q_j a_j & i = k, \end{cases}$$

where $q_1, \ldots, q_k \in \mathbb{Z}$ are to be specified. It is not hard to check that $(s_1, \ldots, s_k)$ is also a co-factor for $(a_1, \ldots, a_k)$. We now define the q_i's inductively. Pick q_1 to be the symmetric quotient (§II.3) of t_1 divided by a_k. Then $|s_1| \leq a_k/2$, as desired. Now consider the partial sums

$$S_i = \sum_{j=1}^{i} s_j a_j.$$

Thus $S_1 = a_1 s_1$ and $|S_1| \leq a_1 a_k/2$. Inductively, assume S_{i-1} has been defined so that $|S_{i-1}| \leq a_{i-1}a_k/2$. For $i = 2, \ldots, k-1$, define q_i so that $|S_i| \leq a_i a_k/2$. This is clearly possible since $S_i = S_{i-1} + a_i s_i = S_{i-1} + a_i(t_i - q_i a_k)$. It follows that

$$\begin{aligned} |a_i s_i| &= |S_i - S_{i-1}| \leq \frac{a_i a_k}{2} + \frac{a_{i-1} a_k}{2} \\ |s_i| &< \frac{a_k}{2} + \frac{a_{i-1}}{2}, \end{aligned}$$

as desired. Finally, we bound $|s_k|$. By definition of S_k, we have $S_k = d$, the GCD of $a_1, \ldots, a_k$. So $|s_k a_k| = |S_k - S_{k-1}| \leq d + \frac{a_{k-1}a_k}{2}$ or $|s_k| \leq \frac{d}{a_k} + \frac{a_{k-1}}{2}$. **Q.E.D.**

Note that for $k = 2$, this lemma gives a well-known bound

$$|s_1| \leq \frac{a_2}{2}, \qquad |s_2| \leq 1 + \frac{a_1}{2}.$$

Suppose we further assume that

$$a_1 > a_2 > \cdots > a_k.$$

Then we may further infer the bounds $|s_i| \leq a_{i-1} - (k-i+1)/2$ for $i = 2, \ldots, k$ and $|s_1| \leq a_k/2$. This yields:

Corollary 12 *For all $a_1 > a_2 > \cdots > a_k \geq 2$, there exists a co-factor $(s_1, \ldots, s_k)$ for $(a_1, \ldots, a_k)$ such that*

$$\prod_{i=1}^{k} |s_i| < \frac{a_k}{2} \prod_{i=1}^{k-1} \left(a_i - \frac{k-i+1}{2} \right) \leq \prod_{i=1}^{k} (a_i - 1).$$

This says that the output size of the co-GCD algorithm need not be larger than the input size.

We now address the question of computing a co-factor $(s_1, \ldots, s_k)$ satisfying the lemma. We will use a divide and conquer approach. We split $a_1, \ldots, a_k$ into two subsets according to the parity of their subscripts, and assume inductively that we have computed $c, c', t_1, \ldots, t_k$ such that

$$\begin{aligned} c &:= \texttt{GCD}(a_2, a_4, \ldots, a_{2\lfloor k/2 \rfloor}) = \sum_{i=1}^{\lfloor k/2 \rfloor} t_{2i} a_{2i}, \\ c' &:= \texttt{GCD}(a_1, a_3, \ldots, a_{2\lfloor (k-1)/2 \rfloor + 1}) = \sum_{i=1}^{\lfloor (k-1)/2 \rfloor} t_{2i+1} a_{2i+1}. \end{aligned}$$

By a call to the simple extended GCD on c, c', we obtain d, t, t' such that

$$d = tc + t'c'.$$

By induction, we may assume that $|t_i|$ is bounded according to lemma 11. In particular, $|t_i| < a_1$ for $i = 1, \ldots, k$. Similarly, $|t| < a_1$ and $|t'| < a_1$. Define

$$s'_1 = \begin{cases} t_i t & \text{if } i = \text{even}, \\ t_i t' & \text{if } i = \text{odd}. \end{cases}$$

Thus $\sum_{i=1}^{k} s'_i a_i = d$ and $|s'_i| < a_1^2$ for all i. Following the proof of lemma 11, we may now reduce $s'_1, \ldots, s'_k$ to $s_1, \ldots, s_k$:

Reduction Step:
1. $s_1 \leftarrow s'_1 \bmod a_k$ and $S_1 \leftarrow s_1 a_1$.
2. for $i = 2, \ldots, k-1$ do
   $S_i \leftarrow (S_{i-1} + a_i s'_i) \bmod a_i a_k$;
   $s_i \leftarrow (S_i - S_{i-1})/a_i$;
   $q_i \leftarrow (s'_i - s_i)/a_k$.
3. $s_k \leftarrow s'_k + \sum_{i=1}^{k-1} q_i a_i$.

The **mod** operator here is the symmetric version (§II.3). Note that both divisions in step 2 are exact.

Let

$$L = \lg \max\{a_1, \ldots, a_k\}. \tag{18}$$

The bit complexity of this reduction step is $O(k\mathrm{M}_B(L))$. Let $T(k, L)$ be the bit complexity of the overall recursive procedure. Clearly

$$T(k, L) = 2T(k/2, L) + O(k\mathrm{M}_B(L) + \mathrm{M}_B(L) \log L) \tag{19}$$

where '$k\mathrm{M}_B(L)$' comes from the reduction step and '$\mathrm{M}_B(L) \log L$' comes from the simple co-GCD computation for c, c' (Lecture II). The solution is easily seen to be $T(k, L) = O(k(\log k + \log L)\mathrm{M}_B(L))$. Thus we have:

Theorem 13 *There is a multiple co-GCD algorithm with bit complexity*

$$T(k, L) = O(k(\log k + \log L)\mathrm{M}_B(L)).$$

On input $(a_1, \ldots, a_k)$, the output co-factor $(s_1, \ldots, s_k)$ satisfy the bounds of lemma 11.

Application to multiple LCM. There are many applications of multiple GCD. An obvious application is for computing the primitive factorization (§III.1) of an integer polynomial. Bareiss [10] states that "there is no question that for maximum efficiency in any integer-preserving elimination code, the elements of all the rows and columns respectively should be made relative prime to each other before starting the elimination process". Here we consider another application: the computation of multiple LCM. Simple GCD and simple LCM are closely related: $\mathtt{LCM}(a, b) = ab/\mathtt{GCD}(a, b)$. It is slightly more involved to relate multiple GCD with multiple LCM. Our multiple GCD algorithm computes intermediate information that can be used to obtain the multiple LCM. Specifically, on

input $(a_1, \ldots, a_k)$ the information can be organized as a binary tree T of height $\lceil \lg k \rceil$ with k leaves, each associated with an a_i. At each internal node u of T, let $S_u \subseteq \{a_1, \ldots, a_k\}$ denote the a_i's associated to the leaves of the subtree rooted at u. Assume we store at u the value $\texttt{GCD}(S_u)$. It is then simple to extend the multiple GCD algorithm so that we recursively compute at node u the LCM of S_u. If v, w are the children of u, we use the formula

$$\texttt{LCM}(S_u) = \texttt{LCM}(\texttt{LCM}(S_v), \texttt{LCM}(S_w)) = \frac{\texttt{LCM}(S_v) \cdot \texttt{LCM}(S_w)}{\texttt{GCD}(S_u)}.$$

Let $L' = \lg \texttt{LCM}(a_1, \ldots, a_k)$. With L as in (18), we see that $L' \leq kL$. The cost of additional computation at u is $O(\mathrm{M}_B(L'))$. Overall, the additional cost is $O(k\mathrm{M}_B(L'))$. Combined with theorem 13, we conclude:

Theorem 14 *There is an algorithm to compute both the GCD and LCM of $a_1, \ldots, a_k$ with bit complexity $O(k(\log k \cdot \mathrm{M}_B(L) + \log L \cdot \mathrm{M}_B(L) + \mathrm{M}_B(L')))$ which is*

$$O(k(\log L \cdot \mathrm{M}_B(L) + k\mathrm{M}_B(L))).$$

Remarks. Iliopoulos [89] obtains the co-factor bound $|s_i| = O(a_i^k)$ by using a balanced binary tree and co-factor bounds for simple GCD's. Lüneburg [121] gives the bound

$$|s_i| \leq a_1/2$$

for all i except when $i = i_0$ (for some i_0). Moreover, $|s_{i_0}| \leq (1 + a_1(k-1))/2$. The dependence on k in these bounds is somewhat unsatisfactory. Hongwei Cheng[1] shows the uniform bound $|s_i| < a_1$ for all i; our lemma 11 follows his argument.

Exercises

Exercise 5.1: Suppose we apply lemma 11 with

$$a_{k-1} > a_{k-2} > \cdots > a_2 > a_1 > a_k.$$

How does this compare to the bound in the corollary? □

Exercise 5.2: Verify the solution to recurrence (19) by an appropriate induction. □

Exercise 5.3: Suppose $a_1, \ldots, a_k \in \mathbb{Z}[X]$. Give an efficient method for computing the extended GCD of $a_1, \ldots, a_k$. □

Exercise 5.4: (Open: Odlyzko, Sims) Is there a function $f(k) > 0$ that goes to infinity as $k \to \infty$ such that for all $a_1 > a_2 > \cdots > a_k \geq 2$, a cofactor $(s_1, \ldots, s_k)$ exists where $|s_i| \leq |a_1|/f(k)$? [Can we take $f(k) = \lg k$?] □

§6. Hermite Reduction Step

[1] Private communication, 1991.

Let $a \in \mathbb{Z}^{1\times n}$ be a non-zero row n-vector, $n \geq 2$. The goal in this section is to construct a unimodular matrix $U \in \mathbb{Z}^{n\times n}$ such that $a \cdot U$ is zero except for its first entry. Since U is invertible, the non-zero entry of $a \cdot U$ must be equal to $\texttt{GCD}(a)$. We will call the transformation

$$a \mapsto U \tag{20}$$

a "Hermite reduction step problem". The reason for this name is clear. For, if a is the first row in a $m \times n$ matrix, then $A \cdot U$ is essentially the first step of the generic HNF algorithm. Repeating this process suitably, we finally obtain a column echelon form of A, which is easily converted into the Hermite normal form.

First let us illustrate our approach for the case $n = 2$. Suppose $a = (a_1, a_2)$ and we wish to find a 2×2 unimodular U such that $aU = (g, 0)$ where $g = \texttt{GCD}(a_1, a_2)$. So there exist integers s, t such that $g = sa_1 + ta_2$. Moreover, s, t are relatively prime. Hence there exist integers s', t' such that $ss' + tt' = 1$. In fact, we may choose $s' = a_1/g$ and $t' = a_2/g$. Then it is easy to see that

$$(a_1, a_2)\begin{bmatrix} s & -t' \\ t & s' \end{bmatrix} = (g, g') \tag{21}$$

for some g'. But g divides g' (since g divides a_1 and a_2). If $e = g'/g$ then we see that the desired U can be taken to be $P \cdot P^*$ where

$$P = \begin{bmatrix} s & -t' \\ t & s' \end{bmatrix}, \qquad P^* = \begin{bmatrix} 1 & -e \\ 0 & 1 \end{bmatrix}.$$

It turns out that U can always be written as the product of matrices P and P^*. In fact, we will see that it is more useful to represent U as the pair (P, P^*). We begin with a key lemma which shows how to construct P (see [86, p. 375] or [145]).

Lemma 15 *Let $u = (u_1, \ldots, u_n)^T$ be a column of integers with $d = \texttt{GCD}(u)$. There exists an $n \times n$ integer matrix P with the following properties:*

(i) *The first column of P equals u.*

(ii) $\det P = d$.

(iii) $\|P\|_\infty < \|u\|_\infty$.

Proof. We use induction on n. The result is trivial for $n = 1$ so let $n \geq 2$. Let $u' = (u_1, \ldots, u_{n-1})^T$ with $d' = \texttt{GCD}(u')$. By induction, there is a matrix P' with first column u' and $\det P' = d'$. So there are integers s, t such that

$$sd' + tu_n = d. \tag{22}$$

We may assume that $|s| < |u_n|$ and $|t| < |d'|$. We claim (§5) that the desired matrix can be taken to be

$$P = \left[\begin{array}{ccccc} & & P' & & \frac{u_1}{r} \\ & & & & \frac{u_2}{r} \\ & & & & \vdots \\ & & & & \frac{u_{n-1}}{r} \\ \hline u_n & 0 & \cdots & 0 & s \end{array}\right],$$

where $r \in \mathbb{Q}$ is to be determined. Part (i) is clearly satisfied. By expanding the determinant along the last row, we have

$$\det P = s \det P' + u_n \det P'',$$

where P'' is obtained by omitting the first column and last row of P. We want to choose r so that $\det P'' = t$ (so that (ii) is satisfied). We observe that P'' is rather similar to P': the last column of P'' is just $1/r$ times the first column of P'. In fact,

$$P'' = P' \cdot C = P' \cdot \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \frac{1}{r} \\ 1 & 0 & 0 & & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

Here C has a subdiagonal of 1's and a top-right entry of $1/r$. Note that C simply circulates the columns of P', moving the first column (multiplied by $1/r$) to the last place and for the other columns, one place to the left. When $n = 2$, C is simply the matrix $[\frac{1}{r}]$. Then $\det P'' = \det C \det P' = (-1)^n \frac{1}{r} d'$. Thus part (ii) is satisfied with $r = (-1)^n d'/t$. To see that P is correct, note that the entries of P are integers since $u_i/r = u_i t/d' \in \mathbb{Z}$ for $i = 1, \ldots, n-1$.

Finally, part (iii) claims that each entry of P is bounded by $\|u\|_\infty$. This is true of the entries of P' (by induction) and also true of the non-zero entries in the rest of P, namely, s, u_n and u_i/r (in the last case, we use the fact that $|t| < |d'|$). **Q.E.D.**

Note that P is not unique since s, t are not unique. In our application, the vector $u = (u_1, \ldots, u_n)^T$ satisfies $\texttt{GCD}(u) = 1$ so that the matrix P is unimodular.

Now we implement the above lemma.

Lemma 16 *Let $a = (a_1, \ldots, a_n) \neq 0$ be an integer row vector and $\log \|a\|_\infty \leq L$.*

(i) *We can construct the unimodular matrix P of the previous lemma 15 in bit complexity*

$$O(n(n + \log L)\mathrm{M}_B(L)). \tag{23}$$

(ii) *Suppose $b = (b_1, \ldots, b_n) \in \mathbb{Z}^n$ satisfies $\log \|b\|_\infty \leq L$ then we can compute $b \cdot P$ from b, P in bit complexity*

$$O(n\mathrm{M}_B(L + \log n)). \tag{24}$$

Proof. (i) Let $u = (u_1, \ldots, u_n)^T$ such that $\sum_{i=1}^n u_i a_i = \texttt{GCD}(a_1, \ldots, a_n) = d$. By theorem 13, we can compute u in time

$$O(n(\log n + \log L)\mathrm{M}_B(L)). \tag{25}$$

Note that $\texttt{GCD}(u_1, \ldots, u_n) = 1$. As in the proof of lemma 15, we recursively compute the $(n-1) \times (n-1)$ matrix P' with $\det P' = d'$ with first column $(u_1, \ldots, u_{n-1})^T$. Then we compute s, t with $sd' + tu_n = d$, in time $\mathrm{M}_B(L) \log L$. Finally, we compute the last row of P in time $O(n\mathrm{M}_B(L))$. Hence if $T(n, L)$ is the complexity of computing P

$$T(n, L) = T(n-1, L) + O((n + \log L)\mathrm{M}_B(L)).$$

Hence $T(n, L) = O(n(n + \log L)\mathrm{M}_B(L))$, and this dominates the complexity in (25).

(ii) Given a and P, a straightforward multiplication gives bP in time $O(n^2\mathrm{M}_B(L))$. But a better bound is needed. For $i = 2, \ldots, n$, the ith column of P has the form

$$p_i = (\frac{u_1}{r_i}, \frac{u_2}{r_i}, \ldots, \frac{u_{i-1}}{r_i}, s_i, 0, \ldots, 0)^T$$

where r_i, s_i are the elements described in the proof of lemma 15. So

$$bP = (\langle b, p_1\rangle, \langle b, p_2\rangle, \ldots, \langle b, p_n\rangle)$$

where $\langle b, p_i\rangle$ indicates scalar product. We will compute the entries of bP in a particular order. Notice that

$$\log|\langle b, p_i\rangle| = O(L + \log n). \tag{26}$$

Clearly $\langle b, p_2\rangle$ can be computed in time $O(\mathrm{M}_B(L))$. For $i = 2, \ldots, n-1$, $\langle b, p_{i+1}\rangle$ can be obtained from $\langle b, p_i\rangle$ in time $O(\mathrm{M}_B(L + \log n))$ using the formula

$$\langle b, p_{i+1}\rangle = \frac{(\langle b, p_i\rangle - b_i s_i) r_i + b_i u_i}{r_{i+1}} + b_{i+1} s_{i+1}.$$

Finally, the first entry $\langle b, p_1\rangle$ in bP can be computed from the last entry in time $O(\mathrm{M}_B(L + \log n))$. The entire computation costs $O(n\mathrm{M}_B(L + \log n))$ as claimed. **Q.E.D.**

Continuing in our pursuit of the matrix U, we now need to compute the matrix P^* such that $(aP)P^*$ will have zero everywhere except the first entry. Clearly, if $aP = (g_1, g_2, g_3, \ldots, g_n)$ then $g_1 = d$ and

$$P^* = \begin{bmatrix} 1 & -g_2/d & -g_3/d & \cdots & -g_{n-1}/d & -g_n/d \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Note that P^* is integer unimodular since d divides each g_i. The entries in the first row of P^* have bit-size of $O(L + \log n)$. We return to the main problem (20):

Lemma 17 (Hermite Reduction Step) *Let $a \neq 0$ be the first row of $A \in \mathbb{Z}^{m\times n}$ and $L = \lg\|A\|_\infty$. Define the matrices P, P^*, U relative to a, as before.*

(i) *We can compute the matrices P, P^*, U from a in time $O(n(n + \log L)\mathrm{M}_B(L + \log n))$.*

(ii) *If b is any row of A, we can compute bU from b, P, P^* in time $O(n\mathrm{M}_B(L + \log n))$.*

(iii) *We can compute AU from A, P, P^* in time $O(mn\mathrm{M}_B(L + \log n))$.*

(iv) *We can compute P, P^*, AU from A in time $O(n(n + m + \log L)\mathrm{M}_B(L + \log m))$.*

Proof. (i) We use the previous lemma to compute P in time (23). Similarly, we can compute $aP = (g_1, \ldots, g_n)$ in time (24). Assuming that we only represent the first row of P^* explicitly, we can also construct P^* in time (24). Next, each entry of $U = PP^*$ can be computed in time $O(\mathrm{M}_B(L + \log n))$ or

$$O(n^2\mathrm{M}_B(L + \log n))$$

for the entire matrix U. Thus the three matrices P, P^*, U can be computed in time

$$O(n(n + \log L)\mathrm{M}_B(L)) + n\mathrm{M}_B(L + \log n) + O(n^2\mathrm{M}_B(L + \log n)) = O(n(n + \log L)\mathrm{M}_B(L + \log n)).$$

(ii) To compute bU, we first compute bP in time (24). Then compute $(bP)P^*$ within the same time bound (24).
(iii) This amounts to repeating part (ii) m times.
(iv) This just adds up parts (i) and (iii). **Q.E.D.**

Application. Although we could use this lemma repeatedly to compute the Hermite normal form, it seems that the best bounds for bit sizes of intermediate values are exponential in $\min\{m, n\}$. So this application is only useful for $m \gg n$ or $n \gg m$. We describe another application here. Suppose the columns of $A = [a_1, \ldots, a_n] \in \mathbb{Z}^{m\times n}$ are not linearly independent. Consider the problem of computing $B = [b_1, \ldots, b_{n-1}] \in \mathbb{Z}^{m\times(n-1)}$ such that $\Lambda(A) = \Lambda(B)$. For instance, this is useful in preprocessing a basis before calling the LLL algorithm, since the LLL algorithm requires the input columns to be linearly independent.

Note that there exists $x \in \mathbb{Z}^{n\times 1}$ such that $Ax = 0$. Finding such an x from A is easily reduced to Gram-Schmidt orthogonalization. In the notations of §IX.1, if $A^* = [a_1^*, \ldots, a_n^*]$ is the Gram-Schmidt version of A, then $a_i^* = 0$ for some i. This means

$$a_i^* = 0 = a_i - \sum_{j=1}^{i-1} \mu_{ij} a_j^*.$$

Recall that μ_{ij} are rational numbers. It is then easy to find integers $t_1, \ldots, t_i$ such that

$$0 = t_i a_i + \sum_{j=1}^{i-1} t_j a_j. \tag{27}$$

We may set $x = (t_1, \ldots, t_i, 0, \ldots, 0)^T$. Clearly, we may assume that $\texttt{GCD}(x) = 1$. To find B, we can use the Hermite reduction step to give us a unimodular matrix $U \in \mathbb{Z}^{n\times n}$ such that $Ux = (1, 0, \ldots, 0)^T$. Since $(AU^{-1})(Ux) = \mathbf{0}$, we conclude that the first column of AU^{-1} is zero. The desired matrix B may be taken to comprise all but the first column of AU^{-1}.

Algebraic complexity of the Hermite Reduction Step.

We have given an upper bound on the bit complexity of the Hermite Reduction Step (20). But suppose we want its complexity in an algebraic model of computation (§0.6). It is clear from the preceding that problem (20) over $\mathbb{Z}$ can be solved in polynomial time in an algebraic model M if the elementary operations or *basis* of M includes the following:

$$+, \quad -, \quad , \times, \quad \texttt{cGCD}(x, y). \tag{28}$$

Here cGCD denotes the co-GCD primitive (§II.2) that returns a co-factor (s, t) of an input pair $(x, y) \in \mathbb{Z}^2$: $sx + ty = \texttt{GCD}(x, y)$. Hafner and McCurley [77, p. 1075] suggested that there may be no solution in case our basis comprises only the ring operations $(+, -, \times)$, *i.e.*, if we omit cGCD. Let us show this. If (20) can be solved in the algebraic model M then $\texttt{GCD}(x, y)$ can be solved in M in constant time by reduction to the $n = 2$ case. This assumes (as we may) that the ring operations are available in M. Suppose $\pi(x, y)$ is a branching program in M for computing the $\texttt{GCD}(x, y)$. The inputs x, y and constants used in π are all integers, and at each decision node, there is an integer polynomial $f(X, Y)$ whose sign at the input (x, y) determines the flow of computation. The finiteness of π means that there are finitely many leaf nodes in the branching program. At each leaf ℓ, there is an integer polynomial $P_\ell(X, Y)$ such that if input (x, y) terminates at ℓ the $P_\ell(x, y) = \texttt{GCD}(x, y)$. Let S_ℓ denote the set of all $(x, y) \in \mathbb{R}^2$ that terminate at ℓ. Note that it makes sense to feed pairs (x, y) of real numbers to π and hence the set S_ℓ is well-defined. Clearly, S_ℓ is a semi-algebraic set (*i.e.*, defined by a finite set of polynomial inequalities). By basic properties of semi-algebraic sets, there is some leaf ℓ such that S_ℓ has the following properties: S_ℓ contains an infinite cone C which in turn contains infinitely many vertical rays of the form $R_i = \{(x_i, y) : y \geq c_i\}$ where $x_i \in \mathbb{Z}$ is prime and c_i is arbitrary, for $i = 0, 1, 2, \ldots$. Focus on the output polynomial $P_\ell(X, Y)$ at such an ℓ. We may pick a ray R_i such that such that none of the non-zero coefficients of the Y's in $P_\ell(X, Y)$ vanish when X is replaced by x_i. Since there are infinitely many prime y's such that $(x_i, y) \in R_i \subseteq S_\ell$, we conclude that $P_\ell(x_i, Y)$ is the constant 1 and $P_\ell(X, Y)$ does not depend on Y. Next suppose

P_ℓ has X-degree $< d$. Pick a disc $D \subseteq S_\ell$ large enough so that there are prime numbers y_0 and u_i $(i = 1, \ldots, d)$ such that D contains (u_i, y_0) for $i = 1, \ldots, d$ and none of the non-zero coefficients of X in $P_\ell(X, Y)$ vanishes when y_0 replaces Y. Again, we argue that $P_\ell(X, y_0)$ is the constant 1 and $P_\ell(X, Y)$ does not depend on X. Thus $P_\ell(X, Y)$ must be the constant 1. But, clearly S_ℓ contains a pair (a, b) of integers such that $\texttt{GCD}(a, b) > 1$. This is a contradiction since $1 = P_\ell(a, b) = \texttt{GCD}(a, b) > 1$.

Exercises

Exercise 6.1: (i) Show that the matrix P in lemma 15 is a product of elementary integer unimodular matrices (§VIII.1). HINT: use induction on n.
(ii) Do the same for P^*. Hence conclude that U in (20) is a product of elementary integer unimodular matrices. □

Exercise 6.2: (i) What is the algebraic complexity of the Hermite reduction step assuming the basis (28)?
(ii) Show that if 2×2 Smith normal form (see §8) can be solved in constant time in an algebraic model M then the general Smith normal form problem can be solved in polynomial time in M.
(iii) Assume an algebraic computation model M whose basis is given by (28). Show that the 2×2 Smith normal form problem is equivalent to the following: given $a, b, c \in \mathbb{Z}$ find $U, V, W, Z \in \mathbb{Z}$ such that $aUV + bVW + cWZ = \texttt{GCD}(a, b, c)$. HINT: first reduce the 2×2 matrix to Hermite normal form.
(iv) (Open) Prove that the problem in (iii) cannot be solved in constant time in model M. □

Exercise 6.3: In the above application, work out efficient algorithms for:
(i) Computing the integers $t_1, \ldots, t_n$ in (27) from μ_{ij}'s.
(ii) Computing the inverse U^{-1} from U. □

Exercise 6.4: Suppose $a = (a_1, \ldots, a_n)^T$ and $b = (b_1, \ldots, b_n)^T$ $(n \geq 2)$ be two columns. Under what circumstances is there a unimodular matrix U whose first and second columns are a and b? A necessary condition is that $\texttt{GCD}(a) = \texttt{GCD}(b) = 1$ and the GCD of all minors of order 2 of $[a, b]$ is 1. □

Exercise 6.5: (Bass) Let $M = [a_{ij}]$ be an $n \times n$ matrix over a ring R. Say M is invertible iff its determinant is invertible in R. But $\det M = a_{11}A_{11} + ... + a_{1n}A_{1n}$ where A_{ij} is the co-factor of a_{ij}. Write a_i for a_{1i}. Then invertibility of M implies $R = \texttt{Ideal}(a_1, \ldots, a_n)$. We also call $(a_1, \ldots, a_n)$ a unimodular row in this case. The converse asks: is every unimodular row the row of an invertible matrix? This is related to a conjecture of Serre's which was answered affirmatively by Quillen and Suslin. Verify the counterexample: $R = \mathbb{R}[X, Y, Z]/(X^2 + Y^2 + Z^2 = 1)$. The unimodular row $(\overline{X}, \overline{Y}, \overline{Z})$ is not the first row of any invertible matrix. Here $\overline{u}$ denotes the image of u in the canonical map from $\mathbb{R}[X, Y, Z]$ to R. □

§7. Bachem-Kannan Algorithm

We present a polynomial time algorithm for Hermite normal form. It is essentially that of Bachem and Kannan [7], extended to non-square matrices of arbitrary rank. *We assume that the input matrix*

$A \in \mathbb{Z}^{m\times n}$ has been preprocessed so that the ith principal minor is non-zero for $i = 1, \ldots, \rho$ where $\rho > 0$ is the rank of A. The Extended Bareiss Algorithm (§2) can convert any matrix A into this form. Moreover, the complexity of this conversion is dominated by the subsequent computation.

Algorithm. The algorithm is relatively straightforward to describe. For any matrix M, let

$$\langle M\rangle_i := M(1, 2, \ldots, i; 1, 2, \ldots, i)$$

using the general matrix notations in §0.9. So $\langle M\rangle_i$ is just the ith principal submatrix of M. On input A, the algorithm has n stages where in the sth ($s = 1, \ldots, n$) stage, we seek to put $\langle A\rangle_s$ into the Hermite normal form. Stage 1 requires no action. Suppose that we have just completed the $(s-1)$st stage. The sth stage has two *phases*. *Elimination phase:* we eliminate (*i.e.*, zero out) the first $s-1$ entries in the sth column. This will make $\langle A\rangle_s$ a lower triangular matrix. *Reduction phase:* we then reduce the off-diagonal entries of $\langle A\rangle_s$ so that each such entry is non-negative and less than the corresponding diagonal entry in its row. This completes the sth stage.

Bachem-Kannan Algorithm
Input: $A \in \mathbb{Z}^{m\times n}$ and $\rho \geq 1$ the rank of A.
Assume the ith principal minor is non-zero for $i = 1, \ldots, \rho$.
Output: $H \in \mathbb{Z}^{m\times n}$, the HNF of A, and $U \in \mathbb{Z}^{n\times n}$, a unimodular matrix such that $H = AU$.
INITIALIZATION:
1. $H := A$; $U := I$, the identity matrix.
MAIN LOOP:
2. for $s \leftarrow 2$ to n do
ELIMINATION PHASE:
3. for $i \leftarrow 1$ to $\min\{s-1, \rho\}$ do
4. By postmultiplying H with a suitable unimodular matrix K, eliminate the (i, s)th entry of H; Update $U \leftarrow UK$.
REDUCTION PHASE:
5. if $s > \rho$, skip this phase; else continue.
6. for $j \leftarrow s-1$ downto 1 do
7. for $i \leftarrow j+1$ to s do
8. By postmultiplying H with a suitable unimodular matrix K, reduce the (i, j)th entry of H; Update $U \leftarrow UK$.

Call step 4 an "elimination step" and step 8 a "reduction step". Note that when $s > \rho$, the reduction phase is omitted since column s would be zero after the elimination phase. The order of reduction represented by the double for-loop (steps 6 and 7) is important for the analysis: this is an improvement from Chou and Collins [42]. The reduction step is rather obvious: the entry x to be reduced is replaced by $x \bmod d$ where d is the diagonal element in the same row. The rest of the column of x is modified accordingly. We now illustrate the elimination step: it is basically the 2×2 version of the Hermite reduction step (§6). For instance, suppose $H \in \mathbb{Z}^{5\times 6}$ already has its 3rd principal submatrix in HNF. The following shows the first elimination step of stage 4:

$$H = [h_1, h_2, h_3, h_4, h_5, h_6] \quad \rightarrow \quad [h'_1, h_2, h_3, h'_4, h_5, h_6] = H',$$

$$\begin{bmatrix} a_{1,1} & 0 & 0 & a_{1,4} & * & * \\ a_{2,1} & a_{2,2} & 0 & a_{2,4} & * & * \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & * & * \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & * & * \\ * & * & * & * & * & * \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} a'_{1,1} & 0 & 0 & 0 & * & * \\ a'_{2,1} & a_{2,2} & 0 & a'_{2,4} & * & * \\ a'_{3,1} & a_{3,2} & a_{3,3} & a'_{2,4} & * & * \\ a'_{4,1} & a_{4,2} & a_{4,3} & a'_{2,4} & * & * \\ * & * & * & * & * & * \end{bmatrix}.$$

As in (21) (§6), this amounts to replacing columns h_1, h_4 (respectively) by h'_1, h'_4 which is defined as follows:

$$[h'_1, h'_4] \leftarrow [h_1, h_4] \begin{bmatrix} s & a_{1,4}/g \\ t & -a_{1,1}/g \end{bmatrix}$$

where $g = \texttt{GCD}(a_{1,1}, a_{1,4}) = s \cdot a_{1,1} + t \cdot a_{1,4}$. We may assume that (see §5)

$$|s| < |a_{1,4}|, \qquad |t| \leq |a_{1,1}|. \tag{29}$$

We may write this elimination step as $H' = HK$ where

$$K = \begin{bmatrix} s & 0 & 0 & a_{1,4}/g & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ t & 0 & 0 & -a_{1,1}/g & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{30}$$

If U is the unimodular matrix such that $H = AU$ then we update U via $U \leftarrow UK$.

Analysis. There are subtleties in proving a polynomial running time. In this analysis, we simply write $\|M\|$ instead of $\|M\|_\infty$ for the ∞-norm of a matrix M. With

$$\lambda_0 := \|A\|, \quad L := \lg \lambda_0,$$

the analysis amounts to showing that throughout the algorithm, $\lg \|H\|$ and $\lg \|U\|$ are polynomially-bounded in terms of m, n and L.

Bound between stages. Let $H^{(s)}$ denote the H-matrix just after the sth stage. Thus $H^{(1)}$ is equal to the initial matrix A. Also let the unimodular matrix that transforms A to $H^{(s)}$ be denoted $U^{(s)}$:

$$H^{(s)} = AU^{(s)}.$$

So $U^{(1)}$ is the identity matrix. Letting

$$\lambda_1 := (\rho\lambda_0)^\rho, \tag{31}$$

we see that every minor of A is bounded by λ_1. From this we infer

$$\|\langle H^{(s)} \rangle_s\| \leq \lambda_1. \tag{32}$$

This is because each step (elimination or reduction) in the first s stages does not change the sth principal minor of H, and H is initially equal to A. Since $\langle H^{(s)} \rangle_s$ is lower triangular, this means that each diagonal entry of $\langle H^{(s)} \rangle_s$ is bounded by λ_1. But the off-diagonal entries are also bounded by λ_1, since $\langle H^{(s)} \rangle_s$ is in HNF. Thus (32) is verified.

First assume $s \leq \rho$. Note that $U^{(s)}$ has the form

$$U^{(s)} = \begin{bmatrix} \langle U^{(s)} \rangle_s & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}. \tag{33}$$

So

$$\langle U^{(s)} \rangle_s = \langle A \rangle_s^{-1} \langle H^{(s)} \rangle_s. \tag{34}$$

Each entry of $\langle A \rangle_s^{-1}$ is bounded by $\lambda_1 / \det(\langle A \rangle_s)$ (see §1). So a typical element $U_{ij}^{(s)}$ of $\langle U^{(s)} \rangle_s$ is bounded by

$$|U_{ij}^{(s)}| \leq \sum_{k=1}^{s} \frac{\lambda_1}{|\det \langle A \rangle_s|} |H_{kj}^{(s)}| \leq \lambda_1,$$

since

$$\begin{aligned}\sum_{k=1}^{s} |H_{kj}^{(s)}| &\leq 1 + \sum_{k=j}^{s} (|H_{kk}^{(s)}| - 1) &(35)\\ &\leq \prod_{k=1}^{s} |H_{kk}^{(s)}| &(36)\\ &= |\det\langle A\rangle_s|.\end{aligned}$$

Inequality (35) is a consequence of Hermite normal form of $\langle H^{(s)}\rangle_s$. Inequality (36) exploits the elementary bound

$$1 + \sum_{k=1}^{\ell} (x_k - 1) \leq \prod_{k=1}^{\ell} x_k \tag{37}$$

which holds for any $\ell \geq 1$ positive integers $x_1, \ldots, x_\ell$ (see Exercise). From (33), we infer that

$$\|U^{(s)}\| \leq \|\langle U^{(s)}\rangle_s\| \leq \lambda_1. \tag{38}$$

From $H^{(s)} = AU^{(s)}$ and the special form (33), we conclude

$$\|H^{(s)}\| \leq \lambda_2 \tag{39}$$

where

$$\lambda_2 := \rho\lambda_0\lambda_1 = (\rho\lambda_0)^{1+\rho}. \tag{40}$$

We have therefore given a bound on $\|U^{(s)}\|$ and $\|H^{(s)}\|$ for $s \leq \rho$.

Now let $s > \rho$. Clearly $H^{(s)}$ still satisfies the bound (39) since in stage s we eliminated column s while the remaining columns are unchanged from $H^{(s-1)}$. We conclude that $\|H\| \leq \lambda_2$ holds in *the transition* between any two successive stages.

What about $U^{(s)}$? Eliminating the (r, s)th entry amounts to adding some multiple $c_{r,s}$ of column r ($r \leq \rho$) to column s. The multiple $c_{r,s}$ is bounded by λ_0. Therefore, it increases $\|U^{(s)}\|$ by a factor of $(\lambda_0 + 1)$. We perform ρ such elimination steps to entirely eliminate column s. Thus

$$\|U^{(s)}\| \leq (\lambda_0 + 1)^\rho \|U^{(s-1)}\|.$$

The maximum size bound is when $s = n$:

$$\|U^{(n)}\| \leq (\lambda_0 + 1)^{\rho(n-\rho)} \|U^{(\rho)}\| \leq \lambda_3, \tag{41}$$

where

$$\lambda_3 := (\lambda_0 + 1)^{\rho(n-\rho)} \lambda_1. \tag{42}$$

Bounds on H during a stage. What remains is to bound $\|H\|$ and $\|U\|$ during a stage. In other words, although the entries in H and U are nicely bounded between two successive stages, we must ensure that they do not swell up excessively within a stage. In our analysis, we shall use the observation that once the (s, s)th element is "fixed" at the end of the sth stage, it is hereafter bounded by λ_1. If it changes at all, it can only become smaller (replaced by a divisor). In fact, the product of all the "fixed" diagonal elements is bounded by λ_1. Let us now focus on stage s for some $s = 1, \ldots, n$. Of course, the columns of interest are the first $\min\{s - 1, \rho\}$ columns and column s.

ELIMINATION PHASE: let $H^{(r,s)}$ be the H matrix just before the (r, s)th entry is eliminated ($r = 1, \ldots, \min\{s - 1, \rho\}$). Let $h_j^{(r,s)}$ be the jth column of $H^{(r,s)}$ and $H_{i,j}^{(r,s)}$ be the (i, j)th entry of $H^{(r,s)}$. Initially, we have

$$\|h_j^{(1,s)}\| \leq \begin{cases} \lambda_2, & \text{for } j = 1, \ldots, \min\{s - 1, \rho\}, \\ \lambda_0, & \text{for } j = s. \end{cases}$$

When we eliminate the (r,s)th entry, column s is transformed (cf. equation (30)) according to the rule

$$h_s^{(r+1,s)} \leftarrow \frac{H_{r,s}^{(r,s)} h_r^{(r,s)} - H_{r,r}^{(r,s)} h_s^{(r,s)}}{g} \tag{43}$$

where $g = \texttt{GCD}(H_{r,r}^{(r,s)}, H_{r,s}^{(r,s)})$. At the same time, column r is transformed according to the rule

$$h_r^{(r+1,s)} \leftarrow c \cdot h_r^{(r,s)} - c' \cdot h_s^{(r,s)} \tag{44}$$

where $|c| < |H_{r,s}^{(r,s)}|$ and $|c'| \leq |H_{r,r}^{(r,s)}|$ (cf. equation (29)). Define

$$\beta_r := \lambda_2(\lambda_1 + \lambda_2)^{r-1}.$$

Inductively assume that column s is bounded as follows:

$$\|h_s^{(r,s)}\| \leq \beta_r. \tag{45}$$

This is true for $r = 1$. From (43), we extend the inductive hypothesis to $r+1$:

$$\begin{aligned} \|h_s^{(r+1,s)}\| &\leq |H_{r,s}^{(r,s)}| \cdot \|h_r^{(r,s)}\| + |H_{r,r}^{(r,s)}| \cdot \|h_s^{(r,s)}\| \\ &\leq \beta_r \cdot \lambda_2 + \lambda_1 \cdot \beta_r \\ &= \beta_r(\lambda_1 + \lambda_2) = \beta_{r+1}. \end{aligned}$$

From (44), we similarly obtain a bound on column s:

$$\|h_r^{(r,s)}\| \leq \beta_r. \tag{46}$$

Let $H^{(s,s)}$ be the H matrix just *after* the $(s-1,s)$th entry of H is eliminated. Then the bounds (45) and (46) extend to

$$\|h_{s-1}^{(s,s)}\| \leq \beta_s, \qquad \|h_s^{(s,s)}\| \leq \beta_s.$$

We conclude that throughout an elimination phase, each entry of H is bounded by

$$\beta_s = \lambda_2(\lambda_1 + \lambda_2)^{s-1} < \lambda_4$$

where

$$\lambda_4 := (2\lambda_2)^\rho = 2^\rho(\rho\lambda_0)^{\rho(1+\rho)}. \tag{47}$$

REDUCTION PHASE: So $H^{(s,s)}$ is the H matrix just before the start of the reduction phase. Note that we may assume $s \leq \rho$. Let $h_j^{(s)}$ be the jth column of $H^{(s,s)}$. Also let $\widehat{h}_j^{(s)}$ be the jth column at the *end* of the reduction phase: these are called the "reduced vectors". Note that the reduced vectors are columns of $H^{(s)}$ and hence satisfy

$$\|\widehat{h}_j^{(s)}\| \leq \lambda_2. \tag{48}$$

Exploiting the special sequencing of reduction steps in the algorithm (following Chou-Collins), we see that

$$\widehat{h}_j^{(s)} = h_j^{(s)} - \sum_{r=j+1}^{s} b_{r,j,s} \widehat{h}_r^{(s)} \tag{49}$$

for suitable constants $b_{r,j,s}$. The point is that reduced vectors appear on the right-hand side of (49). The entries of column j in H are bounded by λ_4 at the start of the reduction phase. To reduce column j ($j = 1, \ldots, s-1$), we first reduce its $j+1$st entry by subtracting the column $b_{j+1,j,s}\widehat{h}_{j+1}^{(s)}$ (cf. equation (49)). Clearly $b_{j+1,j,s} \leq \lambda_4$ so that entries of column j are bounded by $\lambda_4(1+\lambda_2)$ after this reduction step. Inductively, it is easy to see that after the $(j+k)$th ($k = 1, 2, \ldots$) entry of column j is reduced, the entries of column j are bounded by

$$\lambda_4(1+\lambda_2)^k.$$

So just after the $s-1$st entry is reduced, its entries are bounded by

$$\lambda_4(1+\lambda_2)^{s-1-j} < \lambda_4(2\lambda_2)^{\rho} = \lambda_4^2.$$

Finally, we reduce the sth entry of column j. But the result is a reduced column bounded as in (48). It follows that the bound

$$\|H\| \leq \lambda_4^2$$

holds throughout the stage.

Bounds on U during a stage. First assume $s \leq \rho$. It suffices to use the bound

$$\|U\| \leq n\|A^{-1}\| \cdot \|H\| \leq n\lambda_1\lambda_4^2$$

since the relation $U = A^{-1}H$ holds throughout the stage. Suppose $s > \rho$. There is no reduction phase and the argument showing $\|U^{(s)}\| \leq \lambda_3$ in (41) actually shows that $\|U\| \leq \lambda_3$ *throughout* stage s. This concludes our analysis.

We summarize the foregoing analysis: the entries of H and U matrices are uniformly bounded by

$$\lambda_5 := \lambda_3 + \lambda_4 < \lambda_1((2\lambda_0)^{\rho(n-\rho)} + n\lambda_4^2)$$

throughout the algorithm. Since $L = \lg \lambda_0$, we get

$$L' := \lg \lambda_5 = O(\rho n L + \rho^2 \lg \rho)$$

as a bound on the bit-size of entries. There are $O(\rho^2 n)$ column operations and $O(\rho^2)$ co-GCD operations on matrix entries. Each column operation takes $O(m)$ ring operations on matrix entries. Hence the cost of co-GCD operations is dominated by the cost of column operations, which amounts to:

Theorem 18 *With $L = \lg \lambda_0$, the matrix entries have bit-size that are uniformly bounded by*

$$L' = O(\rho[nL + \rho \lg \rho)])$$

in the Bachem-Kannan algorithm. The bit-complexity of the algorithm is $O(mn\rho^2 \mathrm{M}_B(L'))$.

Remarks. Our complexity analysis is somewhat more involved than that in Bachem and Kannan [7], in part because the input matrix is non-square and may have dependent rows and columns. For instance, if $n = \rho$ then $\lg(\lambda_3) = O(n(L + \lg n)$ and not $O(n^2 L)$.

Suppose we want to compute the HNF $H(A)$ of an arbitrary matrix A. Accordingly, we submit A to the Extended Bareiss algorithm (§2) to obtain $B = PAQ$ where P and Q are permutation matrices. Now we submit B to the Bachem-Kannan algorithm which outputs $H = H(B)$, the HNF of B. We leave it as an exercise to show:

$$H(A) = P^{-1}H. \tag{50}$$

Exercises

Exercise 7.1: (Chou-Collins) Verify inequality (37) □

Exercise 7.2: Show equation (50). HINT: this depends on the particular structure of our Extended Bareiss algorithm. □

Exercise 7.3: Analyze the algorithm assuming A is square and non-singular. □

§8. Smith Normal Form

H. J. S. Smith (1861) introduced the normal form bearing his name. Let $A \in \mathbb{Z}^{m \times n}$. We say A is in *Smith normal form* (SNF) if it is diagonal, that is, $(A)_{i,j} = 0$ for $i \neq j$, and its diagonal entries are non-negative satisfying

$$(A)_{i-1,i-1} | (A)_{i,i}, \qquad \text{for } \; i = 2, \ldots, \min\{m, n\}. \tag{51}$$

Since every number divides 0, but 0 divides only itself, we conclude from (51) that if $(A)_{i,i} = 0$ for some i, then $(A)_{j,j} = 0$ for all $j > i$. The multi-set of non-zero diagonal entries of a Smith normal form matrix A,

$$\{(A)_{1,1}, (A)_{2,2}, \ldots, (A)_{r,r}\}$$

where $(A)_{r,r} > 0$ and $(A)_{r+1,r+1} = 0$, is called the *set of Smith invariants* of A. We also call $(A)_{i,i}$ the ith Smith invariant ($i = 1, \ldots, r$). [In the literature, the Smith invariants of A are also called "invariant factors" of A.]

By *elementary operations* in this section, we mean elementary integer row or column operations. We say that two matrices are *equivalent* if they are inter-transformable by elementary operations.

Lemma 19 *Every integer matrix A can be brought into a Smith normal form S by a sequence of elementary operations.*

We leave the proof to an Exercise. The algorithm below implies this, of course, but it is instructive for the student to give a direct proof. We will show that S is unique for A, and so S is *the* Smith normal form of A, usually denoted $S(A)$. For $k = 1, \ldots, \min\{m, n\}$, let $\delta_k(A)$ denote the GCD of all the order k minors of A. In particular, $\delta_1(A)$ is the GCD of all the entries of A.

Lemma 20

(i) *The elementary operations on a matrix A preserve $\delta_k(A)$.*

(ii) *The set of Smith invariants of A is unique.*

(iii) *The Smith normal form of A is unique.*

Proof. (i) This is immediate from our treatment of the γ-invariants for the HNF.
(ii) Let the rank of A be r. Then $\delta_k(A) \neq 0$ iff $k = 1, \ldots, r$. From the definition of the δ's, it is clear that

$$\delta_k(A) | \delta_{k+1}(A)$$

for $k = 1, \ldots, r-1$. Let S be a Smith normal form of A. Since A and S have the same rank, we conclude that $(S)_{k,k} \neq 0$ iff $k = 1, \ldots, r$. Note that $\delta_k(S) = (S)_{1,1}(S)_{2,2} \cdots (S)_{k,k}$. In view of part (i), we conclude

$$(S)_{1,1}(S)_{2,2} \cdots (S)_{k,k} = \delta_k(A).$$

It follows that $(S)_{1,1} = \delta_1(A)$ and $(S)_{k+1,k+1} = \delta_{k+1}(A)/\delta_k(A)$ $(k = 2, \ldots, r-1)$.
(iii) This follows from (ii) since a Smith normal form is determined by the set of its Smith invariants.
Q.E.D.

Polynomial-time Algorithm. Computing the Smith normal form of a matrix A is equivalent to computing the set of Smith invariants. For some applications, it is desirable to also know the unimodular matrices U, V such that

$$S(A) = UAV.$$

For instance, it is easy to compute the 1st Smith invariant – it is just the GCD of all the matrix entries. But computing U and V such that $(UAV)_{1,1}$ has this invariant is non-trivial. As usual, the difficulty in giving a polynomial time algorithm is the possibility of exponential size intermediate entries. We describe an algorithm from Bachem-Kannan [7], based on a Hermite normal form algorithm.

It suffices to show how to perform a *Smith Reduction Step* (in analogy to the Hermite Reduction Step of §6): given a matrix $A \in \mathbb{Z}^{m\times n}$, compute two unimodular matrices $U \in \mathbb{Z}^{m\times m}, V \in \mathbb{Z}^{n\times n}$ such that UAV is "Smith-reduced". In general, a matrix M is said to be *Smith-reduced* if:
(i) The first row and first column of M are zero except for the top-left corner entry $(M)_{1,1}$.
(ii) $(M)_{1,1}$ divides all the remaining entries of M.

Our Hermite normal form is based on column operations. We now need the "row version" of the normal form: a matrix A is in *row Hermite normal form* (abbr. RHNF) if its transpose A^T is in Hermite normal form. Using elementary row operations, or multiplication by unimodular matrices on the left, we can transform any matrix into its RHNF. Algorithms for RHNF are trivially obtained from HNF algorithms, simply by interchanging the roles of rows and columns.

```
Smith Reduction Step
    Input:   A ∈ Z^{m×n}. Assume (A)_{1,1} ≠ 0.
    Output:  Unimodular matrices U ∈ Z^{m×m} and V ∈ Z^{n×n}, and matrix S
                 such that S is Smith-reduced and HNF, and S = UAV.
INITIALIZATION:
    1.  S ← A.
    2.  U ← I_m; V ← I_n (identity matrices).
MAIN LOOP:
        {Loop Invariant: S = UAV}
    3.  while S is not Smith-reduced do
    4.      if (S)_{1,1} is the only non-zero entry in the first row
                and first column, then (S)_{1,1} does not divide some (S)_{i,j}.
    5.          In this case, add column j to column 1, and update V.
    6.      Apply a RHNF algorithm to S and update U accordingly.
    7.      Apply a HNF algorithm to S and update V accordingly.
```

Analysis. The bit-sizes of entries remain polynomially bounded between successive while-iterations: this is because S is in HNF on exit from an iteration, and so the largest entry lies on the diagonal. But the diagonal entries of S are bounded by the determinant of the input matrix A, since the elementary operations preserve $\delta_k(A)$ for each k. Since the RHNF and HNF algorithms are polynomial time, we conclude that each while-iteration is polynomial-time.

It remains to show that the number of iterations is polynomial. Note that whenever $(S)_{1,1}$ changes after an iteration, it is being replaced by a factor of itself. We claim that $(S)_{1,1}$ must change at least in every other iteration. To see this, consider the two possibilities: either step 5 (adding column j to column 1) is executed or it is not. *Step 5 is executed*: then column 1 contains an entry not divisible by $(S)_{1,1}$ before the RHNF algorithm. After the RHNF algorithm, $(S)_{1,1}$ will contain the the GCD of entries in column 1, and this will be a proper factor of its previous value. *Step 5 is not executed*: then row 1 or column 1 must have at least some other non-zero entry. Again there are two cases. (i) If all of these non-zero entries are divisible by $(S)_{1,1}$ then after this iteration, $(S)_{1,1}$ is the only non-zero entry in row 1 and column 1. If this is not the last iteration, then we already saw that $(S)_{1,1}$ will change in the next iteration. (ii) If there is an entry that is not divisible by $(S)_{1,1}$ then either the RHNF or HNF algorithm will change $(S)_{1,1}$ to a smaller value. Hence the number of iterations is at most $1 + 2\lg\|A\|$.

One other remark: the output matrices U, V are of polynomial bit-size. This is because each matrix is the product of $O(\lg\|A\|)$ many component matrices (produced by the call to HNF or RHNF or by Step 5). But each component matrix is of polynomial bit-size. This concludes our analysis of the Smith Reduction Step algorithm. Clearly the Smith normal form of an $m \times n$ matrix can be reduced to $\min\{m, n\}$ Smith Reduction steps. Moreover, the result of Smith Reduction Step is an HNF, and hence polynomially bounded in terms of the original input. This proves:

Theorem 21 *There is a polynomial-time algorithm to compute the Smith normal form $S(A)$ of a matrix A. This algorithm simultaneously computes two unimodular matrices U, V such that $S(A) = UAV$.*

Exercises

Exercise 8.1: Show lemma 19 by a direct argument (*i.e.*, without reference to the existence of the SNF algorithm). □

Exercise 8.2: Let d_i $(i = 1, \ldots, r)$ be the ith Smith invariant of A. Write

$$d_i = p_1^{e_{i,1}} p_2^{e_{i,2}}, \ldots, p_{\ell_i}^{e_{i,\ell_i}}, \qquad (\ell_i \geq 1)$$

where p_i is the ith prime number. Call the prime power $p_j^{e_{i,j}}$ an *elementary divisor* of A. Show that two matrices are equivalent iff they have the same rank and the same set of elementary divisors. □

Exercise 8.3: Analyze the complexity of the Smith Reduction Step and the associated SNF algorithm. □

Exercise 8.4: Let $A, B, C \in \mathbb{Z}^{n\times n}$. We say A is *irreducible* if, whenever $A = BC$ then either B or C is unimodular. Otherwise A is *reducible*. If $A = BC$, we call C a *right divisor* of A or C *right-divides* A. Write $C|A$ in this case. Similarly, there is a notion of *left divisor*.
(i) An irreducible matrix is equivalent to $\texttt{diag}(1, \ldots, 1, p)$, the diagonal matrix whose main diagonal has all ones except the last entry, which is a prime p.
(ii) A necessary and sufficient condition for a square matrix to be irreducible is that its determinant is prime.
(iii) A reducible matrix can be factored into a finite product of irreducible matrices. Formulate a uniqueness property for this factorization. □

Exercise 8.5: Let $A, B, C, D \in \mathbb{Z}^{n\times n}$. Assume A and B are not both zero. Call D a (right) *greatest common divisor* (GCD) of A, B if $D|A$ and $D|B$ and for any other C that divides both A and B, then $C|D$. (See definitions in previous exercise.)
(i) Show that GCDs of A, B exist. Moreover, if D is a GCD then $D = PA + QB$ for some P, Q. HINT: consider the SNF of $[A|B]$.
(ii) If D, D' are two GCD's of A then $D = UD'$ for some unimodular matrix U. □

§9. Further Applications

Linear Diophantine equations. Let $A \in \mathbb{Z}^{m\times n}$ and $b = (b_1, \ldots, b_n) \in \mathbb{Z}^{1\times n}$. Consider the problem of solving the linear system

$$x \cdot A = b \tag{52}$$

for an unknown $x = (x_1, \ldots, x_m) \in \mathbb{Z}^{1\times m}$. Such a system is also called a *Diophantine linear system*. For simplicity, assume $m \geq n$; otherwise, the n equations in (52) are not independent and some may be omitted. Let $S = UAV$ be the Smith normal form of A and let the diagonal elements of S be $d_1, \ldots, d_n$. Then (52) implies

$$\begin{aligned} (xU^{-1})(UAV) &= bV, \\ \widehat{x}S &= \widehat{b}. \end{aligned}$$

where $\widehat{x} = (\widehat{x}_1, \ldots, \widehat{x}_m) = xU^{-1}$ and $\widehat{b} = (\widehat{b}_1, \ldots, \widehat{b}_n) = bV$. The last equation amounts to

$$\widehat{x}_i d_i = \widehat{b}_i, \qquad i = 1, \ldots, n. \tag{53}$$

Suppose $d_1, \ldots, d_r$ are non-zero and $d_{r+1} = \cdots = d_n = 0$. Then the system (52) has solution iff
(i) $d_i|\widehat{b}_i$ for $i = 1, \ldots, r$ and
(ii) $\widehat{b}_i = 0$ for $i = r+1, \ldots, n$.
If these conditions are satisfied then a *general solution* to (53) is given by setting $\widehat{x}_i = \widehat{b}_i/d_i$ for $i = 1, \ldots, r$ and arbitrary assignments to $\widehat{x}_i$ for $i = r+1, \ldots, m$. For instance, we may choose

$$\widehat{x} = (\widehat{b}_1/d_1, \ldots, \widehat{b}_r/d_r, 0, \ldots, 0).$$

From any such *particular solution* we obtain a solution $x = \widehat{x}U$ to the original system (52).

Homogeneous Case. The important special case of (52) where $b = \mathbf{0}$ is said to be *homogeneous*. A solution to this homogeneous system $xA = \mathbf{0}$ is called a *null-vector* of A. The set $N(A) \subseteq \mathbb{Z}^m$ of these null-vectors forms a $\mathbb{Z}$-module. By the Hilbert basis theorem for modules (see Lecture XI), $N(A)$ has a finite basis. Now $e_{r+1}, \ldots, e_m$ is a basis for the set of solutions to (53) in the homogeneous case. Here e_i denotes the m-vector with 1 in the ith position and 0 everywhere else. We conclude that the set

$$e_{r+1}U, e_{r+2}U, \ldots, e_mU$$

is a basis for $N(A)$. Therefore, we may consider a *complete solution* of (52) to have the form $(\widehat{x}U, e_{r+1}U, \ldots, e_mU)$ where $\widehat{x}$ is any particular solution to (53). Notice that $N(A)$ is a lattice, which we may regard as the "dual" of the lattice $\Lambda(A)$ (§VIII.1).

If $A \in \mathbb{Z}^{m\times n}$ and $m > n$ then we have just shown that the homogeneous Diophantine system $x \cdot A = \mathbf{0}$ has non-trivial solutions. An interesting question is whether there exist small integer solutions. Siegel (1929) shows: *there is an* $x \in N(A)$ *satisfying*

$$\|x\| < 1 + (m\|A\|)^{n/(m-n)}. \tag{54}$$

We leave the demonstration to an exercise.

Finitely-generated Abelian groups. Smith normal forms are intimately related to the theory of finitely generated Abelian groups. Let G be a *finitely-presented* Abelian group, that is, G is represented by n *generator* $x_1, \ldots, x_n$ and m *relations* of the form

$$\sum_{j=1}^{n} a_{ij} x_j = 0, \qquad (a_{ij} \in \mathbb{Z})$$

for $i = 1, \ldots, m$. (Note the convention of writing the operations of an Abelian group "additively".) The corresponding *relations matrix* is an $m \times n$ integer matrix A where $a_{ij} = (A)_{i,j}$. We may rewrite the relations in the form $Ax = \mathbf{0}$ where $x = (x_1, \ldots, x_n)^T$. In the special case where $m = 0$ (alternatively, the matrix A is all zero) deserves mention: the group G in this case is called the *free Abelian group of rank n*. Clearly, $G \approx \mathbb{Z}^n$ where $\approx$ indicates group isomorphism.

Let the Smith normal form of A be $S = S(A) = UAV$ for some unimodular matrices U, V. This amounts to transforming the generators of G to $(y_1, \ldots, y_n)^T = V^{-1}(x_1, \ldots, x_n)^T$. Then $S \cdot (y_1, \ldots, y_n)^T = \mathbf{0}$. If S has rank r and the diagonal elements of S are $d_1, \ldots, d_{\min(m,n)}$ then we see that $d_i y_i = 0$ for $i = 1, \ldots, r$ and $y_{r+1}, \ldots, y_n$ satisfy no relations whatsoever. Each y_i generates the subgroup $G_i = \mathbb{Z}y_i$ of G. Moreover, $G_i \approx \mathbb{Z}_{d_i}$ for $i = 1, \ldots, r$ and $G_i \approx \mathbb{Z}$ for $i = r+1, \ldots, n$. Clearly G is a direct sum of the G_i's: $G = \oplus_{i=1}^{n} G_i$. There are three kinds of subgroups:
$d_i = 0$: these correspond to torsion-free subgroups $G_i \approx \mathbb{Z}$. The number β of these subgroups is called the *Betti number* of G.
$d_i = 1$: these are trivial subgroups, and may be omitted in the direct sum expression.
$d_i \geq 2$: these give rise to finite cyclic groups G_i. These d_i's are called *torsion coefficients* of G.

We have just proven the "fundamental theorem of finitely generated Abelian groups": *every finitely presented Abelian group G on n generators can be written as a direct sum $G = \oplus_{i=0}^{r} H_i$ where H_0 is a free Abelian group of rank β, and each H_i $(i = 1, \ldots, r)$ is a finite cyclic group of order $d_i \geq 2$ satisfying such that $d_1 | d_2 | \cdots | d_r$. The numbers $\beta, d_1, \ldots, d_r$ are uniquely determined by G.*

It follows that a polynomial time algorithm for SNF implies that we can check for isomorphism between two finitely generated Abelian groups in polynomial time. A slightly different group isomorphism problem arises if we assume that finite groups are represented by their multiplication tables instead of by a set of relations. An observation of Tarjan implies that we can check isomorphism of two such groups in $O(n^{\log n + O(1)})$ time. For the Abelian case, Vikas [212] has shown an $O(n \log n)$ isomorphism algorithm.

Exercises

Exercise 9.1: Let $a, b, c \in \mathbb{Z}$ where a, b are relatively prime. Suppose $sa + tb = 1$. Show that the general solution of the Diophantine equation $ax + by = c$ is $(x, y) = (sc + nb, tc - na)$ where n is any parameter. □

Exercise 9.2: Consider $N(A)$ in case $n = 1$. Say $A = (a_1, \ldots, a_m)^T$.
(i) Let $s = (s_1, \ldots, s_m)$ be a co-factor of $(a_1, \ldots, a_m)$. Show that $\mathbb{Z} \cdot N(A) + \mathbb{Z} \cdot s$ is the unit lattice $\mathbb{Z}^m$.
(ii) For $1 \leq i < j \leq m$ let $T(i, j)$ be the m-vector that is zero everywhere except $a_i / \texttt{GCD}(a_i, a_j)$ at the jth position and $-a_j / gcd(a_i, a_j)$ at the ith position. The set of these $T(i, j)$'s generates $N(A)$. □

Exercise 9.3: (i) Show the bound of Siegel (54). HINT: for H a parameter to be chosen, let $C = C_H$ be the cube comprising points $x \in \mathbb{R}^m$ where $\|x\| \leq H$. Let $\alpha : \mathbb{R}^m \to \mathbb{R}^n$ be the

linear map given by $\alpha(x) = x \cdot A$. Give a cube C' in $\mathbb{R}^n$ the contains $\alpha(C)$. Use a pigeon hole argument to show that α is not $1-1$ on the integer points of C. See [179] for several versions of Siegel's bound.
(ii) Show that the exponent $n/(m-n)$ cannot be improved. □

Exercise 9.4: Show:
(i) If d_1, d_2 are co-prime then $\mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \approx \mathbb{Z}_{d_1 d_2}$.
(ii) Every finite cyclic group is a direct sum of cyclic groups of prime power.
(iii) Every finitely generated Abelian group written as a direct sum $G = \oplus_{i=0}^{\ell} H_i$ where H_0 is a free Abelian group of rank β, and $H_i \approx \mathbb{Z}_{q_i}$ where q_i is a prime power $(i = 1, \ldots, \ell)$. Moreover, the numbers $\beta, q_1, \ldots, q_\ell$ are uniquely determined by G. (These q_i's are called the *invariant factors* of G.) □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture XI
Elimination Theory

Algebraic geometry has been reformulated several times and, by Dieudonné's account [54], contains seven epochs leading to the present time. In an earlier incarnation, elimination theory and more generally, invariant theory, played a central role. Early invariant theory (under Cayley, Sylvester and Gordan) relies on constructive methods, as epitomized by the "master computer" Gordan. Hilbert's non-constructive proof[1] in 1888 of his basis theorem is said to have killed invariant theory as an active area. Nevertheless, Emmy Noether wrote a thesis on the subject in 1907 under Gordan. Recent interest in constructive methods have revived interest in these topics. See Sturmfels [198] for a modern computational perspective on invariant theory.

This lecture concerns elimination theory. Elimination theory is the study of conditions on coefficients of a system of polynomials that are necessary and sufficient for the system to have a solution. We are familiar with this idea in two important special cases:

- When there are n linear homogeneous equations in n unknowns, the vanishing of the determinantal function on the coefficients is a necessary and sufficient condition for solvability.

- When we have two univariate polynomials, the vanishing of the Sylvester resultant of these polynomials gives a necessary and sufficient condition for solvability.

These are both special cases of the fact that we can define a "resultant polynomial" R for any system Σ of n homogeneous polynomials in n variables such that R is a polynomial in the coefficients of these polynomials, and the vanishing of R is a necessary and sufficient condition for the solvability of the system (we prove this in §6). Elimination theory is useful in other unexpected ways. For instance, if $F(x, y, \lambda)$ is a family of plane curves parameterized by λ, then the envelop curve of the family can be obtained by eliminating the parameter from the equations $F = 0$ and $\partial F/\partial\lambda = 0$. This is a form of the *implicitization problem*, to compute a system of equations that defines an algebraic set given in some other form. The converse problem is the *parameterization problem* (or the construction of "generic solutions", see §4).

The results in this lecture are updates of classical work that goes back to Macaulay, Hurwitz and others. In recent years, Canny [36, 37] revisited this classical literature, showing its usefulness for efficient algorithms. This lecture is clearly influenced by his work. Because the classical background is scattered and perhaps unfamiliar to the modern reader, we have tried to be self-contained in this exposition (except in §10,11). There is a newer body of work on sparse elimination theory (e.g., Sturmfels [199], Emiris [64]) which we unfortunately omit. A deep investigation of elimination theory is found in the book of Gelfand, Kapranov and Zelevinsky [70]. Furthermore, there has been a large output of recent material on techniques for solving algebraic systems.

Most results in this lecture, unless otherwise noted, pertain to a Noetherian UFD D. We use $\mathbf{X}$ to denote the set of indeterminates $\{X_1, \ldots, X_n\}$.

§1. Hilbert Basis Theorem

[1] Which led Gordan to his famous remark "this is not mathematics, this is theology". Gordan later admitted that "theology has its uses". See Kline [103] for a fuller account of the period.

Hilbert proved three fundamental results that can be regarded as the starting point of modern algebraic geometry: the Basis Theorem, the Zero Theorem (Nullstellensatz) and the Syzygy Theorem. This section proves the Basis Theorem for polynomial rings. The Zero Theorem is treated in the next section while the Syzygy Theorem is found in §XII.8 and §XIII.2.

Let $S \subseteq R[\mathbf{X}]$ be an arbitrary set of polynomials, R is a ring. A *basis* for S is any subset $B \subseteq S$ such that $S \subseteq \texttt{Ideal}(B) \subseteq R[\mathbf{X}]$. Alternatively, B is a basis if each $f \in S$ can be expressed as a linear combination of elements of B:

$$f = \sum_{i=1}^{m} f_i b_i, \qquad (f_i \in R[\mathbf{X}], b_i \in B).$$

Note that S is not necessarily an ideal. In case S is an ideal, modern terminology prefers to call B a *generator set* for S. A basis or generator set is *finite* if it has finitely many elements. A ring R is *Noetherian* if every ideal in R has a finite generator set. For example, if R is a field then it is Noetherian since R has only two ideals (0) and (1). It is also easy to see that $\mathbb{Z}$ is Noetherian.

Theorem 1 (Hilbert Basis Theorem) *If R is Noetherian then $R[\mathbf{X}]$ is Noetherian.*

Proof. It is enough[2] to prove this for $\mathbf{X} = \{X\}$. The theorem then follows by induction on the number of variables in $\mathbf{X}$: if $R[X_1, \ldots, X_{n-1}]$ is Noetherian then so is $R[X_1, \ldots, X_{n-1}][X_n]$. So we want to show that any given ideal $I \subseteq R[\mathbf{X}]$ has a finite basis. Let us construct a sequence of polynomials

$$f_1, f_2, f_3, \ldots,$$

where f_1 is a smallest degree polynomial in I, and in general f_{i+1} is a smallest degree polynomial in $I \setminus \texttt{Ideal}(f_1, \ldots, f_i)$. The result is proved if this process stops in a finite number of steps, *i.e.*, when $I = \texttt{Ideal}(f_1, \ldots, f_i)$ for some i. So by way of contradiction, assume that the sequence $f_1, f_2, \ldots$, is infinite. Consider the corresponding sequence

$$a_1, a_2, a_3, \ldots,$$

where $a_i = \texttt{lead}(f_i)$. Since R is Noetherian, there exists a first value a_{k+1} such that

$$a_{k+1} \in \texttt{Ideal}(a_1, \ldots, a_k) \subseteq R. \tag{1}$$

Hence a_{k+1} has the form

$$a_{k+1} = \sum_{j=1}^{k} c_j a_j \quad (c_j \in R).$$

If the degree of f_j is n_j then $n_1 \leq n_2 \leq n_3 \leq \ldots$. Note that polynomial

$$g := \sum_{j=1}^{k} c_j X^{n_{k+1} - n_j} f_j$$

has leading coefficient a_{k+1}. Hence the degree of $f_{k+1} - g$ is less than n_{k+1}. Since $f_{k+1} - g \in I \setminus \texttt{Ideal}(f_1, \ldots, f_k)$, this means f_{k+1} is not of minimum degree in $I \setminus \texttt{Ideal}(f_1, \ldots, f_k)$, a contradiction. **Q.E.D.**

[2] This proof of Heidrun Sarges (1976), like the original proof of Hilbert, is non-constructive. Hilbert's proof was a *tour de force* in his time, much to the chagrin of constructivists like Gordan and Kronecker. Gordan, by involved arguments, only managed to construct finite bases for ideals in two variables. Some historians of mathematics have pin-pointed this proof as the genesis of the modern tendency to use non-constructive proofs. Hilbert subsequently returned to his proof to make it constructive [165]. For a modern constructive treatment see [198].

Corollary 2 *If R is Noetherian, then every set $S \subseteq R[\mathbf{X}]$ has a finite basis.*

Proof. By the theorem, $\texttt{Ideal}(S)$ has a finite basis $B' \subseteq \texttt{Ideal}(S)$. Each $b' \in B'$ is a linear combination of some finitely many elements of S. Hence there is a finite subset $B \subseteq S$ such that each $b' \in B'$ is a linear combination of B. Clearly B is a basis for S. **Q.E.D.**

This corollary is the original basis theorem shown by Hilbert. Modern texts usually only treat the case where S is an ideal. In an application below, we need this more general form.

The basis theorem has the following implication for solving polynomial equations in $D[X_1, \ldots, X_n]$ where D is a Noetherian domain. Note that if B is a basis for a set $S \subseteq D[X_1, \ldots, X_n]$ then $(x_1, \ldots, x_n) \in \overline{D}^n$ is a zero of S iff $(x_1, \ldots, x_n)$ is a zero of B. So without loss of generality, it is enough to solve only finite systems of equations.

We now prove Hilbert's basis theorem for modules. Let R be a ring and M an R-module. If $x \in M$, then xR denotes the set $\{xa : a \in R\}$. If $M_i \subseteq M$ ($i \in J$ where J is an index set) then $\sum_{i\in J} M_i$ denotes the set of all sums of the form $\sum_{i\in J} x_i$ where $x_i \in M_i$ or $x_i = 0$, and moreover only finitely many x_i's are non-zero. Say a set $S \subseteq M$ *generates* M if $\sum_{x\in S} xR = M$. M is *finitely generated* if it has a finite set S as generator. M is *Noetherian* if every R-submodule of M is finitely generated. (Note that the concept of Noetherian modules generalizes the concept of Noetherian rings.)

Let us illustrate these definitions as well as the theorem to be stated: $R = \mathbb{Z}$ is a Noetherian ring and $M = \mathbb{Z}^n$ is a $\mathbb{Z}$-module. Note that a $\mathbb{Z}$-submodule of M is also called a lattice (§VIII.1). Clearly, M is finitely generated. The next theorem implies that every $\mathbb{Z}$-submodule of M is finitely generated.

One more definition. For R-modules M and N, a map $\varphi : M \to N$ is an *R-module homomorphism* if $\varphi(x + y) = \varphi(x) + \varphi(y)$, $\varphi(ax) = a\varphi(x)$ for all $x, y \in M, a \in R$.

Theorem 3 (Hilbert's Basis Theorem for modules) *If R is a Noetherian ring and M is a finitely generated R-module then M is Noetherian.*

Proof. Let $M = x_1R + \cdots + x_nR$. Then we have a canonical R-homomorphism

$$\varphi : R^n \longrightarrow M$$

where $(a_1, \ldots, a_n) \in R^n$ is mapped to $\sum_{i=1}^n a_i x_i \in M$. Let $N \subseteq M$ be any submodule. We have to show that N has a finite set of generators. Let

$$U = \{\mathbf{a} \in R^n : \varphi(\mathbf{a}) \in N\}.$$

One checks that U is a submodule of R^n. If $U = \mathbf{a}_1R + \cdots + \mathbf{a}_mR$ $(\mathbf{a}_i \in R^n)$ then $N = \varphi(\mathbf{a}_1)R + \cdots + \varphi(\mathbf{a}_n)R$. So it suffices to show U is finitely generated. We first observe that the set of first components of members of U form an ideal I in R. Since R is Noetherian, $I = (u_1, \ldots, u_k)$ for some $u_1, \ldots, u_k$. Pick $\mathbf{a}_1, \ldots, \mathbf{a}_k \in U$ such that the first component of $\mathbf{a}_i$ is u_i $(i = 1, \ldots, k)$. Let $V \subseteq U$ be the set of those n-vectors with zero as first component. Then

$$U = \mathbf{a}_1R + \cdots + \mathbf{a}_kR + V.$$

But V is isomorphic to a submodule of R^{n-1}. If $n = 1$, then we are done. Otherwise by induction, V has a finite set of generators and so does U. **Q.E.D.**

Exercises

Exercise 1.1: The following are standard facts:

(i) For $I \subseteq R$ an ideal in a ring R, I is maximal (resp., prime, primary) iff R/I is a field (resp., a domain, a ring in which all zero divisors are nilpotent). An element x is nilpotent if some power x^m is zero.

(ii) An ideal is said to be irreducible if it is not the proper intersection of two ideals. Show that prime ideals are irreducible and under the ascending chain condition, irreducible ideals are primary. □

Exercise 1.2:

(i) The homomorphic image of a Noetherian ring is Noetherian.

(ii) Every non-unit in a Noetherian domain is a product of irreducible elements. □

§2. Hilbert Nullstellensatz

Henceforth, let [3] D be a Noetherian UFD. This section gives several forms of the Nullstellensatz of Hilbert for D. Basically, the Nullstellensatz is a theorem about the existence of zeros in the algebraic closure $\overline{D}$ of D. Accordingly, for a set $S \subseteq D[X_1, \ldots, X_n]$, a *zero* of S is an element $(x_1, \ldots, x_n) \in \overline{D}^n$ such that $p(x_1, \ldots, x_n) = 0$ for each polynomial $p \in S$. We also say S *vanishes at* $(x_1, \ldots, x_n)$. Denote the set of zeros of S by

$$\textsc{Zero}(S) \subseteq \overline{D}^n.$$

We begin with what is known as the field-theoretic version of Hilbert's Nullstellensatz [107, 111]. By way of motivation, note that if $D(\xi_1, \ldots, \xi_n)$ is an algebraic extension of D then $D(\xi_1, \ldots, \xi_n)$ is obtained as a ring adjunction of the quotient field Q_D of D, *i.e.*, $D(\xi_1, \ldots, \xi_n) = Q_D[\xi_1, \ldots, \xi_n]$. We show the converse.

Theorem 4 (Nullstellensatz – field-theoretic form) *Let D be a Noetherian UFD and E a field extension of D. If $E = Q_D[\xi_1, \ldots, \xi_n]$ for some $\xi_1, \ldots, \xi_n$, then E is algebraic over D.*

We first prove two lemmas.

Lemma 5 (Artin-Tate) *Let $R \subseteq S \subseteq T$ be rings, R Noetherian and T be finitely generated as an S-module. If T is finitely generated as a ring over R, i.e., $T = R[\xi_1, \ldots, \xi_n]$, then S is finitely generated as a ring over R.*

Proof. Since T is a finitely generated S-module, let $\omega_1, \ldots, \omega_m \in T$ such that

$$T = S\omega_1 + S\omega_2 + \cdots + S\omega_m. \tag{2}$$

We may assume $\xi_1, \ldots, \xi_n$ is contained among $\omega_1, \ldots, \omega_m$. Let

$$M = \{a_k^{i,j} : i, j, k = 1, \ldots, m\}$$

be the multiplication table of $\omega_1, \ldots, \omega_m$, that is,

$$\omega_i \omega_j = \sum_{k=1}^{m} a_k^{i,j} \omega_k.$$

[3] It is known that a Noetherian domain is a UFD iff all its height one prime ideals are principal.

We may assume $M \subseteq S$ because of (2). Consider the ring

$$S' := R[M] \subseteq S.$$

From $T = R[\xi_1, \ldots, \xi_n]$ we conclude that T is generated as an R-module by the set of power products of $\xi_1, \ldots, \xi_n$. But since the multiplication table of $\xi_1, \ldots, \xi_n$ is in S' we get

$$T = S'\omega_1 + S'\omega_2 + \cdots + S'\omega_m.$$

Since R is Noetherian, Hilbert's basis theorem (for polynomials) implies S' is Noetherian. By Hilbert's basis theorem (for modules) T is a Noetherian S'-module, and since S is a S'-submodule of T, S must be finitely generated over S': for some $\{u_1, \ldots, u_t\} \in T$,

$$S = S'u_1 + S'u_2 + \cdots + S'u_t.$$

Hence $S = R[M, u_1, u_2, \ldots, u_t]$. **Q.E.D.**

Lemma 6 *Let $S = D(Z_1, \ldots, Z_t)$ be a rational function field over D with $t \geq 1$ indeterminates. Then S is not finitely generated as a ring over Q_D.*

Proof. Suppose $S = Q_D[\xi_1, \ldots \xi_n]$ where $\xi_i = \frac{f_i(Z_1,\ldots,Z_t)}{g_i(Z_1,\ldots,Z_t)}$ and $f_i, g_i \in D[Z_1, \ldots, Z_t]$, $(i = 1, \ldots, n)$. Then each element of $Q_D[\xi, \ldots, \xi_n]$ has the form $\frac{f}{g} = \frac{f(Z_1,\ldots,Z_t)}{g(Z_1,\ldots,Z_t)}$ where

$$g(Z_1, \ldots, Z_t) = a \prod_{i=1}^{n} g_i^{e_i}(Z_1, \ldots, Z_t), \quad a \in D, e_i \geq 0. \tag{3}$$

There are infinitely many non-associated irreducible polynomials in $Q_D[Z_1, \ldots, Z_t]$ (see appendix B). Pick an irreducible polynomial $p \in Q_D[Z_1, \ldots, Z_t]$ that does not divide any g_i. Then $1/p \in S$ implies that it has a representation of the form f/g where g is given by (3). From $1/p = f/g$, we obtain $g = f \cdot p$. Hence p divides some g_i, since $Q_D[Z_1, \ldots, Z_t]$ is a UFD. This contradicts our choice of p. **Q.E.D.**

We now prove the field version of the Nullstellensatz: If $E = Q_D[\xi_1, \ldots, \xi_n]$ is not algebraic, let $Z_1, \ldots, Z_t$ $(t \geq 1)$ be the maximal subset of $\{\xi_1, \ldots, \xi_n\}$ that is algebraically independent over Q_D. Set $S = Q_D(Z_1, \ldots, Z_t) = D(Z_1, \ldots, Z_t)$. Then E is a finite algebraic extension of S. Since S is a field, E is finitely generated as an S-module. Applying the first lemma using $R = D$, $S = D(Z_1, \ldots, Z_t)$ and $T = E = Q_D[\xi_1, \ldots, \xi_n]$ implies $D(Z_1, \ldots, Z_t)$ is finitely generated as a ring over Q_D. This contradicts the second lemma, completing the proof.

