Algebra 3: algorithms in algebra

Hans Sterk

2003-2004

Contents

Chapter 1

Polynomials, Gröbner bases and Buchberger's algorithm

1.1 Introduction

This chapter deals with the algebraic approach to systems of polynomial equations. Rather than manipulating polynomial equations directly, this approach focuses on studying ideals in polynomial rings and finding generators for these ideals suitable for various types of computations. The suitable sets of generators we are looking for are the so–called Gröbner bases (introduced in Section 1.5). In Section 1.6 we discuss Buchberger's algorithm to construct such Gröbner bases. The algorithm can be seen as a common generalization of the Euclidean algorithm for gcd computations (for polynomials in one variable) and the Gauss–Jordan procedure for solving systems of linear equations.

To analyse ideals we need a bit of the machinery of rings in the context of polynomial rings, and, most significantly, an ordering on the set of monomials in polynomial rings to enable us to generalize division with remainder to polynomials in several variables. These items are explained in Sections 1.2, 1.3 and 1.4.

1.2 Polynomial rings and systems of polynomial equations

1.2.1 Instead of considering a system of poynomial equations

$$f_1 = 0,\ f_2 = 0, \ldots, f_m = 0,$$

it turns out to be more fruitful to study the ideal $(f_1, \ldots, f_m)$ in the corresponding polynomial ring. This section introduces the terminology and notations regarding polynomial rings over a field. Let $k[X_1, \ldots, X_n]$ be a polynomial ring in n indeterminates over the field k.

1.2.2 Definition. A *monomial* is an element of $k[X_1, \ldots, X_n]$ of the form

$$X_1^{m_1} X_2^{m_2} \cdots X_n^{m_n}.$$

The *(multi)degree* of this monomial is the vector $\mathbf{m} = (m_1, \ldots, m_n)$; the *total degree* is the sum $m_1 + \cdots + m_n$ and is often denoted by $|\mathbf{m}|$. If $n = 1$, the notions coincide with the usual notion of degree. We often shorten the notation by writing $X^{\mathbf{m}}$ for $X_1^{m_1} X_2^{m_2} \cdots X_n^{m_n}$.

1.2.3 Every polynomial is a finite k–linear combination of monomials.

The first result states that every ideal in a polynomial ring is finitely generated. This is a consequence of a more general result, which we derive in a moment.

1.2.4 Definition. A ring R is called *noetherian* if every ideal in R is finitely generated.

Here is an equivalent way of phrasing the property.

1.2.5 Lemma. *A ring R is noetherian if and only if every ascending chain of ideals $I_1 \subset I_2 \subset I_3 \subset \cdots$ in R stabilizes (i.e., there is an index m such that $I_m = I_{m+1} = I_{m+2} = \cdots$).*

1.2.6 Theorem. (Hilbert basis theorem) *If R is noetherian, then so is $R[X]$.*

Proof. Suppose $I \subset R[X]$ is not finitely generated. Choose $f_1 \in I \setminus \{0\}$ of minimal degree d_1. Then choose $f_2 \in I \setminus (f_1)$ of minimal degree n_2 (this is possible since f_1 cannot generate I by assumption), etc. Now suppose $f_i = a_{d_i} X^{d_i} + \cdots$ and consider the chain of ideals $(a_1) \subset (a_1, a_2) \subset (a_1, a_2, a_3) \subset \cdots$ in R. This chain stabilizes at some point by the lemma. Let's say that $a_{k+1} \in (a_1, \ldots, a_k)$ and write $a_{k+1} = b_1 a_1 + \cdots + b_k a_k$ for some $b_1, \ldots, b_k \in R$. Then the polynomial

$$g := f_{k+1} - \sum_{i=1}^{k} b_i f_i X^{d_{k+1} - d_i}$$

has, by construction, degree less than the degree of f_{k+1}, but is, like f_{k+1}, not an element of $(f_1, \ldots, f_k)$. This is a contradiction. □

1.2.7 Since a field k has only two ideals, (0) and k itself, every field is a noetherian ring. By applying the Hilbert basis theorem several times we find that the polynomial ring in n indeterminates over a field is noetherian. Every ideal in such a ring is therefore finitely generated.

1.2.8 Corollary. *Every polynomial ring over a field is noetherian. If I is an ideal in such a ring then there exist elements $f_1, \ldots, f_s \in I$ such that $I = (f_1, f_2, \ldots, f_n)$.*

1.2.9 The importance of this result is that every system of polynomial equations in n variables can be replaced by an equivalent finite system. With this observation, the problem of finding the zeros of a system of polynomial equations is equivalent to the problem of finding the common zeros of an ideal. If I is an ideal in $k[X_1, \ldots, X_n]$, then we define $V(I)$, the *zeroset* of I as

$$V(I) = \{(a_1, \ldots, a_n) \in k^n \mid f(a_1, \ldots, a_n) = 0 \text{ for all } f \in I\}.$$

If $I = (f_1, \ldots, f_s)$, then it is easy to see that

$$V(I) = \{(a_1, \ldots, a_n) \in k^n \mid f_i(a_1, \ldots, a_n) = 0,\ i = 1, \ldots, s\}.$$

In particular, if $I = (f_1, \ldots, f_s) = (g_1, \ldots, g_t)$, then the systems of equations

$$f_1 = 0, \ldots, f_s = 0$$

and

$$g_1 = 0, \ldots, g_t = 0$$

are equivalent. With the appearance of ideals, the whole machinery of commutative algebra is at our disposal to analyse their structure and properties.

1.3 Monomial orderings

1.3.1 In this section we discuss various ways to order the monomials of a polynomial ring. This is needed in order to set up a division algorithm. This algorithm imitates the one for polynomials in one variable. For polynomials in one variable X, there is one ordering: $1 < X < X^2 < X^3 < \cdots$ that makes sense for division purposes. In the case of several variables, there are more possibilities.

1.3.2 Definition. A *partial order* on a set S is a relation $\geq$ on S such that

(i) $a \geq a$ for every $a \in S$ (the relation is *reflexive*);

(ii) if $a \geq b$ and $b \geq c$ then $a \geq c$ (the relation is *transitive*);

(iii) $a \geq b$ and $b \geq a$ imply $a = b$ (the relation is *antisymmetric*).

A partial order is called a *total order* if, in addition,

(iv) for all $a, b \in S$, either $a \leq b$ or $b \leq a$.

A total order is called a *well–ordering* if moreover the following holds:

(v) Every nonempty subset T of S contains a smallest element: there is a $t \in T$ such that $t \leq s$ for all $t \in T$.

1.3.3 Remark. In dealing with monomials $X^{\mathbf{a}} = X_1^{a_1} \cdots X_n^{a_n}$, we will often just use the exponent vector $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}_{\geq 0}^n$ instead of the whole monomial.

1.3.4 Definition. (Monomial ordering) A *monomial ordering* on $k[X_1, \ldots, X_n]$ is a well–ordering on the set of monomials $X^{\mathbf{a}}$ such that

$$\mathbf{a} > \mathbf{b} \text{ and } \mathbf{c} \in \mathbb{Z}_{\geq 0}^n \quad \Rightarrow \quad \mathbf{a} + \mathbf{c} > \mathbf{b} + \mathbf{c}.$$

The condition means that the ordering behaves well with respect to multiplication by monomials. In the following we will define three orderings.

1.3.5 Definition. (Lexicographic order) $\mathbf{a} >_{\text{lex}} \mathbf{b}$ (or $X^{\mathbf{a}} >_{\text{lex}} X^{\mathbf{b}}$) if the first nonzero entry from the left in $\mathbf{a} - \mathbf{b}$ is positive. We often abbreviate lexicographic order to 'lex order'.

1.3.6 Definition. (Graded lex order) $\mathbf{a} >_{\text{grlex}} \mathbf{b}$ (or $X^{\mathbf{a}} >_{\text{grlex}} X^{\mathbf{b}}$) if $|\mathbf{a}| > |\mathbf{b}|$ or $|\mathbf{a}| = |\mathbf{b}|$ and $\mathbf{a} >_{\text{lex}} \mathbf{b}$. Graded lex order orders by total degree first and breaks ties using lex order.

1.3.7 Definition. (Graded reverse lex order) $\mathbf{a} >_{\text{grevlex}} \mathbf{b}$ (or $X^{\mathbf{a}} >_{\text{grevlex}} X^{\mathbf{b}}$) if $|\mathbf{a}| > |\mathbf{b}|$ or $|\mathbf{a}| = |\mathbf{b}|$ and the first nonzero entry from the right in $\mathbf{a} - \mathbf{b}$ is negative.

1.3.8 Proposition. *The lexicographic order, graded lex order and graded reverse lex order are monomial orderings.*

Proof. (Sketch for lex order) Most of the conditions to be verified are straightforward. To check that lex order is a well–ordering we use the observation that a total order on $\mathbb{Z}_{\geq 0}^n$ is a well–ordering if and only if every decreasing sequence $\mathbf{a}(1) > \mathbf{a}(2) > \cdots$ terminates. To check this condition for the lex order, one first considers the first (from the left) coordinates in such a sequence $\mathbf{a}(1) > \mathbf{a}(2) > \cdots$; since they all come from $\mathbb{Z}_{\geq 0}$, they become constant from some point onwards. Next consider from that point on the second coordinates, etc. □

1.3.9 Now that we have the notion of an ordering on monomials, we can refine our definitions regarding the terms in a polynomial.

1.3.10 Definition. If $f = \sum_{\mathbf{a}} c_{\mathbf{a}} X^{\mathbf{a}}$ is a polynomial in $k[X_1, \ldots, X_n]$ and $>$ is a monomial ordering, then we define

- the *multidegree* of f to be the maximum degree of the nonzero terms of f;
- the *leading term* $\mathbf{lt}(f)$ of f to be the nonzero term $c_{\mathbf{a}} X^{\mathbf{a}}$ of f of maximum degree and the *leading monomial* to be the monomial $X^{\mathbf{a}}$;
- the *leading coefficient* $lc(f)$ of f to be the coefficient of the leading term of f.

1.4 A division algorithm for polynomials

1.4.1 Given a monomial ordering on a polynomial ring, we can mimic the division algorithm for polynomials in one variable. However, in the general case some complications arise when doing a division. Before we turn to the division algorithm, we remark that division is related to the question of deciding whether a given element belongs to an ideal. In the case of one variable this is clear: if an ideal is generated by $f_1, \ldots, f_s$, say, then it is also generated by the gcd (greatest common divisor) f of these elements. To decide if an element h belongs to the ideal, perform a division with remainder: $h = qf + r$. Then h is in the ideal if and only if $r = 0$.

Fix a monomial ordering on the polynomial ring $k[X_1, \ldots, X_n]$. We will describe how to divide a given polynomial f by the polynomials $f_1, \ldots, f_s$. The result will consist of a list of s 'quotients' $q_1, \ldots, q_s$ and a 'remainder' r. These polynomials will be constructed along the way. First they are all set to 0. In each stage the polynomial f will change: this changing polynomial is called p; at the beginning it is set equal to f.

In each stage the algorithm works roughly as follows.

- Look for the first polynomial among the f_i (starting from f_1) whose leading term divides the leading term of p. If such a division occurs for f_i, then subtract
$$\frac{\mathbf{lt}(p)}{\mathbf{lt}(f_i)} f_i$$
from p and add
$$\frac{\mathbf{lt}(p)}{\mathbf{lt}(f_i)}$$
to the i–th quotient q_i.
- If for no i division occurs, then subtract the term $\mathbf{lt}(p)$ from p and add this term to the remainder r.

Since in each step the leading term of p decreases, this process must terminate. Upon termination, we have an equality

$$f = q_1 f_1 + \cdots + q_s f_s + r,$$

where $r = 0$ or no term of r is divisible by any of the leading terms of the f_i $(i = 1, \ldots, s)$. The remainder and the quotients need not be unique (as in the one variable case). In fact, the results even depend in general on the order of the f_i. We will come back to these matters soon.

1.5 Monomial ideals and Gröbner bases

In this section we assume that a monomial ordering is specified on the polynomial ring $k[X_1, \ldots, X_n]$. Before we can state the definition of a Gröbner basis, we need some preparations involving monomials.

1.5.1 Definition. (Monomial ideals) A *monomial ideal* in $k[X_1, \ldots, X_n]$ is an ideal generated by monomials. Note that it is harmless to replace 'monomial' by 'term', since monomials and terms differ by a constant.

1.5.2 Lemma. *Let I be a monomial ideal generated by the monomials $X^{\mathbf{a}}$, $\mathbf{a} \in A$.*

(a) $f \in I \Leftrightarrow$ every term/monomial of f is in I.

(b) $X^{\mathbf{b}} \in I \Leftrightarrow X^{\mathbf{a}} \mid X^{\mathbf{b}}$ for some $\mathbf{a} \in A$.

(c) $I = (X^{\mathbf{a}(1)}, X^{\mathbf{a}(2)}, \ldots, X^{\mathbf{a}(m)})$ for some m.

Proof. (a) and (b) follow from writing out an expressions for f and $X^{\mathbf{b}}$, respectively. Item (c): Apply the Hilbert basis theorem to get finitely many (possibly non–monomial) generators. Then apply the previous items to replace these generators by finitely many monomials from the original generating set. □

1.5.3 Definition. For an ideal $(0) \neq I \subset k[X_1, \ldots, X_n]$ we let $\mathbf{lt}(I)$ be the set of leading terms in I:

$$\mathbf{lt}(I) = \{\mathbf{lt}(f) \mid 0 \neq f \in I\}.$$

The *leading term ideal* is the ideal $(\mathbf{lt}(I))$ generated by $\mathbf{lt}(I)$.

1.5.4 Lemma. *Let $(0) \neq I \subset k[X_1, \ldots, X_n]$ be an ideal, then $(\mathbf{lt}(I))$ is a monomial ideal and there exist finitely many elements $f_1, \ldots, f_s \in I$ such that $\mathbf{lt}(f_1), \ldots, \mathbf{lt}(f_s)$ generate this ideal.*

Proof. This follows from the definition and the previous lemma. □

1.5.5 Definition. (Gröbner basis) A finite subset $\{g_1, \ldots, g_s\}$ of the ideal I is called a *Gröbner basis* of I if the leading term ideal is generated by the leading terms of the g_i:

$$(\mathbf{lt}(g_1), \ldots, \mathbf{lt}(g_s)) = (\mathbf{lt}(I)).$$

Here is the first important result about Gröbner bases.

1.5.6 Theorem. *Let $I \neq (0)$ be an ideal in the polynomial ring $k[X_1, \ldots, X_n]$.*

(a) The ideal I has a Gröbner basis.

(b) A Gröbner basis $\{g_1, \ldots, g_s\}$ of I generates I (as an ideal):

$$(g_1, \ldots, g_s) = I.$$

(c) If $\{g_1, \ldots, g_s\}$ is a Gröbner basis for I, then division by $g_1, \ldots, g_s$ leaves a unique remainder r independent of the order of the g_i. In fact, r is characterized as the unique polynomial such that

(i) $r = 0$ or no term of r is divisible by any of the leading terms of the g_i ($i = 1, \ldots, s$);

(ii) $f - r \in I$.

Proof. (a) follows from Lemma 1.5.2 since the leading term ideal $(\mathbf{lt}(I))$ of I is generated by the leading terms of elements $\neq 0$ of I.

(b) It is obvious that $(g_1, \ldots, g_s) \subset I$ as all the g_j are in I and therefore also the ideal generated by them. For the converse inclusion let $f \in I$ and use division with remainder to write $f = q_1 g_1 + \cdots + q_s g_s + r$, where either $r = 0$ or no term of r is divisible by any of the leading terms $\mathbf{lt}(g_j)$ $(j = 1, \ldots, s)$. Suppose $r \neq 0$. From $r = f - (q_1 g_1 + \cdots + q_s g_s)$ we conclude that $r \in I$ and so $\mathbf{lt}(r) \in (\mathbf{lt}(I))$. Again by Lemma 1.5.2 we find that $\mathbf{lt}(r)$ is divisible by one of the leading terms $\mathbf{lt}(g_j)$ $(j = 1, \ldots, s)$, a contradiction. So r must be 0 and $f \in (g_1, \ldots, g_s)$.

(c) Division with remainder as discussed before shows that the remainder r satisfies the two properties mentioned. So existence of such an r is clear. As for uniqueness, if $\tilde{r}$ also satisfies the two properties, then $r - \tilde{r} \in I$ (subtract $f - \tilde{r}$ and $f - r$) and so the leading term of $r - \tilde{r}$ belongs to $(\mathbf{lt}(I))$, if $r - \tilde{r} \neq 0$. In this case, as in (b), we arrive at a contradiction, because no term of $r - \tilde{r}$ is divisible by any of the $\mathbf{lt}(g_j)$, wheras Lemma 1.5.2 implies that one of them is. So $r - \tilde{r} = 0$, i.e., $r = \tilde{r}$. □

1.6 Buchberger's algorithm

Apart from the definition given in the previous section, there are various ways of characterizing Gröbner bases. Some of these characterizations together with further properties of Gröbner bases lead to an algorithmic approach to computing Gröbner bases: *Buchberger's algorithm.*

We begin with an application of the division algorithm. It settles the 'ideal membership problem'.

1.6.1 Proposition. (Ideal membership test) *Let G be a Gröbner basis for the ideal $I \subset k[X_1, \ldots, X_n]$ and let $f \in k[X_1, \ldots, X_n]$. Then $f \in I$ if and only if the remainder on division of f by G is zero.*

Proof. The implication 'If' is trivial. Conversely, if $f \in I$, then $f = f + 0$ satisfies the properties of Theorem 1.5.6, so 0 is the remainder upon division. □

1.6.2 In a given Gröbner basis there may be elements of redundancy. For example, if $G = \{g_1, \ldots, g_s\}$ is a Gröbner basis for I and if $\mathbf{lt}(f)$ is contained in the ideal $(\mathbf{lt}(G - \{f\})$ for $f \in G$, then $G - \{f\}$ is also a Gröbner basis for I. Given the definition of Gröbner basis, this is almost a triviality: Since

$\mathbf{lt}(f) \in (\mathbf{lt}(G-\{f\})$, we find $(\mathbf{lt}(G-\{f\}) = (\mathbf{lt}(G)) = (\mathbf{lt}(I))$. The resulting equality of the first and third term imply that $G - \{f\}$ is a Gröbner basis.

The following definitions and results are intended to produce 'unique' Gröbner bases in some sense.

1.6.3 Definition. A *minimal Gröbner basis* for an ideal I is a Gröbner basis G for I satisfying

(i) $lc(f) = 1$ for all $f \in G$;

(ii) $\mathbf{lt}(f) \notin (\mathbf{lt}(G - \{f\}))$ for all $f \in G$.

A *reduced Gröbner basis* for an ideal I satisfies (i) and the following condition, which is stronger than (ii):

(ii') no (nonzero) term of f is in $(\mathbf{lt}(G - \{f\}))$ for all $f \in G$.

1.6.4 Theorem. *Every nonzero ideal $I \subset k[X_1, \ldots, X_n]$ has a unique reduced Gröbner basis (for a given monomial ordering).*

Proof. Given a Gröbner basis, it is easy to construct a minimal one: just apply the observation we made above in 1.6.2 and replace leading coefficients.

To construct a reduced one is less obvious. □

1.6.5 (Equality of ideals) Once reduced Gröbner bases can be effectively computed, one has a method to decide whether two ideals are equal: they are equal if and only if they the same reduced Gröbner basis.

1.6.6 Definition. For nonzero polynomials $f, g \in k[X_1, \ldots, X_n]$ of multidegree $\mathbf{a}$ and $\mathbf{b}$, respectively, we define their *S–polynomial* as the polynomial

$$S(f,g) = \frac{X^{\mathbf{c}}}{\mathbf{lt}(f)} \cdot f - \frac{X^{\mathbf{c}}}{\mathbf{lt}(g)} \cdot g,$$

where $\mathbf{c} = (\max(a_1, b_1), \ldots, \max(a_n, b_n))$. The monomial $X^{\mathbf{c}}$ is called the *least common multiple* (lcm) of the leading terms of f and g.

1.6.7 S–polynomials are vital ingredients in Buchberger's algorithm for computing Gröbner bases. Here is a characterization of Gröbner bases involving S–polynomials.

1.6.8 Theorem. *A basis $G = \{g_1, \ldots, g_s\}$ for the nonzero ideal I is a Gröbner basis for I if and only if the remainder on division of $S(g_i, g_j)$ by G is zero for all $i \neq j$.*

Proof. The proof of this theorem is not deep, but quite elaborate. It comes down to an analysis of the relation between S–polynomials and cancellation of leading terms. The proof of 'only if' is a trivial consequence of the division algorithm. A proof of the implication 'if' will be sketched below. □

1.6.9 A first indication of the usefulness of S–polynomials is given in the following lemma.

1.6.10 Lemma. *Let $f = \sum_i c_i X^{\mathbf{a}_i} f_i$ be a sum whose multidegree is less than $\mathbf{d}$. If $\mathbf{a}_i + \text{multdeg}(f_i) = \mathbf{d}$ for all i, then f can be written as a sum*

$$f = \sum_{j,k} c_{j,k} X^{\mathbf{d}-\mathbf{c}_{j,k}} S(f_j, f_k),$$

where $X^{\mathbf{c}_{j,k}}$ is the least common multiple of the leading monomials of f_j and f_k, and where the multidegree of each term in the sum is less than $\mathbf{d}$.

Proof (sketch). We will demonstrate the proof in the case of two polynomials f_1 and f_2 of multidegrees $\mathbf{b}_1$ and $\mathbf{b}_2$, respectively. For simplicity we'll also assume that the leading terms of f_1, f_2 have leading coefficient 1. Since the multidegree of f is less than $\mathbf{d}$, we have $c_1 + c_2 = 0$. So we can (re)write

$$\begin{aligned} f &= c_1(X^{\mathbf{a}_1} f_1 - X^{\mathbf{a}_2} f_2) \\ &= c_1\left(\tfrac{X^{\mathbf{d}}}{X^{\mathbf{b}_1}} f_1 - \tfrac{X^{\mathbf{d}}}{X^{\mathbf{b}_2}} f_2\right) \\ &= c_1 X^{\mathbf{d}-\mathbf{c}}\left(\tfrac{X^{\mathbf{c}}}{X^{\mathbf{b}_1}} f_1 - \tfrac{X^{\mathbf{c}}}{X^{\mathbf{b}_2}} f_2\right) \\ &= c_1 X^{\mathbf{d}-\mathbf{c}} S(f_1, f_2), \end{aligned}$$

as desired. In the third equality we use that $X^{\mathbf{c}}$ divides $X^{\mathbf{d}}$; this is the case since $X_i^{\mathbf{a}}$ divides $X^{\mathbf{d}}$ for $i = 1, 2$. For the general proof we refer to the exercises. □

1.6.11 (Proof of Theorem 1.6.8) Again we restrict to the case that the basis consists of two elements g_1, g_2. Among the (possibly many) ways to write $f = h_1 g_1 + h_2 g_2$, choose one in which the maximum of the multidegrees of $h_1 g_1$ and $h_2 g_2$ is minimal, say d. Of course, multdegree$(f) \leq d$.

Now suppose first that the multidegree of f is strictly less than d. Then the leading terms of $h_1 g_1$ and $h_2 g_2$ both have multidegree d and they cancel

each other. By Lemma 1.6.10, the sum $\mathbf{lt}(h_1)g_1 + \mathbf{lt}(h_2)g_2$ can be rewritten using the S–polynomial $S(g_1, g_2)$:

$$\mathbf{lt}(h_1)g_1 + \mathbf{lt}(h_2)g_2 = cX^{d-c_{12}}S(g_1, g_2),$$

where c_{12} is the exponent of the least common multiple of g_1, g_2 and where the multidegree of the terms on the right–hand side is less than d. By hypothesis, the S–polynomial $S(g_1, g_2)$ can be written as a sum $a_1g_1 + a_2g_2$ with multidegrees of the two terms bounded by the multidegree of $S(g_1, g_2)$. Then use this to rewrite the original expression for f in such a way that the degree d goes down. This contradicts our assumption. So, from $f = h_1g_1 + h_2g_2$, we deduce that the multidegree of f equals the multidegree of at least one of the h_ig_i. But this implies that $\mathbf{lt}(f)$ is divisible by the leading term of the corresponding g_i. This shows that $\mathbf{lt}(f) \in (\mathbf{lt}(g_1), \mathbf{lt}(g_1))$. □

1.6.12 (Buchberger's algorithm I) A first version of Buchberger's algorithm is easily described using the above S–polynomials. It is primitive in the sense that no care is taken of efficiency matters.

We start with a basis $(f_1, \ldots, f_t)$ of the nonzero ideal I and transform this basis stepwise into a Gröbner basis. In each step, we find an intermediate finite basis G' for the ideal, form all the possible S–polynomials of elements in G'. If division of such an S–polynomial by G' leaves a nonzero remainder, we add this remainder to our intermediate basis. If all these remainders are zero, we stop and output the basis found so far; by Theorem 1.6.8 it is a Gröbner basis. Of course, we need to show that the algorithm terminates and then produces a Gröbner basis. Here is the algorithm schematically:

`Input`: $F = (f_1, \ldots, f_t)$
`Output`: a Gröbner basis $g_1, \ldots, g_s$

$G = F$
`repeat`

  $G' = G$
  For each pair p, q in G'
    `do` compute the S–polynomial $S(p, q)$ and its remainder $r(p, q)$
      upon division by G'
    `if` $r(p, q) \neq 0$, `then` $G := G \cup \{r(p, q)\}$
`until` $G = G'$

1.6.13 (Termination) To show that the algorithm terminates and upon termination produces a Gröbner basis, consider what happens if a nonzero remainder

R of an S–polynomial is added to G: $G' = G \cup \{R\}$. Then the leading term of R is not divisible by any of the leading terms of the elements in G and so the ideal $(\mathbf{lt}(G))$ is strictly contained in the ideal $(\mathbf{lt}(G'))$ (here we use Lemma 1.5.2 again). Since every ascending chain of ideals must eventually become constant, the algorithm must terminate.

When the algorithm terminates in stage G, say, all remainders of S–polynomials upon division by G are 0, and therefore G must be a Gröbner basis by Theorem 1.6.8.

1.6.14 Many improvements can be made, but we refrain from doing this here. In the computer algebra packages Mathematica and Maple, versions of Buchberger's algorithm are available. From easy examples, it already becomes claer that doing computations by hand is a very unpleasant task. So do try to use computer algebra packages for such computations.

Chapter 2

Applications

2.1 Elimination

For systems of linear equations, the Gauss–elimination algorithm does exactly what the term suggests: in the process variables are eliminated from the consecutive equations. A similar result holds for Gröbner bases if you use the lex order.

2.1.1 Theorem. (Elimination) *Let G be a Gröbner basis for the ideal I in $k[X_1, \ldots, X_n]$ with respect to lex order, where $X_1 > \cdots > X_n$. For $j = 1, \ldots, n$, let $G_j = G \cap k[X_{j+1}, \ldots, X_n]$. Then G_j is a Gröbner basis for the ideal $I_j = I \cap k[X_{j+1}, \ldots, X_n]$.*

Proof. If $g \notin G_j$, then the leading term of g must involve at least one of the variables $X_1, \ldots, X_j$: otherwise, the leading term would be in $k[X_{j+1}, \ldots, X_n]$, and consequently, since we are using lex order with $X_1 > \cdots > X_n$, all the other terms of g must be in $k[X_{j+1}, \ldots, X_n]$. This would imply that $g \in k[X_{j+1}, \ldots, X_n]$, a contradiction.

Suppose $G_j = (g_1, \ldots, g_m) \subset G = (g_1, \ldots, g_t)$ and let $f \in I_j$. Then division by G yields an expression $f = q_1 g_1 + \cdots + q_t g_t$. But since $f, g_1, \ldots, g_m$ do not involve the variables $X_1, \ldots, X_j$, the quotients $q_{m+1}, \ldots, q_t$ must all be 0 (for instance, the leading term of g_{m+1} cannot divide the leading term of f by what we remarked above).

So $f \in (g_1, \ldots, g_m)$ and division by G comes down to division by G_j. Now every S–polynomial of two polynomials in G_j belongs to I_j (check!) and the remainder upon division by G_j equals the remainder upon division by G. The last mentioned remainder is 0 because G is a Gröbner basis (see Theorem 1.6.8). □

2.1.2 Example. Suppose we have a curve in k^2 described parametrically by

$$x = t^2, \quad y = t^3.$$

To find an equation for this curve, we want to eliminate t. In this example this is quite easy to do 'by hand': $x^3 = y^2$. But in the process of computing a Gröbner basis this equation occurs necessarily as a by–product. Fix the lex order with $t > x > y$ and consider the ideal $I = (f_1 = t^2 - x, f_2 = t^3 - y)$. Then the leading term of $S(f_1, f_2) = -(tx - y)$ is not divisible by the leading terms of f_1 and f_2, so we add $f_3 = tx - y$ to the generators of the ideal: $I = (f_1, f_2, f_3)$. Next we compute $S(f_1, f_3) = ty - x^2$. Again, the leading term is not divisible by the leading terms of f_1, f_2, f_3. So we add $f_4 = ty - x^2$ to the generators and turn to $S(f_1, f_4) = tx^2 - xy$. Since this equals $(tx - y)x$, the remainder upon division by $f_1, \ldots, f_4$ is 0; the same holds for $S(f_2, f_3) = t^2y - xy$ as we see from writing it as $(t^2 - x)y$. Then we turn to $S(f_2, f_4) = t^2x^2 - y^2$. Upon division we get $t^2x^2 - y^2 = (t^2 - x)x^2 + (x^3 - y^2)$, leaving $f_5 = x^3 - y^2$ as remainder. Further S–polynomials yield no new generators, so we obtain the following Gröbner basis for I:

$$f_1 = t^2 - x,\ f_2 = t^3 - y,\ f_3 = tx - y,\ f_4 = ty - x^2,\ f_5 = x^3 - y^2.$$

The basis is not reduced, since the leading term of f_2 is still divisible by the leading term of f_1, so we may leave out f_2. (In fact, to speed up the computation we should have done this as soon as we had f_3 at our disposal.) This leaves us with the reduced Gröbner basis

$$f_1 = t^2 - x,\ f_3 = tx - y,\ f_4 = ty - x^2,\ f_5 = x^3 - y^2.$$

The elimination theorem states that $I \cap k[x, y] = (x^3 - y^2)$. So every polynomial $g(x, y)$ that becomes the zero–polynomial upon substituting $x = t^2$ and $y = t^3$ must be a multiple of $x^3 - y^2$.

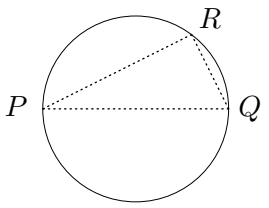
There is a subtle problem here: from the above we see that the zeroset $V(x^3 - y^2)$ contains the set of points of the form (t^2, t^3), but it could be that this last set is strictly contained in the first set. This is one of the reasons for a more detailed analysis of sets of the form $V(I)$. This is done extensively in algebraic geometry.

2.1.3 Other applications of elimination are discussed in Section 2.4 and in Chapter **??**.

2.2 Geometry theorem proving: first glimpse

2.2.1 The framework of ideals and Gröbner bases can be used to explore classical geometry. The set–up works (in principle) when both the hypotheses of a geometric situation and the statement about it can be expressed as polynomial equations. Before such equations can be deduced, a coordinate frame has to be chosen, and the various objects have to be described in terms of these coordinates.

2.2.2 Example. To illustrate this in a simple example, consider the following situation. Draw a circle C and draw a line through the center of C. The points of intersection of the line with C are denoted by P and Q. Choose a third point R on the circumference of the circle. The claim is that PR and QR are perpendicular.



A circle with radius $r > 0$ is described by $x^2 + y^2 - r^2 = 0$. Two antipodal points on the circumference are given by $P = (-r, 0)$ and $Q = (r, 0)$. The third point on the circumference is $R = (u, v)$. Since R is supposed to be on the circle, this produces one hypothesis:

$$\text{Hypothesis: } u^2 + v^2 - r^2 = 0.$$

The statement to be proved is that $(u+r, v)$ and $(u-r, v)$ are perpendicular. Using the standard innerproduct, this means:

$$\text{Thesis: } (u + r) \cdot (u - r) + v \cdot v = 0.$$

Now the thesis is easily seen to reduce to $u^2 + v^2 - r^2 = 0$, an equality that holds because of the assumption we made.

In terms of ideals, we phrase this as follows. The hypothesis describes the subset $\{(u, v, r) \in k^3 \mid u^2+v^2-r^2 = 0\}$, i.e., the vanishing locus $V(I)$ of the ideal $I = (u^2 + v^2 - r^2)$. Now we want our thesis to hold on this subset, i.e., we want the polynomial $T = (u + r) \cdot (u - r) + v \cdot v$ to vanish on $V(I)$.