Theorem 7 (Nullstellensatz – weak form) *An ideal $I \subseteq D[X_1, \ldots, X_n]$ has no zeros iff I contains a non-zero element of D.*

Proof. Clearly if I contains a non-zero element $a \in D$ then I has no zero. Conversely, if $I \cap D = \{0\}$, we must show that $\textsc{Zero}(I)$ is non-empty. We may (by the maximum principle) assume I is maximal. Then $E = D[\mathbf{X}]/I$ is a field. If $b \in D[\mathbf{X}]$ maps to $\overline{b}$ under the canonical homomorphism $D[\mathbf{X}] \longrightarrow E$ then for $a, b \in D$, $a \neq b$ implies $\overline{a} - \overline{b} \neq 0$ (otherwise $a - b \in I$). So we may assume $D \subseteq E$ and indeed, $Q_D \subseteq E$. Since $E = Q_D[\overline{X}_1, \ldots, \overline{X}_n]$, the previous theorem shows that E is algebraic over D. Thus we may assume that $E \subseteq \overline{D}$. Now the canonical homomorphism takes $p(X_1, \ldots, X_n) \in D[\mathbf{X}]$ to

$$\overline{p(X_1, X_2, \ldots, X_n)} = p(\overline{X}_1, \ldots, \overline{X}_n).$$

Hence $p(\overline{X}_1, \ldots, \overline{X}_n) = 0$ iff $\overline{p(X_1, X_2, \ldots, X_n)} = 0$ iff $p(X_1, \ldots, X_n) \in I$. Hence $(\overline{X}_1, \ldots, \overline{X}_n) \in$ Zero(I). **Q.E.D.**

This theorem is equivalent to an apparently stronger version:

Theorem 8 (Nullstellensatz – strong form) *Let $I \subseteq D[X_1, \ldots, X_n]$ be an ideal and $p \in D[X_1, \ldots, X_n]$. Then p vanishes at all the zeros of I iff there is an $m \geq 0$ and a non-zero $a \in D$ such that*

$$a \cdot p^m \in I.$$

The strong Nullstellensatz implies the weak form: to show the non-trivial direction of the weak Nullstellensatz, suppose Zero(I) is empty. Then (vacuously) 1 vanishes at all the zeros of I. The strong form then implies that $a = a \cdot 1^m \in I$ for some $m \geq 0$ and non-zero $a \in D$.

Conversely, the weak Nullstellensatz implies the strong form: again, in the nontrivial direction, we assume that $p \in D[\mathbf{X}]$ vanishes at Zero(I). Suppose $I =$ `Ideal`$(f_1, \ldots, f_r)$. Using the "trick of Rabinowitz", introduce a new variable Z and let

$$g := 1 - Z \cdot p.$$

Then the ideal $(f_1, \ldots, f_r, g)$ has no zeros since g will not vanish at any zero of $f_1, \ldots, f_r$. Hence the weak Nullstellensatz implies the existence of some nonzero $a \in D \cap$ `Ideal`$(f_1, \ldots, f_r, g)$. Let

$$a = \sum_{i=1}^{r} \alpha_i f_i + \beta(1 - Zp)$$

for suitable $\alpha_i, \beta \in D[X_1, \ldots, X_n, Z]$. Substituting $Z = \frac{1}{p}$, we get:

$$a = \sum_{i=1}^{r} \alpha_i' f_i$$

where each $\alpha_i' \in D(X_1, \ldots, X_n)$ is a rational function whose denominator is some power of p. Multiplying by a suitable power $m \geq 0$ of p, we get

$$a \cdot p^m = \sum_{i=1}^{r} (\alpha_i' p^m) f_i$$

where $\alpha_i' p^m \in D[X_1, \ldots, X_n]$. Thus $a \cdot p^m \in I$, proving the non-trivial direction of the strong Nullstellensatz.

Quantitative Nullstellensatz. It is known from Hermann [80] (cf. [129]) that the number m appearing in the strong Nullstellensatz is at most double exponential in n. In recent years, starting from the work of Brownawell [31], single-exponential bounds began to appear. Essentially the best possible bound is from Kollár [106] (see also [190]). Dubé [60] gives a purely combinatorial proof of similar bounds[4]. We quote without proof the bound of Dubé (somewhat simplified).

Theorem 9 (Nullstellensatz – quantitative form) *Let p vanish at all the zeros of an ideal $I \subseteq D[X_1, \ldots, X_n]$. If I is generated by a set of polynomials of degrees at most d then there exists $a \in D$ such that $a \cdot p^N \in I$ where*

$$N = 13d^n.$$

[4]The bound of Dubé's applies more generally to ideals generated by prime sequences.

Let the "Nullstellensatz bound"

$$N(n, d) \tag{4}$$

to be the least value for N for which this theorem holds. Thus $N(n, d) \leq 13d^n$. We remark that Kollár's formulation of the theorem generally gives better bounds (without the constant factor 13) but it is less suited for our purpose of defining $N(n, d)$ because he has a technical requirement that the generators of I have degrees not equal to 2. The following easy extension of the strong Nullstellensatz will be useful:

Theorem 10 (Nullstellensatz – extended form) *Let $A_1, \ldots, A_r \in D[X_1, \ldots, X_n]$ be polynomials such that each A_i vanishes at all the zeros of an ideal $I \in D[X_1, \ldots, X_n]$. If I is generated by polynomials of degrees at most d then there exists a non-zero $a \in D$ such that*

$$a \cdot A_1^{e_1} A_2^{e_2} \cdots A_r^{e_r} \in I$$

whenever $\sum_{i=1}^r e_i \geq 1 + r(N-1)$, $N = N(n, d)$ and $e_i \geq 0$.

Proof. By definition of $N = N(n, d)$, for each i, we have $a_i A_i^N \in I$ for some non-zero $a_i \in D$. Let $a = \prod_{i=1}^r a_i$. If $\sum_{i=1}^r e_i \geq 1 + r(N-1)$ then some $e_i \geq N$ and so $a_i A_i^{e_i} \in I$. Hence $a \cdot A_1^{e_1} A_2^{e_2} \cdots A_r^{e_r} \in I$. **Q.E.D.**

Exercises

Exercise 2.1: (Corollaries to Hilbert's Nullstellensatz)
(i) If $D = \overline{D}$ then I is a maximal ideal iff $I = \texttt{Ideal}(X_1 - \xi_1, \ldots, X_n - \xi_n)$ where $\xi_i \in D$.
(ii) $D[X_1, \ldots, X_n]/I$ is a finite field extension of D iff I is a maximal ideal.
(iii) Let Ideal be the map taking a set $V \subseteq \mathbb{A}^n(\overline{D}) = \overline{D}^n$ to the set

$$\text{Ideal}(V) = \{f \in D[X_1, \ldots, X_n] : f \text{ vanishes on } V\}.$$

Show that Ideal is a bijection between algebraic sets and radical ideals. □

Exercise 2.2: Show that we can find an exponent m in the Strong Nullstellensatz that depends only on I (and not on p). HINT: show that $(\sqrt{I})^e \subseteq I$ for some e and $p \in \sqrt{I}$. □

Exercise 2.3: (Mishra and Gallo, 1992) Assume $D = \mathbb{Z}$. Obtain primitive recursive bounds on $|a|$ in the strong or quantitative form of the Nullstellensatz. □

§3. Specializations

The informal idea of specialization is that of "substitution of indeterminates". In most applications, this naive understanding suffices. We wish to explore this concept in a more general setting. Specialization arises in three ways in solving systems of polynomial equations over D. In illustration, consider the system

$$F_1 = F_2 = 0 \tag{5}$$

where $F_1(X, Y, Z) = X - Y^2$ and $F_2(X, Y, Z) = XY - Z$ are polynomials over D. First, we see that the specialization

$$(X, Y, Z) \longrightarrow (t^2, t, t^3) \tag{6}$$

is a solution to the system (5), for any $t \in \overline{D}$. The "$\longrightarrow$" notation here means the right-hand quantities are substituted for the left-hand quantities. So (6) is a "specialization" of X, Y, Z. Second, suppose t is an indeterminate. Then we can express another important idea: solution (6) is the most general or "generic" solution to (5) in the following sense:

(i) For any "specialization" of t to a value $\alpha \in \overline{D}$, the corresponding substitution $(X, Y, Z) \longrightarrow (\alpha^2, \alpha, \alpha^3)$ is a solution to (5).

(ii) Moreover, every solution $(\alpha_1, \alpha_2, \alpha_3) \in \overline{D}^3$ of (5) can be obtained as in (i).

The third way in which specialization arises is when we consider polynomials with indeterminate coefficients. We are interested in conditions under which specializations of these coefficients lead to solvability.

As discussed in Lecture 0, it is significant to ask where the solutions $(\alpha_1, \alpha_2, \alpha_3)$ come from. There are three natural cases: the α_i's can come from D, or from its quotient field Q_D, or from the algebraic closure $\overline{D}$. These are, respectively, the *Diophantine case*, the *rational case* and *the algebraic case* of solving equations over D. The Nullstellensatz (§2) concerns the algebraic case. Following A. Weil, we may go beyond the algebraic case by asking for solutions in the *universal field* Ω of D. By definition $\Omega = \Omega_D$ is defined to be

$$\Omega_D = \overline{D}(t_1, t_2, \ldots)$$

where the t_i's is an infinite set of indeterminates. Call this the *universal case* of solving equations over D. Since within each normal context of discourse, we only have to deal with a finite number of these t_i's, we may regard the infinite transcendence degree of Ω as a mere convenience. But the existence of transcendental elements in Ω allows us to accomplish more than the algebraic case: it affords the notion of a generic solution as seen in the above example.[5] Finding generic solutions can be considered as another view of the parameterization problem, briefly noted in this lecture's introduction.

As far as the existence of solutions goes, the universal case of solving polynomial equations does not add anything: a system of polynomial equations over D is solvable in $\overline{D}$ iff it is solvable in Ω_D. This is because any transcendental solution can always be "specialized" to a non-trivial algebraic solution.

The concept of specialization. We formalize the above notions. In the following, let $S \subseteq \Omega$. We write

$$D[S] \quad \text{and} \quad D(S)$$

to denote the smallest subring and subfield (respectively) of Ω containing $D \cup S$. This is a natural extension of the standard notations, $D[X]$ and $D(X)$. Suppose

$$\sigma : S \to \Omega$$

is any function. We would like to extend σ into the *canonical homomorphism*

$$h_\sigma : D[S] \to \Omega$$

where, for each $F(X_1, \ldots, X_n) \in D[\mathbf{X}]$ and $x_1, \ldots, x_n \in S$,

$$h_\sigma(F(x_1, \ldots, x_n)) = F(\sigma x_1, \ldots, \sigma x_n).$$

[5] It is possible to avoid this universal field; indeed, contemporary algebraic geometry prefers instead to use the set of prime ideals of $D[X_1, \ldots, X_n]$ for Ω. But as in this example, the language of the universal field seems more intuitive and geometric.

But one must verify that h_σ is well-defined. A necessary condition is that $\sigma(a) = a$ whenever $a \in D$. Clearly if h_σ were well-defined, it would be a homomorphism. In fact, it would be a D-homomorphism (*i.e.*, it is the identity function when restricted to D).

Usually, S is a set of transcendental[6] elements over D. If these transcendental elements are algebraically independent over D, meaning that $F(x_1, \ldots, x_n) \neq 0$ for all $F(X_1, \ldots, X_n) \in D[X_1, \ldots, X_n] \setminus 0$ and $x_1, \ldots, x_n \in S$. We may call S a *set of indeterminates* and elements of S are called indeterminates. So the concept of an "indeterminate" is always relative to some such set S, which is often implicit. This definition of indeterminates agrees with the usual informal uses of the term; of course, we have relied on this informal understanding in the earlier discussions.

Example:

(i) If S is a set of indeterminates, h_σ is clearly well-defined for any σ.
(ii) S may contain algebraic relations among its members, as in our introductory example with $S = \{t, t^2, t^3\}$.
(iii) S may even have algebraic elements. Consider $D = \mathbb{R}$ and $S = \{\mathbf{i}\}$ (where $\mathbf{i} = \sqrt{-1}$). The specialization which maps $\mathbf{i}$ to $-\mathbf{i}$ amounts to complex conjugation.
(iv) Let $S = \{X, XY\}$ where $\{X, Y\}$ is a set of indeterminates. The map h_σ is well-defined provided $\sigma(X) = 0$ implies $\sigma(XY) = 0$. ■

Definition: Let $S \subseteq \Omega$. An *S-specialization* (or simply, *specialization*) is a function $\sigma : S \to \Omega$ such that for all $F(X_1, \ldots, X_n) \in D[\mathbf{X}]$ (the X_i are variables) and for all $x_1, \ldots, x_n \in S$,

$$F(x_1, \ldots, x_n) = 0 \Longrightarrow F(\sigma x_1, \ldots, \sigma x_n) = 0.$$

We claim: *h_σ is well-defined iff σ is a specialization.* To show this in one direction, if $F(x_1, \ldots, x_n) = 0$ and $h_\sigma(F(x_1, \ldots, x_n)) \neq 0$, *i.e.*, σ is not a specialization, then $h_\sigma(0)$ is not well-defined. In the other direction, if σ is a specialization, we must show that for any $F, G \in D[\mathbf{X}]$, if $F(x_1, \ldots, x_n) = G(x_1, \ldots, x_n)$ then $h_\sigma(F(x_1, \ldots, x_n)) = h_\sigma(G(x_1, \ldots, x_n))$. This is equivalent to showing that $F(x_1, \ldots, x_n) = 0$ implies $h_\sigma(F(x_1, \ldots, x_n)) = 0$. But this is immediate since $h_\sigma(F(x_1, \ldots, x_n)) = F(\sigma(x_1), \ldots, \sigma(x_n))$ which equals 0 since σ is a specialization.

We may think of specializations as a type of ring homomorphism. We call h_σ the *canonical σ-homomorphism* from $D[S]$ to $D[\sigma(S)]$ (or, to Ω). We say that $h_\sigma(x)$ is the *specialization* of x (under σ). Since h_σ is an extension of σ, we often continue to use 'σ' instead of 'h_σ'. In case $\sigma(S) = S$, we have a (the older literature calls it a 'substitution'). If $\sigma(S) \subseteq D$ we call σ a *ground specialization*. We say a specialization σ is *partial* if for some x_i, $\sigma(x_i) = x_i$. A *partial ground specialization* is σ such that for each $c \in S$, either $\sigma(c) = c$ or $\sigma(c) \in D$. If $\sigma(S) \subseteq \overline{D}$ we call σ an *algebraic specialization*.

Generic points. Henceforth, assume $S \subseteq \Omega_D$ is a finite set. Fixing any enumeration $\mathbf{x} = (x_1, \ldots, x_n)$ of S, we may represent a S-specialization σ by the sequence

$$\mathbf{y} = (\sigma(x_1), \ldots, \sigma(x_n)).$$

Let us write (following Artin [5])

$$\mathbf{x} \xrightarrow{\sigma} \mathbf{y} \quad \text{or} \quad \mathbf{x} \xrightarrow{S} \mathbf{y} \quad \text{or} \quad \mathbf{x} \longrightarrow \mathbf{y} \tag{7}$$

[6] An element x is *transcendental* over D if $F(x) \neq 0$ holds for any non-zero polynomial $F(X) \in D[X]$. Otherwise, it is *algebraic* over D.

to indicate that $\mathbf{y}$ is an S-specialization. If σ is an S-specialization and τ is a $\sigma(S)$-specialization then the composition $\tau \circ \sigma$ is an S-specialization. In the arrow notation,

$$\mathbf{x} \longrightarrow \mathbf{y} \longrightarrow \mathbf{z} \quad \text{implies} \quad \mathbf{x} \longrightarrow \mathbf{z}.$$

An element $\mathbf{y} \in \Omega^n$ $(n \geq 1)$ is called an *n-point*, or, simply a *point* in the *affine n-space*

$$\mathbb{A}^n(\Omega) := \Omega^n.$$

Thus each $\mathbf{x}$-specialization can be viewed as point. This gives a more geometric language for specializations.

For any subset $U \subseteq \Omega$, the *specialization of U under σ* is the set $\sigma(U) = \{h_\sigma(a) : a \in U\}$, which we also denote by

$$U|_\sigma.$$

In case $U|_\sigma = \{0\}$, we simply write $U|_\sigma = 0$. In this case, we say that U *vanishes* under σ, or[7] σ is an *S-solution* (or simply, *solution*) of U. The solution σ is *non-trivial* if σ is not identically zero. Specializations can be composed: if τ is a another specialization, then we write

$$U|_{\sigma,\tau}$$

instead of the more awkward $(U|_\sigma)|_\tau$.

Let $V \subseteq \mathbb{A}^n(\Omega)$. We call $\mathbf{y}$ a *generic point* of V and say $\mathbf{y}$ *determines V* if

$$V = \{\mathbf{z} \in \mathbb{A}^n(\Omega) : \mathbf{y} \xrightarrow{S} \mathbf{z}\}.$$

For example, if $t_1, \ldots, t_n \in \Omega$ are algebraically independent over D, then $(t_1, \ldots, t_n)$ determines $V = \mathbb{A}^n(\Omega)$. Recalling an earlier example, (t^2, t, t^3) is a generic point of the $\{X, Y, Z\}$-solutions of $U = \{X - Y^2, X^2 - YZ\}$.

Lemma 11
(i) If $x_1, \ldots, x_n \in \Omega$, then the set I of polynomials in $D[X_1, \ldots, X_n]$ that vanishes at $(x_1, \ldots, x_n)$ is a prime ideal.
(ii) If $I \in D[X_1, \ldots, X_n]$ is a prime ideal and $1 \notin I$ then I has a generic zero.

Proof. Part (i) is immediate. To see (ii), consider the ring $E = D[\mathbf{X}]/I$. This is a domain (§1, Exercise). As in the proof of the weak Nullstellensatz (§2), we can assume D is embedded in E via the canonical homomorphism $p \mapsto \overline{p}$ from $D[\mathbf{X}]$ to E. We may also assume E is embedded in Ω_D. For any $p(X_1, \ldots, X_n) \in D[\mathbf{X}]$, we saw that $p(\overline{X}_1, \ldots, \overline{X}_n) = 0$ iff $p(X_1, \ldots, X_n) \in I$. Hence $\mathbf{x} = (\overline{X}_1, \ldots, \overline{X}_n) \in \Omega_D^n$ is a zero of I. We claim that $\mathbf{x}$ is a generic zero. Let $\mathbf{a} = (a_1, \ldots, a_n) \in \Omega_D^n$. We must show that $\mathbf{a}$ is a zero of I iff

$$\mathbf{x} \longrightarrow \mathbf{a},$$

i.e., the function taking $\overline{X}_i$ to a_i $(i = 1, \ldots, n)$ is a specialization. Suppose $\mathbf{a}$ is a zero of I. If $p(X_1, \ldots, X_n) \in D[\mathbf{X}]$ and $p(\mathbf{x}) = 0$ then $p(X_1, \ldots, X_n) \in I$ (property of canonical homomorphism) and hence $p(\mathbf{a}) = 0$ (as $\mathbf{a}$ is a zero of I). This proves $\mathbf{x} \longrightarrow \mathbf{a}$. Conversely, assume $\mathbf{x} \longrightarrow \mathbf{a}$. If

[7] When we call σ a "solution" of U, it is understood that we are considering the universal case of solving the system $U = 0$. In contrast, as in §3, we shall call σ a "zero" of U when we consider the algebraic case of solving the system $U = 0$.

$p(X_1, \ldots, X_n) \in I$ then $p(\mathbf{x}) = 0$ (property of canonical homomorphism) and so $p(\mathbf{a}) = 0$ (definition of $\mathbf{x} \longrightarrow \mathbf{a}$). This means $\mathbf{a}$ is a zero of I. **Q.E.D.**

This lemma justifies the treatment of the zero set of prime ideals as "points" in an abstract space.

Exercises

Exercise 3.1: Let $F_1(X, Y, Z) = X - Y^2$, and $F_2(X, Y, Z) = XY - Z$.
i) Show that $\texttt{Ideal}(F_1, F_2)$ is prime.
ii) Let $F_0(X, Y, Z) = X^2 - YZ$. Show that $I_0 := \texttt{Ideal}(F_0, F_1)$ is not prime.
iii) Show that I_0 has two generic points, (t^2, t, t^3) and $(0, 0, t)$. (See above for a definition of generic points.)
iv) Consider the homogeneous polynomials $f_1(X, Y, Z, U) = UX - Y^2$ and $f_2(X, Y, Z, U) = XY - UZ$. Is $\texttt{Ideal}(f_1, f_2)$ prime? □

Exercise 3.2: If $I \subseteq D[\mathbf{X}]$ is the set of polynomials that vanish at two distinct points in Ω_D^n, then I is not prime. □

§4. Resultant Systems

We introduce the concept of a resultant system for a system of homogeneous polynomials with indeterminate coefficients (this will be precisely defined). Using the extended Nullstellensatz (§2), we show the existence of resultant systems.

Forms. Let $\mathbf{C}, \mathbf{X}$ be disjoint sets of indeterminates. We write $D[\mathbf{C}][\mathbf{X}]$ instead of $D[\mathbf{C} \cup \mathbf{X}]$ to signal our intention to view $F \in D[\mathbf{C} \cup \mathbf{X}]$ as a polynomial in $\mathbf{X}$ with coefficients in $D[\mathbf{C}]$. Hence the *degree*, $\deg(F)$, of F is understood to refer to the $\mathbf{X}$-degree. If $e = (e_1, \ldots, e_n) \in \mathbb{N}^n$, we shall write $\mathbf{X}^e$ for the power product $\prod_{i=1}^n X_i^{e_i}$. The set $\mathbf{C}$ will be used in a very special way as captured next:

Definition:
(i) A polynomial $F \in D[\mathbf{C}][\mathbf{X}]$ is an *indeterminate polynomial* if it is the sum of terms such as

$$c_0 \mathbf{X}^e \qquad (c_0 \in \mathbf{C}, e \in \mathbb{N}^n).$$

We call c_0 and $\mathbf{X}^e$ (respectively) the *coefficient* and *power product* of this term. *degree* of this term. *Moreover, this association between coefficients c_0 and power products $\mathbf{X}^e$ among the terms is a bijection.*
(ii) A *form* is an indeterminate polynomial that is homogeneous, *i.e.*, each term has the same degree.
(iii) If σ is a partial ground $\mathbf{C}$-specialization, and F is an indeterminate polynomial, then $F|_\sigma$ is called a *partially indeterminate polynomial*. So the coefficient of each power product in such a polynomial belongs to D or to $\mathbf{C}$. A *partial form* is similarly defined.
(iv) A *system of indeterminate polynomials* (of partial forms, etc) is a set of indeterminate polynomials (of partial forms, etc) such that distinct polynomials in the set have disjoint sets of indeterminate coefficients.

For instance, let $\mathbf{X} = \{X, Y\}, \mathbf{C} = \{c_0, c_1, c_2\}$. Then $c_0X^2 + c_0XY = c_0(X^2 + XY)$ and $c_1X + c_2X = (c_1 + c_2)X$ are not indeterminate polynomials as they fail the bijectivity requirement. Although c_0X^2Y and $c_0X + c_1Y$ are both indeterminate polynomials, together, they do not

constitute a "system" of indeterminate polynomials. The reason is, of course, because c_0 occurs in both polynomials. Note that the underlying domain D is irrelevant in this definition.

Two indeterminate polynomials are *equivalent* if they are identical after a suitable renaming of the coefficients in the polynomials. A form F of degree d is *generic* if every power product of degree d occurs in F; clearly F has $\binom{d+n-1}{n-1}$ terms. For any n, d, all generic forms of degree d on n variables are equivalent. Caveat: "forms" in the literature sometimes refers to homogeneous polynomials and sometimes refers to generic forms, neither agreeing with our definition.

More examples. The following are indeterminate polynomials:

$$c_1X + c_2XY, \quad c_0 + c_2X^2, \quad c_1X^2 + c_2Y^2.$$

The following are not indeterminate polynomials, but they are partially indeterminate:

$$X + 4X(= 5X),\ c_1X + XY,\ 4 + c_0X,\ 3X^2 - 5XY + c_2Y^2.$$

The following are not even partially indeterminate:

$$c_0X - c_1Y,\ 4X^3 + c_0X^3 = (4 + c_0)X^3,\ 1 + c_0,\ c_1^2XY,\ 8c_1Y.$$

Both c_0 and $c_1X^2 + c_0XY + c_2Y^2$ are generic forms while c_0X^2Y is a non-generic form.

We come to the main definition:

Definition: Let $\Sigma \subseteq D[\mathbf{C}][\mathbf{X}]$. A set $\Gamma \subseteq D[\mathbf{C}]$ is called a *resultant system* of Σ if for any ground specialization

$$\sigma : \mathbf{C} \longrightarrow D,$$

the vanishing of $\Gamma|_\sigma$ is a necessary and sufficient condition for the existence of a non-trivial $\mathbf{X}$-solution of the set $\Sigma|_\sigma$. In case Γ is a singleton set $\{R\}$, then we call the polynomial R a *resultant polynomial* of Σ.

The resultant system Γ can be the empty set $\emptyset$ in which case $\Gamma|_\sigma$ always vanishes, by definition. Two other special cases are $\Gamma = \{0\}$ and $\Gamma = \{1\}$: then Γ_σ always vanishes and never vanishes (respectively). We say Γ is **trivial**, and simply write "$\Gamma = 0$", if Γ always vanishes.

The classic example of a resultant polynomial is the determinant. More precisely, if Σ is a system of n generic linear forms in n variables, and $\det(\Sigma)$ is the determinant of a matrix whose ith row contains the coefficients of the ith form, then $\det(\Sigma)$ is a resultant polynomial for Σ. If σ is a ground specialization, it is conventional to call $\det(\Sigma)|_\sigma$ the "resultant of $\Sigma|_\sigma$".

We note a simple fact: a set Γ is a resultant system iff any basis (§1) for Γ is a resultant system. In particular, if $\mathbf{C}$ is finite (the usual case), and there exists a resultant system, then the Hilbert basis theorem (§1) assures us there is a finite resultant system.

The following is immediate:

Lemma 12 *Let $\Sigma \subseteq D[\mathbf{C}][\mathbf{X}]$ be any set polynomials. If $\Gamma \subseteq D[\mathbf{C}]$ is a resultant system for Σ and σ is a partial $\mathbf{C}$-specialization then $\Gamma|_\sigma$ is a resultant system for $\Sigma|_\sigma$.*

It will be shown that any system Σ of forms has a resultant system Γ. For any partial specialization $\Sigma|_\sigma$ of Σ, we can of course compute its resultant system by first computing Γ and then specializing

it to $\Gamma|_\sigma$. However, it may be more efficient to directly construct a resultant system for $\Sigma|_\sigma$. Recent results on sparse elimination are able to do this to some extent.

Construction of Resultant Systems. We now construct a resultant system for a system

$$\Sigma = \{A_1, \ldots, A_r\} \subseteq D[\mathbf{C}][X_1, \ldots, X_n]$$

of forms. To avoid trivial cases, let us assume that $n \geq 2$, $r \geq 1$ and $d \geq 1$ where $d = \max_{i=1}^{r} \deg(A_i)$.

Recall $\mathtt{PP} = \mathtt{PP}(X_1, \ldots, X_n)$ is the set of power products in the X_i's. There are

$$N_m := \binom{m+n-1}{n-1} \tag{8}$$

power products in $\mathtt{PP}$ of total degree m. Let $\mathtt{PP}^m$ be a column vector of length N_m that enumerates all these power products. When convenient, we also view $\mathtt{PP}^m$ as a set. A homogeneous polynomial $F \in D[\mathbf{C}][\mathbf{X}]$ of degree m can be identified with a row vector $\overline{F} \in D[\mathbf{C}]^{N_m}$ such that F is equal to the scalar product

$$\langle \overline{F}, \mathtt{PP}^m \rangle.$$

We are interested in the multiset Σ_m of homogeneous polynomials of degree m that are formed by[8] multiplying each $A_i \in \Sigma$ by various power products:

$$\Sigma_m := \{uA : \deg(uA) = m, A \in \Sigma, u \in \mathtt{PP}\}.$$

Let N_m^* denote the number (with multiplicity counted) of polynomials in Σ_m (we estimate N_m^* below). Let Q_m be the matrix that represents Σ_m: Q_m has N_m^* rows and N_m columns, with the rows corresponding to the vectors $\overline{F}$, $F \in \Sigma_m$. So

$$Q_m \cdot \mathtt{PP}^m$$

is just a column vector enumerating all the members of the multiset Σ_m. Now let

$$\Gamma_m \subseteq D[\mathbf{C}]$$

denote the set of $N_m \times N_m$ subdeterminants of Q_m (if $N_m^* < N_m$ then Γ_m is empty). Finally, let Γ be the union of the Γ_m for all $m \geq 1$.

Example: Let $\Sigma = \{A_1, A_2\}$ where $A_i = a_i X + b_i Y + c_i Z$. Let $\mathtt{PP}^2 = (X^2, XY, Y^2, YZ, Z^2, ZX)^T$.
Viewed as a set, $\Sigma_2 = \{XA_1, YA_1, ZA_1, XA_2, YA_2, ZA_2\}$. Then

$$Q_2 = \begin{array}{c} \begin{array}{cccccc} X^2 & XY & Y^2 & YZ & Z^2 & ZX \end{array} \\ \left[\begin{array}{cccccc} a_1 & b_1 & & & c_1 & \\ & a_1 & b_1 & c_1 & & \\ & & & b_1 & c_1 & a_1 \\ a_2 & b_2 & & & c_2 & \\ & a_2 & b_2 & c_2 & & \\ & & & b_2 & c_2 & a_2 \end{array}\right] \end{array} \begin{array}{l} \\ A_1X \\ A_1Y \\ A_1Z \\ A_2X \\ A_2Y \\ A_2Z \end{array}.$$

[8] Macaulay calls these the *elementary polynomials* with respect to Σ.

Note that we have labeled the rows and columns of the matrix in the obvious way. Here $N_2^* = N_2 = 6$ and so Γ_2 consists of just the determinant of Q_2. This determinant turns out to be 0. ■

Theorem 13 *If $\Sigma \subseteq D[\mathbf{C}][X_1, \ldots, X_n]$ is a system of forms of degrees at most d then Γ_m is a resultant system for Σ whenever*

$$m \geq 1 + n(N-1)$$

where $N = N(n,d)$ is the Nullstellensatz bound (§2).

Proof. Fix any ground specialization σ of the coefficients in Σ, and write $\widetilde{\Sigma}$ for $\Sigma|_\sigma$, $\widetilde{\Gamma}_m$ for $\Gamma_m|_\sigma$, etc. We must show that $\widetilde{\Sigma}$ has a non-trivial solution iff $\widetilde{\Gamma}_m = 0$. (Recall that Γ_m may be empty in which case $\widetilde{\Gamma}_m = 0$ is automatic.)

First suppose $\widetilde{\Sigma}$ has only trivial solutions. By the extended form of the Nullstellensatz (§2), as each X_i $(i = 1, \ldots, n)$ vanishes at all the zeros of $\widetilde{\Sigma}$, there exists $a \in D$ such that for all $p_i \in \mathtt{PP}^m$,

$$a \cdot p_i \in \mathtt{Ideal}(\widetilde{\Sigma}).$$

This follows from our choice of m. This means that there exists an $N_m \times N_m^*$ matrix U with entries in D such that

$$a \cdot I = U \cdot \widetilde{Q}_m, \tag{9}$$

where I is the identity matrix. [To see this, if $\mathtt{PP}^m = (p_1, p_2, \ldots)^T$ then on translating $a \cdot p_i \in \mathtt{Ideal}(\widetilde{\Sigma})$ to the matrix form

$$a \cdot p_i = (u_{i1}, u_{i2}, \ldots) \cdot \widetilde{Q}_m \cdot \mathtt{PP}^m,$$

we make take the ith row of U to be $(u_{i1}, u_{i2}, \ldots)$.] This means that the rank of $\widetilde{Q}_m$ is N_m and some determinant in $\widetilde{\Gamma}_m$ is non-zero. (In particular, Γ_m is non-trivial.)

Conversely, suppose $\widetilde{\Sigma}$ has a nontrivial solution $\tau : \mathbf{X} \to \Omega_D$. By way of contradiction, let us assume that $\widetilde{\Gamma}_m \neq 0$. Then there exists a non-zero matrix U and $a \in D$ such that equation (9) holds. [To see this, note that Γ_m is non-trivial implies that $N_m^* \geq N_m$. We may assume the first N_m rows of $\widetilde{Q}_m$ form a square matrix V with nonzero determinant a. Then we set $U = [\mathtt{adj}(V)|\mathbf{0}]$ where $\mathtt{adj}(V)$ denotes the adjoint of V and $\mathbf{0}$ is a matrix of zeros.] Then

$$a \cdot \mathtt{PP}^m = U \cdot \widetilde{Q}_m \cdot \mathtt{PP}^m. \tag{10}$$

The right hand side of this equation evaluates to a column vector: each element in this vector is a D-linear combination of the polynomials in $\widetilde{\Sigma}$. Hence if we specialize this equation using τ, then the right-hand side vanishes (as τ is a solution of $\widetilde{\Sigma}$). But the left-hand side does not vanish (as τ is a non-trivial solution), contradiction. **Q.E.D.**

Kapferer [100] already shows in 1929 that there is a resultant system whose polynomials have degree $2^{n-1}d^{2^{n-1}-1}$.

Estimates for N_m^*. For simplicity, assume $m \geq d$ in the following. First let us dispose of a simple case, when $r = 1$: then $N_m^* = N_{m-d} < N_m$ and Γ_m is empty. (So the above theorem shows that a single polynomial always has non-trivial solutions.) Henceforth, let $r \geq 2$. The number of polynomials in Σ_m is

$$N_m^* = \sum_{i=1}^{r} N_{m-d_i} \geq r N_{m-d}.$$

Note that

$$\frac{N_m}{N_{m-d}} = \frac{(m+n-1)(m+n-2)\cdots(m+1)}{(m-d+n-1)(m-d+n-2)\cdots(m-d+1)} < \left(1+\frac{d}{m-d}\right)^{n-1} < r$$

for $r \geq 2$ and m large enough. Hence $N_m^* \geq rN_{m-d} > N_m$ and so Γ_m is non-empty. However, as a previous example shows, this does not guarantee that Γ contains a non-zero element. Precisely when this happens will be clarified shortly.

Non-existence of Resultant Systems. We give an example of van der Waerden showing that if we drop the homogeneity requirement for the polynomials of $\Sigma \subseteq D[\mathbf{X}]$, then Σ may have no resultant systems. Assume D is an infinite field. Suppose $\Sigma = \{A, B\} \subseteq D[\mathbf{a}, \mathbf{b}][x, y]$ where $\mathbf{a} = (a_0, a_1, a_2), \mathbf{b} = (b_0, b_1, b_2)$ and

$$\begin{aligned} A &= a_0 + a_1 x + a_2 y, \\ B &= b_0 + b_1 x + b_2 y. \end{aligned}$$

Consider any specialization σ of $\mathbf{a}, \mathbf{b}$ such that

$$G_1 = a_0^2 + b_0^2, \quad G_2 = a_1 b_2 - a_2 b_1$$

either both vanish, or both do not vanish under σ. Then by linear algebra, $\Sigma|_\sigma$ has a nontrivial solution. It can be verified that converse also holds. We claim that Σ does not have a resultant system Γ. Assume to the contrary that $\Gamma \subseteq D[\mathbf{a}, \mathbf{b}]$ is a resultant system. Now there exists a ground specialization σ_0 of $\mathbf{a}, \mathbf{b}$ such that

$$G_1|_{\sigma_0} = 0, \qquad G_2|_{\sigma_0} \neq 0.$$

Thus $\Sigma|_{\sigma_0}$ has no non-trivial solutions and hence $\Gamma|_{\sigma_0} \neq 0$. But there are also infinitely many σ' such that σ' and σ_0 agree at all values except that $\sigma'(a_0) \neq \sigma_0(a_0)$, and

$$G_1|_{\sigma'} \neq 0, \qquad G_2|_{\sigma'} \neq 0. \tag{11}$$

Hence $\Gamma|_{\sigma'} = 0$. Since a_1, a_2, b_0, b_1, b_2 are held fixed when considering assignments such as σ', we can view elements of Γ as polynomials in the variable a_0 with constant coefficients. But such polynomials can have infinitely many solutions only if they are identically zero. But this contradicts the fact that $\Gamma|_{\sigma_0} \neq 0$. Thus Γ does not exist.

Resultant systems for a system of indeterminate (possibly non-homogeneous) polynomials exist if we modify our definition of a "resultant system" to exclude specializations that cause the leading coefficients to all vanish. See an exercise in the next section for the non-homogeneous version of the Sylvester resultant.

Exercises

Exercise 4.1:
(i) $n^m \geq N_m = \binom{m+n-1}{n-1}$, with strict inequality if $n \geq 2$ and $m \geq 2$.
(ii) $N_m^* > N_m$ if $m > d(1 + n/\ln r)$. □

Exercise 4.2: Construct the set Γ_2 for $\Sigma = \{A_0, A_1, A_2\}$ where $A_0 = aX^2 + bY^2 + cX^2$ and

$$A_i = a_i X + b_i Y + c_i Z, \qquad i = 1, 2.$$

□

§5. Sylvester Resultant Revisited

It is instructive to explicitly construct a resultant polynomial for a system of two forms in two variables. Unlike the general setting of the previous section, we do not use the Nullstellensatz. The reader will recognize that this is essentially the Sylvester matrix construction in §III.3.

Consider two generic forms $A, B \in D[\mathbf{a}, \mathbf{b}][X, Y]$,

$$\begin{aligned} A(X,Y) &= \sum_{i=0}^{m} a_i \, X^i Y^{m-i} \\ B(X,Y) &= \sum_{i=0}^{n} b_i \, X^i Y^{n-i}. \end{aligned}$$

The Sylvester matrix of A, B is the $(m+n)$-square matrix

$$S(A, B)$$

formed from the coefficients of A, B, in analogy to the non-homogeneous case in §III.3. The matrix in equation (13) below is $S(A, B)$. The *resultant* $\mathtt{res}(A, B)$ of A, B is the determinant of $S(A, B)$. Our main goal is to prove the following:

Theorem 14 *The polynomial* $\mathtt{res}(A, B)$ *is a resultant polynomial for* $\{A, B\}$.

Let σ be any ground specialization of $\mathbf{a}, \mathbf{b}$. Write $\widetilde{A}, \widetilde{B}$ and $\widetilde{R}$ for $A|_\sigma, B|_\sigma$ and $\mathtt{res}(A, B)|_\sigma$, respectively.

Lemma 15 *The following are equivalent:*
(a) $\widetilde{R} = 0$.
(b) There exist homogeneous polynomials $U, V \in D[X, Y]$ *of degrees* $n-1$ *and* $m-1$ *(respectively) such that*

$$U \cdot \widetilde{A} + V \cdot \widetilde{B} = 0. \tag{12}$$

Proof. The equation $0 = U \cdot A + V \cdot B$ can be written as a matrix equation involving the Sylvester matrix $S(A, B)$,

$$0 = (u_{n-1}, u_{n-2}, \ldots, u_0, v_{m-1}, v_{m-2}, \ldots, v_0) \cdot \begin{bmatrix} a_0 & a_1 & \cdots & a_m & & & \\ & a_0 & a_1 & \cdots & a_m & & \\ & & \ddots & & & \ddots & \\ & & & a_0 & \cdots & & a_m \\ b_0 & b_1 & \cdots & & b_n & & \\ & \ddots & & & & \ddots & \\ & & b_0 & \cdots & & & b_n \end{bmatrix} \tag{13}$$

where the u_i's and v_j's are the coefficients of U, V, respectively. Thus equation (12) holds for U, V iff the determinant of $S(A, B)|_\sigma$ vanishes, *i.e.*, iff $\widetilde{R} = 0$. **Q.E.D.**

Lemma 16 $\widetilde{R} = 0$ *implies* $\{\widetilde{A}, \widetilde{B}\}$ *has a non-trivial solution.*

Proof. If $\widetilde{A}$ is the zero polynomial, then clearly any non-trivial solution for $\widetilde{B}$ is a non-trivial solution for $\{\widetilde{A}, \widetilde{B}\}$. So assume that $\sigma(a_k) \neq 0$ for some $k = 0, \ldots, m$, and $\sigma(a_i) = 0$ for $i \geq k+1$. Then $\widetilde{A}$ has the factorization

$$\widetilde{A} = \sigma(a_k) Y^{m-k} \prod_{i=1}^{k} (X - \alpha_i Y), \qquad \alpha_i \in \overline{D}.$$

From equation (12), we get $U\widetilde{A} = -V\widetilde{B}$. Since $\overline{D}[X,Y]$ is a UFD and $\deg(V) = m-1$, at least one of the m linear factors, Y or $X - \alpha_i Y$, divides $\widetilde{B}$. Then either $(1,0)$ or $(\alpha_i, 1)$ would be a non-trivial zero $\{\widetilde{A}, \widetilde{B}\}$. **Q.E.D.**

Finally, our main result, theorem 14, is just a consequence of the following result:

Theorem 17 *The following are equivalent:*
(a) $\widetilde{R} \neq 0$,
(b) There exists a non-zero $c \in D$ *such that for* $r = 1, \ldots, m+n$,

$$c \cdot X^{r-1} Y^{m+n-r} \in \texttt{Ideal}(\widetilde{A}, \widetilde{B}). \tag{14}$$

(c) $\{\widetilde{A}, \widetilde{B}\}$ *has only the trivial solution.*

Proof. (a)⇒(b): We set up $m+n$ equations involving A and B:

$$\begin{bmatrix} a_0 & a_1 & \cdots & a_m & & & \\ & a_0 & a_1 & \cdots & a_m & & \\ & & \ddots & & & \ddots & \\ & & & a_0 & \cdots & & a_m \\ b_0 & b_1 & \cdots & & b_n & & \\ & \ddots & & & & \ddots & \\ & & b_0 & \cdots & & & b_n \end{bmatrix} \begin{bmatrix} Y^{m+n-1} \\ XY^{m+n-2} \\ \\ \vdots \\ \\ X^{m+n-2}Y \\ X^{m+n-1} \end{bmatrix} = \begin{bmatrix} Y^{n-1}A \\ XY^{n-2}A \\ \vdots \\ X^{n-2}Y \cdot A \\ X^{n-1}A \\ Y^{m-1}B \\ \vdots \\ X^{m-1}B \end{bmatrix} \tag{15}$$

Note that the Sylvester matrix is again involved. Fix $r = 1, \ldots, m+n$. Let $C_1, \ldots, C_{m+n}$ be the co-factors of the entries in the rth column. Multiply both sides of equation (15) on the left by the row $(C_1, \ldots, C_{m+n})$. Equating the left and right side of this multiplication,

$$X^{r-1} Y^{m+n-r} \texttt{res}(A,B) = U_r(X,Y) \cdot A(X,Y) + V_r(X,Y) \cdot B(X,Y). \tag{16}$$

where $U_r, V_r \in D[\mathbf{a}, \mathbf{b}][X,Y]$ are forms of degrees $n-1$ and $m-1$. Assertion (b) follows with the choice $c = \texttt{res}(A,B)|_\sigma$.
(b)⇒(c): Any non-trivial solution $(X,Y) \longrightarrow (x,y)$ of $\{A,B\}|_\sigma$ will make the right-hand side of equation (16) vanish, but the left-hand side will not vanish for either $r = 1$ or $r = m+n$, contradiction.
(c)⇒(a): This is the contrapositive of the previous lemma. **Q.E.D.**

Remarks. The element c in (14) is called an *inertial element* of $\{A,B\}$ (see next section). An important application of resultants is in the elimination of variables from a system of equations. Although the multivariate resultant of Macaulay may be more efficient in an asymptotic sense, it seems that for actual examples, eliminating one variable at a time using Sylvester resultants may be faster. An exercise below illustrates this application.

Exercises

Exercise 5.1: (Resultant system for non-homogeneous polynomials)
Let $A(X) = \sum_{i=0}^{m} a_i X^i, B(X) = \sum_{i=0}^{n} b_i X^i \in D[\mathbf{a}, \mathbf{b}][X]$. For all specializations σ of $\mathbf{a}, \mathbf{b}$, such that $\sigma(a_m) \neq 0$ or $\sigma(b_n) \neq 0$, the system $\{A, B\}|_\sigma$ has a solution iff $R|_\sigma = 0$, where R is the Sylvester resultant of A, B. □

Exercise 5.2: How many terms are there in $\mathtt{res}(A, B)$? □

Exercise 5.3: Elimination of variables.
(a) If F, G are polynomials in $X_1, X_2, \ldots, X_n$, we can compute the Sylvester resultant of F and G with respect to the variable X_1 (so the resultant is a polynomial in $X_2, \ldots, X_n$). Interpret this as a projection of the intersection of two hypersurfaces.
(b) Suppose $C_1 : F(X, Y) = 0$ and $C_2 : G(X, Y) = 0$ are two plane curves. We want to define the locus V of points $v = (v_x, v_y)$ such that the largest disk D_v whose interior avoids these curves simultaneously touches both C_1 and C_2. This locus V is called the *Voronoi diagram* defined by C_1, C_2. Let $a = (a_x, a_y)$ and $b = (b_x, b_y)$ denote points on C_1 and C_2 respectively where D_v touches C_1 and C_2. The relevant equations are

$$\begin{aligned}
F_0 : &\quad |v - a|^2 - |v - b|^2 = 0,\\
F_1 : &\quad F(a) = 0,\\
F_2 : &\quad G(b) = 0,\\
F_3 : &\quad (v_x - a_x)F_x(a) - (v_y - a_y)F_y(a) = 0,\\
F_4 : &\quad (v_x - b_x)G_x(b) - (v_y - b_y)G_y(b) = 0,
\end{aligned}$$

Suppose that F, G are polynomials of degrees m. Show by pairwise elimination that V is, in general, an algebraic curve of degree at most $4m^5$. HINT: eliminate variables successively in carefully chosen order.
(c) Compute the Voronoi diagram for the following pair of curves, $C_1 : X + Y - 9 = 0$ and $C_2 : 2X^2 + Y^2 - 3 = 0$.
(d) Why do you get different degree bounds with different order of elimination? What happens if your resultant vanishes? Can you prove that this is the best possible using "this method"?
□

§6. Inertial Ideal

In this section, fix

$$\Sigma = \{A_1, \ldots, A_r\}, \quad r \geq 1 \tag{17}$$

to be a system of partial forms in $D[\mathbf{C}][\mathbf{X}]$ where $\deg A_i = d_i \geq 1$. Partial forms are convenient here because they are preserved by partial specializations. We say that a partial form A of degree m is *regular in variable* X_j if the coefficient of its power X_j^m in A is an indeterminate. We say Σ is *regular* in X_j if each $A \in \Sigma$ is regular in X_j. Finally, we say Σ is *regular* if Σ is regular in each variable X_j.

We will assume that Σ is regular unless a less strict condition is explicitly mentioned (e.g., when we explicitly say "Σ is regular in X_1" then the more general assumption is dropped).

In the following, whenever Σ is regular in variable X_1, we assume that the indeterminate coefficients

$$\mathbf{C} = (c_1, c_2 \ldots)$$

are relabeled so that c_i is the coefficient of $X_1^{d_i}$ in A_i. We further write

$$A_i = A_i^* + c_i X_1^{d_i}, \tag{18}$$

where $A_i^* \in D[\mathbf{C}][\mathbf{X}]$.

Let us reëxamine our proof that any system Σ of forms has a resultant system Γ_m (§4). If

$$\Gamma_m \setminus \{0\} = \{a_1, a_2, a_3, \ldots\}$$

is non-empty, then as in the proof, there are $N_m \times N_m$ matrices $U_1, U_2, U_3, \ldots$ with entries in $D[\mathbf{C}]$, such that

$$a_i \cdot I = U_i \cdot Q_m.$$

Therefore

$$a_i \cdot \mathtt{PP}^m = U_i \cdot Q_m \cdot \mathtt{PP}^m.$$

The entries on the right-hand side are $D[\mathbf{C}]$-linear combinations of polynomials from the set Σ_m. In particular, this shows that

$$a_i X_1^m \in \mathtt{Ideal}(\Sigma).$$

In general, following Hurwitz, for any set of polynomials $\Sigma' \subseteq D[\mathbf{C}][\mathbf{X}]$, we call $R \in D[\mathbf{C}]$ an *inertial element* of Σ' if

$$R \cdot X_1^m \in \mathtt{Ideal}(\Sigma') \tag{19}$$

for some $m \geq 0$. Note that $\mathtt{Ideal}(\Sigma')$ here is generated in the ring $D[\mathbf{C}][\mathbf{X}]$, and also note the special role of X_1 in this definition. The following is immediate:

Lemma 18 *For any Σ:*
(i) The set of inertial elements is an ideal.
(ii) If R is an inertial element of Σ and σ is a partial specialization of $\mathbf{C}$ then $R|_\sigma$ is an inertial element of $\Sigma|_\sigma$.

Hence, we may speak of the *inertial ideal* of Σ. When Σ is a system of forms, we constructed in §4 a resultant system Γ consisting of inertial elements. We conclude from this lemma and lemma 12, §4:

Corollary 19 *Every system of partial forms has a resultant system consisting of inertial elements.*

We next give a characterization of inertial elements:

Lemma 20 *Let Σ be regular in variable X_1 and $R(\mathbf{C}) \in D[\mathbf{C}]$. Then R is an inertial element of Σ iff under the partial specialization*

$$\sigma^* : c_i \longmapsto \frac{-A_i^*}{X_1^{d_i}} \qquad (i = 1, \ldots, r) \tag{20}$$

the polynomial $R(\mathbf{C})$, regarded as an element of $D(\mathbf{C})(\mathbf{X})$, vanishes:

$$R|_{\sigma^*} = R\left(-\frac{A_1^*}{X_1^{d_1}}, -\frac{A_2^*}{X_1^{d_2}}, \ldots, -\frac{A_r^*}{X_1^{d_r}}, c_{r+1}, \ldots\right) = 0. \tag{21}$$

Proof. Suppose R is an inertial element. Then (18) and (20) show that $\texttt{Ideal}(\Sigma|_{\sigma^*}) = (0)$. So $R|_{\sigma^*} = 0$ by (19). Conversely assume that (21) holds. Then

$$\begin{aligned} R(c_1, \ldots, c_r, c_{r+1}, \ldots) &= R\left(\frac{A_1 - A_1^*}{X_1^{d_1}}, \ldots, \frac{A_r - A_r^*}{X_1^{d_r}}, c_{r+1}, \ldots\right) \\ &= R\left(-\frac{A_1^*}{X_1^{d_1}}, \ldots, -\frac{A_r^*}{X_1^{d_r}}, c_{r+1}, \ldots\right) \\ &\quad + \sum_{i=1}^{r} \frac{A_i}{X_1^{d_i}} \cdot B_i\left(-\frac{A_1^*}{X_1^{d_1}}, \ldots, -\frac{A_r^*}{X_1^{d_r}}, c_{r+1}, \ldots\right), \end{aligned}$$

for some $B_i(c_1, c_2, \ldots)$. The last expression for $R(c_1, \ldots, c_r, \ldots)$ represents an expansion of R into 2 parts: the first part containing terms that are not divisible by any $A_i/X_1^{d_i}$ ($i = 1, \ldots, r$ and treating $A_i/X_1^{d_i}$ as a new symbol), and the second part for the remaining terms. But the first part is just $R|_{\sigma^*}$, which is assumed to vanish. Multiplying the last equation by a suitable power of X_1, it follows that R satisfies (19). Hence R is an inertial element. **Q.E.D.**

The special role of X_1 in the definition of inertial elements can be replaced by another X_j under the following conditions:

Lemma 21 *Suppose Σ is regular in variable X_1 and in X_j (for some $1 < j \leq n$). Then $X_1^m R \in \texttt{Ideal}(\Sigma)$ (for some $m \geq 0$) iff $X_j^k R \in \texttt{Ideal}(\Sigma)$ (for some $k \geq 0$).*

Proof. Let σ^* be the specialization in equation (20). If $X_j^k R \in \texttt{Ideal}(\Sigma)$ then $R|_{\sigma^*} = 0$ (since $\Sigma|_{\sigma^*}$ vanishes). Then $X_1^m R \in \texttt{Ideal}(\Sigma)$ as in the proof of lemma 20. The reverse implication is similarly shown. **Q.E.D.**

Theorem 22 *If Σ is a system of partial forms that is regular in X_1 then the set of inertial elements of Σ is a prime ideal of $D[\mathbf{C}]$.*

Proof. To see that the inertial ideal is prime, consider $R, S \in D[\mathbf{C}]$. If $R \cdot S$ is an inertial element, then by lemma 20, $R \cdot S$ vanishes under the specialization (20). Hence either R or S vanishes under the specialization. Again by lemma 20, this means R or S is an inertial element. **Q.E.D.**

Theorem 23 *Let Σ be a system of partial forms that is regular. Then the inertial ideal I of Σ is a resultant system for Σ.*

Proof. Let $\Gamma \subseteq I$ be a resultant system for Σ. For any ground specialization σ of $\mathbf{C}$, if $I|_\sigma$ vanishes then $\Gamma|_\sigma$ vanishes. As Γ is a resultant system for Σ, we conclude that $\Sigma|_\sigma$ has a non-trivial solution. Conversely, suppose $\Sigma|_\sigma$ has a non-trivial $\mathbf{X}$-solution, say τ. Suppose $\tau(X_j) \neq 0$. Now for each $R \in I$, for some $m \geq 0$,

$$X_j^m R \in \texttt{Ideal}(\Sigma).$$

Since $\Sigma|_{\sigma,\tau} = 0$, and $X_j^m|_\tau \neq 0$, it follows that $R|_{\sigma,\tau} = 0$. But R does not depend on $\mathbf{X}$, so $R|_\sigma = 0$. This proves that $I|_\sigma = 0$. **Q.E.D.**

Lemma 24 *Let Σ be regular in X_1 and $c_1, \ldots, c_r$ be given by (18). Let $c \in \mathbf{C}$ be different from $c_1, \ldots, c_r$ and σ be a partial $\mathbf{C}$-specialization that maps c to some $\alpha \in D$ (leaving all other indeterminates in $\mathbf{C}$ fixed). Suppose the inertial ideal I of Σ is non-trivial.*
(a) The inertial ideal of $\Sigma|_\sigma$ is also non-trivial.
(b) Every non-zero element $R \in I$ contains a factor P that does not vanish under σ.

Proof. Part (a) follows immediately from part (b). To show (b), suppose $R \in I$ is a non-zero element such that $R|_\sigma = 0$. By the pseudo-division property (§III.2), $\beta^m R = (c-\alpha)P + Q$ for some $m \geq 1$, $\beta \in D[\mathbf{C}]$, and $P, Q \in D[\mathbf{C}]$ such that Q and β do not depend on c (actually $\beta = 1$ here). But $R|_\sigma = 0$ implies $Q|_\sigma = Q = 0$. Thus $(c-\alpha)P$ is an inertial element. Note that $c-\alpha$ is not an inertial element of Σ, by lemma 20. Hence P must be a non-zero inertial element of Σ. We can repeat this argument (choosing P in place of R) until eventually $R|_\sigma \neq 0$. **Q.E.D.**

Lemma 25 *If R is a non-zero inertial element of Σ then R depends on at least n of the coefficients among $c_1, \ldots, c_r$. In particular, $r \geq n$.*

Proof. By way of contradiction, suppose that R depends on $m < n$ elements among $c_1, \ldots, c_r$. Without loss of generality, assume R depends on $c_1, \ldots, c_m$ only. Consider the partial specialization

$$(c_{n+1}, c_{n+2}, \ldots) \xrightarrow{\sigma} (\alpha_{n+1}, \alpha_{n+2}, \ldots), \qquad (\alpha_j \in D). \tag{22}$$

Repeated application of the previous lemma shows that $R|_\sigma \neq 0$. Write $\widetilde{R}$ for $R|_\sigma$. So $\widetilde{R} = \widetilde{R}(c_1, \ldots, c_m)$ is an inertial element of $\Sigma|_\sigma$. Since Σ is regular, we may choose the α_j's in (22) so that $A_i^*|_\sigma = -X_{i+1}^{d_i}$ for $i = 1, \ldots, m$ (using the fact that $i \leq m < n$). Let σ^* be the $\{c_1, \ldots, c_n\}$-specialization for $\Sigma|_\sigma$, defined as in (20). It follows that

$$\sigma^*(\sigma(c_i)) = \sigma^*(c_i) = (X_{i+1}/X_1)^{d_i}.$$

Since $\widetilde{R}|_{\sigma^*}$ vanishes, we have

$$\widetilde{R}|_{\sigma^*} = \widetilde{R}\left(\left(\frac{X_2}{X_1}\right)^{d_1}, \left(\frac{X_3}{X_1}\right)^{d_2}, \ldots, \left(\frac{X_{m+1}}{X_1}\right)^{d_m}\right) = 0.$$

But renaming each $(X_{i+1}/X_1)^{d_i}$ as a new indeterminate Y_i, we conclude that $\widetilde{R}(Y_1, \ldots, Y_m) = 0$. But a non-zero polynomial cannot vanish by renaming, contradiction. **Q.E.D.**

This yields:

Theorem 26
(i) If the number r of polynomials in Σ is less than the number n of variables then the inertial ideal of Σ is trivial.
(ii) Any system of homogeneous equations in $D[\mathbf{X}]$ with fewer equations than unknowns has a non-trivial zero.

Proof. (i) If Σ has a non-zero inertial element then the previous lemma shows that $r \geq n$. So $r < n$ implies such elements do not exist.
(ii) Let $\Sigma' \subseteq D[\mathbf{X}]$ be the homogeneous system under consideration. If $\Sigma' = \Sigma|_\sigma$ where Σ is the system in part (i), then any resultant system Γ for Σ specializes to a resultant system $\Gamma|_\sigma$ for Σ'. But

such a Γ is necessarily trivial (we may always assume Γ is comprised of inertial elements). **Q.E.D.**

An ideal $I \in D[\mathbf{C}]$ is *pseudo-principal* if there exist $R \in I$ and $\alpha \in D$ such that for all $S \in I$, R divides $\alpha \cdot S$. Of course, if $\alpha = 1$, then I is principal in the usual sense.

Theorem 27 *If Σ has as many polynomials as variables, then its inertial ideal I is a non-trivial pseudo-principal ideal. In particular, Σ has a resultant polynomial.*

Proof. First let us show that $I \neq (0)$. If $I = (0)$ then under every $\mathbf{C}$-specialization σ, the system $\Sigma|_\sigma$ has a non-trivial solution. In particular, since $n = r$, we may specialize Σ to

$$\Sigma|_\sigma = \{X_i^{d_i} : i = 1, \ldots, n\}.$$

But it is patently false that $\Sigma|_\sigma$ has a non-trivial solution. This shows $I \neq (0)$. By lemma 25, any non-zero element of I must depend on $c_1, \ldots, c_n$. We now operate in the UFD $D' = Q_D[\mathbf{C}]$ where Q_D is the quotient field of D. Choose $R_0 \in \mathtt{Ideal}_{D'}(I)$ to be an irreducible element whose c_n-degree, say $d \geq 1$, is minimum. We claim that

$$\mathtt{Ideal}_{D'}(R_0) = \mathtt{Ideal}_{D'}(I).$$

The forward inclusion is immediate. In the reverse direction, pick any non-zero $S \in \mathtt{Ideal}_{D'}(I)$ and let e be the c_n-degree of S. Then $e \geq d$ and by the pseudo-division property, $\gamma^m S = AR_0 + B$ for some $m \geq 1$ and $\gamma, A, B \in Q_D[\mathbf{C}]$ where the c_n-degree of B is less then d and γ does not depend on c_n. But $B = \gamma^m S - AR_0 \in \mathtt{Ideal}_{D'}(I)$. So $B = 0$ since the c_n-degree of B is less than d; this in turn means R_0 divides $\gamma^m S$. As R_0 is irreducible in D' which is a UFD, it must divide γ or S. Since the c_n-degree of γ is 0, R_0 cannot divide γ and so it divides S. This proves our claim. Note that for some $\alpha \in D$, $\alpha \cdot R_0 \in I$. This means that for all $S \in I$, αR_0 divides αS, proving that I is pseudo-principal.

This immediately implies that $\alpha \cdot R_0$ can serve as a resultant polynomial for Σ (since the inertial ideal is a resultant system for Σ). **Q.E.D.**

Definition of Macaulay Resultant. We call $\alpha \cdot R_0$ in the preceding proof the *Macaulay resultant* of Σ, and denote it by

$$\mathtt{res}(\Sigma).$$

Although $\alpha \cdot R_0$ in the proof is only defined up to associates in D, we will next show that α can be taken to be 1 for a regular system of n forms. Thus $\mathtt{res}(\Sigma)$ is unique with this convention.

Suppose now $\Sigma = \{A_1, \ldots, A_n\}$ is a system of partial forms, not necessarily regular. Its Macaulay resultant can be defined naturally to be an appropriate specialization: namely, if $\Sigma' = \{A'_1, \ldots, A'_n\}$ is any regular system such that $\deg A'_i = \deg A_i$ and there is a partial specialization σ such that $\Sigma'|_\sigma = \Sigma$ then define the Macaulay resultant $\mathtt{res}(\Sigma)$ of Σ to be $\mathtt{res}(\Sigma')|_\sigma$. We leave it as an exercise to show that this does not depend on the choice of Σ'.

Exercises

Exercise 6.1: (Bloemer) Let Γ be the inertial ideal of Σ. Show by an example that the inertial ideal of $\Sigma|_\sigma$ may be a proper superset of $\Gamma|_\sigma$. HINT: let Γ be trivial. □

§7. The Macaulay Resultant

We derive properties of the Macaulay resultant

$$R_0 := \mathtt{res}(\Sigma)$$

where

$$\Sigma = \{A_1, \ldots, A_n\} \subseteq D[\mathbf{C}][X_1, \ldots, X_n]$$

is a regular system of $n \geq 2$ forms.

We can already infer from theorem 22 that R_0 is irreducible. For further properties, we give an explicit construction which van der Waerden [208] attributed to Hurwitz. Let

$$d_i := \deg(A_i) \geq 1, \qquad (i = 1, \ldots, n).$$

Unlike the previous section, we now let c_i denote the coefficient of $X_i^{d_i}$ in A_i, and call c_i the *main coefficient* of A_i. In §4, we constructed the matrix Q_m of shape $N_m^* \times N_m$. We revisit this construction in more detail. As usual, $\mathbf{X}^e$ denotes the power product $\prod_{i=1}^n X_i^{e_i}$ where $e = (e_1, \ldots, e_n) \in \mathbb{N}^n$. We first partition the set $\mathtt{PP}^d$ of power products of degree d into the following sets:

$$\begin{aligned}
\mathtt{PP}_1^d &:= \{\mathbf{X}^e \in \mathtt{PP}^d \,:\, e_1 \geq d_1\},\\
\mathtt{PP}_2^d &:= \{\mathbf{X}^e \in \mathtt{PP}^d \,:\, e_1 < d_1 \text{ and } e_2 \geq d_2\},\\
&\vdots\\
\mathtt{PP}_i^d &:= \{\mathbf{X}^e \in \mathtt{PP}^d \,:\, e_1 < d_1, e_2 < d_2, \ldots, e_{i-1} < d_{i-1} \text{ and } e_i \geq d_i\},\\
&\vdots\\
\mathtt{PP}_n^d &:= \{\mathbf{X}^e \in \mathtt{PP}^d \,:\, e_1 < d_1, e_2 < d_2, \ldots, e_{n-1} < d_{n-1} \text{ and } e_n \geq d_n\}.
\end{aligned}$$

Finally, let

$$\mathtt{PP}_{n+1}^d := \{\mathbf{X}^e \in \mathtt{PP}^d \,:\, e_i < d_i \text{ for } i = 1, \ldots, n\}.$$

These $n+1$ sets constitute a partition of $\mathtt{PP}^d$ because not only is $\mathtt{PP}_i^d$ disjoint from $\mathtt{PP}_j^d$ for $i \neq j$, but every $\mathbf{X}^e \in \mathtt{PP}^d$ must fall into some $\mathtt{PP}_i^d$.

We say $\mathbf{X}^e$ is *reduced in* X_i if $e_i < d_i$. We say $\mathbf{X}^e$ is *reduced* if it is reduced in every X_i. Finally, if $\mathbf{X}^e$ is reduced in all but one of the variables, we say it is *almost-reduced*.

E.g., $\mathtt{PP}_i^d$ comprises those power products (in $\mathtt{PP}^d$) that are reduced in $X_1, \ldots, X_{i-1}$ but not reduced in X^i. $\mathtt{PP}_{n+1}^d$ comprises the reduced power products. Every element in $\mathtt{PP}_n^d$ is almost-reduced. Note that

$$\mathtt{PP}_i^d \cup \mathtt{PP}_{i+1}^d \cup \cdots \cup \mathtt{PP}_{n+1}^d$$

comprises the power products that are reduced in $X_1, \ldots, X_{i-1}$.

It is easy to verify that $\mathtt{PP}_{n+1}^d$ is empty iff $d \geq 1 + \sum_{i=1}^n (d_i - 1)$. Henceforth, assume

$$d := 1 + \sum_{i=1}^n (d_i - 1). \tag{23}$$

Further let

$$\widehat{d_i} := \left(\prod_{j=1}^n d_j \right) / d_i. \tag{24}$$

Then for $i = 1, \ldots, n$, it is easy (Exercise) to see that

$$|\mathtt{PP}_i^d| \geq \widehat{d}_i, \quad \text{with equality when } i = n. \tag{25}$$

Next, for each $i = 1, \ldots, n$, let

$$S_i := \frac{\mathtt{PP}_i^d}{X_i^{d_i}} \cdot A_i = \{\mathbf{X}^e \cdot A_i \ : \ \mathbf{X}^e \cdot X_i^{d_i} \in \mathtt{PP}_i^d\}.$$

So the set

$$S := S_1 \cup \cdots \cup S_n$$

has $N_d = \binom{d+n-1}{n-1}$ polynomials. Since the polynomials in S have degree d, each polynomial corresponds to a row in the matrix Σ_d. The corresponding $N_d \times N_d$ submatrix M of Σ_d made up of these rows will be called a *Macaulay matrix* of Σ. Of course, the determinant of M is an element of Γ_d (§5) and so is an inertial element of Σ.

Example: Let

$$\Sigma = \{A_1, A_2, A_3\} \tag{26}$$

where $A_3 = a_3X^2 + b_3Y^2 + c_3Z^2$ and $A_i = a_iX + b_iY + c_iZ$ for $i = 1, 2$. Then $(d_1, d_2, d_3) = (1, 1, 2)$. So $d = 3$ and

$$\mathtt{PP}_1^2 = \{X^2, XY, XZ\}, \quad \mathtt{PP}_2^2 = \{Y^2, YZ\}, \quad \mathtt{PP}_3^2 = \{Z^2\},$$
$$S_1 = \{XA_1, YA_1, ZA_1\}, \quad S_2 = \{YA_2, ZA_2\}, \quad S_3 = \{A_3\}.$$

Finally, M is given by

$$M = \left[\begin{array}{ccc|cc|c}
a_1 & b_1 & c_1 & & & \\
 & a_1 & & b_1 & c_1 & \\
 & & a_1 & & b_1 & c_1 \\
\hline
 & a_2 & & b_2 & c_2 & \\
 & & a_2 & & b_2 & c_2 \\
\hline
a_3 & & & b_3 & & c_3
\end{array}\right] \cdot \begin{array}{c} X^2 \\ XY \\ XZ \\ Y^2 \\ YZ \\ Z^2 \end{array} \quad \begin{array}{l} \\ XA_1 \\ YA_1 \\ ZA_1 \\ YA_2 \\ ZA_2 \\ A_3 \end{array}$$

Notice that the main coefficients are a_1, b_2, c_3 and these occur along the main diagonal of M. We have labeled the columns of M by elements of $\mathtt{PP}^d$ and the rows by elements of S. The matrix is also partitioned into blocks. All these illustrate a general convention. ■

Labeling convention for Macaulay Matrix. It is expedient for the subsequent exposition to construct M by first listing the rows corresponding to S_1, followed by the rows corresponding to S_2, and so on. The columns of M are labeled by elements of $\mathtt{PP}^d$ according to the following scheme: for each $F \in S_j$, call the power product $\mathbf{X}^e$ in F whose coefficient is c_j the *main power product*. Observe that distinct polynomials in S have distinct main power products. Thus the c_j's occur in distinct columns of M, and we may arrange so $c_1, \ldots, c_n$ appear along the main diagonal of M. We therefore label a column by the unique main power product $\mathbf{X}^e$ that appears in that column (and in a main diagonal position). Notice that the main power products of elements in S_j comprise the

set $\mathtt{PP}_j^d$. This means that the columns of M are labeled first by the elements of $\mathtt{PP}_1^d$, followed by elements of $\mathtt{PP}_2^d$, etc. Thus M can be given the block structure:

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & & M_{2n} \\ \vdots & & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nn} \end{bmatrix}, \tag{27}$$

where M_{ij} is $|\mathtt{PP}_i^d|$ by $|\mathtt{PP}_j^d|$ and contains coefficients from A_i only. This is schematically shown in figure 1.

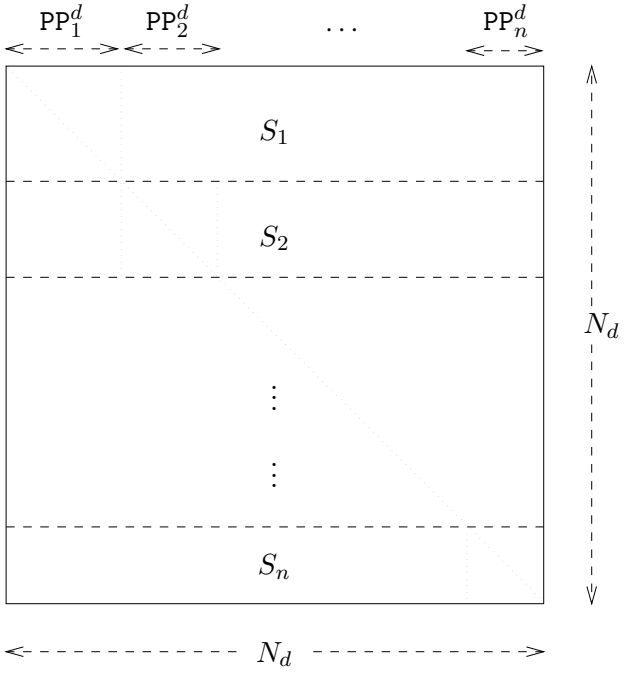


Figure 1: The Macaulay matrix M for Σ.

Lemma 28 *For each $j = 1, \ldots, n$, the determinant* $\det(M)$ *is homogeneous of degree h_j in the coefficients of $A_j \in \Sigma$ where $h_j = |\mathtt{PP}_j^d|$. In fact,* $\det(M)$ *contains the monic monomial*

$$c_1^{h_1} c_2^{h_2} \cdots c_n^{h_n} \tag{28}$$

where c_j is the main coefficient of A_j.

Proof. It is clear that $\det(M)$ is homogeneous in the coefficients of each A_j. If $\det(M) \neq 0$ then it has the claimed degree h_j in the coefficients of A_j since there are exactly h_j rows in M containing only coefficients of A_j. So it remains to show that $\det(M) \neq 0$. This follows if $\det(M)$ contains the monomial (28). Let σ be the specialization yielding $A_j|_\sigma = c_j X_j^{d_j}$ for each j. The non-zero entries of $M|_\sigma$ are therefore confined to its main diagonal (by construction of M), and hence $\det(M|_\sigma)$ is given by (28). **Q.E.D.**

Our preceding construction of the sets $\mathtt{PP}_i^d$ is arbitrary in two ways:

(a) We associated the variable X_i with A_i. (Macaulay suggests that we regard A_i as a polynomial in X_i.)

(b) The definition of $\texttt{PP}_i^d$ (and of S_i) depended on a fixed enumeration of the variables, from X_1 to X_n.

The assertion (25) depends on these choices only to the extent that *we enumerated the variable (viz., X_n) associated with A_n last*. To indicate this special role of A_n in our construction of S and M, let us introduce the notations $S^{(n)}$ and $M^{(n)}$ to refer to them. Let us now vary the construction of S and M by preserving (a) but varying (b) so that the variable X_i associated with A_i is enumerated last. Call the corresponding set $S^{(i)}$ and matrix $M^{(i)}$. Set

$$G^{(i)} := \det(M^{(i)}).$$

We have the following analogue of (25) and lemma 28:

Corollary 29 *Fix $i = 1, \ldots, n$. Then $G^{(i)}$ is homogeneous in the coefficients of each $A_j \in \Sigma$. If $G^{(i)}$ is of degree $d_j^{(i)}$ in the coefficients of A_j then*

$$d_j^{(i)} \geq \widehat{d}_j, \qquad \text{with equality when } j = i.$$

Moreover, $G^{(i)}$ contains the monic monomial

$$\pm c_1^{d_1^{(i)}} c_2^{d_2^{(i)}} \cdots c_n^{d_n^{(i)}}. \tag{29}$$

Since $Q_D[\mathbf{C}]$ is a UFD, we may define the $\texttt{GCD}$ of $G^{(1)}, \ldots, G^{(n)}$ in $Q_D[\mathbf{C}]$. Our goal is to show that R_0 is (up to similarity, §III.1) equal to

$$G := \texttt{GCD}(G^{(1)}, \ldots, G^{(n)}). \tag{30}$$

Let B_1 and B_2 be forms in the variables $\mathbf{X}$ but involving indeterminate coefficients $\mathbf{C}'$ and such that $\deg(B_1 B_2) = \deg(A_1)$. We assume $\mathbf{C}'$ is disjoint from $\mathbf{C}$. We need an unusual specialization:

Lemma 30 *Suppose σ is a partial specialization such that*

$$A_1|_\sigma = B_1 B_2$$

and which leaves the coefficients of the other A_j's untouched. Then some associate of $R_0|_\sigma$ is divisible by $R_1 R_2$ where $R_j := \texttt{res}(B_j, A_2, A_3, \ldots, A_n)$ $(j = 1, 2)$.

Proof. Note that $X_1^e R_0 \in \texttt{Ideal}(A_1, A_2, \ldots, A_n)$ for some $e \geq 0$. Hence $X_1^e R_0|_\sigma \in \texttt{Ideal}(B_1 B_2, A_2, \ldots, A_n)$ where this ideal is generated in the ring $D[\mathbf{C}, \mathbf{C}'][\mathbf{X}]$. Hence $X_1^e R_0|_\sigma \in \texttt{Ideal}(B_j, A_2, \ldots, A_n)$, $j = 1, 2$. This means $R_0|_\sigma$ is divisible by R_j in the ring $Q_D[\mathbf{C}, \mathbf{C}']$. But R_1 and R_2 are each irreducible, and so $R_0|_\sigma$ is in fact divisible by the product $R_1 R_2$ in $Q_D[\mathbf{C}, \mathbf{C}']$. **Q.E.D.**

Theorem 31
(a) The Macaulay resultant R_0 is (up to similarity) equal to G.
(b) R_0 is homogeneous of degree $\widehat{d}_i$ in the coefficients of A_i.
(c) R_0 contains a monomial of the form

$$\alpha \cdot c_1^{\widehat{d}_1} c_2^{\widehat{d}_2} \cdots c_n^{\widehat{d}_n},$$

where α is a unit in D. W.l.o.g., assume $\alpha = 1$.

Proof. We first assume that Σ consists of generic polynomials.
(a) Clearly R_0 divides G, since R_0 divides each $G^{(i)}$ in $Q_D[\mathbf{C}]$. To show (a), it suffices (in view of corollary 29) to show that the degree of R_0 in the coefficients of A_i is equal to $\widehat{d}_i$. By symmetry, it suffices to prove that the degree of R_0 in the coefficients of A_n is equal to $\widehat{d}_n$. To see this, choose a specialization σ such that each

$$A_i|_\sigma, \qquad i = 1, \ldots, n-1$$

factors into d_i linear forms that are regular in all variables. We assume these linear forms involve brand new indeterminate coefficients. For instance, we may choose σ such that

$$A_1|_\sigma = L_1 L_2 \cdots L_{d_1}$$

where $L_i = \sum_{j=1}^n a_{ij} X_j$ and the a_{ij}'s are new indeterminates. This is possible since A_1 is generic. If B_i $(i = 1, \ldots, n-1)$ is a linear factor of $A_i|_\sigma$, then the Macaulay resultant $R' = \texttt{res}(B_1, B_2, \ldots, B_{n-1}, A_n)$ divides some associate of $R_0|_\sigma$, by the obvious extension to the previous lemma. But there are $\widehat{d}_n$ such resultants R', and they all divide some associate of $R_0|_\sigma$. Since each R' is irreducible, their product must divide some associate of $R_0|_\sigma$. By corollary 29, we know that each R' is of degree at most 1 in the coefficients of A_n. By lemma 25, R' must depend on the coefficients of A_n, and so it has degree exactly 1. Hence their product is of degree exactly $\widehat{d}_n$ in the coefficients of A_n. So $R_0|_\sigma$ is of degree (at least, and hence equal to) $\widehat{d}_n$ in the coefficients of A_n. Since σ does not affect the coefficients of A_n, we conclude that R_0 also has degree $\widehat{d}_n$ in the coefficients of A_n.
(b) With respect to the coefficients of A_i, $G^{(i)}$ is homogeneous and R_0 divides $G^{(i)}$ implies that R_0 is also homogeneous; the argument in part (a) shows that R_0 has degree $\widehat{d}_i$.
(c) Let σ be the specialization such that $A_i|_\sigma = c_i X^{d_i}$ for $i = 1, \ldots, n$. Then $R_0|_\sigma$ divides $G^{(i)}|_\sigma = c_1^{d_1^{(i)}} \cdots c_n^{d_n^{(i)}}$ implies that $R_0|_\sigma = \alpha \cdot c_1^{e_1} \cdots c_n^{e_n}$ for some e_i's and α is a unit. Now $\sigma(c_i) = c_i$ for all $i = 1, \ldots, n$ implies that $R_0|_\sigma$ is a monomial in R_0. Since R_0 is homogeneous of degree $\widehat{d}_i$ in the coefficients of A_i (from part (b)), this means $e_i = \widehat{d}_i$ for each $i = 1, \ldots, n$.

Finally, we remove the assumption that Σ has only generic polynomials: since G and R_0 are obtained from the generic case by a common specialization, any equality would be preserved. In particular, (a) holds generally. Next (b) holds since the degree $\widehat{d}_i$ in the coefficients of A_i is preserved under any specialization in which Σ remains a set of forms (such specializations sends each c_i to 0 or c_i). Finally (c) is preserved by specializations in which Σ remains regular. **Q.E.D.**

Macaulay's Theorem. At this point, we have a method for computing the Macaulay resultant, using the GCD formula of (30). Actually, it turns out that we can avoid GCD computation and obtain R_0 as a sequence of $2n$ divisions involving minors of the $M^{(i)}$'s (§9). But a better method follows from an important result of Macaulay [122]. Let M be a Macaulay matrix of Σ (for instance, take $M = M^{(n)}$) and let L be the principal submatrix of M obtained by deleting all the columns that are labeled by almost-reduced power products, and also deleting the rows corresponding to the deleted columns. Note that an almost-reduced power product can be characterized as one that labels some column in the last block of the matrix $M^{(i)}$, for some $i = 1, \ldots, n$. Hence there are exactly $\widehat{d}$ almost-reduced power products in $\texttt{PP}^d$ where (see equation (24))

$$\widehat{d} := \sum_{i=1}^n \widehat{d}_i. \tag{31}$$

We cite without proof:

Theorem 32 (Macaulay)

$$R_0 = \det(M)/\det(L). \tag{32}$$

For example, if $M = M^{(3)}$ is the Macaulay matrix for the system (26) above, then the only power product in $\mathtt{PP}^2$ that is not almost-reduced is XY. In this case, L is a 1×1 matrix containing the entry a_1. Note that it is hardly obvious from an examination of the matrix M that such a factor a_1 exists in $\det(M)$.

Remark. The basic properties of the Macaulay resultant goes back to Mertens [130]. It was also treated by Hurwitz [87] and Perron [156]. The only proof in print of Macaulay's theorem seems to be his original paper [122]. This paper contains the construction of the matrix M above. These ideas apparently trace back to Bézout (1779). See also [123]. For improved constructions when all the polynomials have the same degree, see [124].