A sufficient condition for T to vanish on $V(I)$ is that it belongs to I. This condition is trivially satisfied in our situation.

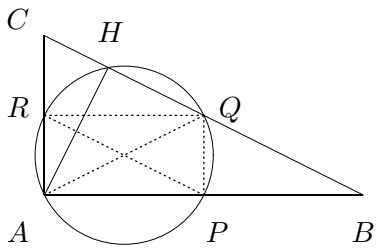
It is also trivial to conclude that an arbitrary point $R = (u, v)$ such that the two segments PR and QR are perpendicular lies on the circle with radius r passing through P and Q.

2.2.3 In general, if the hypotheses are described by the polynomials $h_1, \ldots, h_l$ and the thesis is described by the polynomial T, then the thesis holds if $T \in (h_1, \ldots, h_l)$. This way the problem is shifted to the ideal membership problem.

Unfortunately, this condition is only sufficient and not necessary. For example, if $T \notin I$ but $T^2 \in I$, then T still vanishes on $V(I)$. A precise statement concerning this problem is given in the next section.

The following example shows another type of difficulty, which can arise if one deals with geometry statements.

2.2.4 Example. Consider a right triangle ABC, whose angle at A is $90°$. The foot of the altitude from A on BC is called H. The statement is that H and the midpoints P (of AB), Q (of BC), R (of AC) of the three sides lie on a circle (this is called the *circle theorem of Apollonius*).



Before we go into a proof using Gröbner bases, we sketch the classical approach. The vertex A and the three midpoints P, Q, R form a rectangle and lie on a circle (AQ and PR pass through the center of the circle) by the previous example. Again by the previous example, since the triangle AHQ has a right angle at H, the point H is on the circle with diameter AQ.

Now we turn to the description in terms of polynomials. We choose coordinates in such a way that A is in the origin, that $B = (2x, 0)$ and $C = (0, 2y)$. Then $P = (x, 0)$, $Q = (x, y)$ and $R = (0, y)$. The circle passing

through P, Q, R has equation

$$(X - \frac{1}{2}x)^2 + (Y - \frac{1}{2}y)^2 - \frac{1}{4}(x^2 + y^2) = 0.$$

(For the sake of the example, we assume this as an 'obvious' step.) Now the point $H = (p, q)$ satisfies two conditions:

a) $(p, q) \perp (-x, y)$, i.e., $yq - px = 0$;

b) (p, q) is on the line BC with equation $yX + xY - 2xy = 0$ yielding $yp + xq - 2xy = 0$.

Therefore our hypotheses ideal becomes

$$\text{Hypotheses: } (yq - px, yp + xq - 2xy).$$

Next we compute a Gröbner basis with respect to lex order with $x > y > p > q$ to find

$$G := [\, -y\,p - x\,q + 2\,x\,y, -y\,q + p\,x, -y\,q^2 - y\,p^2 + 2\,y^2\,q\,].$$

We want to check whether $p^2 - xp + q^2 - yq$ is in the ideal. Doing a division with remainder shows that this is not the case. But we do come close. Consider

$$-q(-y\,q + p\,x) - p(-y\,p - x\,q + 2\,x\,y) = p^2y + q^2y - 2xyp = (p^2 + q^2 - 2xp)y.$$

Adding $-y(y\,q - p\,x)$ from the ideal yields $(p^2 + q^2 - yq - xp)y$. This is the one we need, except for a factor y. We would like to cancel out the factor y, assuming that it is never 0. Algebraically, this can be done as follows: extend the polynomial ring to $k[x, y, p, q, t]$ and extend the ideal to

$$(yq - px, yp + xq - 2xy, 1 - yt) \subset k[x, y, p, q, t].$$

This time we have more luck: $p^2 + q^2 - yq - xp$ turns out to belong to this ideal: a Gröbner basis (with respect to lex order, $x > y > p > q > t$) for the ideal

$$I_1 := (yq - px, yp + xq - 2xy, 1 - yt)$$

is

$$[-y\,p - x\,q + 2\,x\,y, 2\,p\,x - p^2 - q^2, t\,x\,q - 2\,x + p, -q^2 + 2\,y\,q - p^2,$$
$$-1 + y\,t, -2\,q + t\,q^2 + t\,p^2],$$

and the remainder upon division turns out to be 0, confirming that $p^2 + q^2 - yq - xp$ belongs to the new ideal. The interpretation is that we have to assume that $y \neq 0$ to get a valid statement. This is not surprising once you realize that in classical geometry the pictures drawn often implicitly assume that certain quantities are nonzero.

Note also that $f = f(1 - yt) + fyt$ with both terms on the right-hand side in the extended ideal once fy belongs to I. This is the algebraic reason why the above works.

2.2.5 The example shows that Gröbner bases can be of help in proving geometry theorems, but also that things are not fully automatic. The reader has lots of choice in assigning coordinates, but may run into various types of problems. One of them is that the theorem one wants to prove tacitly requires the situation to be 'nondegenerate'.

The various projects with this course deal with some of these aspects.

2.3 The Nullenstellensatz

2.3.1 There is one aspect in the above considerations that we would like to elaborate on, since it is related to the famous Hilbert Nullstellensatz, a cornerstone of modern algebraic geometry. It works only over algebraically closed fields, like $\mathbb{C}$. Usually the theorem is split in two parts. In the second statement we come across the notion of the *radical* of an ideal. Given an ideal I in a ring R, the radical, denoted by $\sqrt{I}$, is the ideal $\{f \in R \mid f^N \in I \text{ for some } N > 0\}$. It is straightforward to check that $\sqrt{I}$ is indeed an ideal.

2.3.2 Theorem. (Weak Nullstellensatz) *If I is an ideal in the polynomial ring $k[X_1, \ldots, X_n]$ over an algebraically closed field k such that $V(I) = \emptyset$, then $I = k[X_1, \ldots, X_n]$.*

Proof. We refer to the exercises for the proof of this theorem. □

2.3.3 Theorem. (Hilbert's Nullstellensatz) *Let I be an ideal in $k[X_1, \ldots, X_n]$, where k is algebraically closed. If $f \in k[X_1, \ldots, X_n]$ vanishes identically on the set $V(I)$, then $f \in \sqrt{I}$, i.e., there exists an $m > 0$ such that $f^m \in I$.*

Proof. Let $I = (f_1, \ldots, f_m)$ and assume for simplicity that $f \neq 0$. Consider the ideal $J = I + (1 - fY) \subset k[X_1, \ldots, X_n, Y]$. If $(a_1, \ldots, a_n, a_{n+1}) \in k^{n+1}$ is in $V(J)$, then $f_1(a_1, \ldots, a_n) = \cdots = f_m(a_1, \ldots, a_n) = 0$, so that $(a_1, \ldots, a_n) \in V(I)$ and $1 - f(a_1, \ldots, a_n)a_{n+1} = 0$, so that $f(a_1, \ldots, a_n) \neq 0$. This contradicts the assumption that f vanishes on $V(I)$. So $V(J) = \emptyset$.

By the Weak Nullstellensatz $J = k[X_1, \ldots, X_n, Y]$ and, consequently, there exists a relation of the form

$$1 = g_1 f_1 + \cdots + g_m f_m + g(1 - fY)$$

for some $g, g_1, \ldots, g_m \in k[X_1, \ldots, X_n, Y]$. Now substitute $1/f$ for Y in the above relation to find a relation of the form

$$1 = g_1(X_1, \ldots, X_n, 1/f) f_1 + \cdots + g_m(X_1, \ldots, X_n, 1/f) f_1.$$

Multiplying through by a sufficiently high power of f, we find

$$f^N = \tilde{g}_1 f_1 + \cdots + \tilde{g}_m f_m$$

as claimed. □.

2.3.4 The weak Nullstellensatz generalizes the fact that over an algebraically closed field every nonconstant polynomial in one variable has (at least) one zero. Let $I \neq (0)$ be an ideal in $k[X]$. Then I is generated by one element, say $I = (f)$. If $I \neq k[X]$, i.e., if f is not a constant, then the weak Nullstellensatz claims that f has a zero.

2.4 Algebraic numbers

2.4.1 In this section we discuss algebraic numbers and a way to compute their minimal polynomials.

2.4.2 Definition. Let $\alpha \in \mathbb{C}$. Then α is called *algebraic* if there is a nonconstant polynomial $f(X) \in \mathbb{Q}[X]$ such that $f(\alpha) = 0$. The subset of algebraic numbers is denoted by $\mathbb{A}$. In lemma 2.4.11 we show that $\mathbb{A}$ is in fact a subfield of $\mathbb{C}$.

2.4.3 Example. Every rational number r is of course algebraic: r is a zero of $X - r \in \mathbb{Q}[X]$. The nonrational number $\sqrt{2}$ is algebraic, since it is a zero of $X^2 - 2 \in \mathbb{Q}[X]$.

The complex number $\alpha = \sqrt{2} + i$ is also algebraic. To find a polynomial f such that $f(\alpha) = 0$, we proceed as follows. From $\alpha - i = \sqrt{2}$ we deduce that $(\alpha - i)^2 = 2$. Working out this expression and rewriting a bit, we find

$$\alpha^2 - 3 = -2i\alpha.$$

Squaring again yields

$$(\alpha^2 - 3)^2 = -4\alpha^2.$$

From this equality we conclude that α is a zero of $X^4 - 2X^2 + 9 \in \mathbb{Q}[X]$.

The numbers e and π are known to be non–algebraic.

2.4.4 Given $\alpha \in \mathbb{C} \setminus \{0\}$, the ideal $I_\alpha = \{f \in \mathbb{C}[X] \mid f(\alpha) = 0\}$ is generated by one element $p(X)$. If $p(X)$ is not the zero polynomial, then p is necessarily irreducible, for if $p(X) = p_1(X)\, p_2(X)$, then from $p_1(\alpha)\, p_2(\alpha) = 0$ it follows that either p_1 or p_2 (or both) belongs to I_α, so that one of the two factors is constant. The unique polynomial with leading coefficient 1 that generates I_α is called the *minimal polynomial* of α.

The case $I_\alpha = (0)$ corresponds to the situation that α is not an algebraic number.

2.4.5 Example. The minimal polynomial of $i + \sqrt{2}$ is $p(X) = X^4 - 2X^2 + 9$. In the previous example we noted that α is a zero of p. It remains to show that p is irreducible in $\mathbb{Q}[X]$. The four roots of the polynomial are $\pm i \pm \sqrt{2}$ and it is easy to see from these roots that any linear or quadratic factor of p is not in $\mathbb{Q}[X]$.

2.4.6 Given an algebraic number $\alpha \in \mathbb{C}$, the field $\mathbb{Q}(\alpha)$ consists of all expressions of the form

$$\frac{g(\alpha)}{h(\alpha)},$$

where g, h are polynomials and where $h(\alpha) \neq 0$. The relation with the ideal I_α is as follows. Define the map $F : \mathbb{Q}[X] \to \mathbb{Q}(\alpha)$ by $q(X) \mapsto q(\alpha)$. Then the kernel is precisely I_α and we get an induced injective map

$$\overline{F} : \mathbb{Q}[X]/I_\alpha \to \mathbb{Q}(\alpha).$$

Since I_α is generated by an irreducible polynomial, the ideal is maximal and the quotient is a field. Therefore, the image is also a field contained in $\mathbb{Q}(\alpha)$. But this image contains α (the image of the class of X). As $\mathbb{Q}(\alpha)$ is the smallest field containing α, this implies that $\overline{F}$ is an isomorphism. It also implies that $\mathbb{Q}(\alpha) = \mathbb{Q}[\alpha]$. The argument given here is a special case of the fact that $k[X]/(f(X))$ is a field and can be identified with $k(\overline{X})$, where $\overline{X}$ is the class of X, if $f(X)$ is irreducible and k is a field.

Now suppose $I_\alpha = (p(X))$ where $p(X)$ has degree n. As a $\mathbb{Q}$–vectorspace, $\mathbb{Q}[X]/I_\alpha$ has the basis $1, X, X^2, \ldots, X^{n-1}$. In terms of $\mathbb{Q}[\alpha]$ this means that $1, \alpha, \ldots, \alpha^{n-1}$ is a $\mathbb{Q}$–basis of $\mathbb{Q}[\alpha]$. The integer n is called the *degree of the extension* $\mathbb{Q} \subset \mathbb{Q}(\alpha)$ and is often denoted by $[\mathbb{Q}(\alpha) : \mathbb{Q}]$. It is also common usage to denote the extension as $\mathbb{Q}(\alpha) : \mathbb{Q}$.

In general, if $L : K$ is an extension of fields, we denote by $[L : K]$ the dimension of L as K–vectorspace. Again, this dimension is called the *degree* of the extension.

The discussion shows the following proposition:

2.4.7 Proposition. *If α is algebraic with minimal polynomial of degree n, then $\mathbb{Q}(\alpha) = \mathbb{Q}[\alpha]$, and this field is a finite dimensional $\mathbb{Q}$–vectorspace of dimension n, with $\mathbb{Q}$–basis $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$.*

2.4.8 The equality $\mathbb{Q}[\alpha] = \mathbb{Q}(\alpha)$ implies that every element of the field $\mathbb{Q}(\alpha)$ can be expressed as a polynomial in α. For instance, if $\alpha^2 \neq -1$, there are rationals $a_0, \ldots, a_{n-1}$ such that

$$1/(\alpha^2 + 1) = a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1},$$

although we have not indicated a way to compute these coefficients.

If α and β are bot algebraic, then a similar equality,

$$\mathbb{Q}(\alpha, \beta) = \mathbb{Q}[\alpha, \beta],$$

holds. The proof of this statement is similar to the proof in the case of one algebraic number and is based on considering the map $\mathbb{Q}(\alpha)[X] \to \mathbb{Q}(\alpha)(\beta) = \mathbb{Q}(\alpha, \beta)$, determined by sending X to β.

In general, if $\alpha_1, \ldots, \alpha_n$ are algebraic, then $\mathbb{Q}(\alpha_1, \ldots, \alpha_n) = \mathbb{Q}[\alpha_1, \ldots, \alpha_n]$.

2.4.9 Theorem. *Let $L : K$ be subfields of $\mathbb{C}$ and suppose $\dim_{\mathbb{Q}}(K) = m$ and $\dim_K(L) = n$. Then $\dim_{\mathbb{Q}}(L) = mn$.*

Proof. Suppose $\beta_1, \ldots, \beta_n$ is a K–vectorspace basis of L and that $\alpha_1, \ldots, \alpha_m$ is a $\mathbb{Q}$–vectorspace basis of K. We claim that the set of mn products $\alpha_i\,\beta_j$ is a $\mathbb{Q}$–vectorspace basis of L.

First we show that the set spans L over $\mathbb{Q}$. If $\beta \in L$, then there exist elements $\lambda_1, \ldots, \lambda_n \in K$ with

$$\beta = \lambda_1\beta_1 + \cdots + \lambda_n\beta_n,$$

since the β_j's span L over K. Now every λ_i can be written as a $\mathbb{Q}$–linear combination of the α's:

$$\lambda_i = a_{i1}\alpha_1 + \cdots + a_{im}\alpha_m \quad (i = 1, \ldots, n).$$

Merging these expressions yields

$$\beta = \sum_i (\sum_j a_{ij}\alpha_j)\beta_i = \sum_{i,j} a_{ij}(\alpha_j\beta_i).$$

Next we turn to the linear independence of the elements. So suppose

$$\sum_{i,j} \mu_{ij}\alpha_i\,\beta_j = 0,$$

where alle the $\mu_{ij} \in \mathbb{Q}$. Upon rewriting this as

$$\sum_j (\sum_i \mu_{ij}\alpha_i)\,\beta_j = 0$$

and using the independence of the β_j over K we conclude

$$\sum_i \mu_{ij}\alpha_i = 0 \quad \text{for all } i.$$

Using the independence of the α_i over $\mathbb{Q}$, we finally conclude that $\mu_{ij} = 0$ for all i and j. □

2.4.10 Remark. We have stated the theorem in the context of subfields of $\mathbb{C}$, but it holds in general for field extensions $K \subset L \subset M$: if M is a L–vectorspace of dimension m and L is a K–vectorspace of dimension n, then M is a K–vectorspace of dimension mn. The proof given above, with minor modifications, also works in this general context. The following result is again a special version of a result valid in a more general context.