Exercises

Exercise 7.1:
(i) Verify (25).
(ii) Give an exact expression for $|\mathtt{PP}_i^d|$. □

Exercise 7.2: Construct the partition of $\mathtt{PP}^d = \uplus_{i=1}^n \mathtt{PP}_i^d$ in the following cases:
(i) Let $n = 3$, $(d_1, d_2, d_3) = (2, 4, 2)$. So $d = 6$.
(ii) Let $(d_1, \ldots, d_n) = (1, 1, \ldots, 1, m)$. So $d = m$. □

Exercise 7.3: Consider the Macaulay resultant for the following:
(i) A system of n generic linear forms in n variables. Verify that this is the standard determinant.
(ii) A system of 2 generic forms in 2 variables. Verify that this is the Sylvester resultant.
(iii) A system of $n-1$ generic linear forms and 1 arbitrary generic form, in n variables. □

Exercise 7.4: Consider the system in equation (26).
(i) Compute its Macaulay resultant using formula (32). Show intermediate results.
HINT: the Macaulay resultant is

$$R_0 = a_2^2b_1^2c_3 - 2a_1a_2b_1b_2c_3 + a_1^2b_2^2c_3 + a_2^2b_3c_1^2 + a_3b_2^2c_1^2 - 2a_1a_2b_3c_1c_2 - 2a_3b_1b_2c_1c_2 + a_1^2b_3c_2^2 + a_3b_1^2c_2^2.$$

(ii) What are the elements of Γ_2 (see §4, exercise 4.2) expressed in terms of R_0? □

Exercise 7.5: Consider the system $\Sigma = \{A_1, A_2, A_3\}$ where $A_1(X, Y, Z)$ is generic of degree 1 and $A_2(X, Y, Z)$, $A_3(X, Y, Z)$ are generic of degree 2 (*i.e.*, quadrics). Write down a Macaulay matrix M for Σ and its associated matrix L. Compute its Macaulay resultant. HINT: do not expect to explicitly write down the final resultant by hand — use some computer algebra system. □

Exercise 7.6: The Macaulay matrix M of Σ has shape $N_d \times N_d$ where d depends only on the degrees $d_1, \ldots, d_n$. We want to exploit the sparse structure of the actual forms in Σ. Describe

an infinite family of systems Σ such that we can set up similar matrices M whose sizes are smaller than that of the corresponding Macaulay matrix. E.g., suppose each $A_i \in \Sigma$ is a sum of powers: $A_i = \sum_{j=1}^{n} c_{ij} X_j^{d_i}$. □

p Last Update *Date* : 1999/09/0622 : 26 : 47

§8. U-Resultant

Let $\Sigma = \{A_1, \ldots, A_r\}$ be a system of homogeneous polynomials in $D[X_1, \ldots, X_n]$ with a finite number $s \geq 1$ of non-trivial zeros in the algebraic closure $\overline{D}$, say,

$$\xi^{(j)} = \left(\xi_1^{(j)}, \ldots, \xi_n^{(j)}\right), \quad (j = 1, \ldots, s), \tag{33}$$

such that every other zero of Σ is proportional to one of these zeros. Such a set $\{\xi^{(1)}, \ldots, \xi^{(s)}\}$ is called a representative set *of solutions for Σ. We consider the problem of computing a representative set. A useful tool for this purpose is the U-resultant.*

The above "finiteness condition" is clearly the best that one can impose on the zero set of Σ, since if $\xi = (\xi_1, \ldots, \xi_n)$ is any zero, so is any multiple $\alpha\xi = (\alpha\xi_1, \ldots, \alpha\xi_n)$, $\alpha \in D$. The set

$$D \cdot \xi = \{\alpha\xi : \alpha \in D\}$$

is called a *solution line*. So our assumption on the zero set of Σ amounts to saying that it has finitely many solution lines. Such a system Σ is also called a *projective zero-dimensional system*. Two solutions are said to be *proportional* if they belong to the same solution line.

We first introduce the polynomial

$$A_0 = U_1 X_1 + \cdots + U_n X_n$$

in the new indeterminates $\mathbf{U} = (U_1, \ldots, U_n)$ and in $\mathbf{X}$. Let $\Gamma \subseteq D[\mathbf{U}]$ be a resultant system of Σ', where

$$\Sigma' := \{A_0, A_1, \ldots, A_r\}.$$

We may assume that Γ is an ideal. Note that Γ is non-trivial (otherwise, for any j, we get that $\sum_{i=1}^{n} \alpha_i \xi_i^{(j)} = 0$ for any specialization $U_i \to \alpha_i$, $i = 1, \ldots, n$ which is clearly false). Consider

$$G(\mathbf{U}) := \prod_{j=1}^{s} g_j(\mathbf{U}), \qquad \text{where } \; g_j(\mathbf{U}) := U_1 \xi_1^{(j)} + \cdots + U_n \xi_n^{(j)}. \tag{34}$$

We shall say that $g_j(\mathbf{U})$ "encodes" the solution line corresponding to $\xi^{(j)}$.

Lemma 33 *For any ground $\mathbf{U}$-specialization σ, $\Gamma|_\sigma = 0$ iff $G(\mathbf{U})|_\sigma = 0$.*

Proof. ($\Rightarrow$) If $\Gamma|_\sigma$ vanishes, then $\Sigma'|_\sigma$ has an $\mathbf{X}$-solution. But since any $\mathbf{X}$-solution is proportional to some solution $\xi^{(j)}$ in (33), this means for some j,

$$A_0(\xi^{(j)}, \mathbf{U})|_\sigma = g_j(\mathbf{U})|_\sigma = 0,$$

which implies $G(\mathbf{U})|_\sigma = 0$.
($\Leftarrow$) If $G|_\sigma$ vanishes then for some j, $g_j(\mathbf{U})|_\sigma = 0$. Then $\Sigma'|_\sigma$ has the non-trivial solution $\xi^{(j)}$. Hence $\Gamma|_\sigma$ vanishes. **Q.E.D.**

So Γ and G have precisely the same set of zeros. By the Nullstellensatz (in the coefficient ring $\overline{D}$), for each $R \in \Gamma$, there is some $e \in \mathbb{N}$ such that

$$R^e \in \mathtt{Ideal}_{\overline{D}}(G(\mathbf{U})) \tag{35}$$

and there is some $d \in \mathbb{N}$ such that

$$G(\mathbf{U})^d \in \mathtt{Ideal}_{\overline{D}}(\Gamma). \tag{36}$$

Note that

$$R_0 := \mathtt{GCD}(\Gamma)$$

is well-defined, assuming the GCD takes place in the UFD $Q_D[\mathbf{U}]$. We will construct R_0 explicitly as follows: for each $i = 1, \ldots, s$, let $e_i \geq 0$ be the maximum value such that $g_i^{e_i}$ divides each $R \in \Gamma$. Since $G(\mathbf{U})$ is a product of the irreducible linear forms $g_1, \ldots, g_s$, it follows from (35) that each R is divisible by each g_i. This proves that each $e_i \geq 1$. Next we pick $H_1, \ldots, H_s \in \Gamma$ such that $g_i^{1+e_i}$ does not divide H_i (by definition, $g_i^{e_i}$ divides H_i). From equation (36), there exists $G_1, \ldots, G_t \in \Gamma$ such that $G(\mathbf{U})^d \in \mathtt{Ideal}_{\overline{D}}(G_1, \ldots, G_t)$. Consider

$$R_1 := \mathtt{GCD}(H_1, \ldots, H_s, G_1, \ldots, G_t)$$

where the GCD takes place in the UFD $Q_D[\mathbf{U}]$. Clearly $R_1 = \beta g_1^{e_1} \cdots g_n^{e_n}$ for some $\beta \in Q_D[\mathbf{U}]$. We show that, in fact, $\beta = 1$. From our choice of $G_1, \ldots, G_t$, we have $\mathtt{GCD}(G_1, \ldots, G_t) | G(\mathbf{U})^d$. Hence $R_1 | G(\mathbf{U})^d$, and so β is a power product in $g_1, \ldots, g_n$. But no positive power of any g_i can divide β (otherwise $\beta g_i^{e_i} | H_i$ implies $g_i^{1+e_i} | H_i$). This proves $\beta \in Q_D$, but this means $\beta = 1$, by the usual conventions for the GCD function. We define R_0 as an element of $D[\mathbf{U}]$ by multiplying R_1 with a suitable $\alpha \in D$,

$$R_0 := \alpha \cdot R_1 \in D[\mathbf{U}].$$

Clearly R_1 (and hence R_0) is a GCD of Γ over $Q_D[\mathbf{U}]$.

We call R_0 the *U-resultant* of Σ, and this is defined up to similarity (multiplication by elements of D). The resultant terminology for R_0 is justified because R_0 is a resultant polynomial of Σ: if $\Sigma|_\sigma$ has a solution iff $\Gamma|_\sigma = 0$, iff $G|_\sigma = 0$, iff $R_0|_\sigma = 0$. We summarize all this:

Theorem 34 *The U-resultant R_0 of Σ is a resultant polynomial of $\{A_0, A_1, \ldots, A_r\}$. It factors into linear factors of the form*

$$g_j = U_1 \xi_1^{(j)} + \cdots + U_n \xi_n^{(j)}$$

(with some multiplicity $e_j \geq 1$) in $\overline{D}[\mathbf{U}]$. The distinct linear factors in this factorization correspond to the s solution lines $D \cdot \xi^{(j)}$ $(j = 1, \ldots, s)$ of Σ.

Solving Projective Zero-dimensional Systems. Thus solving Σ is reduced to factoring the U-resultant over $\overline{D}$. Let us now assume $D = \mathbb{Z}$. We indicate how to use univariate real root isolation techniques and avoid factoring over $\mathbb{C}$. Our goal is to compute a representative set (33) of solutions for Σ in case $r = n - 1$. Let us first assume that $\xi_1^{(j)} \neq 0$ for all j. We may proceed as follows.
(1) Compute R_0 as the Macaulay resultant of Σ' $(= \Sigma \cup \{A_0\})$.
(2) Define the partial $\mathbf{U}$-specialization σ_k (for $2 \leq k \leq n$) as follows:

$$\sigma_k(U_\ell) = \begin{cases} U_1, & \text{if } \ell = 1, \\ -1 & \text{if } \ell = k, \\ 0 & \text{else.} \end{cases} \tag{37}$$

Now isolate the roots of the specialized U-resultant,

$$R_0|_{\sigma_k} = \prod_{j=1}^{s} (U_1 \xi_1^{(j)} - \xi_k^{(j)})^{e_j}.$$

Then each $\xi_k^{(j)}/\xi_1^{(j)}$ $(j = 1, \ldots, s)$ appears as a root of $R_0|_{\sigma_k}$. Let T_k denote this set of roots.
(3) Finally, we need to check for each element $\xi = (1, \xi_2, \ldots, \xi_n) \in \{1\} \times \prod_{k=2}^{n} T_k$ whether ξ is a zero of Σ. The ξ_i's are algebraic numbers and so, in principal, we know how to check this. For instance, ξ_i may be represented by isolating intervals or it may be represented by a sufficiently good approximation (cf. §IX.6). In any case, we need some root bounds on ξ_i. This is derived below.

Note that if the 1st and kth components of any root $\xi^{(j)}$ both vanish, then R_0 vanishes under the specialization σ_k and the method fails. In this case, we may be able to replace the special role of the first coordinate by the ith coordinate if the following holds: for all k, $k \neq i$, and for all $j = 1, \ldots, s$, either $\xi_k^{(j)}$ or $\xi_i^{(j)}$ is non-zero. Otherwise, we need further techniques. Since $D = \mathbb{Z}$, we may assume that $R_0(\mathbf{U})$ is a primitive polynomial (§III.1). Hence the U-resultant is determined up to sign. We first factor R_0 over the $\mathbb{Z}$:

Lemma 35 *Let $W \subseteq \{1, \ldots, n\}$. Then $R_0(\mathbf{U}) \in \mathbb{Z}[\mathbf{U}]$ factors over $\mathbb{Z}$ into two polynomials*

$$R_0 = R_W \overline{R}_W$$

with the following property: for all $j = 1, \ldots, s$,

$$\frac{R_W}{g_j} \in \mathbb{C}[\mathbf{U}] \text{ iff } (\exists i \in W)[\xi_i^{(j)} \neq 0].$$

Proof. We may choose $\overline{R}_W(\mathbf{U}) \in \mathbb{Z}[\mathbf{U}]$ to be the U-resultant of

$$\Sigma_W := \Sigma \cup \{X_i : i \in W\}.$$

Note that $\overline{R}_W | R_0$ since each zero of Σ_W is a zero of Σ.

Q.E.D.

Choosing $W = \{1, k\}$, we may apply now the substitution (37) to $R_W(\mathbf{U})$, etc., as before. But how do we know $R_W(\mathbf{U})$? After factoring $R_0(\mathbf{U})$ into its irreducible factors over $\mathbb{Z}$, we can easily discover which of these factors are factors of $R_W(\mathbf{U})$: these are precisely the factors that do not vanish under the specialization (37). In case one has $r > n-1$ equations in n variables, then a set of $n-1$ linear combination of these r equations (with coeficients randomly chosen from a suitable set) can be shown to result in a zero-dimensional system that can be solved as before. See also [113, 37].

Exercises

Exercise 8.1: We may assume that Γ consists of inertial elements. Show that Γ is a homogeneous ideal in $\mathbf{U}$. □

Exercise 8.2: The U-resultant of $X^2 + Y^2 - Z^2$ and $(X - Y - Z)(X - Y + Z)$ is $U_1^2 U_2^2 - U_1^2 U_3^2 - U_2^2 U_3^2 + U_3^4$. What are the linear factors of this U-resultant? Give a geometric interpretation of this system. □

Exercise 8.3: Complete the above outline of a method to solve a projective zero-dimensional system of $n-1$ equations in n variables. □

§9. Generalized Characteristic Polynomial

Let

$$\Sigma = \{A_1, \ldots, A_n\} \subseteq \mathbb{Z}[\mathbf{C}][X_1, \ldots, X_n] \tag{38}$$

be a system of n partial forms in n variables. Let $d_i = \deg A_i \geq 1$. The "main coefficient" of A_i is again the coefficient $a_i \in \mathbb{Z} \cup \mathbf{C}$ of $X_i^{d_i}$ in A_i. Note that a_i may be 0. We are interested in computing the Macaulay resultant $R_0 = \mathtt{res}(\Sigma)$, and its applications for solving zero-dimensional polynomial systems.

First consider the problem of computing R_0. In the two extreme cases, Σ may be a system of generic polynomials or it may be a system of homogeneous polynomials in $\mathbb{Z}[X_1, \ldots, X_n]$. In case Σ is a system of generic forms, we may use Macaulay's formula $\det(M)/\det(L)$ (equation (32)) to compute the resultant. If Σ is not generic, we can still use compute Macaulay's resultant for the generic case and then specialize. The problem is that the generic case is not computationally feasible, except for very small systems. Ideally, we should try to compute directly with the specializations $\widetilde{M}, \widetilde{L}$ of the matrices M, L. Unfortunately, the specialized determinants $\det(\widetilde{M})$ and $\det(\widetilde{L})$ may both vanish so that the division cannot be carried out.

To avoid this problem, Canny [37] introduced the *(generalized) characteristic polynomial* $\chi^\Sigma(\lambda)$ of Σ where λ is a new indeterminate. This is defined to be the Macaulay resultant of the system

$$\Sigma_\lambda := \{\lambda X_i^{d_i} - A_i : A_i \in \Sigma\}. \tag{39}$$

Let M be a Macaulay matrix of Σ_λ (this can be one of the $M^{(i)}$'s in §7). Let $\chi^M(\lambda)$ be the characteristic polynomial of M. Recall that in general, for any square matrix M, the *characteristic polynomial* of M (in the indeterminate λ) may be defined to be

$$\chi^M(\lambda) := \det(\lambda I - M)$$

where I is the identity matrix. In analogy to equation (32), we see that

$$\chi^\Sigma(\lambda) = \frac{\chi^M(\lambda)}{\chi^L(\lambda)}. \tag{40}$$

Note that while (32) assumes a regular system of forms, the formula (40) is also valid[9] for any M, L constructed from partial forms because of the presence of λ. There are well-known methods for computing the characteristic polynomials of matrices. Indeed, we can adapt Bareiss' algorithm (§X.2) if we want a polynomial-time algorithm. An efficient parallel algorithm[10] by Berkowitz [21] is also available for this computation. Once we have computed $\chi^\Sigma(\lambda)$, we can recover the Macaulay resultant $R_0 = \mathtt{res}(\Sigma)$ via the formula

$$R_0 = \pm\chi^\Sigma(0).$$

In other words, up to sign, R_0 is just the constant term in the polynomial χ^Σ.

[9] More precisely, the only concern when we partially specialize the numerator and denominator in formula (40) is that the denominator may vanish. But this cannot happen because of the introduction of λ.

[10] The method of Berkowitz explicitly avoids division in the underlying domain. A comparable algorithm of Csanky does not have this property (but may be adapted to yield an extra factor of $n!$, see [167, appendix]).

Division Scheme for $\chi^\Sigma(\lambda)$. We derive an alternative method suggested by Canny with the attractive feature that it only involves a sequence of $2n$ exact divisions. Let

$$\Sigma' = \{A'_1, \ldots, A'_n\} \subseteq \mathbb{Z}[\mathbf{C}, c_1, \ldots, c_n][X_1, \ldots, X_n]$$

where A'_i is obtained from A_i by replacing a_i with a new indeterminate c_i. We first develop the scheme to compute the Macaulay resultant of Σ' since its correctness is more transparent. Then we specialize Σ' to $\Sigma(\lambda)$ by replacing each c_i with $\lambda - a_i$.

As in §7, let $G^{(i)}$ $(i = 1, \ldots, n)$ denote the determinant of the Macaulay matrix $M^{(i)}$ with respect to the system Σ'. Writing χ_0 for the Macaulay resultant of Σ', we let

$$H^{(i)} := \frac{G^{(i)}}{\chi_0}. \tag{41}$$

Clearly, $H^{(i)}$ is a polynomial in $\mathbb{Z}[\mathbf{C}, c_1, \ldots, c_n]$. We make the important observation:

$$H^{(i)} \text{ is independent of } c_i. \tag{42}$$

This is because the c_i-degree of $G^{(i)}$, and of χ_0, are both equal to $\widehat{d}_i$. Next, define

$$G^{(i)}_j, H^{(i)}_j, \chi_j, \qquad (j = 0, \ldots, n)$$

to be the leading coefficients (respectively) of $G^{(i)}, H^{(i)}, \chi_0$ when viewed as polynomials in $c_1, \ldots, c_j$. Strictly speaking, it is nonsense to speak of the "leading coefficient" of a multivariate polynomial. But $G^{(i)}, \chi_0$ are "rectangular" and hence has a "leading coefficient" in the following sense:

> **Definition.** A polynomial $P = P(c_1, \ldots, c_n) \in D[c_1, \ldots, c_n]$ is *rectangular* if it contains a monomial of the form $\alpha c_1^{d_1} \cdots c_n^{d_n}$ for some $\alpha \in D$ where each d_i is the degree of P in the indeterminate c_i. We say P is *monic rectangular* if, in addition, $\alpha = 1$. The *leading coefficient* of a rectangular P, when viewed as an element of $D'[c_1, \ldots, c_i]$ where $D' = D[c_{i+1}, \ldots, c_n]$, refers to the element $\beta \in D'$ such that $\beta c_1^{d_1} \cdots c_i^{d_i}$ is a monomial of P.

Now each $G^{(i)}_j, H^{(i)}_j, \chi_j$ is, in turn, a rectangular monic polynomial in $c_{j+1}, \ldots, c_n$. The notation "χ_j" agrees with our original subscript "0" in χ_0. From (41), we obtain at once the corresponding equations

$$H^{(i)}_j = \frac{G^{(i)}_j}{\chi_j}, \qquad (j = 0, \ldots, n). \tag{43}$$

Note that $H^{(i)}_j$'s are principal minors of $M^{(i)}$. In view of (42), we obtain a series of equations and recurrences for the χ_i's:

$$\begin{aligned} H^{(i)}_i &= H^{(i)}_{i-1}, \qquad (i = 1, \ldots, n), \\ \frac{G^{(i)}_i}{\chi_i} &= \frac{G^{(i)}_{i-1}}{\chi_{i-1}}, \\ \chi_{i-1} &= \frac{G^{(i)}_{i-1}}{G^{(i)}_i}\chi_i. \end{aligned}$$

Telescoping,

$$\chi_0 = \frac{G^{(1)}_0}{G^{(1)}_1}\frac{G^{(2)}_1}{G^{(2)}_2} \cdots \frac{G^{(n)}_{n-1}}{G^{(n)}_n}\chi_n. \tag{44}$$

Since $\chi_n = 1$ and each $G^{(i)}_j$ is a minor of $M^{(i)}$, this formula leads to scheme to compute χ_0 by a sequence of exact divisions as follows.

Division Scheme:
Input: $G_i^{(i)}$ and $G_{i-1}^{(i)}$ for $i = 1, \ldots, n$.
Output: $\chi_0(\lambda)$, the generalized characteristic polynomial.
Method:
Initialization: $\chi_n \leftarrow 1$.
for $i \leftarrow n$ downto 1, do
1. $H_{i-1}^{(i)} \leftarrow \frac{G_i^{(i)}}{\chi_i}$.
2. $\chi_{i-1} \leftarrow \frac{G_{i-1}^{(i)}}{H_{i-1}^{(i)}}$.

Now let σ be the specialization that takes c_i to $\lambda - a_i$. We could have carried out the above computation for indeterminate c_i's and then specialize using σ. But this is inefficient. If $\widetilde{G}_j^{(i)}, \widetilde{H}_j^{(i)}$, etc., denote the σ-specialization of $G_j^{(i)}, H_j^{(i)}$, etc., then notice that $\widetilde{G}_j^{(i)}$ can be obtained as an appropriate minor of $\widetilde{M^{(i)}}$. Then the $\widetilde{H}_j^{(i)}$'s and $\widetilde{R}_j$'s can be computed using the same scheme. Of course, $\widetilde{R}_0$ is just the desired characteristic polynomial $\chi^{\Sigma}(\lambda)$.

Computing the Constant Term of $\chi^{\Sigma}(\lambda)$. Suppose we only want to compute R_0, without computing the entire characteristic polynomial χ_0. We observe that $\chi^M(\lambda)$ and $\chi^L(\lambda)$ are monic polynomials in λ. As such, the coefficients of their quotient can be directly obtained as suitable subdeterminants according to §III.4. To be explicit, suppose

$$C(X) = \sum_{i=0}^{m-n} c_i X^i \tag{45}$$

is the quotient of A divided by B where

$$A(X) = \sum_{i=0}^{m} a_i X^i, \quad B(X) = \sum_{i=0}^{n} b_i X^i, \qquad m \geq n. \tag{46}$$

are monic polynomials. Then (§III.4) the coefficient c_i is, up to sign, the $(m-n-i+1)$st principal minor of the following matrix:

$$M := \begin{bmatrix} a_m & a_{m-1} & a_{m-2} & \cdots & a_{m-n} & \cdots & a_1 & a_0 \\ b_n & b_{n-1} & b_{n-2} & \cdots & b_0 & \cdots & 0 & 0 \\ & b_n & b_{n-1} & \cdots & b_1 & \cdots & 0 & 0 \\ & & \ddots & & & & & \vdots \\ & & & & b_n & b_{n-1} & \cdots & b_1 & b_0 \end{bmatrix}. \tag{47}$$

For instance, the leading coefficient c_{m-n} of C is given by the 1st principal minor, which is $a_m = 1$. The next coefficient c_{m-n-1} is given by $\det \begin{bmatrix} a_m & a_{m-1} \\ b_n & b_{n-1} \end{bmatrix}$. For $i = 1, \ldots, m-n$, the coefficient c_{m-n-i} depends only on $a_{m-1}, \ldots, a_{m-i}$ and $b_{n-1}, \ldots, b_{\min\{0,n-i\}}$. In particular, the constant term c_0 depends only on the leading $m-n+1$ coefficients of polynomials A and B. Note that these remarks do not assume that B divides A exactly.

Now return to our goal of computing the constant term R_0 of χ^M/χ^L. First, recall that M is an $N_d \times N_d$ matrix where $N_d = \binom{d+n-1}{d}$ and $d = 1 + \sum_{i=1}^n d_i$ (see §7). Also L is a $(N_d - \widehat{d}) \times (N_d - \widehat{d})$ submatrix of M where $\widehat{d} = \sum_{i=1}^n \widehat{d}_i$ and $\widehat{d}_i = (\prod_{j=1}^n d_j)/d_i$. Hence χ^M/χ^L is a polynomial of degree $\widehat{d}$. Therefore to compute the constant term of χ^M/χ^L, it is sufficient to compute the leading $\widehat{d}+1$ coefficients of χ^M and χ^L.

Solving Zero-dimensional Systems. In the rest of this section, assume the system

$$\Sigma = \{A_1, \ldots, A_n\} \subseteq \mathbb{Z}[X_1, \ldots, X_n], \tag{48}$$

not necessarily homogeneous, has a finite number of zeros in $\mathbb{C}^n$. Our goal is to find these zeros. We may first homogenize each A_i to $\widehat{A}_i$ using a new variable X_0. Let $\widehat{\Sigma} = \{\widehat{A}_1, \ldots, \widehat{A}_n\}$. Assuming that $\widehat{\Sigma}$ has finitely many solution lines in $\mathbb{C}^n$, then we may compute its U-resultant $R(U_1, \ldots, U_n)$, which is just the Macaulay resultant of

$$\widehat{\Sigma} \cup \{A_0\}, \qquad \text{where } A_0 = X_0U_0 + X_1U_1 + \cdots + X_nU_n. \tag{49}$$

Using $R(\mathbf{U})$, we may compute a representative set of $\widehat{\Sigma}$, as in §8. From among this representative set, any zero

$$\xi' = (\xi'_0, \xi'_1, \ldots, \xi'_n)$$

where $\xi'_0 \neq 0$ will yield a zero $(\xi'_1/\xi'_0, \ldots, \xi'_n/\xi'_0)$ of Σ. We say ξ' is a "zero at infinity" or a "finite zero" for $\widehat{\Sigma}$, depending on whether $\xi'_0 = 0$ or not. In terms of solution lines, we speak of "solution lines at infinity" or "finite solution lines". Conversely, any zero $\xi = (\xi_1, \ldots, \xi_n)$ of Σ is proportional to a finite zero in the representative set. Hence, the system (48) is solved.

The problem is that $\widehat{\Sigma}$ may have infinitely many solution lines at infinity (Exercise). That is, its zero set has dimension at least 1 in the projective n-space, $\mathbb{P}^n(\mathbb{C})$. For any $\mathbf{U}$-specialization σ, the equation $A_0|_\sigma = 0$ defines a hyperplane, which is a zero set of dimension $n-1$, in projective n-space. If the zero set of $\widehat{\Sigma}$ is of dimension at least 1, it must intersect this hyperplane. This is because of the general result[11] that two zero sets in projective n-space of dimensions i and j (respectively) have non-empty intersection provided $i + j \geq n$. This means that the system (49) has a common solution when specialized by σ, and so $R(\mathbf{U})|_\sigma = 0$. Since σ is arbitrary, we conclude that $R(\mathbf{U})$ is identically zero. Combined with §9, we conclude:

Lemma 36 *The U-resultant of $\widehat{\Sigma}$ does not vanish iff $\widehat{\Sigma}$ is projective zero-dimensional.*

This result can be the basis of a method for testing if $\widehat{\Sigma}$ is projective zero-dimensional. If the U-resultant of $\widehat{\Sigma}$ vanishes, then U-resultants are useless for our purposes. The next section shows how to overcome this.

Exercises

Exercise 9.1: Verify that the following system has finitely many solutions, and solve it.

$$\Sigma = \{X^3 - X^2 + X - 1, XY - Y - X^2 + X, Y^2 - X^2\}$$

□

Exercise 9.2: (Hinternaus) Compute the U-resultant of $A_1 = Y^3 - 2t^3$ and $A_2 = Y^3 + 3XY^2 + 3X^2Y + X^3 - 2t^3$, and solve this system. HINT: there are 9 solutions. □

[11] See, e.g., Mumford [143]. This is a generalization of a more familiar fact on the intersection of linear subspaces: if S, T are linear subspaces of Euclidean n-space, then $\dim(S \cap T) = \dim(S) + \dim(T) - \dim(S \cup T)$. If $\dim(S) + \dim(T) \geq n$ then the intersection $S \cap T$ is necessarily non-empty since $\dim(S \cup T) \leq n$.

Exercise 9.3: Construct a system Σ with finitely many zeros, but where $\widehat{\Sigma}$ has infinitely many solution lines.

□

Exercise 9.4: Try to adapt Berkowitz' algorithm [21] to compute only the first k (for any given $k \geq 1$) leading coefficients of the characteristic polynomial of a matrix. □

Exercise 9.5: Study conditions under which one can directly use the Macaulay quotient formula (without introducing extra indeterminates). □

§10. Generalized U-resultant

Let $\Sigma = \{A_1, \ldots, A_n\} \subseteq \mathbb{Z}[X_1, \ldots, X_n]$ have finitely many zeros in $\mathbb{C}^n$ and let $\widehat{\Sigma} = \{\widehat{A}_1, \ldots, \widehat{A}_n\} \subseteq \mathbb{Z}[X_0, X_1, \ldots, X_n]$, as in equations (48) and (49). Assume the U-resultant of $\widehat{\Sigma}$ vanishes. We now introduce a non-vanishing substitute for the U-resultant.

Let λ be a new indeterminate and compute the U-resultant of

$$\Sigma(\lambda) := \{\lambda X_1^{d_1} - \widehat{A}_1, \lambda X_2^{d_2} - \widehat{A}_2, \ldots, \lambda X_n^{d_n} - \widehat{A}_n\} \tag{50}$$

where $d_i = \deg A_i$. The result, denoted

$$\chi(\lambda, \mathbf{U}),$$

is almost the characteristic polynomial of the system (49). In fact, under the partial specialization taking U_0 to $\lambda - U_0$, $\chi(\lambda, \mathbf{U})$ is transformed to the characteristic polynomial of (49). This implies $\chi(\lambda, \mathbf{U})$ is non-zero. Viewing $\chi(\lambda, \mathbf{U})$ as a polynomial in λ, let $\widehat{R}(\mathbf{U})$ be its tail coefficient. Alternatively, if $i_0 \geq 0$ is the largest index such that λ^{i_0} divides $\chi(\lambda, \mathbf{U})$, then

$$\chi'(\lambda, \mathbf{U}) := \chi(\lambda, \mathbf{U})\lambda^{-i_0} \tag{51}$$

is a polynomial (the "reduced characteristic polynomial") with $\widehat{R}(\mathbf{U})$ as constant term. It turns out that $\widehat{R}(\mathbf{U})$ can now play the role of the U-resultant. Note that the constant term of $\chi(\lambda, \mathbf{U})$ is just the usual U-resultant of $\widehat{\Sigma}$. When this constant term is non-zero, $\widehat{R}(\mathbf{U})$ is equal to this U-resultant. It is therefore appropriate to call $\widehat{R}(\mathbf{U})$ the *generalized U-resultant* of $\widehat{\Sigma}$.

Lemma 37 *$\widehat{R}(\mathbf{U})$ is homogeneous of degree $D = \prod_{i=1}^n d_i$ in $U_0, \ldots, U_n$.*

Proof. Suppose $\Sigma^\wedge$ denotes the system (50) where we replace each non-zero coefficient of a power product of $\mathbf{X}$ by a new indeterminate. Let u_i be the new indeterminate coefficient of $X_i^{d_i}$. Let σ be the specialization such that $\Sigma^\wedge|_\sigma = \Sigma(\lambda)$. [For instance the coefficient of $X_1^{d_1}$ in $\widehat{A_1}$ is $c_1 \in \mathbb{Z}$, then we replace $\lambda - c_1$ by the new indeterminate u_1, and $\sigma(u_1) = \lambda - c_1$.] The U-resultant of $\Sigma^\wedge$, denoted $R^\wedge$, is the Macaulay resultant of $\Sigma^\wedge \cup \{A_0\}$. By theorem 31 (§7), $R^\wedge$ is homogeneous of degree $D = \prod_{i=1}^n d_i$ in the coefficients $\mathbf{U}$ of A_0. Since $R^\wedge|_\sigma = \chi(\lambda, \mathbf{U})$ and since σ does not affect the U_i's, it follows that $\chi(\lambda, \mathbf{U})$ is homogeneous of degree D in $\mathbf{U}$. The lemma follows since $\widehat{R}\lambda^i$ (for some $i \geq 0$, perhaps with an integer coefficient) is a monomial of $\chi(\lambda, \mathbf{U})$ and λ does not occur in A_0. **Q.E.D.**

Lemma 38 *For all $\alpha \in \mathbb{C} \setminus \{0\}$, $\chi(\alpha, \mathbf{U}) \neq 0$ and factors completely into $D = \prod_{i=1}^{n} d_i$ linear homogeneous factors in $\mathbf{U}$.*

Proof. Pick any $\mathbf{u} = (u_0, \ldots, u_n) \in \mathbb{C}^{n+1}$ such that $\widehat{R}(\mathbf{u}) \neq 0$. Then as $\alpha \to 0$, the non-constant monomials in the "reduced characteristic polynomial" (see (51)) $\chi'(\alpha, \mathbf{U})$ approach 0. Hence for α small enough, $\chi'(\alpha, \mathbf{u}) \neq 0$ and so $\chi'(\alpha, \mathbf{U}) \neq 0$. This implies $\chi(\alpha, \mathbf{U}) \neq 0$. But in fact, this argument works for any $\alpha \neq 0$ because we can make $\widehat{R}(\mathbf{u})$ arbitrarily large by multiplication with a large constant K, since $\widehat{R}(K\mathbf{u}) = K^D \widehat{R}(\mathbf{u})$.

To see that $\chi(\alpha, \mathbf{U})$ completely factors, note that $\chi(\alpha, \mathbf{U})$ is the U-resultant of the system (50) under the specialization $\lambda \longrightarrow \alpha$. It follows from lemma 36 that this specialized system has finitely many solution lines. These solution lines correspond to the complete factorization of $\chi(\alpha, \mathbf{U})$ into linear homogeneous factors in $\mathbf{U}$. By lemma 37, the number (with multiplicity of factors counted) of linear factors is $D = \prod_{i=1}^{n} d_i$. **Q.E.D.**

Consider the zero set of the characteristic polynomial,

$$V := \textsc{Zero}(\chi(\lambda, \mathbf{U})) \subseteq \mathbb{C} \times \mathbb{C}^{n+1}.$$

Clearly

$$V = V_0 \cup V_1 \tag{52}$$

where $V_0 := \textsc{Zero}(\lambda^{i_0})$ and $V_1 := \textsc{Zero}(\chi'(\lambda, \mathbf{U}))$ (see (51)). Note that if $i_0 = 0$ then V_0 is the empty set, otherwise it is the hyperplane $\{0\} \times \mathbb{C}^{n+1}$. By lemma 38, we have $V_1 \not\subseteq V_0$. It follows that (see [107, p. 131])

$$\dim(V_1 \cap V_0) \leq n.$$

Also, let

$$W := \textsc{Zero}(\widehat{R}(\mathbf{U})) \subseteq \mathbb{C}^{n+1}.$$

The following simple observation is a useful tool.

Lemma 39 *Let $\mathbf{u} \in \mathbb{C}^{n+1}$. If there exists a sequence $\{(\lambda_i, \mathbf{u}_i)\}_{i \geq 1}$ in $V_1 \setminus V_0$ that converges to $(0, \mathbf{u})$ then $\mathbf{u} \in W$.*

Proof. By choice of $\lambda_i, \mathbf{u}_i$, $\chi'(\lambda_i, \mathbf{u}_i) = 0$. Since $\{(\lambda_i, \mathbf{u}_i)\}_{i \geq 0}$ converges to $(0, \mathbf{u})$, we conclude by continuity that $\chi'(0, \mathbf{u}) = 0$. But $\chi'(0, \mathbf{u}) = 0$ implies $\widehat{R}(\mathbf{u}) = 0$. **Q.E.D.**

Let

$$\pi_{\mathbf{U}} : \mathbb{C} \times \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$$

be the projection map onto the $\mathbf{U}$-coordinates. Thus $\pi(\alpha, u_0, u_1, \ldots, u_n) = (u_0, \ldots, u_n)$. In view of lemma 38, for all $\alpha \neq 0$ small enough, the set

$$W(\alpha) := \pi_{\mathbf{U}}(V \cap (\{\alpha\} \times \mathbb{C}^{n+1}))$$

is a union of D (not necessarily distinct) hyperplanes in $\mathbb{C}^{n+1}$. Now we may view $W(\lambda)$ as a parametrized system of D hyperplanes. In a suitable sense, the function $W(\lambda)$ is continuous in the neighborhood of $\lambda = 0$. To see this, view the D hyperplanes dually, as the multiset of D solution lines of a homogeneous polynomial $\chi_\lambda(\mathbf{U})$ in $\mathbf{U}$ with coefficients depending on λ. The multiset of D solution lines varies continuously with λ provided the (homogeneous) degree is constant. In our case, this provision amounts to $\lambda \neq 0$. We may take the distance between two solution lines to be

the Euclidean distance between their respective intersection points with the unit sphere. Continuity of the D lines means that for any $\epsilon > 0$ *small enough*, there is a $\delta > 0$ such that if λ varies by less than δ then for each solution line ℓ with multiplicity $\mu \geq 1$, there are exactly μ perturbed solutions lines (counting multiplicity) that are within ϵ of ℓ. Take any sequence $\overline{\lambda} = \{\lambda_i\}_{i\geq 1}$ that converges to 0. The systems $W(\lambda_i)$ converge as $i \to \infty$. By compactness of the space of hyperplanes (this is a Grassmanian space, or the fact that the space of solution lines is topologically the unit sphere with antipodal points identified), there exists a limit system of D hyperplanes which we may denote by

$$W(0).$$

We show $W(0)$ is unique, in fact, it is equal to W, the zero set of $\widehat{R}(\mathbf{U})$. To see this, consider any $\mathbf{u} = (u_0, \ldots, u_n) \in W(0) \setminus \mathbf{0}$. Note that $\mathbf{u}$ occurs in $W(0)$ with a given multiplicity. From our definition of $W(0)$, there is a sequence $\{\zeta_i\}_{i\geq 1}$ in $V_1 \setminus V_0$ that approaches $(0, \mathbf{u})$. In fact, we may choose $\zeta_i \in (\{\lambda_i\} \times W(\lambda_i)) \subseteq V_1$ for all $i \geq 1$. By lemma 39, this implies $\mathbf{u} \in W$ (with its multiplicity intact). Thus W contains the D hyperplanes of $W(0)$. Since $W = \textsc{Zero}(\widehat{R}(\mathbf{U}))$ and $\widehat{R}(\mathbf{U})$ has degree D, we conclude:

Lemma 40 *We have the equality* $W(0) = W = \textsc{Zero}(\widehat{R}(\mathbf{U}))$.

Unfortunately, this result does not establish that the zeros of Σ are encoded (in the sense of U-resultants) among the linear factors of $\widehat{R}(\mathbf{U})$. We need a more careful analysis of the behaviour of the system as λ goes to 0, using an argument similar to Renegar [167].

Pick any $\mathbf{u} \in \mathbb{R}^{n+1}$ such that $\widehat{R}(\mathbf{u}) \neq 0$. Let $\mathbf{e}_i \in \mathbb{R}^{n+1}$ $(i = 0, \ldots, n)$ be the elementary unit vector with 1 at the $(i+1)$st coordinate and 0 elsewhere. With t a new indeterminate, consider the polynomial

$$r_i(\lambda, t) := \chi'(\lambda, t\mathbf{u} - \mathbf{e}_i),$$

where we view $r_i(\lambda, t)$ as a polynomial in t with coefficients parametrized by λ. To avoid excessive notations, we may by reason of symmetry focus on $i = 1$, in which case we simply write $r_1(\lambda, t)$ as $r(\lambda, t)$.

Lemma 41 *In a small enough neighborhood* N_0 *of* 0*, for all* $\lambda \in N_0$*:*

(i) *The polynomial* $r(\lambda, t)$ *is non-vanishing of degree* D *in* t.

(ii) *We have* $\chi'(\lambda, \mathbf{u}) \neq 0$. *In particular,* $\chi'(0, \mathbf{u}) \neq 0$.

Proof. (i) Note that

$$\widehat{R}(t\mathbf{u} - \mathbf{e}_1) = t^D \widehat{R}(\mathbf{u}) + \widehat{R}(-\mathbf{e}_1) + t \cdot q(t, \mathbf{u})$$

where $q(t, \mathbf{u})$ is a polynomial whose t-degree is less than $D - 1$. Hence $\widehat{R}(t\mathbf{u} - \mathbf{e}_1)$ is non-vanishing of degree t. Then, arguing as in lemma 38, we conclude that $\chi'(\lambda, t\mathbf{u} - \mathbf{e}_1)$ does not vanish for small enough λ. In fact, $r(\lambda, t)$ has the form $t^D(\widehat{R}(\mathbf{u}) + \lambda h(\lambda, \mathbf{u}))$+l.o.t., for some $h(\lambda, \mathbf{u})$ and "l.o.t" indicates lower order terms in t. Thus the leading coefficient of $r(\lambda, t)$ tends to $\widehat{R}(\mathbf{u})$ when λ is small enough. This establishes the existence of exactly D roots.
(ii) Furthermore, when λ is small enough, $\chi'(\lambda, \mathbf{u})$ tends to $\widehat{R}(\mathbf{u})$, a non-zero value. **Q.E.D.**

By lemma 38,

$$\chi'(\lambda, \mathbf{U}) = \prod_{j=1}^{D} g_j(\lambda, \mathbf{U}), \qquad \text{where } \; g_j(\lambda, \mathbf{U}) = \sum_{\ell=0}^{n} U_\ell \xi_\ell^{(j)}(\lambda),$$

where $\xi^{(j)}(\lambda) = (\xi_0^{(j)}(\lambda), \ldots, \xi_n^{(j)}(\lambda))$. It is immediate from

$$r(\lambda, t) = \prod_{j=1}^{D} g_j(\lambda, t\mathbf{u} - \mathbf{e}_1),$$

that the zeros of $r(\lambda, t)$ are precisely those t such that

$$g_j(\lambda, t\mathbf{u} - \mathbf{e}_1) = t g_j(\lambda, \mathbf{u}) - \xi_1^{(j)}(\lambda) = 0 \tag{53}$$

for some $j = 1, \ldots, D$. Since $\chi'(\lambda, \mathbf{u}) \neq 0$ (for $\lambda \in N_0$), we conclude that $g_j(\lambda, \mathbf{u}) \neq 0$ for all j. It follows that $r(\lambda, t)$ has exact D roots (possibly non-distinct) that can be expressed (via (53)) in the form

$$t_1^{(j)}(\lambda) = \frac{\xi_1^{(j)}(\lambda)}{g_j(\lambda, \mathbf{u})}, \qquad (j = 1, \ldots, D). \tag{54}$$

We now use a well-known fact: if $r(\lambda, t)$ is a polynomial in t whose coefficients vary continuously with λ within an open interval N_0, and the degree D of $r(\lambda, t)$ is invariant for $\lambda \in N_0$, then the D roots (with multiplicity counted) of $r(\lambda, t)$ vary continuously in the following precise sense:

> *For all λ_0, there exist $\varepsilon > 0$ and $\delta > 0$ such that for all λ', if $|\lambda' - \lambda_0| < \delta$ then for each root t_0 of multiplicity $\mu \geq 1$ of $r(\lambda_0, t)$, there are exactly μ roots (with multiplicity counted) of $r(\lambda', t)$ that are within distance ε of t_0.*

We may therefore define the limits of these root functions,

$$\tau_1^{(j)} := \lim_{\lambda \to 0} t_1^{(j)}(\lambda).$$

Note that $\tau_1^{(j)}$ is finite because the denominator in (54) approaches a non-zero value $g_j(0, \mathbf{u})$. Returning to the general notation $r_i(\lambda, t)$ (for $i = 0, \ldots, n$, not just $i = 1$), we have thus established the existence of the continuous root functions

$$t_i^{(j)}(\lambda), \qquad (i = 0, \ldots, n; j = 1, \ldots, D)$$

for $\lambda \in N_0$. Again, we can take their limits as λ approaches 0. But we need something more. We really want to ensure that the "j" superscripts are assigned consistently, so that

$$(t_0^{(j)}(\lambda), \ldots, t_n^{(j)}(\lambda)) \cdot g_j(\lambda, \mathbf{u}) = (t_0^{(j)}(\lambda) \cdot g_j(\lambda, \mathbf{u}), \ldots, t_n^{(j)}(\lambda) \cdot g_j(\lambda, \mathbf{u})) \tag{55}$$

are indeed zeros of $\Sigma(\lambda)$ (cf. (54)).

Each time a multiple root splits, we have to decide which of the branches should a particular $t_i^{(j)}(\lambda)$ follow. There can be a problem if these D roots merge and split indefinitely. More precisely, let us temporarily fix $i = 0, 1, \ldots, n$. For any $\lambda \in N_0$, let us say that the multiset of roots

$$S_i(\lambda) := \{t_i^{(j)}(\lambda) : j = 1, \ldots, D\}$$

has *generalized multiplicity* (or simply, *multiplicity*) $\mu = (\mu_1, \mu_2, \ldots, \mu_D) \in \mathbb{N}^D$ if there are exactly μ_1 simple roots, exactly μ_2 roots of multiplicity 2, and in general, μ_i roots of multiplicity i. So $\sum_{i=1}^{D} i \cdot \mu_i = D$. If the generalized multiplicity is constant for $\lambda \in N_0$, then we clearly can define $t^{(j)}(\lambda)$ unambiguously, and thus the limiting $\xi^{(j)}$ are unique. We will show that the generalized multiplicity of $S_i(\lambda)$ can only change a finite number of times as λ approaches 0. Again, it is sufficient to focus on case $S_1(\lambda)$, and let $\mu(\lambda)$ denote the generalized multiplicity of $S_1(\lambda)$. We call

$\lambda' \in \mathbb{C}$ a *critical value* for $S_1(\lambda)$ if for all neighborhoods N' of λ', there exists $\lambda'' \in N'$ such that $\mu(\lambda'') \neq \mu(\lambda')$. Let the derivative of r be

$$r'(\lambda, t) = \frac{dr(\lambda, t)}{dt}.$$

Let $p_j(\lambda)$ denote the pseudo-principal subresultant coefficient $\texttt{ppsc}_j(r, r')$ (for $j = 0, \ldots, D-1$) of $r(\lambda, t)$ and $r'(\lambda, t)$ (§III.7). For any λ', the degree of $\texttt{GCD}(r, r')$ is the smallest j such that $p_j(\lambda') \neq 0$ (§III.8, Exercise). Hence the number of distinct roots of r is equal to $D - j$ (property C5, §VI.1). Let j_1 be the least index such that $p_{j_1}(\lambda)$ is non-vanishing. Define $C_1(\lambda) := p_{j_1}(\lambda)$. Clearly j_1 and $C_1(\lambda)$ are well-defined.

Lemma 42

(i) *A value λ' is a critical for $S_1(\lambda)$ iff λ' is a root of $C_1(\lambda)$.*

(ii) *If $\lambda_0 < \lambda_1$ and $\mu(\lambda_0) \neq \mu(\lambda_1)$ then there is a critical value λ', $\lambda_0 \leq \lambda' \leq \lambda_1$.*

(iii) *For a small enough neighborhood N_0 of 0, the polynomial $r(\lambda, t)$ has exactly $D - j_1$ distinct roots when $\lambda \in N_0$.*

Proof. (i) Let us note that if λ' is critical then every neighborhood of λ' contains a λ'' such that $r(\lambda'', t)$ has a different number of distinct roots than $r(\lambda', t)$. Hence, if $C_1(\lambda') = 0$ then λ' is critical. Conversely, if $C_1(\lambda') \neq 0$ then there is a neighborhood N' of λ' such that for all $\lambda'' \in N'$, $p_j(\lambda'') \neq 0$ for $j \in I$. Then the number of distinct roots of $r(\lambda, t)$ is constant for $\lambda \in N'$. Such a λ' is not critical.
(ii) If either λ_0 or λ_1 is critical then we are done. Otherwise, let $\lambda' = \sup\{\lambda \in (\lambda_0, \lambda_1) : \mu(\lambda) = \mu(\lambda_0)\}$. If $\mu(\lambda') = \mu(\lambda_0)$ then for all $\varepsilon > 0$ sufficiently small, $\mu(\lambda' + \varepsilon) \neq \mu(\lambda')$. If $\mu(\lambda') \neq \mu(\lambda_0)$ then for all $\varepsilon > 0$ sufficiently small, $\mu(\lambda' - \varepsilon) \neq \mu(\lambda')$. In either case, λ' is critical.
(iii) Choose N_0 small enough to avoid all the roots of $C_1(\lambda)$. By (i) and (ii), $\mu(\lambda)$ is invariant for $\lambda \in N_0$. **Q.E.D.**

The above focused on the multiset $S_1(\lambda)$. Now assume N_0 is small enough to ensure that the previous two lemmas hold for each $S_i(\lambda)$ $(i = 0, 1, \ldots, n)$. Restricting λ to N_0, we can now define unambiguously the continuous functions $t_i^{(j)}(\lambda)$ for each $i = 0, \ldots, n$ and $j = 1, \ldots, D$. Moreover, we will choose the indexes so that for each $\lambda \in N_0$, the point in (55) is a zero of $\Sigma(\lambda)$. Now we may define

$$\xi^{(j)} = (\xi_0^{(j)}, \ldots, \xi_n^{(j)}) \qquad \text{where } \xi_i^{(j)} := \tau_i^{(j)} g_j(0, \mathbf{u}). \tag{56}$$

It is immediate that each $\xi^{(j)}$ is a zero of $\Sigma(0)$. But $\Sigma(0) = \widehat{\Sigma} \cup \{A_0\}$. By a suitable renumbering, let us assume that $\{\xi^{(i)} : i = 1, \ldots, s\}$ comprises the multiset of *finite* zeros, *i.e.*, whose first component $\xi_0^{(j)} \neq 0$. Write

$$\overline{\xi}^{(i)} = (\overline{\xi}_1^{(i)}, \ldots, \overline{\xi}_n^{(i)}) := (\xi_1^{(i)}/\xi_0^{(i)}, \ldots, \xi_n^{(i)}/\xi_0^{(i)}), \qquad (i = 1, \ldots, s). \tag{57}$$

We have therefore shown that each $\overline{\xi}^{(i)}$ is a solution of Σ (the original system). We still need to show that every solution to Σ is among the $\overline{\xi}^{(i)}$'s in (57). Towards this end, and in analogy to (34), define the polynomial

$$G(\mathbf{U}) := \prod_{j=1}^{s} \overline{g}_j(\mathbf{U}), \qquad \text{where } \overline{g}_j(\mathbf{U}) := U_0 + \sum_{i=1}^{n} U_i \overline{\xi}_i^{(j)}.$$

We say that $\overline{g}_j(\mathbf{U})$ "encodes" the solution $\overline{\xi}^{(j)}$. Finally, here is the main result on generalized U-resultants (cf. §8, theorem 34):

Theorem 43 *The generalized U-resultant $\widehat{R}(\mathbf{U})$ of Σ has the following properties:*

(i) *$\widehat{R}(\mathbf{U})$ factors completely into linear factors over the complex numbers $\mathbb{C}$.*

(ii) *The polynomial $G(\mathbf{U})$ divides $\widehat{R}(\mathbf{U})$.*

(iii) *The linear factors of $G(\mathbf{U})$ encode all, and only, solutions of Σ.*

Proof.
(i) This is just a restatement of lemma 40.
(ii) Each $\overline{\xi}^{(j)}$ defines the hyperplane $\overline{g}_j(\mathbf{U}) = 0$ in $\mathbf{U}$-space. Suppose $\overline{g}_j(\mathbf{u}) = 0$ for some $\mathbf{u}$. Then the specialization

$$\sigma_0 : (\lambda; \mathbf{X}; \mathbf{U}) \longrightarrow (0; 1, \overline{\xi}^{(j)}; \mathbf{u}), \qquad \mathbf{X} = (X_0, X_1, \ldots, X_n) \tag{58}$$

is a zero of $A_0(\mathbf{X}, \mathbf{U})$, and hence of $\Sigma(0)$. Take any sequence $\{\lambda_\ell\}_{\ell \geq 1}$ that approaches zero. Then we get a corresponding sequence $\{(1, \overline{\xi}^{(j)}(\lambda_\ell))\}_{\ell \geq 1}$ that approaches $\mathbf{X} = (1, \overline{\xi}^{(j)})$. CLAIM: we can construct a corresponding sequence

$$\{(\lambda_\ell, \mathbf{u}_\ell)\}_{\ell \geq 1}$$

in $V_1 \setminus V_0$ (see (52)) that approaches $(\lambda, \mathbf{U}) = (0, \mathbf{u})$.

Before verifying the CLAIM, note that the claim implies $\widehat{R}(\mathbf{u}) = 0$, by lemma 39. In other words, $\overline{g}_j(\mathbf{U})$ divides $\widehat{R}(\mathbf{U})$ and part (ii) is proved.

It remains to show the CLAIM. It suffices to choose a sequence $\mathbf{u}_\ell$ ($\ell \geq 1$) so that

$$A_0(1, \overline{\xi}^{(j)}(\lambda_\ell), \mathbf{u}_\ell) = 0.$$

and $\mathbf{u}_\ell$ approaches $\mathbf{u}$. This would mean $\Sigma(\lambda_\ell)$ has the solution

$$\sigma_\ell : (\lambda; \mathbf{X}; \mathbf{U}) \longrightarrow (0; 1, \overline{\xi}^{(j)}(\lambda_\ell); \mathbf{v}_\ell).$$

and hence $\chi'(\lambda_\ell, \mathbf{v}_\ell) = 0$. This shows $(\lambda_\ell, \mathbf{v}_\ell) \in V_1 \setminus V_0$, as desired by the claim. By going to a subsequence of $\{\mathbf{u}_\ell\}_\ell$ if necessary, we assume that $\|\overline{\xi}^{(j)}(\lambda_\ell) - \overline{\xi}^{(j)}\|_\infty$ is at most $1/\ell$. Initially, $\mathbf{u}_1$ is arbitrary. It is not hard to see that we can ensure that $\|\mathbf{u}_\ell - \mathbf{u}\|_\infty$ is $O(1/\ell)$. This implies $\mathbf{u}_\ell$ approaches $\mathbf{u}$.
(iii) We already noted that each $\overline{\xi}^{(j)}$ is a zero of Σ. Conversely, let $\zeta = (\zeta_1, \ldots, \zeta_n)$ be a zero of Σ. Then $(1, \zeta)$ is a zero of $\widehat{\Sigma}$. This $(1, \zeta)$ defines the $\mathbf{U}$-hyperplane $g(\mathbf{U}) = U_0 + \sum_{i=1}^n \zeta_i U_i = 0$. By the same argument as part (ii), $g(\mathbf{U})$ divides $\widehat{R}(\mathbf{U})$. By definition, $G(\mathbf{U})$ contains all linear factors of $\widehat{R}(\mathbf{U})$ in which U_0 has a non-zero coefficient. Hence $g(\mathbf{U})$ divides $G(\mathbf{U})$. **Q.E.D.**

Finally, let us consider how to extract the zeros of Σ without factoring over $\mathbb{C}$. We need an analog of lemma 35. As usual, $g_j(\mathbf{U}) = \sum_{i=0}^n \xi_i^{(j)} U_i$.

Lemma 44 *Let $W \subseteq \{1, \ldots, n\}$. Then $\widehat{R}(\mathbf{U})$ factors over $\mathbb{Z}$ into $R_W \overline{R}_W$ where, for all $j = 1, \ldots, s$, $R_W / g_j \in \mathbb{C}[\mathbf{U}]$ iff for some $i \in W$, $\xi_i^{(j)} \neq 0$.*

Proof. Consider the system $\Sigma_W := \Sigma \cup \{X_i : i \in W\}$. This system has a finite number of zeros. We cannot directly construct the generalized U-resultant of Σ_W since this is only defined for a system

of n homogeneous polynomials in $n+1$ variables. But by a result of Eisenbud and Evans [107, p. 126, corollary 1.5], there are n polynomials $B_1, \ldots, B_n \in \mathbb{Z}[X_1, \ldots, X_n]$ whose zero set comprises precisely these zeros (not necessarily with the same multiplicities). We homogenize these B_i's and construct their generalized U-resultant R'_W. By taking a large enough power $(R'_W)^k$ $(k \geq 1)$, we see that $\texttt{GCD}((R'_W)^k, \widehat{R}) \in \mathbb{Z}[\mathbf{U}]$ has the property we desire of $\overline{R}_W$. **Q.E.D.**

To find the kth components of $\overline{\xi}^{(j)}$ $(j = 1, \ldots, s)$, set $W = \{k\}$. We factor $\widehat{R}$ into irreducible polynomials over $\mathbb{Z}$. Such a polynomial F divides R_W iff F depends on U_k. The product of all these F's that depend on U_k therefore constitute R_W. We may now proceed as before (§8) to find the kth components of zeros by root isolation, etc.

Finally, the notion of U-resultant can be generalized to the concept of Chow forms of projective varieties (e.g., see Gelfand, Kapranov and Zelevinsky [70]).

Exercises

Exercise 10.1: Prove that the roots of a polynomial $A(\lambda, t)$ in t with coefficients parameterized continuously by λ vary continuously with λ. □

§11. A Multivariate Root Bound

Let $\Sigma = \{A_1, \ldots, A_n\} \subseteq \mathbb{Z}[X_1, \ldots, X_n]$ be a system of n polynomials, not necessarily homogeneous. Suppose Σ has finitely many complex zeros and $(\xi_1, \ldots, \xi_n) \in \mathbb{C}^n$ is one of these zeros. We provide bounds on $|\xi_i|$. Assume $d_i = \deg(A_i)$ and let

$$K := \max\{\sqrt{n+1}, \max\{\|A_i\|_2 : i = 1, \ldots, n\}\}. \tag{59}$$

The main result of this section is:

Theorem 45 *If $(\xi_1, \ldots, \xi_n)$ is a zero of Σ and $|\xi_i| \neq 0$, $i = 1, \ldots, n$, then*

$$|\xi_i| > (2^{3/2} N K)^{-D} 2^{-(n+1) d_1 \cdots d_n}$$

where

$$N := \binom{1 + \sum_{i=1}^n d_i}{n}, \qquad D := \left(1 + \sum_{i=1}^n \frac{1}{d_i}\right) \prod_{i=1}^n d_i. \tag{60}$$

Before we present the proof, it is worthwhile understanding the basic order of growth of this root bound. First, we claim:

$$N = (c\widetilde{d})^n, \qquad (\text{for some } \; 1 < c < e), \tag{61}$$

where $\widetilde{d} = (\sum_{i=1}^n d_i)/n$ is the "average degree" and $e = 2.718\ldots$ is the base of the natural logarithm. To see this, it is easy to check that $N > (\widetilde{d})^n$. Conversely, we see that $N < (\sum_{i=1}^n d_i)^n/(n!)$ for $n \geq 3$. Since $n! > (n/e)^n$, this implies $N < ((\sum_{i=1}^n d_i)e/n)^n = (\widetilde{d}e)^n$. In fact, the last inequality

holds even when $n = 2$. This verifies our claim (61). We next want to show that $N \gg D$. More precisely,

$$D = o(N). \tag{62}$$

First let us write $D = C\prod_{i=1}^n d_i$ for some $1 < C \leq n+1$. Note that $\prod_{i=1}^n d_i \leq \widetilde{d}^n$ with equality iff all d_i's are equal. [This is easily shown by induction on n.] It follows that $D/N \leq C/c^n \to 0$ as $n \to \infty$, as we wanted to show in (62). Our bound can be simplified to

$$|\xi_i| > \left(2^{5/2}(\widetilde{d}e)^n K\right)^{-(n+1)(\prod_{i=1}^n d_i)}.$$

In case all the d_i's are equal to d, this gives

$$|\xi_i| > \left(2^{5/2}(de)^n K\right)^{-(n+1)d^n}.$$

In terms of bit-sizes, we take logs, leading to a single exponential bit-size bound.

Basic Approach. The strategy is to express $1/\xi_i$ as the root of a suitably specialized U-resultant. As in the previous section, we first homogenize Σ to $\widehat{\Sigma} = \{\widehat{A_1}, \ldots, \widehat{A_n}\} \subseteq \mathbb{Z}[X_0, X_1, \ldots, X_n]$, then define $\Sigma(\lambda)$, as in (50). Let $R(U_0, U_1, \ldots, U_n) = R(\mathbf{U})$ be the generalized U-resultant of $\Sigma(\lambda)$: this is the tail coefficient of the Macaulay resultant of $\widehat{\Sigma} \cup \{A_0\}$ where $A_0 = U_0X_0 + \cdots + U_nX_n$. Let $\sigma = \sigma_i$ be the $\mathbf{U}$-specialization such that

$$\sigma(U_\ell) = \sigma_i(U_\ell) = \begin{cases} -U_i & \text{if } \ell = i, \\ 1 & \text{if } \ell = 0, \\ 0 & \text{else.} \end{cases} \tag{63}$$

By lemma 44, with $W = \{i\}$, the polynomial $R_W(\mathbf{U})|_\sigma$ is non-vanishing. Moreover, $1/\xi_i$ is a root of $R_W(\mathbf{U})|_\sigma$. We need a bound on multivariate polynomials[12] from Mahler [126], stated here without proof:

Lemma 46 *Let $A, B \in \mathbb{C}[X_1, \ldots, X_n]$. Suppose $B|A$ and the degree of A in X_i is δ_i. Then $\|B\|_1 \leq 2^{\delta_1+\cdots+\delta_n}\|A\|_1$.*

Since $R_W|R$ and the total degree of R is $d_1d_2\cdots d_n$, we conclude

$$\|R_W\|_1 \leq 2^{(n+1)d_1d_2\cdots d_n}\|R\|_1.$$

It follows that

$$\|(R_W)_\sigma\|_1 \leq 2^{(n+1)d_1d_2\cdots d_n}\|R\|_1. \tag{64}$$

Hence theorem 45 follows from Cauchy's root bound and the following:

Lemma 47 *The 1-norm, $\|R(\mathbf{U})\|_1$, is at most $(2^{3/2}NK)^{D-1}$.*

Bounding $R(\mathbf{U})$. Recall that $R(\mathbf{U})$ is the tail coefficient of the exact quotient $\chi^M(\lambda)/\chi^L(\lambda)$ for a suitable Macaulay matrix M, with L a submatrix of M (cf. (40)). *Let us assume that $R(\mathbf{U})$ is,*

[12] Note that the k-norms $\|A\|_k$ for univariate polynomials extend to multivariate polynomials in the obvious way. In particular, when $k = 1$ this is just the sum of the absolute values of the coefficients of A.

in fact, the constant term of χ^M/χ^L, since we will see that this is the worst case for our bounding method. In fact, one can easily verify that (see §7) M is $N \times N$ and L is $(N-D) \times (N-D)$. Write

$$\chi^M(\lambda, \mathbf{U}) = \sum_{i=0}^{N} a_i(\mathbf{U})\lambda^i.$$

Each $a_i(\mathbf{U})$ is the sum of all the principal minors of M of order $N-i$. There are $\binom{N}{i}$ such minors, so that we may write $a_i(\mathbf{U}) = \sum_{j=1}^{\binom{N}{i}} a_{i,j}(\mathbf{U})$. Now we need a generalization of the Goldstein-Graham Hadamard bound (§VI.8) to multivariate polynomials.

Lemma 48 *Let $M(\mathbf{U})$ be a square matrix whose (i,j)th entry $M_{i,j}(\mathbf{U})$ is a polynomial in $\mathbb{C}[\mathbf{U}]$. If $W = [w_{i,j}]_{i,j}$ is the matrix where $w_{i,j} = \|M_{i,j}(\mathbf{U})\|_1$ then $\|\det M(\mathbf{U})\|_2 \leq H(W)$ where $H(W)$ is the product of the 2-norms of the rows of W.*

Proof. Suppose $\det(M(\mathbf{U})) = \sum_\alpha c_\alpha \mathbf{U}^\alpha$ where $\alpha = (\alpha_0, \ldots, \alpha_n)$ ranges over a suitable subset of $\mathbb{Z}^{n+1}$ and $c_\alpha \in \mathbb{C}$. Then

$$\begin{aligned}
\|\det M(\mathbf{U})\|_2^2 &= \sum_\alpha |c_\alpha|^2 \qquad \text{(by definition)} \\
&= \int_0^1 dt_1 \cdots \int_0^1 |\det(M(e^{2\pi \mathbf{i} t_1}, \ldots, e^{2\pi \mathbf{i} t_n})|^2 \qquad \text{(Parseval's identity)} \\
&\leq \int_0^1 dt_1 \cdots \int_0^1 H(W)^2 \qquad \text{(Hadamard's bound)} \\
&= H(W)^2.
\end{aligned}$$

Q.E.D.

To apply this bound, we construct the "W-matrix" corresponding to M: let $W := [w_{ij}]_{i,j=1}^N \in \mathbb{Z}^{N \times N}$ where w_{ij} is the 1-norm of the (i,j)th entry of M. If the non-zero entries in a row of W are the coefficients from a polynomial in Σ, then the 2-norm of the row is bounded by K, by assumption. If the row arises from the polynomial A_0 then its 2-norm is $\sqrt{n+1}$, which is also $\leq K$. Let $a_{i,j}(\mathbf{U})$ be one of the minors of M. If $W_{i,j}$ is the corresponding submatrix of W, and $H(W_{i,j})$ denotes (as in §VI.8) the product of the 2-norms of all the rows of $W_{i,j}$, then the generalized Hadamard bound says that

$$\|a_{i,j}(\mathbf{U})\|_2 \leq H(W_{i,j}) \leq K^{N-i}.$$

Since $a_i(\mathbf{U})$ has degree at most $N-i$, we obtain

$$\|a_{i,j}(\mathbf{U})\|_1 \leq \|a_{i,j}(\mathbf{U})\|_2 \sqrt{N-i+1} \leq K^{N-i}\sqrt{N-i+1}.$$

The first inequality is a general relation between 1-norms and 2-norms (§0.10). Next, since generally

$$\|a(\mathbf{U}) + b(\mathbf{U})\|_1 \leq \|a(\mathbf{U})\|_1 + \|b(\mathbf{U})\|_1,$$

we infer that

$$\|a_i(\mathbf{U})\|_1 \leq \binom{N}{i} K^{N-i}\sqrt{N-i+1}.$$

Similarly, if we write

$$\chi^L(\lambda) = \sum_{i=0}^{N-D} b_i(\mathbf{U})\lambda^i.$$

then we may deduce that

$$\|b_i(\mathbf{U})\|_1 \leq \binom{N-D}{i} K^{N-D-i}\sqrt{N-D-i+1}.$$

Now R is equal (see §9) to the determinant of the following $(D+1)\times(D+1)$ matrix:

$$M_0 := \begin{bmatrix} a_N & a_{N-1} & a_{N-2} & \cdots & a_{N-D} \\ b_{N-D} & b_{N-D-1} & b_{N-D-2} & \cdots & b_{N-2D} \\ & b_{N-D} & b_{N-D-1} & \cdots & b_{N-2D+1} \\ & & \ddots & & \vdots \\ & & & b_{N-D} & b_{N-D-1} \end{bmatrix}. \tag{65}$$

Note that $a_N = b_{N-D} = 1$. Instead of using the generalized Hadamard bound again, we now use the bound

$$\|R\|_1 \leq \mathtt{per}(W_D)$$

where $\mathtt{per}(W_D)$ denotes[13] the permanent of W_D, and W_D is analogous to the W-matrix of M_0 (65):

$$W_D := \begin{bmatrix} 1 & \binom{N}{1}K\sqrt{2} & \binom{N}{2}K^2\sqrt{3} & \cdots & \binom{N}{D-1}K^{D-1}\sqrt{D} & \binom{N}{D}K^D\sqrt{D+1} \\ 1 & \binom{N}{1}K\sqrt{2} & \binom{N}{2}K^2\sqrt{3} & \cdots & \binom{N}{D-1}K^{D-1}\sqrt{D} & \binom{N}{D}K^D\sqrt{D+1} \\ & 1 & \binom{N}{1}K\sqrt{2} & \cdots & \binom{N}{D-2}K^{D-2}\sqrt{D-1} & \binom{N}{D-1}K^{D-1}\sqrt{D} \\ & & \ddots & & & \vdots \\ & & & 1 & \binom{N}{1}K\sqrt{2} & \binom{N}{2}K^2\sqrt{3} \\ & & & & 1 & \binom{N}{1}K\sqrt{2} \end{bmatrix}. \tag{66}$$

Note that the first two rows of W_D are identical. It remains to evaluate this permanent. If we delete the first row and first column of W_D, we obtain the following $D\times D$ matrix

$$W'_D := \begin{bmatrix} \binom{N}{1}K\sqrt{2} & \binom{N}{2}K^2\sqrt{3} & \binom{N}{3}K^3\sqrt{4} & \cdots & \binom{N}{D-1}K^{D-1}\sqrt{D} & \binom{N}{D}K^D\sqrt{D+1} \\ 1 & \binom{N}{1}K\sqrt{2} & \binom{N}{2}K^2\sqrt{3} & \cdots & \binom{N}{D-2}K^{D-2}\sqrt{D-1} & \binom{N}{D-1}K^{D-1}\sqrt{D} \\ & 1 & \binom{N}{1}K\sqrt{2} & \cdots & \binom{N}{D-3}K^{D-3}\sqrt{D-2} & \binom{N}{D-2}K^{D-2}\sqrt{D-1} \\ & & \ddots & & & \vdots \\ & & & 1 & \binom{N}{1}K\sqrt{2} & \binom{N}{2}K^2\sqrt{3} \\ & & & & 1 & \binom{N}{1}K\sqrt{2} \end{bmatrix}. \tag{67}$$

If we expand the permanent of W_D by the first column, we obtain

$$\mathtt{per}(W_D) = 2\cdot\mathtt{per}(W'_D). \tag{68}$$

But if we divide the first row of W'_D by $NK\sqrt{2}$ we see that the resulting row is component-wise upper-bounded by the first row of W_{D-1}. This is because the ith entry in the resulting row is

$$\frac{\binom{N}{i}K^i\sqrt{i+1}}{NK\sqrt{2}}$$

[13] The permanent of a n-square matrix $[a_{ij}]_{i,j=1}^n$ is defined rather like as a determinant, being a sum of $n!$ terms. Each term in the permanent has the form $+\prod_{i=1}^n a_{i,\pi(i)}$ where π is a permutation of $\{1,2,\ldots,n\}$. For instance, $\mathtt{per}(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = ad+bc$. Note that each term is given a positive sign; in contrast, for determinants, we may attach a negative sign to the term.

and this is upper bounded by the ith entry of the first row of W_{D-1}, thus:

$$\frac{\binom{N}{i}K^i\sqrt{i+1}}{NK\sqrt{2}} \leq \binom{N}{i-1}K^{i-1}\sqrt{i}.$$

It follows that that $\mathtt{per}(W'_D) \leq (NK\sqrt{2})\mathtt{per}(W_{D-1})$. Substituted into (68), we obtain the recurrence $\mathtt{per}(W_D) = (2^{3/2}NK)\mathtt{per}(W_{D-1})$. Since $\mathtt{per}(W_2) = 2^{3/2}NK$, we obtain

$$\mathtt{per}(W_D) \leq (2^{3/2}NK)^{D-1}.$$

This concludes our proof of lemma 47.

Upper Bound on Roots. It is clear that we can also make ξ_i a root of a suitable specialization of $R_W(\mathbf{U})$, and this specialization has the same bound on its 1-norm as in (64). Therefore:

Corollary 49 *With the notation of theorem 45, we also have*

$$|\xi_i| < (2^{3/2}NK)^D 2^{(n+1)d_1\cdots d_n}$$

Remarks. Our basic approach is similar to Canny's [36]. Instead of using Macaulay's formula, he uses the division scheme (§9). Canny only treats the special where $R(\mathbf{U})$ is the standard U-resultant.

Exercises

Exercise 11.1: Simplify the above root bounds but give sharper estimates than in the text. E.g., use the Stirling approximation of Robbins: $n! = (n/e)^n\sqrt{2\pi n}e^\alpha$ where $(12n+1)^{-1} < \alpha < (12n)^{-1}$. A simpler approximation is $n! = \Theta((n/e)^{n+(1/2)})$. □

Exercise 11.2: Assuming the case where the generalized U-resultant is the standard U-resultant, improve our root bound above. □

Exercise 11.3: (Canny) Define the *Z-height* of a polynomial $A(X) = \sum_{i=0}^m a_i X^i$, $a_m \neq 0$, to be

$$h_Z(A) := \max\left\{\frac{|a_{m-1}|}{|a_m|}, \sqrt{\frac{|a_{m-2}|}{|a_m|}}, \sqrt[3]{\frac{|a_{m-3}|}{|a_m|}}, \ldots, \sqrt[m]{\frac{|a_0|}{|a_m|}}\right\}.$$

Alternatively, $h_Z(A)$ is the smallest value β such that $\beta^i \geq |a_{m-i}/a_m|$ for all $i = 1, \ldots, m$. [14] In the following proofs, it is often easiest to assume the polynomials are already monic, since this does not affect the Z-height.
(i) If $K = \|M\|_\infty$ where of a matrix $M \in \mathbb{R}^{n\times n}$, then $h_Z(\chi^M(\lambda)) \leq nK$.
(ii) $h_Z(AB) \leq h_Z(A) + h_Z(B)$.
(iii) Assuming A, B are monic, $h_Z(A+B) \leq h_Z(A) + h_Z(B)$.
(iv) $h_Z(C) \leq h_Z(A) + 2h_Z(B)$ where $C(X)$ is the quotient of $A(X)$ divided by $B(X)$
(v) Bound $h_Z(\widehat{R})$ and conclude with an alternative multivariate root bound. □

[14] This terminology recalls the notion of "height" for a polynomial. The "Z" here refers to a lemma attributed to Zassenhaus (§VI.2) showing that if $A(\alpha) = 0$ then $|\alpha| < 2h_Z(A)$.

§A. APPENDIX A: Power Series

We introduce the ring of power series over a ring R and describe some basic properties, mainly to facilitate the proof in appendix B. In particular, we describe what it means to form infinite sums and products.

A *(formal) power series* in the variable Z and with coefficients from a ring R is a formal infinite sum of the form

$$G(Z) = \sum_{i=0}^{\infty} a_i Z^i, \qquad (a_i \in R).$$

Let $R[[Z]]$ denote the set of such power series. These power series are said to be "formal" because, unlike their use in mathematical analysis, we do not evaluate them to a definite value and so there is no discussion of convergence properties. Two power series are deemed equal if their corresponding coefficients are equal. The polynomial ring $R[Z]$ is embedded in $R[[Z]]$ in a natural way. We can add and multiply two power series in the obvious term-wise fashion. Hence $R[[Z]]$ is a ring.

In the rest of this appendix, assume that R is a domain D. We proceed to state some easily verifiable properties of $D[[Z]]$. Then it is easily seen that $G(Z)H(Z) = 0$ iff $G(Z) = 0$ or $H(Z) = 0$. This shows that $D[[Z]]$ is a domain, with 0 and 1 as the zero and unity elements. The *inverse* of $G(Z)$ is the power series $H(Z)$ such that $G(Z)H(Z) = 1$. As in any domain, inverses are unique when they exist. The inverse of G is denoted by $H(Z) = 1/G(Z)$ or $G^{-1}(Z)$. It is not hard to see that $G(Z)$ has an inverse iff a_0 is a unit in D: this amounts to an inductive construction of each of the coefficients of $G^{-1}(Z)$.

*Formal differentiation*differentiation of $G(Z)$ is also defined,

$$G'(Z) = \sum_{i \geq 1} i \cdot a_i Z^{i-1}.$$

The usual rules of differentiation hold. For instance, the *product rule* can be verified:

$$(G(Z)H(Z))' = G'(Z)H(Z) + G(Z)H'(Z).$$

We now derive an important formula using differentiation. If A is a finite product of power series,

$$A(Z) = \prod_{i=1}^{k} G_i(Z)$$

then extension of the above product rule gives

$$A'(Z) = \sum_{i=1}^{k} G_i'(Z) \prod_{j \neq i} G_j(Z).$$

Assuming that $A(Z)$ is invertible, then each $G_i(Z)$ is also seen to be invertible. So we can divide both sides by $A(Z)$ to give the "logarithmic derivative formula" for $A(Z)$:

$$\frac{A'(Z)}{A(Z)} = \sum_{i=1}^{k} \frac{G_i'(Z)}{G_i(Z)}. \tag{69}$$

Recall that $A'(Z)/A(Z)$ is the "logarithmic derivative" of $A(Z)$ (§VI.1). The power of this formula to turn a product into a sum will be useful.

The smallest i such that $a_i \neq 0$ is called the *order* of $G(Z)$, denoted $\texttt{ord}(G)$. When G is a polynomial, $\texttt{ord}(G)$ is just its tail degree (§0.10). Hence we will call a_i the *tail coefficient* of G when $i = \texttt{ord}(G)$. By definition, $\texttt{ord}(0) = \infty$. It is easily seen that $\texttt{ord}(G \cdot H) = \texttt{ord}(G) + \texttt{ord}(H)$ and $\texttt{ord}(G \pm H) \geq \min\{\texttt{ord}(G), \texttt{ord}(H)\}$ with equality whenever $\texttt{ord}(G) \neq \texttt{ord}(H)$. The theory of divisibility is extremely simple for power series: $G|H$ iff $\texttt{ord}(G) \leq \texttt{ord}(H)$ and the tail coefficient of G divides the tail coefficient of H.

Infinite Sums and Products. The domain $D[[Z]]$ only guarantees that finite sums and finite products are well-defined. The fact that a power series is an infinite sum will tempt us to use infinite sums and products. Especially in combinatorial enumeration, infinite sums and infinite products are extremely useful. Clearly they are not always well-defined. Our goal here is to define an unambiguous meaning for the following infinite product and sum

$$P = \prod_{i=0}^{\infty} G_i(Z), \qquad S = \sum_{i=0}^{\infty} G_i(Z), \qquad (G_i(Z) \in D[[Z]]). \tag{70}$$

Towards this end, for each $n \geq 1$, we define $G(Z)$ and $H(Z)$ to be *equivalent modulo* Z^n if Z^n divides $G(Z) - H(Z)$. This is written

$$G(Z) \equiv H(Z) \pmod{Z^n}.$$

Note that $G = H$ iff for all $n \geq 1$,

$$G \equiv H \pmod{Z^n}. \tag{71}$$

We also write

$$G(Z) \,\mathbf{mod}\, Z^n$$

for the polynomial obtained by eliminating all terms in $G(Z)$ that are divisible by Z^n. So $G(Z) \,\mathbf{mod}\, Z^n$ is a polynomial of degree at most $n-1$ and $G \equiv H \pmod{Z^n}$ iff $(G \,\mathbf{mod}\, Z^n) = (H \,\mathbf{mod}\, Z^n)$.

Definition 1
(a) The infinite product P in equation (70) is defined *if for all $n \geq 1$, there exists a finite bound $b(n)$ such that for all $i > b(n)$, $G_i(Z) \,\mathbf{mod}\, Z^n = 1$. In this case P is the power series $P(Z) \in D[[Z]]$ such that for each $n \geq 0$,*

$$P(Z) \equiv \prod_{i=0}^{b(n)} G_i(Z) \pmod{Z^n}. \tag{72}$$

(b) The infinite sum S in equation (70) is defined *if for all $n \geq 1$, there exists a finite bound $b(n)$ such that for all $i > b(n)$, $G_i(Z) \,\mathbf{mod}\, Z^n = 0$. In this case S is the power series $S(Z) \in D[[Z]]$ such that for each $n \geq 0$,*

$$S(Z) \equiv \sum_{i=0}^{b(n)} G_i(Z) \pmod{Z^n}. \tag{73}$$

Clearly $P(Z), S(Z)$ are unique when defined; in particular, they do not depend on the choice of the bounds $b(n)$.

Lemma 50 *If $\prod_{i=0}^{\infty} G_i(Z)$ is defined and $H_0, H_1, \ldots$ is any rearrangement of the factors $G_0, G_1, \ldots$ then $\prod_{i=0}^{\infty} H_i(Z)$ is defined and equal to $\prod_{i=0}^{\infty} G_i(Z)$.*

The proof is left as an exercise. In view of this rearrangement lemma, we can define the product of a set of power series: if $W = \{G_0, G_1, \ldots\} \subseteq D[[Z]]$, the *product* $\prod W$ or $\prod_{G \in W} G$ can be defined to be $\prod_{i=0}^{\infty} G_i$, provided this is defined. We now extend the logarithmic derivative formula to infinite products:

Lemma 51 *If $P(Z) = \prod_{i=0}^{\infty} G_i(Z)$ is defined, and $P(Z)$ is invertible then*

$$\frac{P'(Z)}{P(Z)} = \sum_{i=0}^{\infty} \frac{G_i'(Z)}{G_i(Z)}. \tag{74}$$

Proof. First we show that the infinite sum in equation (74) is defined. Since $P(Z)$ is invertible, we conclude that G_i is also invertible for all i [In proof: for any n, if $i > b(n)$ then the constant term in G_i is 1, and if $i \leq b(n)$ then equation (72) shows that the constant term in G_i is invertible.] Next for each $n \geq 1$ and $i > b(n)$, $G_i(Z) \bmod Z^n = 1$ implies $G_i'(Z) \bmod Z^{n-1} = 0$; then $G_i'(Z)/G_i(Z) \bmod Z^{n-1} = 0$, and so the infinite sum of equation (74) is defined.

It remains to show that the infinite sum equals $P'(Z)/P(Z)$. For each n, equation (72) holds; applying the logarithmic derivative formula yields

$$\frac{P'(Z)}{P(Z)} \equiv \sum_{i=0}^{b(n)} \frac{G_i'(Z)}{G_i(Z)} \pmod{Z^{n-1}}. \tag{75}$$

On the other hand, if $S(Z)$ is defined to be the infinite sum $\sum_{i=0}^{\infty} G_i'(Z)/G_i(Z)$, then since $G_i'(Z)/G_i(Z) \bmod Z^{n-1} = 0$ for $i > b(n)$, we conclude that $S(Z) \equiv \sum_{i=0}^{b(n)} \frac{G_i'(Z)}{G_i(Z)} \pmod{Z^{n-1}}$. This, with equation (75), shows that $S(Z) \equiv P'(Z)/P(Z) \pmod{Z^{n-1}}$ for all $n \geq 0$. So $S(Z) = P'(Z)/P(Z)$. **Q.E.D.**

Exercises

Exercise A.1: Verify the unproved assertions about $D[[Z]]$. □

Exercise A.2:
(a) If D is a field then $D[[Z]]$ is an Euclidean domain (§II.3) with $\varphi(G) = \texttt{ord}(G)$.
(b) If D is a UFD, is $D[[Z]]$ a UFD? □

Exercise A.3: Show lemma 50. □

Exercise A.4: If $F_1 \equiv F_2$ and $G_1 \equiv G_2$ then $F_1 \pm G_1 \equiv F_2 \pm G_2$ and $F_1 G_1 \equiv F_2 G_2$, all modulo Z^n. □

§B. APPENDIX B: Counting Irreducible Polynomials

Let D be a domain. We prove the following result:

Theorem 52 *$D[X]$ has infinitely many non-associated irreducible polynomials.*

This result is used in §2. It is also of independent interest for applications in coding theory and cryptography. In case D is an infinite set, this theorem is immediate since the set $\{X - a : a \in D\}$ is clearly an infinite set of pairwise non-associated irreducible polynomials. Henceforth, let us assume that D is finite. In fact, we have: *a finite domain D is a field.* To see this, for any non-zero $a \in D$, consider the sequence

$$1, a, a^2, \ldots, a^r, a^{r+1}, \ldots, a^{r+s}$$

where a^{r+s} is the first time that the sequence repeats, and $a^{r+s} = a^r$. Then $a^r(1 - a^s) = 0$ and so $a^s = 1$. Hence a^{s-1} is the inverse of a. So D must be a Galois field with q elements where q is a prime power.

Assuming we have picked a distinguished element in each equivalence class of associates of $D[X]$ (§II.1), our problem amounts to counting the number of distinguished irreducible polynomials in $D[X]$. We follow Berlekamp's approach [22] of using generating functions. This approach yields much more information than just the stated theorem.

Generating Functions. The power series $G(Z) = \sum_{i \geq 0} a_i Z^i$ is called a *generating function* for the sequence.

$$a_0, a_1, a_2, \ldots, \qquad (a_i \in D).$$

For any set $S \subseteq D[X]$, the *counting function* $G_S(Z)$ for S is the generating function for the sequence

$$|S_0|, |S_1|, |S_2|, \ldots$$

where $S_i \subseteq S$ is the set elements in S of degree i. Let I be the set of distinguished irreducible polynomials in $D[X]$. According to the convention in §II.1, the distinguished polynomials are precisely the monic ones. Denote the counting function of I by

$$G_I(Z) := \sum_{i=0}^{\infty} d_i Z^i. \tag{76}$$

Our original goal of showing that I is infinite follows from the following stronger claim:

Theorem 53 *For all $i \geq 0$, $d_i \geq 1$.*

Towards a proof of theorem 53, note that if $f \in D[X]$ has degree $d \geq 0$ then

$$H_d(Z) := \frac{1}{1 - Z^d} = 1 + Z^d + Z^{2d} + \cdots$$

is the counting function for the set

$$S(f) := \{f^i : i = 0, 1, 2, \ldots\}.$$

Next, if $f, g \in D[X]$ are relatively prime and have degrees $d, e \geq 0$, then similarly

$$H_d(Z)H_e(Z) = \frac{1}{1 - Z^d} \cdot \frac{1}{1 - Z^e}$$

is the counting function for the set

$$S(f, g) := \{f^i g^j : i, j = 0, 1, 2, \ldots\}.$$

Note that this remark depends on the fact that $D[X]$ is a UFD (recall that D is a Galois field). By induction on k, if $f_1, \ldots, f_k$ are pairwise relatively prime of degrees $d_1, \ldots, d_k \geq 0$ then

$$H_{d_1}(Z)H_{d_2}(Z)\cdots H_{d_k}(Z) = \prod_{i=1}^{k} \frac{1}{1-Z^{d_i}}$$

is the counting function for the set

$$S(f_1, \ldots, f_k)$$

of power products of $f_1, \ldots, f_k$. Now we make an essential leap to the infinite case:

Lemma 54 *Let $S(I)$ be the set of power products of elements in I. Its counting function $G_{S(I)}$ is given by*

$$G_{S(I)} = \prod_{u\in I} H_{\deg(u)}(Z) \tag{77}$$

$$= \prod_{i=0}^{\infty} (H_i(Z))^{d_i}. \tag{78}$$

Proof. It is easy to see that the infinite product

$$P = \prod_{u\in I} H_{\deg(u)}(Z) = \prod_{u\in I} \frac{1}{1-Z^{\deg(u)}}.$$

is defined (cf. lemma 50). We can verify routinely that each coefficient of P is equal to the corresponding coefficient of $G_{S(I)}$, by looking at a finite number of terms in P. Since there are d_i (see equation (76)) elements in I of degree i, P can also be written as in equation (78) (this also needs verification). **Q.E.D.**

Note that the set $S(I)$ is precisely the set of distinguished elements in $D[X]$. On the other hand, the number of distinguished elements of degree n is just $|D|^n = q^n$. Thus the counting function of $S(I)$ is equal to

$$G_{S(I)} = 1 + qZ + q^2Z^2 + \cdots = \frac{1}{1-qZ}.$$

Combining this equation for $G_{S(I)}(Z)$ with the previous lemma, we have shown:

Lemma 55 *The counting function for the set of distinguished elements in $D[X]$ is*

$$G_{S(I)}(Z) = \prod_{i=0}^{\infty} \left(\frac{1}{1-Z^i}\right)^{d_i} \tag{79}$$

$$= \frac{1}{1-qZ}. \tag{80}$$

Taking the logarithmic derivative $G'_{S(I)}(Z)/G_{S(I)}(Z)$ of the two expressions in the lemma and equating them,

$$\frac{q}{1-qZ} = \sum_{i=0}^{\infty} d_i \frac{i\cdot Z^{i-1}}{1-Z^i},$$

$$\begin{aligned}
\frac{qZ}{1-qZ} &= \sum_{i=0}^{\infty} d_i \frac{i \cdot Z^i}{1-Z^i} \\
&= \sum_{i=0}^{\infty} i \cdot d_i \left(\sum_{j \geq 1} Z^{ji} \right) \\
&= \sum_{i=0}^{\infty} i \cdot d_i \left(\sum_{n \geq 1; i|n} Z^n \right) \qquad \text{(putting } n = ji\text{)} \\
&= \sum_{n=1}^{\infty} Z^n \left(\sum_{i \geq 1; i|n} i \cdot d_i \right).
\end{aligned}$$

Equating coefficients of Z^n, we have

Theorem 56 *The number d_i of distinguished irreducible polynomials of degree i in $D[X]$ satisfies the equation*

$$q^n = \sum_{i \geq 1; i|n} i \cdot d_i, \qquad q = |D|. \tag{81}$$

This theorem immediately shows that $d_1 = q$ (interprete this). Also, $d_2 = \frac{q^2-q}{2} \geq 1$. For general d_n, we have the following upper and lower bounds:

Corollary 57

$$\frac{q^n - q^{1+(n/2)}}{n} \leq d_n \leq \frac{q^n - q}{n}. \tag{82}$$

Proof. For the upper bound, we ignore all but the first and last terms in the sum in equation (81), giving $q^n \geq n \cdot d_n + d_1$. For the lower bound, we extract the term involving d_n in equation (81) and note that the remaining terms involve d_i for $i \leq n/2$:

$$q^n \leq n \cdot d_n + \sum_{i=1}^{n/2} i d_i < n \cdot d_n + \sum_{i=1}^{n/2} q^i < n \cdot d_n + q^{1+(n/2)}.$$

Q.E.D.

Now theorem 53 is immediate since the lower bound in equation (82) shows that $d_n > 0$ for $n \geq 3$. This corollary also shows that

$$\left| d_n - \frac{q^n}{n} \right| = o\left(\frac{q^n}{n} \right).$$

Exercises

Exercise B.1: (Nijmeijer and Staring, 1988) For any prime p, $X^p - X - 1$ is irreducible in $\mathbb{Z}_p[X]$. □

Exercise B.2:
(a) Apply Möbius inversion to the equation (81).
(b) Use the inversion formula to compute d_n for $n = 1, \ldots, 10$, when $D = GF(2)$. □

Exercise B.3: Count the number of irreducible polynomials in $D[X, Y]$ using the same method. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture XII
Gröbner Bases

The subject of *Gröbner bases* or *standard bases* has generated considerable recent interest. Hironaka (Thesis, Harvard 1963) introduced the concept of standard bases for ideals in the ring of power series; independently, Buchberger (Thesis, Innsbruck 1965) gave an algorithm to construct a rather similar notion in polynomial rings. Buchberger coined the name Gröbner bases and popularized it [32]. Ideas germane to Gröbner bases appeared independently in the work of numerous authors including Richmond, Sims, Spear, Trinks and Zacharias. Gröbner bases are special generators for ideals with amiable computational properties. Many problems in polynomial algebra can be solved using such bases. The basic algorithm to construct such bases (in polynomial rings over a field) can be viewed as a simultaneous generalization of the Euclidean algorithm (for two univariate polynomials) and Gaussian elimination (for a system of linear multivariate polynomials). In the term rewriting literature, the algorithm is a special case of the Knuth-Bendix procedure. It is related to the "straightening law" of invariant theory and to the Wu-Ritt procedure (see [216]). These connections give a hint of the richness of these ideas. Generalizations and extensions of Gröbner bases to more general rings have been studied. This is only a brief introduction, as there are many extensions of these ideas (Gröbner bases for integer polynomials, universal Gröbner basis, special classes of ideals such as binomial ideals, etc). Our exposition is essentially an expansion of [140]. Some recent books on the subject include Becker and Weispfenning [17], Mishra [137], Cox, Little and O'Shea [50], Sturmfels [198], Adams and Loustaunau [1] and also a collection of research papers [170].

In this lecture, we fix an arbitrary field K and let $R = K[X_1, \ldots, X_n] = K[\mathbf{X}]$.

§1. Admissible Orderings

Let $\mathtt{PP} = \mathtt{PP}(X_1, \ldots, X_n)$ be the set of power products (§0.10) of $\mathbf{X}$. A partial ordering $\underset{\mathtt{A}}{\leq}$ on $\mathtt{PP}$ is *compatible* if for all $p, q, r \in \mathtt{PP}$, $p \underset{\mathtt{A}}{\leq} q$ implies $rp \underset{\mathtt{A}}{\leq} rq$. It is *semi-admissible* if it is a compatible total ordering. It is *admissible* if it is semi-admissible and $1 \underset{\mathtt{A}}{\leq} p$ for all $p \in \mathtt{PP}$. If $p \underset{\mathtt{A}}{\leq} q$ but $p \neq q$ then we write $p \underset{\mathtt{A}}{<} q$ or $q \underset{\mathtt{A}}{>} p$. Sometimes, a power product is called a *term* and an admissible orderings is also called a *term ordering*.

Example: Three important admissible orderings are the *(pure) lexicographic ordering* $\underset{\mathtt{LEX}}{\leq}$, *(total) degree ordering* $\underset{\mathtt{TOT}}{\leq}$, and *reverse lexicographic ordering* $\underset{\mathtt{REV}}{\leq}$. These orderings are completely specified once we choose a total ordering on the variables $X_1, \ldots, X_n$. Typically, we pick

$$X_1 \underset{\mathtt{A}}{>} X_2 \underset{\mathtt{A}}{>} \cdots \underset{\mathtt{A}}{>} X_n \tag{1}$$

where $\underset{\mathtt{A}}{>}$ is $\underset{\mathtt{LEX}}{>}$, $\underset{\mathtt{TOT}}{>}$ or $\underset{\mathtt{REV}}{>}$. Let

$$p = X_1^{d_1} X_2^{d_2} \cdots X_n^{d_n}, \qquad q = X_1^{e_1} X_2^{e_2} \cdots X_n^{e_n}.$$

(a) Define $p \underset{\mathtt{LEX}}{\geq} q$ if $p = q$ or else, the first non-zero element in the sequence

$$(d_1 - e_1, d_2 - e_2, \ldots, d_n - e_n) \tag{2}$$

is positive. Thus, writing (X, Y, Z) for (X_1, X_2, X_3), we have

$$1 \underset{\texttt{LEX}}{<} Z \underset{\texttt{LEX}}{<} Z^2 \underset{\texttt{LEX}}{<} \cdots \underset{\texttt{LEX}}{<} Y \underset{\texttt{LEX}}{<} YZ \underset{\texttt{LEX}}{<} Y^2 \cdots \underset{\texttt{LEX}}{<} \underset{\texttt{LEX}}{<} X \underset{\texttt{LEX}}{<} XZ \underset{\texttt{LEX}}{<} XY \underset{\texttt{LEX}}{<} X^2 \underset{\texttt{LEX}}{<} \cdots.$$

Note that there are an infinity of terms to be inserted at the three ellipses (describe them).

(b) Define $p \underset{\texttt{TOT}}{\geq} q$ if $\deg(p) > \deg(q)$ or else, $p \underset{\texttt{LEX}}{\geq} q$.

$$1 \underset{\texttt{TOT}}{<} Z \underset{\texttt{TOT}}{<} Y \underset{\texttt{TOT}}{<} X \underset{\texttt{TOT}}{<} Z^2 \underset{\texttt{TOT}}{<} YZ \underset{\texttt{TOT}}{<} Y^2 \underset{\texttt{TOT}}{<} XZ \underset{\texttt{TOT}}{<} XY \underset{\texttt{TOT}}{<} X^2 \underset{\texttt{TOT}}{<} \cdots.$$

(c) Define $p \underset{\texttt{REV}}{\geq} q$ iff either $\deg(p) > \deg(q)$ or else, the last non-zero element in the sequence (2) is negative.

$$1 \underset{\texttt{REV}}{<} Z \underset{\texttt{REV}}{<} Y \underset{\texttt{REV}}{<} X \underset{\texttt{REV}}{<} Z^2 \underset{\texttt{REV}}{<} YZ \underset{\texttt{REV}}{<} XZ \underset{\texttt{REV}}{<} Y^2 \underset{\texttt{REV}}{<} XY \underset{\texttt{REV}}{<} X^2 \underset{\texttt{REV}}{<} \cdots.$$

This ordering is less familiar but has important computational properties (Bayer and Stillman [12]). Suppose $\deg(p) = \deg(q)$ and we wish to decide if $p \underset{\texttt{REV}}{\geq} q$. If $\deg_{X_n}(p) < \deg_{X_n}(q)$ then declare $p \underset{\texttt{REV}}{\geq} q$; otherwise $\deg_{X_n}(p) = \deg_{X_n}(q)$ and we apply this rule recursively.

■

We need an oft-rediscovered little lemma of Dickson [55] (also attributed to Gordan).

Theorem 1 (Dickson's Lemma) *Given any subset $T \subseteq \texttt{PP}(X_1, \ldots, X_n)$, there is a finite subset $F \subseteq T$ such that every power product in T is a multiple of some element of F.*

Proof. We use induction on the number n of variables. If $n = 1$ then we let F consist of the unique term in T of minimum degree. Next assume $n > 1$. Pick any $p_0 \in T$ and say

$$p_0 = X_1^{e_1} X_2^{e_2} \cdots X_n^{e_n}.$$

Then every $m \in T$ that is not divisible by p_0 belongs to one of $\sum_{i=1}^n e_i$ different sets: let $i = 1, \ldots, n$ and $v = 0, 1, \ldots, e_i - 1$. Then the set $T_{i,v}$ consists of those monomials $m \in T$ such that $deg_{X_i}(m) = v$. Let $T'_{i,v}$ denote the set of monomials obtained by omitting the factor X_i^v from monomials in $T_{i,v}$. By inductive hypothesis, there exists finite subsets $F'_{i,v} \subseteq T'_{i,v}$ such that each monomial in $T'_{i,v}$ is a multiple of some monomial in $F'_{i,v}$. Let $F_{i,v}$ be the set $\{m \cdot X_i^v \ : \ m \in F'_{i,v}\}$. It is then clear that every monomial in T is a multiple of some monomial in the finite set

$$\{p_0\} \cup \bigcup_{i,v} F_{i,v}.$$

Q.E.D.

The set F in the lemma is known as a *Dickson basis* of T. Recall that a total ordering is *well-founded* if there is no infinite sequence of strictly decreasing elements. We show that any admissible ordering $\underset{\texttt{A}}{\leq}$ is well-founded. For, if $p_1 \underset{\texttt{A}}{>} p_2 \underset{\texttt{A}}{>} p_3 \underset{\texttt{A}}{>} \cdots$ then by Dickson's lemma, the set $T = \{p_1, p_2, \ldots, \}$ has a Dickson basis F. Without loss of generality, let $F = \{p_1, \ldots, p_k\}$ for some k. If T has more than k elements, then p_{k+1} is divisible by some $p_i \in F$. This contradicts $p_i \underset{\texttt{A}}{>} p_{k+1}$.

Corollary 2 *An admissible ordering is well-founded.*

Linear quasi-ordering on polynomials. We need to extend the admissible ordering $\underset{\mathtt{A}}{\leq}$ to polynomials. First we extend it to mononomials: if $c \cdot p$ and $d \cdot q$ ($c, d \in K; p, q \in \mathtt{PP}$) are two monomials, then we declare $c \cdot p \underset{\mathtt{A}}{\leq} d \cdot q$ if $p \underset{\mathtt{A}}{\leq} q$. Note that this is[1] a linear quasi-ordering. For any polynomial $p = c_1 p_1 + c_2 p_2 + \cdots + c_m p_m$, written as a sum of monomials with distinct p_i's, introduce the *term sequence of* p

$$\overline{p} = (p_1, \ldots, p_m), \tag{3}$$

where $p_1 \underset{\mathtt{A}}{\geq} p_2 \underset{\mathtt{A}}{\geq} \cdots \underset{\mathtt{A}}{\geq} p_m$. We compare two polynomials as follows: $p \underset{\mathtt{A}}{\leq} q$ iff in their term sequences $(p_1, \ldots, p_m)$ and $(q_1, \ldots, q_k)$, either (1) $m \leq k$ and $p_i = q_i$ for each $i = 1, \ldots, m$, or (2) for some $i \leq \min\{m, k\}$, $p_i \underset{\mathtt{A}}{<} q_i$ and $p_j = q_j$ for $j = 1, \ldots, i-1$.

It is easy to check that this defines a linear quasi-ordering for polynomials. From a linear quasi-ordering $\underset{\mathtt{A}}{\leq}$, we obtain a *strict* partial order $\underset{\mathtt{A}}{<}$ where $x \underset{\mathtt{A}}{<} y$ holds provided $x \underset{\mathtt{A}}{\leq} y$ holds but not $y \underset{\mathtt{A}}{\leq} x$. We leave it as an exercise to show:

Lemma 3 *There is no strictly descending infinite sequence of polynomials under the ordering* $\underset{\mathtt{A}}{\leq}$.

Weight Schemes. For a quantitative approach to admissible ordering, we introduce another view of such orderings. Define a *weight function* $w : \mathtt{PP} \to \mathbb{R}$ to be a real-valued function with the properties $w(1) = 0$ and $w(pq) = w(p) + w(q)$. To avoid triviality, we assume that $w(p) \neq 0$ for some $p \in \mathtt{PP}$. It is easy to see w is completely characterized by the n-vector

$$\overline{w} := (w(X_1), w(X_2), \ldots, w(X_n)). \tag{4}$$

Let us extend $\mathtt{PP}$ to the set $\overline{\mathtt{PP}}$ defined to consist of all rational power products of $X_1, \ldots, X_n$. Thus a typical element of $\overline{\mathtt{PP}}$ is $\prod_{i=1}^{n} X_i^{e_i}$ where $e_i \in \mathbb{Q}$. For instance, $X_1^{-3} X_2 X_4^{1/5}$. The definition of "weight function" extends to $\overline{\mathtt{PP}}$, *mutatis mutandis*.

Lemma 4 *For any weight function* $w : \mathtt{PP} \to \mathbb{R}$*, there is a unique extension of* w *to a weight function* $w' : \overline{\mathtt{PP}} \to \mathbb{R}$.

Proof. Let $p = X^e$ where $X \in \{X_1, \ldots, X_n\}$ and $e \in \mathbb{Q}$. It suffices to show that $w'(p)$ is uniquely determined by w, since every element in $\overline{\mathtt{PP}}$ is a product of such p's. If $e \in \mathbb{N}$ then $w'(p) = w(p)$. If $e \in \mathbb{Z} \setminus \mathbb{N}$ then $w'(p) + w'(p^{-1}) = w'(1) = 0$ implies $w'(p) = -w(X^{-e})$. If $e = c/d$ (where $c \in \mathbb{Z}$, $d \in \mathbb{N}$) then $d \cdot w'(p) = w'(p^d) = w'(X^c)$. Hence $w'(p) = w'(X^c)/d$. We have completely determined w'. **Q.E.D.**

In some sense, extending the concept of weight functions from $\mathtt{PP}$ to $\overline{\mathtt{PP}}$ does not add anything essential. We wish to identify $\mathtt{PP}$ with $\mathbb{N}^n$, and $\overline{\mathtt{PP}}$ with $\mathbb{Q}^n$. To freely switch between these two views as convenient, it is useful to introduce the analogue of the logarithm function (but still calling it "log"):

$$\begin{aligned} \log : \quad & \overline{\mathtt{PP}} \to \mathbb{Q}^n, \\ & \mathbf{X}^e \mapsto e. \end{aligned}$$

[1] A binary relation $\leq$ on a set S is a *quasi-ordering* if for all $x, y, z \in S$, we have $x \leq x$ and if $x \leq y$ and $y \leq z$ then $x \leq z$. This is like a partial-order except that the symmetry law is omitted: so $x \leq y$ and $y \leq x$ need not imply $x = y$. A quasi-ordering in which every two elements are comparable is called a *linear quasi-ordering*. Thus linear quasi-orderings satisfy the usual laws of linear orderings, save the symmetry law.

So a weight function $w : \mathbb{Q}^n \to \mathbb{R}$ can be expressed as $w(p) = \langle \overline{w}, \log(p) \rangle$ where $\overline{w}$ is given in (4) and $\langle \cdot, \cdot \rangle$ indicates inner product. A sequence

$$W = (w_1, \ldots, w_n) \tag{5}$$

of n weight functions is called a *weight scheme*. W can be viewed as a function $W : \mathbb{Q}^n \to \mathbb{R}^n$ where

$$W(p) = (w_1(p), \ldots, w_n(p)), \qquad p \in \mathbb{Q}^n.$$

The *weight matrix* $\overline{W}$ associated with W is the $n \times n$ real matrix whose ith row is $\overline{w_i} = (w_i(X_1), \ldots, w_i(X_n))$, then we may write $W(p)$ as the matrix-vector product

$$W(p) = \overline{W} \cdot \log(p).$$

Every weight scheme W (or its associated matrix $\overline{W}$) *induces* a partial order $\underset{\mathtt{W}}{\leq}$ on $\overline{\mathtt{PP}}$ as follows. It is determined by the relation:

$$p \underset{\mathtt{W}}{<} q \text{iff } \; W(p) \underset{\mathtt{LEX}}{<} W(q).$$

This is easily seen to be a partial ordering (e.g., transitivity of $\underset{\mathtt{W}}{\leq}$ follows from the transitivity of $\underset{\mathtt{LEX}}{\leq}$). Let us see some examples.

Example: (i) Let $\overline{W}$ be the $n \times n$ identity matrix I_n. Then W induces the pure lexicographic ordering $\underset{\mathtt{LEX}}{\leq}$

(ii) A weight matrix inducing the total degree ordering $\underset{\mathtt{TOT}}{\leq}$ is given by

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

(iii) A weight matrix inducing the reverse lexicographic ordering $\underset{\mathtt{REV}}{\leq}$ is given by

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

■

It should be noted that the ordering induced by W may only depend on a prefix of $(w_1, \ldots, w_n)$. For instance, if the entries of $\overline{w_1}$ are n transcendental elements linearly independent over $\mathbb{Q}$ then W does not depend on the $w_2, \ldots, w_n$. On the other hand, we are usually interested in $\overline{W}$ with integer entries in which case the function will necessarily depend on all w_i's.

A row operation on $\overline{W}$ is *admissible* if it is one of the following two types:

(1) Multiplying a row by a positive constant.

(2) Adding an arbitrary constant multiple of row i to row j for some $j > i$.

It is not hard to see that admissible row operations do not change the partial ordering induced by $\overline{W}$. Moreover, such operations are invertible. Say two weight matrices are *equivalent* if they are equal after a finite sequence of admissible row operations. We say W (or $\overline{W}$) is *non-negative* if all the entries of $\overline{W}$ are non-negative. We say W *semi-admissible* provided W is non-singular, *i.e.*,

$$W(p) = \mathbf{0} = (0, 0, \ldots, 0) \text{iff } \; p = 1. \tag{6}$$

Lemma 5 *Let $\underset{\mathtt{W}}{\leq}$ be the partial ordering on $\overline{\mathtt{PP}}$ induced by a weight scheme W.*

(i) *The partial ordering $\underset{\mathtt{W}}{\leq}$ is a compatible ordering.*

(ii) *If W is semi-admissible, then $\underset{\mathtt{W}}{\leq}$ is a semi-admissible ordering.*

(iii) *If W is non-negative and semi-admissible then $\underset{\mathtt{W}}{\leq}$ is an admissible ordering.*

(iv) *If $\underset{\mathtt{W}}{\leq}$ is admissible then the first non-zero entry of each column of $\overline{W}$ is positive. Moreover, $\overline{W}$ is equivalent a non-negative semi-admissible weight matrix.*

Proof. (i) We have already noted that $\underset{\mathtt{W}}{\leq}$ is a partial ordering. Compatibility follows from the linearity of weight functions: $W(p) \underset{\mathtt{LEX}}{\leq} W(q)$ implies $W(rp) = W(r) + W(p) \underset{\mathtt{LEX}}{\leq} W(r) + W(q) = W(rq)$.
(ii) $\underset{\mathtt{W}}{\leq}$ is a total ordering because $p \neq q$ implies $W(p) \neq W(q)$ (otherwise, by linearity, we get $W(p/q) = \mathbf{0}$ which contradicts $p/q \neq 1$).
(iii) We must show that $p \underset{\mathtt{W}}{\geq} 1$ for all $p \in \mathtt{PP}$. If W is non-negative, then $W(p) \underset{\mathtt{LEX}}{\geq} \mathbf{0} = W(1)$. Thus $p \underset{\mathtt{W}}{\geq} 1$.
(iv) If $\underset{\mathtt{W}}{\leq}$ is admissible, it is easy to see that the first non-zero entry of each column is non-negative. (Trivially, no columns can be entirely zero.) Let us say that a column "belongs to i" if the first non-zero entry of the column is in the ith row. Clearly the first row must be non-negative. So by adding a suitable multiple of the first row to the other rows, we can make all the columns belonging to 1 non-negative. At this point, row 2 becomes non-negative. We repeat this process with row 2, turning all the columns belonging to 2 non-negative. Notice that this does not affect the non-negativity of any column. Clearly we can finally turn the entire matrix non-negative. **Q.E.D.**

Robbiano has shown that (Exercise) every admissible ordering is induced by some weight scheme (see also [56]).

Normal Form. There are obvious parallels between admissible row operations and elementary row operations (§X.4). Let us say that a weight matrix is in *normal form* if (1) the rows are mutually orthogonal, and (2) in each non-zero row, the first non-zero entry is ± 1.

Theorem 6 (i) *Every weight matrix is equivalent to a normal form weight matrix. If the original matrix is rational, so is the normal form.*

(ii) *Two semi-admissible weight matrices with rational entries induce the same semi-admissible ordering on $\overline{\mathtt{PP}}$ iff they have the same normal form.*

Proof. (i) We can make the rows mutually orthogonal using the Gram-Schmidt procedure (§IX.1). It is trivial to divide each row by the absolute value of its first non-zero entry to ensure that the first non-zero entry is ± 1. This procedure preserves rationality of the matrix.
(ii) If two weight matrices have the same normal form, then clearly they induce the same partial ordering on $\overline{\mathtt{PP}}$. Conversely, suppose $\overline{W}, \overline{W'}$ are semi-admissible and rational, and the the orderings $\underset{\mathtt{W}}{\leq}, \underset{\mathtt{W'}}{\leq}$ that they induce on $\overline{\mathtt{PP}}$ are identical. By part (i), we may assume that they are both already in normal form. Suppose the first $i-1$ rows of $\overline{W}$ and $\overline{W'}$ are identical, and the ith rows are not. Let $e, e' \in \mathbb{Q}^n$ be the ith rows of $\overline{W}$ and $\overline{W'}$, respectively. By semi-admissibility, the matrices $\overline{W}, \overline{W'}$ must have full rank. Hence e, e' are both non-zero. First suppose $\langle e, e' \rangle \leq 0$. Then observe that $\langle e - e', e \rangle = e^2 - \langle e, e' \rangle > 0$ while $\langle e - e', e' \rangle = \langle e, e' \rangle - e'^2 < 0$. This shows that $\mathbf{X}^{e-e'} \underset{\mathtt{W}}{>} 1$ and and $\mathbf{X}^{e-e'} \underset{\mathtt{W'}}{\leq} 1$, contradiction. (Note that the rationality e, e' depends on the rationality of the original matrices.) Now suppose $\langle e, e' \rangle > 0$. We choose $\alpha \in \mathbb{Q}$ so that

$$\langle e - \alpha e', e \rangle = e^2 - \alpha \langle e, e' \rangle = 0. \tag{7}$$

That is, $\alpha = e^2 / \langle e, e' \rangle$. Consider $E = \langle e - \alpha e', e' \rangle = \langle e, e' \rangle - \alpha e'^2$. If $E \neq 0$, then by perturbing α slightly, we can make $\langle e - \alpha e', e \rangle$ have a different sign than E. This again gives a contradiction. Hence we must have

$$E = \langle e - \alpha e', e' \rangle = \langle e, e' \rangle - \alpha e'^2 = 0. \tag{8}$$

Combining equations (7) and (8), we obtain $\langle e, e' \rangle^2 = e^2 e'^2$. Hence e, e' are parallel. Since the first non-zero entries of both e and e' are ± 1, and $\langle e, e' \rangle > 0$, this implies $e = e'$, contradiction. **Q.E.D.**

The following justifies the "normal form" terminology:

Corollary 7

(i) *The normal forms of semi-admissible rational weight matrices are unique.*

(ii) *Two semi-admissible rational weight matrices are equivalent iff they have the same normal form.*

Proof. (i) This is immediate from the preceding theorem.
(ii) If two such matrices are equivalent, then they induce the same semi-admissible ordering. So they must have the same normal form. Conversely, since they both are equivalent to a unique normal form, they are equivalent to each other. **Q.E.D.**

Final Remarks.
1. This corollary is useful for constructing new semi-admissible orderings with rational matrices – we just have to make sure that their normal forms are distinct.
2. Our normal form can be defined for any real matrix. In case of rational matrices, a *variant normal form* may be preferable: still keeping the rows mutually orthogonal, we now insist that all entries are integer and the GCD of each row is 1. The preceding results hold for this variant.
3. Further treatment of the subject may be found in Robbiano [168, 169]. In particular, we can generalize weight functions to be $w : R \setminus \{0\} \to \Gamma$ where Γ is an ordered Abelian group playing the role of $\mathbb{R}$. All of Gröbner basis theory can be carried out in this general setting.

Exercises

Exercise 1.1: **(i)** Order the monomials of degree 3 in $\mathtt{PP}(X, Y)$ using the three types of admissible ordering (assuming (1)).

(ii) Similar, order the monomials of degree 3 in $\mathtt{PP}(X, Y, Z)$ for the three admissible orderings.

(iii) Suppose we declare $p \underset{\mathtt{A}}{\geq} q$ if $p = q$ or else, in (2), last nonzero component is negative. (This is rather like reverse lexicographic ordering). Is this admissible? Semi-admissible?

(iv) Show by admissible row operations that the weight matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ -2 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

induces an admissible ordering that is one of the three standard examples $(\underset{\mathtt{LEX}}{\leq}, \underset{\mathtt{TOT}}{\leq}, \underset{\mathtt{REV}}{\leq})$. □

Exercise 1.2: Show the normal form matrices for total degree and reverse lexicographical ordering on n variables. □

Exercise 1.3: (Robbiano) Show that every admissible ordering $\underset{\mathtt{A}}{\leq}$ on $\mathtt{PP}$ arises from some weight scheme W. □

Exercise 1.4: Derive the Hilbert basis theorem (§XI.1) from Dickson's lemma. □

Exercise 1.5: The following implies lemma 3. If U is any set totally ordered by $\leq'$, let $S(U)$ denote the set of finite ordered sequences over U. That is, the elements of $S(U)$ have the form $(u_1, \ldots, u_k)$ where $u_1 >' u_2 >' \cdots >' u_k$ and $k \geq 0$. We extend the ordering $\leq'$ to $S(U)$ where

$$u = (u_1, \ldots, u_k) \leq' (v_1, \ldots, v_\ell) = v$$

if either u is a prefix of v (possibly $u = v$) or else there is some $1 \leq i \leq \min\{k, \ell\}$ such that $u_i <' v_i$ and for $j = 1, \ldots, i-1$, $u_j = v_j$. Show: If U is well-ordered by $\leq'$ then $S(U)$ is well-ordered under the induced ordering. □

Exercise 1.6: We want to construct weight matrices with integer entries that induce an admissible ordering different from our standard examples:
(i) If we permute the rows of the normal forms for reverse lexicographic ordering, do we get new admissible orders?
(ii) What about the total degree ordering?
(iii) Construct examples different from (i) and (ii). □

Exercise 1.7: In analogy to Hermite Normal form, let us define an alternative normal form for a weight matrix $\overline{W}$ as follows: Let us call the first non-zero entry of each row a *critical entry* of the matrix. Say a weight matrix is in *normal form* if (1) the critical entries are ± 1, and (2) there are only zero entries below each critical entry.
(i) Show that every weight matrix is equivalent to one of this form.
(ii) Show the analogues of the above normal form results for this new definition. □

§2. Normal Form Algorithm

We fix some admissible ordering $\underset{\mathtt{A}}{\leq}$ in this section.

The heart of the Gröbner bases algorithm is the normal form algorithm which is described here. We also provide an explicit bound on the number of reduction steps in the normal form algorithm.

Concept of Reduction. The *head term* $\mathtt{hterm}(f)$ of a polynomial $f \in R$ refers to the $\underset{\mathtt{A}}{\leq}$-largest power product (=term) in f. The *head coefficient* $\mathtt{hcoef}(f)$ and *head monomial* $\mathtt{hmono}(f)$ of f refers to the coefficient and monomial associated to the head term of f. Thus

$$\mathtt{hterm}(f) \cdot \mathtt{hcoef}(f) = \mathtt{hmono}(f).$$

For example, if $f = -2X^2Y + X + 1$ then $\mathtt{hterm}(f) = X^2Y$, $\mathtt{hmono}(f) = -2X^2Y$ and $\mathtt{hcoef}(f) = -2$. These assertions did not depend on the choice of admissible ordering; we often choose examples with this property. On the other hand, the head term of $g = X + Y^3$ depends on the admissible ordering.

If $F \subseteq R$ is a set of polynomials, we define $\mathtt{hterm}(F)$ in the natural way, $\mathtt{hterm}(F) = \{\mathtt{hterm}(f) : f \in F\}$. We then define the *head ideal* $\mathtt{Head}(F)$ of any set $F \subseteq R$ to be the ideal generated by $\mathtt{hterm}(F)$:

$$\mathtt{Head}(F) := \mathtt{Ideal}(\mathtt{hterm}(F)).$$

A basic concept in Gröbner bases is that of *reduction* Given two polynomials $f, g \in R$, we say f is *reducible* by g if $\mathtt{hmono}(g)$ divides some monomial m in f. In that case $m = c \cdot \mathtt{hmono}(g)$ for some monomial c. We then call $h = f - c \cdot g$ a *reduct* of f by g, and denote this 3-way relationship by

$$f \xrightarrow{g} h.$$

Necessarily, $f \neq h$. If G is any set of polynomials and there exists a g in G such that $f \xrightarrow{g} h$, then we write

$$f \xrightarrow{G} h,$$

and call this a *G-reduction step*. The reflexive transitive closure of the binary relation $\xrightarrow{G}$ is denoted $\xrightarrow[*]{G}$. If f is not reducible by G, we call f a *G-normal form*, and denote this in the suggestive manner: $f \xrightarrow{G}$. A *normal form of $f \in R$* is any G-normal form $\widehat{f}$ such that

$$f \xrightarrow[*]{G} \widehat{f} \xrightarrow{G}.$$

Write $\mathtt{NF}_G(f)$ for the set of G-normal forms of f. The following is an (essentially trivial) algorithm for computing a member of $\mathtt{NF}_G(f)$.

Normal Form Algorithm
Input: a polynomial $f \in R$, a finite set $G \subseteq R$.
Output: a G-normal form h of f.

$h \leftarrow f$.
while h is reducible by G do
1. Pick $g \in G$ such that g divides some monomial m of f.
2. $h \leftarrow h - c \cdot g$ where $m = \mathtt{hmono}(c \cdot g)$.

We remark that this algorithm is non-deterministic because step 1 may have several choices of g and m. A variant of this algorithm is where G is fixed, so that the input is only f. In this case we call it the *G-normal form algorithm*. We shall write $\mathtt{nf}_G(f)$ for the (non-unique) output of this procedure.

There are two fundamental questions concerning this algorithm:

(A) Termination: does every reduction sequence eventually reach a normal form?

(B) Is the normal form always unique?

Note that a positive answer to (A) implies that $\texttt{NF}_G(f)$ is non-empty. A positive answer to (B) implies that $|\texttt{NF}_G(f)| = 1$. The answer to (A) is easily deduced from lemma 3, once we remark that

$$f \xrightarrow{G} h \text{implies} \quad f \underset{\texttt{A}}{>} h.$$

We may now give our first definition of a Gröbner basis: A finite set G of polynomials is a *Gröbner basis* (for the ideal $\texttt{Ideal}(G)$) if for every polynomial f, the G-normal form of f is unique.

Length of Reduction Sequences. We provide explicit upper bounds on the length of such sequences.

For the rest of this section, we fix a finite set $F \subseteq R$ and $g \in R$. For simplicity, assume that $F \cap K = \emptyset$, i.e., F has no constant polynomials. Our goal is to bound the worst case length

$$k = k(F, g)$$

of any sequence of F-reductions that begins with g:

$$g = h_1 \xrightarrow{F} h_2 \xrightarrow{F} h_3 \xrightarrow{F} \cdots \xrightarrow{F} h_k. \tag{9}$$

We assume that our admissible ordering $\underset{\texttt{A}}{\leq}$ arises from the weight scheme

$$W = (w_1, \ldots, w_n)$$

(cf. (5)). Moreover, we may assume that W is non-negative: $w_i(p) \geq 0$ for all $p \in \texttt{PP}$.

Let $\overline{f} = (f_1, f_2, \ldots, , f_m)$ be the term sequence (see (3)) for a polynomial f. Then we define the *extended term sequence* of f as follows. It is equal to $\overline{f}$ if $f_m = 1$; otherwise, the extended term sequence is $(f_1, f_2, \ldots, , f_m, 1)$. So extended term sequences always end with 1.

Let $(f_1, f_2, \ldots, , f_m)$ be the extended term sequence of f. Assuming f is not a constant (i.e., $f \notin K$), we have $m \geq 2$. Then we define:

$$\mu(f) := \min\{w_i(f_j) - w_i(f_{j+1}) \ : \ w_i(f_j) > w_i(f_{j+1}), i = 1, \ldots, n \text{and} \ \ j = 1, \ldots, m-1\}.$$

Since $f_j > f_{j+1}$, there exists an $i = i(j)$ such that $w_i(f_j) > w_i(f_{j+1})$. Thus $\mu(f)$ is well-defined and $\mu(f) > 0$. We also define

$$\begin{aligned} M(f) &:= -\min\{w_i(f_j) - w_i(f_{j+1}) \ : \ i = 1, \ldots, n \text{and} \ \ j = 1, \ldots, m-1\} \\ &= \max\{w_i(f_{j+1}) - w_i(f_j) \ : \ i = 1, \ldots, n \text{and} \ \ j = 1, \ldots, m-1\}. \end{aligned}$$

Unlike $\mu(f)$, $M(f)$ could be negative.

Our bounds on $k = k(F, g)$ will be in terms of the following parameters. Let the term sequence of g be $\overline{g} = (g_1, g_2, \ldots, g_\ell)$.

$$\begin{array}{lcl} \ell & := & \text{the length of } g \\ \Delta & := & \max\{w_i(g_j) : i = 1, \ldots, n \text{and} \ \ j = 1, \ldots, \ell\} \\ \mu_0 & := & \min\{\mu(f) : f \in F\} \\ M_0 & := & \max\{1, \max\{M(f) : f \in F\}\} \end{array}$$

Example: Let $\underset{\mathtt{A}}{\leq}$ be the pure lexicographic ordering $\underset{\mathtt{LEX}}{\leq}$. Then W can be taken to be the $n \times n$ identity matrix (§1). Then $w_i(f_j)$ gives the X_i-degree of the monomial f_j. If $\mathtt{mdeg}(f) = d$ (§0.10), then $1 \leq \mu(f) \leq d$ and $1 \leq M(f) \leq d$.

The main result of this section is:

Theorem 8 *Any sequence of F-reduction steps on g has length at most*

$$k(F, g) \leq \ell 2^{\frac{\Delta \rho^n}{M_0}},$$

where $\rho := \frac{M_0}{\mu_0} + 1$.

In the above pure lexicographic ordering example, we have $\rho \leq d + 1$. Hence $k(F, g) \leq \ell 2^{\Delta(d+1)^n}$. Note that $\Delta = \mathtt{mdeg}(g)$ in this case.

For the proof, first introduce a "cost function" $C_F : \mathtt{PP} \to \mathbb{R}$ where

$$C_F(p) = \frac{1}{\mu_0} \sum_{i=1}^{n} \rho^{n-i} w_i(p).$$

Also let $C_F(cp) = C_F(p)$ for $p \in \mathtt{PP}$ and $c \in K$. Note that $C_F(pq) = C_F(p) + C_F(q)$. Moreover, $C_F(p) \geq 0$ with equality iff $p = 1$.

Lemma 9 *Let $p \underset{\mathtt{A}}{>} q$ be consecutive power products appearing in the extended term sequence of $f \in F$. Then $C_F(p) \geq 1 + C_F(q)$.*

Proof. Since $p \underset{\mathtt{A}}{>} q$, there exists a k_0 such that

$$w_{k_0}(p) > w_{k_0}(q) \qquad \text{and} \qquad w_i(p) = w_i(q) \quad (\text{for } \; i = 1, \ldots, k_0 - 1).$$

Hence

$$\begin{aligned}
C_F(p) - C_F(q) &= \frac{1}{\mu_0} \sum_{i=1}^{n} \rho^{n-i}(w_i(p) - w_i(q)) \\
&= \frac{1}{\mu_0} \rho^{n-k_0}(w_{k_0}(p) - w_{k_0}(q)) + \frac{1}{\mu_0} \sum_{i=k_0+1}^{n} \rho^{n-i}(w_i(p) - w_i(q)) \\
&\geq \rho^{n-k_0} - \frac{1}{\mu_0} \sum_{i=k_0+1}^{n} \rho^{n-i} M_0 \quad (\text{since } \; \max\{\mu_0, -M_0\} \leq w_i(p) - w_i(q)) \\
&= \rho^{n-k_0} - \frac{M_0}{\mu_0} \cdot \frac{\rho^{n-k_0} - 1}{\rho - 1} = 1,
\end{aligned}$$

where the last line is valid since $M_0 \neq 0$ and hence $\rho \neq 1$. **Q.E.D.**

Inductively, we obtain:

Corollary 10 *If $f \in F$ and $\overline{f} = (f_1, f_2, \ldots, f_k)$ then*

$$C_F(f_1) - i + 1 \geq C_F(f_i) \geq 1, \qquad (i = 1, \ldots, k).$$

Now define a cost function on polynomials. For any polynomial h, let

$$\widehat{C}_F(h) := \sum_{i=1}^{m} 2^{C_F(h_i)} = \sum_{i=1}^{m} \widehat{C}_F(h_i).$$

where $\overline{h} = (h_1, \ldots, h_m)$.

Lemma 11 *For any $g, h \in R$, if $g \xrightarrow{F} h$ then $\widehat{C}_F(g) \geq 2 + \widehat{C}_F(h)$.*

Proof. Suppose $h = g - u \cdot f$ for some $f \in F$ and monomial u. Let $\overline{f} = (f_1, f_2, \ldots, f_k)$. The reduction by f removes the monomial uf_1 from g and 'replaces' it with $uf_2 + uf_3 + \ldots + uf_k$. Note that uf_i $(i = 2, \ldots, k)$ may combine with monomials in g and even vanish. If $k = 1$ then we have simply removed uf_1 and clearly $\widehat{C}_F(g) - \widehat{C}_F(h) \geq 1$. Assuming $k \geq 2$,

$$\begin{aligned}
\widehat{C}_F(g) - \widehat{C}_F(h) &\geq 2^{C_F(uf_1)} - \sum_{i=2}^{k} 2^{C_F(uf_i)} \\
&\geq \widehat{C}_F(u) \left[2^C - \sum_{i=2}^{k} 2^{C-i+1} \right], \qquad (C := C_F(f_1)) \\
&= \widehat{C}_F(u) 2^{C-k+1} \\
&\geq \widehat{C}_F(uf_k), \qquad \text{(by previous corollary).}
\end{aligned}$$

But $\widehat{C}_F(uf_k) \geq 2$ since $C_F(uf_k) \geq 1$ (again by previous corollary). **Q.E.D.**

Conclusion of proof. Since $\widehat{C}_F(h) \geq 1$ for any non-zero polynomial h, the last lemma implies that the length of the F-reduction sequence (9) is at most $\widehat{C}_F(g)$ (actually $\widehat{C}_F(g)/2$, but we give up the factor to 2 for the sake of simplicity). With $\overline{g} = (g_1, \ldots, g_\ell)$, we have

$$\begin{aligned}
C_F(g_j) &= \frac{1}{\mu_0} \sum_{i=1}^{n} \rho^{n-i} w_i(g_j) \\
&\leq \frac{\Delta}{\mu_0} \sum_{i=0}^{n-1} \rho^i \\
&= \frac{\Delta}{M_0} (\rho^n - 1).
\end{aligned}$$

Hence $\widehat{C}_F(g) = \sum_{i=1}^{\ell} 2^{C_F(g_i)} < \ell 2^{\Delta \rho^n / M_0}$, and main result follows.

A variant of this bound is derived in [56]. We remark that the bound makes no assumptions about the order of reductions. With some simple restrictions, we can improve the bound from double exponential in n to single exponential (Exercise and [56]).

Exercises

Exercise 2.1:

(i) Let $G = \{X^2Y, Y^2\}$ and $f = 2Y^4 + X^3Y - 3XY + 1$. Show that $\mathtt{NF}_G(f) = \{-3XY + 1\}$.

(ii) Show that any finite set G of monomials is a Gröbner basis. Further, the G-normal forms are independent of the choice of admissible ordering.

□

Exercise 2.2: Let $G = \{g_1, g_2\}$ where $g_1 = XY - 2$ and $g_2 = 3X^2 - 1$. Show that G is not a Gröbner basis. HINT: consider $\mathtt{NF}_G(X^2Y^2)$. □

Exercise 2.3: We want efficient implementations of the normal form algorithm.
(i) Give an efficient data structure $D_1(S)$ for any finite subset $S \in \mathbb{Z}^n$ such that given any $e \in \mathbb{Z}^n$, you can use $D_1(S)$ to quickly retrieve any $d \in S$ such that $\mathbf{X}^d | \mathbf{X}^e$.
(ii) Modify $D_1(S)$ into a semi-dynamic data structure, *i.e.*, where S can have elements inserted and $D_1(S)$ is quickly updated.
(ii) Use the above data structure to implement the normal form algorithm. Assume a sparse representation of polynomials. □

Exercise 2.4: Let n be the number of variables, d the maximum degree of any variable in F, Δ the maximum degree of any variable in g. Show the following:
(i) For total degree ordering, $k(F, g) \leq 2^{(\Delta+1)^n}$.
(ii) The bound in (i) is $k(F, g) \leq (\Delta + 1)^n$ if we insist that the normal form algorithm always chooses to eliminate the $\underset{\mathtt{A}}{\leq}$-largest monomial in g that can be eliminated. □

Exercise 2.5: We show that the above upper bound is, in some sense, the best possible ([56]). The admissible ordering is the usual $\underset{\mathtt{LEX}}{\leq}$ with

$$X_1 \underset{\mathtt{LEX}}{>} X_2 \underset{\mathtt{LEX}}{>} \cdots \underset{\mathtt{LEX}}{>} X_n.$$

Let d, Δ, ℓ, L be integers satisfying $d \geq \ell - 2 > 0$ and $\Delta > L$. The polynomial g is

$$g := X_1^\Delta X_n^L + X_1^\Delta X_n^{L-1} + \cdots + X_1^\Delta X_n$$

and the reducing set F comprises the following polynomials:

$$\begin{aligned}
f_1 &= X_1 - (X_2^d X_3^d \cdots X_{n-1}^d)(X_n^d + X_n^{d-1} + \cdots + X_n^{d-\ell+2}) \\
f_2 &= X_2 - (X_3^d X_4^d \cdots X_{n-1}^d)(X_n^d + X_n^{d-1} + \cdots + X_n^{d-\ell+2}) \\
&\vdots \\
f_{n-1} &= X_{n-1} - (X_n^d + X_n^{d-1} + \cdots + X_n^{d-\ell+2}) \\
f_n &= X_n^\ell - X_n^{\ell-1} - \cdots - X_n \\
f_{n+1} &= X_n^{\ell-1} - X_n^{\ell-2} - \cdots - X_n \\
&\vdots \\
f_{n+\ell-2} &= X_n^2 - X_n \\
f_{n+\ell-1} &= X_n - 1
\end{aligned}$$

(i) What is the upper bound on $k(F, g)$ based on our theorem?
(ii) Consider the reduction sequence that always chooses the $\underset{\overline{\mathtt{LEX}}}{\leq}$-least monomial to reduce. Let $s(f)$ be the length of the entire F-reduction sequence starting from f. Show that if $\overline{f} = (f_1, \ldots, f_m)$ then $s(f) = \sum_{i=1}^m s(f_i)$.
(iii) Show that $s(X_1^{e_1} \cdots X_n^{e_n})$ is at least

$$2^{\frac{\ell-2}{\ell-1}((e_1(d+1)^{n-1} + e_2(d+1)^{n-2} + \cdots + e_n - 1)}.$$

(iv) Conclude that $s(g) \geq 2^{\frac{\ell-2}{\ell-1}(d+1)^{n-1}\Delta}L$. □

§3. Characterizations of Gröbner Bases

We now give several characterizations of Gröbner bases. Note that the following discussion is entirely relative to some fixed admissible ordering $\underset{\mathtt{A}}{\geq}$.

The Normal Form Characterization. Our original definition (§2) of Gröbner basis was based on the concept of reduction and the normal form algorithm: *a finite set $G \subseteq R$ is Gröbner iff every $f \in R$ has a unique G-normal form.* We also say G is a *Gröbner basis* for the $\mathtt{Ideal}(G)$ that it generates.

The Standard Characterization. Perhaps the simplest characterization is this: *a finite set $G \subseteq R$ is Gröbner iff*

$$\mathtt{Head}(G) = \mathtt{Head}(\mathtt{Ideal}(G)). \tag{10}$$

This definition is elegant but highly non-constructive as it does not suggest any procedure to verify if a given set is Gröbner. We call this the *standard characterization* of Gröbner bases because some literature defines "standard bases" precisely this way, being the closest to Hironaka's notion of standard bases.

Buchberger's Characterization. Computationally, this is the most important characterization. The *least common multiple* (LCM) of two power products $p = X_1^{e_1}X_2^{e_2}\cdots X_n^{e_n}$ and $q = X_1^{d_1}X_2^{d_2}\cdots X_n^{d_n}$ is given by

$$\mathtt{LCM}(p,q) := X_1^{\max(d_1,e_1)}X_2^{\max(d_2,e_2)}\cdots X_n^{\max(d_n,e_n)}.$$

The *LCM* of two monomials ap and bq ($a, b \in K$; $p, q \in \mathtt{PP}$) is[2] $ab \cdot \mathtt{LCM}(p,q)$. For example, the LCM of $6X^3Y$ and $4X^2Y^2$ is $24X^3Y^2$.

A critical notion in the constructive approach to Gröbner basis due to Buchberger is the following: for $f, g \in R$ where $m = \mathtt{LCM}(\mathtt{hterm}(f), \mathtt{hterm}(g))$, we define the *S-polynomial* of f and g to be

$$S(f,g) := \frac{m}{\mathtt{hmono}(f)} \cdot f - \frac{m}{\mathtt{hmono}(g)} \cdot g.$$

It is easy to understand $S(f,g)$ once we note that the head terms in $\frac{m}{\mathtt{hmono}(f)} \cdot f$ and $\frac{m}{\mathtt{hmono}(g)} \cdot g$ both equal m, and hence they cancel each other in the defining expression for $S(f,g)$. For example, with $g_1 = XY - 2$, $g_2 = 3X^2 - 1$, we get

$$m = \mathtt{LCM}(XY, 3X^2) = 3X^2Y, \qquad S(g_1, g_2) = 3X \cdot g_1 - Y \cdot g_2 = -6X + Y.$$

Buchberger's characterization is this: *G is Gröbner iff $S(f,g) \xrightarrow[*]{G} 0$ for all $f, g \in G$.*

[2] If K is the quotient field of a UFD, then we could replace ab by $\mathtt{LCM}(a,b)$ (§III.1).

Term Rewriting Connection. The Gröbner bases literature is closely related to the term-rewriting literature. For any fixed set $F \subseteq R$, we can view the binary relation $\xrightarrow[*]{F}$ on R as a partial order. In general, consider a partial order $\xrightarrow[*]{}$ on an arbitrary set U. Let us write $f \longrightarrow g$, if $f \xrightarrow[*]{} g$ and $f \neq g$. The partial order is *Noetherian* if it has no infinite descending sequence in the sense of

$$f_1 \longrightarrow f_2 \longrightarrow f_3 \longrightarrow \cdots.$$

The partial order is said to be *Church-Rosser* (alternatively, *confluent*) if for all $f, g, g' \in U$, if $f \xrightarrow[*]{} g$ and $f \xrightarrow[*]{} g'$ then there exists $h \in U$ such that $g \xrightarrow[*]{} h$ and $g' \xrightarrow[*]{} h$. (Such an h is called a *common successor* of g, g'.) A minimal element f in this partial order is called an *normal form*, or *a normal form of* g if, in addition, $g \xrightarrow[*]{} f$. A powerful induction principle for Noetherian structures is the following:

Proposition 12 (Principle of Noetherian Induction) *For a Noetherian relation $\xrightarrow[*]{}$, to establish the validity of a predicate $P(x)$ for all $x \in U$, it is sufficient that:*
(1) $P(x)$ holds for all normal form elements $x \in U$.
(2) For $x \in U$: if $P(y)$ holds whenever $x \longrightarrow y$, then we can conclude that $P(x)$ holds.

In proof, suppose that (1) and (2) hold. For the sake of contradiction, suppose $P(x_0)$ fails. Then x_0 is not a normal form. So there exists some x_1 such that $x_0 \longrightarrow x_1$ and $P(x_1)$ fails. Continuing in this fashion, we obtain an infinite descending sequence $x_0, x_1, x_2, \ldots$, which contradicts the Noetherian property. The following is a basic result in the theory of term-rewriting:

Theorem 13 *If $\xrightarrow[*]{}$ is a Noetherian partial order on U then $\xrightarrow[*]{}$ is Church-Rosser iff every $g \in U$ has a unique normal form.*

This is easy to show by Noetherian induction (Exercise).

The Church-Rosser Property Characterization. We now apply this terminology to the partial order $\xrightarrow[*]{F}$ induced by F-reductions. From §2, we know that this partial order is Noetherian. Combined with theorem 13, we conclude that the Church-Rosser property for $\xrightarrow[*]{F}$ is equivalent to F being Gröbner: *a set $G \subseteq R$ to be Gröbner iff the relation $\xrightarrow[*]{G}$ is Church-Rosser.*

Extended Standard Characterization. The following is a useful variant of the standard characterization:

Lemma 14 (Extended standard characterization) *G is Gröbner in the standard sense of (10) iff for all $f \in \texttt{Ideal}(G)$, there are elements $\alpha_i \in R, g_i \in G$, $(i = 1, \ldots, r)$ such that*

$$f = \sum_{i=1}^{r} \alpha_i g_i \tag{11}$$

and $\texttt{hterm}(f) \underset{\texttt{A}}{\geq} \texttt{hterm}(\alpha_i g_i)$ for all i.

Proof. ($\Rightarrow$) Let $f \in \texttt{Ideal}(G)$. If $f = 0$ then f has the desired form (11). Otherwise, since $\texttt{hmono}(f) \in \texttt{Head}(\texttt{Ideal}(G)) = \texttt{Head}(G)$ there is some $g_1 \in G$ and monomial p_1 such that

$\texttt{hmono}(p_1 g_1) = \texttt{hmono}(f)$. By the principle of Noetherian induction on $\texttt{Ideal}(G)$, we may assume that $f - p_1 g_1$ can be represented in a form of equation (11). Then clearly f can be so represented. [The Noetherian induction refers to the partial order $\xrightarrow[*]{G}$ on $\texttt{Ideal}(G)$ with 0 as the only normal form.]

($\Leftarrow$) Conversely, assume every $f \in \texttt{Ideal}(G)$ can be expressed as in Equation (11). We want to show $\texttt{Head}(G) = \texttt{Head}(\texttt{Ideal}(G))$. It is clear that $\texttt{Head}(G) \subseteq \texttt{Head}(\texttt{Ideal}(G))$. For the opposite inclusion, suppose $p \in \texttt{Head}(\texttt{Ideal}(G))$ and $p \neq 0$. Then $p = \sum_i \beta_i \texttt{hterm}(f_i)$ where $\beta_i \in R$ and $f_i \in \texttt{Ideal}(G)$. From (11), we see that each f_i can be written as $\sum_j \alpha_{ij} g_{ij}$ where $g_{ij} \in G$. We may assume that α_{ij} are monomials. Therefore

$$p = \sum_i \beta_i \left(\sum_j \texttt{hterm}(\alpha_{ij} g_{ij}) \right) = \sum_i \sum_j (\beta_i \alpha_{ij}) \texttt{hterm}(g_{ij}),$$

implying $p \in \texttt{Head}(G)$. **Q.E.D.**

Ideal Membership Characterization. It is clear that if g is the F-normal form of h then $g \equiv h (\mathrm{mod}\, \texttt{Ideal}(F))$. If F is Church-Rosser, then we have a unique representative for each equivalence class of polynomials modulo $\texttt{Ideal}(F)$. Since 0 is clearly a normal form, it must be the unique representative for each element in $\texttt{Ideal}(F)$. In particular, for all $h \in \texttt{Ideal}(F)$, $h \xrightarrow[*]{F} 0$. The *ideal membership characterization* of Gröbner bases says that the converse also holds: *F is Gröbner iff*

$$\forall h \in \texttt{Ideal}(F), \quad h \xrightarrow[*]{F} 0. \tag{12}$$

Note that the condition (12) is equivalent to $0 \in \texttt{NF}_F(h)$ for all $h \in \texttt{Ideal}(F)$. But in fact (12) is equivalent to the stronger assertion

$$\forall h \in \texttt{Ideal}(F), \quad \texttt{NF}_F(h) = \{0\}. \tag{13}$$

To see this, it suffices to show that $g \in \texttt{NF}_F(h)$ implies $g = 0$. If $g \neq 0$ then, since $g \in \texttt{Ideal}(F)$, we have $g \xrightarrow[*]{F} 0$. So g is F-reducible, contradiction.

We summarize the above discussions in the following:

Theorem 15 (Characterization of Gröbner bases) *Let $G \subseteq I \subseteq R$ where I is an ideal and G is finite. The following are equivalent:*

1. (Normal Form) *Every $f \in R$ has a unique G-normal form.*
2. (Standard bases) $\texttt{Head}(G) = \texttt{Head}(I)$.
3. (Buchberger's Criterion) *For all $f, g \in G$, we have $S(f,g) \xrightarrow[*]{G} 0$.*
4. (Church-Rosser) *The relation $\xrightarrow[*]{G}$ is Church-Rosser.*
5. (Extended Standard) *Every $f \in \texttt{Ideal}(G)$ has the expression*

$$f = \sum_{i=1}^{r} \alpha_i g_i, \qquad (\text{where } \texttt{hterm}(f) \underset{\texttt{A}}{\geq} \texttt{hterm}(\alpha_i g_i), \alpha_i \in R, g_i \in G).$$

6. (Ideal Membership) *For all $f \in I$, we have $f \xrightarrow[*]{G} 0$.*

Exercises

Exercise 3.1: Prove theorem 13 using Noetherian induction. □

Exercise 3.2: Define $H(f)$ to mean the highest degree homogeneous component of a polynomial $f \in R$. If $F \subseteq R$, let $H(F) = \{H(f) : f \in H\}$. The *H-base* or *Macaulay base* of an ideal I is a finite set F such that $\texttt{Ideal}(H(I)) = \texttt{Ideal}(H(F))$. Let $f^\wedge$ denote the homogenization of f with respect to a new variable X_0, and for a set $S \subseteq R$, $S^\wedge = \{f^\wedge : f \in S\}$. Prove: $\{f_1, \ldots, f_m\}$ is a H-base of I iff $I^\wedge = \texttt{Ideal}({f_1}^\wedge, \ldots, {f_m}^\wedge)$. □

Exercise 3.3:

(i) Is the following set

$$F_1 = \{X^3 - X^2 + X - 1, XY - Y - X^2 + X, Y^2 - X^2\}$$

a Gröbner basis under the lexicographic ordering where $X \underset{\texttt{LEX}}{<} Y$?

(ii) Show that the following is a Gröbner basis under the total degree ordering (does not matter whether $X \underset{\texttt{TOT}}{<} Y$ or $X \underset{\texttt{TOT}}{>} Y$):

$$F_2 = \{XY^4 - 1, X^3 - Y, Y^5 - X^2\}.$$

(iii) Give a finite procedure to check if any set F is Gröbner. HINT: use Buchberger's characterization.

□

Exercise 3.4: (Sturmfels-Eisenbud) Let $\Lambda \subseteq \mathbb{Z}^n$ is a lattice (§VIII.1). The *lattice ideal* $I \subseteq K[X_1, \ldots, X_n] = K[\mathbf{X}]$ of Λ is the ideal with the generator set of elements of the form

$$\mathbf{X}^{e^+} - \mathbf{X}^{e^-}, \qquad e \in \Lambda.$$

Here e^+ is obtained from e by replacing any negative component by 0, and e^- is obtained from $-e$ by a similar process. E.g., $e = (-2, 2, 0, 3, -1)$, $e^+ = (0, 2, 0, 3, 0)$, $e^- = (2, 0, 0, 0, 1)$.

(i) Show that the ideal generated by F_2 (previous exercise) is a lattice ideal, where the lattice is generated by the vectors $(1, 4)$ and $(4, 3)$.

(ii) If $A = \{a_1, \ldots, a_m\} \subseteq \mathbb{Z}^n$ is a generating set of Λ and all the components of a_i are positive, then I is generated by the set $\{\mathbf{X}^{a^+} - \mathbf{X}^{a^-} : a \in A\}$.

(iii) Show that $\mathbb{Z}^n/\Lambda$ is a direct sum $G_0 \oplus G_1$ where G_0 is a free Abelian group and G_1 is the direct sum of a finite number of cyclic groups. HINT: Use Hilbert's basis theorem for modules and the fundamental theorem of finitely generated Abelian groups (see §XI.1 and §X.9).

(iv) Show that if G_1 is trivial then I is a prime ideal. This is called a *toric ideal* [Sturmfels and Eisenbud].

(v) Show that if G_0 is trivial, the order of the group G_1 is equal to the determinant of Λ, and I is zero-dimensional. In fact, $K[\mathbf{X}]/I$ has dimension $\det(\Lambda)$.

(vi) Relative to an admissible ordering $\underset{\texttt{A}}{\leq}$, $\texttt{Head}(I)$ is generated by the monomials $\{\mathbf{X}^{e^+} : e \in \Lambda, e^+ \underset{\texttt{A}}{\geq} e^-\}$.

(vii) A binomial $\mathbf{X}^e - \mathbf{X}^d$ belongs to I iff $e - d \in \Lambda$.

□

Exercise 3.5: Complete the proof of theorem 15. □

§4. Buchberger's Algorithm

Despite Buchberger's criterion, it is not immediately obvious how to construct a Gröbner basis for any ideal (given by a set of generators, as always). We now present Buchberger's algorithm to construct Gröbner bases. There are various heuristics to improve the algorithm but the basic version of the algorithm given below is easy to describe.

```
Buchberger's Algorithm
    Input:   F ⊆ R a finite set of polynomials
    Output:  G a Gröbner basis for Ideal(F).
    G ← F;
    B ← {S(f,g) : f,g ∈ F, f ≠ g};
    while B ≠ ∅ do begin
        Remove f from B;
        h ← nf_G(f);
        if h' ≠ 0 then begin
            B ← B ∪ {S(f,h') : f ∈ G};
            G ← G ∪ {h'}
        end {if}
    end {while}
```

Correctness: If this algorithm terminates, it is easy to see from Buchberger's criterion that G is a Gröbner basis for $\texttt{Ideal}(F)$. To see that this algorithm terminates, let

$$h_1, h_2, \ldots$$

be the sequence of non-zero polynomials produced by the normal form algorithm in the while-loop. Since each h_i is not reducible by any of the previous polynomials (since h_i is not reducible by G which contains $\{h_1, \ldots, h_{i-1}\}$) we see that $\texttt{hterm}(h_i)$ is not divisible by $\texttt{hterm}(h_j)$ for all $j = 1, \ldots, i-1$. By Dickson's lemma, the set $\{\texttt{hterm}(h_i) : i \geq 1\}$ must be finite, *i.e.*, the loop terminates.

Exercises

Exercise 4.1: **(i)** Compute the Gröbner basis of $\{XY^4 - 1, X^3 - Y, Y^5 - X^2\}$ under the pure lexicographic ordering assuming $X \underset{\texttt{LEX}}{>} Y$.

(ii) Now assume $X \underset{\texttt{LEX}}{<} Y$.

(iii) Now use the reverse lexicographic ordering with $X \underset{\texttt{REV}}{>} Y$. □

Exercise 4.2: Compute the Gröbner basis of $\{X^3 - Y, XY - Z\}$ relative to the reverse lexicographic order (assuming $Z \underset{\texttt{REV}}{>} Y \underset{\texttt{REV}}{>} X$). □

Exercise 4.3: (Trinks) Compute the Gröbner basis for the following polynomials in $\mathbb{Q}[w, p, z, t, s, b]$ using the pure lexicographic order determined by $w \underset{\texttt{LEX}}{>} p \underset{\texttt{LEX}}{>} z \underset{\texttt{LEX}}{>} t \underset{\texttt{LEX}}{>} s \underset{\texttt{LEX}}{>} b$:

$$\begin{aligned} A_1 &= 45p + 35s - 165b - 36, \\ A_2 &= 35p + 40z + 25t - 27s, \\ A_3 &= 15w + 25ps + 30z - 18t - 165b^2, \\ A_4 &= -9w + 15pt + 20zs, \\ A_5 &= wp + 2zt - 11b^3, \\ A_6 &= 99w - 11sb + 3b^2. \end{aligned}$$

NOTE: The Gröbner basis has a univariate polynomial in b of degree 10, coefficients up to 60 digits. Christoph Koegl informs us that this can be solved in tenths of a second on Sparc 10 or 20 (ca. 1997). □

§5. Uniqueness

Now that we have established the existence of Gröbner bases, we turn to the question of their uniqueness. It is clear that we need to impose some additional conditions to get any kind of uniqueness.

We define a non-empty set $F \subseteq K[X_1, \ldots, X_n]$ to be *self-reduced* if each $f \in F$ is not reducible by $F - \{f\}$. Of course, this definition is relative to some admissible ordering.

To treat some extreme cases of this definition, let us define the zero polynomial 0 to be reducible by F, for all F. Hence, no self-reduced set contains 0. By definition, if $f \neq 0$ then f is not reducible by the empty set. Hence if F is a singleton set and $F \neq \{0\}$, then it is self-reduced. Also, if $c \in K$ and $c \in F$ then F is self-reduced implies $F = \{c\}$.

Definition: A finite non-empty set $G \subseteq R$ is a *reduced (Gröbner) basis* if (1) it is a Gröbner basis (2) it is self-reduced and (3) each $f \in G$ is monic (*i.e.*, $\texttt{hcoef}(f) = 1$).

Lemma 16 *Every ideal has a reduced basis (relative to any choice of an admissible ordering).*

Proof. Begin with any Gröbner basis G. For each $g \in G$, reduce it with respect to the remaining polynomials of G. If the result g' is 0, then discard g from G; otherwise replace g by g'. After such a step, it is easy to see that G remains a Gröbner basis for the original ideal (use the standard characterization). We systematically cycle through all choices of g, and we continue this as long as G continues to change. More precisely, we stop if no $g \in G$ can be reduced by the remaining polynomials. Clearly this process must terminate because each replacement polynomial g' is $\underset{\texttt{A}}{<}$ the original g. Finally, we replace each polynomial g by $g/\texttt{hcoef}(g)$. **Q.E.D.**

Monomial Ideals. A *monomial ideal* is one generated by a set of monomials. It is easy to show:

An ideal I is monomial iff for all $f \in I$, each monomial of f also belongs to I. (14)

We prove a strong uniqueness property for reduced bases of monomial ideals:

Theorem 17 *Reduced bases for monomial ideals are comprised of power products; moreover, they are unique and this is independent of the admissible ordering.*

We shall omit the proof because it is similar to, but simpler than, the proof for the homogeneous case (Theorem 20 below).

Lemma 18 *If G is a reduced basis then*

$$\texttt{hterm}(G) = \{\texttt{hterm}(f) : f \in G\}$$

is a reduced basis for $\texttt{Head}(G)$.

Proof. By the standard characterization of Gröbner bases, $\texttt{hterm}(G)$ is a basis for $\texttt{Head}(G)$. Clearly $\texttt{hterm}(G)$ is self-reduced because this amounts to saying that no term in $\texttt{hterm}(G)$ divides another term in $\texttt{hterm}(G)$. **Q.E.D.**

Theorem 19 (Buchberger) *Reduced bases are unique, up to choice of admissible orderings.*

Proof. By way of contradiction, suppose G, G' are distinct reduced bases for the same ideal. By the previous two lemmas, we conclude that $\texttt{hterm}(G) = \texttt{hterm}(G')$; this equality holds only because we assume the admissible ordering in G and G' are the same. In particular, $|G| = |G'|$. Since G, G' are distinct, without loss of generality, let $f \in G \setminus G'$. Then there is $f' \in G'$ such that $\texttt{hterm}(f) = \texttt{hterm}(f')$. Then $f - f'$ is a non-zero polynomial. Suppose that $\texttt{hterm}(f - f')$ occurs in f (the argument if it occurs in f' is similar). Since $f - f' \xrightarrow[*]{G} 0$, there is some $g \in G$ such that $\texttt{hterm}(g)$ divides $\texttt{hterm}(f - f')$. Clearly $g \neq f$. Then f is reducible by $G \setminus \{f\}$, contradicting the self-reduced property. **Q.E.D.**

Homogeneous Ideals. A *homogeneous ideal* is one generated by a set of homogeneous elements. As for monomial ideals, it is easy to show:

$$I \text{ is homogeneous iff for all } f \in I, \text{ each homogeneous component of } f \text{ belongs to } I. \tag{15}$$

We have a strong uniqueness result similar to the case of monomial ideals:

Theorem 20 *Reduced bases for homogeneous ideals are comprised of homogeneous polynomials; moreover, such bases are unique and independent of the admissible ordering.*

Proof. Let G be a reduced basis for a homogeneous ideal I. If $f \in G$ is not homogeneous, let h be any homogeneous component of f of degree less than $\deg(f)$. Since $h \in I$, there is some $g \in G$ such that h is reducible by g. Hence f is reducible by g and $g \neq f$. This contradicts the self-reduced property.

To show uniqueness, suppose G' is another reduced basis, $G' \neq G$. We do not assume that the admissible orderings implicit in the definitions of G and G' are the same. Hence, for any polynomial f, we shall write $\texttt{hterm}(f)$ and $\texttt{hterm}'(f)$ to denote the head terms of f relative to the two admissible orderings. We may now assume that G, G' have only homogeneous polynomials. Without loss of

generality, let $f \in G \setminus G'$. Then $\texttt{hterm}'(f')$ divides $\texttt{hterm}(f)$ for some $f' \in G'$. Similarly, $\texttt{hterm}(g)$ divides $\texttt{hterm}'(f')$ for some $g \in G$. Hence f is reducible by g. This is a contradiction unless $g = f$. So assume $g = f$. Then $\texttt{hterm}(f) = \texttt{hterm}'(f')$. Consider the non-zero polynomial $f - f'$. Consider $\texttt{hterm}(f - f')$ and we may suppose it belongs to f. Since $f - f' \xrightarrow[*]{G} 0$, we conclude that $\texttt{hterm}(f - f')$ is reducibile by some $h \in G$. This means f is reducible by h, contradicting the self-reduced property. Thus f does not exist and $G = G'$. **Q.E.D.**

Exercises

Exercise 5.1: Show the characterizations of monomial and homogeneous ideals in (14) and (15). □

Exercise 5.2: Modify Buchberger's algorithm (§4) to output a reduced basis. □

§6. Elimination Properties

In this and the following sections, we discuss of applications of Gröbner bases. The number of applications are numerous. But two basic applications are: computing in residue class rings (the original motivation of Buchberger) and elimination. This section considers the elimination applications.

Let $\mathbf{X}, \mathbf{Y}$ be disjoint sets of variables and suppose $\underset{\mathbf{X}}{\leq}$ and $\underset{\mathbf{Y}}{\leq}$ are admissable orderings on $\texttt{PP}(\mathbf{X})$ and $\texttt{PP}(\mathbf{Y})$, respectively. Then we define the *lexicographic product* of $\underset{\mathbf{X}}{\leq}$ and $\underset{\mathbf{Y}}{\leq}$,

$$(\underset{\mathbf{X}}{\leq}, \underset{\mathbf{Y}}{\leq})$$

which is an admissible ordering on $\texttt{PP}(\mathbf{X} \cup \mathbf{Y})$: If $p, p' \in \texttt{PP}(\mathbf{X}), q, q' \in \texttt{PP}(\mathbf{Y})$ then

$$pq \; (\underset{\mathbf{X}}{\leq}, \underset{\mathbf{Y}}{\leq}) \; p'q'$$

iff $p \underset{\mathbf{X}}{\leq} p'$ or else, $p = p'$ and $q \underset{\mathbf{Y}}{\leq} q'$.

Clearly a pure lexicographic ordering $\underset{\texttt{LEX}}{\leq}$ on n variables in which $X_i \underset{\texttt{LEX}}{\leq} X_{i+1}$ is obtained as a lexicographic product of the unique admissible ordering on each of the variables, $(\underset{\mathtt{X}_1}{\leq}, \ldots, \underset{\mathtt{X}_n}{\leq})$.

Theorem 21 *Let* $\underset{\mathbf{X},\mathbf{Y}}{\leq}$ *denote the lexicographic product of admissible orderings* $\underset{\mathbf{X}}{\leq}$ *and* $\underset{\mathbf{Y}}{\leq}$ *on* $\texttt{PP}(\mathbf{X})$ *and* $\texttt{PP}(\mathbf{Y})$*, respectively. If* $G \subseteq K[\mathbf{X}, \mathbf{Y}]$ *is a Gröbner basis with respect to* $\underset{\mathbf{X},\mathbf{Y}}{\leq}$ *then*

(a) $\texttt{Ideal}_{K[\mathbf{X},\mathbf{Y}]}(G) \cap K[\mathbf{Y}] = \texttt{Ideal}_{K[\mathbf{Y}]}(G \cap K[\mathbf{Y}])$

(b) $G \cap K[\mathbf{Y}]$ *is a Gröbner basis for* $\texttt{Ideal}_{K[\mathbf{Y}]}(G \cap K[\mathbf{Y}])$.

Proof. For part (a), note that the inclusion

$$\texttt{Ideal}_{K[\mathbf{Y}]}(G \cap K[\mathbf{Y}]) \subseteq \texttt{Ideal}_{K[\mathbf{X},\mathbf{Y}]}(G) \cap K[\mathbf{Y}]$$

is immediate. The converse inclusion uses a simple observation:

$$\mathtt{hterm}(f) \in \mathtt{PP}(\mathbf{Y}) \Longleftrightarrow f \in K[\mathbf{Y}].$$

If $f \in \mathtt{Ideal}_{K[\mathbf{X},\mathbf{Y}]}(G) \cap K[\mathbf{Y}]$, then by applying a sequence of G-reductions to reduct f to 0, we see that

$$f = \sum_{i=1}^{m} f_i g_i$$

where $f_i \in K[\mathbf{X},\mathbf{Y}]$, $g_i \in G$, $\mathtt{hterm}(f) \underset{\mathbf{X},\mathbf{Y}}{\geq} \mathtt{hterm}(f_i g_i)$. But this implies that $\mathtt{hterm}(f_i g_i) \in \mathtt{PP}(\mathbf{Y})$ which means $f_i, g_i \in K[\mathbf{Y}]$. This proves part (a). But the proof also shows that $G \cap K[\mathbf{Y}]$ is a Gröbner basis for the ideal $\mathtt{Ideal}_{K[\mathbf{Y}]}(G \cap K[\mathbf{Y}])$ (use the extended standard characterization). **Q.E.D.**

If $I \subseteq K[\mathbf{X}]$ and $\mathbf{Y}$ is any subset of $\mathbf{X}$ then we call $I \cap K[\mathbf{Y}]$ an *elimination ideal* of I. Note that $I \cap K[\mathbf{Y}]$ is clearly an ideal in $K[\mathbf{Y}]$. The theorem just proved shows how to obtain elimination ideals — by constructing a Gröbner basis with respect to any admissible ordering which is a lexicographic product of two admissible orderings over $\mathtt{PP}(\mathbf{X} \setminus \mathbf{Y})$ and $\mathtt{PP}(\mathbf{Y})$, respectively. Note that the variables in $\mathbf{X} \setminus \mathbf{Y}$ are to be "eliminated" in order to get $I \cap K[\mathbf{Y}]$, and these variables are given "greater priority" in the admissible ordering. In particular, with the pure lexicographic ordering $\underset{\mathtt{A}}{<}$ in which

$$\mathbf{X}_1 \underset{\mathtt{A}}{<} \cdots \underset{\mathtt{A}}{<} \mathbf{X}_n,$$

we get a Gröbner basis $G \subseteq K[\mathbf{X}_1, \ldots, \mathbf{X}_n]$ such that, for each $i = 1, \ldots, n$, $G \cap K[\mathbf{X}_1, \ldots, \mathbf{X}_i]$ generates an elimination ideal of (G). Of course, $G \cap K[\mathbf{X}_1, \ldots, \mathbf{X}_i]$ may be trivial, *i.e.*, $\{0\}$.