2.4.11 Lemma. *a) Let α, β be algebraic numbers. Then*

$$[\mathbb{Q}(\alpha,\beta) : \mathbb{Q}(\alpha)] \leq [\mathbb{Q}(\beta) : \mathbb{Q}].$$

In particular, $[\mathbb{Q}(\alpha,\beta) : \mathbb{Q}] \leq [\mathbb{Q}(\alpha) : \mathbb{Q}] \cdot [\mathbb{Q}(\beta) : \mathbb{Q}]$.

b) Let K be a subfield of $\mathbb{C}$ with $\dim_{\mathbb{Q}}(K) < \infty$ and let $\gamma \in K$. Then γ is algebraic.

c) $\mathbb{A}$ is a subfield of $\mathbb{C}$.

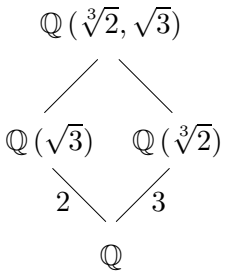
Proof. a) Suppose $[\mathbb{Q}(\beta) : \mathbb{Q}] = n$. Let $x \in \mathbb{Q}[\alpha,\beta] = \mathbb{Q}(\alpha,\beta)$. Then x can be written as a polynomial in α and β. Rewrite this polynomial as $f_1(\alpha) + f_2(\alpha)\beta + \cdots + f_m(\alpha)\beta^m$, where the f_i are polynomials in one variable. Since any β^j $(j \geq n)$ can be rewritten as a $\mathbb{Q}$–linear combination of $1, \beta, \ldots, \beta^{n-1}$, we see that x can be written as

$$x = g_1(\alpha) + g_2(\alpha)\beta + \cdots + g_{n-1}(\alpha)\beta^{n-1},$$

i.e., a $\mathbb{Q}(\alpha)$–linear combination of the $n-1$ elements $1, \beta, \ldots, \beta^{n-1}$. The dimension as $\mathbb{Q}(\alpha)$–vectorspace is therefore at most n.

The last statement in a) is a consequence of Theorem 2.4.9.
b) Suppose $\dim_{\mathbb{Q}}(K) = m$. Then the $m+1$ elements $1, \gamma, \gamma^2, \ldots, \gamma^m$ must be linearly dependent over $\mathbb{Q}$. So there exist rational numbers $a_0, \ldots, a_m$, not all of them zero, such that $a_0 + a_1\gamma + \cdots + a_m\gamma^m = 0$. This expresses that γ is an algebraic number.
c) Given algebraic numbers α and β, we must show that $\alpha - \beta$ and $1/\alpha$ (if $\alpha \neq 0$) are also algebraic. Now all these elements belong to $\mathbb{Q}(\alpha, \beta)$, which is a finite extension of $\mathbb{Q}$ by a). Then b) implies that every element in $\mathbb{Q}(\alpha, \beta)$ is algebraic. □

2.4.12 It is common practice to draw field extensions in pictures like the following.



The edges are often labelled with the degrees (known so far) of the corresponding extensions.

2.4.13 Example. To compute the degree $[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}]$ use the lemma to deduce that $[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}(\sqrt[3]{2})] \leq 2$. Theorem 2.4.9 implies that $[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}]$ is divisible by 2 and 3 and that the total degree is bounded by 6:

$$[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}] = [\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}(\sqrt[3]{2})] \cdot [\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] \leq 2 \cdot 3 = 6.$$

In conclusion, $[\mathbb{Q}(\sqrt[3]{2}, \sqrt{3}) : \mathbb{Q}] = 6$. This implies, for example, that $\sqrt[3]{2}+\sqrt{3}$ has minimal polynomial of degree at most 6 (in fact, it turns out to have degree exactly 6, see Example 2.4.16).

2.4.14 Next we turn to the computation of minimal polynomials of polynomial expressions of two given algebraic numbers. Suppose α has minimal polynomial $p(X)$ and β has minimal polynomial $q(X)$. Suppose that we want to compute the minimal polynomial of some polynomial expression $f(\alpha, \beta)$

of α and β. Note that $f(\alpha,\beta) \in \mathbb{Q}(\alpha,\beta)$. (Lemma 2.4.11a) implies that $[\mathbb{Q}(\alpha,\beta) : \mathbb{Q}] < \infty$ and b) of the same lemma implies that $f(\alpha,\beta)$ is algebraic.)

The abstract way to compute with α is to compute modulo $p(X)$. Similarly, computing with β comes down to computing modulo $q(X)$. To separate these two, we do the computations in $\mathbb{Q}[X,Y]/(p(X),q(Y))$. To compute with $f(\alpha,\beta)$ we introduce a third variable Z and add the generator $Z - f(X,Y)$ to the ideal:

$$\mathbb{Q}[X,Y,Z]/(p(X),q(Y),Z-f(X,Y)).$$

So the class of Z represents our algebraic number $f(\alpha,\beta)$. To find its minimal polynomial we compute a reduced Gröbner basis G for the ideal $I = (p(X),q(Y),Z-f(X,Y))$ with respect to lex order and such that Z is the smallest variable. Then we consider $G \cap \mathbb{Q}[Z]$. By the elimination theorem 2.1.1 this produces a generator for $I \cap \mathbb{Q}[Z]$. This generator is the minimal polynomial of $f(\alpha,\beta)$. The formal statement (there is one condition in the theorem that we neglected so far) and its proof are as follows.

2.4.15 Theorem. *Let α and β be algebraic numbers with minimal polynomials $p(X)$, $q(X)$, respectively. Suppose furthermore that $q(X)$ is irreducible over $\mathbb{Q}(\alpha)$. Let $f(X,Y) \in \mathbb{Q}[X,Y]$. Then the minimal polynomial of $f(\alpha,\beta)$ is the generator of the ideal $((p(X),q(Y),Z-f(X,Y)) \cap \mathbb{Q}[Z]$ in $\mathbb{Q}[Z]$.*

Proof. We split the proof in several parts. As an auxiliary part, we first establish that

$$\mathbb{Q}[X,Y]/(p(X),q(Y)) \cong \mathbb{Q}(\alpha,\beta),$$

under the natural map determined by $X \mapsto \alpha$, $Y \mapsto \beta$. Start with the isomorphism $\mathbb{Q}[X]/(p(X)) \to \mathbb{Q}(\alpha)$ sending the class of X to α and use it to define $\Phi : \mathbb{Q}(\alpha)[Y] \to \mathbb{Q}(\alpha,\beta)$, determined by $\phi_2(Y) = \beta$. This morphism is clearly surjective. The kernel consists precisely of those polynomials $h(Y) \in \mathbb{Q}(\alpha)[Y]$ such that $h(\beta) = 0$. We were given that $q(Y) \in \ker(\Phi)$. Now the ideal $\ker(\Phi)$ is generated by a single element. From the irreducibility of $q(Y)$ we deduce that $\ker(\Phi) = (q(Y))$. So

$$\mathbb{Q}(\alpha)[Y]/(q(Y)) \cong \mathbb{Q}(\alpha,\beta).$$

Now use $\mathbb{Q}(\alpha) \cong \mathbb{Q}[X]/(p(X))$ to conclude that $\mathbb{Q}[X,Y]/(p(X),q(Y)) \cong \mathbb{Q}(\alpha,\beta)$ (further details in the exercises).

Having established this, we turn to the actual proof. Consider the ring morphism determined by

$$\phi : \mathbb{Q}[Z] \to \mathbb{Q}[X,Y]/(p(X),q(Y)), \quad Z \mapsto f(X,Y) + (p(X),q(Y)).$$

Then $g(Z)$ is in the kernel of this map if and only if $f(\alpha, \beta) = 0$. We need to show that this kernel equals $((p(X), q(Y), Z - f(X,Y)) \cap \mathbb{Q}[Z]$. Consider the following maps

$$\mathbb{Q}[Z] \overset{j}{\hookrightarrow} \mathbb{Q}[X,Y,Z] \xrightarrow{\psi} \mathbb{Q}[X,Y]/(p(X), q(Y)) \cong \mathbb{Q}(\alpha, \beta),$$

where ψ maps X and Y to their respective classes and maps Z to the class of $f(X,Y)$. The kernel of the map ψ evidently contains $p(X)$, $q(Y)$, $Z - f(X,Y)$. It is in fact precisely I. To see this, start with $F(X,Y,Z) \in \ker(\psi)$, and rewrite F as a polynomial in Z with coefficients in $\mathbb{Q}[X,Y]$:

$$a_0(X,Y) + a_1(X,Y)Z + \cdots + a_m(X,Y)Z^m.$$

Now replace every occurrence of Z by $(Z - f(X,Y)) + f(X,Y)$ and expand accordingly (formally, replace Z by $U + V$ and expand; then substitute $Z - f(X,Y)$ for U and $f(X,Y)$ for V and do not expand any further). The result is that F is rewritten as the sum of a multiple of $Z - f(X,Y)$ and and a polynomial in X, Y:

$$\begin{aligned}&(Z - f(X,Y))h(X,Y,Z) + \\ &\quad a_0(X,Y) + a_1(X,Y)f(X,Y) + \cdots + a_m(X,Y)f(X,Y)^m.\end{aligned}$$

Since F and $(Z - f(X,Y))h(X,Y,Z)$ are in the kernel of ψ, we conclude that $a_0(X,Y) + a_1(X,Y)f(X,Y) + \cdots + a_m(X,Y)f(X,Y)^m$ is also in the kernel of ψ. But this polynomial maps to the class of

$$a_0(X,Y) + a_1(X,Y)f(X,Y) + \cdots + a_m(X,Y)f(X,Y)^m$$

in $\mathbb{Q}[X,Y]/(p(X), q(Y))$. So

$$a_0(X,Y) + a_1(X,Y)f(X,Y) + \cdots + a_m(X,Y)f(X,Y)^m \in (p(X), q(Y)).$$

Altogether we find that

$$\begin{aligned}F(X,Y,Z) &= (Z - f(X,Y))h(X,Y,Z) + a_0(X,Y) + \\ &\quad a_1(X,Y)f(X,Y) + \cdots + a_m(X,Y)f(X,Y)^m \\ &\quad \in ((p(X), q(Y), Z - f(X,Y)).\end{aligned}$$

Now the kernel of ϕ is the kernel of the composition $\psi \circ j$, where j denotes the inclusion map. This means that $g(Z) \in \ker(\phi)$ if and only if $g(Z) \in \ker(\psi)$, i.e., $g(Z) \in I$ (note that we identify $\mathbb{Q}[Z]$ with its image under j). $\square$

2.4.16 Example. The minimal polynomial of $\sqrt{2}$ is X^2-2 and the minimal polynomial of $\sqrt[3]{2}$ is X^3-2. To find the minimal polynomial of $\sqrt{2}+\sqrt[3]{2}$ we compute a Gröbner basis for the ideal $(X^2-2, Y^3-2, Z-X-Y)$ in $\mathbb{Q}[X,Y,Z]$ with respect to the lex order, where $X>Y>Z$. The result is

$$\begin{aligned}&[310\,X-24\,Z^5-462\,Z+156\,Z^2-9\,Z^4+364+160\,Z^3,\\&152\,Z-156\,Z^2+9\,Z^4+310\,Y-364-160\,Z^3+24\,Z^5,\\&-24\,Z+12\,Z^2-6\,Z^4-4-4\,Z^3+Z^6].\end{aligned}$$

The last polynomial is one involving only Z and by the above must be the minimal polynomial of $\sqrt{2}+\sqrt[3]{2}$:

$$-24\,Z+12\,Z^2-6\,Z^4-4-4\,Z^3+Z^6.$$

2.4.17 Remark. There are many variants on this theorem for computing minimal polynomials of various combinations of algebraic numbers.

Chapter 3

Factorisation of polynomials

3.1 Introduction

This chapter deals with the problem of factoring polynomials into irreducible factors. We will sketch approaches to the the problem of factoring polynomials in $\mathbb{F}_q[X]$, where q is a prime power, and of polynomials with integer coefficients. We begin with a section on generalities.

3.2 Polynomials with integer coefficients

3.2.1 If k is a field, then every polynomial in $k[X]$ factors into irreducibles in a (up to the order of the factors and up to constant factors) unique way. This holds in particular for polynomials in $\mathbb{Q}[X]$. In the following it will be essential to work with polynomials with integer coefficients. Fortunately, factors of a polynomial with integer coefficients can always be taken in $\mathbb{Z}[X]$. To explain this, we denote by $c(f)$ the gcd of the coefficients of $f \in \mathbb{Z}[X]$, the so–called *content* of f. The first statement in the following proposition is usually called the Lemma of Gauss.

3.2.2 Proposition. *a) The content is multiplicative in the sense that for non-zero polynomials $f, g \in \mathbb{Z}[X]$ the relation $c(fg) = c(f)\,c(g)$ holds.*

b) If $f \in \mathbb{Z}[X]$ factors as $f = gh$ with $g, h \in \mathbb{Q}[X]$, then there exist $a, b \in \mathbb{Q}$ such that ag, $bh \in \mathbb{Z}[X]$ and $f = (ag)\,(bh)$.

Proof. a) The crucial case to consider is when $c(f) = c(g) = 1$. Reducing modulo any prime p we get $\overline{fg} = \overline{f} \cdot \overline{g}$ with $\overline{f}$, $\overline{g} \neq 0$. As $\mathbb{F}_p[X]$ is an integral

domain, we conclude that $\overline{fg} \neq 0$ showing that $c(fg)$ has no prime divisors, hence equals 1.

b) The proof is easily reduced to the case where $c(f) = 1$. In that case multiply g and h by rational numbers a and b, respectively, in such a way that ag and bh have integer coefficients and content 1. Then $c(abf) = c(ag)c(bh) = 1$ by a), so that $ab = \pm 1$. Altering the sign of a if necessary, we can arrange it so that $ab = 1$ and $f = (ag)(bh)$. □

3.2.3 Now an obvious operation on polynomials with integer coefficients is to reduce their coefficients modulo a prime. Let p be a prime and let $f = a_m X^m + a_{m-1}X^{m-1} + \cdots + a_1 X + a_0$ be a polynomial with integer coefficients. Then the reduction mod p of f is the polynomial

$$\overline{f} = \overline{a_m} X^m + \overline{a_{m-1}} X^{m-1} + \cdots + \overline{a_1} X + \overline{a_0}.$$

3.2.4 Proposition. *Let $f \in \mathbb{Z}[X]$ of positive degree whose leading coefficient is not divisible by the prime p. If the reduction of f mod p is irreducible, then f is irreducible as a polynomial in $\mathbb{Q}[X]$.*

Proof. Suppose f factors as $f = gh$ with g and h of positive degree. By the Lemma of Gauss, we may assume that g and h have integer coefficients. Upon reducing mod p we find the equality $\overline{f} = \overline{g} \cdot \overline{h}$ in $\mathbb{F}_p[X]$, contradicting the irreducibility of $\overline{f}$. □

3.2.5 The next criterion is equally simple to prove, yet is much subtler.

3.2.6 Proposition. (Eisenstein's criterion) *Let p be a prime and let $f = a_m X^m + a_{m-1}X^{m-1} + \cdots + a_1 X + a_0$ be a polynomial with integer coefficients satisfying:*

a) p does not divide a_m;

b) p divides $a_{m-1}, a_{m-2}, \ldots, a_0$;

c) p^2 does not divide a_0.

Then f is irreducible in $\mathbb{Q}[X]$.

Proof. If f factors as $f = gh$ with g and h both in $\mathbb{Z}[X]$ and of positive degree, then reduction mod p yields $\overline{f} = \overline{g} \cdot \overline{h}$. Since $\overline{f} = \overline{a_m}\, X^m \neq 0$ and since we have unique factorization in $\mathbb{F}_p[X]$, we conclude that $\overline{g}$ and $\overline{h}$ each consists of its leading term only. In particular, their constant terms are 0, so that the constant terms of g and h are divisible by p. But then a_0 is divisible by p^2, a contradiction. □

3.2.7 We leave the following as an exercise for the reader. Let k be a field and let $a \in k$. If $f \in k[X]$, then f is irreducible if and only if $f(X+a)$ is irreducible.

3.2.8 Example. For p prime, define $\Phi_p(X) = X^{p-1} + X^{p-2} + \cdots + X + 1$. Then

$$\Phi_p(X) = \frac{X^p - 1}{X - 1}.$$

Replace X by $X + 1$ and we find

$$\Phi_p(X + 1) = \frac{(X + 1)^p - 1}{X}.$$

The right–hand side works out as

$$X^{p-1} + \binom{p}{p-1} X^{p-2} + \cdots + \binom{p}{2} X + p.$$

This polynomial is suitable for the application of Eisenstein's criterion for the prime p. We conclude that $\Phi_p(X)$ is irreducible.

$\Phi_p(X)$ is part of a family of polynomials, the *cyclotomic polynomials* $\Phi_m(X)$ for $m \in \mathbb{Z}$, $m > 0$. These are the minimal polynomials of

$$e^{\frac{2\pi i}{m}}$$

and of fundamental importance in number theory.

3.3 Factoring polynomials modulo a prime

3.3.1 Factoring a polynomial in $\mathbb{F}_q[X]$ (with q a power of the prime p) is a finite job. The purpose of this section is to demonstrate Berlekamp's algorithm, a more efficient way of factoring. It is based on the following observation.

3.3.2 Lemma. *Let $f \in \mathbb{F}_q[X]$ and suppose $u \in \mathbb{F}_q[X]$ satisfies $u^q \equiv u \pmod{f}$. Then*

$$f = \Pi_{a \in \mathbb{F}_q} \gcd(f, u - a).$$

Proof. Substituting u in the equality $X^q - X = \Pi_{a \in \mathbb{F}_q}(X - a)$, we find $u^q - u = \Pi_{a \in \mathbb{F}_q}(u - a)$. Since $(u - a) - (u - b) = b - a$ is a constant, The factors on the right–hand side are relatively prime. As f divides $u^q - u$, we find the following string of equalities

$$f = \gcd(f, u^q - u) = \gcd(f, \Pi_{a \in \mathbb{F}_q}(u - a)) = \Pi_{a \in \mathbb{F}_q} \gcd(f, u - a).$$

□

3.3.3 Example. Let $f = X^4 - 1 \in \mathbb{F}_5[X]$ and let $u = X$. Then the conditions of the lemma are satisfied and we find the factors

$$\begin{array}{lll} \gcd(f, X-0) = 1, & \gcd(f, X-1) = X-1, & \gcd(f, X+1) = X+1, \\ \gcd(f, X-2) = X-2, & \gcd(f, X+2) = X+2. & \end{array}$$

In this example we happen to find the full factorisation. This need not be the case in general.

3.3.4 Of course, when one u doesn't work, another one may. So it makes sense to vary the choice of u. To do this, we first investigate the structure of the set of all u satisfying $u^q \equiv u \pmod f$. It turns out to be an m–dimensional vector space over $\mathbb{F}_q$, where m is the number of distinct irreducible divisors of f. Before we state this and give the proof, we first explain an ingredient of the proof. Suppose f factors as $f = f_1^{e_1} \cdots f_m^{e_m}$. Consider the map

$$\begin{array}{l} \phi : \mathbb{F}_q[X] \to \mathbb{F}_q[X]/(f_1^{e_1}) \times \cdots \times \mathbb{F}_q[X]/(f_m^{e_m}) \\ g \mapsto (g + (f_1^{e_1}), \ldots, g_m + (f_m^{e_m})) \end{array}$$

It is easy to verify that this map is a morphism of rings. The kernel consists of the polynomials $h \in \mathbb{F}_q[X]$ such that $f_j^{e_j}$ divides h for $j = 1, \ldots, m$. Since the $f_j^{e_j}$ are relatively prime, we conclude that h is in the kernel if and only if $f \mid h$. But this implies that we get a well-defined injective morphism

$$\begin{array}{l} \overline{\phi} : \mathbb{F}_q[X]/(f) \to \mathbb{F}_q[X]/(f_1^{e_1}) \times \cdots \times \mathbb{F}_q[X]/(f_m^{e_m}), \\ g + (f) \mapsto (g + (f_1^{e_1}), \ldots, g_m + (f_m^{e_m})). \end{array}$$

The number of elements in $\mathbb{F}_q[X]/(f)$ is $q^{\deg(f)}$ and the number of elements in $\mathbb{F}_q[X]/(f_1^{e_1}) \times \cdots \times \mathbb{F}_q[X]/(f_m^{e_m})$ is $q^{e_1 \deg(f_1)} \cdots q^{e_m \deg(f_m)}$. Since these numbers are equal the map $\overline{\phi}$ is an isomorphism.

This isomorphism (a polynomial version of the Chinese remainder theorem) enables us to translate statements about $\mathbb{F}_q[X]/(f)$ into statements about $\mathbb{F}_q[X]/(f_1^{e_1}) \times \cdots \times \mathbb{F}_q[X]/(f_m^{e_m})$ and conversely. Note that it is not evident how to describe the inverse of $\overline{\phi}$.

The class of an element $u \in \mathbb{F}_q[X]$ such that $f \mid u^q - u$ behaves as follows under the isomorphism. The decomposition

$$u^q - u = \Pi_{a \in \mathbb{F}_q}(u - a)$$

into relatively prime factors shows that each f_j divides exactly one of these factors, say $u - a$. But then $f_j^{e_j}$ divides $u - a$ because $f | u^q - u$. So modulo $f_j^{e_j}$, the class of u corresponds to a constant a. So the image of an element

u satisfying $f \mid u^q - u$ belongs to the subring $(\mathbb{F}_q)^m$ of $\mathbb{F}_q[X]/(f_1^{e_1}) \times \cdots \times \mathbb{F}_q[X]/(f_m^{e_m})$.

Conversely, since every element $a \in \mathbb{F}_q$ satisfies $a^q = a$, every element of the subring $(\mathbb{F}_q)^m$ comes from an element $u \in \mathbb{F}_q[X]/(f)$ satisfying $u^q = u$. This proves the following proposition.

3.3.5 Proposition. *The set $S_f = \{u \in \mathbb{F}_q[X]/(f) \mid u^q = u\}$ is an m–dimensional $\mathbb{F}_q$–vector space, where m is the number of distinct irreducible divisors of f. Under the isomorphism*

$$\overline{\phi} : \mathbb{F}_q[X]/(f) \to \mathbb{F}_q[X]/(f_1^{e_1}) \times \cdots \times \mathbb{F}_q[X]/(f_m^{e_m})$$

the set S_f corresponds to the subring $(\mathbb{F}_q)^m$ of constants.

3.3.6 To find the set S_f we need to solve a system of equations in $\mathbb{F}_q[X]/(f)$. Since $u(X)^q = u(X^q)$, solving $u^q = u \pmod f$ comes down to solving a system of linear equations.

Berlekamp's algorithm to find the factors $f_j^{e_j}$ of f runs as follows: first determine a vector space basis $u_1 = 1, \ldots, u_m$ of S_f and let $D = \{f\}$. In each stage of the algorithm the set D contains polynomials whose product equals f. If $m = 1$, there is only one prime power in f and we are done. If $m > 1$, then we use Lemma 3.3.2 and replace for $j = 2, \ldots, m$ every element g of D by the nontrivial elements among the $\gcd(g, u_j - a)$ with $a \in \mathbb{F}_q$. Since only finitely many steps are involved, the algorithm terminates and it remains to check that at the end $D = \{f_1^{e_1}, \ldots, f_m^{e_m}\}$. We do this in two steps.

a) First note that in every step of the algorithm every element of D is either divisible by $f_i^{e_i}$ or is relatively prime to f_i (for every i): if this property holds for g, then it is inherited by every $\gcd(g, u - a)$, where $a \in \mathbb{F}_q$, since every $u - a$ is either relatively prime to f_i or is divisible by $f_i^{e_i}$ (see 3.3.4).

b) If g does not decompose further at the end, then for every j there exists an $s_j \in \mathbb{F}_q$ such that $g \mid u_j - s_j$. Since the u_j make up a $\mathbb{F}_q$–basis of S_f, a similar statement holds for arbitrary $u \in S_f$: for every $u \in S_f$ there exists an $s_u \in \mathbb{F}_q$ such that $g \mid u - s_u$. If g is divisible by the factors $f_k^{e_k}$ and $f_l^{e_l}$ then the k–th and l–th components of $\overline{\phi}(u)$ equal s_u, contradicting the surjectivity of $\overline{\phi}$.

3.3.7 The algorithm so far produces the factors $f_i^{e_i}$ of f. The last step consists of finding the f_i from these powers. Before we proceed, we recall the following

equality for polynomials in $\mathbb{F}_q[X]$:

$$g(X)^p = g(X^p) \quad \text{and consequently} \quad g(X)^{p^k} = g(X^{p^k}).$$

For example, in characteristic 2, we have

$$X^8 + X^4 + 1 = (X^2)^4 + X^4 + 1^4 = (X^2 + X + 1)^4.$$

Now let's suppose we are given g^n and we wish to find g. The first approach is to differentiate g^n, this yields $(g^n)' = ng'\,g^{n-1}$, and divide g^n by this derivative. (We remark that we haven't defined differentiation in characteristic p and the following discussion shows exactly some of the pitfalls.) This goes well if p does not divide n and if $g' \neq 0$. If $g' = 0$, then all exponents in g are divisible by p. By using the above equality repeatedly, we can absorb all these factors p and rewrite g^n as

$$g(X)^n = h(X^{p^k})^t,$$

with p relatively prime with t and with at least one of the exponents of $h(X)$. Then compute h/h'.

3.4 Factoring polynomials over the integers

3.4.1 As we saw above, factoring over the rationals is essentially the same as factoring over the integers, so we will focus on the latter. We started the discussion on factoring polynomials mod q by remarking that it is a finite problem anyway. It takes a little consideration to show that factoring over the integers or rationals is also a finite job: if $g \in \mathbb{Z}[X]$ divides $f \in \mathbb{Z}[X]$, then it is easy to see that the coefficients of g can be bounded in terms of those of f. Just start with the leading coefficient and work through the coefficients one by one, bookkeeping the bounds found so far. Without proof we state below a nice bound (in fact sharp), the *Landau–Mignotte bound*. To state it, we need the so–called l_2–norm of a polynomial $f = \sum_{i=0}^{n} f_i\,X^i \in \mathbb{Z}[X]$:

$$||f|| = \sqrt{\sum_{i=0}^{n} f_i^2}.$$

Of course, our algorithmic approach will be better than just an exhaustive search using this bound. In this section we restrict ourselves to a discussion of an algorithm based on the *Hensel lift*.

3.4.2 Theorem. (Landau–Mignotte) *If $f, g \in \mathbb{Z}[X]$ are of degrees n and m, respectively, and if $g \mid f$, then*

$$|g_i| \leq \binom{m}{i} ||f||$$

for $i = 0, 1, \ldots, m$. A bound for the l_2–norm of g is given by

$$||g||^2 \leq \binom{2m}{m}^2 ||f||^2.$$

3.4.3 The starting point for the Hensel lift is a factorization of the reduction of a polynomial modulo a prime p. The Hensel lift aims at lifting a factorisation modulo p^k to a factorisation modulo p^{k+1}. Together with the Landau–Mignotte bound on the coefficients of possible factors, these ingredients can be put together into an algorithm.

3.4.4 Theorem. (Hensel lift) *Let p be a prime and let $f, g, h \in \mathbb{Z}[X]$ be monic polynomials of positive degree such that $f \equiv g\,h \pmod{p^m}$ for some $m \in \mathbb{N}$ with $\gcd(g \pmod p, h \pmod p) = 1$. Then there exist monic polynomials $\tilde{g}$, $\tilde{h}$ such that $\tilde{g} \equiv g \pmod{p^m}$, $\tilde{h} \equiv h \pmod{p^m}$ and $f \equiv \tilde{g}\,\tilde{h} \pmod{p^{m+1}}$. Moreover, the lifts $\tilde{g}$ and $\tilde{h}$ are uniquely determined modulo p^{m+1}.*

Proof. The polynomials $\tilde{g}$ and $\tilde{h}$ are required to be of the form

$$\tilde{g} = g + p^m u, \quad \tilde{h} = h + p^m v,$$

with the polynomials u and v satisfying $\deg(u) < \deg(g)$ and $\deg(v) < \deg(h)$. Using these two explicit forms, the equation $f \equiv \tilde{g}\,\tilde{h} \pmod{p^m}$ can now be rewritten as

$$\frac{f - gh}{p^m} - (uh + vg) \equiv 0 \pmod p.$$

Since $\gcd(g \pmod p, h \pmod p) = 1$, there exist polynomials u of degree less than $\deg(g)$ and v of degree less than $\deg(h)$ such that

$$\frac{f - gh}{p^m} \equiv uh + vg \pmod p.$$

Moreover, modulo p these polynomials u and v are unique. Consequently, the resulting $\tilde{g}$ and $\tilde{h}$ are unique modulo p^{m+1}. □

3.4.5 (Factorization) Let $f \in \mathbb{Z}[X]$ be square–free of degree $n > 0$ and suppose $c(f) = 1$. Choose un upper bound C for the coefficients of possible divisors of f, for instance the Landau–Mignotte bound. Also choose a prime p that does not divide the discriminant of f (so that the reduction of f mod p is also square–free) and does not divide the leading coefficient a_n of f (so that the reduction mod p does not drop in degree) and an integer m so that $p^m > 2\,|a_n|\,C$. The algorithm then goes through the following steps:

a) Factor f mod p.

b) Lift this factorization to a factorization

$$f \equiv a_n g_1 \cdots g_k \pmod{p^k},$$

with monic g_i.

c) For every subset S of $\{1, 2, \ldots, k\}$ compute a polynomial $h \in \mathbb{Z}[X]$ such that

(i) $h \equiv a_n \Pi_{i \in S} g_i \pmod{p^m}$;

(ii) h has degree at most $[n/2]$;

(iii) the absolute values of the coefficients of h are less than $|a_n|C$.

Finally, test whether h divides f and assemble the divisors of f.

The number of tests in the final step can be exponential in n, for instance if f factors into linear factors modulo p.

Chapter 4

Symbolic integration

4.1 Introduction

4.1.1 In this chapter we are concerned with the problem of finding exact *antiderivatives* or *indefinite integrals*: given a function $f(x)$, find $F(x)$ such that

$$F'(x) = f(x) \quad \text{or} \quad \int f(x)\,dx = F(x).$$

In basic calculus one learns a range of methods for determining indefinite integrals, most notably integration by parts and the substitution rule. But in most cases a fully algorithmic approach is not clearly available. One is often guided by intuition and experience. Some integrals don't seem to yield to any method; for instance,

$$\int e^{-x^2}\,dx$$

is such an integral. In such cases, sometimes there do exist methods to determine specific *definite* integrals, like

$$\int_0^\infty e^{-x^2}\,dx,$$

but that will not be the topic of this chapter. We will be solely interested in the problem of finding antiderivatives in the following two senses:

- given an expression $f(x)$ in terms of 'elementary' functions, determine (algorithmically) an antiderivative in terms of 'elementary functions', or

- show that there is no antiderivative of $f(x)$ in such elementary terms.

The term 'elementary function' will be made more precise below, but for the moment it suffices to know that it refers to combinations of the usual functions, like polynomials, sine, cosine, exponential, where combination is meant in the sense of composition of functions and in the sense of applying the usual arithmetical operations like addition, multiplication, taking n–th roots. An example of an elementary expression is

$$\frac{\sqrt[5]{e^{x^2}+\sin(x)}}{\log(3+x^2-\sin(\sqrt[4]{e^x})}.$$

4.1.2 Remark. In this chapter we shall be interested in formal properties of differentiation and integration; *domains of definition of a given expression are of no importance for us*, i.e., we will be studying expressions rather than functions, although our terminology will be sloppy in this respect and the terms function and expression are both used.

4.1.3 *From the following section on, all fields in this chapter have characteristic zero.*

4.2 Differential fields

4.2.1 From the point of view of differentiation, polynomials and rational functions are quite simple. In the algebraic context, differentiation can be viewed as a certain operator on the field $K(x)$ of rational functions over the field K. In this algebraic setting more complicated expressions, like the logarithm, can be treated by adjoining them to the field $K(x)$. For this to make sense we need to be able to algebraically characterize such expressions and to define how differentiation extends to this larger field.

The idea of using fields is somewhat reminiscent of Galois theory and there is a striking resemblance between two of the corner stones in both theories: solving polynomial equations by radicals in Galois theory versus solving indefinite integrals in symbolic integration. We note that there is an obvious extension of symbolic integration to differential equations, which is often called differential Galois theory.

In the sequel we will work in suitable field extensions. Here, suitable means that the field is relevant to our specific class of functions, but is also adapted to the process of differentiation. The formal notion is the following.

4.2.2 Definition. A *differential field* consists of a field K of characteristic 0 and a map $D : K \to K$ satisfying the rules

a) $D(f+g) = Df + Dg$ for all $f, g \in K$;

b) (*Leibniz' rule*) $D(f \cdot g) = f\,Dg + g\,Df$ for all $f, g \in K$.

The map D is a so–called *derivation* or *differential operator*. If D is understood, then we simply say 'the differential field K'. If we need to be more precise we write for example (K, D). Again, if no confusion arises, we also write f' instead of $D(f)$.