A set $S \subseteq R$ is *zero-dimensional* iff its zero set $\textsc{Zero}(S) \subseteq \overline{K}^n$ is finite. For example, if $S = \{X_1^2 - 1, X_2^2, \ldots, X_n^2\}$ then $\textsc{Zero}(S)$ has two zeros $(\pm 1, 0, \ldots, 0)$.

Theorem 22 *Let $\mathbf{X}_i$ $(i = 1, \ldots, n)$ be disjoint groups of variables, and $I \subseteq K[\mathbf{X}_1, \ldots, \mathbf{X}_n]$. Then I is zero-dimensional iff for each $i = 1, \ldots, n$, $I \cap K[\mathbf{X}_i] \neq \emptyset$.*

Proof. ($\Leftarrow$) If $f_i \in I \cap K[\mathbf{X}_i]$ for each $i = 1, \ldots, n$ then for any zero $\xi = (\xi_1, \ldots, \xi_n) \in \textsc{Zero}(I)$, ξ (and hence ξ_i) must be a root of f_i. So $\textsc{Zero}(I)$ is clearly finite.

($\Rightarrow$) Suppose $\textsc{Zero}(I) = \{\xi^{(1)}, \ldots, \xi^{(m)}\}$ where $\xi^{(j)} = (\xi_1^{(j)}, \ldots, \xi_n^{(j)})$. Then for each $i = 1, \ldots, n$, there is a polynomial $f_i \in K[\mathbf{X}_i]$ such that $f_i(\xi_i^{(j)}) = 0$ for each $j = 1, \ldots, m$. Hence f_i vanishes on $\textsc{Zero}(I)$. By Hilbert's Nullstellensatz, some power m_i of f_i belongs to I: $f_i^{m_i} \in I$. This proves $I \cap K[\mathbf{X}_i] \neq \emptyset$.

Q.E.D.

Corollary 23 *Let G be a Gröbner basis. Then G is zero-dimensional, iff for each $i = 1, \ldots, n$, there is a $g_i \in G$ with $\mathtt{hterm}(g_i) \in \mathtt{PP}(\mathbf{X}_i)$.*

Proof. This is because $\mathtt{hterm}(G) = \{\mathtt{hterm}(g) : g \in G\}$ generates $\mathtt{Head}(\mathtt{Ideal}(G))$ and $\mathtt{Ideal}(G) \cap K[\mathbf{X}_i]$ is non-empty implies there exists $g_i \in \mathtt{Head}(\mathtt{Ideal}(G)) \cap K[\mathbf{X}_i]$. **Q.E.D.**

Theorem 24 *Let G be a Gröbner basis with respect to the pure lexicographic order $\underset{\texttt{LEX}}{\leq}$ where $\mathbf{X}_1 \underset{\texttt{LEX}}{<} \mathbf{X}_2 \underset{\texttt{LEX}}{<} \cdots \underset{\texttt{LEX}}{<} \mathbf{X}_n$. Then G is zero-dimensional if and only if for each $i = 1, \ldots, n$, there is a $p_i \in G$ such that $p_i \in K[\mathbf{X}_1, \ldots, \mathbf{X}_i] - K[\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}]$.*

Proof. If G is zero dimensional, the existence of p_i follows from the above corollary (if $\texttt{hterm}(g_i) \in \texttt{PP}(\mathbf{X}_i)$ then $g_i \in K[\mathbf{X}_1, \ldots, \mathbf{X}_i] \setminus K[\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}]$). Conversely, suppose for each i, we have some $p_i \in G$ where $p_i \in K[\mathbf{X}_1, \ldots, \mathbf{X}_i] - K[\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}]$. Then we see that every $\xi = (\xi_1, \ldots, \xi_n)$ in $\textsc{Zero}(G)$ must satisfy $p_i(\xi_1, \ldots, \xi_i) = 0$. If $i = 1$, this shows there are finitely many ξ_1 in $\textsc{Zero}(G)$. Inductively, if there are only finitely many values for the first $i-1$ components of $\xi \in \textsc{Zero}(G)$, we see from $p_i(\xi_1, \ldots, \xi_{i-1}, \xi_i) = 0$ that there are only finitely many possible values for ξ_i.

Q.E.D.

Application to Solving Polynomial Equations. Suppose we want to solve the system $f_1 = f_2 = \cdots = f_m = 0$ of polynomial equations:

- The question whether the system is inconsistent (has no solutions) amounts to whether $1 \in \texttt{Ideal}(f_1, \ldots, f_m)$. We can determine the membership of any element g in $\texttt{Ideal}(f_1, \ldots, f_m)$ by first computing a Gröbner basis G of $\texttt{Ideal}(f_1, \ldots, f_m)$ and then checking if $\texttt{nf}_G(g) = 0$. But for $g = 1$, it is sufficient to see if G contains a non-zero constant.
- If the system is consistent, we can check whether it has finitely many solutions. Compute a Gröbner basis G relative to any $\underset{\texttt{A}}{\leq}$. Check for each i whether there is a polynomial $p_i \in G$ with $\texttt{hterm}(p_i) \in \texttt{PP}(X_i)$.
- Finally, suppose the system has finitely many solutions. Suppose we answered the previous question of zero-dimensionality by choosing $\underset{\texttt{A}}{\leq}$ according to theorem 24. Then we can continue from the Gröbner basis G computed there: we solve for all possible values for X_1 (considering the polynomials in $G \cap K[X_1]$). Then we back back-solve in the natural way: for each possible value of X_1, we solve the polynomials in $G \cap K[X_1, X_2]$ for the associated values of X_2. This can be continued in the obvious way.

Application in Geometric Projections. Geometrically, the elimination of variables amounts to projection. Let

$$\mathbf{X} = \{X_1, \ldots, X_m\}, \qquad \mathbf{Y} = \{Y_1, \ldots, Y_n\} \tag{16}$$

and for $I \subseteq K[\mathbf{X}, \mathbf{Y}]$, let V be the algebraic set $\texttt{Zero}(I) \subseteq \overline{K}^{m+n}$. We are interested in the projection map

$$\pi : \overline{K}^{m+n} \to \overline{K}^{m}$$

given by $\pi(\mathbf{x}, \mathbf{y}) = \mathbf{x}$. Intuitively, $I \cap K[\mathbf{X}]$ is the zero set of the projection $\pi(V)$. There is a subtlety: the projection may not be an algebraic set. A simple example is $V = \texttt{Zero}(XY - 1)$ where $\pi(V) = \overline{K} - \{0\}$, which is not an algebraic set. Thus we define the *Zariski closure* of a subset $S \subseteq \overline{K}^m$ to be the smallest algebraic set that contains S. Then we have the following theorem, whose proof we leave to an exercise.

Theorem 25 . *Let $I \subseteq K[\mathbf{X}, \mathbf{Y}]$ be an ideal. The Zariski closure of $\pi(\texttt{Zero}(I))$ is equal to $\texttt{Zero}(I \cap K[\mathbf{X}])$.*

An instance of projection is the problem of "implicitization". In computer-aided design, a surface $S \subseteq \mathbb{R}^3$ is often represented *parametrically* in the form:

$$X = \frac{f(s,t)}{d(s,t)}, \quad Y = \frac{g(s,t)}{d(s,t)}, \quad Z = \frac{h(s,t)}{d(s,t)}, \tag{17}$$

where $f, g, h, d \in \mathbb{Q}[s,t]$. That is, $S = \{(X(s,t), Y(s,t), Z(s,t)) : (s,t) \in \mathbb{R}^2\}$. Implicitization of these parametric equations is the problem of computing a polynomial $A(X,Y,Z)$ such that $S = \texttt{Zero}(A)$. This polynomial can be obtained by computing

$$\texttt{Ideal}(d(s,t)X - f(s,t), d(s,t)Y - g(s,t), d(s,t)Z - h(s,t)\} \cap \mathbb{Q}[X,Y,Z],$$

i.e., eliminating s, t. In fact, by homogenizing the equations using another indeterminate u and treating X, Y, Z as the indeterminates, $A(X,Y,Z)$ can be computed as a Macaulay resultant, (see [8, 9] for this and other applications). Note that the parametric form (17) is useful when we have to render the surface, as in computer graphics. But the implicit form $A(X,Y,Z)$ is more useful when we need to know if a given point (a,b,c) is on the surface, or on one side of the surface).

Exercises

Exercise 6.1: For the following systems, indicate whether it is solvable, and if so, whether it is zero-dimensional, and if so, compute the finitely many solutions:
(i) $\{X^2 + Y^2 - 1,\ XY - 1,\ Y^2 - X\}$
(ii) $\{X^2Y - Y + X^2 - 1,\ XY + X - 1,\ XY^3 + Y^3 + Y + 2\}$.
(iii) $\{XY^2 - Y^2 + X - 1,\ X^3 + X^2,\ Y^2 - Y\}$.
(iv) $\{XY^2 - Y^2 - X + 1,\ XY^2 - Y,\ X^3 - X^2 - X + 1\}$. □

Exercise 6.2: (Macaulay's quartic curve) Consider the curve $X = t, Y = t^3, Z = t^4$.
(i) Use the Gröbner basis algorithm to show that its ideal is $\texttt{Ideal}(X^3 - Y, XY - Z)$ or $\texttt{Ideal}(Y - X^3, Z - X^4)$ (by eliminating t).
(ii) Re-do (i) but by using Macaulay's resultant (as suggested in the text).
(iii) Show by computing the appropriate Gröbner basis that its projective closure on the XY plane, the YZ plane and the XZ-plane are (respectively) $Y - X^3$, $Z^3 - Y^4$ and $Z - X^4$. □

Exercise 6.3: Show how Gröbner basis algorithms can be used to compute the resultant (with respect to X_1) of two polynomials in $K[X_1, \ldots, X_n]$. □

Exercise 6.4: Prove theorem 25. □

Exercise 6.5: With $\mathbf{X}, \mathbf{Y}$ as in (16), let $f_1, \ldots, f_m \in K[\mathbf{Y}]$.
(i) Show that the kernel of the map $\phi : K[\mathbf{X}] \to K[\mathbf{Y}]$ where $\phi(X_i) = f_i$ $(i = 1, \ldots, m)$ is given by

$$\texttt{Ideal}(X_1 - f_1, \ldots, X_m - f_m)_{K[\mathbf{X},\mathbf{Y}]} \cap K[\mathbf{X}].$$

(ii) With $I \subseteq K[\mathbf{Y}]$ an ideal, give an interpretation of

$$\texttt{Ideal}(I, X_1 - f_1, \ldots, X_n - f_n)_{K[\mathbf{X},\mathbf{Y}]} \cap K[\mathbf{X}].$$

□

Exercise 6.6: (Bayer)
Let $G = (V, E)$ be an undirected graph on the vertex set $V = \{1, \ldots, n\}$. We want to test if G is 3-colorable, *i.e.*, there is a 3-coloring of G such that adjacent vertices have distinct colors. Associate vertex i with the variable X_i and consider the system of polynomials

$$\{X_i^3 - 1,\ X_i^2 + X_i X_j + X_j^2 \ :\ i, j = 1, \ldots, n, (i, j) \in E\}.$$

Show that G is 3-colorable iff this system is consistent. □

Exercise 6.7:
Let $P(X)$ be the minimal polynomial of the algebraic number α, and let $\beta \in \mathbb{Q}(\alpha)$. Say, $\beta = A(\alpha)/B(\alpha)$ where $A, B \in \mathbb{Z}(X)$. Show that the minimal polynomial of β appears as a reduced Gröbner basis of an ideal intersection:

$$\texttt{Ideal}(P(X), B(X)Y - A(X)) \cap \mathbb{Q}[Y].$$

□

§7. Computing in Quotient Rings

Gröbner bases provide a simple tool for computing in residue class rings (equivalently, quotient rings). For an ideal $I \subseteq R$, let $S = R/I$ be the quotient ring and the canonical homomorphism from R to S be denoted

$$a \in R \longmapsto \overline{a} = a + I \in S.$$

The element $\overline{a}$ is called the "residue class" of a. Let $G \subseteq I$ be a reduced Gröbner basis for I, relative to some admissible ordering. First let us address the question of representing elements of S. For $a \in R$, we let $\texttt{nf}_G(a)$ represent $\overline{a}$. We write $\texttt{nf}_G(a) \cong \overline{a}$ to indicate this relation between a particular polynomial $\texttt{nf}_G(a)$ and an element $\overline{a}$ of R/I. Note that $\overline{a} = \overline{b}$ iff $a - b \in I$ iff $\texttt{nf}_G(a - b) = 0$ iff $\texttt{nf}_G(a) = \texttt{nf}_G(b)$. So there is a bijection between the elements of S and the set

$$\{\texttt{nf}_G(a) : a \in R\}.$$

Ring operations in S are easily simulated:

$$\begin{aligned} \overline{a} \pm \overline{b} &\cong \texttt{nf}_G(a) \pm \texttt{nf}_G(b) \\ \overline{a} \cdot \overline{b} &\cong \texttt{nf}_G(\texttt{nf}_G(a) \cdot \texttt{nf}_G(b)) \end{aligned}$$

where the $\pm$ and $\cdot$ on the right hand side are ordinary polynomial operations.

Lemma 26 *Let $I \subseteq R = K[X_1, \ldots, X_n]$ be an ideal with G as Gröbner basis.*

(i) *$S = R/I$ is a K-vector space.*

(ii) *Let $B = \{p \in \texttt{PP} : p = \texttt{nf}_G(p)\}$. Then $\overline{B} = \{\overline{p} : p \in B\}$ forms a basis for this vector space.*

(iii) *S is a finite dimensional K-vector space iff I is zero-dimensional.*

Proof. (i) is routine. (ii) follows from the above representation of S by the G-normal forms, since each normal form is of the form $\sum_{i=1}^{k} \alpha_i p_i$ with $\alpha_i \in K, p_i \in \texttt{PP}(X_1, \ldots, X_n)$ and $p = \texttt{nf}_G(p)$. To see (iii), recall that I is zero-dimensional iff for each i, there is a $f_i \in G$ with $\texttt{hterm}(f_i) \in \texttt{PP}(X_i)$. So if I

is zero-dimensional, we easily see that the X_i-degree of each $p \in B$ is less than $\deg(\mathtt{hterm}(f_i))$. This implies B is a finite set and S is finite dimensional. Conversely, if I is not zero-dimensional, there exists a variable X_i such that $X_i^\ell \in B$ for all $\ell \geq 0$. Then B is infinite and S is infinite dimensional.
Q.E.D.

Therefore, if I is zero-dimensional then we can construct the basis $\overline{B}$ (as represented by B) for the K-vector space S. Furthermore, we can construct the so-called *multiplication table* for this basis: for each $p, q \in B$, we only need to compute the G-normal form of pq. We are thus set up for computation in $S = R/I$.

Dimension of an Ideal. In general, for a prime ideal $P \subseteq K[\mathbf{X}] = R$, its *dimension* $\dim(P)$ is the transcendence degree[3] of R/I over K. If I is a general ideal, its *dimension* $\dim(I)$ is defined to be the maximum dimension of a prime ideal P that contains I. Let us relate dimension to another quantity: define (following Gröbner)

$$\delta(I) := \max\{|\mathbf{Y}| \ : \ \mathbf{Y} \subseteq \mathbf{X},\ I \cap K[\mathbf{Y}] = (0)\}.$$

It is immediate that $\delta(P) \leq \dim(P)$ since $P \cap K[\mathbf{Y}] = (0)$ implies that $\mathbf{Y}$ is a transcendental set over R/P. The following theorem (from Gröbner [74]) shows that they are in fact equal for all I.

Theorem 27 *The dimension of I is the largest cardinality of a set $\mathbf{Y} \subseteq \mathbf{X}$ such that $I \cap K[\mathbf{Y}] = (0)$. That is,* $\dim(I) = \delta(I)$.

This leads to a method to compute a subset of $\mathbf{X}$ that is a transcendental base for R/I: for each subset $\mathbf{Y} \subseteq \mathbf{X}$, compute a Gröbner basis G for I relative to an admissible ordering which is the lexicographic product $(\underset{\mathbf{X}-\mathbf{Y}}{\leq}, \underset{\mathbf{Y}}{\leq})$. Then $\mathbf{Y}$ is a transcendental set iff $G \cap K[\mathbf{Y}]$ is empty. We may systematically do this computation for $\mathbf{Y}$'s of larger cardinality before those of smaller cardinality, stopping at the first transcendent set $\mathbf{Y}$ we find. Note that the alternative "bottom-up" approach is to find a set transcendental $\mathbf{Y}$ and then try to extend it to $\mathbf{Y} \cup \{X_i\}$ for some $X_i \in \mathbf{X} \setminus \mathbf{Y}$. Unfortunately, this fails for non-prime ideals I. The following is an example from Cavaliere (unpublished lecture notes):

$$I = \mathtt{Ideal}(X_1^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4) \subseteq K[X_1, \ldots, X_4] \tag{18}$$

Then $I \cap K[X_2] = (0)$ and $I \cap [X_2, X_j] \neq (0)$ for $j \neq 2$. However, $I \cap K[X_3, X_4] = (0)$, and in fact $\dim(I) = 2$. See Weispfenning (in [170]) for further aspects of this problem.

Exercises

Exercise 7.1: Construct the multiplication table for $R/\mathtt{Ideal}(G)$ where $G = \{X^2 - X + 1, Y^2 - Y + 2X, XY - 2\}$. □

Exercise 7.2: Given g a polynomial and a set F of polynomials generating an ideal $I \subseteq R$, show how to decide if $\overline{g}$ is an invertible element in $S = R/I$. NOTE: g is invertible if there is some h such that $gh + I = R = \mathtt{Ideal}(1)$. □

[3] A subset $T \subseteq R/I$ is said to be *algebraically independent* over K if there is no polynomial relation $A(t_1, \ldots, t_m) \in K[t_1, \ldots, t_m]$ and elements $a_1, \ldots, a_m \in T$ such that $A(a_1, \ldots, a_m) = 0$. We say T is a *transcendental set* over K if every finite subset of T is algebraically independent over K. If T is a maximal cardinality transcental set, we call it a *transcendental base* of R/I over K. Although a transcendental base is non-unique, its cardinality is unique. This cardinality is called the *transcendence degree* of R/I over K.

Exercise 7.3: Verify the assertions on the example (18) of Cavaliere. □

§8. Syzygies

In this section, let $\mathbf{g} = [g_1, \ldots, g_m] \in R^m$ *be a fixed but arbitrary sequence of polynomials in* $R = K[X_1, \ldots, X_n]$. *We regard* $\mathbf{g}$ *as an* ordered basis *for* $\texttt{Ideal}(g_1, \ldots, g_m)$.

Definition: Let $\mathbf{a} = [a_1, \ldots, a_m] \in R^m$. The $\mathbf{g}$-*grade* of $\mathbf{a}$ is given by

$$\max\{\texttt{hterm}(a_i g_i) : i = 1, \ldots, m\}$$

where the maximum is with respect to the implicit admissible ordering $\underset{\texttt{A}}{<}$. We call $\mathbf{a}$ a *syzygy* of $\mathbf{g}$ if $\sum_{i=0}^{m} a_i g_i = 0$.

Let us motivate some syzygy computations that might be desired.

Let $\texttt{Syz}(\mathbf{g}) \subseteq R^m$ denote the set of syzygies of $\mathbf{g}$. Observe that $\texttt{Syz}(\mathbf{g})$ is a R-submodule of R^m.
By Hilbert's basis theorem for Noetherian modules (§XI.1), $\texttt{Syz}(\mathbf{g})$ is finitely generated. So the question to construct a finite basis of $\texttt{Syz}(\mathbf{g})$ naturally arises. *syzygy* Why would we be interested in syzygies in the first place? Well, it would tell us about any non-trivial relations among $g_1, \ldots, g_m$. If particular, if g_1 can be expressed as a combination of $g_2, \ldots, g_m$, then clearly any basis for $\texttt{Syz}(\mathbf{g})$ must contain a syzygy whose first component is a non-zero constant! Let us say that a finite set $F \subseteq R$ is a *minimal basis* (for the ideal it generates) if no proper subset of F generates $\texttt{Ideal}(F)$. Thus computing a syzygy basis yields a method (possibly overkill) to decide if F is a minimal basis.

Next, let the map $\varphi : R^m \to R$ given by $\varphi(\mathbf{a}) = \sum_{i=1}^{m} a_i g_i$ be called the *canonical map* determined by $\mathbf{g}$: it is clearly an R-module homomorphism (§XI.1). So $\texttt{Syz}(\mathbf{g})$ is simply the kernel of φ. Then the following sequence[4] is *exact*:

$$R^m \xrightarrow{\varphi} R \xrightarrow{\psi} R/I \longrightarrow 0$$

where I is the ideal generated by $g_1, \ldots, g_m$ and ψ is the canonical map from R to R/I. Again, $\texttt{Syz}(\mathbf{g})$ has a finite basis, say $\mathbf{h} = [h_1, \ldots, h_t]$ ($h_i \in R^m$). We can thus look at the set $\texttt{Syz}(\mathbf{h})$ of syzygies of $\mathbf{h}$, where $\texttt{Syz}(\mathbf{h})$ is called the *second syzygy module* of I. Again, why would we be interested in the second syzygy module? As before, it would tell us of any non-trivial relations among the generators of the first syzygy module (and similar to the minimal basis application, might be used for a similar purpose).

Again, we may introduce another canonical map $\varphi' : R^t \longrightarrow R^m$ such that the following extended sequence is exact:

$$R^t \xrightarrow{\varphi'} R^m \xrightarrow{\varphi} R \xrightarrow{\psi} R/I \longrightarrow 0.$$

But the $\texttt{Syz}(\mathbf{h})$ also has a finite basis (say with s generators) with canonical map φ'', and we can extend this exact sequence again. The Hilbert syzygy theorem says that such extensions of the exact sequence terminate in that we eventually reach a final canonical map φ'' with the trivial kernel (0) (hence only trivial syzygies). So we have a finite exact sequence,

$$0 \longrightarrow R^s \xrightarrow{\varphi''} \cdots \longrightarrow R^t \xrightarrow{\varphi'} R^m \xrightarrow{\varphi} R \xrightarrow{\psi} R/I \longrightarrow 0,$$

[4] A (finite or infinite) sequence of R-module homomorphisms

$$\cdots \xrightarrow{\varphi_{i-1}} M_{i-1} \xrightarrow{\varphi_i} M_i \xrightarrow{\varphi_{i+1}} \cdots$$

is *exact* if the image of φ_i equals the kernel of φ_{i+1} for all i. (The kernel of a homomorphism is the set elements that maps to 0.)

called a *free resolution* of I. There are clearly arbitrary choices in the definition of this sequence (the ordering in $\mathbf{g} = [g_1, \ldots, g_m]$, the choice of syzygy bases, etc). Nevertheless, it contains certain invariants that depend only on I (e.g., [222, ch.VII] and also lecture XIII). It is therefore important to develop computational tools for computing with syzygies.

Syzygy basis for a Gröbner basis. We will initially assume that $\mathbf{g}$ is a Gröbner basis.

For $1 \leq i < j \leq m$, the S-polynomial of g_i and g_j can be written in the form

$$S(g_i, g_j) = \alpha_{i,j} g_i + \beta_{i,j} g_j$$

where $\alpha_{i,j}, \beta_{i,j}$ are monomials. Then by the extended standard characterization of Gröbner bases,

$$S(g_i, g_j) = \sum_{\mu=1}^{m} h_\mu^{i,j} g_\mu$$

where $\mathtt{hterm}(S(g_i, g_j)) \underset{\mathtt{A}}{\geq} \mathtt{hterm}(h_\mu^{i,j} g_\mu)$ for each μ.

Define

$$T(i,j) := [h_1^{i,j}, h_2^{i,j}, \ldots, h_i^{i,j} - \alpha_{i,j}, \ldots, h_j^{i,j} - \beta_{i,j}, \ldots, h_m^{i,j}].$$

In other words, the μ-th entry of $T(i,j)$ is $h_\mu^{i,j}$ except when $\mu = i$ or $\mu = j$, in which case the entries are $h_j^{i,j} - \alpha_{i,j}$ and $h_j^{i,j} - \beta_{i,j}$, respectively. Clearly, $T(i,j)$ is a syzygy of $\mathbf{g}$. The grade of $T(i,j)$ is equal to

$$\mathtt{hterm}(\alpha_{ij} g_i) = \mathtt{hterm}(\beta_{ij} g_j)$$

which is equal to $\mathrm{LCM}(\mathtt{hterm}(g_i), \mathtt{hterm}(g_j))$. It is also important for the following proof to notice that the grade of $T(i,j)$ is attained at the ith and jth components *and nowhere else*. grade!of a vector of polynomials

Theorem 28 (Spear, Schreyer) *If* $\mathbf{g}$ *is an ordered Gröbner basis then*

$$\{T(i,j) : 1 \leq i < j \leq m\}$$

forms a basis for the module $\mathtt{Syz}(\mathbf{g})$.

Proof. We have noted that the the $T(i,j)$'s belong to $\mathtt{Syz}(\mathbf{g})$. To show that they generate $\mathtt{Syz}(\mathbf{g})$, let $\mathbf{a} = [a_1, \ldots, a_m] \in \mathtt{Syz}(\mathbf{g})$. Let p be the grade of $\mathbf{a}$. Without loss of generality, suppose that k of the components of $\mathbf{a}$ achieve this grade so that, for some $1 \leq j_1 < j_2 < \cdots < j_k \leq m$, we have

$$p = \mathtt{hterm}(g_{j_1} a_{j_1}) = \mathtt{hterm}(g_{j_2} a_{j_2}) = \cdots = \mathtt{hterm}(g_{j_k} a_{j_k}).$$

Clearly $k \geq 2$ since $\mathtt{hterm}(g_{j_1} a_{j_1})$ must be cancelled in the expression $\sum_{i=1}^{m} a_i g_{j_i} = 0$. We will show that for some monomial γ,

$$\mathbf{b} = \mathbf{a} - \gamma \cdot T(j_1, j_2)$$

such that either $\mathbf{b}$ has grade $\underset{\mathtt{A}}{<} p$ or else less than k components of $\mathbf{b}$ attain the grade p. Repeating this, we must eventually achieve the trivial syzygy $[0, 0, \ldots, 0]$ since $\underset{\mathtt{A}}{<}$ is well-founded. This would prove that the set of $T(i,j)$ generates $\mathtt{Syz}(\mathbf{g})$.

In fact, since $p = \mathtt{hterm}(a_{j_1} g_{j_1}) = \mathtt{hterm}(a_{j_2} g_{j_2})$, we see that

$$\mathrm{LCM}(\mathtt{hterm}(g_{j_1}), \mathtt{hterm}(g_{j_2}))$$

must divide p. Hence $\gamma \cdot T(j_1, j_2)$ has grade p, for some monomial γ. Moreover, we may choose γ so that the j_1th component of

$$\mathbf{b} = \mathbf{a} - \gamma \cdot T(j_1, j_2)$$

is $\underset{\mathtt{A}}{<} p$. Since the grade p is attained *only* at the j_1th and j_2th components of $T(j_1, j_2)$, this means that the number of components of $\mathbf{b}$ that attains grade p is strictly reduced. Further, since the grade of $\mathbf{b}$ is $\underset{\mathtt{A}}{\leq} p$, our inductive step is complete. This completes our proof.

Q.E.D.

With $s = \binom{m}{2}$, let

$$U \in R^{s \times m} \tag{19}$$

denote matrix whose rows are the syzygies $T(i, j)$. Note that $T(i, j)$ can be computed by running any normal form algorithm on the S-polynomial $S(g_i, g_j)$, and doing some straightforward bookkeeping.

Syzygy basis for an arbitrary ordered basis. Suppose now $\mathbf{f} = [f_1, \ldots, f_r] \in R^r$ is an arbitrary ordered ideal basis. We again want to construct a module basis for $\mathtt{Syz}(\mathbf{f})$.

Let $\mathbf{g} = [g_1, \ldots, g_m]$ be a Gröbner basis for the ideal generated by $\{f_1, \ldots, f_r\}$. We may assume $m \geq r$ and $g_i = f_i$ for $i = 1, \ldots, r$. Let

$$G = [I_r | O]$$

be an $r \times m$ matrix consisting of an $r \times r$ identity matrix I_r and O is a $r \times (m - r)$ matrix of zeros. Then

$$\mathbf{f}^T = G\mathbf{g}^T$$

where $(\cdot)^T$ denotes matrix transpose. Since each g_i is a linear combination of $f_1, \ldots, f_r$, there is an $m \times r$ matrix F with entries in R such that

$$\mathbf{g}^T = F\mathbf{f}^T.$$

We may further assume the first r rows of F form the identity matrix I_r. Hence

$$GF = I_r.$$

With U as in (19), we obtain

$$U \cdot \mathbf{g}^T = 0.$$

Theorem 29 *The matrix $B := U \cdot F$ is an $s \times r$ matrix whose rows form a basis for* $\mathtt{Syz}(\mathbf{f})$.

Proof. Suppose $\mathbf{a} \in \mathtt{Syz}(\mathbf{f})$. Then $\mathbf{a}G$ is a syzygy of $\mathbf{g}$ (by our assumption on G, $\mathbf{a}G$ is just the padding out of $\mathbf{a}$ with 0's). Since the rows of U form a basis for these syzygies, there exists an s-tuple $\mathbf{b}$ such that $\mathbf{a}G = \mathbf{b}U$. Hence:

$$\mathbf{a} = \mathbf{a}GF = \mathbf{b} \cdot UF = \mathbf{b}B.$$

This proves the rows of B are indeed a basis for $\mathtt{Syz}(\mathbf{f})$. **Q.E.D.**

Note that the matrix F, and hence B, can be constructed from $\mathbf{f}$: the entries of F can be built up during the construction of the Gröbner basis $\mathbf{g}$ from $\mathbf{f}$.

Application to Ideal Intersection. Let $f_1, \ldots, f_r, g_1, \ldots, g_s$ be given. We want the ideal intersection

$$I = \texttt{Ideal}(f_1, \ldots, f_r) \cap \texttt{Ideal}(g_1, \ldots, g_s).$$

Write $\mathbf{h} := [f_1, \ldots, f_r, g_1, \ldots, g_s]$. We construct B, the matrix whose rows form a basis for $\texttt{Syz}(\mathbf{h})$. If C is the matrix consisting of the first r columns of B, and for each row $\mathbf{a}$ of C, we form the dot product $\mathbf{a} \cdot \mathbf{f}^T$ (where $\mathbf{f} = [f_1, \ldots, f_r]$), then we claim that the set S of these dot products $\mathbf{a} \cdot \mathbf{f}^T$ generates I. To see that each $\mathbf{a} \cdot \mathbf{f}^T \in S$ is in I, note that for some $\mathbf{b}$, the concatenation $\mathbf{a}; \mathbf{b}$ is a row of B. Hence $\mathbf{a} \cdot \mathbf{f}^T + \mathbf{b} \cdot \mathbf{g}^T = 0$ where $\mathbf{g} = [g_1, \ldots, g_s]$. Hence $\mathbf{a} \cdot \mathbf{f}^T \in \texttt{Ideal}(g_1, \ldots, g_s) \cap \texttt{Ideal}(f_1, \ldots, f_r)$. Conversely, any element of I has the form $\mathbf{a}' \cdot \mathbf{f}^T$ and also the form $\mathbf{b}' \cdot \mathbf{g}^T$. Hence $[\mathbf{a}'; -\mathbf{b}']$ is a syzygy of $\mathbf{h}$. Thus $\mathbf{a}'$ is a linear combination of the rows of C, and so $\mathbf{a}' \cdot \mathbf{f}^T$ is a linear combination of elements of the form $\mathbf{a} \cdot \mathbf{f}^T \in S$.

Final Remarks. This section touches on an important direction in the development of Gröbner bases: the whole subject is capable of generalization to R-submodules of R^m, for some $m \geq 1$. (It is a generalization because an ideal of R is just a submodule of R^m with $m = 1$.) Let us briefly indicate how it goes. We may define a "generalized power product" to be an element of the form $p \cdot e_i$ where $p \in \texttt{PP}(\mathbf{X})$ and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the ith elementary m-vector with a 1 in the ith position. Write $\texttt{PP}_m(\mathbf{X})$ for these generalized power products. (So the original notation PP is just $\texttt{PP}_1$.) We say a total ordering $\underset{\texttt{A}}{\leq}$ on $\texttt{PP}_m(\mathbf{X})$ is *admissible*ordering!admissible if for all $\mathbf{g}, \mathbf{h} \in \texttt{PP}_m(\mathbf{X})$ and all $p \in \texttt{PP}(\mathbf{X})$, we have (1) $\mathbf{g} \underset{\texttt{A}}{\leq} p\mathbf{g}$ and (2) $\mathbf{g} \underset{\texttt{A}}{\leq} \mathbf{h}$ implies $p\mathbf{g} \underset{\texttt{A}}{\leq} p\mathbf{h}$. One then shows that this is a well-ordering, and extends this total ordering to a linear quasi-ordering on R^m. We may define the notion of the headterm of $\mathbf{g}$ and reduction $\mathbf{f} \xrightarrow{\mathbf{g}} \mathbf{h}$, and hence a notion of normal forms and Gröbner basis. Alternatively, the standard characterization can be used to define Gröbner basis. Finally, an analogue of Buchberger's algorithm can be shown.

EXERCISES

Exercise 8.1: Describe the necessary modifications to the normal form algorithm and Buchberger's algorithm in order to carry out the computation of a general syzygy basis above. □

Exercise 8.2: Compute a syzygy basis for the following ordered basis

$$[g_1, g_2, g_3] = [X_1^2 - X_2X_4, X_1X_2 - X_3X_4, X_1X_3 - X_2^2],$$

say, assuming a pure lexicographic ordering. □

Exercise 8.3: Describe some efficient method to solve the following: given F, compute a subset $H \subseteq F$ that is a minimal basis for $\texttt{Ideal}(F)$. □

Exercise 8.4: (Gianni-Trager-Zacharias)

(i) If $I, J \subseteq R$ are ideals and t is a new indeterminate, show that

$$(I \cap J) = (\texttt{Ideal}(tI, (t-1)J) \cdot R[t]) \cap R.$$

(ii) Use this to give an alternative algorithm for computing ideal intersection.

□

Exercise 8.5: If $\underset{\mathtt{A}}{\leq}$ is an admissible ordering on $\mathtt{PP}$, show that the following (still denoted $\underset{\mathtt{A}}{\leq}$) are admissible orderings on $\mathtt{PP}_m$. Let $p, q \in \mathtt{PP}$.
(i) Define $p \cdot e_i \underset{\mathtt{A}}{\leq} q \cdot e_j$ iff $i < j$ or else, $p \underset{\mathtt{A}}{\leq} q$.
(ii) Define $p \cdot e_i \underset{\mathtt{A}}{\leq} q \cdot e_j$ iff $p \underset{\mathtt{A}}{<} q$ or else, $i \leq j$. □

Exercise 8.6: Carry out the above outline of a generalization of Gröbner basis theory to submodules of R^m. □

Exercise 8.7: Generalize the above syzygy algorithm. If $G = [\mathbf{g}_1, \ldots, \mathbf{g}_k]$ where $\mathbf{g}_i \in RR^m$, we may define $\mathbf{a} = [a_1, \ldots, a_k] \in \mathbb{R}^k$ to be a syzygy of G if $\sum_{i=1}^{k} a_i \mathbf{g}_i = \mathbf{0}$ (as in the second syzygy module). Then the set $\mathtt{Syz}(G)$ of syzygies of G is a submodule of R^k. Develop an algorithm to compute a basis for $\mathtt{Syz}(G)$. □

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents

Lecture XIII
Bounds in Polynomial Ideal Theory

Despite Hilbert's acceptance of non-constructive proofs and non-finitistic arguments in mathematics[1], the constructive approach to mathematics was not entirely ignored by the Göttingen school. Thus, Grete Hermann's highly constructive doctoral dissertation (see the quote from Seidenberg above) was completed under Emmy Noether. In constructive approaches to a problem, the derivation of explicitly bounds is often a first step towards an algorithm. Or, from the viewpoint of complexity theory, such bounds are essential towards pinning down the inherent complexity of a computational problem. To illustrate this, consider an example from Seidenberg [12]: for given polynomial ideals A and B, suppose we want to compute a number ρ such that $A : B^\rho = A : B^{\rho+1}$. An obvious procedure is to compute the sequence

$$A_1 < A_2 < \cdots < A_k < \cdots$$

of ideals where $A_k = A : B^k$ until the first time $A_k = A_{k+1}$. This k is our desired ρ. This procedure always halts, by the ascending chain condition. Seidenberg and Hermann do not consider this procedure "effective" until we derive a bound on ρ, expressed as a computable function of numerical parameters in (the representation of) A, B. Such a bound is provided in [12].

In this lecture, we address another such bound that is fundamental in constructive polynomial ideal theory.

Let $G(n, d)$ denote the minimum degree of Gröbner bases of ideals generated by polynomials of degree at most d in the ring $R = \mathbb{Q}[X_1, \ldots, X_n]$ of rational polynomials. We will show double-exponential upper and lower bounds on $G(n, d)$: for large n,

$$G(n, d) = d^{2^{\beta n}}, \qquad 0.16 < \beta \leq 1. \tag{1}$$

There is a close connection between $G(n, d)$ and two other important bounds $I(n, d)$ (ideal membership bound) and $S(n, d)$ (syzygy basis bound) in the theory of polynomial ideals.

Sharper bounds than (1) are known: for any $\varepsilon > 0$ and n large enough, we have

$$0.5 - \varepsilon < \beta < 0.79.$$

This upper bound is from Lazard [7], but our lecture presents the weaker result $\beta \leq 1$ from Dubé [3, 4]. See also Giusti [5] and Möller and Mora [10]). The original lower bound is from Mayr and Meyer [9], but here we use a simpler construction from Yap [13]. Bayer and Stillman [1] first observed that the construction of Mayr-Meyer leads to lower bounds on $S(n, d)$. Huynh [6] gives other applications of these lower bound constructions.

> *Unless otherwise noted, R refers to the ring $K[\mathbf{X}] = K[X_1, \ldots, X_n]$ for some field K. For the lower bound proof, we further assume $K = \mathbb{Q}$. We normally assume the standard grading of $K[\mathbf{X}]$, so "homogeneous polynomials" have the standard sense (see §2).*

§1. Some Bounds in Polynomial Ideal Theory

[1] Hilbert's program in metamathematics and the axiomatic approach in his later life may be seen as attempts to justify the non-constructive approaches, and a direct response to to the criticisms of Kronecker and Brouwer.

Three bounds arise naturally in the constructive theory of ideals in R:

$$I(n,d), \quad S(n,d), \quad D(n,d).$$

We call them (respectively) the *ideal membership bound*, *Syzygy basis bound* and *grob basis bound*. They are defined as follows: let $f_1, \ldots, f_m \in R = \mathbb{Q}[X_1, \ldots, X_n]$ with (total) degree $\deg f_i \leq d$. Then the *ideal membership degree bound* $I(n,d)$ is the least value such that for all such f_i's whenever $f_0 \in \texttt{Ideal}(f_1, \ldots, f_m)$ then

$$f_0 = \sum_{i=1}^{m} a_i f_i, \qquad a_i \in R,\ \deg(a_i f_i) \leq I(n,d) + \deg(f_0).$$

The *syzygy basis bound* $S(n,d)$ is the least value such that for all such f_i's there exists a syzygy basis for the module of syzygies of $(f_1, \ldots, f_m)$, where each syzygy in the basis has degree at most $S(n,d)$. The degree of a syzygy $[h_1, \ldots, h_m]$ is the maximum degree of the h_i's. Finally, the *Gröbner basis bound* $D(n,d)$ is the least value such that for all such f_i's and any admissible ordering $\underset{\texttt{A}}{\leq}$, there is a Gröbner basis with respect to $\underset{\texttt{A}}{\leq}$ whose members are polynomials of degree at most $D(n,d)$.

All the three bounds are closely related, and are essentially doubly exponential [13]. For instance, $D(n,d) \geq S(n,d)$. Lazard (1982,1991 [7]) has shown

$$\begin{aligned} S(n,d) &\leq d^{2^{\beta n}}, \\ I(n,d) &\leq d^{2^{\beta n + O(\log n)}}, \end{aligned}$$

where $\beta = \log_4 3 < 0.79$. Lazard indicated a similar upper bound for $G(n,d)$, superceding the upper bound shown in this lecture. Nevertheless, the proof to be presented has independent interest. These bounds are all tight up to constant factors in the second exponent. Yap [13] shows that $I(n,d)$, $D(n,d)$ and $S(n,d)$ are each lower-bounded by the function

$$d^{2^{\beta n}}, \quad \beta \sim 0.5. \tag{2}$$

The notation "$\beta \sim 0.5$" means that the right-hand side has an implicit dependence on n and β approaches 0.5 as $n \to \infty$. The lower bound was originally[2] shown by Mayr-Meyer [9] with $\beta \sim 0.1$. In this lecture, we give a simple construction to achieve $\beta \sim 0.2$. This construction can be sharpened to the cited result.

Remarks.
1. Essentially all known doubly exponential lower bounds proofs in polynomial ideal theory use variants of the construction in the proof of (2). The construction allows us to simulate an exponential space Turing machine, and leads to this conclusion: *any Turing machine for deciding the ideal membership problem must use space c^n for some $c > 0$ for infinitely many n.* The reason is that such systems can simulate counter-machines with double-exponentially large counters, and this corresponds exactly to single-exponential space-bounded Turing machines. See [9] for details.
2. A double-exponential degree bound immediately leads to a double-exponential complexity bound on ideal membership problem: to test if f_0 belongs to the ideal generated by $f_1, \ldots, f_m$, we can set up a suitable large linear system to be solved (Exercise). Similarly, this leads to a double-exponential time algorithm for computing Gröbner bases (Exercise).
3. The fact the coefficients comes from $\mathbb{Q}$ is only needed in the lower bound argument. The lower

[2] Several authors (including Bayer-Stillman, Lazard, etc) noted that the original Mayr-Meyer construction can be improved from using about $15n$ variables to using about $10n$ variables for counting up to d^{2^n}. This implies the stated bound.

bound applies to any field K.
4. If we consider the polynomial ring $R_0[X_1, \ldots, X_n]$ where R_0 is a general ring, such universal bounds $G(n, d)$, etc, may not exist. This is the case when $R_0 = \mathbb{Z}$ or if $R_0 = K[Z]$ where K is a field and Z an indeterminate. See Li [8].

Exercises

Exercise 1.1: By writing $f_0 = \sum_{i=1}^m a_i f_i$ as a linear system in the unknown coefficients of $a_i \in R$, show that ideal membership in $\mathbb{Q}[\mathbf{X}]$ can be solved in double-exponential time. □

Exercise 1.2: Suppose we have a Gröbner degree bound $G(n, d, m)$ where n, d are as above and m is the number of generators. Show how to eliminate the dependence on m. □

.

Exercise 1.3: Use the bound on $G(n, d)$ to give a double-exponential bound on the bit-complexity of a *suitable variation* of Buchberger's algorithm. HINT: let $B = \binom{G(n,d)+n-1}{n-1}$. Restrict the polynomials in the Gröbner basis H that is being constructed to satisfy: (1) $|H| \leq B$, (2) each $f \in H$ has degree at most B. Use an efficient normal form algorithm (see exercise, §XII.3). □

Exercise 1.4: (Buchberger, Lazard) Show that $G(2, d) \leq d^2$. □

§2. The Hilbert-Serre Theorem

By way of proving the Hilbert-Serre theorem, we introduce the Hilbert polynomial for graded modules.

The concept of homogeneous elements can be generalized as follows. (These definitions are slightly more general than we strictly need.) Let $(G, +, 0)$ be an Abelian semigroup (so $+$ is associative, commutative and 0 is the identity for $+$). Typically, $G = \mathbb{N}$, $G = \mathbb{Z}$ or $G = \mathbb{Z}^n$. A *G-grading* (or, simply *grading*) of a ring R is a (weak) direct sum decomposition of R,

$$R = \bigoplus_{d \in G} R_d$$

where each R_d is an additive subgroup of R and $R_d R_e \subseteq R_{d+e}$ for all $d, e \in G$. Here, the decomposition is weak because in the decomposition of an element $u = \sum_d u_d \in R$ ($u_d \in R_d$) only finitely many of the u_d's are non-zero. A *graded ring* is a ring together with a grading. A *G-graded module* M is an R-module with a weak direct sum decomposition, $M = \bigoplus_{d \in G} M_d$ such that M_d is an additive subgroup of M and $R_d M_e \subseteq M_{d+e}$. Elements of M_d are said to be *homogeneous* of *degree* d. Note that 0 is homogeneous of every degree d. In fact, $M_d \cap M_e = \{0\}$ for all $d \neq e$. If an element $u \in M$ is decomposed as $u = \sum_d u_d$, then each non-zero u_d is called a *homogeneous component* of u. A *graded submodule* or *homogeneous submodule* N of M is a submodule $N \subseteq M$ that is generated by homogeneous elements. Equivalently, for all $u \in I$, each homogeneous component of u is in I (cf. §XII.5). Note that all these terms (homogenous element, graded submodule, etc) for modules apply to the ring R, since we may view R as an R-module.

Example: (1) Let $G = \mathbb{R}$ and w be a weight function on $\mathtt{PP}(\mathbf{X})$ (§XII.1). Say a polynomial f is *homogeneous* of degree d if each power product u in f has weight $w(u) = d$. Let R_d comprise all w-homogeneous polynomials of degree d. This gives rise to the *w-grading* of R. If $w = (1, 1, \ldots, 1)$ then we get the *standard grading* of R. In this case, elements of R_d is said to be homogeneous in the *standard sense*.
(2) Let $G = \mathbb{N}^n$ and fix an admissible ordering on $\mathtt{PP}(\mathbf{X})$ (so that $\mathtt{hterm}(f)$ is defined for $f \in R$). Then $f \in R_e$ iff $\log(\mathtt{hterm}(f)) = e \in \mathbb{N}^n$ (see §XII.1 for definition of $\log(\cdot)$).
(3) (Matsumura) Let $G = \mathbb{N}$ and $M = R[X, Y, Z]/\mathtt{Ideal}(f)$ where $f = aX^\alpha + bY^\beta + cZ^\gamma$. Then the weight function $w(\overline{X}) = \beta\gamma$, $w(\overline{Y}) = \alpha\gamma$ and $w(\overline{Z}) = \alpha\beta$ yields a grading on M.
(4) If $I \subseteq R$ is an ideal, then $\mathtt{gr}_I(R) = \bigoplus_{n\geq 0} I^n/I^{n+1}$ is an $\mathbb{N}$-graded ring. ■

We introduce homomorphisms on graded modules. Given two G-graded R-modules M, M', we say a module homomorphism

$$\phi : M \to M'$$

is *homogeneous* of degree d if $\phi(M_e) \subseteq M'_{e+d}$ for each $e \in G$. (By module homomorphism, we mean $\phi(u + v) = \phi(u) + \phi(v)$ and $\phi(au) = a\phi(u)$ where $u, v \in M$, $a \in R$.) In particular, if $M = R$ and suppose $M' = R'$ is another ring then ϕ becomes a homogeneous ring homomorphism of degree d. For instance, if $R = R' = K[X_1, \ldots, X_n]$ with the standard grading, then the map ϕ defined by $\phi(u) = uX_n$ for all $u \in M$ is homogeneous of degree 1. (This map will be used in the proof below.) The following is straightforward and left to an exercise.

Lemma 1 *Let $M = \bigoplus_{d\in G} M_d$ be a G-graded R-module.*

(i) *A homogeneous submodule N of M is a G-graded R-module with grading $\bigoplus_{d\in G}(N \cap M_d)$.*

(ii) *If N is homogeneous, the module difference $M - N$ has grading $\bigoplus_{d\in G}(M_d - N)$.*

(iii) *If $\phi : M \to M'$ is homogeneous, then the kernel of ϕ is a homogeneous submodule of M and the image of ϕ is a homogeneous submodule of M'.*

(iv) *Also, the co-kernel $\mathtt{coKer}\,\phi = M' - \phi(M)$ (module difference) is a graded module.*

We state the Hilbert-Serre theorem:

Theorem 2 *Let $R = K[X_1, \ldots, X_n]$ have the standard grading, and M be a finitely generated graded module over R, $M = \bigoplus_{d=-\infty}^{+\infty} M_d$. Then M_d is a K-vector space of finite dimension $dim_K(M_d)$, and there is a polynomial $\Phi_M(z)$ of degree $\leq n - 1$ with integer coefficients and a constant d_0 such that for all $d \geq d_0$, $\Phi_M(d) = dim_K(M_d)$.*

Remarks.

1. The polynomial $\Phi_M(z)$ is called the *Hilbert polynomial* of the graded module M. The smallest constant d_0 for which the theorem holds is called the *Hilbert constant* of M and denoted h_M. This polynomial contains important geometric data as noted next. See Bayer-Stillman [2] for computations of this polynomial.
2. Let I be a homogeneous ideal of the polynomial ring R. Two main applications of this theorem are when $M = I$ and when $M = R/I$. For $M = I$, each additive group M_d is the K-vector space I_d comprising the homogeneous polynomials in I of degree d. The degree of the Hilbert polynomial $\Phi_M(z)$ in this case is the dimension of the projective zero set $\textsc{Zero}(I) \subseteq \mathbb{P}^{n-1}(K)$.

3. Again let M be an ideal $I \subseteq R$. The *degree* of the zero set $\text{Zero}(I)$ is defined to be $d!$ times the leading coefficient of $\Phi_I(z)$. The "degree" of a plane algebraic curve V_1 is intuitively clear: it is the maximum number of points (multiplicity counted) obtainable by intersecting V_1 with an arbitrary line L. In general, the degree of an algebraic set $V_d \subseteq \mathbb{A}^n(K)$ of dimension $d \leq n$ may be defined as the maximum number of points obtainable by intersecting V_d with an arbitrary $(n-d)$-dimensional hyperplane L. The precise definition here is complicated because of the phenomenon of multiplicity; historically, there have been many erroneous definitions. The algebraic notion of degree via the Hilbert polynomial has no ambiguity although the geometric intuition is lost.

We make some preparatory remarks for the proof.

1. *Every polynomial $p(z)$ of degree $\leq d$ can be written in the form*

$$p(z) = c_0\binom{z}{d} + c_1\binom{z}{d-1} + \cdots + c_{d-1}\binom{z}{1} + c_d \tag{3}$$

for suitable coefficients c_i. If $p(z)$ has integer coefficients then each c_i is integer. This is clear for $d = 0$ or 1. For $d > 1$, we note that

$$z^d = d!\binom{z}{d} + l.o.t.$$

where '*l.o.t.*' (lower order terms) refers to terms in z of degree less than d. Then, writing $p(z) = a_0 z^d + \cdots + a_{d-1}z + a_d$, we obtain at once

$$p(z) = a_0 d!\binom{z}{d} + l.o.t.$$

So we can choose $c_0 = a_0 d!$. By induction, the polynomial in the *l.o.t.* of $p(z)$ can be put in the desired form, giving the constants $c_1, \ldots, c_d$. This proves that $p(z)$ has the form in equation (3). Moreover, if $p(z)$ has integer coefficients, then c_0 is an integer and the *l.o.t.*'s of $p(z)$ also has integer coefficients. By induction, we conclude that $c_1, \ldots, c_d$ are also integers.

2. *If $p(z)$ is a polynomial and $p(z')$ is an integer for all large enough integers z', then $p(z)$ has integer coefficents.* The result is true for $d = 0$. We now apply the identity

$$\binom{n+1}{d} - \binom{n}{d} = \binom{n}{d-1}$$

to equation (3) to obtain

$$p(z+1) - p(z) = c_0\binom{z}{d-1} + c_1\binom{z}{d-2} + \cdots + c_{d-2}\binom{z}{1} + c_{d-1}.$$

But $p(z+1) - p(z)$ is a polynomial of degree less than d which evaluates to an integer for z large enough. By induction, therefore, $c_0, \ldots, c_{d-1}$ are integers. Pick z to be a large enough multiple of $d!$ so that $p(z)$ is integer. Then equation (3) shows that $p(z) - c_d$ is an integer. We conclude that c_d is also integer.

3. Suppose N is a proper submodule of M. A *composition series* from N to M is a strictly increasing sequence of submodules of the form

$$N = M_0 < \cdots < M_k = M$$

which cannot be extended in the sense that there is no submodule M' such that $M_i < M' < M_{i+1}$ for any i. We call k the *length* of this series. If $N = \{0\}$, we call it a composition series of M.

By definition, composition series are finite, and some modules may have no composition series. For instance, the module $\mathbb{Z}$ of integers has no composition series (why?). By the Jordan theorem [15, p. 159], every composition series of M has the same length, which we define to be the *length* $\ell(M)$ of M. For instance, the module $\{0\}$ has length zero. By definition, $\ell(M) = \infty$ if M has no composition series. If $N < M$ and $\ell(M) < \infty$, it follows easily that

$$\ell(M - N) = \ell(M) - \ell(N).$$

[Just as modules are generalizations of vector spaces, so dimension of vector spaces becomes length of modules.]

4. *Let*

$$0 \xrightarrow{f_0} E_1 \xrightarrow{f_1} E_2 \xrightarrow{f_2} \cdots \xrightarrow{f_{n-1}} E_n \xrightarrow{f_n} 0$$

be an exact sequence of R-modules where each E_i has length $\ell(E_i)$. Then the alternating sum

$$\ell(E_1) - \ell(E_2) + \cdots (-1)^{n-1}\ell(E_n) \tag{4}$$

is equal to zero. Note that $\mathtt{Ker}\, f_{i+1} = \mathtt{Im} f_i$ because the sequence is exact at E_{i+1}. But $\mathtt{Im} f_i$ is isomorphic to $E_i/\mathtt{Ker}\, f_i$, so $\ell(\mathtt{Im} f_i) = \ell(E_i) - \ell(\mathtt{Ker}\, f_i)$. Therefore

$$\ell(E_i) = \ell(\mathtt{Ker}\, f_i) + \ell(\mathtt{Ker}\, f_{i+1}).$$

The desired sum (4) becomes

$$\ell(\mathtt{Ker}\, f_1) + \ell(\mathtt{Ker}\, f_2) - \ell(\mathtt{Ker}\, f_2) - \ell(\mathtt{Ker}\, f_3) + \cdots + (-1)^{n-2}\ell(\mathtt{Ker}\, f_n) + (-1)^{n-1}\ell(E_n). \tag{5}$$

The non-extreme terms of this sum cancel out, so it is equal to $\ell(\mathtt{Ker}\, f_1) + (-1)^{n-2}\ell(\mathtt{Ker}\, f_n) + (-1)^{n-1}\ell(E_n)$. But $\mathtt{Ker}\, f_n = E_n$ and $\ell(\mathtt{Ker}\, f_1) = 0$ (since f_1 is injective). Thus the entire sum is 0, as we wanted shown.

5. For any R-module M, we define its *order* or *annihilator* to be the set

$$\mathtt{Ann}\, M = \{a \in R \,:\, aM = \{0\}\}.$$

Note that $\mathtt{Ann}\, M$ is a homogeneous ideal of R. Now let I be any subideal of $\mathtt{Ann}\, M$ and let $\overline{R} = R/I$. We view M as a $\overline{R}$-module in the natural way: if $\overline{a} \in \overline{R}$, $u \in M$, we define $\overline{a}u$ as the element au, where $a \in R$ and $\overline{a} = a + I$. (Check that this is well-defined.) Any R-submodules of M can be viewed as a $\overline{R}$-submodule of M, and vice-versa. Thus we can reduce the study of M as a R-module to the study of M as an $\overline{R}$-module. This important technique will be used in our proof. Typically, we take $I = \mathtt{Ann}\, M$. See [15, p. 141].

Proof of the Hilbert-Serre Theorem. We first claim that each M_d has finite dimension as a K-vector space. Since M is finitely generated over R, let $\{y_1, \ldots, y_k\}$ be a basis for M. Since M is graded, each y_i may be assumed to be homogeneous (if not, we can replace y_i be its set of homogeneous components). Then every element of M_d has the form

$$\sum_{i=1}^{k} u_i y_i, \quad (u_i \in R)$$

where $u_i \neq 0$ implies y_i has degree $\deg(y_i) \leq d$ and u_i is a homogeneous polynomial of degree $d - \deg(y_i)$. Hence M_d can be generated as a K-vector space by elements of the form py where $y \in \{y_1, \ldots, y_k\}$ and p is a power product of degree $d - \deg(y)$. Our claim about M_d follows since there are only finitely many elements in this form py.

Now we prove the theorem by induction on the number n of variables. First, assume $n = 0$. We claim that the polynomial $\Phi_M(z)$ can be chosen to be identically zero. It suffices to show that $\dim_K(M_d) = 0$ for all d sufficiently large. Again, let M be generated over R by a basis of homogeneous elements $\{y_1, \ldots, y_k\}$. Let d_0 be the maximum degree of these basis elements. Since every element u of M has the form $u = \sum_{i=1}^k \alpha_i y_i$, $\alpha_i \in R = K$, we see that $\deg(u) \leq d_0$. Thus if $d > d_0$ then $M_d = \{0\}$ and $\dim_K(M_d) = 0$, as desired.

Now let $n \geq 1$. Consider the homomorphism $\phi : M \to M$ defined by

$$\phi(u) = uX_n$$

for all $u \in M$. This homomophism is homogeneous of degree 1. Now consider the modules

$$N = \texttt{Ker}\,\phi, \qquad P = \texttt{coKer}\,\phi = M - \phi(M)$$

where $M - \phi(M)$ denotes module difference. Note that X_n is in $\texttt{Ann}\,N$ (since $X_nN = \{0\}$) and X_n is in $\texttt{Ann}\,P$ (since $X_n \cdot (u + \phi(M)) = 0 + \phi(M)$ for any $u \in M$). Thus, by our preceding remarks, N and P can be regarded as modules over $K[X_1, \ldots, X_{n-1}, X_n]/(X_n)$ which is isomorphic to $K[X_1, \ldots, X_{n-1}]$. Thus our induction hypothesis can be applied to N and P.

Now consider the sequence of homomorphisms:

$$0 \longrightarrow N \xrightarrow{i} M \xrightarrow{\phi} M \xrightarrow{j} P \longrightarrow 0$$

where ϕ is as above, i is the inclusion map and j is the natural homomorphism from M to its difference module P. This is an exact sequence. By lemma 1, $N = \bigoplus_d N_d$ is a graded module since it is the kernel of a homogeneous homomorphism ϕ, and $P = \bigoplus_d P_d$ is also a graded module since it is the co-kernel of ϕ. For each d, we may restrict the above sequence to the following:

$$0 \longrightarrow N_d \xrightarrow{i} M_d \xrightarrow{\phi} M_{d+1} \xrightarrow{j} P_{d+1} \longrightarrow 0.$$

By the formula (4) for the alternating sum of module lengths, we get

$$\dim(N_d) - \dim(M_d) + \dim(M_{d+1}) - \dim(P_{d+1}) = 0$$

$$\dim(M_{d+1}) - \dim(M_d) = \dim(P_{d+1}) - \dim(N_d)$$

But P and N are isomorphic to modules over $K[X_1, \ldots, X_{n-1}]$. So by induction on $n-1$, $\dim(P_d)$ (resp. $\dim(N_d)$) is a polynomial in d of degree $\leq n-2$ for all $d \geq d_0$ (for some d_0). [This remark holds even when $n = 1$ as the said polynomial is identically zero, which has degree $-\infty$. The following argument is modified accordingly.] But this means that for $d \geq d_0$,

$$\dim(M_{d+1}) - \dim(M_d) = a_0\binom{d-1}{n-2} + a_1\binom{d-1}{n-3} + \cdots + a_{n-2}$$

where $a_0, \ldots, a_{n-2}$ are integers. For $d = 1, 2, \ldots, d_0 - 1$, we may write

$$\dim(M_{d+1}) - \dim(M_d) = a_0\binom{d-1}{n-2} + a_1\binom{d-1}{n-3} + \cdots + a_{n-2} + c_d$$

where the c_d are integers. Also, we set

$$\dim(M_1) = c_1.$$

Finally, using the binomial identity

$$\binom{d}{s} = \sum_{i=1}^{d-1} \binom{i}{s-1},$$

we sum up the above expressions for $\dim(M_i) - \dim(M_{i-1})$ for $i = 1, \ldots, d$ and $d \geq d_0$, giving

$$\dim(M_d) = a_0\binom{d}{n-1} + a_1\binom{d}{n-2} + \cdots + a_{n-2}\binom{d}{1} + a_{n-1}$$

for some integer a_{n-1}. This proves the Hilbert-Serre theorem.

Exercises

Exercise 2.1: Suppose $I \subseteq R$ is an ideal where $I = Ra_1 + Ra_2 + \cdots + Ra_k$. Let $\overline{a}_i$ denote the image of a_i in I/I^2. Show that $\mathtt{gr}_I(R) = (R/I)[\overline{a}_1, \ldots, \overline{a}_k]$. □

Exercise 2.2: Verify lemma 1. □

§3. Homogeneous Sets

In the following, let $\mathbf{X}' = \{X_0, X_1, \ldots, X_n\} = \{X_0\} \cup \mathbf{X}$. *Some admissible ordering* $\underset{\mathtt{A}}{\leq}$ *on* $\mathtt{PP}(\mathbf{X})$ *will be assumed.*

Homogenization. Homogeneous ideals have nice combinatorial properties (somewhat akin to its well-known nice geometric motivations) which will be exploited in our proof. We first discuss the relation between general ideals and homogeneous ideals (in the standard sense). Given $f \in K[\mathbf{X}]$, there is a well-known procedure to obtain a homogeneous polynomial $F \in K[\mathbf{X}']$: if f has total degree d then F is obtained by multiplying each monomial of degree e in f by X_0^{d-e}. We shall denote the *homogenized form* F of f by $f^\wedge$ and call X_0 the *homogenizing variable*. For example,

$$f = 2X_1^4X_2 - X_2^2 + 3, \qquad f^\wedge = 2X_1^4X_2 - X_0^3X_2^2 + 3X_0^5.$$

There is a corresponding procedure to *dehomogenize* any polynomial $F \in K[\mathbf{X}']$, simply by specializing X_0 to 1. We denote the *dehomogenized form* of F by $F^\vee$. If $S \subseteq K[\mathbf{X}]$ then $S^\wedge = \{f^\wedge : f \in S\}$, and similarly for $S^\vee$. Clearly $(f^\wedge)^\vee = f$. It is also easy to check that

$$(f \cdot g)^\wedge = f^\wedge \cdot g^\wedge, \qquad (f \cdot g)^\vee = f^\vee \cdot g^\vee, \qquad (f + g)^\vee = f^\vee + g^\vee.$$

If $I = \mathtt{Ideal}(f_1, \ldots, f_m)$ then clearly

$$\mathtt{Ideal}({f_1}^\wedge, \ldots, {f_m}^\wedge) \subseteq \mathtt{Ideal}(I^\wedge).$$

The reverse inclusion may be false: let $f_1 = X_1, f_2 = X_1^2 + X_2$. Then $I = \mathtt{Ideal}(f_1, f_2) = \mathtt{Ideal}(X_1, X_2)$ so $\mathtt{Ideal}(I^\wedge) = \mathtt{Ideal}_{K[X_0,X_1,X_2]}(X_1, X_2)$, which properly contains $\mathtt{Ideal}({f_1}^\wedge, {f_2}^\wedge) = \mathtt{Ideal}(X_1, X_1^2 + X_0X_2) = \mathtt{Ideal}(X_1, X_0X_2)$. This situation cannot arise if $\{f_1, \ldots, f_m\}$ is a suitable Gröbner basis.

Let us say that an admissible ordering $\underset{\mathtt{A}}{\leq}$ is *degree-compatible* if it is induced by a weight matrix $\overline{W}$ whose first row is all 1. Thus, the total-degree $\underset{\mathtt{TOT}}{\leq}$ and reverse lexicographic $\underset{\mathtt{REV}}{\leq}$ orderings are degree-compatible, but pure lexicographic $\underset{\mathtt{LEX}}{\leq}$ is not. For any admissible ordering $\underset{\mathtt{A}}{\leq}$ on $\mathtt{PP}(\mathbf{X})$, let

$$\underset{\mathtt{A}}{\leq}^\wedge$$

denote the admissible ordering on $\mathtt{PP}(\mathbf{X}')$ where $u(\underset{\mathtt{A}}{\leq}^{\wedge})v$ iff $\deg(u) < \deg(v)$ or else, $(u^{\vee})\underset{\mathtt{A}}{\leq}(v^{\vee})$. Notice that $\underset{\mathtt{A}}{\leq}^{\wedge}$ is a degree-compatible admissible ordering. In fact, reverse lexicographic ordering can be obtained by recursive application of this operator.

Lemma 3 *Let $\underset{\mathtt{A}}{\leq}$ be an admissible ordering on $\mathtt{PP}(\mathbf{X})$.*

(i) *If $G = \{f_1, \ldots, f_m\}$ is a Gröbner basis for $I \subseteq K[\mathbf{X}]$ relative to $\underset{\mathtt{A}}{\leq}$, and $\underset{\mathtt{A}}{\leq}$ is degree-compatible, then $\mathtt{Ideal}(G^{\wedge}) = \mathtt{Ideal}(I^{\wedge})$. In particular, $G^{\wedge}$ is a Gröbner basis.*

(ii) *If $\{F_1, \ldots, F_m\}$ is a Gröbner basis for $\mathtt{Ideal}(I^{\wedge}) \subseteq K[\mathbf{X}']$ relative to $\underset{\mathtt{A}}{\leq}^{\wedge}$, then $\{F_1, \ldots, F_m\}^{\vee}$ is a Gröbner basis for I relative to $\underset{\mathtt{A}}{\leq}$.*

Proof. (i) The non-trivial inclusion is $\mathtt{Ideal}(I^{\wedge}) \subseteq \mathtt{Ideal}(G^{\wedge})$. For this, it suffices to show that if $F \in I^{\wedge}$ then $F \in \mathtt{Ideal}(G^{\wedge})$. We may assume that $F = f^{\wedge}$ where $f = \sum_i \alpha_i f_i$, $f_i \in G$, $\alpha_i \in K[\mathbf{X}]$ and $\mathtt{hterm}(f) \underset{\mathtt{A}}{\geq} \mathtt{hterm}(\alpha_i f_i)$. Hence $\deg(f) \geq \deg(\alpha_i f_i)$. Then $F = \sum_i \alpha_i{}^{\wedge} f_i{}^{\wedge} X_0^{d-d_i}$ where $d = \deg f$ and $d_i = \deg(\alpha_i f_i)$. This shows $F \in \mathtt{Ideal}(G^{\wedge})$. The fact that $G^{\wedge}$ is a Gröbner basis follows from the standard characterization of Gröbner bases.
(ii) Let $f \in I$. It is sufficient to show that $f = \sum_{i=1}^m \alpha_i \cdot F_i{}^{\vee}$ for some $\alpha_i \in K[\mathbf{X}]$ and where $\mathtt{hterm}(f) \underset{\mathtt{A}}{\geq} \mathtt{hterm}(\alpha_i \cdot F_i{}^{\vee})$. Since the F_i's form a Gröbner basis, we have $f^{\wedge} = \sum_{i=1}^m \beta_i F_i$ where $\beta_i \in K[\mathbf{X}']$ and $\mathtt{hterm}(f^{\wedge})(\underset{\mathtt{A}}{\geq}^{\wedge})\mathtt{hterm}(\beta_i F_i)$. Applying the dehomogenizing operator to this expression for $f^{\wedge}$, and letting $\beta_i{}^{\vee} = \alpha_i$, the result follows. In particular, we check that $\mathtt{hterm}(f^{\wedge})(\underset{\mathtt{A}}{\geq}^{\wedge})\mathtt{hterm}(\beta_i F_i)$ and $\deg(f^{\wedge}) = \deg(\beta_i F_i)$ implies $\mathtt{hterm}(f) \underset{\mathtt{A}}{\geq} \mathtt{hterm}(\alpha_i \cdot F_i{}^{\vee})$. **Q.E.D.**

Homogeneous Sets. A subset $T \subseteq K[\mathbf{X}]$ is a *homogeneous set* if T is a K-vector space and $f \in T$ implies each homogeneous component of f belongs to T. Note that T is a K-vector space means that $au + v \in T$ for all $u, v \in T$, $a \in K$. An important special case of homogeneous sets is where $f \in T$ implies each monomial in f belongs to T; we call these *monomial sets*.

For instance, homogeneous ideals are homogeneous sets and monomial ideals are monomial sets. Gröbner bases theory provides other examples of such sets: for any ideal $I \subseteq K[\mathbf{X}]$ and $f \in K[\mathbf{X}]$, let $\mathtt{nf}_I(f)$ denote the unique normal form of f with respect to the Gröbner basis of I. Let

$$\mathtt{NF}(I) := \{\mathtt{nf}_I(f) : f \in K[\mathbf{X}]\}$$

denote[3] the set of all normal forms.

Lemma 4
(i) $\mathtt{NF}(I)$ is a monomial set.
(ii) If $H = \mathtt{Head}(I)$ then $\mathtt{NF}(H) = \mathtt{NF}(I)$.

Proof. (i) If $f, g \in \mathtt{NF}(I)$, then it is easy to see that $af + g \in \mathtt{NF}(I)$ for any $a \in K$. Thus $\mathtt{NF}(I)$ is a K-vector space. Also $\mathtt{NF}(I)$ is a monomial set because if m is a monomial in $f \in \mathtt{NF}(I)$ then $m \in \mathtt{NF}(I)$.
(ii) Recall $\mathtt{Head}(I) = \mathtt{Ideal}(\{\mathtt{hterm}(f) : f \in I\})$. Clearly f is reducible by G iff f is reducible by

[3] This should not be confused with the notation $\mathtt{NF}_G(f)$ for the set of G-normal forms of f.

$\texttt{hterm}(G)$. But, by the standard characterization of Gröbner bases, G is a Gröbner basis for I iff $\texttt{hterm}(G)$ is a Gröbner basis for H. Hence $f \in \texttt{NF}(I)$ iff $f \in \texttt{NF}(H)$. **Q.E.D.**

Hilbert Function of Homogeneous Sets. For each natural number $z \in \mathbb{N}$, let

$$T_z := \{f \in T : f \text{ is homogeneous of degree } z\}. \tag{6}$$

Note that T_z is also a K-vector space, and so we may speak of its vector space dimension, denoted $\dim_K(T_z)$. The *Hilbert function* $\phi_T : \mathbb{N} \to \mathbb{N}$ of T is defined by

$$\phi_T(z) := \dim_K(T_z).$$

As before, if the Hilbert function is equal to a polynomial $f(z)$ for z sufficiently large, we shall call $f(z)$ the *Hilbert polynomial* of T and denote it by $\Phi_T(z)$.

Example: The Hilbert function should be viewed combinatorially as a counting function. If T is a monomial set, then $\phi_T(z)$ just counts the number of power products of total degree z in T_z. In figure 1, we illustrate this for $K[X, Y]$: each lattice point corresponds to a power product in $\texttt{PP}(X, Y)$, where moving to the north-east (resp., north-west) corresponds to increasing powers of Y (resp., X).

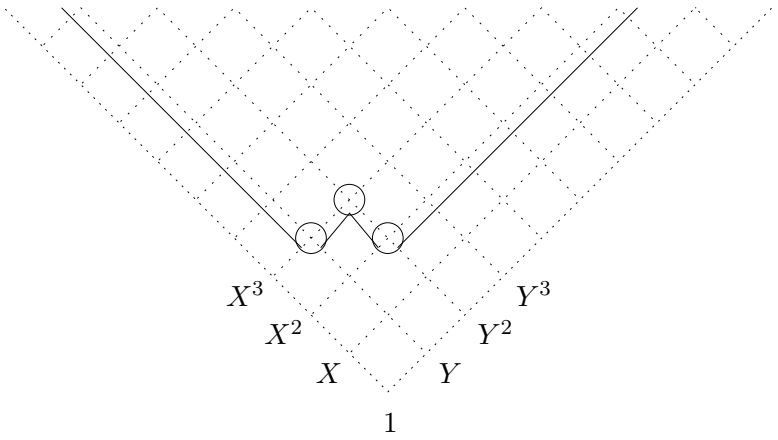


Figure 1: Hilbert function of $\texttt{Ideal}(X^3Y, X^2Y^2)$.

The origin corresponds to $1 = X^0Y^0$. The three circled lattice points are X^3Y, X^2Y^2 and X^3Y^2. The ideal $I = \texttt{Ideal}(X^3Y, X^2Y^2)$ is represented in figure 1 by the set of lattice points lying above the solid curve. The Hilbert function $\phi_I(z)$ counts the number of monomials in I that are of total degree z; these are the monomials at level z in the figure. For instance, $\phi_I(z) = 0$ for $z \leq 3$, $\phi_I(4) = 2$. In general, $\phi_I(z) = z - 2$ for $z \geq 4$. ■

If I is a homogeneous ideal, the Hilbert-Serre theorem shows that Hilbert function $\phi_I(z)$ is equal to the Hilbert polynomial $\Phi_I(z)$ for z larger than the Hilbert constant h_I. Note that with $T = K[\mathbf{X}]$ we have

$$\phi_T(z) = \binom{z+n-1}{n-1},$$

a polynomial of degree $n-1$ that counts the number of terms of total degree z. The set $\mathtt{NF}(I)$, being monomial, has a Hilbert function $\phi_{\mathtt{NF}(I)}(z)$. It is easy to see that

$$\phi_I(z) + \phi_{\mathtt{NF}(I)}(z) = \binom{z+n-1}{n-1}. \tag{7}$$

So we conclude from the Hilbert-Serre theorem:

Corollary 5 *For any ideal I, the Hilbert function $\phi_{\mathtt{NF}(I)}(z)$ of $\mathtt{NF}(I)$ is a polynomial of degree $\leq n-1$ for z large enough.*

Again, let $\Phi_{\mathtt{NF}(I)}(z)$ denote the polynomial in this corollary. Note that for $z \geq h_I$ (the Hilbert constant), we also have

$$\phi_{\mathtt{NF}(I)}(z) = \Phi_{\mathtt{NF}(I)}(z).$$

Our immediate goal is to find an upper bound on h_I. It turns out that h_I yields an upper bound on the function $G(n,d)$ (§1). Monomial sets have a key role in this development:

Lemma 6 *Let I be a homogeneous ideal and $H = \mathtt{Head}(I)$. Then I and H have the same Hilbert functions:*

$$\phi_I = \phi_H$$

Proof. We first show $\phi_I(z) \leq \phi_H(z)$. Let $f_1, \ldots, f_m \in I$ be homogeneous of degree z and suppose they are linearly independent over K. We must show that there exists $h_1, \ldots, h_m \in H$ of degree z and linearly independent over K. Suppose $f_1' = f_1$ is the polynomial in $\{f_1, \ldots, f_m\}$ with the $\underset{\mathtt{A}}{\leq}$-largest head term. Now use f_1' to reduce each polynomial in $\{f_2, \ldots, f_m\}$. This just ensures that the resulting polynomials have head terms distinct from $h_1 := \mathtt{hterm}(f_1')$. Also, note that none of reduced polynomials are zero – otherwise, we have found a linear dependence in $\{f_1, \ldots, f_m\}$. Let f_2' be the polynomial with the $\underset{\mathtt{A}}{\leq}$-largest head term in the reduced set of polynomials and let $h_2 := \mathtt{hterm}(f_2')$. Now repeat this process to find $h_3, \ldots, h_m$. But $h_1, \ldots, h_m$, being pairwise distinct, are linearly independent. Since they are of degree d and belong to H, this proves $\phi_I(z) \leq \phi_H(z)$.

To show $\phi_I(z) \geq \phi_H(z)$, suppose $f_1, \ldots, f_m \in I$ such that $\mathtt{hterm}(f_1), \ldots, \mathtt{hterm}(f_m)$ are distinct and of degree z. Then clearly the above process yields $f_1', \ldots, f_m' \in I$ that are linearly independent over K. **Q.E.D.**

From (7), we obtain $\phi_{\mathtt{NF}(I)} = \phi_{\mathtt{NF}(H)}$, since $\mathtt{NF}(I) = \mathtt{NF}(H)$. We are mainly interested in Hilbert functions ϕ_T where T has the form I or $\mathtt{NF}(I)$; by the last two lemmas, we may assume that T are monomial sets in these cases.

Homogeneous Decompositions. We want to decompose a homogeneous set T as a direct sum

$$T = S_1 \oplus S_2 \oplus \cdots \oplus S_i \oplus \cdots \tag{8}$$

where the $S_i \subseteq T$ are K-vector spaces. In case the S_i's are homogeneous sets, we call the set $\{S_1, S_2, S_3, \ldots\}$ a *homogeneous decomposition* of T. An example of homogeneous decomposition is $T = T_0 \oplus T_1 \oplus T_2 \oplus \cdots$ where T_z is defined in (6). If equation (8) is a homogeneous decomposition, then

$$\phi_T(z) = \sum_{i \geq 1} \phi_{S_i}(z). \tag{9}$$

This is because each $u \in T_z$ has a unique decomposition $u = \sum_{i \geq 1} u_i$ where $u_i \in (S_i)_z = S_i \cap T_z$. Thus computing the Hilbert function of T is reduced to computing those of S_i.

Lemma 7 (Dubé)

(i) *For any ideal $I \subseteq K[\mathbf{X}]$, we obtain the following direct decomposition:*

$$K[\mathbf{X}] = I \oplus \mathtt{NF}(I).$$

(ii) *Let J be a proper subideal of $I \in K[\mathbf{X}]$ and $I = \mathtt{Ideal}(f_0, J)$ for some $f_0 \in K[\mathbf{X}]$.*

$$I = J \oplus f_0 \cdot \mathtt{NF}(J : f_0).$$

Proof.
(i) For $f \in K[\mathbf{X}]$, $f = \mathtt{nf}_I(f) + g$ where $g \in I$. But $\mathtt{nf}_I(f)$ is unique.
(ii) First we show that $J + f_0 \cdot \mathtt{NF}(J : f_0) = I$. Since $J : f_0$ is the ideal $\{g \in R : g \cdot f_0 \in J\}$, the containment $J + f_0 \cdot \mathtt{NF}(J : f_0) \subseteq I$ is immediate. To show the reverse inclusion, note that $I = \mathtt{Ideal}(f_0, J)$ implies every $f \in I$ has the form

$$f = \alpha f_0 + \beta$$

where $\alpha \in R, \beta \in J$. But $\alpha = \mathtt{nf}_{J:f_0}(\alpha) + \gamma$ where $\gamma \in J : f_0$. Hence

$$f = f_0 \cdot \mathtt{nf}_{J:f_0}(\alpha) + (\beta + \gamma f_0).$$

This shows $f \in f_0 \cdot \mathtt{NF}(J : f_0) + J$. Finally, we must show that $f_0 \cdot \mathtt{NF}(J : f_0) + J$ is indeed a direct sum. Suppose $f = \alpha f_0 + \beta = \alpha' f_0 + \beta'$ where $\alpha, \alpha' \in \mathtt{NF}(J : f_0)$ and $\beta, \beta' \in J$. Then $(\alpha - \alpha')f_0 = \beta' - \beta \in J$. This implies $\alpha - \alpha' \in J : f_0$. But $\alpha - \alpha' \in \mathtt{NF}(J : f_0)$. The last two assertions imply $\alpha - \alpha' = 0$. **Q.E.D.**

Note that if I, J and f_0 are homogeneous then the decompositions in the above lemma are homogeneous. Moreover, we can repeatedly apply the decomposition in part (ii), so that if $I = \mathtt{Ideal}(f_1, \ldots, f_m)$ and $I_i = \mathtt{Ideal}(f_1, \ldots, f_{i-1})$ $(i = 1, \ldots, m)$ then

$$I = \mathtt{Ideal}(f_1) \oplus \left(\bigoplus_{i=1}^{m} f_i \cdot \mathtt{NF}(I_i : f_i) \right). \tag{10}$$

Exercises

Exercise 3.1: Suppose that I, J are homogeneous ideals of a graded ring R.
(i) Show that $I + J$, $I \cap J$, IJ, $I : J$ and $\sqrt{I}$ are each homogeneous.
(ii) Show that I is prime if and only if for all homogeneous elements $u, v \in R$, $uv \in I$ implies $u \in I$ or $v \in I$.
(iii) Show that I is primary if and only if for all homogeneous elements $u, v \in R$, $uv \in I$ implies $u \in \sqrt{I}$ or $v \in \sqrt{I}$. □

Exercise 3.2: What is the Hilbert function of $\mathtt{Ideal}(f)$ where $\deg(f) = d$ and $n = 3$? of $\mathtt{Ideal}(X^2 + Y^2 + Z^2, XY, Z)$? □

Exercise 3.3: Show that $\phi_T(z)$ is the number of distinct power products of total degree z that appear as head terms of polynomials in T. (Thus $\phi_T(z)$ is independent of the admissible order $\underset{\mathtt{A}}{\leq}$.) □

Exercise 3.4: Carry out the preceding development for a general K-vector space $T \subseteq K[X_1, \ldots, X_n]$, not necessarily a homogeneous set. HINT: for the affine version of the Hilbert function, $\phi_T(z) = \dim_K(T_z)$ let T_z be the set of polynomials of total degree at most z. □

Exercise 3.5: Given $\{f_1, \ldots, f_m\}$, give an algorithm to test if $\mathtt{Ideal}(f_1, \ldots, f_m)$ is homogeneous. HINT: reduce the problem to computing a Gröbner basis. □

§4. Cone Decomposition

Let $f \in K[\mathbf{X}]$ be homogeneous and $\mathbf{Y} \subseteq \mathbf{X}$ be a set of variables. A *cone* C is a set of the form

$$\mathtt{Cone}(f, \mathbf{Y}) := f \cdot K[\mathbf{Y}].$$

For instance, $\mathtt{Cone}(1, \mathbf{X}) = K[\mathbf{X}]$. A principal ideal is a cone: $\mathtt{Ideal}(f) = \mathtt{Cone}(f, \mathbf{X})$. Define

$$\deg(C) := \deg(f), \qquad \dim(C) := |\mathbf{Y}|.$$

If $f = 0$ then $\mathtt{Cone}(f, \mathbf{Y}) = \{0\}$ is the *trivial cone*; by definition, $\dim(C) = \deg(C) = -\infty$ for the trivial cone. If $T = C_1 \oplus \cdots \oplus C_r$ where C_i are cones, then we call $D = \{C_1, \ldots, C_r\}$ a *cone decomposition* of T. Note that the C_i's are non-trivial unless $T = \{0\}$, in which case we have the *trivial decomposition* $D = \{\mathtt{Cone}(0, \emptyset)\}$.