4.2.3 Example. The standard example is the field of rational functions in one variable over a field, say, the complex numbers or the rational numbers,

$$K = \mathbb{C}(x) \quad \text{or} \quad \mathbb{Q}(x),$$

with derivative D determined by:

$$D(a_n x^n + \cdots + a_2 x^2 + a_1 x + a_0) = n a_n x^{n-1} + \cdots + 2a_2 x + a_1,$$

$$D(\tfrac{f}{g}) = \tfrac{g\,Df - f\,Dg}{g^2}.$$

The coefficients a_i are constants in $\mathbb{C}$ or $\mathbb{Q}$, respectively, the elements f, g are polynomials over $\mathbb{C}$ or $\mathbb{Q}$, respectively, with $g \neq 0$.

4.2.4 For the sake of practicality (think of implementations in computer algebra systems), we usually focus on extensions of the field $\mathbb{Q}$. There is a price to pay for this starting point, namely that we have to accept that sometimes algebraic integers have to be adjoined.

The proof of the following properties is left as an exercise.

4.2.5 Proposition. *In the differential field (K, D), the following properties hold:*

a) $D(0) = D(1) = 0$, more generally, $D(r) = 0$ for all $r \in \mathbb{Q} \subset K$.

b) The set $\{c \in K \mid D(c) = 0\}$ is a subfield of K. This field is called the field of constants with respect to D.

c) $D(af + bg) = aD(f) + bD(g)$ for all $f, g \in K$ and all $a, b \in K$ satisfying $D(a) = D(b) = 0$.

d) $D(\tfrac{f}{g}) = \tfrac{g\,D(f) - f\,D(g)}{g^2}$ for all $f, g \in K$ with $g \neq 0$.

e) $D(f^n) = n\, f^{n-1}\, D(f)$ *for all* $f(\neq 0) \in K$ *and* $n \in \mathbb{Z}$.

4.2.6 Similar to field extensions in ordinary field theory are the differential field extensions in our context. The important aspect is of course the role of the differential operator.

4.2.7 Definition. Let (L, D_L) and (K, D_K) be differential fields, where L is a field extension of K. If $D_L(f) = D_K(f)$ for all $f \in K$, then (L, D_L) is called a *differential extension field* of (K, D_K). If no confusion arises, we simply say that L is a differential extension field of K.

4.2.8 For our purposes three types of extensions will play a crucial role; one type of extension is of the sort discussed in the previous chapter; the two other types correspond to adjoining a 'logarithm' or 'exponential', respectively. The definitions are inspired by the formal differentiation properties of logarithms and exponentials. Again, we note that in the sequel it will be irrelevant for us to view expressions as functions; domains of definition play no role whatsoever.

4.2.9 Definition. (Logarithm, exponential) A differential field extension $L : K$ is said to *contain a logarithm* of $u \in K$ with $u \neq 0$ if L contains an element θ such that

$$D(\theta) = \frac{D(u)}{u}.$$

We often write $\theta = \log(u)$.

The extension $L : K$ is said to *contain an exponential* of $u \in$ if it contains an element θ such that

$$D(\theta) = D(u) \cdot \theta.$$

We often write e^u or $\exp(u)$ for such θ.

4.2.10 Note that we haven't proved that given $u \in K$ an extension containing a logarithm or exponential exists. Also, we are not claiming that such logarithms or exponentials necessarily bring us outside the field K.

The notation $\log(u)$ and e^u suggests that such expressions share more of the properties of the logarithm and exponential, respectively. Indeed, as regards the arithmetic structure of the field, we expect properties like $\log(u) + \log(v) = \log(uv)$ and $e^u \cdot e^v = e^{u+v}$.

In the sequel we will be a bit sloppy about the precise construction of logarithmic and exponential extensions, and simply use the various properties whenever convenient.

4.2.11 Before we turn to a specific class of functions in the next section, we end this section by showing that logarithms occur more often in the process of integration than one is inclined to think at first glance. Here, it will be an advantage that we need not be concerned about domains of definition of our expressions.

The integral

$$\int \frac{1}{x} = \log(x)$$

shows that we need the logarithm, an example of a non–rational function (see 4.3.2), to integrate a relatively easy expression like $1/x$. In the integral

$$\int \frac{1}{x^2+1} = \arctan(x)$$

the right–hand side suggests that we need again a different type of non–rational function to express the antiderivative of $1/(x^2+1)$. From the analyst's point of view, the arctan may be the best way of writing the integral, but for an algebraist the arctan obscures the structure of the integral in the following sense. The identity

$$\frac{1}{x^2+1} = -\frac{i}{2}\left(\frac{1}{x+i} - \frac{1}{x-i}\right)$$

enables us to rewrite the integral with the help of logarithms as

$$\int \frac{1}{x^2+1}\,dx = -\frac{i}{2}(\log(x+i) - \log(x-i)),$$

at the expense of extending the coefficients to $\mathbb{Q}(i)$. In the exercises you will show that more integrals, which are usually not expressed in terms of logarithms, can be expressed in terms of logarithms.

In fact, for rational functions we will see in the next section that logarithmic extensions suffice in the integration process.

4.3 Rational functions

4.3.1 A class of functions where the collection of classical integration rules is most successfull is the class of rational functions. In fact, partial fraction expansion and the standard rules of integration applied to rational functions come close to an algorithmic description. The purpose of this section is to explain this algorithmic approach. We first show that integrating $1/x$ necessarily brings us outside the realm of rational functions.

4.3.2 Theorem. *Let $\mathbb{Q}(x)$ be the function field in one variable with the usual differentiation D.*

a) The field of constants in $\mathbb{Q}(x)$ is $\mathbb{Q}$.

b) There exists no $f \in \mathbb{Q}(x)$ such that $D(f) = 1/x$.

Proof. We leave the proof of a) as an exercise and turn to the proof of b). If p and q are polynomials such that $D(p/q) = 1/x$, then the rules for differentiation imply

$$x\,q\,D(p) - x\,p\,D(q) = q^2,$$

implying $x|q^2$ and therefore $x|q$. Upon rewriting q as $x^n q_1$ with x and q_1 relatively prime and substituting we find

$$x^{n+1}\,q_1 D(p) - nx^n\,pq_1 - x^{n+1}p\,D(q_1) = x^{2n}q_1^2.$$

From this equality we obtain $x|p$ contradicting the assumption that p and q are relatively prime. □

4.3.3 In integrating a rational function, the first step is usually to apply partial fraction expansion. This reduces the problem to a set of usually simpler problems. Below we describe the various types of fractions we would like to distinguish and how to handle them. Following this discussion we show how to use partial fraction expansion to put everything together.

The next lemma is used below; its proof is left as an exercise.

4.3.4 Lemma. (Integration by parts) *For every u, v in a differential field, the following equality holds*

$$\int u\,D(v) = uv - \int v\,D(u).$$

4.3.5 (Square–free denominator) If the denominator b of the rational function $a/b \in \mathbb{Q}(x)$ (with a and b relatively prime, b monic, and $\deg(a) < \deg(b)$) is squarefree, and if b factors as

$$b = (x - c_1)(x - c_2)\cdots(x - c_n)$$

(the c_i are distinct elements in $\mathbb{C}$), then a/b can be written as a sum

$$\frac{a}{b} = \sum_{i=1}^{n} \frac{\alpha_i}{x - c_i},$$

with unique constants α_i. Therefore, we have

$$\int \frac{a}{b} = \sum_{i=1}^{n} \alpha_i \log(x - c_i).$$

To be able to write down this expression, we need to adjoin the roots $c_1, \ldots, c_n$ to the field $\mathbb{Q}$ and adjoin the various $\log(x - c_i)$. Since the roots are algebraic, adjoining them involves at most a finite extension of $\mathbb{Q}$. If two or more of the coefficients α_i coincide, then we can rearrange the sum of the logarithms. For instance, if $\alpha_1 = \alpha_2$, then

$$\alpha_1 \log(x - c_1) + \alpha_2 \log(x - c_2) = \alpha_1 \log((x - c_1)(x - c_2)).$$

This rearranging may reduce the degree of the algebraic extension we need to write down the integral as the following example shows.

$$\int \frac{2x}{x^2 + 1} = \int \frac{1}{x + i} + \frac{1}{x - i} = \log(x + i) + \log(x - i) = \log(x^2 + 1).$$

In this case no constants outside the rationals are necessary. But we still need a logarithm.

4.3.6 (Denominator is a pure power) Here we are dealing with the case a/b^m with

- $\deg(a) < \deg(b)$,
- b is squarefree,
- $m > 1$.

The idea in this case is to apply integration by parts to reduce the exponent in a/b^m as far as possible, in fact to 1. Since b is squarefree, $\gcd(b, D(b)) = 1$, and the Euclidean algorithm produces an identity of the form

$$ub + vD(b) = a.$$

Dividing both sides of the equation by b^m and integrating we get

$$\int \frac{a}{b^m} = \int \frac{u}{b^{m-1}} - \int \frac{v\, D(b)}{b^m}.$$

To reduce the exponent occurring in the second term, we apply integration by parts, noting that

$$D(b^{-(m-1)}) = \frac{-(m-1)\, D(b)}{b^m}.$$

This yields

$$\int v \cdot D(\frac{-b^{-(m-1)}}{m-1}) = \frac{-v\, b^{-(m-1)}}{m-1} + \int D(v) \cdot \frac{b^{-(m-1)}}{m-1}.$$

The integral of a/b^m then reduces to

$$\int \frac{a}{b^m} = \frac{-v}{(m-1)b^{m-1}} + \int \frac{u + D(v)/(m-1)}{b^{m-1}}.$$

So the exponent has been reduced by 1 at the cost of introducing an extra rational function. By repeating this process we can get down to exponent 1 in the denominator.

4.3.7 To put the above ingredients in a systematic scheme, we need one more item, viz., a suitable partial fraction expansion. The expansion we have in mind is not the usual one from calculus. In calculus one uses the following expansion to deal with a fraction a/b. Division with remainder reduces to the case $a/b = u + p/b$ with $p = 0$ or $\deg(p) < \deg(b)$, and with u a polynomial. Expand b as product of irreducible factors: $b = p_1^{m_1} \cdots p_s^{m_s}$. Then there exist polynomials a_{ij} each of which is either zero or satisfies $\deg(a_{ij}) < \deg(p_j)$ such that

$$\frac{p}{b} = \sum_j \left(\frac{a_{1j}}{p_j} + \frac{a_{2j}}{p_j^2} + \cdots + \frac{a_{m_j j}}{p_j^j} \right).$$

The bottle–neck in computations is the decomposition into irreducible factors. To avoid this, one uses a partial fraction expansion based on so–called *squarefree factorizations*. This is more efficient in computations, see the exercises for an algorithmic approach. A polynomial b is called *squarefree* if there exists no polynomial d of positive degree such that d^2 divides b.

4.3.8 Lemma. (Squarefree factorization) *Let $b \in K[x]$ be a monic polynomial of positive degree. Then there exist relatively prime monic squarefree polynomials $b_1, b_2, \ldots, b_k$ with b_k of positive degree such that*

$$b = b_1\, b_2^2 \cdots b_k^k.$$

Proof. To produce such a factorization, factor b into irreducible factors and regroup according to exponent. □

4.3.9 Example. A squarefree decomposition of $x^2(1+x^2)^2(1+x+x^2)^5 \in \mathbb{Q}[x]$ is

$$[x(1+x^2)]^2 \cdot (1+x+x^2)^5.$$

Squarefree decompositions are not unique in general. For example, $x^3(1+x)^6$ admits the squarefree factorizations $x^3(1+x)^6$ and $(x(1+x)^2)^3$.

4.3.10 The proof of the above lemma on the existence of squarefree factorizations uses the decomposition into irreducible factors. There is, however, a more efficient way to obtain such a factorization which only uses gcd computations. This approach is worked out in the exercises.

4.3.11 If the polynomial b (of positive degree) is represented as a product $f_1 \cdots f_m$ of relatively prime polynomials f_i, then every fraction a/b admits a partial fraction decomposition of the form

$$\frac{a}{b} = a_0 + \frac{a_1}{f_1} + \cdots + \frac{a_m}{f_m},$$

where every a_i $(= 1, \ldots, m)$ is either 0 or satisfies $\deg(a_i) < \deg(f_i)$.

Similarly, every fraction a/b^m with $\deg(a) < m \deg(b)$ admits a partial fraction decomposition of the form

$$\frac{a}{b^m} = \frac{a_1}{b} + \frac{a_2}{b^2} + \cdots + \frac{a_m}{b^m},$$

with each a_i satisfying either $a_i = 0$ or $\deg(a_i) < \deg(b)$.

4.3.12 (Hermite's method) Here is how we reduce the integral of a given rational expression to one in which only integrals of rational expressions with squarefree denominators occur.

Let $p, q \in K(x)$ be relatively prime and let q be monic. Hermite's method consists of reducing the integral of p/q to an expression of the form

$$\frac{r}{s} + \int \frac{a}{b},$$

where $a, b, r, s \in K[x]$, $\deg(a) < \deg(b)$, $\gcd(a, b) = 1$, b monic and squarefree.

- Using division we can write $p = q\,u + v$ with $v = 0$ or $\deg(v) < \deg(q)$. This reduces the problem to the integration of v/q.

- Use the squarefree factorization

$$q = \Pi_{k=1}^{m} q_k$$

to produce a partial fraction expansion of the form

$$\frac{u}{v} = \sum_{k=1}^{m} \sum_{l=1}^{k} \frac{t_{kl}}{q_k^l},$$

where $\deg(t_{kl}) < \deg(q_k)$ if $\deg(q_k) > 0$ and $t_{kl} = 0$ otherwise.

By applying 4.3.6 to terms of the form

$$\int \frac{t_{kl}}{q_k^l}$$

with $l > 1$ repeatedly, reduce to the form asserted above.

4.3.13 Once an integral is reduced to integrals in which the denominators of the rational expressions are squarefree, we can apply 4.3.5 to integrate at the expense of having to introduce logarithmic extensions. To deal with these integrals a bit more precise, we state the following result (for resultants, see the appendix)

4.3.14 Theorem. (Rothstein–Trager) *Let $K(x)$ be a differential field in one variable with field of constants K. Let $a, b \in K[x]$ be relatively prime with $\deg(a) < \deg(b)$ and with b monic. Then the minimal algebraic extension of K such that the integral of a/b can be expressed as*

$$\int \frac{a}{b} = \sum_i c_i \log(v_i),$$

with $c_i \in K$ and $v_i \in K[x]$, is the splitting field $\overline{K}$ of the resultant

$$R(z) = \operatorname{res}_x(a - z\, D(b), b) \in K[z].$$

In fact, the c_i are the roots of $R(z)$ and $v_i = \operatorname{res}_x(a - c_i\, D(b), b)$ for every i.

Proof. First we determine the c_i as roots of the resultant mentioned. We start with the equality

$$\frac{a}{b} = \sum_i c_i \frac{v_i'}{v_i}$$

and clear denominators. In doing this, it is useful to write $u_i = \Pi_{j\neq i} v_j$:

$$a\,\Pi_{j=1}^{n} v_j = b\,\sum_i c_i v_i'\,u_i.$$

From this equality we deduce that $b|v_1\cdots v_n$. The converse also holds: since $v_j|u_i$ for $j\neq i$, the same equality implies that $v_j|bv_j'\,u_j$. But $\gcd(v_j, v_j') = 1$ since v_j is squarefree and $\gcd(v_j, u_j) = 1$ by construction, and so $v_j|b$. Again using that the v_i are relatively prime, we conclude that the product $v_1\cdots v_n$ divides b.

Using the equality $b = v_1\cdots v_n$, we find $a = \sum_i c_i v_i'\,u_i$. Then

$$a - c_j b' = \sum_i c_i v_i'\,u_i - c_j \sum_i v_i' u_i = \sum_i (c_i - c_j) v_i'\,u_i.$$

Now v_j divides each term in the last sum (for $i = j$ the term vanishes!), so that v_j is a common divisor of $a - c_j b'$ and b. But that implies that c_j is a root of $\mathrm{res}(a - zb', b)$.