Figure 2 illustrates a cone decomposition for $n = 2$. The set

$$\mathtt{Cone}(X^2Y^2, \emptyset), \quad \mathtt{Cone}(X^3Y, \{X\}), \quad \mathtt{Cone}(X^3Y^2, \{X\}), \quad \mathtt{Cone}(X^2, Y^3, \{X, Y\})$$

is a cone decomposition for the ideal $\mathtt{Ideal}(X^3Y, X^2Y^2)$.

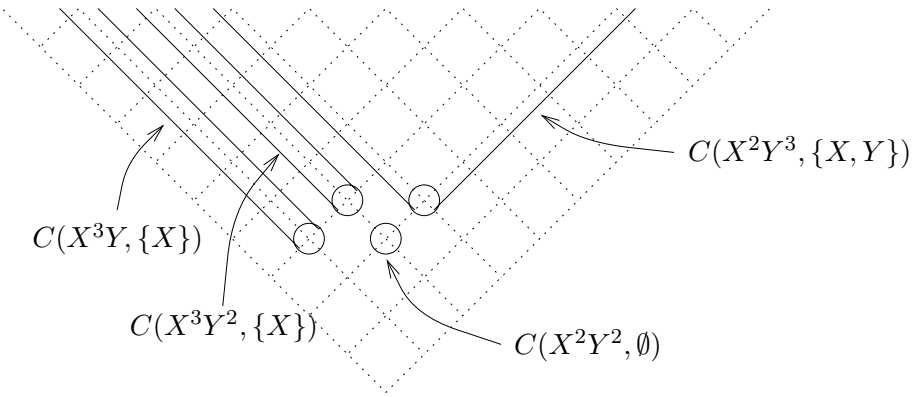


Figure 2: Cone decomposition for $\mathtt{Ideal}(X^3Y, X^2Y^2)$.

Cones are homogeneous sets with very simple Hilbert functions:

$$\dim(C) = 0: \quad \phi_C(z) \quad = \quad \begin{cases} 1 & \text{if } \ z = \deg(C), \\ 0 & \text{else.} \end{cases}$$

$$\dim(C) > 0: \quad \phi_C(z) \quad = \quad \begin{cases} 0 & \text{if } \ z < \deg(C), \\ \binom{z-\deg(C)+\dim(C)-1}{\dim(C)-1} & \text{else.} \end{cases}$$

Definition 1 *We call a cone decomposition*

$$D = \{C_1, \ldots, C_r, C_{r+1}, \ldots, C_{r+s}\}, \quad (r, s \geq 0) \tag{11}$$

exact *if, after relabelling, we have*

$$\dim(C_1) \geq \dim(C_2) \geq \cdots \geq \dim(C_r) > 0 = \dim(C_{r+1}) = \cdots = \dim(C_{r+s})$$

and

$$\deg(C_1) < \deg(C_2) < \cdots < \deg(C_r) = \deg(C_1) + r - 1.$$

The positive part *of D be defined as*

$$D^+ := \{C \in D : \dim(C) > 0\}.$$

Thus $\deg(C_{i+1}) = \deg(C_i) + 1$ for $i = 1, \ldots, r-1$. We provide some intuition for this definition using a graphic representation. In figure 3, we represent each cone as a lattice point in a deg-dim axes system.

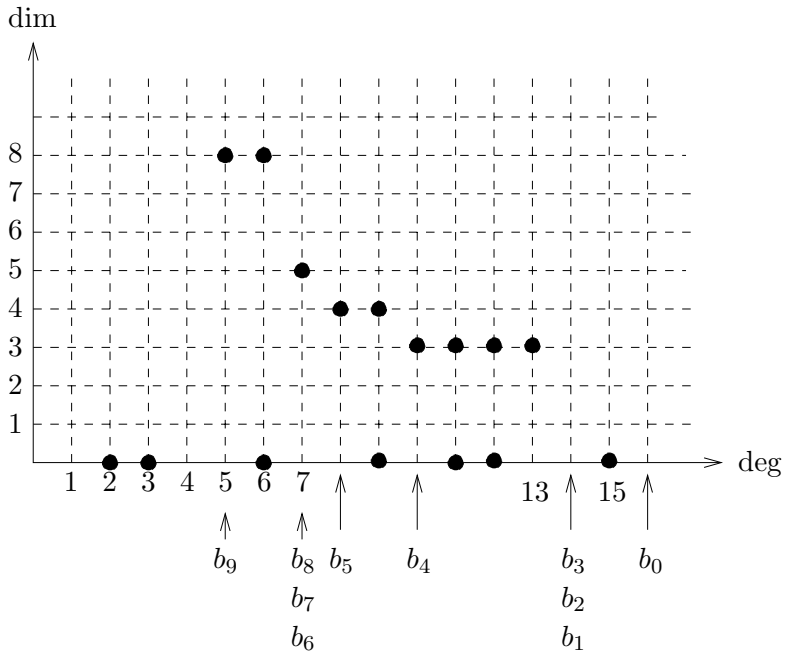


Figure 3: Illustrating exact decomposition

A cone decomposition D is then represented by a multiset of lattice points, one lattice point for each cone in D. If D is exact, then only lattice points of dimension 0 can have multiplicity greater than 1. For any set D of cones, let

$$\dim(D) := \max\{\dim(C) : C \in D\}$$

and

$$\texttt{maxdeg}(D) := \max\{\deg(C) : C \in D\}, \qquad \texttt{mindeg}(D) := \min\{\deg(C) : C \in D\}.$$

If $D = \emptyset$, we set $\texttt{maxdeg}(D) = \texttt{mindeg}(D) = 0$. For any cone decomposition D, we are mostly interested in the quantities $\texttt{maxdeg}(D)$ and $\texttt{mindeg}(D^+)$. Notice that if D and D' are cone decompositions of a set T then $\dim(D) = \dim(D')$. Hence, we may speak of the *cone dimension* $\dim(T)$ of T.

Let D_0 be the cone decomposition illustrated in figure 3. Then $\dim(D_0) = 8$, $\texttt{maxdeg}(D_0) = 15$ and $\texttt{mindeg}(D_0^+) = 5$. Assuming no multiplicity of lattice points in this figure, we may verify that D_0 is exact: this amounts to saying that the lattice points of D_0^+ represent *a discrete non-increasing function in the range from* $\texttt{mindeg}(D_0^+)$ *to* $\texttt{maxdeg}(D_0^+)$.

Macaulay Constants. The *Macaulay constants* of the exact cone decomposition D in equation (11) is a set $b_0, \ldots, b_{n+1}$ of constants defined as follows: For $i = 0, \ldots, n$, define b_i to be

$$b_i := \begin{cases} 1 + \max\{\deg(C) : C \in D, \dim(C) \geq i\}, & \text{if } \dim(D) \geq i \\ \texttt{mindeg}(D^+), & \dim(D) < i. \end{cases}$$

In particular, we always have the relation $b_{n+1} = \texttt{mindeg}(D^+)$. Hence,

$$1 + \texttt{maxdeg}(D) = b_0 \geq b_1 \geq \cdots \geq b_n \geq b_{n+1} = \texttt{mindeg}(D^+).$$

In the example D_0 of figure 3, we have

$$b_0 = 16, \quad b_1 = b_2 = b_3 = 14, \quad b_4 = 10, \quad b_5 = 8, \quad b_6 = b_7 = b_8 = 7, \quad b_9 = 5.$$

The following derivation motivates the definition of the Macaulay constants. First, there are exactly

$$b_i - b_{i+1}$$

cones of dimension i ($i = 1, 2, \ldots, n$). In fact, these cones have degrees

$$b_{i+1}, b_{i+1} + 1, \ldots, b_i - 1.$$

So for $z \geq b_0$, by (9),

$$\begin{aligned}
\phi_T(z) &= \sum_{C \in D} \phi_C(z) \\
&= \sum_{i=1}^{n} \sum_{d=b_{i+1}}^{b_i - 1} \binom{z - d + i - 1}{i - 1} \\
&= \sum_{i=1}^{n} \left[\binom{z - b_{i+1} + i}{i} - \binom{z - b_i + i}{i} \right] \\
&= \binom{z - b_{n+1} + n}{n} + \sum_{i=1}^{n-1} \left[\binom{z - b_{i+1} + i}{i} - \binom{z - b_{i+1} + i + 1}{i + 1} \right] - \binom{z - b_1 + 1}{1} \\
&= \binom{z - b_{n+1} + n}{n} - \sum_{i=1}^{n-1} \binom{z - b_{i+1} + i}{i + 1} - \binom{z - b_1}{1} - 1 \\
&= \binom{z - b_{n+1} + n}{n} - \sum_{i=1}^{n} \binom{z - b_i + i - 1}{i} - 1. \qquad (12)
\end{aligned}$$

This form of the Hilbert polynomial is due to Macaulay, and we will refer to it as the "Macaulay form" of the Hilbert polynomial of T. This polynomial is of at most degree $n-1$ since the leading terms of $\binom{z-b_{n+1}+n}{n}$ and $\binom{z-b_n+n-1}{n}$ cancel. Also notice that b_0 does not appear in the equation. Since these constants depend on the exact decomposition which is by no means unique, the following result is interesting.

Lemma 8 *The Macaulay constants $b_0, b_1, \ldots, b_n$ for a homogeneous set T are uniquely determined once b_{n+1} is fixed.*

Proof. As a homogeneous set, T has a Hilbert polynomial. Write this in the standard form

$$\Phi_T(z) = \sum_{i=0}^{n-1} a_i z^i. \tag{13}$$

The constants $a_0, \ldots, a_{n-1}$ are unique in this form. Note that only b_{n+1} and b_n affect the term of degree $n-1$ in equation (12). Equating the coefficients of z^{n-1} in the two forms, we express a_{n-1} in terms of b_{n+1} and b_n. Thus b_n is completely determined by b_{n+1} and a_{n-1}. In general, a_{i-1} can be expressed in terms of $b_i, \ldots, b_{n+1}$ for $i = n, n-1, \ldots, 2, 1$. Hence b_i is determined from $a_{i-1}, b_{i+1}, \ldots, b_{n+1}$. By induction, $b_{i+1}, \ldots, b_{n+1}$ have been determined, so b_i is determined. So the Macaulay form of the polynomial is completely determined. What about the constant b_0? We claim that b_0 is the smallest value $z_0 \geq b_1$ such that for all $z \geq z_0$, the expression 12 equals the Hilbert function $\phi_T(z)$. To see this, note that for $z \geq b_1$, the Macaulay formula accurately counts the contribution of cones of positive dimension. So any error is due to cones of dimension zero; by definition of b_0, it is correct for $z \geq b_0$. On the other hand, if $b_0 > b_1$ then the Macaulay formula is wrong at $z = b_0 - 1$. **Q.E.D.**

We may also define the *dimension* $\dim(T)$ of T to be $\dim(D)$ for any cone decomposition D of T. It is clear that this definition does not depend on the choice of D.

Exercises

Exercise 4.1: If $T \subseteq K[\mathbf{X}]$ is a K-vector space, not necessarily homogeneous, define the affine version of the Hilbert function $\varphi_T(z)$ to be the dimension of the set $T_{\leq z} := \{u \in T : \deg(u) \leq z\}$. Carry out the analogous development. □

§5. Exact Decomposition of NF(I)

This section gives an exact decomposition of NF(I) where I is any ideal. We may assume that I is a monomial ideal. Note that if u is a monomial then $I : u$ is a monomial ideal. We introduce the convenient if somewhat unusual notation

$$a : b = \frac{a}{\mathtt{GCD}(a, b)}$$

where $a, b \in K[\mathbf{X}]$. Note that $a : b \in K$ if and only if $a | b$, and $a : b, b : a$ are relatively prime. This notation is justified on the grounds that $\mathtt{Ideal}(a : b) = \mathtt{Ideal}(a) : b$, where the right-hand side refers to ideal quotient.

Lemma 9 *Let u be a monomial. Then $I : u = \texttt{Ideal}(u_1 : u, u_2 : u, \ldots, u_m : u)$, where I is generated by the monomials $u_1, \ldots, u_m$.*

Proof. Let $J = \texttt{Ideal}(u_1 : u, u_2 : u, \ldots, u_m : u)$. To show $I : u \subseteq J$, let $f \in I : u$ where $f = \sum_i f_i$ and each f_i is a monomial. Then $f \cdot u = \sum_i f_i u \in I$. Since I is a monomial ideal, each $f_i u$ is in I. So $f_i u$ is divisible by some u_j. This means f_i is divisible by $u_j : u$. We conclude that $f \in J$. Conversely, each $u_i : u \in I : u$, so $(u_i : u) \cdot u \in \texttt{Ideal}(u_i) \subseteq I$. This shows $u_i : u \in I : u$. So $J \subseteq I : u$. **Q.E.D.**

The key idea of the decomposition is simple: let $X_0 \in \mathbf{Y}$ and $\mathbf{Y}' = \mathbf{Y} \setminus \{X_0\}$. Then we have the cone decomposition

$$\begin{aligned} \texttt{Cone}(u, \mathbf{Y}) &= \texttt{Cone}(uX_0, \mathbf{Y}) \oplus \texttt{Cone}(u, \mathbf{Y}') \\ &= C_1 \oplus C_2 \end{aligned} \tag{14}$$

where C_1, C_2 are implicitly defined here. We may call this the *shift decomposition* of $\texttt{Cone}(u, \mathbf{Y})$. For example, let $u = X^2Y$, $\mathbf{Y} = \{X, Z\}$ and $X_0 = X$. Then $C_1 = \texttt{Cone}(X^3Y, \{X, Z\})$ and $C_2 = \texttt{Cone}(X^2Y, \{Z\})$. Note that $\deg(C_1)$ is 1 more than $\deg(C)$ and $\dim(C_2)$ is 1 less than $\dim(C)$. This is illustrated in figure 4:

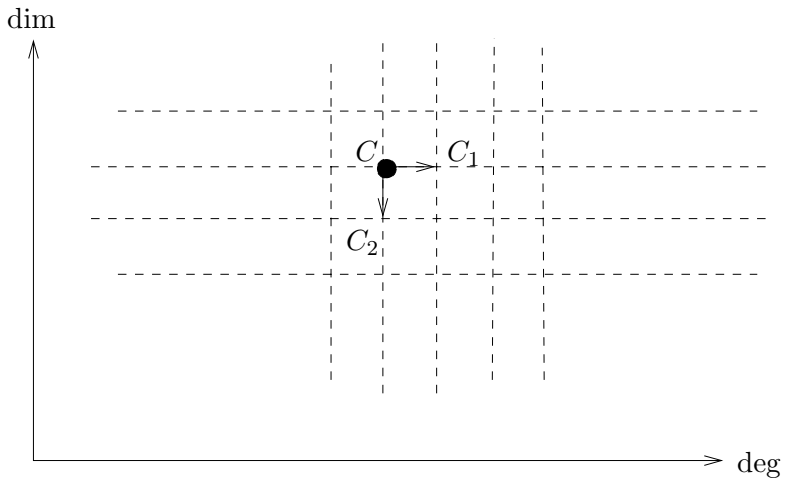


Figure 4: Shifting a cone.

We repeat this operation and obtain for any cone C the cone decomposition

$$C = \bigoplus_{i=0}^{d} C'_i \tag{15}$$

where $d = \dim C$, $\dim C'_i = i$ $(i = 0, \ldots, d)$ and $\deg(C'_i) = \deg(C) + 1$ $(i = 1, \ldots, d)$. Let us call this the "complete shift decomposition". We now describe an application of this. We will say that C is "violating" in a set D of cones if $\dim(C) > 0$ and for some other $C' \in D$, one of the following three conditions holds:

(I) $\deg C' = \deg C$ and $\dim C' \geq \dim C$.

(II) $\deg C' > \deg C$ and $\dim C' > \dim C$.

(III) $\deg C' > 1 + \deg C$ but there does not exist $C'' \in D^+$ with $\deg C'' = 1 + \deg C$.

We refer to these as *violations of types* (I), (II) and (III), respectively. For any cone decomposition D, it is easy to see that D is exact iff D contains no violating cones. The following is a procedure to remove violating cones:

Exact Conversion
Input: D, a cone decomposition of a set T.
Output: D, an exact cone decomposition of T.
Method:
  while there exists a "violating" cone $C \in D$ do
    Replace C by its "complete shift decomposition", $C_0, C_1, \ldots, C_d$ as in equation (15).

If this procedure halts, it is clear that the result is correct.

Lemma 10 *The exact conversion procedure halts.*

Proof. We use induction on $|D^+|$. If $|D^+| \leq 1$, the result is clear. So let $|D^+| \geq 2$. Let $d_1 := \mathtt{mindeg}\{C \in D^+ : \dim C = \dim D\}$ and $D_1 := \{C \in D^+ : \deg C \leq d_1\}$. If D_1 has no violating cones then $|D_1| = 1$ and D is exact iff $D \setminus D_1$ is exact. By induction, the exact conversion applied to $D \setminus D_1$ halts. The lemma holds in this case. Suppose D_1 contains a violating cone. At each step, the procedure replaces a cone C by its complete shift decomposition. If $C \in D_1$, then this reduces the size of D_1 and the procedure halts, by double induction on $|D_1|$ and $|D^+|$. If $C \notin D_1$, then the procedure can only choose C from $D \setminus D_1$. Again by induction, it cannot do this infinitely often. This completes the proof. **Q.E.D.**

Now we describe a method to exactly decompose a set of the form

$$\mathtt{NF}(I) \cap \mathtt{Cone}(u, \mathbf{Y})$$

where $\mathbf{Y} \subseteq \mathbf{X} = \{X_1, \ldots, X_n\}$ and $u \in \mathtt{PP}(\mathbf{X})$. This is a slight generalization of the original goal of exactly decomposing $\mathtt{NF}(I)$.

Exact Decomposition Algorithm
Input: $(u, \mathbf{Y}, F)$ where $u \in \mathtt{PP}(\mathbf{X})$,
  $\mathbf{Y} \subseteq \mathbf{X}$ a set of variables and $F \subseteq \mathtt{PP}(\mathbf{X})$ is a monomial basis for an ideal I.
Output: D, an exact cone decomposition for $\mathtt{NF}(I) \cap \mathtt{Cone}(u, \mathbf{Y})$.
Method:
1. (BASIS 1) For each $v \in F$ do: if $v : u \in \mathtt{PP}(\mathbf{Y})$ then goto step 2.
   Return $D = \{\mathtt{Cone}(u, \mathbf{Y})\}$.
2. (BASIS 2) If $v : u = 1$ for any $v \in F$
   then return the trivial cone decomposition $D = \{\mathtt{Cone}(0, \emptyset)\}$.
3. (DIVIDE) Let $X_0 \in \mathbf{Y}$ such that $X_0 | v : u$ for some $v \in F$.
   Let $\mathbf{Y}' = \mathbf{Y} \setminus \{X_0\}$ and $\mathtt{Cone}(u, \mathbf{Y}) = C_1 \oplus C_2$
     with $C_1 = \mathtt{Cone}(uX_0, \mathbf{Y})$ and $C_2 = \mathtt{Cone}(u, \mathbf{Y}')$, as in equation (14).
4. (RECURSE) Call the exact decomposition algorithm recursively (twice) to decompose
   $\mathtt{NF}(I) \cap C_1$ and also $\mathtt{NF}(I) \cap C_2$.
5. (CONQUER) Let D_i $(i = 1, 2)$ be the exact decomposition for $\mathtt{NF}(I) \cap C_i$.
   Return the "exact conversion" of $D_1 \cup D_2$ using the procedure above.

Correctness. We justify each step in the algorithm.

Step 1: (BASIS 1) Suppose we returned in this step. This means for all $v \in F$, $v : u \notin \mathtt{PP}(\mathbf{Y})$. We need to show that $D = \{\mathtt{Cone}(u, \mathbf{Y})\}$ is an the exact decomposition of $\mathtt{NF}(I) \cap \mathtt{Cone}(u, \mathbf{Y})$. This is equivalent to $\mathtt{Cone}(u, \mathbf{Y}) \subseteq \mathtt{NF}(I)$. The following characterization justifies this step.

Lemma 11 $\mathtt{Cone}(u, \mathbf{Y}) \subseteq \mathtt{NF}(I)$iff (for all $f \in F) f : u \notin \mathtt{PP}(\mathbf{Y})$.

Proof. ($\Rightarrow$) If $f : u \in \mathtt{PP}(\mathbf{Y})$ then there exists $\alpha \in \mathtt{PP}(\mathbf{Y})$ such that $f : u = \alpha$, *i.e.*, $f | u\alpha$, or equivalently, $u\alpha \notin \mathtt{NF}(I)$. This contradicts our assumption that $\mathtt{Cone}(u, \mathbf{Y}) \subseteq \mathtt{NF}(I)$. ($\Leftarrow$) If $\alpha u \in \mathtt{Cone}(u, \mathbf{Y})$ then for all $f \in F$, f does not divide αu since $f : u \notin \mathtt{PP}(\mathbf{Y})$. Hence $\alpha u \in \mathtt{NF}(I)$. **Q.E.D.**

Step 2: (BASIS 2) Suppose we returned this step. Then clearly $v | u$ and $\mathtt{Cone}(u, \mathbf{Y}) \subseteq I$. Hence $\mathtt{Cone}(u, \mathbf{Y}) \cap \mathtt{NF}(I) = \{0\}$. This justifies the output of the trivial decomposition.

Step 3: (DIVIDE) The choice $X_0 \in \mathbf{Y}$ exists because we did not return in steps 1 and 2. For now, any choice of X_0 will do. We describe later a more careful choice of X_0 in order to ensure a certain property of the decomposition.

Step 4: (RECURSE) The recursive calls to decompose $\mathtt{NF}(I) \cap C_i$ $(i = 1, 2)$ must be shown to eventually return. To show this, we must indicate the sense in which the new arguments are "smaller" than the original inputs. Towards this end, we define the "size" of the input $(u, \mathbf{Y}, F)$ to be

$$|\mathbf{Y}| + \sum_{v \in F} \deg(v : u).$$

Note that we use $\deg(v : u)$ instead of $\deg(v)$. In the case of C_1, our arguments are $(uX_0, \mathbf{Y}, F)$. Note that

$$\deg(v : uX_0) \leq \deg(v : u)$$

is always true; moreover, our choice of X_0 ensures that the inequality is strict for some $v \in F$. Hence the size has gone down. In the case of C_0, the arguments are $(u, \mathbf{Y}', F)$ clearly has reduced size by virtue of $|\mathbf{Y}'| < |\mathbf{Y}|$.

Step 5: (CONQUER) Notice that $D_1 \cup D_2$ is a cone decomposition of $\mathtt{Cone}(u, \mathbf{Y}) \cap \mathtt{NF}(I)$. Then applying the "exact conversion" procedure above to $D_1 \cup D_2$ gives us the desired output. This concludes our justification of the algorithm.

Modified Exact Decomposition Algorithm. We wish to modify step 3 in the above algorithm. But we first develop some insights. Let D be the exact decomposition for $\mathtt{NF}(I)$ as returned by the above algorithm on input $(u, \mathbf{Y}, F)$ where F is a monomial basis for I. Let

$$T = \mathtt{Cone}(u, \mathbf{Y}) \cap \mathtt{NF}(I).$$

Note that for any $w \in \mathtt{Cone}(u, \mathbf{Y})$ and $\mathbf{Y}' \subseteq \mathbf{Y}$,

$$\mathtt{Cone}(w, \mathbf{Y}') \subseteq T \text{iff } \mathtt{Cone}(u, \mathbf{Y}') \subseteq T.$$

Of course, $\dim(T)$ is equal to the largest subset $\mathbf{Y}' \subseteq \mathbf{Y}$ such that $\mathtt{Cone}(u', \mathbf{Y}') \subseteq \mathtt{NF}(I)$. The preceding remark implies that we can in fact take $u = u'$. For lack of a better name, let us call any subset $\mathbf{Y}' \subseteq \mathbf{Y}$ of maximal cardinality such that $\mathtt{Cone}(u, \mathbf{Y}') \subseteq T$ a "maximum set" relative to u, I. This is equivalent to saying $\mathbf{Y}'$ has maximum cardinality subject to

$$\mathtt{Cone}(u, \mathbf{Y}') \subseteq \mathtt{NF}(I). \tag{16}$$

Example: Let $\mathbf{Y} = \mathbf{X} = \{W, X, Y, Z\}$, $u = Y^2$ and $F = \{W^2XZ, X^2Y^3, Y^4Z, WY^3\}$. Let F' comprise the elements $f : u$ for all $f \in F$. So $F' = \{W^2XZ, X^2Y, Y^2Z, WY\}$. Using the characterization in lemma 11, we see that the maximal (not maximum) sets that satisfy (16) are

$$\{W, Z\}, \quad \{W, X\}, \quad \{X, Z\}, \quad \{Y\}.$$

The first three sets are maximum. ■

We want our algorithm on input $(u, \mathbf{Y}, F)$ to return a decomposition D that contains a cone of the form $\texttt{Cone}(u, \mathbf{Y}^*)$ where $\mathbf{Y}^*$ is a maximum set in the sense of (16). Note that D will always contain a cone of the form $\texttt{Cone}(u', \mathbf{Y}^*)$ — so what we seek is $u = u'$. To this end, we modify step 3 of the algorithm as follows:

> 3. (DIVIDE) Choose X_0 such that $\mathbf{Y}' := \mathbf{Y} \setminus \{X_0\}$ contains a maximum set in the sense of (16). Let $C_1 = \texttt{Cone}(uX_0, \mathbf{Y})$ and $C_2 = \texttt{Cone}(u, \mathbf{Y}')$.

This choice of X_0 exists since $\mathbf{Y}$ is not a maximum set by the time we reach step 3 in the algorithm. Furthermore, X_0 still has the property that X_0 divides $f : u$ for some $f \in F$ (this property ensures termination). To see this, let $\mathbf{Y}^* \subseteq \mathbf{Y} \setminus \{X_0\}$ be a maximum set. By the characterization in lemma 11, $f : u \notin \texttt{PP}(\mathbf{Y}^*)$ for all $f \in F$. If X_0 does not divide $f : u$ for all $f \in F$ then $\{X_0\} \cup \mathbf{Y}^*$ has the C-property, contradicting the assumption that $\mathbf{Y}^*$ is maximum.

Lemma 12 *Consider the modified algorithm on input $(u, \mathbf{Y}, F)$. The output cone decomposition D, if non-trivial, contains a cone of the form $\texttt{Cone}(u, \mathbf{Y}^*)$ where $\mathbf{Y}^*$ is maximum and $\texttt{mindeg}(D) = \deg u$.*

Proof. The lemma is clearly true in case of "basis 1", and by assumption, "basis 2" does not occur. So assume D is the "exact conversion" of $D_1 \cup D_2$. Notice that D_2 is nontrivial since some maximum set $\mathbf{Y}^*$ is contained in $\mathbf{Y}' = \mathbf{Y} \setminus \{X_0\}$. So by induction, D_2 contains some cone $C^* = \texttt{Cone}(u, \mathbf{Y}^*)$. Such a cone is clearly unique (since u is given). If D_1 is trivial, then $D = D_2$ and the lemma is true. Otherwise, $\texttt{mindeg}(D_2) = \deg u$ and $\texttt{mindeg}(D_1) = 1 + \deg u$. It follows that C^* is non-violating and our exact conversion procedure will not replace C^*. Hence $C^* \in D$. **Q.E.D.**

Applying the modified algorithm to input $(1, \mathbf{X}, F)$, we conclude that for every monomial ideal $I = \texttt{Ideal}(F) \neq K[\mathbf{X}]$, the set $\texttt{NF}(I)$ has an exact cone decomposition D for which

$$\texttt{mindeg}(D^+) = 0. \tag{17}$$

(Recall that if $D^+ = \emptyset$, then $\texttt{mindeg}(D^+) = 0$ by definition.)

Recall that the reduced basis of a monomial ideal is unique and consists of power products (§XII.5).

Lemma 13 *Let F be a reduced monomial basis, and D be the output of the modified algorithm on input $(u, \mathbf{Y}, F)$, where $\mathbf{Y} \subseteq \mathbf{X}$, $|\mathbf{Y}| \geq 1$. If D is non-trivial then for each $f \in F \cap \texttt{Cone}(u, \mathbf{Y})$, there is some $C \in D$ with $\deg C = \deg(f) - 1$.*

Proof. Let $f \in F \cap \texttt{Cone}(u, \mathbf{Y})$. First note that $\deg(f : u) \geq 1$ since otherwise $f|u$ and D is trivial. We now use double induction on $\deg(f : u)$ and $|\mathbf{Y}|$. If $\deg(f : u) = 1$ then $f = uX_0$ for some $X_0 \in \mathbf{Y}$ and by the previous lemma, D contains a cone of degree $\deg(u) = \deg(f) - 1$. If $|\mathbf{Y}| = 1$ then it is easy to see that there is at most one element f in $F \cap \texttt{Cone}(u, \mathbf{Y})$. Moreover, if such an f exists, then $f = uX_0^k$ for some $k \geq 1$ where $\mathbf{Y} = \{X_0\}$. Then clearly D contains the $\texttt{Cone}(uX_0^{k-1}, \emptyset)$ which has degree $\deg(f) - 1$. Finally, assume $\deg(f : u) \geq 2$ and $|\mathbf{Y}| \geq 2$. Then step 3 decomposes $\texttt{Cone}(u, \mathbf{Y})$ into $C_1 \oplus C_2$ where $C_1 = \texttt{Cone}(uX_0, \mathbf{Y})$ and $C_2 = \texttt{Cone}(u, \mathbf{Y} \setminus \{X_0\})$. The result then follows by inductive hypothesis. **Q.E.D.**

Here is a cone decomposition of $\texttt{NF}(I)$ that does not have the property in this lemma: let $I = \texttt{Ideal}(XYZ)$ and

$$D = \{\texttt{Cone}(1, \emptyset), \texttt{Cone}(X, \{X, Y\}), \texttt{Cone}(Y, \{Y, Z\}), \texttt{Cone}(Z, \{Z, X\})\}$$

Of course, this decomposition is not exact.

Corollary 14 *Let F be a reduced basis for an arbitrary ideal I. If $b_0, \ldots, b_{n+1}$, ($b_{n+1} = 0$), is a set of Macaulay constants for $\texttt{NF}(I)$ then $\deg(\texttt{hterm}(f)) \leq 1 + b_0$ for each $f \in F$.*

Proof. The constants $b_0, \ldots, b_n$ are uniquely determined by the requirement $b_{n+1} = 0$. But we can also obtain b_0 as $\texttt{maxdeg}(D)$ where D is is an exact decomposition for $\texttt{NF}(I)$ obtained by applying the algorithm to the input $(1, \mathbf{X}, \texttt{hterm}(F))$, since $\texttt{hterm}(F)$ is a basis for $\texttt{Head}(I)$. For $f \in F$, the preceding lemma implies that $\texttt{maxdeg}(D) \geq \deg(\texttt{hterm}(f)) - 1$ Hence $b_0 \geq \deg(\texttt{hterm}(f)) - 1$. **Q.E.D.**

Remark: If the implicit admissible ordering here is degree-compatible, we can simplify the bound in this corollary to $\deg(f) \leq 1 + b_0$. It is interesting to note that the exact decomposition algorithm gives us a constructive proof of the Hilbert-Serre theorem.

Exercises

Exercise 5.1: Let $I = \texttt{Ideal}(XY^2Z, X^2YZ^2)$. Give an exact cone decomposition of $\texttt{NF}(I)$, and determine its Macaulay constants, and Hilbert function. □

Exercise 5.2: Derive an expression for $I : (u_1, \ldots, u_k)$ involving $I : u_i$'s. Further, derive an expression involving $\texttt{LCM}(v_i, v_j)$'s for suitable v_i, v_j's. HINT: $\texttt{Ideal}(u, u') \cap \texttt{Ideal}(v, v') = \texttt{Ideal}(\texttt{LCM}(u, v), \texttt{LCM}(u, v'), \texttt{LCM}(u', v), \texttt{LCM}(u', v'))$. □

Exercise 5.3: Derive a bound on the length of the exact conversion process. □

§6. Exact Decomposition of Ideals

We construct a cone decomposition of an arbitrary ideal I. Let

$$I = \texttt{Ideal}(f_1, \ldots, f_m)$$

and

$$d = \max\{\deg(f_i) : i = 1, \ldots, m\}.$$

So I has the decomposition (§4)

$$I = S_1 \oplus S_2 \oplus \cdots \oplus S_m$$

where $S_1 = \texttt{Ideal}(f_1)$ and for $i = 2, \ldots, m$,

$$I_i = \texttt{Ideal}(f_1, \ldots, f_{i-1}), \qquad S_i = f_i \cdot \texttt{NF}(I_i : f_i).$$

We describe exact cone decompositions D_i for S_i: we may choose $D_1 = \{S_1\}$. For $i = 2, \ldots, m$, we already know how to obtain an exact decomposition $\widehat{D}_i$ of $\texttt{NF}(I_i : f_i)$. Then

$$D_i := \{f_i C : C \in \widehat{D}_i\}$$

is an exact decomposition for S_i. Hence, $\texttt{mindeg}(D_i^+) = \deg f_i$ since we may assume $\texttt{mindeg}(\widehat{D}_i^+) = 0$. Let D be the result of applying the "exact conversion" procedure to the set

$$D_2 \cup D_3 \cup \cdots \cup D_m.$$

Note that the cones in D have the form $\texttt{Cone}(f, \mathbf{Y})$ where f are no longer power products. We claim that

$$\texttt{mindeg}(D^+) \leq d. \tag{18}$$

Before the exact conversion procedure, this is clearly true. In each step of the conversion, we replace a violating cone C by its "complete shift decomposition". If this is a violation of type (I), it will not affect $\texttt{mindeg}(D^+)$. In case of a violation of type (II) or type (III), then $\deg(C) < d$, and the replacement cones have degrees $\deg(C) + 1$. Again the new $\texttt{mindeg}(D^+)$ is bounded by d. This proves (18). It is easy to modify D so that we have equality in equation (18). We will assume this in the following.

The above decomposition applies for any ideal. But in the rest of the section, we assume $f_1, \ldots, f_m$ are homogeneous polynomials. Then D is an exact cone decomposition of some homogeneous set $T \subseteq I$. The corresponding Hilbert polynomial of T has the Macaulay form (cf. equation (12))

$$\Phi_T(z) = \binom{z - d + n}{n} - 1 - \sum_{i=1}^{n} \binom{z - a_i + i - 1}{i}$$

for suitable Macaulay constants

$$a_0 \geq a_1 \geq \cdots \geq a_n \geq a_{n+1} = d. \tag{19}$$

In fact we claim that

$$a_0 = a_1. \tag{20}$$

This claim amounts to saying that if $C \in D$ is zero-dimensional then there exists $C' \in D$ with $\dim C' > 0$ and $\deg C' \geq \deg C$. This property holds vacuously for each D_i $(i = 1, \ldots, m)$ since D_i has no zero-dimensional cones. Next note that the conversion procedure applied to $D_1 \cup \cdots \cup D_m$ preserves this property because for each zero-dimensional cone C that the procedure introduces, it also produces a cone C' of positive dimension with $\deg C' = \deg C$.

We have freedom in choosing the polynomial f_1 for forming S_1, so we may assume $\deg f_1 = d$. Since $I = S_1 \oplus T$, its Hilbert polynomial can now be written as

$$\begin{aligned} \Phi_I(z) &= \Phi_{S_1}(z) + \Phi_T(z) \\ &= \binom{z - d + n - 1}{n - 1} + \binom{z - d + n}{n} - 1 - \sum_{i=1}^{n} \binom{z - a_i + i - 1}{i}. \end{aligned} \tag{21}$$

Exercise 6.1: (i) Let $I = \texttt{Ideal}(XY^2Z, X^2YZ^2)$. Give an exact cone decomposition of I, and determine its Macaulay constants, and Hilbert function.
(ii) Repeat part (i) for an ideal I that is generated by linear polynomials ($d = 1$). □

§7. Bounding the Macaulay constants

We now have exact decompositions for $I = (f_1, \ldots, f_m)$ and for $\texttt{NF}(I)$, where I is homogeneous. Let the Macaulay constants for the exact decomposition of $\texttt{NF}(I)$ be

$$b_0 \geq b_1 \geq \cdots \geq b_n \geq b_{n+1} = 0$$

with

$$\Phi_{\texttt{NF}(I)}(z) = \binom{z+n}{n} - 1 - \sum_{i=1}^{n} \binom{z - b_i + i - 1}{i}.$$

We also have the expression for $\Phi_I(z)$ with associated constants $a_0, \ldots, a_{n+1}$ (equations (21) and (19)). Therefore,

$$\begin{aligned}
\binom{z+n-1}{n-1} &= \Phi_I(z) + \Phi_{\texttt{NF}(I)}(z) \\
&= \binom{z-d+n-1}{n-1} + \binom{z-d+n}{n} + \binom{z+n}{n} \\
&\quad -2 - \sum_{i=1}^{n} \left\{ \binom{z-a_i+i-1}{i} + \binom{z-b_i+i-1}{i} \right\}. \qquad (22)
\end{aligned}$$

We now apply the "backwards difference operator" ∇, defined by its effect on real functions $F(z)$,

$$\nabla F(z) := F(z) - F(z-1).$$

For $i \geq 0$, let $\nabla^0 F = F$ and $\nabla^{i+1} F = \nabla(\nabla^i F)$. It is easy to verify the identity $\nabla\binom{z-k}{n} = \binom{z-k-1}{n-1}$. Applied i times,

$$\nabla^i \binom{z-k}{n} = \binom{z-k-i}{n-i}.$$

This formula applies for all values of i, recalling that by definition of the binomial coefficients,

$$\binom{z}{k} = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k < 0. \end{cases}$$

Clearly, $\nabla(F+G) = (\nabla F) + (\nabla G)$, and $\nabla C = 0$ iff C is a constant. Applying ∇^j ($j \geq 0$) to both sides of equation (22),

$$\begin{aligned}
\binom{z+n-j-1}{n-j-1} &= \binom{z-d+n-j-1}{n-j-1} + \binom{z-d+n-j}{n-j} + \binom{z+n-j}{n-j} - 2\delta(j,0) \\
&\quad - \sum_{i=j}^{n} \left\{ \binom{z-a_i+i-j-1}{i-j} + \binom{z-b_i+i-j-1}{i-j} \right\}. \qquad (23)
\end{aligned}$$

Here

$$\delta(x,y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{else,} \end{cases}$$

denotes the Kronecker delta function. Setting $z = 0$ and using the identity

$$\binom{k}{n} = (-1)^n \binom{n-k-1}{n}, \qquad k < 0,$$

equation (23) becomes

$$1 = (-1)^{n-j-1}\binom{d-1}{n-j-1} + (-1)^{n-j}\binom{d-1}{n-j} + 1 - 2\delta(j,0) - \sum_{i=j}^{n}(-1)^{i-j}\left\{\binom{a_i}{i-j} + \binom{b_i}{i-j}\right\}. \tag{24}$$

For $j = n-1$, this reduces to

$$a_n + b_n = d.$$

But we know $a_n \geq d, b_n \geq 0$. Hence $a_n = d, b_n = 0$. It is convenient to define

$$c_j = a_j + b_j, \qquad j = 0, \ldots, n+1.$$

Thus we have just shown $c_n = d$. Next we simplify equation (24) by extracting the terms $\binom{a_i}{i-j}+\binom{b_i}{i-j}$ in the summation corresponding to $i = j$, $i = j+1$ and $i = n$. These three terms (respectively) are

$$2, \quad -c_{j+1}, \quad (-1)^{n-j}\binom{d}{n-j}.$$

This gives us, for $j = 0, \ldots, n-2$,

$$c_{j+1} = 2(1-\delta(j,0)) + 2(-1)^{n-j}\binom{d-1}{n-j-1} + \sum_{i=j+2}^{n-1}(-1)^{i-j}\left\{\binom{a_i}{i-j} + \binom{b_i}{i-j}\right\}, \tag{25}$$

For $j = n-2$, this yields

$$c_{n-1} = 2d.$$

We now use the inequality

$$\binom{a_i}{k} + \binom{b_i}{k} \leq \binom{c_i}{k}$$

(this is obvious from the combinatorial interpretation). Note that equation (25) (for $j \leq n-3$) contains the term

$$2 + (-1)^{n-j}\left\{2\binom{d-1}{n-j-1} - \binom{a_{n-1}}{n-j-1} - \binom{b_{n-1}}{n-j-1}\right\}.$$

If $j = n-3$, this expression is bounded by $2 - 2\binom{d-1}{2} + \binom{2d}{2} = d(d+2)$. From (25), we get

$$c_{n-2} \leq d(d+2).$$

But to get a general bound, proceed as follows. Extracting the term corresponding to $i = j+2$ in (25), and discarding the negative term corresponding to $i = j+3$ and also discarding the Kronecker-delta,

$$c_{j+1} \leq 2 + 2\binom{d-1}{n-j-1} + \binom{c_{j+2}}{2} + \sum_{i=j+4}^{n-1}\binom{c_i}{i-j}. \tag{26}$$

In the following, we assume $d \geq 2$ (see §6, Exercise, for $d = 1$).

Lemma 15 *For $d \geq 2$,*

$$c_j < B_j := d^{2^{n-j}}, \qquad j = 1, \ldots, n-2.$$

Proof. The result is true for $j = n-2$. We may also note that $c_{n-1} \leq B_{n-1} = d^2$. So let $j \leq n-3$. For $i = j+4, \ldots, n-1$, we have $i - j \leq 2^{i-j-2}$ and so

$$\binom{B_i}{i-j} < \frac{B_i^{i-j}}{(i-j)!} \leq \frac{B_i^{2^{i-j-2}}}{(i-j)!} = \frac{B_{j+2}}{(i-j)!}.$$

Also, check that $2 + 2\binom{d-1}{n-j-1} \leq 2B_{j+2}$. Inductively, equation (26) yields

$$\begin{aligned} c_{j+1} &\leq 2B_{j+2} + \binom{B_{j+2}}{2} + \sum_{i=j+4}^{n-1} \binom{B_i}{i-j} \\ &\leq 2B_{j+2} + \frac{B_{j+1}}{2} - \frac{B_{j+2}}{2} + \sum_{i \geq j+4} \frac{B_{j+2}}{(i-j)!} \\ &\leq \frac{B_{j+1}}{2} + B_{j+2}\left\{\frac{3}{2} + \sum_{i \geq 4} \frac{1}{i!}\right\} \\ &< \frac{B_{j+1}}{2} + 2B_{j+2} \\ &\leq B_{j+1}. \end{aligned}$$

Q.E.D.

In particular, we have $a_1 + b_1 = c_1 \leq B_1 = d^{2^{n-1}}$. We already know $a_0 = a_1$. The remaining constant to be bounded is b_0.

Lemma 16

$$b_0 \leq \max\{a_1, b_1\}.$$

Proof. Let D be the cone decomposition for $K[\mathbf{X}]$, constructed relative to I as in the foregoing. The Hilbert function of $K[\mathbf{X}]$ can be expressed as

$$\left\{\sum_{i \geq 0} W_{0,i}\delta(z,i)\right\} + \left\{\sum_{d=0}^{n}\sum_{i \geq 0} W_{d,i} \cdot \binom{z-i+d-1}{d-1}_i\right\}$$

where $W_{d,i}$ is the number of d-dimensional cones of degree i in D. Here, we attach a subscript "i" on the binomial coefficient $\binom{z-i+d-1}{d-1}_i$ to indicate that the expression must be equated to zero for $z < i$. For $z \geq \max\{a_1, b_1\}$, these subscripts can be removed, giving

$$\left\{\sum_{i \geq 0} W_{0,i}\delta(z,i)\right\} + \left\{\sum_{d=0}^{n}\sum_{i \geq 0} W_{d,i} \cdot \binom{z-i+d-1}{d-1}\right\} = \binom{z+n-1}{n-1}. \tag{27}$$

Hence, for $z \geq \max\{a_1, b_1\}$, equation (27) shows that the expression $\sum_{i \geq 0} W_{0,i}\delta(z,i)$ is equal to a sum of binomial coefficients, and so is equal to some integer polynomial $p(z)$. But for $z \geq \max\{a_0, b_0\}$ ($\geq \max\{a_1, b_1\}$) we know that $p(z) = 0$, so $p(z)$ is the zero polynomial. This implies that $b_0 \leq \max\{a_1, b_1\}$. **Q.E.D.**

To obtain our main result about general polynomials, we need to first homogenize the input polynomials (§3).

Theorem 17 *Let $f_1, \ldots, f_m \subseteq K[X_1, \ldots, X_n]$ be polynomials of degree at most d. Relative to any admissible order $\underset{\mathtt{A}}{\leq}$, there is a Gröbner basis G for $I = \mathtt{Ideal}(f_1, \ldots, f_m)$ of degree at most d^{2^n}. In other words,*

$$G(n,d) \leq d^{2^n}.$$

Proof. First we homogenize each f_i to $F_i \in K[X_0, \ldots, X_n]$ where X_0 is a homogenizing variable. Let G be a reduced basis for $J = \texttt{Ideal}(F_1, \ldots, F_m)$ relative to $\leq^{\wedge}_{\texttt{A}}$ (§3). We want an upper bound on $B = \max\{\deg(g) : g \in G\}$. The set $F = \{\texttt{hterm}(f) : f \in G\}$ is a reduced monomial basis for the head ideal $H = \texttt{Head}(G)$. Applying the exact cone decomposition algorithm to $(1, \mathbf{X}, F)$, we obtain an exact decomposition for $\texttt{NF}(H)$. But G is a Gröbner basis means $\texttt{NF}(H) = \texttt{NF}(J)$. We similarly obtain an exact decomposition for J. As above, the associated Macaulay constants are $b_0, \ldots, b_{n+2}$ and $a_0, \ldots, a_{n+2}$, respectively. (Note that there are $n+3$ Macaulay constants for each decomposition since we have $n+1$ variables.) By lemma 13 we know that $B \leq 1 + b_0$. But

$$b_0 \leq \max\{a_1, b_1\} \leq c_1 \leq d^{2^n}.$$

Finally, dehomogenizing the polynomials in G yields a Gröbner basis for I (lemma 3). **Q.E.D.**

Exercises

Exercise 7.1: Obtain the sharpest bound you can on the above constants a_j, b_j. For instance, $c_{n-3} \leq \binom{2d}{3} + \binom{d^2+d}{2}$. In general,

$$c_j \leq 2\left(\frac{d^2}{2} + d\right)^{2^{n-j-1}}.$$

□

§8. Term Rewriting Systems

The rest of this lecture proves a lower bound on $G(n,d)$, $I(n,d)$ and $S(n,d)$. We now assume the coefficient field K is $\mathbb{Q}$. The underlying set of variables will now be denoted $\Sigma = \{X_1, \ldots, X_n\}$.

The context for the lower bound construction comes from the term-rewriting literature. We had a brief encounter with term-rewriting systems in §XII.3. Such systems deal with syntactic objects ("free terms") and their behavior under a set of transformations. These transformations are governed by finitistic or local rules. For example, let the terms be words over an alphabet Σ. A simple rule might be the transposition of two adjacent variables occurring in a word: $\ldots XY \ldots \to \ldots YX \ldots$. This particular rule amounts to saying that the variables are *commutative*. If the rules are reversible, then this defines an equivalence relation among the terms. Questions of recognizing equivalence are important issues, and, as in Gröbner bases, this can be studied via normal forms. Indeed, many general questions of interest in this subject are already reflected in our study of Gröbner bases: termination for a set of rules, normal forms, Church-Rosser property.

We will now study polynomial transformations from this "syntactic" view-point. In fact, the lower bound results use a very special class of polynomials which we call *Thue polynomials*, and the corresponding term-rewriting systems are called *Thue systems*. Many concepts of term-rewriting systems have analogues in polynomial rings. In such cases, it is a simple matter of introducing the corresponding vocabulary. Generally, we will try to adopt the new vocabulary.

Let $\Sigma = \{X_1, \ldots, X_n\}$ be a set of variables. In term-rewriting, Σ is a called *alphabet* and each $X_i \in \Sigma$ is a *commutative variable*. Although commutativity of variables is the normal assumption

in polynomial rings, the opposite assumption is more common in term-rewriting. In the following definitions, we put the qualification "commutative" in parenthesis when defining the appropriate concepts. *However, we normally drop the 'commutative' qualifications in our usage, as this will be understood.*

The set of power products $\mathtt{PP}(\Sigma)$ is also called the free commutative monoid generated by Σ. A power product $w \in \mathtt{PP}(\Sigma)$ is also called a (commutative) *word*. Because of commutativity, ww' and $w'w$ both refer to the same word. Thus the *empty word* corresponds to "1". If $\Gamma \subseteq \Sigma$ and $w, v \in \mathtt{PP}(\Sigma)$, then $\deg_\Gamma(w)$ denotes the degree of w in the variables of Γ. For instance, we say w is *linear* in Γ if $\deg_\Gamma(w) = 1$. If $\Gamma = \Sigma$, then $\deg_\Gamma(w)$ is the usual degree, $\deg(w)$. We say w is a *subword* of w' if $w|w'$ (w divides w').

A *(commutative) semi-Thue system* over Σ is a pair (S, Σ) where S is a finite set of pairs in $\mathtt{PP}(\Sigma)$. Each pair $(\alpha, \beta) \in S$ is called a *rule*. We call α and β (respectively) the *precondition* and *postcondition* of the rule (α, β). The *reverse* of a rule (α, β) is the rule (β, α). We say S is a *(commutative) Thue system* if (α, β) is in S implies its reverse (β, α) is in S. The *degree* of a rule (α, β) is $\max\{\deg(\alpha), \deg(\beta)\}$, and the *degree* of S is the maximum degree of its rules.

Derivations. Given words v, w, we write

$$v \to w \pmod{S}$$

if for some $(\alpha, \beta) \in S$ and $\gamma \in \mathtt{PP}(\Sigma)$, we have $v = \gamma\alpha$ and $w = \gamma\beta$. We also call "$v \to w$" a *transition* (of S). The reflexive, transitive closure of $\to \pmod{S}$ is denoted $\xrightarrow[*]{} \pmod{S}$. A sequence

$$D = (w_1, w_2, \ldots, w_k)$$

(also written, $D : w_1 \to \cdots \to w_k$, or, $D : w_1 \xrightarrow[*]{} w_k$) where $w_i \in \mathtt{PP}(\Sigma)$ ($k \geq 1$) is called a *derivation* of S from w_1 to w_k if $w_i \to w_{i+1} \pmod{S}$ for $i = 1, \ldots, k-1$. The derivation has a *repetition* if $w_i = w_j$ for some $1 \leq i < j \leq k$; this instance of repetition is *trivial* if $i = j - 2$. A derivation is *repetition-free* if it has no repetition; it is *simple* if it has no trivial repetition. Thus simple derivations are allowed to have non-trivial repetitions. Clearly only simple derivations are of interest, and this is often implicit in our discussion. A repetition-free derivation $D : w \xrightarrow[*]{} w'$ is said to be *unique* if it is the only repetition-free derivation from w to w'. We say D is *strongly unique* if for all repetition-free derivations $D' : w \xrightarrow[*]{} u$, if $w' \,|\, u$ then $u = w'$ and $D' = D$. A word w is *recursive* if there is a non-trivial (i.e., at least one transition step) simple derivation from w to some w' such that $w \,|\, w'$; otherwise it is *non-recursive*. Note that possibly $w = w'$ here.

For any set $G = \{g_1, \ldots, g_m\}$ or sequence $\overline{g} = (g_1, \ldots, g_m)$ of polynomials, let $\mathtt{maxdeg}(G)$ (resp., $\mathtt{maxdeg}(\overline{g})$) be the maximum degree of a polynomial in G (resp., in $\overline{g}$). Here, $\overline{g}$ may represent a syzygy or a derivation and G a Gröbner basis.

We are often interested in derivations of a Thue system S from some distinguished word w_0. We call w_0 the *initial assertion* of S and words derivable from the initial assertion are called *assertions of S*. (This terminology is from Post [11].)

Main Task. The main task before us (following [9]) involves constructing a Thue system $S_{n,d}$ that "counts to d^{2^n}" in the following sense: there are distinguished variables $Q_0, Q_\infty, A \in \Sigma$ such that if D is any simple derivation

$$D : Q_0 \xrightarrow[*]{} Q_\infty w \pmod{S_{n,d}}$$

then w is equal to $A^{d^{2^n}}$. Here A is called an "accumulator" variable since it acts as a unary counter. Moreover, the derivation D is strongly unique. For the lower bound application, this construction needs to be "efficient" in the sense that $S_{n,d}$ involves only $O(n)$ variables and has degree $O(n+d)$.

Connection to Polynomial Ideals. Such a construction has implications for polynomials because we can view each rule (α, β) as the polynomial $\alpha - \beta$. Let us call a polynomial of this form a *Thue polynomial* and an ideal generated by a set of Thue polynomials is called a *Thue ideal* [13, 14]. Let

$$F_S := \{\alpha - \beta : (\alpha, \beta) \in S\} \tag{28}$$

be the set of Thue polynomials corresponding to rules of S. For $w, w' \in \mathtt{PP}(\Sigma)$, notice that if $w \longrightarrow w' (\mathrm{mod}\, S)$ then $w = \alpha\gamma$ and $w' = \beta\gamma$ for some rule $(\alpha, \beta) \in S$. Then $w - w' \in \mathtt{Ideal}(F_S)$. This argument can be repeated for derivations of any length, giving:

Lemma 18 *If* $w \xrightarrow{*} w' (\mathrm{mod}\, S)$ *then* $w - w' \in \mathtt{Ideal}(F_S)$.

Let us now show a partial converse. For this, we need to assume that the field of coefficients is $K = \mathbb{Q}$. For $f \in K[\Sigma]$, define the *support* of f be the set $\mathtt{support}(f) \subseteq \mathtt{PP}(\Sigma)$ comprising those words whose coefficients in f are non-zero.

Lemma 19 (Mayr-Meyer) *Suppose* $w - w' \in \mathtt{Ideal}(F_S)$ *where the ideal is generated over* $R = \mathbb{Q}[X_1, \ldots, X_n]$ *then there exists a derivation* $D : w \xrightarrow{*} w' (\mathrm{mod}\, S)$. *Moreover, if*

$$w - w' = \sum_{i=1}^{m} a_i f_i \tag{29}$$

where $a_i \in R$ *and* $f_i \in F_S$, *then* $\mathtt{maxdeg}(D) \leq \mathtt{maxdeg}\{a_i f_i : i = 1, \ldots, m\}$.

Proof. We may assume that each a_i in (29) is a monomial. By multiplying both sides by a suitable positive integer d, we obtain an expression

$$d(w - w') = \sum_{i=1}^{m} b_i f_i$$

where each b_i is now a power product. We want to construct a derivation

$$D : w = w_0 \longrightarrow w_1 \longrightarrow \cdots \longrightarrow w_k = w' (\mathrm{mod}\, S)$$

for some $k \leq m$. This is done by induction on m. It is easiest to see the argument graphically. Let us construct a digraph G on the vertex set $V = \cup_{i=1}^{m} \mathtt{support}(b_i f_i)$, where we introduce an edge from u to v whenever $b_i f_i = u - v$ $(i = 1, \ldots, m)$. Note that G may have multiple edges (*i.e.*, there may be more than one edge going from any vertex u to any v). Let $d_G(u)$ be equal to the number of outgoing edges from u minus the number of incoming edges into u. Clearly

$$d_G(u) = \begin{cases} 0 & \text{if } u \notin \{w, w'\}, \\ d & \text{if } u = w, \\ -d & \text{if } u = w'. \end{cases}$$

It easily follows by induction on m that there exists a path from w to w'. This is immediate if $m = 1$. Otherwise, assume that we have constructed a path $(w_0, w_1, \ldots, w_i)$ where $w_i \neq w'$ $(i \geq 0)$.

Moreover, we have removed the edges of this path from G (so G now has $m - i$ edges). Then it is easy to see that there remains an outgoing edge from w_i, and so this construction can proceed. We leave the formal argument to the reader. But any path from w to w' corresponds to a derivation from w to w'. It is also not hard to see that the corresponding derivation has degree bounded by $\texttt{maxdeg}\{a_i f_i : i = 1, \ldots, m\}$. **Q.E.D.**

Exercises

Exercise 8.1: Let $\Sigma = \{Q_0, A, B, Q_1\}$. Construct a Thue system S such that for any $w \in \texttt{PP}(A, B)$, we have $Q_0 w \xrightarrow{*} Q_1 A^m B^n (\mathrm{mod}\, S)$ where $m = \deg_A(w)$ and $n = \deg_B(w)$. Moreover, this derivation is strongly unique. S may use a larger alphabet than Σ. □

Exercise 8.2: The reduced Gröbner basis (with respect to any admissible ordering) of a Thue ideal is comprised of Thue polynomials. HINT: analyze Buchberger's algorithm and note that the S-polynomial of two Thue polynomials is a Thue polynomial. □

§9. A Quadratic Counter

We construct a Thue system S_0 that can be viewed as a "counter".

Constants. Throughout this construction, we fix the integers $n \geq 1$ and $d \geq 2$. We also define $e_d(k) = e(k) := d^{2^k}$ for all $k \geq 0$. Observe that $e(k+1) = e(k)^2$.

Variables. There are two types of variables.

1. *Accumulator variables:* A_k, B_k, $(k = 0, 1, \ldots, n)$
2. *Flag variables:* $F_k[\texttt{color}]$ $(k = 1, \ldots, n;\ \texttt{color} \in \{\texttt{inc}, \texttt{dec}, \texttt{pass}\})$.

Let Σ_0 denote this set of $5n+2$ variables. Each variable belongs to some *level* k (its subscript) which is an integer between 0 and n. The colors are read 'increment', 'decrement' and 'pass', respectively. We are mainly interested in commutative words of the following form:

$$w = A_0^{m_0} B_0^{n_0} \prod_{k=1}^{n} A_k^{m_k} B_k^{n_k} F_k[\texttt{color}_k]$$

where $m_0, n_0, m_k, n_k \geq 0$ and $m_0 + n_0 = d$. We call such words *well-formed*.

So a well-formed word is linear in $\{F_k[\texttt{inc}], F_k[\texttt{dec}], F_k[\texttt{pass}]\}$ for each $k = 1, \ldots, n$.

For $1 \leq k \leq n$ and $k \leq \ell \leq n+1$, we use the abbreviation:

$$F_{k,\ell}[\texttt{color}] \equiv \prod_{i=k}^{\ell-1} F_i[\texttt{color}]$$

Thus $F_{k,k}[\texttt{color}] = 1$. The following well-formed word is designated the *initial assertion*

$$w_0 := A_0^d F_{1,n+1}[\texttt{inc}]. \tag{30}$$

We now present the rules, which naturally fall under two groups:

Start Rules. $(k = 1, 2, \ldots, n)$

$$(\mathrm{S1})_k \quad A_0 \xrightarrow{F_{1,k}[\mathtt{pass}]F_k[\mathtt{inc}]} B_0 A_k^d \qquad (\text{"increment rule"})$$

$$(\mathrm{S2})_k \quad A_0 B_k^d \xrightarrow{F_{1,k}[\mathtt{pass}]F_k[\mathtt{dec}]} B_0 \qquad (\text{"decrement rule"})$$

Finish Rules. $(k = 1, 2, \ldots, n-1)$

$$(\mathrm{F1})_k \quad B_0^d F_{1,k}[\mathtt{dec}]F_k[\mathtt{inc}] \longrightarrow A_0^d F_{1,k}[\mathtt{inc}]F_k[\mathtt{pass}] \qquad (\text{"inc} \Rightarrow \text{pass rule"})$$

$$(\mathrm{F2})_k \quad B_0^d F_{1,k}[\mathtt{dec}]A_k \xrightarrow{F_k[\mathtt{pass}]} A_0^d F_{1,k}[\mathtt{inc}]B_k \qquad (\text{"pass} \Rightarrow \text{pass rule"})$$

$$(\mathrm{F3})_k \quad B_0^d F_{1,k}[\mathtt{dec}]F_k[\mathtt{pass}]A_k \longrightarrow A_0^d F_{1,k}[\mathtt{inc}]F_k[\mathtt{dec}]B_k \qquad (\text{"pass} \Rightarrow \text{dec rule"})$$

Let (S_0, Σ_0) denote the Thue system corresponding to these rules.

Remark: We write the rule (α, β) as $\alpha \to \beta$ above to be suggestive of the "forward" direction of applying the rules. But one must remember that we are describing a Thue system, so the reverse rule $\beta \to \alpha$ is also implied. A derivation that only uses the forward (resp. reverse) rules will be called a *forward* (resp. *reverse*) derivation; otherwise the derivation is *mixed*. Furthermore, a rule of the form $\alpha\gamma \to \beta\gamma$ where $\mathtt{GCD}(\alpha, \beta) = 1$ may be written as

$$\alpha \overset{\gamma}{\longrightarrow} \beta$$

as in the rules[4] (S1) and (S2). Since γ is unchanged, we call it the "catalyst" for the rule. In this and the next section, all transitions $\to$, $\overset{*}{\to}$ are understood to be (mod S_0).

Example: We illustrate a derivation of S_0. Note that any word $w \in \mathtt{PP}(\Sigma_0)$ can be expressed as $w = w_0 w_1 \cdots w_n$ where w_k $(k = 0, \ldots, n)$ is the subword of w comprising all variables of level k. In the following derivation, we will represent w by stacking w_{k-1} above w_k. Here $d = 2, n = 3$.

$$w_0 = \begin{pmatrix} A_0^2 \\ F_1[\mathtt{inc}] \\ F_2[\mathtt{inc}] \\ F_3[\mathtt{inc}] \end{pmatrix} \xrightarrow{rule(\mathrm{S1})_1} \begin{pmatrix} A_0 B_0 \\ F_1[\mathtt{inc}]A_1^2 \\ F_2[\mathtt{inc}] \\ F_3[\mathtt{inc}] \end{pmatrix} \xrightarrow{rule(\mathrm{S1})_1}$$

$$\begin{pmatrix} B_0^2 \\ F_1[\mathtt{inc}]A_1^4 \\ F_2[\mathtt{inc}] \\ F_3[\mathtt{inc}] \end{pmatrix} \xrightarrow{rule(\mathrm{F1})_1} \begin{pmatrix} A_0^2 \\ F_1[\mathtt{pass}]A_1^4 \\ F_2[\mathtt{inc}] \\ F_3[\mathtt{inc}] \end{pmatrix} \xrightarrow{rule(\mathrm{S1})_1} \begin{pmatrix} A_0 B_0 \\ F_1[\mathtt{pass}]A_1^4 \\ F_2[\mathtt{inc}]A_2^2 \\ F_3[\mathtt{inc}] \end{pmatrix}$$

$$\longrightarrow \quad \cdots \quad \longrightarrow \begin{pmatrix} B_0^2 \\ F_1[\mathtt{dec}] \\ F_2[\mathtt{dec}] \\ F_3[\mathtt{inc}]A_3^{256} \end{pmatrix}.$$

The reader should try to see how to eventually derive the last indicated word. ■

Let us briefly comment on the rules: the forward start rules at level k are all dependent on the catalyst $F_{1,k}[\mathtt{pass}]$ and in each case, convert an occurrence of A_0 to B_0. Rule (S1), in the presence

[4]By a harmless abuse of language, we call, for instance, (S1) a 'rule' even though it is really a family of n rules parameterized by k.

of the catalyst $F_k[\texttt{inc}]$, increments the accumulator A_k by d: in this, we see that the pass flags $F_i[\texttt{pass}]$ at levels $i = 1$ to $i = k-1$ each signal that the accumulator A_i should be ignored. Rule (S2) is the counterpart of (S1) where the flag $F_k[\texttt{dec}]$ signals the decrementing of accumulator A_k by d.

We similarly note some salient features of the forward finish rules: in each case, the subword B_0^d transforms to A_0^d. Furthermore, the initial block of "decrement flags" $F_{1,k}[\texttt{dec}]$ is converted to $F_{1,k}[\texttt{inc}]$. Hence differences among the finish rules hinge on the flag at level k: in case of (F1), the flag is $F_k[\texttt{inc}]$ and we convert it to $F_k[\texttt{pass}]$. In the case of (F2), the flag $F_k[\texttt{pass}]$ acts only as a catalyst for converting an A_k into a B_k. In (F3), we convert $F_k[\texttt{pass}]$ to $F_k[\texttt{dec}]$, and at the same time convert an A_k to a B_k. Thus the flag variable at level k is transformed by these rules in a cyclic fashion

$$\texttt{inc} \Rightarrow \texttt{pass} \Rightarrow \texttt{dec} \Rightarrow \texttt{inc} \Rightarrow \cdots.$$

More precisely: *in any forward derivation, the flag variable at each level transforms cyclically as indicated.*

Clearly the rules preserve well-formedness of words. In particular, all assertions are well-formed. To see these rules in action, it is best to follow the proof of the following lemma:

Lemma 20 (Standard Derivation) *Let $1 \leq k \leq \ell \leq n$ and w be any commutative word.*

$(\mathbf{a})_{\mathbf{k},\ell}$ *(Increment) Let $u_1 = wA_0^d F_{1,k}[\texttt{inc}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}]$. Then there is a forward derivation $u_1 \stackrel{*}{\to} u_1'$ where*

$$u_1' = wB_0^d A_\ell^{e(k)} F_{1,k}[\texttt{dec}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}].$$

The first and last rules in this derivation are start rules.

$(\mathbf{b})_{\mathbf{k},\ell}$ *(Decrement) Let $v_1 = wA_0^d B_\ell^{e(k)} F_{1,k}[\texttt{inc}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{dec}]$. Then there is a forward derivation $v_1 \stackrel{*}{\to} v_1'$ where*

$$v_1' = wB_0^d F_{1,k}[\texttt{dec}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{dec}].$$

The first and last rules in this derivation are start rules.

Proof. Let $k = 1$. Part (a) consists of d applications of rule $(\text{S1})_\ell$, and part (b) consists of d applications of rule $(\text{S2})_\ell$. Now assume $k > 1$.

$\mathbf{(a)}_{k,\ell}$

$$
\begin{array}{lll}
 & wA_0^d F_{1,k}[\texttt{inc}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_1 \\
\xrightarrow[*]{\texttt{ind}(a)_{k-1,k-1}} & wB_0^d A_{k-1}^{e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{inc}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_2 \\
\xrightarrow{\text{(F1)}} & wA_0^d A_{k-1}^{e(k-1)} F_{1,k-1}[\texttt{inc}]F_{k-1}[\texttt{pass}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_3 \\
\xrightarrow[*]{\texttt{ind}(a)_{k-1,\ell}} & wB_0^d A_{k-1}^{e(k-1)} A_\ell^{e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_4 \\
\xrightarrow{\text{(F2)}} & wA_0^d A_{k-1}^{e(k-1)-1} B_{k-1} A_\ell^{e(k-1)} F_{1,k-1}[\texttt{inc}]F_{k-1,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_5 \\
\xrightarrow[*]{\texttt{ind}(a)_{k-1,\ell}} & wB_0^d A_{k-1}^{e(k-1)-1} B_{k-1} A_\ell^{2e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_6 \\
\vdots & \text{(by } e(k-1)-2 \text{ more applications of last 2 steps)} & \\
\xrightarrow[*]{\texttt{ind}(a)_{k-1,\ell}} & wB_0^d A_{k-1} B_{k-1}^{e(k-1)-1} A_\ell^{e(k)} F_{1,k-1}[\texttt{dec}]F_{k-1,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_7 \\
\xrightarrow{\text{(F3)}} & wA_0^d B_{k-1}^{e(k-1)} A_\ell^{e(k)} F_{1,k-1}[\texttt{inc}]F_{k-1}[\texttt{dec}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_8 \\
\xrightarrow[*]{\texttt{ind}(b)_{k-1,k-1}} & wB_0^d A_\ell^{e(k)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{dec}]F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] & = u_1'
\end{array}
$$

The inductive invocations of the lemma are labeled "ind(a)" or "ind(b)".

(b)$_{k,\ell}$: We omit the similar derivation. The only difference is that all the inductive invocations of "ind(a)$_{k-1,\ell}$" are replaced by "ind(b)$_{k-1,\ell}$". **Q.E.D.**

We shall call any prefix of the derivations (a) or (b) in the proof of this lemma a *standard derivation* (at level k). To understand how the system may deviate from standard derivations, we consider the 'ambiguities' that arise when the precondition of a rule R subsumes the precondition of another rule R$'$. This would mean that whenever R is applicable, so is R$'$, causing non-determinism in derivations. In general, non-determinism arises even when the rules are not ambiguous in this sense. But in S_0, such ambiguities are the only cause of deviation from our intended or standard derivations. Of course, we must consider the reverse rules in describing ambiguities.

Forward Ambiguities Among the forward rules, the only source of ambiguity arises from the fact that both (F2) and (F3) have the same preconditions.

Reverse Ambiguities Among the reverse rules, the only ambiguity arises because the precondition of reverse (F2) subsumes the precondition of reverse (F1).

Mixed Ambiguities Finally, consider ambiguities involving a forward and a reverse rule. The problem here is essentially caused by the case $k = 1$. Thus the precondition for rule $(\text{S1})_1$ is subsumed by the postcondition of the finish rules $(\text{F}m)_\ell$ (for $m = 1, 2, 3$ and $\ell \geq 2$). Similarly, the postcondition of rule $(\text{S2})_1$ is subsumed by the precondition of the finish rules $(\text{F}m)_\ell$ (for $m = 1, 2, 3$ and $\ell \geq 2$). It turns out that these mixed ambiguities are harmless for our purposes.[5]

[5]The fact that the postcondition of $(\text{F}m)_\ell$ subsumes the precondition of $(\text{S1})_1$ means that following an application of $(\text{F}m)_\ell$, we can immediately apply rule $(\text{S1})_1$. But the reader may check that in standard derivations, we always apply rule $(\text{S1})_1$ right after rule $(\text{F}m)_\ell$! Similarly, that the precondition of rule $(\text{F}m)_\ell$ subsumes the postcondition

The following properties of S_0 may now be noted. Let w be any well-formed word.

- Rule (F2) is applicable to w if and only if rule (F3) is applicable to w. If both are applicable, then no other rules apply; if both are non-applicable, then there is at most one forward rule applicable to w.
- If reverse (F2) is applicable to w, then reverse (F1) (and no other reverse rules) is applicable to w. If reverse (F2) is not applicable, then there is at most one reverse rule applicable to w.

These properties will be assumed when we prove the Basic Lemma next.

§10. Uniqueness Property

If we apply lemma 20(a) to the initial assertion w_0 with the parameters $k = \ell = n$, we get the word

$$w_\infty := B_0^d A_n^{e(n)} F_{1,n}[\texttt{dec}] F_n[\texttt{inc}]. \tag{31}$$

We call w_∞ the *final assertion*. The next lemma shows that this derivation is unique.

For any word w, define the *level* of w to be the smallest $k = 0, 1, \ldots, n$ such that if $\Gamma := \{A_k, B_k, F_k[\texttt{inc}], F_k[\texttt{pass}], F_k[\texttt{dec}]\}$ then $\deg_\Gamma(w) > 0$.

Lemma 21 (Basic Lemma) *Let $1 \le k \le \ell \le n$ and let w_1, w_1' be commutative words of level at least k and the following are well-formed words:*

$$u_1 := w_1 A_0^d F_{1,k}[\texttt{inc}], \quad u_1' := w_1' B_0^d F_{1,k}[\texttt{dec}].$$

$(\mathbf{a})_{\mathbf{k},\ell}$ *(Forward Derivation) Let*

$$D_1 : u_1 \xrightarrow{*} u_1'$$

be a simple derivation such that the first transition in D_1 is a forward one and u_1' is the only word in D_1 divisible by $B_0^d F_{1,k}[\texttt{dec}]$. Then D_1 is unique and a standard derivation, and moreover:

$(\mathbf{a}.\mathbf{1})_{\mathbf{k},\ell}$ *If $F_{k,\ell}[\texttt{pass}] F_\ell[\texttt{inc}] \,|\, w_1$ then $w_1 A_\ell^{e(k)} = w_1'$, i.e., the accumulator A_ℓ has increased by $e(k)$.*

$(\mathbf{a}.\mathbf{2})_{\mathbf{k},\ell}$ *If $F_{k,\ell}[\texttt{pass}] F_\ell[\texttt{dec}] \,|\, w_1$ then $w_1 = w_1' B_\ell^{e(k)}$, i.e., the accumulator B_ℓ has decreased by $e(k)$.*

$(\mathbf{b})_{\mathbf{k},\ell}$ *(Backward Derivation) Let*

$$D_2 : u_1' \xrightarrow{*} u_1$$

be a simple derivation such that the first transition in D_2 is a reverse one and u_1 is the only word in D_2 divisible by $A_0^d F_{1,k}[\texttt{inc}]$. Then D_2 is unique and the reverse of a standard derivation, and moreover:

$(\mathbf{b}.\mathbf{1})_{\mathbf{k},\ell}$ *If $F_{k,\ell}[\texttt{pass}] F_\ell[\texttt{inc}] \,|\, w_1'$ then $w_1' = w_1 A_\ell^{e(k)}$, i.e., the accumulator A_ℓ has decreased by $e(k)$.*

$(\mathbf{b}.\mathbf{2})_{\mathbf{k},\ell}$ *If $F_{k,\ell}[\texttt{pass}] F_\ell[\texttt{dec}] \,|\, w_1'$ then $w_1' A_\ell^{e(k)} = w_1$, i.e., the accumulator B_ℓ has increased by $e(k)$.*

of $(\mathrm{S}2)_1$ means that, whenever rule $(\mathrm{F}m)_\ell$ is applicable, so is the reverse of $(\mathrm{S}2)_1$. But in standard derivations, we only apply rule $(\mathrm{F}m)_\ell$ after an application of rule $(\mathrm{S}2)_1$. Hence it is impossible to deviate from standard behavior using this ambiguity: it would mean that we apply reverse $(\mathrm{S}2)_1$ instead of $(\mathrm{F}m)_\ell$. But this would give a non-simple derivation.

We defer the proof to the appendix. The following shows that S_0 is (with some simple modifications) counting up to the double-exponential number $e(n)$.

Corollary 22 *If $D : w_0 \xrightarrow{*} w$ and $B_0^d F_{1,n}[\texttt{dec}]F_n[\texttt{inc}] \,|\, w$ then D is unique and $w = w_\infty$. Hence, the standard derivation $w_0 \xrightarrow{*} w_\infty$ is strongly unique.*

Proof. Only a forward rule can be applied to w_0, and there is a first word x in D such that $B_0^d F_{1,n}[\texttt{dec}]F_n[\texttt{inc}] \,|\, x$. By the basic lemma $(\text{a})_{1,n}$, x has the form of the final assertion,

$$w_\infty = B_0^d A_n^{e(n)} F_{1,n}[\texttt{dec}]F_n[\texttt{inc}].$$

Moreover, the derivation cannot be extended from x (recall that the finish rules does not include the case $k = n$). So x is indeed equal to w. The uniqueness of D follows from the uniqueness of standard derivations. **Q.E.D.**

Lemma 23 *The initial assertion $w_0 = A_0^d F_{1,n+1}[\texttt{inc}]$ is non-recursive.*

Again, the proof is technical and deferred to the appendix.

§11. Lower Bounds

Modified System S_1. Recall that S_0 has a unique derivation from the initial assertion w_0 (30) to the final assertion w_∞ (31). Let us modify S_0 to a new system S_1 so that it has, instead, a unique derivation of the form:

$$Q_0 \xrightarrow[*]{} w_\infty \xrightarrow[*]{} Q_\infty \tag{32}$$

where Q_0 and Q_∞ are new variables. Intuitively, the derivation from w_∞ to Q_∞ should be the reverse of the standard derivation. (So calling w_∞ the "final assertion" is slightly misleading in this setting, but we stick to this terminology.)

More precisely, we first expand Σ_0 to Σ_1 by adding three new variables

$$Q_0, Q_{\texttt{mid}}, Q_\infty.$$

We now treat Q_0 as the initial assertion of S_1. Next, we augment the rules of S_0 with three new rules.

Augmented Rules.

(S0)	Q_0	$\longrightarrow$	$A_0^d F_{1,n+1}[\texttt{inc}] = w_0$	("initialization rule")
(T0)	1	$\xrightarrow{B_0^d F_{1,n}[\texttt{dec}]F_n[\texttt{inc}]}$	$Q_{\texttt{mid}}$	("transition rule")
(F0)	$Q_{\texttt{mid}} A_0^d F_{1,n+1}[\texttt{inc}]$	$\longrightarrow$	Q_∞	("termination rule")

Intended Derivation. We intend to use these rules as follows. The initialization rule (S0) simply converts Q_0 to w_0 in one step. Then the standard derivation takes w_0 to w_∞. At this point, the only applicable rule that preserves simplicity is rule (T0), and this simply introduces the new variable $Q_{\texttt{mid}}$. At this point, the reverse standard derivation will take $Q_{\texttt{mid}}w_\infty$ back to $Q_{\texttt{mid}}w_0$. Now the final termination rule (F0) converts $Q_{\texttt{mid}}w_0$ to Q_∞. To summarize:

$$Q_0 \xrightarrow{(S0)} w_0 \xrightarrow[*]{std} w_\infty \xrightarrow{(T0)} Q_{\texttt{mid}}w_\infty \xrightarrow[*]{rstd} Q_{\texttt{mid}}w_0 \xrightarrow{(F0)} Q_\infty. \tag{33}$$

It is clear that this derivation has all the properties we previously asserted for S_0:

Lemma 24 *The derivation $Q_0 \xrightarrow[*]{} Q_\infty \pmod{S_1}$ is strongly unique. The initial assertion Q_0 is non-recursive.*

Modified System S_2. We need to make one more modification. The degree of the rules of S_1 can be as large as $n+d+2$. We now modify them so that their degree is at most $d+O(1)$. Introduce n new "level" variables

$$L_1, \ldots, L_n$$

and also

$$Q_{\texttt{init}}, Q_{\texttt{end}}.$$

We use them to simulate the rules of S_1. Thus the start rules (S1)-(S2) are now replaced by:

$$\begin{array}{llll}
(\text{T0})_k & L_k & \xrightarrow{F_k[\texttt{pass}]A_0} & L_{k+1} \qquad (k=1,\ldots,n-1)\\
(\text{T1})_k & L_kA_0 & \xrightarrow{F_k[\texttt{inc}]} & L_1B_0A_k^d \qquad (k=1,\ldots,n)\\
(\text{T2})_k & L_kA_0B_k^d & \xrightarrow{F_k[\texttt{dec}]} & L_1B_0 \qquad (k=1,\ldots,n).
\end{array}$$

Intuitively, we see that the start rules can be simulated by the new rules (T0)-(T2). Roughly speaking, rule $(\text{S1})_k$ is simulated by the sequence

$$L_1A_0 \xrightarrow[(\text{T0})_1]{F_1[\texttt{pass}]} L_2A_0 \xrightarrow[(\text{T0})_2]{} \cdots \xrightarrow[(\text{T0})_{k-1}]{} L_kA_0 \xrightarrow[(\text{T1})_k]{F_k[\texttt{inc}]} L_1B_0A_k^d.$$

The rule (S0) can be simulated using

$$\begin{array}{llll}
(\text{C1}) & Q_0 & \longrightarrow & Q_{\texttt{init}}L_1A_0^d\\
(\text{C2}) & L_k & \xrightarrow{Q_{\texttt{init}}} & L_{k+1}F_k[\texttt{inc}] \qquad (k=1,\ldots,n-1)\\
(\text{C3}) & Q_{\texttt{init}}L_n & \longrightarrow & L_1F_n[\texttt{inc}]
\end{array}$$

We leave it for the reader to find substitutes for the rules (F0)-(F3) and for (T0). These modifications should not modify the basic properties of the original system. The upshot is a modified system S_2 with $6n+7$ variables and whose rules have degree at most $d+3$.

Lower Bound on $I(n,d)$.

Lemma 25 *The ideal membership problem in $\mathbb{Q}[X_1,\ldots,X_n]$ has bound $I(n,d+3) \geq d^{2^p}$, $p \sim n/6$.*

Proof. Let D_∞ be the unique derivation of S_2 from Q_0 to Q_∞. Let $F = \{f_1,\ldots,f_m\}$ be the set of Thue polynomials corresponding to S_2 and let $f_0 = Q_0 - Q_\infty$. Then

$$\texttt{maxdeg}(F) = d+3, \qquad \deg(f_0) = 1, \qquad \texttt{maxdeg}(D_\infty) \geq e(n) + d + 3.$$

Suppose $f_0 = \sum_{i=1}^m a_i f_i$ ($a_i \in \mathbb{Q}[X_1,\ldots,X_n]$). By lemma 19, there is a derivation D' from Q_0 to Q_∞ such that $\texttt{maxdeg}(D') \leq \deg(a_i f_i)$ for all $i = 1,\ldots,m$. After omitting repetitions from D', we must get the unique derivation D_∞. Since $\texttt{maxdeg}(D_\infty) \leq \texttt{maxdeg}(D')$, we conclude that $\deg(a_i f_i) \geq \texttt{maxdeg}(D_\infty) \geq e(n) + d + 3$. But by definition of $I(n,d)$, we have

$$I(6n+7, d+3) + \deg(f_0) \geq \deg(a_i f_i) \geq e(n) + d + 3.$$

Hence $I(n,d+3) \geq e(p)$ where $p \sim n/6$. **Q.E.D.**

Assuming $n \geq \ln d$, we may simplify this bound to $I(n,d) \geq e(p)$ for some $p \sim n/6$.

Lower Bound on $S(n,d)$. First we show a property of Thue polynomials. If $f = \sum_{i=1}^k h_i$ where each $h_i \in R$ satisfies $\texttt{support}(h_i) \subseteq \texttt{support}(f)$ then we say that f is a *non-canceling sum* of $h_1,\ldots,h_k$, and write

$$f = \bigoplus_{i=1}^k h_i.$$

Lemma 26 *Let $f_1,\ldots,f_m$ be Thue polynomials and $f = \sum_{i=1}^m \alpha_i f_i$, ($\alpha_i \in R$) and $d_0 := \max\{\deg(\alpha_i f_i) : i = 1,\ldots,m\}$. Then $f = \bigoplus_{j=1}^k h_j$ ($k \geq 1$) where h_j are Thue polynomials that can be expressed as*

$$h_j = \sum_{i=1}^m \alpha_{j,i} f_i$$

with $\alpha_{i,j} \in R$ and $\deg(\alpha_{j,i} f_i) \leq d_0$.

Proof. First we rewrite f in the form

$$f = \sum_{i=1}^\ell \beta_i g_i$$

where each β_i is a monomial, each $g_i \in \{f_1,\ldots,f_m\}$ (possibly repeated) and $\deg(\beta_i g_i) \leq d_0$. We use induction on ℓ. If $\ell = 1$ then the result is immediate. Assume $\ell > 1$. Without loss of generality, assume that $\texttt{support}(\beta_1 g_1) \cap \texttt{support}(f)$ is non-empty. By induction, $f' := f - \beta_1 g_1$ can be written as a non-canceling sum,

$$f' = \bigoplus_{j=1}^k h_j$$

where h_j can be expressed as $h_j = \sum_{i=1}^m \alpha_{j,i} f_i$, $\deg(\alpha_{j,i} f_i) \leq d_0$.

There are two cases. Case 1: $\texttt{support}(\beta_1 g_1) \subseteq \texttt{support}(f)$. Then we can write $f = \bigoplus_{j=0}^k h_j$ where $h_0 = \beta_1 g_1$. This satisfies the lemma. Case 2: there is a word $w \in \texttt{support}(\beta_1 g_1) - \texttt{support}(f)$.

Now let c_j be the coefficient of w in h_j for $j = 1, \ldots, k$, and let c be the coefficient of w in f'. Since $w \in \texttt{support}(f')$, $c \neq 0$ and the sum of all the c_j's is equal to c. Now consider the expression

$$\begin{aligned} f &= f' + \beta_1 g_1 \\ &= \sum_{j=1}^{k} h_j + \sum_{j=1}^{k} \frac{c_j}{c} \beta_1 g_1 \\ &= \sum_{j=1}^{k} h'_j \end{aligned}$$

where $h'_j := h_j + (c_j/c)\beta_1 g_1$. One verifies: h'_j is Thue: this is obvious if $c_j = 0$, and otherwise, the coefficient of w in h'_j is zero because the coefficient of w is $\beta_1 g_1$ is $-c$. Hence $\texttt{support}(h'_j) \subseteq \texttt{support}(f)$, and h'_j can be expressed as a linear sum of the $f_1, \ldots, f_m$ with degree at most d_0. **Q.E.D.**

Lemma 27 *Let S be a Thue system with a strongly unique derivation $D_0 : w_0 \xrightarrow{*} w_\infty$ and w_0 is non-recursive. Let $F = \{f_1, \ldots, f_m\}$ be the set of Thue polynomials corresponding to S and $f_0 = w_\infty - w_0$. Then any syzygy basis for $S = \texttt{Syz}(f_0, f_1, \ldots, f_m)$ has degree at least $\texttt{maxdeg}(D_0) - \texttt{maxdeg}(F)$.*

Proof. Let $B \subseteq S$ be a basis for S where each element of B has degree at most d_0. Note that S contains a syzygy $(g_0, \ldots, g_m)$ where $g_0 = 1$. This implies that there is a basis element $(h_0, \ldots, h_m) \in B$ such that the constant term of h_0 is non-zero; without loss of generality, assume the constant term is 1. Now we have that

$$h_0 f_0 = -\sum_{i=1}^{m} h_i f_i.$$

By lemma 26, $h_0 f_0$ can be written as a non-canceling sum $h_0 f_0 = \bigoplus_{j=1}^{k} p_j$ where each p_j is Thue. Also we may assume that $p_1 = \sum_{i=1}^{m} \alpha_i f_i$ where $\deg(\alpha_i f_i) \leq \texttt{maxdeg}(\{h_1 f_1, \ldots, h_m f_m\}) \leq d_0 + \texttt{maxdeg}(F)$. Each monomial of $h_0 f_0$ is of the form $w_0\beta$ or $w_\infty\beta$ where $\pm\beta$ is a monomial of h_0. In particular, w_0 is a monomial of $h_0 f_0$. Without loss of generality, assume $-w_0$ is a monomial of p_1. If the other monomial in p_1 is of the form $w_0\beta$, then $p_1 = w_0\beta - w_0$. By lemma 19, there is a derivation from w_0 to $w_0\beta$, contradicting the non-recursiveness of w_0. Hence $p_1 = w_\infty\beta - w_0$. So there is a simple derivation D from w_0 to $w_\infty\beta$ such that $\texttt{maxdeg}(D) \leq d_0 + \texttt{maxdeg}(F)$. By the strong uniqueness of D_0, $\beta = 1$ and $D = D_0$. Hence $d_0 \geq \texttt{maxdeg}(D_0) - \texttt{maxdeg}(F)$. **Q.E.D.**

Applying this lemma to the system S_2 yields the lower $S(n, d) \geq d^{2^p}$, $p \sim n/6$. Finally, our lower bound for $G(n, d)$ follows from the fact (see [13]) that

$$G(n, d) \geq S(n, d).$$

Exercises

Exercise 11.1: Observe that each of the level variables (L_i's) and the flag variables ($F_i[\texttt{inc}], F_i[\texttt{dec}], F_i[\texttt{pass}]$) has degree at most one in assertions. Show how to replace all these variables by just $2n$ new variables, say G_i, H_i $(i = 1, \ldots, n)$. This improves the lower bounds to $\sim d^{2^p}$ where $p \sim n/4$. □

Exercise 11.2:

(i) Let $K \subseteq R$ be any set of polynomials and $I \subseteq R$ be an ideal. We say I is *of type* K if for all $f \in I$, there are polynomials $f_1, \ldots, f_m \in I \cap K$ $(m \geq 1)$ such that $f = \bigoplus_{i=1}^{m} f_i$. Show that an ideal is Thue iff it is generated by a finite subset of Thue polynomials.

(ii) The reduced Gröbner basis of a Thue ideal is Thue.

Remark: Monomial and homogeneous ideals exhibit similar properties. □

§A. APPENDIX: Properties of S_0

We prove two technical properties of the system S_0: the Basic Lemma (§10) and the non-recursiveness of w_0 (lemma 23).

Proof of Basic Lemma. The result is easy to check for $k = 1$. So assume $k > 1$.

(a.1)$_{\mathbf{k},\ell}$ *We shall be referring to the words u_i ($i = 1, \ldots, 8$) defined in the standard derivation in the proof of lemma 1(a).* By assumption, $F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] \mid w_1$, and the word u_1 in this lemma and u_1 in lemma 1(a) are identified by assuming

$$w_1 = wF_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}].$$

Since $B_0^d F_{1,k}[\texttt{dec}] \mid u_1'$, there must be a first word x_1 in the derivation D_1 in which $B_0^d F_{1,k-1}[\texttt{dec}] \mid x_1$. By the induction hypothesis (a.1)$_{k-1,k-1}$, x_1 has the form of u_2:

$$u_2 = w_1 B_0^d A_{k-1}^{e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{inc}].$$

So the prefix of D_1 that derives u_2 is standard and unique. In particular, the last rule applied to get u_2 is a forward start rule. Now two rules are applicable to u_2: (F1)$_{k-1}$ and some reverse start rule. But this reverse start rule is excluded since otherwise we get a non-simple derivation (it is easily checked: a forward start rule followed by a reverse start rule gives a non-simple derivation). Thus the word after u_2 in D_1 must be obtained by applying rule (F1)$_{k-1}$. This word is u_3. Since $B_0^d F_{1,k-1}[\texttt{dec}] \mid u_1'$ but not u_3, there is a first word x_2 after u_3 in D_1 such that $B_0^d F_{1,k-1}[\texttt{dec}] \mid x_2$. Also the rule applied to u_3 must be a forward one. By induction hypothesis (a.1)$_{k-1,\ell}$, we conclude that x_2 is the same as u_4.

Now $u_4 = x_3(0)$ where

$$x_3(i) := w_1 B_0^d A_{k-1}^{e(k-1)-i} B_{k-1}^i A_\ell^{(i+1)e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{pass}],$$

$i = 0, \ldots, e(k-1)$. Assume in general that $x_3(i)$ has just been derived by a forward start transition. In that case, if $i = e(k-1)$ then there are no rules applicable to $x_3(i)$, so assume $i < e(k-1)$. Then we could apply either rule (F2)$_{k-1}$ or rule (F3)$_{k-1}$ to $x_3(i)$.

Case 1: If we apply (F3) to $x_3(i)$ we get

$$x_4 := w_1 A_0^d A_{k-1}^{e(k-1)-i-1} B_{k-1}^{i+1} A_\ell^{(i+1)e(k-1)} F_{1,k-1}[\texttt{inc}]F_{k-1}[\texttt{dec}].$$

But then we may apply induction (a.2)$_{k-1,k-1}$ (since only a forward rule now applies to x_4) which implies $B_{k-1}^{e(k-1)} \mid x_4$. This implies $i = e(k-1) - 1$.

Case 2: If we apply (F2) to $x_3(i)$, we get

$$x_5 := w_1 A_0^d A_{k-1}^{e(k-1)-i-1} B_{k-1}^{i+1} A_\ell^{(i+1)e(k-1)} F_{1,k-1}[\texttt{inc}]F_{k-1}[\texttt{pass}].$$

If the next rule applied is a forward rule, then by induction (a.1)$_{k-1,\ell}$, we get $x_3(i+1)$. We claim that this is the only possibility. Suppose for the sake of contradiction that the rule applied to x_5 is a reverse rule. Then it must be reverse (F1)$_{k-1}$, giving us

$$x_6 := w_1 B_0^d A_{k-1}^{e(k-1)-i-1} B_{k-1}^{i+1} A_\ell^{(i+1)e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{inc}].$$

Observe that no forward rule is applicable to x_6. Since the flag $F_{k-1}[\texttt{inc}]$ must subsequently change, and this can only occur as a result of applying a finish rule $(\mathrm{F}m)_j$ or its reverse (for some $m = 1, 2, 3$ and $j \geq k-1$). This means that there is a first word x_7 after x_6 such that $A_0^d F_{1,k-1}[\texttt{inc}] \mid x_7$. By induction $(\mathrm{b}.1)_{k-1,k-1}$, this means $A_{k-1}^{e(k-1)}$ divides x_6. Clearly a contradiction.

We conclude from the analysis of these two cases that starting from $u_4 = x_3(0)$, we must repeatedly apply the sequence of rules,

$$[(\mathrm{F}2)_{k-1},\ \text{induction}\ (\mathrm{a}.1)_{k-1,\ell}]$$

for $e(k-1)-1$ times, yielding $x_3(e(k-1)-1)$; finally, we apply rule $(\mathrm{F}3)_{k-1}$, giving us u_8. From this, we can invoke induction $(\mathrm{a}.2)_{k-1,k-1}$ to get u_1', exactly as in the standard derivation. Since all the steps are forced, this is unique.

$(\mathbf{a.2})_{\mathbf{k},\ell}$ Similar to part (a.1).

$(\mathbf{b})_{\mathbf{k},\ell}$ We only prove case $(\mathrm{b}.1)_{k,\ell}$. So $F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}] \mid w_1'$, and we will try to show that D_2 is the reverse of the standard derivation in lemma 1. Instead of $u_1, \ldots, u_8$ of the proof in lemma 1, we define $v_1, \ldots, v_8$ by

$$v_i := u_i w_1' (w A_\ell^{e(k)} F_{k,\ell}[\texttt{pass}]F_\ell[\texttt{inc}])^{-1}.$$

For instance,

$$v_8 = w_1' A_0^d B_{k-1}^{e(k-1)} F_{1,k-1}[\texttt{inc}]F_{k-1}[\texttt{dec}].$$

It is not clear that the v_i's are well-defined words; they would all be clearly well-defined if

$$A_\ell^{e(k)} \mid w_1',$$

which we will show. But v_8, v_7 are well-defined in any case.

Starting from $u_1' = w_1' B_0^d F_{1,k}[\texttt{dec}]$, we may invoke induction $(\mathrm{b}.2)_{k-1,k-1}$ to get to v_8. Then reverse $(\mathrm{F}3)_{k-1}$ is forced and we get to

$$v_7 = w_1' B_0^d A_{k-1} B_{k-1}^{e(k-1)-1} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{pass}].$$

Now either rule $(\mathrm{F}2)_{k-1}$ or some reverse rule applies to v_7.

First assume that rule $(\mathrm{F}2)_{k-1}$ is applied to v_7, giving

$$x_8 := w_1' A_0^d B_{k-1}^{e(k-1)} F_{1,k-1}[\texttt{inc}]F_{k-1}[\texttt{pass}].$$

If a reverse rule were applied to x_8 then this is reverse $(\mathrm{F}1)_{k-1}$ giving us

$$x_9 := w_1' B_0^d B_{k-1}^{e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{inc}].$$

Now only a reverse rule is applicable to x_9, and we can apply induction $(\mathrm{b}.1)_{k-1,k-1}$ which implies $A_{k-1}^{e(k-1)}$ divides x_9, contradiction. So a forward rule is applied to x_8. Since $F_{k-1}[\texttt{pass}]$ in x_8 must change as some later point, we argue as before that $B_0^d F_{1,k-1}[\texttt{dec}]$ divides some subsequent word x_{10}. But then induction $(\mathrm{a}.1)_{k-1,\ell}$ implies that x_8 leads to

$$x_{11} := w_1' B_0^d B_{k-1}^{e(k-1)} A_\ell^{e(k-1)} F_{1,k-1}[\texttt{dec}]F_{k-1}[\texttt{pass}].$$

But now we are stuck.

Hence we can assume that some reverse rule is applied to v_7. Then by induction $(\text{b.1})_{k-1,\ell}$ we know that $A_\ell^{e(k-1)} \mid v_7$ and we arrive at the word

$$x_{12} := w_1' A_0^d A_{k-1} B_{k-1}^{e(k-1)-1} A_\ell^{-e(k-1)} F_{1,k-1}[\texttt{inc}] F_{k-1}[\texttt{pass}].$$

Note that $x_{12} = x_{13}(1)$ where

$$x_{13}(i) := w_1' A_0^d A_{k-1}^i B_{k-1}^{e(k-1)-i} A_\ell^{-i\cdot e(k-1)} F_{1,k-1}[\texttt{inc}] F_{k-1}[\texttt{pass}].$$

In general assume that $x_{13}(i)$ occurs in D_2. In particular, this means that $A_\ell^{i\cdot e(k-1)} \mid w_1'$ (i.e., $x_{13}(i)$ is well-defined). Further assume that $x_{13}(i)$ had been obtained by applying a reverse start rule (this is true of $x_{12} = x_{13}(1)$). So the only rules applicable to $x_{13}(i)$ are reverse $(\text{F1})_{k-1}$ or reverse $(\text{F2})_{k-1}$.

(i) If it is known that reverse $(\text{F2})_{k-1}$ is next applied to $x_{13}(i)$ then we can apply induction $(\text{b.1})_{k-1,\ell}$ to conclude that $x_{13}(i+1)$ occurs in D_2. But note that reverse (F2) cannot be applied to $x_{13}(e(k-1))$.

(ii) If reverse $(\text{F1})_{k-1}$ is applied to $x_{13}(i)$ then we get to

$$x_{14}(i) := w_1' B_0^d A_{k-1}^i B_{k-1}^{e(k-1)-i} A_\ell^{-i\cdot e(k-1)} F_{1,k-1}[\texttt{dec}] F_{k-1}[\texttt{inc}].$$

Now only a reverse rule applies to $x_{14}(i)$ and by induction $(\text{b.1})_{k-1,k-1}$, we see that $A_{k-1}^{e(k-1)} \mid x_{14}(i)$. This implies that $i = e(k-1)$. But $x_{14}(e(k-1))$ is v_2.

From (i) and (ii), we conclude: starting from $x_{13}(1)$ in D_2, we must repeatedly apply the sequence of reverse rules

$$[\text{reverse } (\text{F2})_{k-1}, \text{ induction } (\text{b.1})_{k-1,\ell}]$$

for $e(k-1)-1$ times, yielding $x_{13}(e(k-1))$. Finally we apply reverse (F2) to give v_2. From v_2, we invoke induction $(\text{b.1})_{k-1,k-1}$ to get to v_1. This proves the basic lemma.

Proof of the non-recursiveness of w_0. Under the assumption that we have a non-trivial simple derivation $D : w_0 \stackrel{*}{\to} \tilde{w}$ where $w_0 | \tilde{w}$, we force D to trace through the computation path of a standard derivation and derive a contradiction. We will rely heavily on the analysis of the proof of the Basic Lemma. Invocations of parts of the Basic Lemma are denoted "$\text{ind(a.1)}_{k,\ell}$", etc, for appropriate k, ℓ. *We also refer to the words $u_1, \ldots, u_8, x_1, \ldots, x_7$ used in the proof of the Basic Lemma. We assume that u_1 there is equal to w_0.*

We proceed as follows. Any derivation starting from w_0 begins with a forward rule. Suppose that $k = 1, \ldots, n$ is the smallest level such that D does not modify the flag at level k. It is clear that $k \geq 2$. Since the flag at level $k-1$ is initially $F_{k-1}[\texttt{inc}]$, and the only rule that changes $F_{k-1}[\texttt{inc}]$ is rule $(\text{F1})_{k-1}$, we conclude that D must have a first word y_1 such that $B_0^d F_{1,k-1}[\texttt{dec}] F_{k-1}[\texttt{inc}] \mid y_1$. By $\text{ind(a.1)}_{k-1,k-1}$, the prefix of D up to y_1 is a standard derivation and y_1 must be equal to u_2. After applying rule $(\text{F1})_{k-1}$, we get u_3. Since the flag at level $k-1$ of u_3 must subsequently change (since $F_{k-1}[\texttt{inc}] \mid \tilde{w}$), some finish rule at level $j \geq k-1$ must be applied. This means some subsequent word is divisible by $B_0^d F_{1,k-1}[\texttt{dec}]$. Only a forward rule is applicable to u_3. So we may invoke $\text{ind(a.1)}_{k-1,k}$ to get to u_4 (assuming $\ell = k$).

Now u_4 has the general form of $x_3(i)$. The two cases are again applicable: Case 1, where we apply (F3) to $x_3(i)$ to get x_4. Since the flag at level $k-1$ is not colored [inc], we can again invoke

ind(a.2)$_{k-1,k-1}$ which implies $i = e(k-1)-1$. Case 2, where we apply (F2) to $x_3(i)$ to get x_5. If we next apply a backward rule to x_5, it must be reverse (F1), yielding x_6. Now only a backward rule apply to x_6 and since we eventually must reach a word divisible by $A_0^d F_{1,k}[\texttt{inc}]$, we may invoke ind(b.1)$_{k-1,k-1}$ to get a contradiction. This means only a forward rule is applicable to $x_3(i)$ and after invoking ind(a.1)$_{k-1,k}$ we get to $x_3(i+1)$.

Again we conclude from cases 1 and 2 that we must eventually reach u_8. Only a forward rule applies to u_8. Since the flag of u_8 at level $k-1$ must eventually change, we can again invoke ind(b.1)$_{k-1,k-1}$ to yield u_1'. But note that from u_1' there is only rule (F1)$_k$ that applies. But this rule modifies the flag at level k, a contradiction. This concludes the proof.

References

[1] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[2] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[3] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[4] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[5] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[6] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[7] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[8] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[9] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[10] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[11] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[12] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[13] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[14] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[15] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

Contents

Lecture XIV
Continued Fractions

This venerable subject goes back to the Greeks: Archimedes (287–212 B.C.) and Theon of Smyrna (c. 70–135 A.D.) were suspected of using continued fractions to approximate square roots [67]. The first attempt at a general definition of continued fractions was due to Leonardo of Pisa better known as Fibonacci (c. 1170-1250). Pringsheim [162] states that Pietro Antonio Cataldi (1548-1626) is the inventor of continued fractions. Euler uses the term "continued fractions" (*fractio continua*) for the first time. Continued fractions originally arose in the solution of algebraic equations. They are also widely used in analysis under the "analytic theory of continued fractions". The classic in the subject is Oskar Perron's two-volume *Die Lehre von den Kettenbrüchen* [157], a considerable expansion of the original volume [155]. The classic exposition from Olds [150] is highly recommended. Other references include Wall [214], Khovanskii [101], Jones and Thron [94], and Lorentzen and Waadeland [120]. See Brezinski [28] for a comprehensive history on continued fractions and Padé approximation from antiquity until 1939.

Continued fractions are a particular representation of numbers with an implied iterative computational process. It furnishes us with an alternative constructive approach to the subject of computable real numbers (the more obvious approach takes a computable real number to be a computable sequence of bits, say). Lagrange's method (see §7) of approximating roots of polynomial equations using continued fractions is intimately related to Gaussian algorithm in the theory of reduced binary quadratic forms (Lecture VIII) [28, p. 185]. of polynomials, Gosper [18] is an early advocate of the merits of continued fraction arithmetic, and described algorithms for the basic arithmetic operations. Vuillemin [213] expanded on these ideas, advocating the use of redundant continued fraction representations. Zimmer [223] treats the use of continued fractions in representing algebraic numbers. See also Zippel [224, chapter 2].

§1. Introduction

What is a continued fraction? Intuitively, it is a sequence of approximations to a number. For instance, consider the fraction $\frac{343}{284} = 1.20774\ldots$. To the nearest integer, $\frac{343}{284} \sim 1$. More precisely,

$$\frac{343}{284} = 1 + \epsilon_1$$

where $0 \leq \epsilon_1 < 1$ is an error term. The term ϵ_1 is more precisely given by

$$\frac{1}{\epsilon_1} = \frac{284}{59} = 4 + \epsilon_2$$

where $0 \leq \epsilon_2 < 1$. The term ϵ_2 is really

$$\frac{1}{\epsilon_2} = \frac{59}{48} = 1 + \epsilon_3$$

where $0 \leq \epsilon_3 < 1$. Continuing in this spirit,

$$\frac{1}{\epsilon_3} = \frac{48}{11} = 4 + \epsilon_4, \quad \frac{1}{\epsilon_4} = \frac{11}{4} = 2 + \epsilon_5, \quad \frac{1}{\epsilon_5} = \frac{4}{3} = 1 + \epsilon_6, \quad \frac{1}{\epsilon_6} = 3.$$

Thus we see that

$$\frac{343}{284} = 1 + \frac{1}{4 + \frac{1}{1+\frac{1}{4+\frac{1}{2+\frac{1}{1+\frac{1}{3}}}}}}.$$

The expression on the right is called a continued fraction. The reader may notice that we have just carried out the Euclidean algorithm (§II.2) on the pair $343, 284$. This connection is no accident. The sequence $[1, 4, 1, 4, 2, 1, 3]$ of quotients obtained in the Euclidean algorithm can be regarded as a representation of $343/282$. Furthermore, each initial prefix of this sequence is an approximation — it will be shown that these approximations are the best possible in a suitable sense.

Let us next take an irrational number, $\sqrt{2} = 1.414213\ldots$. To the nearest integer, $\sqrt{2} \sim 1$. Well, it is closer to $\sqrt{2} \sim 1 + \frac{1}{2}$. But in fact,

$$1 + \frac{1}{2 + \frac{1}{2}}, \qquad 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}}, \qquad \cdots$$

are successively better (each improvement is obtained by replacing the rightmost occurrence of '2' with '$2 + \frac{1}{2}$'.) Apparently this process could be continued indefinitely. To see why, note that

$$\sqrt{2} - 1 = \frac{1}{\sqrt{2} + 1} = \frac{1}{2 + (\sqrt{2} - 1)}$$

which is a recursive equation in $\sqrt{2} - 1$. Thus the recursion can be expanded as often as we like. This example shows that a continued fraction is generally an infinite expression.

Our third motivation for continued fractions is to view them as approximate solutions to algebraic equations. Suppose x satisfies the equation $X^2 - 2X - 1 = 0$. Then we have

$$x = 2 + \frac{1}{x}.$$

We may further expand the x on the right-hand side,

$$x = 2 + \frac{1}{2 + \frac{1}{x}} = 2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{x}}} = \cdots$$

By ignoring the trailing x in these continued fractions as negligible, we get a succession of approximations to x:

$$2, \quad 2 + \frac{1}{2}, \quad 2 + \frac{1}{2 + \frac{1}{2}}, \quad 2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}}, \quad 2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}}}, \cdots$$

These approximations are

$$2, \quad \frac{5}{2} = 2.5, \quad \frac{12}{5} = 2.4, \quad \frac{29}{12} = 2.4166\ldots, \quad \frac{70}{29} = 2.4138\ldots, \cdots$$

The reader must have noticed the similarity between x and $\sqrt{2}$ above. Indeed, solving the quadratic equation for x we get $x = 1 \pm \sqrt{2}$. Thus we have obtained a sequence of approximations for $x = 1 + \sqrt{2}$. We will return to the subject of using continued fractions to solve algebraic equations.

An advantage of continued fractions is illustrated by the nice continued fraction representation of $\sqrt{2}$: many transcendental numbers such as e, π also have "regularly describable" continued fractions. So for some applications, it may be a more natural representation of numbers than, say, binary or decimal notations.

Exercises

Exercise 1.1: Use the continued fraction expansion approach to find a root of $X^3 - 2X^2 + 1$. Can you find the other roots by a similar approach? HINT: $X = 2 - \frac{1}{X^2} = \cdots$. □

Exercise 1.2: (Bombelli, 1572 [197, p.111]) Generalize the above continued fraction for $\sqrt{2}$: for any positive integer X,

$$\sqrt{X} = a + \frac{r}{2a + \frac{r}{2a + \frac{r}{2a + \cdots}}}.$$

where $a = \left\lfloor \sqrt{X} \right\rfloor$ and $r = X - a^2$. □

§2. Extended Numbers

We make a small but essential detour. The natural setting for discussing continued fractions involves infinity as a number. Let the set

$$\widehat{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$$

of *extended complex numbers* be the set $\mathbb{C}$ of complex numbers augmented with a single new point ∞. It has a compact topology, obtained by the "one-point compactification" of the usual topology of $\mathbb{C}$. When we restrict attention to the set $\mathbb{R} \subseteq \mathbb{C}$ of real numbers, we similarly have the one-point compactification,

$$\widehat{\mathbb{R}} := \mathbb{R} \cup \{\infty\},$$

the set of *extended real numbers*. An extended number is *finite* if it is not equal to ∞ (so finite numbers are just numbers in the usual sense). It is important to realize that we equate ∞ with $-\infty$, in the same way that $0 = -0$. It is convenient to further introduce a new symbol $\perp$ (called *bottom*) that is not a number: we call it the *indefinite* value. Hence we also call an extended number a *definite* value. With all this, the traditional injunction against dividing by zero is removed and given an algebraic expression:

$$0/0 = \perp, \qquad x/0 = \infty$$

where x is any non-zero extended number. Of course, all this begs the question of extending the arithmetic operations to $\perp$ as well as ∞. The rule for $\perp$ is easy: *if any operand equals $\perp$, the result of the arithmetic operation is $\perp$.* For example, $x + \perp = \perp + x = \perp$ for all x. The rule for ∞ is more involved, but could be deduced by using simple limiting arguments. For instance, suppose $\{a_i\}, \{b_i\}$ are monotonic sequences that tend to 0 and to ∞, respectively. Then the sequence $\{a_i + b_i\}$ tends to ∞. This gives the rule

$$0 + \infty = \infty.$$

If the monotonic sequence $\{c_i\}$ also tends to ∞, we see that $\{b_i + c_i\}$ could tend to any value or to nothing at all. Thus

$$\infty + \infty = \perp.$$

We have $\infty - \infty = \infty \div \infty = \perp$ by the same argument. Similarly, $\{a_i b_i\}$ is indefinite, yielding

$$\infty \times 0 = 0 \times \infty = \perp.$$

This exhausts all cases where the result is indefinite. A summary of these operations is given in the following tables.

$+$	0	y	∞
0	0	y	∞
x	x	$x+y$	∞
∞	∞	∞	$\perp$

$-$	0	y	∞
0	0	$-y$	∞
x	x	$x-y$	∞
∞	∞	∞	$\perp$

$\times$	0	y	∞
0	0	0	$\perp$
x	0	xy	∞
∞	$\perp$	∞	∞

$\div$	0	y	∞
0	$\perp$	0	0
x	∞	x/y	0
∞	∞	∞	$\perp$

In these tables, x and y are finite non-zero values. The table for *negation* $x \mapsto -x$ ought to be added to the above. But this table is rather trivial:

$$-x = x \quad \text{iff} \quad x \in \{0, \infty, \perp\}.$$

There is no standard name for this ring-like algebraic structure $\mathbb{C} \cup \{\infty, \perp\}$.

Chordal Metric. In order to avoid any special role for ∞ among the extended complex numbers, we introduce a natural metric on $\widehat{\mathbb{C}}$ via the stereographic projection $\sigma : \widehat{\mathbb{C}} \to S^2$ where $S^2 \subseteq \mathbb{R}^3$ is the unit sphere centered at the origin of an Euclidean 3-space and $\mathbb{C}$ is identified with the xy-plane. S^2 is called the *Riemann sphere*, with *North* and *South poles* given by $N = (0, 0, 1)$ and $S = (0, 0, -1)$, respectively. The complex number $p = p_x + \mathbf{i}p_y$ (by identification with the point $p = (p_x, p_y, 0)$ on the xy-plane) is mapped to the unique point $\sigma(p) \in S^2$ such that $N, p, \sigma(p)$ are collinear. We further define $\sigma(\infty) = N$. This map can be explicitly given by

$$\sigma(p) = \left(\frac{2p_x}{|p|^2+1}, \frac{2p_y}{|p|^2+1}, \frac{|p|^2-1}{|p|^2+1} \right), \qquad |p|^2 = p_x^2 + p_y^2.$$

The *chordal distance* $\Delta(p, q)$ between two points $p, q \in \widehat{\mathbb{C}}$ is equal to the Euclidean distance between $\sigma(p)$ and $\sigma(q)$. More explicitly,

$$\Delta(p,q) = \begin{cases} \frac{2|p-q|}{\sqrt{(1+|p|^2)(1+|q|^2)}} & \text{if } p, q \neq \infty, \\ \frac{2|p|}{\sqrt{1+|p|^2}} & \text{if } q = \infty. \end{cases}$$

Note that $\Delta(p, q) \leq 2$.

Our applications in continued fractions focus on the restriction (still called σ) of this map to the extended real numbers

$$\sigma : \widehat{\mathbb{R}} \to S^1 \tag{1}$$

where S^1 is intersection of S^2 with the xz-plane. *extended real numbers* We may also identify the xz-plane with $\mathbb{C}$ and call S^1 the "unit complex circle". Under this identification, the North and South Pole of S^1 are the complex numbers $\mathbf{i}$ and $-\mathbf{i}$. See figure 1.

Then $\sigma(1) = 1$, $\sigma(-1) = -1$, $\sigma(0) = -\mathbf{i}$ and $\sigma(\infty) = \mathbf{i}$. The inverse of σ is given by

$$\sigma^{-1}(e^{\mathbf{i}\theta}) = \frac{\cos\theta}{1 - \sin\theta}.$$

It is easy to see that $\Delta(0, 1) = \sqrt{2}$ and $\Delta(\sqrt{3}, \infty) = 1$. More generally, $\Delta(0, r) = 2|r|/\sqrt{1+r^2}$, $\Delta(\infty, r) = 2/\sqrt{1+r^2}$.

Exercises

Exercise 2.1:

(i) Verify the formulas for $\sigma(p)$ and $\Delta(p, q)$.

(ii) Verify that $\Delta(p, q)$ is a metric. □

§3. General Terminology

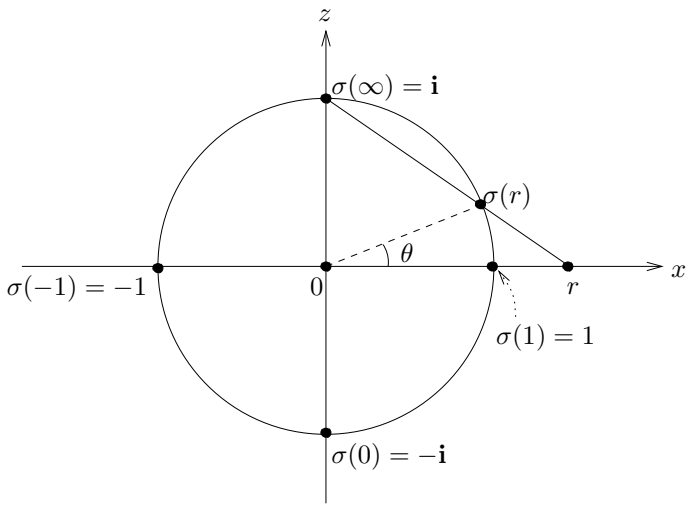


Figure 1: Stereographic projection: $r \mapsto \sigma(r) \in S^1$.

Formally, a *continued fraction* is a possibly infinite expression of the form

$$q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{q_2 + \cfrac{p_3}{q_3 + \cfrac{p_4}{q_4 + \cdots}}}} \tag{2}$$

where p_i, q_i are extended complex numbers or functions of complex variables. We call p_i, q_i the ith *partial numerator* and ith *partial denominator*; p_i/q_i is the ith *partial quotient*. We refer to the partial numerators and partial denominators as *terms* of the continued fraction. For simplicity, we normally assume each $q_i \neq 0$. To avoid the cumbersome form (2), we express the same continued fraction using a linear notation

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\frac{p_3}{q_3+}\cdots \tag{3}$$

or a summation-like notation

$$q_0 + \mathop{\mathbf{K}}_{i=1}^{B}\left(\frac{p_i}{q_i}\right)$$

where $B \geq 0$ is an integer or ∞. The latter form is useful for stating explicit expressions for the general i-th partial denominator and numerator. For example,

$$\begin{aligned} \ln(1+z) &= \frac{z}{1+}\frac{z/2}{1+}\frac{z/6}{1+}\frac{2z/6}{1+}\frac{2z/10}{1+}\frac{3z/10}{1+}\frac{3z/14}{1+}\cdots \\ &= \mathop{\mathbf{K}}_{i\geq 1}\left(\frac{a_i z}{1}\right) \end{aligned}$$

where $a_1 = 1$ and for $i \geq 1$, $a_{2i} = \frac{i}{2(2i-1)}$ and $a_{2i+1} = \frac{i}{2(2i+1)}$. This continued fraction converges for all $z \in \mathbb{C}$ except when $z \in (-\infty, -1]$, part of the negative x-axis. Many such continued fractions are known for common transcendental functions. See the Exercise for a continued fraction for e^z.

The continued fraction is called *terminating* or *non-terminating* according to whether or not it has finitely many terms.[1] The *length* of a terminating continued fraction is the largest n such that both the nth partial numerator and denominator are defined; the length is infinite in case of a non-terminating continued fraction.

[1]Terminating/non-terminating continued fractions are also called finite/infinite in the literature. We avoid this terminology since "finite" and "infinite" also applies to extended numbers.

We are mainly interested in the continued fraction (3) when all p_i, q_i are numbers: in which case we say the continued fraction is *numerical*; otherwise it is *functional*. A typical example of a functional continued fraction is where all the p_i, q_i depends on a single variable X. For instance, Euler shows that the power series $c_0 + c_1X + c_2X^2 + \cdots$ is "equivalent" to

$$c_0 + \frac{c_1X}{1-}\frac{\frac{c_2}{c_1}X}{1+\frac{c_2}{c_1}X-}\frac{\frac{c_3}{c_2}X}{1+\frac{c_3}{c_2}X-}\cdots\frac{\frac{c_n}{c_{n-1}}X}{1+\frac{c_n}{c_{n-1}}X-}\cdots$$

in a suitable sense (cf. below). Note that we write

$$\cdots\frac{p_i}{q_i-}\frac{p_{i+1}}{q_{i+1}+}\cdots$$

as an alternate form for

$$\cdots\frac{p_i}{q_i+}\frac{-p_{i+1}}{q_{i+1}+}\cdots$$

We now define the *value* of a continued fraction. This value is an element of $\hat{\mathbb{C}} \cup \{\perp\}$ when the continued function is numerical; otherwise the value is a function. We will not discuss the value of a functional continued fraction except to say that it is somewhat like the convergence of series, but more subtle (Exercise). When the continued fraction is numerical and terminating, the value is defined in an obvious fashion. For instance, the value of $1 + \frac{1}{2+}\frac{1}{2}$ is $7/5$. For the value of a non-terminating continued fraction, we proceed as follows: if $i \geq 0$ is at most the length of the continued fraction (3), we call the terminating continued fraction

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\cdots\frac{p_{i-1}}{q_{i-1}+}\frac{p_i}{q_i}$$

the ith *convergent* of the continued fraction (3). The value of the ith convergent is called the ith *quotient* or ith *approximant*. For example, the ith quotient of (3) for $i = 0, \ldots, 3$ are:

$$q_0, \quad \frac{p_1 + q_0q_1}{q_1}, \quad \frac{p_1q_2 + p_2q_0 + q_0q_1q_2}{p_1 + q_1q_2}, \tag{4}$$

$$\frac{p_1p_3 + p_1q_2q_3 + p_2q_0q_3 + p_3q_0q_1 + q_0q_1q_2q_3}{p_2q_3 + p_3q_1 + q_1q_2q_3}. \tag{5}$$

The ith quotient is clearly a fraction P_i/Q_i which could be infinite ($P_i \neq 0, Q_i = 0$) or indefinite ($P_i = Q_i = 0$); otherwise the ith quotient is *finite*. We define the *value* of a non-terminating continued fraction (3) to be the extended number r provided all but a finite number of its quotients are definite, and the sequence of definite quotients P_i/Q_i converges to r as $i \to \infty$. Note that r may be ∞. A basic problem in the theory of continued fraction is to study conditions on p_i, q_i to ensure convergence. We prove such a theorem in §6.

We now define the *(nominal) numerator* and *denominator* of the ith quotient P_i/Q_i. The definition should not be taken for granted since there can be accidental cancellations between the nominal numerators and denominators[2]. It is best to view them as polynomials in the partial numerators and denominators p_j, q_j $(j \leq i)$,

$$P_i = P_i(q_0, p_1, q_1, \ldots, p_i, q_i), \qquad Q_i = Q_i(p_1, q_1, \ldots, p_i, q_i). \tag{6}$$

We call P_i and Q_i the ith *numerator* and ith *denominator*, respectively. They are defined as follows, using the compact matrix notation:

$$\begin{bmatrix} P_{-1} \\ Q_{-1} \end{bmatrix} := \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{7}$$

[2]It turns out the systematic cancellation of common factors in the numerator and denominator cannot happen: the polynomials P_i, Q_i that we define are absolutely irreducible in the indeterminates $q_0, p_1, q_1, \ldots, p_i, q_i$ (see [67]).

$$\begin{bmatrix} P_0 \\ Q_0 \end{bmatrix} := \begin{bmatrix} q_0 \\ 1 \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} P_i \\ Q_i \end{bmatrix} := \begin{bmatrix} P_{i-2} & P_{i-1} \\ Q_{i-2} & Q_{i-1} \end{bmatrix} \begin{bmatrix} p_i \\ q_i \end{bmatrix}, \quad i \geq 1 \tag{9}$$

We must show that this definition is valid, i.e., P_i/Q_i gives the expected ith quotient. Clearly (9) is valid for $i = 1$ (we have concocted P_{-1}, Q_{-1} for this purpose). Assume inductively that P_i/Q_i gives the correct ith quotient. From (9) we get

$$\frac{P_i}{Q_i} = \frac{p_i P_{i-2} + q_i P_{i-1}}{p_i Q_{i-2} + q_i Q_{i-1}}. \tag{10}$$

We observe that P_{i+1}/Q_{i+1} is obtained from P_i/Q_i by replacing q_i by $q_i + \frac{p_{i+1}}{q_{i+1}}$. Substituting in (10) we get

$$\begin{aligned} \frac{P_{i+1}}{Q_{i+1}} &= \frac{p_i P_{i-2} + \left(q_i + \frac{p_{i+1}}{q_{i+1}}\right) P_{i-1}}{p_i Q_{i-2} + \left(q_i + \frac{p_{i+1}}{q_{i+1}}\right) Q_{i-1}} \\ &= \frac{q_{i+1}\left(p_i P_{i-2} + q_i P_{i-1}\right) + p_{i+1} P_{i-1}}{q_{i+1}\left(p_i Q_{i-2} + q_i Q_{i-1}\right) + p_{i+1} Q_{i-1}} \\ &= \frac{q_{i+1} P_i + p_{i+1} P_{i-1}}{q_{i+1} Q_i + p_{i+1} Q_{i-1}} \end{aligned}$$

which proves that (9) is correct for $i + 1$.

Complements and Tails. The remaining terms of a continued fraction after we have extracted the ith convergent is

$$c_i = \underset{j \geq i+1}{\mathbf{K}} \left(\frac{p_j}{q_j}\right), \tag{11}$$

which we will call the ith *complement*. Its value is called the *complementary quotient*. In general, a continued fraction (3) where the 0-th term vanishes ($q_0 = 0$) is said to be in "complementary form".

Sometimes we prefer to use an alternative decomposition of a continued fraction. The following continued fractions which comprise a prefix and a suffix of (3),

$$h_i = q_0 + \frac{p_1}{q_1 +} \cdots \frac{p_{i-1}}{q_{i-1} +} \frac{p_i}{1}, \qquad t_i = q_i + \underset{j \geq i+1}{\mathbf{K}} \left(\frac{p_j}{q_j}\right)$$

are called (respectively) the ith *head* and *tail* of (3). The value (if it exists) of t_i is called the ith *tail quotient*.

Exercises

Exercise 3.1: Show that the ith quotients of

$$1 - \frac{1}{2-} \frac{1}{2-} \frac{1}{2-} \cdots$$

are $\frac{1}{i+1}$ for all $i \geq 0$. Hence the value is 0. □

Exercise 3.2: (Lorentzen-Waadeland) The value of $\mathbf{K}_{i=1}^{\infty}\left(\frac{6}{1}\right)$ is 2. HINT: the ith approximant is $-6\frac{(-3)^i - 2^i}{(-3)^{i+1} - 2^{i+1}}$. □

Exercise 3.3: Consider the continued fraction

$$f(Z) = \frac{Z}{1-}\frac{Z}{Z+1-}\cdots = \mathbf{K}\left(\frac{Z}{1-Z}\right).$$

(i) $f(Z)$ converges to Z if $|Z| < 1$.
(ii) $f(Z)$ converges to -1 if $|Z| > 1$. □

Exercise 3.4: Show that

$$(1+X)^r = 1 + \frac{rX}{1-}\frac{\frac{r-1}{2}X}{1+\frac{r-1}{2}X-}\frac{\frac{r-2}{3}X}{1+\frac{r-2}{3}X-}\cdots$$

where the continued fraction terminates (with length r) if and only if r is a non-negative integer. □

Exercise 3.5: (Khovanskii)
(i) For any a, we have the functional continued fraction

$$\sqrt{X} = a + \frac{X-a^2}{2a+}\frac{X-a^2}{2a+}\cdots.$$

(ii) The above continued fraction is convergent for all $X \geq 0$.
(iii) Show

$$\begin{aligned}\sqrt{Y^2+Z^2} &= Y + \frac{Z^2}{2Y+}\frac{Z^2}{2Y+}\frac{Z^2}{2Y+}\cdots \\ &= Y + \frac{Z^2}{Y+Z+}\frac{Y^2}{Y+Z+}\frac{Z^2}{Y+Z+}\frac{Y^2}{Y+Z+}\cdots\end{aligned}$$

□

Exercise 3.6: Consider a continued fraction of the form

$$F(Z) = \frac{1}{b_1+Z-}\frac{a_1^2}{b_2+Z-}\frac{a_2^2}{b_3+Z-}\cdots.$$

(i) Show that the nth denominator is given by

$$B_n(Z) = \det\begin{bmatrix} b_1+Z & -a_1 & & & & \\ -a_1 & b_2+Z & -a_2 & & & \\ & -a_2 & b_3+Z & & & \\ & & & \ddots & & \\ & & & & b_{n-1}+Z & -a_{n-1} \\ & & & & -a_{n-1} & b_n+Z \end{bmatrix}.$$

NOTE: let $B_{-1}(Z) = 0, B_0(Z) = 1$ and $a_0 = 1$.
(ii) If the a_i, b_i are real then $B_n(Z)$ has all real roots. □

§4. Ordinary Continued Fractions

Two continued fractions with the same value are said to be *equivalent*.[3] It is important to study *equivalent transformations* of continued fractions, i.e., transformations that preserve values. The following is a simple example of an equivalent transformation of the continued fraction (3):

$$q_0 + \frac{c_1p_1}{c_1q_1+}\frac{c_1c_2p_2}{c_2q_2+}\frac{c_2c_3p_3}{c_3q_3+} + \cdots = q_0 + \mathop{\mathbf{K}}_{i=1}^{\infty} \frac{c_{i-1}c_ip_i}{c_iq_i} \tag{12}$$

for non-zero c_i's ($c_0 = 1$). It is clear that the ith convergents of (12) and (3) are equal for all i. In general, when this condition holds between two continued fractions, we say that they are *term-wise equivalent*.

It is not hard to see that with suitable choice of c_i's, every continued fraction (3) is term-wise equivalent to one whose partial numerators are 1:

$$q_0 + \frac{1}{q_1+}\frac{1}{q_2+}\frac{1}{q_3+}\cdots. \tag{13}$$

We call such[4] a continued fraction (13) *ordinary*, and its value is given the compact notation

$$[q_0, q_1, q_2, \ldots] = [q_i]_{i=0}^{\infty}. \tag{14}$$

As examples we have $\sqrt{2} = [1, 2, 2, 2, \ldots]$,

$\infty = [r, 0]$ (for any finite number r). The ith quotient of (14) is $[q_0, q_1, q_2, \ldots, q_i]$; the ith tail and ith complementary quotients are $[q_i, q_{i+1}, q_{i+2} \ldots]$ and $[0, q_i, q_{i+1}, q_{i+2} \ldots]$ (respectively). Moreover,

$$[q_0, q_1, q_2 \ldots] = [q_0, q_1, \ldots, q_{i-1}, [q_i, q_{i+1}, \ldots]\,].$$

If the partial denominators q_i ($i \geq 1$) are positive integers in (13) and q_0 is an integer, we call[5] it a *regular continued fraction*.

Empty Ordinary Continued Fraction. It is convenient to introduce the *empty ordinary continued fraction*, denoted by the symbol []. We define the value of [] to be ∞ and its length to be $-\infty$. If n is the length of the continued fraction, we define its nth complement to be [] and so the nth complementary quotient is ∞.

Periodic Continued Fractions. If for some $h \geq 1$ and $n \geq 0$ we have

$$q_i = q_{i+h}$$

whenever $i \geq n$, then we say (14) is *periodic*. This can be indicated either by a bar over the periodic part or by a semicolon separating the periodic part from a suffix:

$$[q_0, q_1, \ldots, q_{n-1}, \overline{q_n, q_{n+1}, \ldots, q_{n+h-1}}] \quad \text{or} \quad [q_0, q_1, \ldots, q_{n-1}; q_n, q_{n+1}, \ldots, q_{n+h-1}]. \tag{15}$$

If h is chosen as small as possible, then h is called the *period* of the ordinary continued fraction. For instance $\sqrt{2} = [1, 2, 2, \overline{2, 2, 2}] = [1, \overline{2}] = [1; 2]$ has period 1. The classic theorem about periodic continued fractions is due to Lagrange: *the value of a real regular continued fraction is an irrational quadratic number if and only if the continued fraction is periodic.* An irrational quadratic number has the form $a + b\sqrt{d}$ where a, b are rational numbers, $b \neq 0$ and $d > 1$ is squarefree. (One direction of this result is easy – see Exercise.) Unfortunately there is no known extension of Lagrange's characterization to higher degree algebraic numbers.

[3] The more usual definition of "equivalence", following L. Seidel (1855), corresponds to what we call term-wise equivalence.

[4] Lagrange concludes from this that there is no interest in the more general form of continued fractions; this is unwarranted for many reasons.

[5] Usually called *simple continued fraction*, but we follow Perron [155] who uses the equivalent German term *regelmäßig*.

The Continued Fraction Algorithm. It is clear that for any finite real number r we can define a regular continued fraction $\texttt{RCF}(r)$ as follows: if r is an integer, then $\texttt{RCF}(r) = r$. Otherwise,

$$\texttt{RCF}(r) := \lfloor r \rfloor + \frac{1}{\texttt{RCF}(1/(r - \lfloor r \rfloor))}. \tag{16}$$

This can be viewed as a process to transform a number into a sequence of integers and is sometimes called the "continued fraction algorithm".

Is the representation of finite reals by regular continued fractions unique? The answer is no because of this easily seen identity:

$$[q_0, \ldots, q_{n-1}, q_n, 1] = [q_0, \ldots, q_{n-1}, q_n + 1], \quad n \geq 1$$

This identity implies that any terminating regular continued fraction is equivalent to one whose last partial quotient q_n is at least 2 whenever its length n is positive. It turns out that, *with this sole exception*, we have uniqueness:

Lemma 1
(i) There is a bijective correspondence between the finite real numbers and regular continued fractions, provided we restrict the last partial quotient of a terminating continued fraction of positive length to be at least 2.
(ii) Under the correspondence of part (i), a finite real number is irrational if and only its regular continued fraction is non-terminating.

We leave the proof as an Exercise. The distinction between a continued fraction and its value is often confused. The preceding lemma justifies the practice in case of regular continued fractions. We sometimes perpetrate the same language abusive by referring to the value "$[q_0, q_1, q_2, \ldots]$" as a continued fraction.

Simple Operations on Ordinary Continued Fractions. For reciprocals, we have

$$\frac{1}{[q_0, q_1, \ldots]} = \begin{cases} [0, q_0, q_1, \ldots], & q_0 > 0 \\ [q_1, q_2, q_3, \ldots], & q_0 = 0. \end{cases} \tag{17}$$

For negation, we have $-[q_0, q_1, q_2, \ldots] = [-q_0, -q_1, -q_2, \ldots]$ or more generally:

$$-\left[q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\cdots\right] = \left[-q_0 - \frac{p_1}{q_1-}\frac{p_2}{q_2-}\cdots\right]. \tag{18}$$

Alternatively, we can proceed as follows. It is easy to verify that

$$[x, y] = [x - 1, 1, -(y + 1)] \quad \text{or} \quad [x + 1, y - 1] = [x, 1, -y]. \tag{19}$$

Applying this identity,

$$\begin{aligned} [-q_0, -q_1, -q_2, \ldots] &= [-q_0, -[q_1, q_2, \ldots]] \\ &= [-q_0 - 1, 1, [q_1, q_2, \ldots] - 1] \\ &= [-q_0 - 1, 1, [q_1 - 1, q_2, \ldots]]. \end{aligned}$$

This yields a negation formula that manipulates only a finite part of the original continued fraction.

$$-[q_0, q_1, q_2, q_3, \ldots] = [-q_0 - 1, 1, q_1 - 1, q_2, q_3, \ldots]. \tag{20}$$

We can absorb partial quotients that vanish

$$[\ldots, q_{n-1}, 0, q_{n+1}, \ldots] = [\ldots, q_{n-1} + q_{n+1}, \ldots], \qquad n \geq 1. \tag{21}$$

Using the identity (19), partial quotients of 1 can also be absorbed.

Finally, we defer the treatment of the arithmetic operations to a later section.

Exercises

Exercise 4.1:
(i) Is the continued fraction

$$\frac{q_0 q_1 + p_1}{q_1+}\frac{p_2}{q_2+}\frac{p_3}{q_3+}\cdots$$

equivalent to the one in (3)?
(ii) What are the numerators and denominators of the continued fraction (12)? □

Exercise 4.2:
(i) What is the number $[\overline{1}] = [1, 1, 1, \ldots]$? Again: $[\overline{0}]$, $[\overline{1,2}]$ and $[\overline{1,2,3}]$?
(ii) Give upper and lower bounds for $[1, 2, 3, 4, \ldots] = \mathbf{K}_{i=1}^{\infty}\left(\frac{1}{i}\right)$. Is this an algebraic number? □

Exercise 4.3:
(i) The continued fraction solution to the equation $X^2 - 3X - 1 = 0$ is $X = [3, 3, 3, \ldots] = [\overline{3}]$. Use this to hand-compute arbitrarily good approximations to $\sqrt{13}$. HINT: what is the relation of X to $\sqrt{13}$?
(ii) Show that $\sqrt{13} = [3, \overline{1, 1, 1, 1, 6}]$. □

Exercise 4.4: (Euler)
(i) For any complex numbers a, b, if $X = a - \frac{b}{2} + \sqrt{1 + \frac{b^2}{4}}$ then

$$X - a = \frac{1}{b+}\frac{1}{b+}\frac{1}{b+}\cdots.$$

(ii) Show that $\sqrt{1 + a^2} = [a, 2a, 2a, \ldots] = [a, \overline{2a}]$. (This gives us the continued fractions of $\sqrt{2}, \sqrt{5}, \sqrt{10}$, etc.) □

Exercise 4.5: Let $C = q_0' + \frac{p_1'}{q_1'+}\frac{p_2'}{q_2'+}\cdots$ be the continued fraction equivalent to (3) such that the ith quotient P_i'/Q_i' of C is the $(2i)$th quotient of (3) for all $i \geq 0$. Determine p_i', q_i'. □

Exercise 4.6: Show the easy direction of Lagrange's result characterizing irrational quadratic numbers, namely, if a real regular continued fraction is periodic, then its value is irrational quadratic. HINT: first assume the purely periodic case. [The other direction of Lagrange's result is an exercise in §7.] A regular continued fraction is *purely periodic* if it has the form $[\overline{q_1, \ldots, q_n}]$; the aperiodic part is empty. □

Exercise 4.7: A real quadratic irrationality $X = a + b\sqrt{d}$ is *reduced* if $X > 1$ and its conjugate $X' = a - b\sqrt{d}$ satisfies $-1 < X' < 0$.
(i) Show that the regular continued fraction of a real quadratic irrationality is purely periodic if and only if it is reduced. See previous exercise for the definition of purely periodic.
(ii) If $\alpha = [\overline{q_1, q_2, \ldots, q_n}]$ then the continued fraction of $-1/\alpha$ is $[\overline{q_n, q_{n-1}, \ldots, q_1}]$.
REMARK: see §8 for the general concept of a "reduced real irrationality". □

Exercise 4.8: Show that

$$\sqrt{3} = 1 + \frac{2}{2+}\frac{2}{2+}\frac{2}{2+}\cdots.$$

Apply the equivalent transformation (12) to convert this into the regular continued fraction $\sqrt{3} = [1, \overline{1, 2}]$. □

Exercise 4.9:
(i) Show that $\sqrt{8} = [2, \overline{1, 4}]$. More generally, $\sqrt{4(1 + a^2)} = [2a, \overline{a, 4a}]$.
(ii) Investigate the continued fractions of period 2: let X satisfy

$$X - a = \frac{1}{b+}\frac{1}{c+}\frac{1}{b+}\frac{1}{c+}\cdots = \frac{1}{b+}\frac{1}{c + X - a}.$$

(iii) Conclude that $\sqrt{a(a+1)} = [a, \overline{2, 2a}]$. (This gives the continued fraction of $\sqrt{3}, \sqrt{6}, \sqrt{12}$, etc.)
(d) Generalize this to $\sqrt{a^2 + b^2}$. □

Exercise 4.10: Let $x = \frac{1}{2}(\sqrt{5} - 1)$.
(i) Show that $x = [0, \overline{1}]$.
(ii) If P_i/Q_i is the ith convergent to x, prove that

$$\left|\frac{P_i}{Q_i} - x\right| < \frac{1}{Q_i^2\sqrt{5}}.$$

(iii) Prove that the constant $\sqrt{5}$ in the right-hand side cannot be improved for this choice of x.
REMARK: for every irrational x there are infinitely many choices of p, q such that $|\frac{p}{q} - x| < (q^2\sqrt{5})^{-1}$. □

Exercise 4.11: (Euler) Consider the connection between power series and continued fractions:
(i) The power series $c_0 + c_1 + c_2 + \cdots$ ($c_i \neq 0$ for $i \geq 1$) and the continued fraction

$$c_0 + \frac{c_1}{1-}\frac{c_2/c_1}{1 + (c_2/c_1)-}\frac{c_3/c_2}{1 + (c_3/c_2)-}\cdots \quad = \quad \mathop{\mathbf{K}}_{i\geq 1}\left(\frac{-c_i/c_{i-1}}{1 + (c_i/c_{i-1})}\right)$$

have the same value.
(ii) Show that for all real or complex x,

$$\begin{aligned} e^x &= 1 + \frac{x}{1-}\frac{x/2}{1 + x/2-}\frac{x/3}{1 + x/3-}\cdots \\ &= 1 + \frac{x}{1-}\frac{1x}{2 + x-}\frac{2x}{3 + x-}\frac{3x}{4 + x-}\frac{4x}{5 + x-}\cdots \end{aligned}$$

HINT: put $c_i = 1/i!$ in part (i). □

§5. Continued fractions as Möbius transformations

We now view continued fractions as transformations on the extended complex numbers $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.

A *Möbius transformation* is a function $f : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ given by

$$f(z) = \frac{az + b}{cz + d}$$

where a, b, c, d are finite complex numbers.[6] It is easy to check that $f(z)$ is a constant function iff $ad - bc = 0$. The constant function is $f(z) = b/d = a/c$. We call $ad - bc$ the *determinant* of $f(z)$. Another special case is when $f(z) = z$ (the identity function). This happens iff $b = c = 0$ and $a = d$. *To avoid special cases, we henceforth restrict f to be neither a constant function nor the identity function.* Since the function is unchanged if we multiply all of a, b, c, d by a common nonzero constant, we further assume

$$ad - bc = 1.$$

The reader can check that $f(\infty) = a/c$ and $f(-d/c) = \infty$. This function is analytic everywhere except for a pole at $z = -d/c$. The function is injective since

$$f(z) - f(y) = \frac{(ad - bc)(z - y)}{(cz + d)(cy + d)}$$

is zero if and only if $z = y$. It is surjective since its inverse function

$$z = \frac{d \cdot f(z) - b}{-c \cdot f(z) + a}$$

is also a Möbius transformation. We leave as an exercise to show that Möbius transformations map circles onto circles (straight lines are special cases of circles that pass through ∞).

Matrix Representation. Basic manipulations and properties of continued fractions are easier to "see" when stated in the language of matrices. Each $x \in \widehat{\mathbb{C}}$ is non-uniquely *represented* in *homogeneous coordinates* by the 2-vector of the form

$$\begin{bmatrix} cx \\ c \end{bmatrix},$$

for each choice $c \in \mathbb{C} - \{0\}$. Conversely, any vector

$$\begin{bmatrix} x \\ y \end{bmatrix}$$

with $x, y \in \widehat{\mathbb{C}}$ *represents* the value $x/y \in \widehat{\mathbb{C}} \cup \{\perp\}$. If either $x = y = 0$ or $x = y = \infty$, then $x/y = \perp$ and the vector is said to be *indefinite*. If the 2-vectors $\mathbf{u}, \mathbf{v}$ represent the same value, we say they are *proportional* and we write

$$\mathbf{u} :: \mathbf{v}.$$

In matrix notation, the function $f(z) = (az + b)/(cz + d)$ is represented by a 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and function application translates into multiplying a 2-vector by such a matrix: if $z = x/y$ then

$$\begin{bmatrix} f(z) \\ 1 \end{bmatrix} :: \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}.$$

The composition of two Möbius transformation becomes multiplication of two such matrices.

[6] Such transformations are also called *linear fractional transformations* or *homographic functions*.

Classification. We call $z^* \in \widehat{\mathbb{C}}$ a *fix-point* of $f(z)$ if $f(z^*) = z^*$. It is easy to see that $Z = z^*$ is a solution to the following equation:

$$cZ^2 + (d-a)Z - b = 0.$$

Since $f(z)$ is not the identity function, this equation does not identically vanish — this implies that there are at most two fixed points. In case $c = 0, d = a$, we must have $b \neq 0$, so there are no fixed points. This corresponds to the translation $f(z) = z + (b/d)$. We can classify Möbius transformations by their actions relative to fixed points. More precisely, their iterated action $f^{(n)}(z)$ (n-fold application of f, as $n \to \infty$) can either move any point $z \in \widehat{\mathbb{C}}$ towards a fixed point or away from it. For simplicity, assume $c \neq 0$ so that f has two fixed points z_1, z_2 (not necessarily distinct). The following occurs:

- (Parabolic) There is only one distinct fixed point z_1. Then $f^{(n)}(z) \to z_1$ as $n \to \infty$ for all $z \in \widehat{\mathbb{C}}$.
- (Elliptic) We have $z_1 \neq z_2$ and $|cz_1 + d| = |cz_2 + d|$. Then $f^{(n)}(z)$ diverges for all $z \neq z_1, z_2$.
- (Loxodromic) We have $z_1 \neq z_2$ and $|cz_1 + d| > |cz_2 + d|$. Then $f^{(n)}(z) \to z_1$ as $n \to \infty$ for all $z \neq z_2$.

Continued Fractions as Transformations. The pair p_i, q_i of partial numerator and denominator in the continued fraction

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\frac{p_3}{q_3+}\cdots \tag{22}$$

can be viewed as the Möbius transformation

$$x \mapsto \frac{p_i}{q_i + x}. \tag{23}$$

In terms of matrix transformations, (23) becomes:

$$\begin{bmatrix} x \\ 1 \end{bmatrix} \mapsto \begin{bmatrix} 0 & p_i \\ 1 & q_i \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} = \begin{bmatrix} p_i \\ q_i + x \end{bmatrix}.$$

Hence we call a matrix of the form

$$\begin{bmatrix} 0 & p \\ 1 & q \end{bmatrix}$$

a *partial quotient matrix*. Similarly, the partial denominator q_0 corresponds to the matrix transformation

$$\begin{bmatrix} x \\ 1 \end{bmatrix} \mapsto \begin{bmatrix} 1 & q_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} = \begin{bmatrix} q_0 + x \\ 1 \end{bmatrix}.$$

$$T(q) := \begin{bmatrix} 1 & q \\ 0 & 1 \end{bmatrix}.$$

is called a *translation matrix*.

The ith quotient ($i \geq 0$) of the continued fraction (22) corresponds to the matrix product

$$M_i := \begin{bmatrix} 1 & q_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & p_1 \\ 1 & q_1 \end{bmatrix} \begin{bmatrix} 0 & p_2 \\ 1 & q_2 \end{bmatrix} \cdots \begin{bmatrix} 0 & p_i \\ 1 & q_i \end{bmatrix}. \tag{24}$$

Call M_i the ith *convergent matrix* of the continued fraction. If the continued fraction (22) has finite length n, then M_n is *the transformation matrix associated with* the continued fraction; if the length

is infinite, the transformation matrix will be an infinite product of matrices, viewed formally. It is immediate that

$$\det M_i = (-1)^i p_1 p_2 \cdots p_i. \tag{25}$$

The ith convergent matrix can be expressed in terms of the the jth numerators P_j and denominators Q_j (6). It is easy to verify (cf. (7–9)) that

$$M_i = \begin{bmatrix} P_{i-1} & P_i \\ Q_{i-1} & Q_i \end{bmatrix}. \tag{26}$$

Hence, we have the so-called *determinant formula*

$$P_{i-1}Q_i - P_iQ_{i-1} = (-1)^i p_0 p_1 p_2 \cdots p_i \qquad (p_0 = 1). \tag{27}$$

This matrix is associated to the Möbius transformation

$$f_i(z) = \frac{zP_i + P_{i-1}}{zQ_i + Q_{i-1}}. \tag{28}$$

Notice that this is the value of the continued fraction

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\cdots\frac{p_i}{q_i+}\frac{1}{z}.$$

That is, we append $(1/z)$ as the $(i+1)$st partial quotient. The reason we use $1/z$ instead of z is because this form is convenient to specialize to ordinary continued fractions where $p_{i+1} = 1$. More explicitly, for ordinary continued fraction $[q_0, q_1, q_2, \ldots, , q_i]$ we have

$$f_i(z) = [q_0, q_1, q_2, \ldots, , q_i, z] = \frac{zP_i + P_{i-1}}{zQ_i + Q_{i-1}}.$$

We shall call $f_i(z)$ in equation (28) the *z-value* of the continued fraction

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\cdots\frac{p_i}{q_i}.$$

Our original definition of the "value" of a continued fraction corresponds to $z = \infty$. Thus, 0-value is P_{i-1}/Q_{i-1} and the ∞-value is the usual value P_i/Q_i. Suppose x_i denotes the $(i-1)$st complementary quotient of the continued fraction (3), and let x_0 denote the value of the continued fraction. Then $x_0 = q_0 + x_1 = q_0 + p_1/(q_1 + x_2)$. In general,

$$\begin{bmatrix} x_0 \\ 1 \end{bmatrix} = \begin{bmatrix} P_{i-1} & P_i \\ Q_{i-1} & Q_i \end{bmatrix} \begin{bmatrix} x_i \\ 1 \end{bmatrix}. \tag{29}$$

Inverting the matrix, we obtain:

$$\begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} P_{i-1} & P_i \\ Q_{i-1} & Q_i \end{bmatrix}^{-1} \begin{bmatrix} x_0 \\ 1 \end{bmatrix} = \frac{(-1)^i}{p_1 p_2 \cdots p_i} \begin{bmatrix} Q_i & -P_i \\ -Q_{i-1} & P_{i-1} \end{bmatrix} \begin{bmatrix} x_0 \\ 1 \end{bmatrix}.$$

Recurrence for Tail Quotients. We carry out a similar analysis for tail quotients. First, let

$$t_i := q_i + \frac{p_{i+1}}{q_{i+1}+}\frac{p_{i+2}}{q_{i+2}+}\cdots$$

denote the ith tail quotient of (3). Then we have $t_i = q_i + \frac{p_{i+1}}{t_{i+1}}$ for $i \geq 0$. By telescopy,

$$\begin{aligned} \begin{bmatrix} t_0 \\ 1 \end{bmatrix} &= \begin{bmatrix} q_0 & p_1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} t_1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} q_0 & p_1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} q_1 & p_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} t_2 \\ 1 \end{bmatrix} \\ &\vdots \\ &= \begin{bmatrix} q_0 & p_1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} q_1 & p_2 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} q_{n-1} & p_n \\ 1 & 0 \end{bmatrix} \begin{bmatrix} t_n \\ 1 \end{bmatrix}. \end{aligned}$$

Expressing this in analogy to (29):

$$\begin{bmatrix} t_0 \\ 1 \end{bmatrix} = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} t_n \\ 1 \end{bmatrix} \tag{30}$$

where A_n, B_n, C_n, D_n are the analogues of P_n, Q_n, satisfying the recurrence relations:

$$\begin{aligned} \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} &= \begin{bmatrix} q_0 & p_1 \\ 1 & 0 \end{bmatrix}, \\ \begin{bmatrix} A_{n+1} & B_{n+1} \\ C_{n+1} & D_{n+1} \end{bmatrix} &= \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} q_n & p_{n+1} \\ 1 & 0 \end{bmatrix}, \\ &= \begin{bmatrix} q_n A_n + B_n & p_{n+1} A_n \\ q_n C_n + D_n & p_{n+1} C_n \end{bmatrix}. \end{aligned} \tag{31}$$

Notice that the recurrence for complementary quotients requires only 2 recurrence sequences (for P_i, Q_i) instead of 4 (for A_i, B_i, C_i, D_i). However, if we assume ordinary continued fractions ($p_i = 1$) then we have

$$B_i = A_{i-1}, \qquad D_i = C_{i-1}, \tag{32}$$

with $A_0 = 1$ and $C_0 = 0$. But A_i, B_i are now related to the numerators and denominators:

$$A_{i+1} = P_i, \qquad C_{i+1} = Q_i, \qquad (i \geq 0). \tag{33}$$

Inverting the matrices,

$$\begin{aligned} \begin{bmatrix} t_n \\ 1 \end{bmatrix} &= \left(\prod_{i=1}^{n} \frac{1}{p_i} \right) \begin{bmatrix} 0 & p_n \\ 1 & -q_{n-1} \end{bmatrix} \cdots \begin{bmatrix} 0 & p_2 \\ 1 & -q_1 \end{bmatrix} \begin{bmatrix} 0 & p_1 \\ 1 & -q_0 \end{bmatrix} \begin{bmatrix} t_0 \\ 1 \end{bmatrix} \\ &= \frac{(-1)^i}{p_1 p_2 \cdots p_i} \begin{bmatrix} D_n & -B_n \\ -C_n & A_n \end{bmatrix} \begin{bmatrix} t_0 \\ 1 \end{bmatrix}. \end{aligned}$$

Beardon [15] gives a general account of Möbius transformations (in any dimension).

EXERCISES

Exercise 5.1: Show that Möbius transformations map circles in the complex plane into circles.
(i) Show that the equation of a circle is

$$Az\overline{z} + Bz + \overline{B}\overline{z} + C = 0$$

for complex constants A, B, C with A, C real. Also, the circle is a straight line iff $A = 0$.
(ii) Show that if $z = f(w) = (aw + b)/(cw + d)$, then the w satisfies another equation of the form

$$A'w\overline{w} + B'w + \overline{B'}\overline{w} + C' = 0$$

for some other A', B', C' with A', C' real. □

Exercise 5.2: Verify the three-fold classification of Möbius transformations and associated properties. □

Exercise 5.3: Two numbers x, y are *equivalent* if they are related by an integer unimodular transformation: $y = \frac{ax+b}{cx+d}$ where a, b, c, d are integers and $ad - bc = \pm 1$. Check that this is a mathematical equivalence.
(i) Any two rational numbers are equivalent.
(ii) If x, y are irrational then they are equivalent iff their regular continued fractions share a common suffix: $x = [a_0, a_1, \ldots]$ and $y = [b_0, b_1, \ldots]$ implies there exists k, ℓ such that for all $i \geq 0$, $a_{i+k} = b_{i+\ell}$.
(iii) Partition the following numbers according to their equivalence classes:

$$\sqrt{2}, \quad \sqrt{3}, \quad \sqrt{5}, \quad (1+\sqrt{5})/2, \quad 1+\sqrt{2}, \quad 1+\sqrt{3}.$$

□

Exercise 5.4: Consider the following matrix representation of a continued fraction of $a\sqrt{2}$ (for any a), corresponding to $a \cdot [1, \overline{2}]$

$$\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} \cdots.$$

Show that if the ith convergent matrix $M_i = \begin{bmatrix} a_i & a_{i+1} \\ b_i & b_{i+1} \end{bmatrix}$ then

$$a_i = \frac{a}{2}\left((1+\sqrt{2})^i + (1-\sqrt{2})^i\right), \quad b_i = \frac{1}{2\sqrt{2}}\left((1+\sqrt{2})^i - (1-\sqrt{2})^i\right).$$

□

§6. Convergence Properties

We investigate the ith numerator P_i and denominator Q_i. From (7–9), we get

$$\begin{bmatrix} P_0 \\ Q_0 \end{bmatrix} = \begin{bmatrix} q_0 \\ 1 \end{bmatrix}, \tag{34}$$

$$\begin{bmatrix} P_1 \\ Q_1 \end{bmatrix} = \begin{bmatrix} p_1 + q_0 q_1 \\ q_1 \end{bmatrix}, \tag{35}$$

$$\begin{bmatrix} P_2 \\ Q_2 \end{bmatrix} = \begin{bmatrix} p_1 q_2 + q_0 p_2 + q_0 q_1 q_2 \\ p_1 + q_1 q_2 \end{bmatrix}, \tag{36}$$

$$\begin{bmatrix} P_3 \\ Q_3 \end{bmatrix} = \begin{bmatrix} p_1 p_3 + p_1 q_2 q_3 + q_0 p_2 q_3 + q_0 q_1 p_3 + q_0 q_1 q_2 q_3 \\ p_2 q_3 + q_1 p_3 + q_1 q_2 q_3 \end{bmatrix}. \tag{37}$$

The polynomials P_i, Q_i are instances of an infinite family of polynomials $K_1, K_2, \ldots$ called *continuant polynomials*. We give a simple rule (cf. Knuth [105, p. 340]) to describe all the terms of P_i:

Lemma 2 *Each term in P_n is obtained from the initial term $q_0 q_1 \cdots q_n$ by substituting zero or more non-overlapping pairs $q_{i-1} q_i$ of consecutive variables by p_i. Moreover, every term obtained in this way appears in P_n*

For instance, $q_0 p_2 q_3 q_4$, $q_0 p_2 p_4$ and $p_1 q_2 p_4$ are terms in P_4. The proof is by a simple induction: the result is trivially true for $P_0 = q_0$ and $P_1 = p_1 + q_0 q_1$. Inductively, the rule follows from the recurrence:

$$P_n = p_n P_{n-2} + q_n P_{n-1}.$$

An analogous rule for the terms in Q_n holds: we only have to use $q_1 q_2 \cdots q_n$ as initial term in the preceding lemma.

Define the polynomial K_n in $2n+1$ variables such that

$$K_n \left(\begin{array}{ccccc} & p_1 & p_2 & \cdots & p_n \\ q_0 & q_1 & q_2 & \cdots & q_n \end{array} \right) := P_n(q_0, p_1, \ldots, q_n).$$

Here we have arranged, following Muir (ca. 1874), the variables of K_n in a suggestive matrix form. With this notation, we see that

$$K_n \left(\begin{array}{ccccc} & p_2 & p_3 & \cdots & p_n \\ q_1 & q_2 & q_3 & \cdots & q_n \end{array} \right) = Q_n(q_1, p_2, \ldots, q_n).$$

The continuant polynomials can also be written in a determinantal form (Exercise).

The difference between two consecutive quotients is easy to determine:

$$\frac{P_{i-1}}{Q_{i-1}} - \frac{P_i}{Q_i} = \frac{\det M_i}{Q_{i-1}Q_i} = \frac{(-1)^i p_1 p_2 \cdots p_i}{Q_{i-1}Q_i}. \tag{38}$$

In case of ordinary continued fractions, $p_j = 1$ for all j, and so

$$\frac{P_{i-1}}{Q_{i-1}} - \frac{P_i}{Q_i} = \frac{(-1)^i}{Q_{i-1}Q_i}.$$

If we telescope the difference (38) from $i = 1$ to n we obtain

$$\begin{aligned} \frac{P_0}{Q_0} - \frac{P_n}{Q_n} &= \frac{-p_1}{Q_0 Q_1} + \frac{p_1 p_2}{Q_1 Q_2} + \cdots + \frac{(-1)^n p_1 p_2 \cdots p_n}{Q_{n-1}Q_n}, \\ \frac{P_n}{Q_n} &= q_0 + \frac{p_1}{q_1} - \frac{p_1 p_2}{Q_1 Q_2} + \cdots + \frac{(-1)^{n+1} p_1 p_2 \cdots p_n}{Q_{n-1}Q_n}. \end{aligned}$$

Next observe that the difference between the nth and $n-2$nd quotients is

$$\begin{aligned} \frac{P_n}{Q_n} - \frac{P_{n-2}}{Q_{n-2}} &= (-1)^n \frac{p_1 \cdots p_{n-1}}{Q_{n-2}Q_{n-1}} + (-1)^{n+1} \frac{p_1 \cdots p_n}{Q_{n-1}Q_n} \\ &= (-1)^n \frac{p_1 \cdots p_{n-1}}{Q_{n-1}} \left(\frac{1}{Q_{n-2}} - \frac{p_n}{Q_n} \right) \\ &= (-1)^n \frac{p_1 \cdots p_{n-1} q_n}{Q_{n-2}Q_n}. \end{aligned} \tag{39}$$

These lead to:

Theorem 3 *Assume the partial numerators p_i and denominators q_i are positive for $i \geq 1$ in the continued fraction (3).*
(i) The sequence of quotients with even indices

$$\frac{P_0}{Q_0}, \frac{P_2}{Q_2}, \frac{P_4}{Q_4}, \cdots$$

is increasing but bounded above by $q_0 + \frac{p_1}{q_1}$, and hence has a limiting value K_e.
(ii) The sequence of quotients with odd indices

$$\frac{P_1}{Q_1}, \frac{P_3}{Q_3}, \frac{P_5}{Q_5}, \cdots$$

is decreasing but bounded below by q_0, and hence has a limiting value K_o.
(iii) $K_e \leq K_o$.
(iv) (Seidel-Stern) In case each $p_i = 1$, then the divergence of the series $\sum_{i=1}^{\infty} q_i$ is a necessary and sufficient condition for $K_e = K_o$ (= value of the continued fraction).

Proof. From equation (38), we have

$$\frac{P_{2i}}{Q_{2i}} < \frac{P_{2i-1}}{Q_{2i-1}}.$$

The inequality of part (iii) follows from this, assuming that K_e and K_o exist. From equation (39), we see that the sequence of even quotients is strictly increasing, the sequence of odd quotients is strictly decreasing. It follows that each P_i/Q_i ($i \geq 2$) is greater than q_0: this is immediate if i is even; otherwise we have $P_i/Q_i > P_{i-1}/Q_{i-1} > q_0$. Likewise, P_i/Q_i is smaller than $q_0 + \frac{p_1}{q_1}$. This proves parts (i) and (ii).