In the next step we show that $v_i = \gcd(a - c_i b', b)$. □

4.4 Beyond rational functions

This section is intended to give an impression of the situation for transcendental expressions. Since a full treatment is beyond the scope of this course, we will focus on special cases. In particular, we will concentrate on integrals containing a logarithmic or exponential expression transcendental over the field of rational functions, like

$$\int e^{x^2} \quad \text{or} \quad \int \frac{x}{\log(x)},$$

but will not touch upon integrals whose integrand is algebraic over this field, like

$$\int \frac{1}{\sqrt{x^8+1}}.$$

The third type seems closer to the rational functions than the first two types, and one is therefore tempted to think that the third type is easier than the other two. As it turns out, the first two types are simpler to deal with. A treatment of the third type requires a working knowledge of the theory of algebraic functions.

4.4.1 Definition. If L is a differential extension field of K, then L is called a *transcendental elementary extension* of K if L is of the form

$$L = K(\theta_1, \ldots, \theta_n),$$

where each θ_i is logarithmic or exponential over $K(\theta_1, \ldots, \theta_{n-1})$ and is transcendental (i.e., non–algebraic) over $K(\theta_1, \ldots, \theta_{n-1})$. If there is no condition on being trancendental, then we simply speak of an *elementary extension* of K.

Usually, K will be $\mathbb{Q}(x)$.

4.4.2 Remark. Note that if α is transcendental over K, then $K[\alpha] \cong K[X]$ (the polynomial ring in one variable over K) and $K(\alpha) \cong K(X)$ under the map sending X to α. This enables us to speak of the degree in α of a polynomial expression in α.

4.4.3 Theorem. (Liouville's principle) *Suppose the integral of $f \in K$ exists in an elementary extension L of K, which has the same field of constants as K. Then*

$$\int f = v_0 + \sum_i c_i \log(v_i) \quad \text{or} \quad f = v_0' + \sum_i c_i \frac{v_i'}{v_i}$$

for some constants c_i and elements $v_i \in K$.

Proof. We will prove this in the special case where $L = K(\theta)$ with θ transcendental and logarithmic, $\theta' = u'/u$. The assumption implies that we can write

$$\int f = \frac{a(\theta)}{b(\theta)},$$

with a and b relatively prime and with b monic. Again simplifying, we assume that $b(\theta) = c(\theta)^m$ is the factorization into irreducible factors. Then applying partial fraction expansion yields

$$\frac{a(\theta)}{b(\theta)} = a_0(\theta) + \sum_{j=1}^{m} \frac{a_j(\theta)}{c(\theta)^j}$$

with $\deg(a_j) < \deg(c(\theta))$. Diferentiating both sides gives us

$$f = a_0(\theta)' + \sum_{j=1}^{m} \left[\frac{a_j(\theta)'}{c(\theta)^j} - \frac{j a_j(\theta) c(\theta)'}{c(\theta)^{j+1}} \right].$$

The left–hand side of this equality is independent of θ, so the right–hand side must be independent of θ. If $c(\theta)$ is of positive degree, then $c(\theta)'$ has degree less than $\deg(c(\theta))$ and so is relatively prime with $c(\theta)$. But then there is no term on the right–hand side to cancel

$$\frac{ma_m(\theta)c(\theta)'}{c(\theta)^{m+1}}.$$

This contradiction implies that we get $f = a_0(\theta)'$. But then $a_0(\theta)$ must be (at most) linear[1] in θ, i.e., $a_0(\theta) = c\,\theta + d$ with c a constant in K and $d \in K$. In other words,

$$\int f = d + c\log(u).$$

The proof in the exponential case is similar, the proof in the case of a purely algebraic extension is simpler. The general case needs some care. □

4.4.4 The following theorem provides a criterion for deciding when an integral is elementary (i.e., exists in an elementary extension), given that the integrand is a rational expression in a single transcendental logarithmic variable.

4.4.5 Theorem. *Let K be a differential field with field of constants C and let $K(\theta)$ be a transcendental logarithmic extension of K with the same field of constants. Suppose $p(\theta)/q(\theta)$ satisfies*

a) $\gcd(p(\theta), q(\theta)) = 1$;

b) $\deg p(\theta) < \deg q(\theta)$;

c) $q(\theta)$ *is monic and squarefree.*

Then

$$\int \frac{p(\theta)}{q(\theta)}$$

is elementary if and only if the zeros of $R(z) = \mathrm{Res}_\theta(p(\theta) - zq(\theta)', q(\theta))$ are constants.

Proof. We restrict ourselves to the main aspects of the proof in a special case. First suppose $\int(p(\theta)/q(\theta))$ is elementary. Liouville's principle produces an expression for $p(\theta)/q(\theta)$. Assume for the sake of simplicity that this form is

$$\frac{p(\theta)}{q(\theta)} = v_0(\theta)' + c\,\frac{v(\theta)'}{v(\theta)},$$

[1] This requires proof, for which we refer to the exercises

where we may assume that $v(\theta) \in K[\theta]$ and is squarefree.

Consider the term $v_0(\theta) = a(\theta)/b(\theta)$ for some relatively prime $a(\theta), b(\theta)$. If the derivative $v_0(\theta)' \neq 0$, it will contribute a square to the denominator of $p(\theta)/q(\theta)$ contradicting the assumption. So $v_0(\theta) = a(\theta)$, a polynomial expression in θ.

From the remaining equality

$$\frac{p(\theta)}{q(\theta)} = c\,\frac{v(\theta)'}{v(\theta)}$$

we deduce $q(\theta) = v(\theta)$. But then $p(\theta) - cq(\theta)' = 0$, so $v(\theta) = \gcd(p(\theta) - cq(\theta)', q(\theta))$ showing that c is a zero of the resultant $R(z) = \mathrm{res}_\theta(p(\theta) - zq(\theta)', q(\theta))$.

If $R(d) = 0$, then it follows that $\deg \gcd(p(\theta) - dq(\theta)', q(\theta)) > 0$. Let g be an irreducible factor. Then $g|q = v$ and $g|p - dq' = cv' - dv' = (c-d)v'$. Since $\gcd(v, v') = 1$, we conclude that $c = d$. □

4.4.6 Using our previous results we can refine the theorem in the case the integral is elementary. Then

$$\int \frac{p(\theta)}{q(\theta)} = \sum_{i=1}^{m} c_i \log(v_i(\theta)),$$

where the c_i are the distinct roots of the resultant $R(z)$ and where v_i equals $\gcd(p(\theta) - c_i q(\theta)', q(\theta))$. In fact, $\overline{K} = K(c_1, \ldots, c_m)$ is the minimal algebraic extension of K such that the integral can be expressed in such a form.

4.4.7 Example. Consider the integral

$$\int \frac{1}{\log(x)},$$

i.e., the integrand is of the form above with $p(\theta) = 1$ and $q(\theta) = \theta$, where $\theta = \log(x)$. So we work in the differential extension field $\mathbb{Q}(x, \theta)$ of $\mathbb{Q}(x)$. This is a transcendental logarithmic extension of $\mathbb{Q}(x)$. The resultant

$$R(z) = \mathrm{res}_\theta(1 - z\,q(\theta)', q(\theta)) = \mathrm{res}_\theta(1 - \frac{z}{x}, \theta) = 1 - \frac{z}{x}$$

has no constant zeros, so we conclude that the integral is not elementary.

4.4.8 Example. The integrand of

$$\int \frac{1}{x\log(x)}$$

is of the form $p(\theta)/q(\theta)$ with $p(\theta) = 1/x$ and $q(\theta) = \theta$. This time the resultant is

$$R(z) = \mathrm{res}_\theta(\frac{1}{x} - \frac{z}{x}, \theta) = \frac{1-z}{x},$$

with zero $z = 1$. This implies that the solution is $1 \cdot \log(v)$ with $v = \gcd(p(\theta) - q(\theta)', q(\theta)) = \theta$. So

$$\int \frac{1}{x \log(x)} = \log(\log(x)).$$

4.4.9 Example. The theorem does not apply to the integral $\int \log(x^2+1)$ since the degree of the numerator (the numerator is $\theta = \log(x^2+1)$ of degree 1) is greater than the degree of the denominator. This example represents another extreme of our problem, namely where the expression is polynomial in θ.

4.4.10 Similar to the case of a single logarithmic extension is the case of a single exponential extension.

4.4.11 Theorem. (Single exponential extension) *Let K be a differential field with field of constants C and let $K(\theta)$ be a transcendental exponential extension of K with the same field of constants. Suppose $p(\theta)/q(\theta)$ satisfies*

a) $\gcd(p(\theta), q(\theta)) = 1$ *and* θ *does not divide* $q(\theta)$*;*

b) $\deg p(\theta) < \deg q(\theta)$*;*

c) $q(\theta)$ *is monic and squarefree.*

Then

$$\int \frac{p(\theta)}{q(\theta)}$$

is elementary if and only if the zeros of $R(z) = \mathrm{Res}_\theta(p(\theta) - zq(\theta)', q(\theta))$ *are constants.*

4.4.12 Again, there is an explicit form if the integral is elementary. Let $\theta'/\theta = u'$, then the explicit form is

$$-\sum_i c_i \deg(v_i(\theta))u + \sum_i c_i \log(v_i),$$

where the c_i are the distinct roots of the resultant $R(z)$, and $v_i = \gcd(p(\theta) - c_i q(\theta)', q(\theta))$. The minimal finite extension of K over which such an explicit form exists is K with the c_i adjoined.

4.4.13 Example. This theorem covers the case

$$\int \frac{1}{e^{x^2}+1},$$

but doesn't cover the case $\int e^{-x^2}$. In the first case we find $R(z) = -1 - 2xz$ with no constant zeros, so the integral is not elementary. For the integral

$$\int \frac{x}{e^{x^2}+1},$$

the resultant $R(z) = -2xz - x = x(-2z-1)$ with zero $z = -1/2$. So this integral is elementary. Using 4.4.12, this leads to the following expression for the integral:

$$\frac{1}{2}\,x^2 - \frac{1}{2}\,\log(e^{x^2}+1),$$

since the gcd equals $\theta + 1$.

4.4.14 The various results above describe integrals where the integrand is a quotient in which the degree of the numerator is less than the degree of the denominator. In particular, this excludes polynomial expressions from the range of application of these results! At first glance, they would seem the easiest.

4.4.15 (Polynomial in an exponential) Here, we outline the procedure for elementary integrals where the integrand is a polynomial in e^u and e^{-u}. Let K be a differential field with field of constants C, and let $K(\theta)$ be a transcendental exponential extension with $\theta'/\theta = u'$, where $u \in K$. Consider a Laurent polynomial

$$p(\theta) = \sum_{i=-k}^{l} p_j \theta^j.$$

Using Liouville's principle one can reduce to the case that $p(\theta)$ is of the form

$$p(\theta) = \left[\sum_{j=-k}^{l} q_j \theta^j\right]' + \sum_i c_i \frac{v_i'}{v_i},$$

with the q_j to be determined. Since $(q_j\theta^j)' = (q_j' + ju'\,q_j)\theta^j$, equating coefficients in our two expressions for $p(\theta)$ yields the following system of equations in the q_i:

$$\begin{aligned} p_j &= q_j' + ju'q_j \quad (j \neq 0) \\ p_0 &= (q_0 + \textstyle\sum_i c_i \log(v_i))'. \end{aligned}$$

The second equation can be solved if and only if $\int p_0$ is elementary. If this is not the case, then the original integral is not elementary; otherwise, we proceed to the equations with $j \neq 0$. These are of the form

$$y' + fy = g,$$

with $f, g \in K$. This differential equation is called the *Risch differential equation*. We need to solve it *within* K. If any of these differential equations fails to have a solution in K, then $\int p(\theta)$ is not elementary.

Solving the Risch equations brings us into the realm of symbolic differential equations, the topic of one of the projects for this course.

4.4.16 Example. Consider $\int e^{x^2}$ and let $\theta = e^{x^2} \in \mathbb{Q}(x, e^{x^2})$. Here the situation reduces to one Risch differential equation

$$1 = q_1' + 2x\, q_1,$$

for $q_1 \in \mathbb{Q}(x)$. Substituting a rational expression in x for q_1 easily leads to an inconsistency demonstrating that the integral not elementary.

4.4.17 (Polynomial in a logarithm) There is a similar discussion for the integration of polynomial expressions in θ, where θ is logarithmic. The expression to integrate is of the form

$$p(\theta) = p_l\theta^l + p_{l-1}\theta^{l-1} + \cdots + p_0,$$

where the $p_i \in K$. If the integral is elementary then Liouville's principle (and a little arguing) shows that p is also of the form

$$p(\theta) = v_0(\theta)' + \sum_i c_i \frac{v_i'}{v_i}$$

with $v_0(\theta) \in K[\theta]$ and $v_i \in K$. The c_i are algebraic over K as usual. From the two expressions we obtain that $\deg(v_0(\theta)) = l + 1$, i.e., $v_0(\theta) = q_{l+1}\theta^{l+1} + \cdots + q_0$. Equating both expressions yields a system of equations in the q_j. Depending on whether these equations admit a solution, the integral is elementary. Here is an example of how this works out.

4.4.18 Example. The integral $\int x \log(x)$ is of the required form with $p(\theta) = x\theta$ (where $\theta = \log(x)$), i.e., of degree 1. So we start with

$$\int x\,\theta = q_2\theta^2 + q_1\theta + q_0,$$

and obtain the system of equations

$$q_2' = 0, \quad x = 2q_2\theta' + q_1', \quad 0 = q_1\theta' + q_0'.$$

For the moment we ignore the first equation (it says that q_2 is a constant). Integrating the second equation gives $\frac{1}{2}x^2+\gamma = 2q_2\theta+q_1$, with γ a constant, so that $q_2 = 0$ and $q_1 = \frac{1}{2}x^2+\gamma$. Substituting this into the third equation we $0 = (\frac{1}{2}x^2+\gamma)\theta'+q_0'$, so that $-\frac{x}{2} = \gamma\theta'+q_0'$. Integrating yields $-\frac{x^2}{4} = \gamma\theta+q_0$, so that $\gamma = 0$ and $q_0 = -\frac{x^2}{4}$. We find

$$\int x\,\log(x) = \frac{1}{2}\,x^2\,\log(x) - \frac{x^2}{4}.$$

4.4.19 Example. The integral

$$\int \frac{\log(x)}{x}$$

works out similarly. The integrand is θ/x, with $\theta = \log(x)$. Set

$$\int \frac{\theta}{x} = q_2\theta^2 + q_1\theta + q_0.$$

Then we get the system

$$\begin{aligned} &q_2' = 0, \\ &(2q_2\theta + q_1)' = 1/x, \\ &q_1\theta' + q_0' = 0. \end{aligned}$$

(The second equation uses the fact that q_2 is a constant, a fact following from the first equation.) From the second equation we conclude that $2q_2\theta + q_1 = \theta+\gamma$ (where γ is a constant), and therefore $q_2 = 1/2$ and $q_1 = \gamma$. Substituting this in the third equation yields $q_0 = -\gamma\theta+\delta$, where δ is a constant, so that $\gamma = 0$ since q_0 is does not depend on θ. So q_0 is a constant and we get

$$\int \frac{\log(x)}{x} = \frac{1}{2}\,\theta^2 + \text{constant} = \frac{1}{2}\,\log(x)^2 + \text{constant}.$$

Appendix A

Algebraic prerequisites

A.1 Groups

A.1.1 Definition. A *group* G is a nonempty set with a binary operation that satisfies:

- associativity: $g(hk) = (gh)k$ for all g, h, $k \in G$.
- existence of an identity element: there exists $e \in G$ such that $eg = ge = g$ for all $g \in G$.
- existence of an inverse element: for every $g \in G$, there exists an inverse element $g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = e$.

The number of elements in G is called the *order* of G and it is denoted $|G|$.

A *subgroup* H of G is a nonempty subset of G satisfying:

(a) H is closed under the binary operation: for all h, k in H, we have $hk \in H$.

(b) H is closed under taking inverses: for all $h \in H$, we have $h^{-1} \in H$.

A subgroup is itself a group.

A.1.2 Example. Two important examples:

- the *symmetric group* S_n. This is the group of all bijections of the set $\{1, \ldots, n\}$. The binary operation is the composition of two bijections and the identity element is the identity map.

- The set of invertible $n \times n$–matrices over a field $\mathbb{K}$ is a group with respect to matrix multiplication; the identity matrix serves as identity element. We denote this group by $\mathrm{GL}(n, \mathbb{K})$.

A.1.3 Definition. A group G is *generated* by $g_1, \ldots, g_n \in G$ if any $g \in G$ can be written as a product of the g_i's and g_i^{-1}'s. We denote this as $G = \langle g_1, \ldots, g_n \rangle$; the g_i's are called *generators* of G. In the case that G is finite, it suffices to impose the condition that every element of G can be written as a product of the g_i's.