It remains to prove (iv). First assume that the series $\sum_{i=1}^{\infty} q_i$ is divergent and we will show that $K_e = K_o$. First observe that $Q_1 = q_1, Q_2 = q_1 q_2 + 1$ and

$$Q_{2n} = q_{2n}Q_{2n-1} + Q_{2n-2}, \quad Q_{2n+1} = q_{2n+1}Q_{2n} + Q_{2n-1} \qquad (n \geq 1)$$

imply $Q_{2n} \geq 1$ and $Q_{2n+1} \geq q_1$ for all $n \geq 1$. From this we get that

$$\begin{aligned} Q_{2n} &\geq q_{2n}q_1 + Q_{2n-2} \\ &\geq \cdots \\ &\geq q_{2n}q_1 + q_{2n-2}q_1 + \cdots + q_4 q_1 + q_2 q_1. \end{aligned}$$

Similarly

$$\begin{aligned} Q_{2n+1} &\geq q_{2n+1} + Q_{2n} \\ &\geq \cdots \\ &\geq q_{2n+1} + q_{2n-1} + \cdots + q_3 + q_1. \end{aligned}$$

Thus, the divergence of $\sum_{i=1}^{\infty} q_i$ implies the divergence of either the odd Q-sequence $Q_1, Q_3, Q_5, \ldots$ or the even Q-sequence $Q_2, Q_2, Q_6, \ldots$. Hence the sequence $Q_1Q_2, Q_3Q_4, Q_5Q_6, \ldots$ diverges. The fact that $K_e = K_o$ then follows from (38).

Conversely, if the series $\sum_{i=1}^{\infty} q_i$ is convergent, we show that $K_e < K_o$. First we show that the even P-sequence

$$P_0, P_2, P_4, \ldots, P_{2n}, \ldots$$

and the odd P-sequence

$$P_1, P_3, P_5, \ldots, P_{2n+1}, \ldots$$

both converge to some values P_* and P^*, respectively. It is not hard to show by induction from $P_n = q_n P_{n-1} + P_{n-2}$ that

$$P_n < \prod_{i=1}^{n}(1 + q_i)$$

and hence

$$P_n < \exp(\sum_{i=1}^{n} q_i) < c_0$$

for some constant c_0. Next, by recursively expanding the term P_{n-2} in $P_n = q_n P_{n-1} + P_{n-2}$, we get

$$P_n = q_n P_{n-1} + q_{n-2}P_{n-3} + q_{n-4}P_{n-5} + \cdots.$$

Hence

$$\begin{aligned} P_{n+2m} - P_n &= q_{n+2m}P_{n+2m-1} + q_{n+2m-2}P_{n+2m-3} + \cdots + q_{n+2}P_{n+1} \\ &< c_0(q_{n+2m} + q_{n+2m-2} + \cdots + q_{n+2}). \end{aligned}$$

The convergence criterion of Bolzano-Cauchy for the series $\sum_{i=1}^{\infty} q_i$ says that the sequence is convergent if and only if for any $\epsilon > 0$ there is a N such that for all $m \geq n \geq N$, $q_n + q_{n+1} + \cdots + q_m < \epsilon$. So for any $\epsilon > 0$, for sufficiently large n, the above derivation implies $P_{n+2m} - P_n < \epsilon$. Another application of the Bolzano-Cauchy criterion shows that the even and the odd P-sequences each converge. A similar argument shows that the even and odd Q-sequences converge to some values Q_* and Q^* respectively. Using the fact that $P_{2n+1}Q_{2n} - P_{2n}Q_{2n+1} = 1$, we conclude that in the limit

$$P^*Q_* - P_*Q^* = 1.$$

But $K_e = P_*/Q_*$ and $K_o = P^*/Q^*$. This shows $K_o - K_e = (Q_*Q^*)^{-1}$ or $K_e < K_o$ as desired.

Q.E.D.

Another classic convergence theorem (from Śleszyński-Pringsheim) is this: if $|q_i| \geq 1 + |p_i|$ for all i then $\mathbf{K}_{i\geq 1}(p_i/q_i)$ converges to a value of modulus ≤ 1.

Approximation of Irrationals by Rationals. The preceding shows that the value K (if it exists) of the continued fraction (3) is equal to the series

$$K = q_0 + \sum_{i\geq 1} (-1)^{i+1} \frac{p_1 \cdots p_i}{Q_{i-1}Q_i}.$$

Moreover, for each $n \geq 2$, it satisfies

$$\left| K - \frac{P_n}{Q_n} \right| < \left| \frac{p_1 \cdots p_n}{Q_{n-1}Q_n} \right|.$$

In particular, the nth approximant of the regular continued fraction of any K satisfies

$$\left| K - \frac{P_n}{Q_n} \right| \leq \frac{1}{Q_{n-1}Q_n} < \frac{1}{c_0 Q_n^2} \tag{40}$$

where $c_0 = 1$. It is not hard to show that for any two consecutive quotients of K, one of them satisfies the inequality with $c_0 = 2$. This turns out to be a defining characteristic of quotients: whenever any fraction p/q has the property $|K - p/q| < 1/(2q^2)$, then p/q is a quotient of the regular continued fraction of K. Borel showed that of any three consecutive quotients, at least one satisfies the above inequality with $c_0 = \sqrt{5}$. Hence *every irrational number K has infinitely many approximations P_n/Q_n satisfying the above inequality with $c_0 = \sqrt{5}$*. This is the best possible in the sense that the statement fails if any larger value of c_0 is used. We also obtain a characterization of irrational numbers:

Corollary 4 *A number α is irrational iff there are infinitely many pairs of relatively prime integers p, q which satisfy the inequality $\alpha - (p/q) < 1/q^2$.*

Proof. If α is a rational a/b then $|\alpha - (p/q)| \geq 1/(bq) \geq 1/q^2$ for all but finitely many q. But if α is irrational then the nth approximant p_n/q_n of its ordinary continued fraction satisfies the desired inequality, following (40). To see that there are infinitely many distinct solutions among these

approximants, note that for any p_n, q_n, we may choose m large enough so that $|\alpha-(p_n/q_n)| > 1/(q_m^2)$. Then p_m, q_m must be a new solution. **Q.E.D.**

Exercises

Exercise 6.1: Prove that the following problem can be solved in polynomial time in the bit-sizes of the integer $N > 1$ and of the rational number α: given N, α, find a rational number p/q such that $|q\alpha - p|$ is minimum, subject to $q \le N$. □

Exercise 6.2: Show that

$$P_n = \det \begin{pmatrix} q_0 & -1 & & & & & & 0 \\ p_1 & q_1 & -1 & & & & & 0 \\ 0 & p_2 & q_2 & -1 & & & & 0 \\ 0 & & p_3 & q_3 & -1 & & & 0 \\ & & & & & \cdots & & \\ 0 & & & & & p_{n-1} & q_{n-1} & -1 \\ 0 & & & & & & p_n & q_n \end{pmatrix}.$$

HINT:

$$\begin{array}{cccccccc} -P_0 & & & & & & = & -q_0, \\ q_1P_0 & -P_1 & & & & & = & -p_1, \\ p_2P_0 & q_2P_1 & -P_2 & & & & = & 0, \\ & p_3P_1 & q_3P_2 & -P_3 & & & = & 0, \\ & & & \cdots & & & & \\ & & & p_{n-1}P_{n-3} & q_{n-1}P_{n-2} & -P_{n-1} & = & 0, \\ & & & & p_nP_{n-2} & q_nP_{n-1} \quad -P_n & = & 0. \end{array}$$

(b) Derive a similar determinantal form for Q_n.
(c) Conclude that

$$Q_n = \frac{\partial P_n}{\partial q_0}$$

and

$$K_n \begin{pmatrix} & p_1 & p_2 & \cdots & p_n \\ q_0 & q_1 & q_2 & \cdots & q_n \end{pmatrix} = K_n \begin{pmatrix} & p_n & p_{n-1} & \cdots & p_1 \\ q_n & q_{n-1} & q_{n-2} & \cdots & q_0 \end{pmatrix}.$$

□

Exercise 6.3: (Gosper) Give a method to compute the rational number with minimum denominator that lies within an interval $[a, b]$ with rational endpoints. □

§7. Real Möbius Transformations

We consider the method of Lagrange and Vincent for computing the regular continued fraction of a real algebraic number. For modern updates on this method see [203] and Cantor, Galyean and Zimmer [38]. Thull reports that his implementation of this approach is faster than Newton's method by a factor of 3 to 4.

For this and the next section, assume α_0 is a real root of a real polynomial $A_0(X)$ of degree $m \geq 2$. We may assume that $A_0(X)$ has only simple non-integer roots, and we are given an isolating interval $[r_0, s_0]$ with rational endpoints for α_0. Our goal is to compute the regular continued fraction $[a_0, a_1, a_2, \ldots]$ of α_0.

The basic method is simple to understand. By binary search, we can compute $a_0 = \lfloor \alpha_0 \rfloor$, *viz.*, a_0 is the unique integer n between $\lfloor r_0 \rfloor$ and $\lfloor s_0 \rfloor$ such that $A_0(n)A_0(n+1) < 0$. Then we could output a_0 as the zeroth partial denominator q_0. We then transform α_0 to $\alpha_1 = \frac{1}{\alpha_0 - a_0} = [a_1, a_2, a_3, \ldots]$. Notice that α_1 is the root of the transformed polynomial

$$A_1(X) = X^m A_0(a_0 + \frac{1}{X}).$$

The "key question" for continuing this process is to find a simple rule for a new isolating interval $[r_1, s_1]$ of α_1 relative to $A_1(X)$. If so, we can repeat this process indefinitely: find $a_1 = \lfloor \alpha_1 \rfloor$, transform $A_1(X)$ to $A_2(X)$ and find an isolating interval for $\alpha_2 := \frac{1}{\alpha_1 - a_2}$, etc. We refer to this generic process as the "continued fraction algorithm for (real) roots".

Transformations of the Complex Plane. To answer the "key question", we study the effects of the map

$$T_0(z) = a_0 + (1/z) \tag{41}$$

on the complex plane, corresponding to the transformation from α_0 to α_1 after we output the term a_0. In general, if the regular continued fraction of α_0 is $[a_0, a_1, \ldots]$, then after outputing the first $n+1$ partial denominators, we have the Möbius transformation (see equation (30))

$$T_n : z \mapsto [a_0, a_1, \ldots, a_n, z] = \frac{zA_{n+1} + A_n}{zC_{n+1} + C_n} = \frac{zP_n + P_{n-1}}{zQ_n + Q_{n-1}}$$

where A_i, C_i, P_i, Q_i are given by equations (31), (32) and (33) (with a_i's in place of q_i's). Note that P_n, Q_n are positive integers such that for $n \geq 2$,

$$P_n \geq \max\{P_{n-1}, Q_n\} \geq \min\{P_{n-1}, Q_n\} \geq Q_{n-1} \geq 0.$$

The first and last inequality is strict for $n \geq 3$. More generally, consider the Möbius transformation $T : z \mapsto w = \frac{az+b}{cz+d}$, or in matrix form

$$T : \begin{bmatrix} z \\ 1 \end{bmatrix} \mapsto \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} w \\ 1 \end{bmatrix}, \tag{42}$$

where $a \geq \max\{b, c\} \geq \min\{b, c\} \geq d \geq 0$ are non-negative with determinant

$$\Delta = ad - bc \neq 0.$$

We said (§5) that circles (in the z-plane) are transformed into circles (in the w-plane) by such transformations. Since T is real, we have $T(\overline{z}) = \overline{T(z)}$ ($\overline{z}$ denotes complex conjugation). This means that the real line is invariant under T and T has reflection symmetry about the real line. A simple calculation for the following 6 points shows

$$T(0) = \frac{b}{d}, \quad T(1) = \frac{a+b}{c+d}, \quad T(\infty) = \frac{a}{c}, \quad T(-d/c) = \infty, \quad T(-b/a) = 0, \quad T\left(-\frac{b+d}{a+c}\right) = -1.$$

See figure 2 for a representation of this mapping.

Let I_T be the interval on the real-axis of the w-plane with end-points a/c and b/d. It follows from the above calculations that T maps the imaginary axis of the z-plane to the circle K_T where K_T is

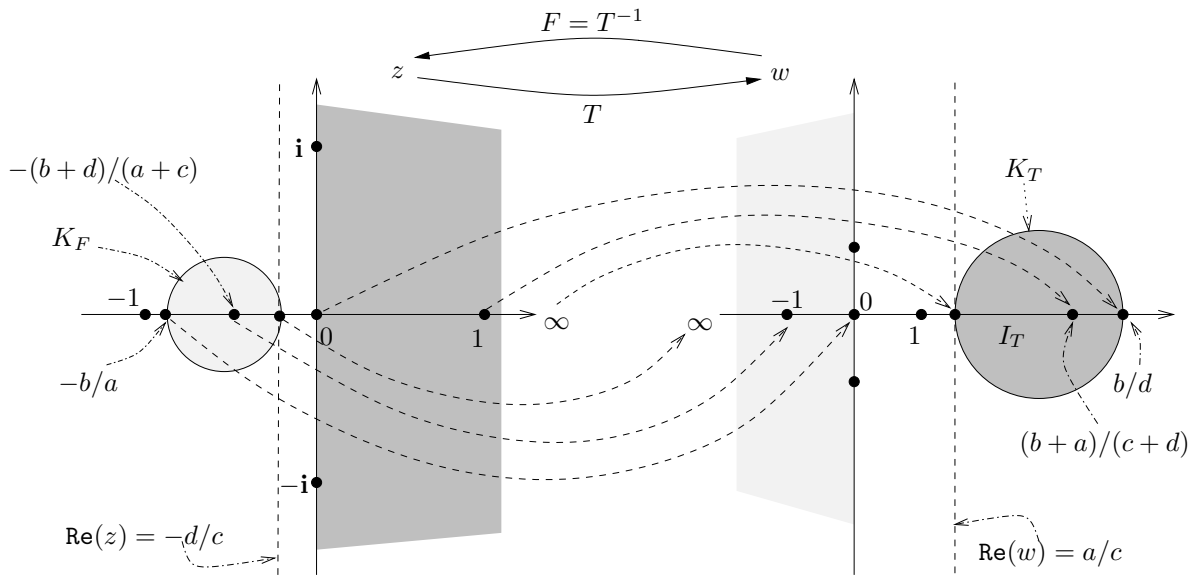


Figure 2: Transformation $T : z \mapsto w = (az + b)/(cz + d)$. Case of $\Delta < 0$.

the circle with diameter I_T. Note that $\frac{a}{c} - \frac{b}{d} = \frac{\Delta}{cd}$. So $a/c > b/d$ iff $\Delta > 0$. Similarly $-b/a > -d/c$ iff $\Delta > 0$. Figure 2 illustrates the transformation for $\Delta < 0$. Moreover, since the distinction between the inside and outside of a circle is preserved by Möbius transformations, and $T(1)$ lies inside K_T while $T(-d/c)$ lies outside K_T, we conclude that the entire half-plane $\mathtt{Re}(z) > 0$ is mapped to the open disc inside K_T, and the positive real axis of the z-plane to I_T.

Similarly, let F denote the inverse of T, and I_F the interval on the real-axis of the z-plane with endpoints $-b/a$ and $-d/c$, and K_F be the circle with diameter I_F. Then F maps K_F to the imaginary axis of the w-plane and I_F to the positive real axis of the w-plane.

Another notable feature is that the line $\mathtt{Re}(z) = -d/c$ is mapped into the line $\mathtt{Re}(w) = a/c$. Summarizing these observations:

Lemma 5 *The transformation T in (42) from the z-plane to the w-plane has these properties:*
(i) The half-plane $\mathtt{Re}(z) > 0$ is mapped to the open disc inside K_T, and $\mathtt{Re}(z) = 0$ is mapped to the circle K_T.
(ii) The circle K_F is mapped to the negative real axis in the w-plane and its interior is mapped to the half-plane $\mathtt{Re}(w) < 0$.
(iii) The half-plane $\mathtt{Re}(z) > -d/c$ is mapped to the half-plane $\mathtt{Re}(w) > a/c$, and the line $\mathtt{Re}(z) = -d/c$ becomes the line $\mathtt{Re}(w) = a/c$.

Now specialize our considerations to the case $T = T_0$ as given by (41). The associated matrix is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Let us assume $a_0 > 0$. The circle K_T turns out to be the line $\mathtt{Re}(w) = a_0$, and hence the half-plane $\mathtt{Re}(z) > 0$ is simply translated to the half-plane $\mathtt{Re}(w) > a_0$. In fact, parts (i) and (iii) in the lemma

become identical. We can also verify that the half-plane $\mathtt{Re}(z) > 1$ becomes the interior of the circle K_{a_0} with diameter $[a_0, a_0 + 1]$. Furthermore, the real intervals $[0, 1]$ and $[\infty, -1/a_0]$ in the z-plane transforms (respectively) into the real intervals $[a_0 + 1, \infty]$ and $[0, a_0]$ in the w-plane. See figure 3.

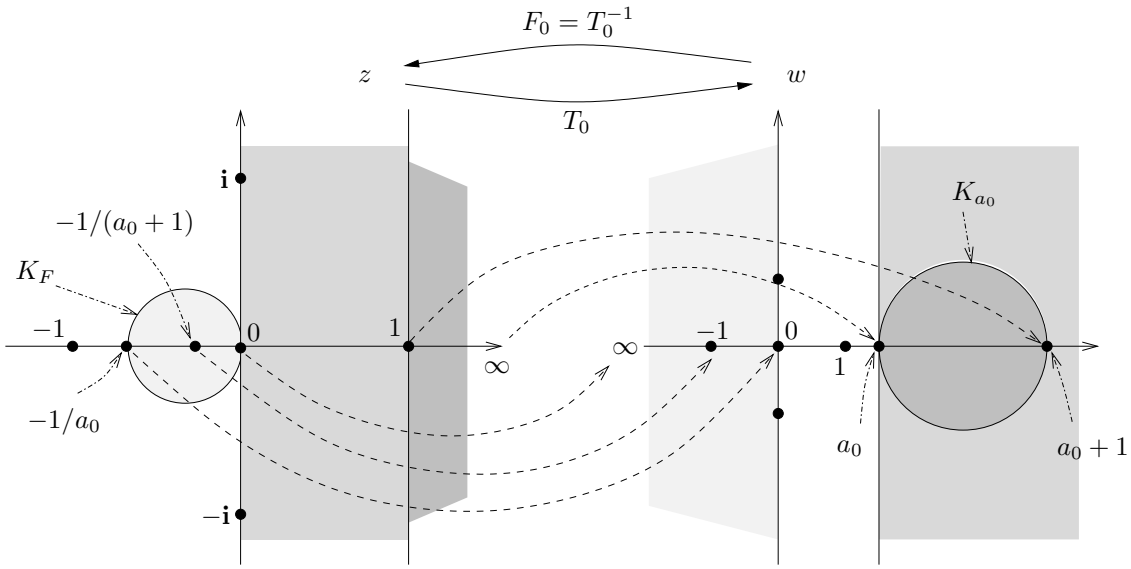


Figure 3: Transformation $T_0 : z \mapsto (a_0 z + 1)/z$.

.

Transformation of Isolating Intervals. Let us see how the roots of $A_0(X)$ are transformed by T. Let its roots be

$$\alpha_0 = \alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(m)}. \tag{43}$$

Then

$$A_1(X) := A_0(T(X))$$

is a real polynomial whose roots,

$$F(\alpha_0^{(1)}), F(\alpha^{(2)}), \ldots, F(\alpha^{(m)}),$$

are obtained by applying the inverse transform $F := T^{-1}$ to the roots of $A_0(X)$. Let us assume $\alpha_0 > 0$ and $[r_0, s_0] \subseteq [0, \infty]$ is an isolating interval for α_0 with respect to $A_0(X)$, and let $T = T_0$ be the transformation (41). Writing $F_0 := T_0^{-1}$ and $\alpha_1 := F_0(\alpha_0)$, and $a_0 = \lfloor \alpha_0 \rfloor$, it follows that

$$\alpha_1 \in F_0([a_0, a_0 + 1]) = [1, \infty].$$

Notice that $F_0([r_0, s_0])$ is an isolating interval for α_1 relative to $A_1(X) = A_0(T_0(X))$. *A fortiori*,

$$I_1 := F_0([r_0, s_0]) \cap [1, \infty]$$

is an isolating interval for α_1. It is easy to see that $I_1 = [r_1, s_1]$ where

$$\begin{aligned} r_1 &:= \begin{cases} F_0(s_0), & \text{if } s_0 < a_0 + 1, \\ 1, & \text{else;} \end{cases} \\ s_1 &:= \begin{cases} F_0(r_0), & \text{if } r_0 > a_0, \\ \infty, & \text{else.} \end{cases} \end{aligned} \tag{44}$$

We have thus answered the "key question" that motivated this section.

We dispose of two final details: (a) We have assumed $\alpha_0 > 0$. The condition $\alpha_i > 0$ is clearly maintained for $i \geq 1$. In case $\alpha_0 < 0$, we can replace α_0 by $\alpha_0 - \lfloor \alpha_0 \rfloor$ before starting the initial iteration, and also modify $A_0(X)$ and $[r_0, s_0]$ accordingly. (b) If $s_1 = \infty$, we may have trouble performing the standard binary search for $a_1 = \lfloor \alpha_1 \rfloor$ in the infinite interval $[r_1, s_1]$. But it is easy to replace s_1 by some finite upper bound on α_1 (for instance, Cauchy's bound in §VI.2).

Exercises

Exercise 7.1: Let $\alpha = (-b + \sqrt{\Delta})/2a$ be a zero of $A(X) = aX^2 + bX + c$ with discriminant $\Delta = b^2 - 4ac$.
(i) Let $a_0 = \lfloor \alpha \rfloor$ and let $\alpha = a_0 + (1/\alpha')$. Let α' be a root of the polynomial $a'X^2 + b'X + c'$. Show that $b'^2 - 4a'c' = \Delta$.
(ii) Let T be the transformation (42) and $A(T(X)) = a^T X^2 + b^T X + c^T$. If α lies between a/c and b/d and $|ad - bc| = 1$ then $|a^T| < 2|a\alpha| + |a| + |b|$, $|c^T| < 2|a\alpha| + |a| + |b|$ and $|b^T| < 4(2|a\alpha| + |a| + |b|)^2 + |\Delta|$.
(iii) Conclude that the regular continued fraction of a quadratic number is eventually periodic. (This is the harder direction of Lagrange's theorem, §4). □

§8. Continued Fractions of Roots

Let us recapitulate the continued fraction algorithm developed in the previous section. Suppose we are given a polynomial $A_0(X)$ of degree m and an isolating interval $[r_0, s_0]$ for a real root $\alpha_0 > 0$ of A_0. To generate the regular continued fraction $[a_0, a_1, a_2, \ldots]$ of α, we compute the following successive members of the following sequence of tuples:

$$(r_i, s_i, A_i(X), a_i), \qquad i \geq 0 \tag{45}$$

where

1. $[r_i, s_i]$ is an isolating interval for a root α_i of $A_i(X)$;
2. $a_i := \lfloor \alpha_i \rfloor$;
3. $A_{i+1}(X) := X^m A_i(a_i + X^{-1})$; and
4. r_i and s_i are computed as in equation (44); we assume that s_i is further replaced by some root-bound if this is better.

Here, $\alpha_i = 1/(\alpha_{i-1} - a_{i-1})$ is just the ith tail quotient of $[a_0, a_1, a_2, \ldots]$. Note that a_i is found by a binary search on the interval $[\lfloor r_i \rfloor, \lfloor s_i \rfloor]$.

Although the sequence (45) is potentially infinite, it is useful to give a definite termination condition. Suppose that we only want to approximate α_0 to within some absolute error $\epsilon > 0$. Then we can compute, in addition to (45), also the ith numerator P_i and denominator Q_i (§3) of the $[a_0, a_1, \ldots]$. We can terminate this computation at the ith iteration where

$$\frac{1}{Q_{i-1}Q_i} \leq \epsilon$$

since the ith approximant P_i/Q_i has the property $|P_i/Q_i - \alpha_0| < 1/(Q_{i-1}Q_i)$ (§6).

The reader can easily turn this description into a more explicit algorithm.

Reduced Numbers. In view of the preceding, it is natural to call a real algebraic number α *reduced* if $\alpha > 1$ and all of whose conjugates β distinct from α lie in the interior of circle in the complex plane with the real line segment $[-1, 0]$ as diameter. This is related[7] to Zassenhaus' notion [38]: a real algebraic number α is said to be in "reduced state" if $\alpha > 1$ and for any conjugate $\beta \neq \alpha$,

$$\mathtt{Re}(\beta) < 0, \qquad |\beta| < 1.$$

Clearly, reduced numbers are in reduced state. Note that the transformation (42) is favorable for reduced numbers: if α_0 is reduced then so is $F(\alpha_0)$. Reduced numbers have the trivial isolating interval $[1, \infty]$, and the tails of their continued fractions remain reduced. We now show that every real algebraic number can be transformed into a reduced number within an explicitly given number of transformations of the form (41).

Let $\delta > 0$ be a root separation bound (§VI.7) for $A_0(X)$, and as usual, let α_i be the ith tail quotient of the regular continued fraction of α_0.

Theorem 6 *If $Q_{i-2}Q_{i-1} \geq 2/\delta$ then α_{i+2} is a reduced number.*

Proof. Let $\beta \neq \alpha_0$ be any conjugate of α_0. Setting $\beta_0 := \beta$, we transform β_i to $\beta_{i+1} = 1/(\beta_i - a_i)$, in analogy to our transformations on the α_i's. Then we have, for $i \geq 1$,

$$\begin{aligned}
\begin{bmatrix} \beta \\ 1 \end{bmatrix} &= \begin{bmatrix} P_i & P_{i-1} \\ Q_i & Q_{i-1} \end{bmatrix} \begin{bmatrix} \beta_i \\ 1 \end{bmatrix} \\
\begin{bmatrix} \beta_i \\ 1 \end{bmatrix} &= (-1)^i \begin{bmatrix} Q_{i-1} & -P_{i-1} \\ -Q_i & P_i \end{bmatrix} \begin{bmatrix} \beta \\ 1 \end{bmatrix} \\
\beta_i &= (-1)^i \frac{\beta Q_{i-1} - P_{i-1}}{-\beta Q_i + P_i} \\
&= (-1)^i \frac{\beta - P_{i-1}/Q_{i-1}}{-\beta + P_i/Q_i} \cdot \frac{Q_{i-1}}{Q_i} \\
&= (-1)^i \frac{\beta - P_{i-1}/Q_{i-1}}{-\beta + P_i/Q_i} \cdot \frac{1}{a_i} \cdot \left(1 - \frac{Q_{i-2}}{Q_i}\right) \qquad (46)
\end{aligned}$$

where the last equation is a consequence of the recurrence $Q_i = a_i Q_{i-1} + Q_{i-2}$. Our goal is to show that

$$|\beta_i| < 1/a_i \qquad (47)$$

for i large enough. Once this goal is reached, all the conjugates of α_i lie in the half-plane $\mathtt{Re}(w) < 1$ and by our observations on the map T_0, these conjugates will be mapped via F_0 to the half-space $\mathtt{Re}(z) < 0$. That is, all the conjugates of α_{i+1} (which are among the roots of $A_{i+1}(X)$) have negative real parts. By another application of the observation about T_0, we conclude that all the conjugates of α_{i+2} lie within the circle with diameter $[-1, 0]$, *i.e.*, α_{i+2} is reduced, proving our theorem.

The goal (47) follows from (46) if we show

$$\left| \frac{\beta - P_{i-1}/Q_{i-1}}{-\beta + P_i/Q_i} \right| \leq 1 + \frac{Q_{i-2}}{Q_i}. \qquad (48)$$

[7] It is also related to the notion of a PV-number or Pisot-Vijayaraghavan number. This is a real algebraic integer α with conjugates $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\alpha > 1$ and $|\alpha_j| < 1$ for $j \neq 1$. See Cassels [39]. The smallest PV-number is the real root of the $X^3 - X - 1$ [C. L. Siegel, 1944].

To this end, note that

$$\begin{aligned}\left|\frac{\beta - P_{i-1}/Q_{i-1}}{-\beta + P_i/Q_i}\right| &= \left|\frac{\beta - (P_i/Q_i) - (-1)^i/(Q_iQ_{i-1})}{-\beta + (P_i/Q_i)}\right| \\ &\leq 1 + \frac{1}{Q_{i-1}Q_i} \cdot \frac{1}{|-\beta + (P_i/Q_i)|}.\end{aligned}$$

Hence it suffices to show that

$$\begin{aligned}\frac{Q_{i-2}}{Q_{i-1}} &\geq \frac{1}{Q_{i-1}Q_i}\frac{1}{|-\beta + (P_i/Q_i)|}, \\ Q_{i-1}Q_{i-2} &\geq \frac{1}{|-\beta + (P_i/Q_i)|}.\end{aligned} \tag{49}$$

By definition of δ, $|\beta - \alpha_0| \geq \delta$. From (40), $|P_i/Q_i - \alpha_0| \leq 1/(Q_iQ_{i-1})$. Hence, if i satisfies the condition of the theorem then

$$\left|-\beta + \frac{P_i}{Q_i}\right| \geq |\beta - \alpha_0| - \left|\alpha_0 - \frac{P_i}{Q_i}\right| \geq \delta - \frac{1}{Q_iQ_{i-1}} > \frac{\delta}{2} \geq \frac{1}{Q_{i-1}Q_{i-2}}.$$

Q.E.D.

A simplified version of this bound goes as follows. Recall that Q_i's satisfies the recurrence

$$Q_i = a_iQ_{i-1} + Q_{i-2}, \qquad (i \geq 2; Q_0 = 1, Q_1 = q_1).$$

Compare this with the Fibonacci numbers

$$F_i = F_{i-1} + F_{i-2}, \qquad (i \geq 2; F_0 = 0, F_1 = 1).$$

It follows that

$$Q_i \geq F_{i+1}, \qquad i \geq 0. \tag{50}$$

It is easy to show that $F_i \geq \phi^{i-1}$ for $i \geq 1$ and $\phi = (\sqrt{5} - 1)/2$. Thus $Q_i \geq \phi^i$ for all $i \geq 0$. Thus:

Corollary 7 *If* $2i - 3 \geq \log_\phi(2/\delta)$, *i.e.,*

$$i \geq \frac{3 + \log_\phi(2/\delta)}{2}, \tag{51}$$

then α_{i+2} *will be reduced.*

For the special case of quadratic numbers, more efficient and specialized algorithms for their continued fractions have been known from antiquity (Exercises).

Exercises

Exercise 8.1: Let $R := \{\lfloor r_0 \rfloor, \lfloor s_1 \rfloor, \lfloor r_2 \rfloor, \lfloor s_3 \rfloor, \ldots\}$ and $S := \{\lfloor s_0 \rfloor, \lfloor r_1 \rfloor, \lfloor s_2 \rfloor, \lfloor r_3 \rfloor, \ldots\}$. Then both R and S eventually become the periodic sequence $\{\ldots, 1, \infty, 1, \infty, \ldots\}$. Before the appearance of this periodic part, R is the regular continued fraction expansion of r_0 and S is the regular continued fraction expansion of s_0. The common "aperiodic" prefix of both sequences must agree except possibly for the last partial quotient. □

Exercise 8.2:
(i) Show that $F_i = (\phi^n - \widehat{\phi}^n)/\sqrt{5}$ where $\phi = (\sqrt{5}-1)/2$ and $\widehat{\phi} = 1 - \phi = -0.618\ldots$. Verify the bound $F_i \geq \phi^{i-1}$.
(ii) If $A_0(X)$ is of degree m and $\|A_0\|_\infty \leq M$, deduce an upper bound on i in terms of m, M that guarantees α_i will be reduced. □

Exercise 8.3: Let $a_0, a_1, \ldots$ be an arbitrary sequence of positive integers and we transform an initial real polynomial $A_0(X)$ by a succession of substitutions, $A_{i+1}(X) = A_i(a_i + X^{-1})$. If i satisfies (51) then $A_i(X)$ has at most one root with positive real part.
NOTE: this is closely related to a theorem of Vincent (1836) which Uspensky [205] uses as a basis for a root separation method. Vincent's theorem concludes that $A_i(X)$ has at most one sign variation if i satisfies a comparable bound to (51). □

Exercise 8.4: (Thull) Suppose α_0 is reduced and $T(\alpha_1) = \alpha_0$ where T is given by (42).
(i) Let $S = \alpha^{(2)} + \cdots + \alpha^{(m)}$ be the sum of all the conjugates of α_0 which are different from α_0. Then S lies in the interval with endpoints $-(m-1)b/a$ and $-(m-1)d/c$.
(ii) Let $A_0(T(X)) = \sum_{i=0}^m b_i X^i$, where $A_0(\alpha_0) = 0$. Then α_1 lies in the interval with endpoints $-(b_{m-1}/b_m) + (m-1)b/a$ $-(b_{m-1}/b_m) + (m-1)d/c$. □

Exercise 8.5: We develop an algorithm for quadratic numbers. Let $N > 1$ be a square-free integer, $\alpha = (a + \sqrt{N})/b$ for integers a, b and $R = \left\lfloor \sqrt{N} \right\rfloor$.
(i) $\left\lfloor (a+\sqrt{N})/b \right\rfloor$ is equal to $\lfloor (a+R)/b \rfloor$ if $b > 0$ and equal to $\lfloor (1+a+R)/b \rfloor$ if $b < 0$.
(ii) Let $q = \lfloor (a+R)/b \rfloor$ and suppose $(a'+R)/b' = ((a+R)/b - q)^{-1}$. Then $a' = bq - a$ and $b' = (N - a'^2)/b$.
(iii) Repeating the transformations $(a, b) \to (a', b')$, let us form the sequence $\{(a_i, b_i) : i \geq 0\}$ where $(a_0, b_0) = (a, b)$. Prove that for i large enough, $|a_i| < R, |b_i| < 2R$. [Hence it becomes periodic.]
(iv) Develop a continued fraction algorithm of α using these facts (assuming that we can compute the floor function).
(v) Show that α is reduced iff $0 < a < \sqrt{N}$ and $\sqrt{N} - a < b < \sqrt{N} + a$.
(vi) Modify your algorithm in (iii) to detect the onset of periodicity. HINT: recall a characterization of reduced quadratic numbers in an exercise of §4. □

§9. Arithmetic Operations

We consider the arithmetic operations $+, -, \times, \div$. The algorithmic idea is a very natural one. Say we want to compute $x+y$ where x, y are continued fractions. The result z is supposed to be a continued fraction as well. The algorithm for z will request successive terms from x and y as needed. So we view a continued fraction x as a *process* that can respond to requests for its next term. Once the process x outputs a term, it transforms itself into a new process for a modified continued fraction x'. We assume that x' has no memory of its previous outputs – so the algorithm for z must remember these outputs (in some form). Thus in general, a process has internal memory in the form of *state variables*. Similarly, the algorithm for adding x, y can be viewed as a process for z. The terms of a continued fraction $q_0 + \mathbf{K}_{i\geq 1}(p_i/q_i)$ are given in the order $q_0, p_1, q_1, p_2, q_2, \ldots$. As terms are *ingested* (*i.e.*, consumed) by the algorithm for $x+y$, the state variables of (the process) for z change. We have no *á priori* requirement on how terms from x or y are to be ingested – this is under the control of the algorithm for z. One obvious possibility is to ingest one term *each* from x and y simultaneously. We expect the algorithm to *egest* (*i.e.*, spit out) terms of $z = x+y$ from time to time. These ideas fit

very well into the concept of "streams" and "lazy evaluation" in the programming milieu. What we call a process can be identified as a stream. Lazy evaluation of z means that the algorithm ingests terms from x, y only when there is a pending request for a term of z. Of course, the stream for z can be fed into other on-going computations.

Following Gosper [18], the processes for arithmetic operations can unified under the process for computing the general function

$$z(x,y) = \frac{axy + bx + cy + d}{a'xy + b'x + c'y + d'} \tag{52}$$

where $a, b, \ldots, c', d'$ are numerical constants called the *state variables* of the process. We use the compact notation

$$z(x,y) = \frac{(a,b,c,d)}{(a',b',c',d')}\begin{pmatrix} x \\ y \end{pmatrix}.$$

The arithmetic operations can be recovered with suitable choices for the state variables. Thus:

$$\begin{aligned}
x + y &= \frac{(0,1,1,0)}{(0,0,0,1)}\begin{pmatrix} x \\ y \end{pmatrix} \\
x - y &= \frac{(0,1,-1,0)}{(0,0,0,1)}\begin{pmatrix} x \\ y \end{pmatrix} \\
xy &= \frac{(1,0,0,0)}{(0,0,0,1)}\begin{pmatrix} x \\ y \end{pmatrix} \\
x/y &= \frac{(0,1,0,0)}{(0,0,1,0)}\begin{pmatrix} x \\ y \end{pmatrix} \\
\frac{ax+b}{cx+d} &= \frac{(0,a,0,b)}{(0,c,0,d)}\begin{pmatrix} x \\ y \end{pmatrix}
\end{aligned}$$

The last operation is a Möbius transformation in the variable x, since it does not depend on y. It comes from the general expression $z(x,y)$ by replacing y with 0, and may arise from the termination of the stream for y.

Ingesting terms. It is easy to give transformations of the state variables when we ingest one term from x: there are two cases, depending on whether we ingest a partial numerator or a partial denominator. To avoid this dichotomy, it is simplest to ingest a pair of terms at a time. Say the 0th partial denominator $q = q_0$ and the 1st partial numerator $p = p_1$ of the current x are ingested. So

$$x = q + \frac{p}{x'}$$

where x' is the 1st tail of x. Let us see how the state variables changes:

$$\begin{aligned}
z(x,y) &= \frac{a(q + p/x')y + b(q + p/x') + cy + d}{a'(q + p/x')y + b'(q + p/x') + c'y + d'} \\
&= \frac{(aq + c, bq + d, ap, bp)}{(a'q + c', b'q + d', a'p, b'p)}\begin{pmatrix} x' \\ y \end{pmatrix}.
\end{aligned}$$

This changes z from a function of x, y to another function of x', y. In general, if x is a real number and x_i is its ith tail, then from (30),

$$\begin{aligned}
\begin{bmatrix} x \\ 1 \end{bmatrix} &= \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix}\begin{bmatrix} x_i \\ 1 \end{bmatrix} \\
&= \frac{(0, A_i, 0, B_i)}{(0, C_i, 0, D_i)}\begin{pmatrix} x_i \\ y \end{pmatrix}.
\end{aligned}$$

Alternatively, if x is in "complementary form" $x = \mathbf{K}_{i\geq 1}(p_i/q_i)$ (*i.e.*, $q_0 = 0$) then we could write

$$x = \frac{p_1}{q_1 + x'}$$

where x' is now the 1st complement. If we ingest the pair $(p, q) = (p_1, q_1)$ the transformed function becomes

$$\begin{aligned} z(x,y) &= \frac{ayp/(q+x') + bp/(q+x') + cy + d}{a'yp/(q+x') + b'p/(q+x') + c'y + d'} \\ &= \frac{(c, d, ap+cq, bp+dq)}{(c', d', a'p+c'q, b'p+d'q)} \begin{pmatrix} x' \\ y \end{pmatrix}. \end{aligned}$$

Egesting terms. Next suppose that we are ready to egest a term from $z(x, y)$. For now, we defer the question of deciding when it is appropriate to egest terms and what these terms should be. Again, it is convenient to assume that we will egest a pair of terms u, v and transform z to z':

$$z = u + \frac{v}{z'}.$$

Then

$$\begin{aligned} z' &= \frac{v}{z-u} \\ &= \frac{v}{\frac{axy+bx+cy+d}{a'xy+b'x+c'y+d'} - u} \\ &= \frac{(a'v, b'v, c'v, d'v)}{(a-ua', b-ub', c-uc', d-ud')} \begin{pmatrix} x \\ y \end{pmatrix}. \end{aligned}$$

Similarly, if z is already in the "complementary form", we can write $z = (v/z') - u$, and derive z' accordingly.

The strategy for deciding when and what terms to egest is simplified by assuming that z is a regular continued fraction (in particular, x, y are real and $v = 1$ above). The state variables of the process for z contains information about the range of possible values for z. If this range of values is narrowed sufficiently so that we know the value of $\lfloor z \rfloor$, then we may egest the pair $(u, v) = (\lfloor z \rfloor, 1)$ and transform z to z' as above. It is intuitively clear that without any restriction on the real numbers represented by x, y, there cannot be any *á priori* bound on how many terms must be ingested before the range is "sufficiently narrow" [213]. But for nice numbers (say algebraic numbers), such bounds exist. In many situations, the range of possible values are can be restricted to intervals of $\widehat{\mathbb{R}}$.

Gosper suggests a method of egesting intermediate information without waiting for $\lfloor z \rfloor$ to be determined. Suppose we have narrowed the range of z so that we know that z lies in $[3000, 4000)$. Then we could egest 3000 and transform z to $z' = z - 3000$. Now we know that z' lies in $[0, 1000)$. If subsequent narrowing of the intervals for x, y tells us that z' lies in $[700, 800)$, we then egest the two terms 0 and 700, and transform z' to $z'' = z - 700$. Now we know z'' lies in $[0, 100)$. Suppose we further discover z'' lies in $[20, 30)$. We then egest the two terms 0 and 20, transforming z'' to $z''' = z'' - 20$. Now $z''' \in [0, 10)$. Say we discover $z''' \in [5, 6)$. We may then output 0 and 5 and transform z''' to $z'''' = 1/(z''' - 5)$ and continue in the normal fashion. Notice that we have egested the sequence

$$\ldots, 3000, 0, 700, 0, 20, 0, 5, \ldots$$

which, by rule (21), is equivalent to egesting one term 3725. The intermediate terms $3000, 700, 20$ may be useful, for instance, in narrowing the interval containing z, without waiting for the eventual discovery of the term 3725 (which may not even be necessary). Of course, by having partial

denominators that vanish, we have slightly violated the condition for a regular continued fraction. Our example is intended to suggest[8] the output of decimal representation for partial denominators, but clearly the rule (21) can be applied in more general situations. We can also modify the egestion algorithm to output the value of a continued fraction in human-friendly decimal notation (Exercise).

Intervals. As suggested above, we would like to restrict the range of possible values of $z(x, y)$ to an interval. Suppose x has already egested $(i + 1)$ pairs of terms:

$$q_0, p_1, q_1, p_2, \ldots, q_i, p_{i+1}.$$

Inductively, we may assume that the numerators and denominators $P_{i-1}, P_i, Q_{i-1}, Q_i$ have been maintained. Therefore x lies in the finite interval with endpoints

$$P_{i-1}/Q_{i-1}, P_i/Q_i.$$

(Recall that i is even iff $P_{i-1}/Q_{i-1} < P_i/Q_i$.) Similarly, we may assume that we know the value of y within some interval. We can then deduce an interval containing z. In general, if $z(x_1, \ldots, x_n)$ is a function where x_i are continued fractions in complementary form, then the set of possible values of z is $z([1, \infty], \ldots, [1, \infty])$. If z is one of the arithmetic operations, this set is an interval. We now give an account of intervals and their basic calculus.

An *interval* I will refer to a connected subset of the extended real numbers $\widehat{\mathbb{R}}$. The special case where $I = \widehat{\mathbb{R}}$ will be given the special notation $\perp$ (this will turn out to be an extension of the use of the symbol $\perp$ in §2). If $I \neq \perp$, we will call I a *proper* interval. We classify proper intervals as follows: I is *infinite* if $\infty \in I$, otherwise it is *finite*. I is *positive* if it comprises only positive numbers (0 and ∞ are neither positive nor negative). Similarly for a *negative* interval. The reader should visualize I using the stereographic projection of I onto the unit complex circle S^1 (§2, figure 1).

A proper interval I has two *endpoints*, a and b which are not necessarily distinct. We now introduce some conventions for representing I in terms of a, b. First assume $a = b$. Then there are two possibilities: either I consists of just the number a or I is equal to $\widehat{\mathbb{R}} \setminus \{a\}$. In the former case, I is a *point interval* denoted $I = [a, a]$; in the latter case, it is a *deleted point interval* denoted $I = (a, a)$. We sometimes write a for $[a, a]$. Next assume the endpoints are distinct and not equal to ∞, say $a < b$. Suppose I is a closed interval (*i.e.*, $a, b \in I$). Then we can use the notation

$$I = \begin{cases} [a, b] & \text{if } I \text{ is finite} \\ [b, a] & \text{if } I \text{ is infinite} \end{cases} \tag{53}$$

If one end point, say b, equals ∞ then we let $I = [a, \infty]$ denote the set of reals that are greater than or equal to a including ∞; and $I = [\infty, a]$ denotes the set of reals less than or equal to a including ∞. If I is open (meaning that both a, b are not in I) then we either write $I = (a, b)$ or $I = (b, a)$ under the same conditions as in (53). We can generalize these notations to half-open intervals $(a, b]$ or $[a, b)$ when $a \neq b$, in the standard way. Note that $(a, a]$ and $[a, a)$ are not defined.

Interval calculus. We define algebraic operations on intervals in a generic fashion: assume that $f(x, y)$ and $g(x)$ are algebraic operations on $\widehat{\mathbb{R}}$. We extend these operations to any subsets $I, J \subseteq \widehat{\mathbb{R}}$ via

$$f(I, J) := \{f(x, y) : x \in I, y \in J\}, \qquad g(I) := \{g(x) : x \in I\}.$$

Now we assume that these operations are continuous and I, J are intervals. Then the result of these operations are intervals. There is an exception to be handled in this generic definition of

[8] A more user-friendly output might be '$3 * **$', subsequently revised to '$32 * *$', then to '$327*$' and finally '3275'.

operations on intervals: it is possible (as usual) that $f(x,y) = \perp$ or $g(x) = \perp$. By definition, if $f(x_0, y_0) = \perp$ for some $x_0 \in I, y_0 \in J$ then we define $f(I,J) = \perp$. But it is also quite possible that $f(I,J) = \perp$ even when $I \times J$ does not contain such a pair (x_0, y_0). For instance, $[1,-1] + [0,2] = \perp$ and $[2,1] \times [1,2] = \perp$.

We consider the case where f, g are rational operations. These are well-known to be continuous. Another useful fact is this: if $f(x,y)$ is monotonic increasing or monotonic decreasing in the finite intervals I, J then the endpoints of $f(I,J)$ (if it is not $\perp$) can be determined by applying $f(x,y)$ to the endpoints of I and J.

Lemma 8 *Let $I = [a,b]$ and $I' = [a',b']$ be proper intervals.*

(i) $1/[a,b] = [1/b, 1/a]$ *provided* $b \neq \infty$. *Otherwise,* $1/[a,b] = [1/a, 1/b]$.

(ii) $-[a,b] = [-b,-a]$.

(iii) *(Addition) If $\infty \in I \cap I'$ then $I + I' = \perp$. If I, I' are both finite, then $I + I' = [a+a', b+b']$. Otherwise,*

$$I + I' = \begin{cases} \perp & \text{if } a + a' \leq b + b' \\ [a+a', b+b'] & \text{else.} \end{cases}$$

(iv) *(Multiplication) Let $E = \{a,b\} \times \{a',b'\} = \{aa', ab', ba', bb'\}$.*
- **(iv.1)** *If $0 \in I, \infty \in I'$ then $I \times I' = \perp$.*
- **(iv.2)** *If $\infty \notin I \cup I'$ then $I \times I' = [\min(E), \max(E)]$.*
- **(iv.3)** *If $0 \notin I \cup I'$ then $I \times I' = [1/\max(1/E), 1/\min(1/E)]$.*
- **(iv.4)** *If none of the above is true, let $\{0, \infty\} \subseteq I'$. So I is either positive or negative.*

$$I \times I' = \begin{cases} [\min(a' \times I), \max(b' \times I)] & \text{if } I \text{ is positive and } \min(a' \times I) > \max(b' \times I), \\ [\min(b' \times I), \max(a' \times I)] & \text{if } I \text{ is negative and } \min(b' \times I) > \max(a' \times I), \\ \perp & \text{else.} \end{cases}$$

Proof. We prove (iv). By definition, $I \times I' = \perp$ if condition (iv.1) holds. Case (iv.2): here, I and I' are both finite. It is easy to check that the result is true in case I is "definite" (i.e., I does not contain both negative and positive numbers). If I is indefinite, write $I = I^+ \uplus I^-$ as the union of two definite intervals and use the fact that $I \times I' = (I^+ \times I') \cup (I^- \times I')$. Case (iv.3): both $1/I$ and $1/I'$ are finite. From part (i), we have $1/I = [1/b, 1/a]$ and similarly for $1/I'$. From case (iv.2), we have $(1/I) \times (1/I') = [\min(1/E), \max(1/E)]$. Using part (i) again, we have $I \times I' = [1/\max(1/E), 1/\min(1/E)]$. Note this expression cannot be simplified to $[\min(E), \max(E)]$, e.g., if $E = \{-2, 1\}$ then $1/\max(1/E) \neq \min(E)$. Finally consider case (iv.4). Since the previous cases do not hold, it follows from $\{0, \infty\} \subseteq I'$ that I is definite (either positive or negative). Also, I' is co-definite (i.e., its complement is definite). Defining "co-negative" or "co-positive" similarly, we verify

$$I \times I' = \begin{cases} [a'a, b'b] & \text{if } I \text{ is positive, } I' \text{ is co-positive, and } b'b < a'a, \\ [a'b, b'a] & \text{if } I \text{ is positive, } I' \text{ is co-negative, and } b'a < a'b, \\ [b'a, a'b] & \text{if } I \text{ is negative, } I' \text{ is co-positive, and } a'b < b'a, \\ [b'b, a'a] & \text{if } I \text{ is negative, } I' \text{ is co-negative, and } a'a < b'b, \\ \perp & \text{else.} \end{cases}$$

This is slightly summarized in the statement of the lemma. **Q.E.D.**

A corollary of this lemma is that the four arithmetic operations preserves intervals. Using these rules, we can implement arithmetic for continued fractions in which we maintain the intervals as terms are egested and ingested. The chordal metric may be used to show that eventually, a new term is egested.

Exercises

Exercise 9.1: Consider the process for $z(x, y)$ above.
(i) Carry out the transformations for ingesting a single term of x; likewise, for a single term each from x and from y.
(ii) Carry out the transformation for ingesting a pair of terms simultaneously from x and from y. (There are 4 cases depending on whether x, y are in complementary form or not.)
(ii) Suppose we only want to ingest one term from either x or y. Discuss heuristics to decide which term should be ingest. □

Exercise 9.2: Modify lemma 8 for open or half-open intervals. □

Exercise 9.3: (Gosper) Let $z(x, y)$ be given by (52). Modify the egestion algorithm above to output the value of z in decimal notation. In particular, we can print the value of any continued fraction in decimal notation. □

Exercise 9.4: Suppose x, y are algebraic numbers of degree at most d and height at most h. Bound (in terms of d, h) the number of terms of x, y that must be egested before we can egest a term of $z = x + y$. □

Exercise 9.5: (Vuillemin)
As an alternative to regular continued fractions we choose the partial denominators q_i as follows: Let $\lfloor r \rceil$ (*rounding* of r) denote the integer nearest to a real number r, breaking ties in some systematic way[9]. For definiteness, we choose $\lfloor r \rceil = \lfloor r \rfloor$ in case of a tie. Then define $\texttt{ZCF}(r) = 1/(r - \lfloor r \rceil)$. The $\mathbb{Z}$-*continued fraction* of r is the sequence

$$[q_0, q_1, \ldots]$$

where $q_i = \lfloor r_i \rceil$ and $r_0, r_1, \ldots$ is the sequence $r_0 = r$ and $r_{i+1} = \texttt{ZCF}(r_i)$.
(i) Show that $|q_i| \geq 2$ for $i \geq 0$.
(ii) If $|q_i| = 2$ then q_{i+1} has the same sign as q_i.
(iii) If the sequence $[q_0, q_1, \ldots]$ terminates then the last term is different from -2.
(iv) Replace the rounding function $\lfloor r \rceil$ in the definition of the $\mathbb{Z}$-continued fraction by a more pragmatic version that can return any value q such that $\Delta(r, q) < 1$ (the chordal metric of §2). Prove that this new continued fraction $[q_0, q_1, \ldots]$ (which is no longer uniquely determined by r) has the value r. Vuillemin calls this the E-continued fraction of r. □

[9]There are four common ways of doing this — always pick the larger, always the smaller, always the smaller magnitude or always the larger magnitude. Thus $\lfloor 1.5 \rceil = 2, 1, 1, 2$ depending on which of these four rules are used. Likewise, $\lfloor -1.5 \rceil = -1, -2, -1, -2$.

References

[1] W. W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics, Vol. 3. American Mathematical Society, Providence, R.I., 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Massachusetts, 1974.

[3] S. Akbulut and H. King. *Topology of Real Algebraic Sets*. Mathematical Sciences Research Institute Publications. Springer-Verlag, Berlin, 1992.

[4] E. Artin. *Modern Higher Algebra (Galois Theory)*. Courant Institute of Mathematical Sciences, New York University, New York, 1947. (Notes by Albert A. Blank).

[5] E. Artin. *Elements of algebraic geometry*. Courant Institute of Mathematical Sciences, New York University, New York, 1955. (Lectures. Notes by G. Bachman).

[6] M. Artin. *Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[7] A. Bachem and R. Kannan. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Computing*, 8:499–507, 1979.

[8] C. Bajaj. Algorithmic implicitization of algebraic curves and surfaces. Technical Report CSD-TR-681, Computer Science Department, Purdue University, November, 1988.

[9] C. Bajaj, T. Garrity, and J. Warren. On the applications of the multi-equational resultants. Technical Report CSD-TR-826, Computer Science Department, Purdue University, November, 1988.

[10] E. F. Bareiss. Sylvester's identity and multistep integer-preserving Gaussian elimination. *Math. Comp.*, 103:565–578, 1968.

[11] E. F. Bareiss. Computational solutions of matrix problems over an integral domain. *J. Inst. Math. Appl.*, 10:68–104, 1972.

[12] D. Bayer and M. Stillman. A theorem on refining division orders by the reverse lexicographic order. *Duke Math. J.*, 55(2):321–328, 1987.

[13] D. Bayer and M. Stillman. On the complexity of computing syzygies. *J. of Symbolic Computation*, 6:135–147, 1988.

[14] D. Bayer and M. Stillman. Computation of Hilbert functions. *J. of Symbolic Computation*, 14(1):31–50, 1992.

[15] A. F. Beardon. *The Geometry of Discrete Groups*. Springer-Verlag, New York, 1983.

[16] B. Beauzamy. Products of polynomials and a priori estimates for coefficients in polynomial decompositions: a sharp result. *J. of Symbolic Computation*, 13:463–472, 1992.

[17] T. Becker and V. Weispfenning. *Gröbner bases : a Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993. (written in cooperation with Heinz Kredel).

[18] M. Beeler, R. W. Gosper, and R. Schroepppel. HAKMEM. A. I. Memo 239, M.I.T., February 1972.

[19] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. of Computer and System Sciences*, 32:251–264, 1986.

[20] R. Benedetti and J.-J. Risler. *Real Algebraic and Semi-Algebraic Sets*. Actualités Mathématiques. Hermann, Paris, 1990.

[21] S. J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Info. Processing Letters*, 18:147–150, 1984.

[22] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill Book Company, New York, 1968.

[23] J. Bochnak, M. Coste, and M.-F. Roy. *Geometrie algebrique reelle*. Springer-Verlag, Berlin, 1987.

[24] A. Borodin and I. Munro. *The Computational Complexity of Algebraic and Numeric Problems*. American Elsevier Publishing Company, Inc., New York, 1975.

[25] D. W. Boyd. Two sharp inequalities for the norm of a factor of a polynomial. *Mathematika*, 39:341–349, 1992.

[26] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *J. Algorithms*, 1:259–295, 1980.

[27] J. W. Brewer and M. K. Smith, editors. *Emmy Noether: a Tribute to Her Life and Work*. Marcel Dekker, Inc, New York and Basel, 1981.

[28] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Series in Computational Mathematics, vol.12. Springer-Verlag, 1991.

[29] E. Brieskorn and H. Knörrer. *Plane Algebraic Curves*. Birkhäuser Verlag, Berlin, 1986.

[30] W. S. Brown. The subresultant PRS algorithm. *ACM Trans. on Math. Software*, 4:237–249, 1978.

[31] W. D. Brownawell. Bounds for the degrees in Nullstellensatz. *Ann. of Math.*, 126:577–592, 1987.

[32] B. Buchberger. Gröbner bases: An algorithmic method in polynomial ideal theory. In N. K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and its Applications, chapter 6, pages 184–229. D. Reidel Pub. Co., Boston, 1985.

[33] B. Buchberger, G. E. Collins, and R. L. (eds.). *Computer Algebra*. Springer-Verlag, Berlin, 2nd edition, 1983.

[34] D. A. Buell. *Binary Quadratic Forms: classical theory and modern computations*. Springer-Verlag, 1989.

[35] W. S. Burnside and A. W. Panton. *The Theory of Equations*, volume 1. Dover Publications, New York, 1912.

[36] J. F. Canny. *The complexity of robot motion planning*. ACM Doctoral Dissertion Award Series. The MIT Press, Cambridge, MA, 1988. PhD thesis, M.I.T.

[37] J. F. Canny. Generalized characteristic polynomials. *J. of Symbolic Computation*, 9:241–250, 1990.

[38] D. G. Cantor, P. H. Galyean, and H. G. Zimmer. A continued fraction algorithm for real algebraic numbers. *Math. of Computation*, 26(119):785–791, 1972.

[39] J. W. S. Cassels. *An Introduction to Diophantine Approximation*. Cambridge University Press, Cambridge, 1957.

[40] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin, 1971.

[41] J. W. S. Cassels. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[42] T. J. Chou and G. E. Collins. Algorithms for the solution of linear Diophantine equations. *SIAM J. Computing*, 11:687–708, 1982.

[43] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[44] G. E. Collins. Subresultants and reduced polynomial remainder sequences. *J. of the ACM*, 14:128–142, 1967.

[45] G. E. Collins. Computer algebra of polynomials and rational functions. *Amer. Math. Monthly*, 80:725–755, 1975.

[46] G. E. Collins. Infallible calculation of polynomial zeros to specified precision. In J. R. Rice, editor, *Mathematical Software III*, pages 35–68. Academic Press, New York, 1977.

[47] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[48] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. of Symbolic Computation*, 9:251–280, 1990. Extended Abstract: ACM Symp. on Theory of Computing, Vol.19, 1987, pp.1-6.

[49] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[50] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1992.

[51] J. H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, New York, 1988.

[52] M. Davis. *Computability and Unsolvability*. Dover Publications, Inc., New York, 1982.

[53] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential Diophantine equations. *Annals of Mathematics, 2nd Series*, 74(3):425–436, 1962.

[54] J. Dieudonné. *History of Algebraic Geometry*. Wadsworth Advanced Books & Software, Monterey, CA, 1985. Trans. from French by Judith D. Sally.

[55] L. E. Dixon. Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors. *Amer. J. of Math.*, 35:413–426, 1913.

[56] T. Dubé, B. Mishra, and C. K. Yap. Admissible orderings and bounds for Gröbner bases normal form algorithm. Report 88, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1986.

[57] T. Dubé and C. K. Yap. A basis for implementing exact geometric algorithms (extended abstract), September, 1993. Paper from URL `http://cs.nyu.edu/cs/faculty/yap`.

[58] T. W. Dubé. *Quantitative analysis of problems in computer algebra: Gröbner bases and the Nullstellensatz*. PhD thesis, Courant Institute, N.Y.U., 1989.

[59] T. W. Dubé. The structure of polynomial ideals and Gröbner bases. *SIAM J. Computing*, 19(4):750–773, 1990.

[60] T. W. Dubé. A combinatorial proof of the effective Nullstellensatz. *J. of Symbolic Computation*, 15:277–296, 1993.

[61] R. L. Duncan. Some inequalities for polynomials. *Amer. Math. Monthly*, 73:58–59, 1966.

[62] J. Edmonds. Systems of distinct representatives and linear algebra. *J. Res. National Bureau of Standards*, 71B:241–245, 1967.

[63] H. M. Edwards. *Divisor Theory*. Birkhauser, Boston, 1990.

[64] I. Z. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, Department of Computer Science, University of California, Berkeley, 1989.

[65] W. Ewald. *From Kant to Hilbert: a Source Book in the Foundations of Mathematics*. Clarendon Press, Oxford, 1996. In 3 Volumes.

[66] B. J. Fino and V. R. Algazi. A unified treatment of discrete fast unitary transforms. *SIAM J. Computing*, 6(4):700–717, 1977.

[67] E. Frank. Continued fractions, lectures by Dr. E. Frank. Technical report, Numerical Analysis Research, University of California, Los Angeles, August 23, 1957.

[68] J. Friedman. On the convergence of Newton's method. *Journal of Complexity*, 5:12–33, 1989.

[69] F. R. Gantmacher. *The Theory of Matrices, volume 1*. Chelsea Publishing Co., New York, 1959.

[70] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[71] M. Giusti. Some effectivity problems in polynomial ideal theory. In *Lecture Notes in Computer Science*, volume 174, pages 159–171, Berlin, 1984. Springer-Verlag.

[72] A. J. Goldstein and R. L. Graham. A Hadamard-type bound on the coefficients of a determinant of polynomials. *SIAM Review*, 16:394–395, 1974.

[73] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York, 1977.

[74] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Vienna, 1949.

[75] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[76] W. Habicht. Eine Verallgemeinerung des Sturmschen Wurzelzählverfahrens. *Comm. Math. Helvetici*, 21:99–116, 1948.

[77] J. L. Hafner and K. S. McCurley. Asymptotically fast triangularization of matrices over rings. *SIAM J. Computing*, 20:1068–1083, 1991.

[78] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, New York, 1959. 4th Edition.

[79] P. Henrici. *Elements of Numerical Analysis*. John Wiley, New York, 1964.

[80] G. Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95:736–788, 1926.

[81] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[82] C. Ho. Fast parallel gcd algorithms for several polynomials over integral domain. Technical Report 142, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1988.

[83] C. Ho. *Topics in algebraic computing: subresultants, GCD, factoring and primary ideal decomposition*. PhD thesis, Courant Institute, New York University, June 1989.

[84] C. Ho and C. K. Yap. The Habicht approach to subresultants. *J. of Symbolic Computation*, 21:1–14, 1996.

[85] A. S. Householder. *Principles of Numerical Analysis*. McGraw-Hill, New York, 1953.

[86] L. K. Hua. *Introduction to Number Theory*. Springer-Verlag, Berlin, 1982.

[87] A. Hurwitz. Über die Trägheitsformem eines algebraischen Moduls. *Ann. Mat. Pura Appl.*, 3(20):113–151, 1913.

[88] D. T. Huynh. A superexponential lower bound for Gröbner bases and Church-Rosser commutative Thue systems. *Info. and Computation*, 68:196–206, 1986.

[89] C. S. Iliopoulous. Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and Hermite and Smith normal form of an integer matrix. *SIAM J. Computing*, 18:658–669, 1989.

[90] N. Jacobson. *Lectures in Abstract Algebra, Volume 3*. Van Nostrand, New York, 1951.

[91] N. Jacobson. *Basic Algebra 1*. W. H. Freeman, San Francisco, 1974.

[92] T. Jebelean. An algorithm for exact division. *J. of Symbolic Computation*, 15(2):169–180, 1993.

[93] M. A. Jenkins and J. F. Traub. Principles for testing polynomial zerofinding programs. *ACM Trans. on Math. Software*, 1:26–34, 1975.

[94] W. B. Jones and W. J. Thron. *Continued Fractions: Analytic Theory and Applications*. vol. 11, Encyclopedia of Mathematics and its Applications. Addison-Wesley, 1981.

[95] E. Kaltofen. Effective Hilbert irreducibility. *Information and Control*, 66(3):123–137, 1985.

[96] E. Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Computing*, 12:469–489, 1985.

[97] E. Kaltofen. Polynomial factorization 1982-1986. Dept. of Comp. Sci. Report 86-19, Rensselaer Polytechnic Institute, Troy, NY, September 1986.

[98] E. Kaltofen and H. Rolletschek. Computing greatest common divisors and factorizations in quadratic number fields. *Math. Comp.*, 52:697–720, 1989.

[99] R. Kannan, A. K. Lenstra, and L. Lovász. Polynomial factorization and nonrandomness of bits of algebraic and some transcendental numbers. *Math. Comp.*, 50:235–250, 1988.

[100] H. Kapferer. Über Resultanten und Resultanten-Systeme. *Sitzungsber. Bayer. Akad. München*, pages 179–200, 1929.

[101] A. N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. P. Noordhoff N. V., Groningen, the Netherlands, 1963.

[102] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. tr. from Russian by Smilka Zdravkovska.

[103] M. Kline. *Mathematical Thought from Ancient to Modern Times*, volume 3. Oxford University Press, New York and Oxford, 1972.

[104] D. E. Knuth. The analysis of algorithms. In *Actes du Congrés International des Mathématiciens*, pages 269–274, Nice, France, 1970. Gauthier-Villars.

[105] D. E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison-Wesley, Boston, 2nd edition edition, 1981.

[106] J. Kollár. Sharp effective Nullstellensatz. *J. American Math. Soc.*, 1(4):963–975, 1988.

[107] E. Kunz. *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser, Boston, 1985.

[108] J. C. Lagarias. Worst-case complexity bounds for algorithms in the theory of integral quadratic forms. *J. of Algorithms*, 1:184–186, 1980.

[109] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Computing*, 14:184–195, 1985.

[110] S. Landau and G. L. Miller. Solvability by radicals in polynomial time. *J. of Computer and System Sciences*, 30:179–208, 1985.

[111] S. Lang. *Algebra*. Addison-Wesley, Boston, 3rd edition, 1971.

[112] L. Langemyr. *Computing the GCD of two polynomials over an algebraic number field*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, January 1989. Technical Report TRITA-NA-8804.

[113] D. Lazard. Résolution des systémes d'équations algébriques. *Theor. Computer Science*, 15:146–156, 1981.

[114] D. Lazard. A note on upper bounds for ideal theoretic problems. *J. of Symbolic Computation*, 13:231–233, 1992.

[115] A. K. Lenstra. Factoring multivariate integral polynomials. *Theor. Computer Science*, 34:207–213, 1984.

[116] A. K. Lenstra. Factoring multivariate polynomials over algebraic number fields. *SIAM J. Computing*, 16:591–598, 1987.

[117] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[118] W. Li. Degree bounds of Gröbner bases. In C. L. Bajaj, editor, *Algebraic Geometry and its Applications*, chapter 30, pages 477–490. Springer-Verlag, Berlin, 1994.

[119] R. Loos. Generalized polynomial remainder sequences. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 115–138. Springer-Verlag, Berlin, 2nd edition, 1983.

[120] L. Lorentzen and H. Waadeland. *Continued Fractions with Applications*. Studies in Computational Mathematics 3. North-Holland, Amsterdam, 1992.

[121] H. Lüneburg. On the computation of the Smith Normal Form. Preprint 117, Universität Kaiserslautern, Fachbereich Mathematik, Erwin-Schrödinger-Straße, D-67653 Kaiserslautern, Germany, March 1987.

[122] F. S. Macaulay. Some formulae in elimination. *Proc. London Math. Soc.*, 35(1):3–27, 1903.

[123] F. S. Macaulay. *The Algebraic Theory of Modular Systems*. Cambridge University Press, Cambridge, 1916.

[124] F. S. Macaulay. Note on the resultant of a number of polynomials of the same degree. *Proc. London Math. Soc*, pages 14–21, 1921.

[125] K. Mahler. An application of Jensen's formula to polynomials. *Mathematika*, 7:98–100, 1960.

[126] K. Mahler. On some inequalities for polynomials in several variables. *J. London Math. Soc.*, 37:341–344, 1962.

[127] M. Marden. *The Geometry of Zeros of a Polynomial in a Complex Variable*. Math. Surveys. American Math. Soc., New York, 1949.

[128] Y. V. Matiyasevich. *Hilbert's Tenth Problem*. The MIT Press, Cambridge, Massachusetts, 1994.

[129] E. W. Mayr and A. R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[130] F. Mertens. Zur Eliminationstheorie. *Sitzungsber. K. Akad. Wiss. Wien, Math. Naturw. Kl. 108*, pages 1178–1228, 1244–1386, 1899.

[131] M. Mignotte. *Mathematics for Computer Algebra*. Springer-Verlag, Berlin, 1992.

[132] M. Mignotte. On the product of the largest roots of a polynomial. *J. of Symbolic Computation*, 13:605–611, 1992.

[133] W. Miller. Computational complexity and numerical stability. *SIAM J. Computing*, 4(2):97–107, 1975.

[134] P. S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[135] G. V. Milovanović, D. S. Mitrinović, and T. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, Singapore, 1994.

[136] B. Mishra. Lecture Notes on Lattices, Bases and the Reduction Problem. Technical Report 300, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, June 1987.

[137] B. Mishra. *Algorithmic Algebra*. Springer-Verlag, New York, 1993. Texts and Monographs in Computer Science Series.

[138] B. Mishra. Computational real algebraic geometry. In J. O'Rourke and J. Goodman, editors, *CRC Handbook of Discrete and Comp. Geom.* CRC Press, Boca Raton, FL, 1997.

[139] B. Mishra and P. Pedersen. Arithmetic of real algebraic numbers is in *NC*. Technical Report 220, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, Jan 1990.

[140] B. Mishra and C. K. Yap. Notes on Gröbner bases. *Information Sciences*, 48:219–252, 1989.

[141] R. Moenck. Fast computations of GCD's. *Proc. ACM Symp. on Theory of Computation*, 5:142–171, 1973.

[142] H. M. Möller and F. Mora. Upper and lower bounds for the degree of Gröbner bases. In *Lecture Notes in Computer Science*, volume 174, pages 172–183, 1984. (Eurosam 84).

[143] D. Mumford. *Algebraic Geometry, I. Complex Projective Varieties*. Springer-Verlag, Berlin, 1976.

[144] C. A. Neff. Specified precision polynomial root isolation is in *NC*. *J. of Computer and System Sciences*, 48(3):429–463, 1994.

[145] M. Newman. *Integral Matrices*. Pure and Applied Mathematics Series, vol. 45. Academic Press, New York, 1972.

[146] L. Nový. *Origins of modern algebra*. Academia, Prague, 1973. Czech to English Transl., Jaroslav Tauer.

[147] N. Obreschkoff. *Verteilung and Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic, 1963.

[148] C. Ó'Dúnlaing and C. Yap. Generic transformation of data structures. *IEEE Foundations of Computer Science*, 23:186–195, 1982.

[149] C. Ó'Dúnlaing and C. Yap. Counting digraphs and hypergraphs. *Bulletin of EATCS*, 24, October 1984.

[150] C. D. Olds. *Continued Fractions*. Random House, New York, NY, 1963.

[151] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[152] V. Y. Pan. Algebraic complexity of computing polynomial zeros. *Comput. Math. Applic.*, 14:285–304, 1987.

[153] V. Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[154] P. Pedersen. Counting real zeroes. Technical Report 243, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1990. PhD Thesis, Courant Institute, New York University.

[155] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Leipzig, 2nd edition, 1929.

[156] O. Perron. *Algebra*, volume 1. de Gruyter, Berlin, 3rd edition, 1951.

[157] O. Perron. *Die Lehre von den Kettenbrüchen*. Teubner, Stuttgart, 1954. Volumes 1 & 2.

[158] J. R. Pinkert. An exact method for finding the roots of a complex polynomial. *ACM Trans. on Math. Software*, 2:351–363, 1976.

[159] D. A. Plaisted. New *NP*-hard and *NP*-complete polynomial and integer divisibility problems. *Theor. Computer Science*, 31:125–138, 1984.

[160] D. A. Plaisted. Complete divisibility problems for slowly utilized oracles. *Theor. Computer Science*, 35:245–260, 1985.

[161] E. L. Post. Recursive unsolvability of a problem of Thue. *J. of Symbolic Logic*, 12:1–11, 1947.

[162] A. Pringsheim. Irrationalzahlen und Konvergenz unendlicher Prozesse. In *Enzyklopädie der Mathematischen Wissenschaften, Vol. I*, pages 47–146, 1899.

[163] M. O. Rabin. Probabilistic algorithms for finite fields. *SIAM J. Computing*, 9(2):273–280, 1980.

[164] A. R. Rajwade. *Squares*. London Math. Society, Lecture Note Series 171. Cambridge University Press, Cambridge, 1993.

[165] C. Reid. *Hilbert*. Springer-Verlag, Berlin, 1970.

[166] J. Renegar. On the worst-case arithmetic complexity of approximating zeros of polynomials. *Journal of Complexity*, 3:90–113, 1987.

[167] J. Renegar. On the Computational Complexity and Geometry of the First-Order Theory of the Reals, Part I: Introduction. Preliminaries. The Geometry of Semi-Algebraic Sets. The Decision Problem for the Existential Theory of the Reals. *J. of Symbolic Computation*, 13(3):255–300, March 1992.

[168] L. Robbiano. Term orderings on the polynomial ring. In *Lecture Notes in Computer Science*, volume 204, pages 513–517. Springer-Verlag, 1985. Proceed. EUROCAL '85.

[169] L. Robbiano. On the theory of graded structures. *J. of Symbolic Computation*, 2:139–170, 1986.

[170] L. Robbiano, editor. *Computational Aspects of Commutative Algebra*. Academic Press, London, 1989.

[171] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois J. Math.*, 6:64–94, 1962.

[172] S. Rump. On the sign of a real algebraic number. *Proceedings of 1976 ACM Symp. on Symbolic and Algebraic Computation (SYMSAC 76)*, pages 238–241, 1976. Yorktown Heights, New York.

[173] S. M. Rump. Polynomial minimum root separation. *Math. Comp.*, 33:327–336, 1979.

[174] P. Samuel. About Euclidean rings. *J. Algebra*, 19:282–301, 1971.

[175] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Trans. on Math. Software*, 8:277–289, 1982.

[176] W. Scharlau. *Quadratic and Hermitian Forms*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1985.

[177] W. Scharlau and H. Opolka. *From Fermat to Minkowski: Lectures on the Theory of Numbers and its Historical Development*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1985.

[178] A. Schinzel. *Selected Topics on Polynomials*. The University of Michigan Press, Ann Arbor, 1982.

[179] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Lecture Notes in Mathematics, No. 1467. Springer-Verlag, Berlin, 1991.

[180] C. P. Schnorr. A more efficient algorithm for lattice basis reduction. *J. of Algorithms*, 9:47–62, 1988.

[181] A. Schönhage. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Informatica*, 1:139–144, 1971.

[182] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[183] A. Schönhage. Factorization of univariate integer polynomials by Diophantine approximation and an improved basis reduction algorithm. In *Lecture Notes in Computer Science*, volume 172, pages 436–447. Springer-Verlag, 1984. Proc. 11th ICALP.

[184] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1985. Manuscript, Department of Mathematics, University of Tübingen.

[185] A. Schönhage and V. Strassen. Schnelle Multiplikation großer Zahlen. *Computing*, 7:281–292, 1971.

[186] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. of the ACM*, 27:701–717, 1980.

[187] J. T. Schwartz. Polynomial minimum root separation (Note to a paper of S. M. Rump). Technical Report 39, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, February 1985.

[188] J. T. Schwartz and M. Sharir. On the piano movers' problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Appl. Math.*, 4:298–351, 1983.

[189] A. Seidenberg. Constructions in algebra. *Trans. Amer. Math. Soc.*, 197:273–313, 1974.

[190] B. Shiffman. Degree bounds for the division problem in polynomial ideals. *Mich. Math. J.*, 36:162–171, 1988.

[191] C. L. Siegel. *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin, 1988. Notes by B. Friedman, rewritten by K. Chandrasekharan, with assistance of R. Suter.

[192] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin (N.S.) of the AMS*, 4(1):1–36, 1981.

[193] S. Smale. On the efficiency of algorithms of analysis. *Bulletin (N.S.) of the AMS*, 13(2):87–121, 1985.

[194] D. E. Smith. *A Source Book in Mathematics*. Dover Publications, New York, 1959. (Volumes 1 and 2. Originally in one volume, published 1929).

[195] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.

[196] V. Strassen. The computational complexity of continued fractions. *SIAM J. Computing*, 12:1–27, 1983.

[197] D. J. Struik, editor. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, Princeton, NJ, 1986.

[198] B. Sturmfels. *Algorithms in Invariant Theory*. Springer-Verlag, Vienna, 1993.

[199] B. Sturmfels. Sparse elimination theory. In D. Eisenbud and L. Robbiano, editors, *Proc. Computational Algebraic Geometry and Commutative Algebra 1991*, pages 377–397. Cambridge Univ. Press, Cambridge, 1993.

[200] J. J. Sylvester. On a remarkable modification of Sturm's theorem. *Philosophical Magazine*, pages 446–456, 1853.

[201] J. J. Sylvester. On a theory of the syzegetic relations of two rational integral functions, comprising an application to the theory of Sturm's functions, and that of the greatest algebraical common measure. *Philosophical Trans.*, 143:407–584, 1853.

[202] J. J. Sylvester. *The Collected Mathematical Papers of James Joseph Sylvester*, volume 1. Cambridge University Press, Cambridge, 1904.

[203] K. Thull. Approximation by continued fraction of a polynomial real root. *Proc. EUROSAM '84*, pages 367–377, 1984. Lecture Notes in Computer Science, No. 174.

[204] K. Thull and C. K. Yap. A unified approach to fast GCD algorithms for polynomials and integers. Technical report, Courant Institute of Mathematical Sciences, Robotics Laboratory, New York University, 1992.

[205] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[206] B. Vallée. Gauss' algorithm revisited. *J. of Algorithms*, 12:556–572, 1991.

[207] B. Vallée and P. Flajolet. The lattice reduction algorithm of Gauss: an average case analysis. *IEEE Foundations of Computer Science*, 31:830–839, 1990.

[208] B. L. van der Waerden. *Modern Algebra*, volume 2. Frederick Ungar Publishing Co., New York, 1950. (Translated by T. J. Benac, from the second revised German edition).

[209] B. L. van der Waerden. *Algebra*. Frederick Ungar Publishing Co., New York, 1970. Volumes 1 & 2.

[210] J. van Hulzen and J. Calmet. Computer algebra systems. In B. Buchberger, G. E. Collins, and R. Loos, editors, *Computer Algebra*, pages 221–244. Springer-Verlag, Berlin, 2nd edition, 1983.

[211] F. Viète. *The Analytic Art*. The Kent State University Press, 1983. Translated by T. Richard Witmer.

[212] N. Vikas. An $O(n)$ algorithm for Abelian p-group isomorphism and an $O(n \log n)$ algorithm for Abelian group isomorphism. *J. of Computer and System Sciences*, 53:1–9, 1996.

[213] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Trans. on Computers*, 39(5):605–614, 1990. Also, 1988 ACM Conf. on LISP & Functional Programming, Salt Lake City.

[214] H. S. Wall. *Analytic Theory of Continued Fractions*. Chelsea, New York, 1973.

[215] I. Wegener. *The Complexity of Boolean Functions*. B. G. Teubner, Stuttgart, and John Wiley, Chichester, 1987.

[216] W. T. Wu. *Mechanical Theorem Proving in Geometries: Basic Principles*. Springer-Verlag, Berlin, 1994. (Trans. from Chinese by X. Jin and D. Wang).

[217] C. K. Yap. A new lower bound construction for commutative Thue systems with applications. *J. of Symbolic Computation*, 12:1–28, 1991.

[218] C. K. Yap. Fast unimodular reductions: planar integer lattices. *IEEE Foundations of Computer Science*, 33:437–446, 1992.

[219] C. K. Yap. A double exponential lower bound for degree-compatible Gröbner bases. Technical Report B-88-07, Fachbereich Mathematik, Institut für Informatik, Freie Universität Berlin, October 1988.

[220] K. Yokoyama, M. Noro, and T. Takeshima. On determining the solvability of polynomials. In *Proc. ISSAC'90*, pages 127–134. ACM Press, 1990.

[221] O. Zariski and P. Samuel. *Commutative Algebra*, volume 1. Springer-Verlag, New York, 1975.

[222] O. Zariski and P. Samuel. *Commutative Algebra*, volume 2. Springer-Verlag, New York, 1975.

[223] H. G. Zimmer. *Computational Problems, Methods, and Results in Algebraic Number Theory*. Lecture Notes in Mathematics, Volume 262. Springer-Verlag, Berlin, 1972.

[224] R. Zippel. *Effective Polynomial Computation*. Kluwer Academic Publishers, Boston, 1993.

Contents