A *cyclic* group G is a group generated by one of its elements. That means that there exists an element $g \in G$ such that every element of G equals g^n for some $n \in \mathbb{Z}$.

A.1.4 Definition. Let G and H be two groups. A *group homomorphism* $\phi\colon G \to H$ is a function satisfying

$$\phi(ab) = \phi(a)\phi(b) \text{ for all } a, b \in G$$

If ϕ is injective, ϕ is said to be a *monomorphism*. If ϕ is surjective, ϕ is said to be a *epimorphism*. If ϕ is bijective, ϕ is said to be an *isomorphism*.

A.2 Rings, ideals and quotient rings

A.2.1 (Rings) In these notes rings are commutative with unit 1. A ring is called a *domain* (or integral domain) if $ab = 0$ implies $a = 0$ or $b = 0$ for all $a, b \in R$. An element $a \in R$ is called a *unit* (or invertible element) if there exists an element $b \in R$ with $ab = 1$. The set of units is a group with respect to multiplication and is denoted by R^*. An element a is called a *zerodivisor* if there exists a nonzero b with $ab = 0$. A ring is a domain if and only if it contains no nonzero zerodivisors. A ring is called a *field* if every element $\neq 0$ is a unit, i.e., if $R^* = R \setminus \{0\}$. Since the sets of units and zerodivisors are always disjoint, every field is a domain.

A *morphism* (of rings) from R to S is a map $f : R \to S$ such that $f(0_R) = 0_S$, $f(1_R) = 1_S$, $f(a + b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$ for all $a, b \in R$. A bijective morphism is called an *isomorphism*. The inverse of an isomorphism is also an isomorphism. In this case the rings R and S are called *isomorphic*; this is denoted by $R \cong S$.

A.2.2 (Ideals) An *ideal* I in the ring R is a nonempty subset such that

a) $a, b \in I \Rightarrow a + b \in I$;

b) $a \in I$ and $r \in R \Rightarrow ra \in I$.

If S is a subset of R, then (S) denotes the ideal *generated by* S. It is the smallest ideal in R containing I. An explicit description is

$$(S) = \{r_1 s_1 + \cdots + r_m s_m \mid r_1, \ldots, r_m \in R,\ s_1, \ldots, s_m \in I \,\text{and}\, m \in \mathbb{Z}_{\geq 1}\}.$$

If I and J are ideals, then $I + J = \{a + b \mid a \in I,\ b \in J\}$ is also an ideal in R, the *sum* of I and J. This notion generalizes to the sum of an arbitrary number of ideals. The ideal I is a *maximal ideal* if $I \neq R$ and if there exists no ideal strictly between I and R, i.e., if J is an ideal in R satisfying $I \subset J \subset R$, then $J = I$ or $J = R$. The ideal I is called a *prime ideal* if

$$a\,b \in I \Rightarrow a \in I \text{ or } b \in I.$$

Every maximal ideal is a prime ideal, but the converse does not hold in general. For instance, $(2) \subset \mathbb{Z}[X]$ is prime, but not maximal.

If J is an ideal in the ring S and if $f : R \to S$ is a morphism, then $f^{-1}(J)$ is an ideal in R. For $J = (0)$, this ideal is called the *kernel* of f and denoted by $\ker(f)$.

A.2.3 (Quotient rings) If I is an ideal in R, then the *quotient ring* R/I is the ring whose elements are the equivalence classes with respect to the equivalence relation

$$a \sim b \Leftrightarrow a - b \in I.$$

The equivalence class containing a is denoted by $a+I$, $\overline{a}$ or just a if confusion is not likely to occur. The ring operations on R/I are

a) (addition) (a+I) + (b+I)=(a+b)+I;

b) (multiplication) (a+I) (b+I)=(ab)+I.

The zero element is $0 + I$ and the unit element $1 + I$. The map

$$p : R \to R/I, \quad p(a) = a + I$$

is a morphism and often called the *canonical map* (or natural map) from R to R/I. The kernel of p is I.

A.2.4 Theorem. *Let I be an ideal in the ring R.*

a) I is a prime ideal if and only if R/I is a domain;

b) I is a maximal ideal if and only if R/I is a field.

A.2.5 Theorem. (First Isomorphism Theorem) *If $f : R \to S$ is a morphism then f induces a unique injective morphism $\overline{f} : R/\ker(f) \to S$ such that $f(a) = \overline{f}(\overline{a})$ for every $a \in R$. If f is surjective, then the induced morphism $\overline{f}$ is an isomorphism.*

A.2.6 If J is an ideal in R then the ideals of R/J are in one–to–one corespondence with the ideals in R containing J under the correspondence $I \leftrightarrow p(I)$, where $p : R \to R/J$ is the canonical map. The image $p(I)$ of an ideal I containing J is often denoted by I/J. The kernel of the composition of the natural maps $R \to R/J \to (R/J)/(I/J)$ is equal to I. Applying the First Isomorphism Theorem A.2.5 yields

A.2.7 Theorem. (Second Isomorphism Theorem) *If $J \subset I$ are ideals in R, then*

$$R/I \cong (R/J)/(I/J).$$

A.3 Finite fields

For p prime, the ring $\mathbb{Z}/p\mathbb{Z}$ is a finite field. But there are many more finite fields.

If F is a field the kernel of the morphism

$$\mathbb{Z} \to F,\ m \mapsto \operatorname{sign}(m)(1 + \cdots + 1) \text{ (with } |m| \text{ terms 1) if } m \neq 0 \text{ and } 0 \mapsto 0$$

has kernel (0) or $p\mathbb{Z}$ for some prime p. If F is a finite field only the latter case can occur and F is said to have *characteristic* p. (If for a field the kernel of the map is (0) the field is said to have characteristic 0.) The First Isomorphism Theorem A.2.5 shows that a finite field contains a copy of the field $\mathbb{Z}/p\mathbb{Z}$ with p elements. In particular, the field F is a finite dimensional vector space over $\mathbb{Z}/p\mathbb{Z}$. If the dimension is m then F has p^m elements. So the number of elements of a finite field is necessarily a prime power.

A.3.1 Theorem. *Let F be a field with p^m elements where p is a prime.*

a) Every element $a \in F$ satisfies $a^{p^m} = a$.

b) The group of units F^* is cyclic of order $p^m - 1$.

c) If L is also a field with p^m elements then F and L are isomorphic.

d) There exists an irreducible polynomial $f \in (\mathbb{Z}/p\mathbb{Z})[X]$ such that $F \cong (\mathbb{Z}/p\mathbb{Z})[X]/(f)$.

A.3.2 Since finite fields with the same number of elements are isomorphic, one uses a single notation for a field with q (with q a prime power) elements: $\mathbb{F}_q$.

A.4 Resultants

A.4.1 Let $f, g \in k[X]$ be two polynomials of positive degree. The resultant of f, g is an element of k that determines if f and g have a common factor or not: f and g have a nonconstant common factor if and only if the resultant is 0. If f and g have a common factor h, then there is a polynomial relation of the form $Af + Bg = 0$ with $\deg(A) < \deg(g)$ and $\deg(B) < \deg(f)$: simply take $A = g/h$ and $B = -f/h$. The following lemma states that the converse also holds.

A.4.2 Lemma. *Let f and g be of positive degree. Then f and g have a factor in common if and only if there exist nontrivial polynomials A and B satisfying* $\deg(A) < \deg(g)$ *and* $\deg(B) < \deg(f)$ *such that*

$$Af + Bg = 0.$$

Proof. One implication was shown above, so we assume that we have a relation $Af + Bg = 0$ as in the statement of the lemma and that f and g are relatively prime. From the last assumption, we deduce from the euclidean algorithm that there exists a relation $uf + vg = 1$. Multiplying this relation by B we obtain $ufB + vgB = B$. Using $Bg = -Af$ we find $B = ufB - vAf = (uB - vA)f$ contradicting the assumption $\deg(B) < \deg(f)$. □

A.4.3 To check the existence of A and B comes down to solving a system of linear equations. If

$$f = f_0 + f_1X + \cdots + f_mX^m \quad (f_0 \neq 0,\ f_m \neq 0) \quad \text{and} \quad g = g_0 + g_1X + \cdots + g_nX^n \quad (g_0 \neq 0,\ g_n \neq 0)$$

and if we write

$$\begin{aligned} A &= a_0 + a_1X + \cdots + a_{n-1}X^{n-1}, \\ B &= b_0 + b_1X + \cdots + b_{m-1}X^{m-1}, \end{aligned}$$

then substitution in the equation $Af + Bg = 0$ gives a system of linear equations, whose coefficient matrix is the transpose of the $(m+n)$–matrix

$$A(f,g) = \begin{pmatrix} f_0 & f_1 & \cdots & f_m & & & \\ & f_0 & \cdots & f_{m-1} & f_m & & \\ & & \vdots & & & & \\ & & f_0 & f_1 & \cdots & f_m & \\ g_0 & g_1 & \cdots & & g_n & & \\ & g_0 & \cdots & & & g_n & \\ & & \vdots & & & & \\ & & g_0 & & & \cdots & g_n \end{pmatrix}$$

This system has a nontrivial solution if and only if the determinant $\det(A(f,g)) = 0$.

Note that if the resultant is 0, the gcd of f and g is a nontrivial common factor.

A.4.4 Definition. The determinant of the matrix in (A.4.3) is called the *resultant* of f and g and denoted by $R(f,g)$. The notation $R(f,g,x)$ is used to emphasise that we view f and g as polynomials in x.

A.4.5 Remark. Although not stated explicitly, the resultant is also usefull for polynomials in several variables. For instance, the polynomials $X^3+XY^2+YX^2+Y^3$ and $X+Y$ can be viewed as polynomials in x with coefficients in $\mathbb{Q}(y)$: $Y^3 + (Y^2)X + Y X^2 + X^3$ and $Y + X$. The resultant is

$$\begin{vmatrix} Y^3 & Y^2 & Y & 1 \\ Y & 1 & 0 & 0 \\ 0 & Y & 1 & 0 \\ 0 & 0 & Y & 1 \end{vmatrix} = 0,$$

showing that the two polynomials have a common factor in $\mathbb{Q}(Y)[X]$. By clearing denominators, a common factor in $\mathbb{Q}[X,Y]$.

A.4.6 Example. Let $f = ax^2 + bx + c$ with $a \neq 0$. Then

$$R(f,f') = \begin{vmatrix} c & b & a \\ b & 2a & 0 \\ 0 & b & 2a \end{vmatrix} = -a(b^2 - 4ac).$$

In this expression, the well-known discriminant $b^2 - 4ac$ of a quadratic polynomial occurs. We regain the result that f has a zero of multiplicity 2 if and only if $b^2 - 4ac = 0$.

A.4.7 Proposition. *Let $f, g \in k[X]$ be polynomials of positive degree.*

a) *f and g have a common factor if and only if $R(f, g) = 0$.*

b) *$R(f, g) \in (f, g)$, i.e., there exist polynomials u and v such that $uf + vg = R(f, g)$.*

Proof. The first item was shown above, so we turn to b). If $R(f, g) = 0$, then we can take $u = v = 0$. If $R(f, g) \neq 0$, then f and g are relatively prime, so there exist $\tilde{u}$ and $\tilde{v}$ with $\tilde{u}f + \tilde{v}g = 1$. Multiplying through by $R(f, g)$ gives the required identity. □

A.4.8 For computations, the determinant in the definition of the resultant is quite inefficient. In the exercises a more efficient way to compute the resultant is described.

A.5 Groups

One of the most important notions in Mathematics is the notion of a group:

A.5.1 Definition. A *group* G is a non empty set with a binary operation that satisfies:

- associativity: $g(hk) = (gh)k$ for all g, h, $k \in G$.
- existence of the identity element: there exists $e \in G$ such that $eg = ge = g$ for all $g \in G$.
- existence of an inverse element: for every $g \in G$, there exists an inverse element $g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = e$.

The number of elements in G is called the *order* of G and it is denoted $|G|$.

A.5.2 Example. The integers $\mathbb{Z}$, the rationals $\mathbb{Q}$ and the real $\mathbb{R}$ are groups with respect to the addition $+$ and identity element 0. However, $\mathbb{Z}$ is not a group with respect to the multiplication.

A.5.3 Example. A very important example of group is the *symmetric group* S_n. This is the group of all bijections of the set $\{1, \ldots, n\}$. The binary operation is the composition of two bijections and the identity element is the identity map.

A.5.4 Example. The set of invertible matrices over a field $\mathbb{K}$ is a group denoted by $\mathrm{GL}(n, \mathbb{K})$.

A.5.5 Definition. A finite group G is *generated* by $g_1, \ldots, g_n \in G$ if any $g \in G$ can be written as a product of the g_i's. We denote $G = \langle g_1, \ldots, g_n \rangle$ and the g_i's are called the *generators* of G.

A *cyclic* group G is a group generated by one of its elements. That means there exists an element $g \in G$ such that $G = \{e, g, g^2, \ldots, g^n\}$ where e denotes the identity element. We can also write $G = \langle g \rangle$.

A.5.6 Definition. Let G and H be two groups. A *group homomorphism* $\phi\colon G \to H$ is a function satisfying

$$\phi(ab) = \phi(a)\phi(b) \text{ for all } a, b \in G$$

If ϕ is injective, ϕ is said to be a *monomorphism.*If ϕ is surjective, ϕ is said to be a *epimorphism.* If ϕ is bijective, ϕ is said to be a *isomorphism.*

A.5.7 Example. Consider the group homomorphism $\phi\colon S_n \to \mathrm{GL}(n, \mathbb{K})$ defined by: to each element $\sigma \in S_n$ we associate the matrix $\phi(\sigma) = (a_{ij})_{i,j=1}^n$ with

$$a_{ij} = \begin{cases} 1 \text{ if } \sigma(j) = i, \\ 0 \text{ otherwise} \end{cases}$$

It is easy to see that S_n and $\phi(S_n)$ are isomorphic. Using this representation of S_n we can say that S_n is a subset of $\mathrm{GL}(n, \mathbb{K})$.

For instance, the permutation $\sigma \in S_3$ with $\sigma(1) = 2$, $\sigma(2) = 1$, $\sigma(3) = 3$ has matrix:

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

A.5.8 Definition. A *subgroup* H of G is a non-empty subset of G satisfying:

(a) H is closed under the binary operation: for all h, k in H, we have $hk \in H$.

(b) H is closed under taking inverses: for all $h \in H$, we have $h^{-1} \in H$.

A.5.9 Example. For every $m \in \mathbb{Z}$, the subset $m\mathbb{Z} = \{mn \mid n \in \mathbb{Z}\}$ is a subgroup of the group $\mathbb{Z}$.

A.5.10 Example. Consider the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

The set $\{I, A, A^2, A^3\}$ is a subgroup of $\mathrm{GL}(2, \mathbb{C})$. Besides, this is a cyclic group generated by A.

A.5.11 Example. An example of a subgroup we will use later is a *finite matrix group*. It is defined as being a non-empty finite subset of $\mathrm{GL}(n, \mathbb{C})$ which is closed under matrix multiplication. In the exercises, you will show that this is actually a subgroup of $\mathrm{GL}(n, \mathbb{C})$. The example A.5.10 is a finite matrix group. The symmetric group S_n seen as permutation matrices is also a finite matrix group.

A.5.12 Example. The set of $\det 1$ matrices of order n over a field $\mathbb{K}$ is a subgroup of $\mathrm{GL}(n, \mathbb{K})$ denoted by $\mathrm{SL}(n, \mathbb{K})$.

Bibliography

[1] W. Adams, P. Loustaunau (1994). *An introduction to Gröbner bases*. Graduate Studies in Mathematics 3, AMS, Providence.

[2] M. Bronstein (1997). *Symbolic integration I*. Algorithms and Computation in Mathematics, Vol. 1, Springer, Berlin etc.

[3] B. Buchberger, F. Winkler (eds.) (1998) *Gröbner Bases and Applications*, London Mathematical Society Lecture Notes 251, Cambridge University Press.

[4] A.M. Cohen, H. Cuypers, H. Sterk (eds.) (1999). *Some tapas of computer algebra*, Springer-Verlag, Berlin etc.

[5] D. Cox, J. Little, D. O'Shea (1992, 1997). *Ideals, varieties and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Undergraduate Texts in Mathematics, Springer Verlag, Berlin etc.

[6] K.O. Geddes, S.R. Czapor, G. Labahn (1992). *Algorithms for computer algebra*. Kluwer Academic Publishers, Boston/Dordrecht/London.

[7] B. Sturmfels (1993) *Algorithms in Invariant Theory*, Text and Monographs in Symbolic Computation, Springer-Verlag, Wien.

Index