Summary.

This report describes techniques for resolving systems of polynomial equations and inequalities. The general technique used is *cylindrical algebraic decomposition*, which decomposes space into a number of regions, on each of which the equations and inequalities have the same sign. Most of the report is spent describing the algebraic and algorithmic pre-requisites (resultants, algebraic numbers, Sturm sequences etc.), and then describing the method, first in two dimensions and then in an arbitrary number of dimensions.

Contents.

Preface.

This report contains the notes from a graduate course given at the Numerical Analysis and Computer Science (NADA) Department. The aim of the course was to cover material from computer algebra relevant to the solution of systems of polynomial equations and inequalities. The ultimate application area envisaged was that of motion planning, but these notes do not address that area directly.

The audience (and potential readers) were assumed to have some knowledge of computer algebra, at the level of polynomial manipulation and the complexity of elementary algorithms. More advanced ideas, such as resultants and algebraic numbers, are developed as required. Theorems whose proof illuminates the algorithm, or whose proofs are not too remote from the subject area, are proved: some others are just stated.

The course culminates in cylindrical algebraic decomposition, after Collins [1975], Arnon *et al.* [1984] and McCallum [1985]. There was no deliberate intention of lecturing on the author's research, but some results seem to be new. These are mainly the inequality in I.5 and some of the complexity analyses in chapters II and IV.

Acknowledgements.

The author is grateful to Professor Stefan Arnborg, for inviting him to NADA and arranging the course, as well as for many useful and stimulating discussions on all parts of the course. Hans Riesel very kindly entertained me in the Archipelago while proposition I.5.8 was proved. The audience's questions and comments have, I hoped, improved the presentation of this report, and I am grateful to them for that, and for their patience.

No-one could lecture on this subject without owing an enormous debt to Professor George Collins, who has worked on cylindrical decomposition for thirty years, and who is responsible, either directly or through his students, for nearly all the material presented in chapters II and IV, as well as much of the supporting technology in chapter I.

Introduction.

In this course, we look at polynomial equations, or systems of polynomial equations, and their solution. We will also treat mixtures of equations and inequalities. By "solution", we will generally (except in chapter III) mean "solution over the real numbers", since this is what the applications mostly require. After all, what use is it to discover that the robot is unstable if the arm is rotated through i radians?

The reader may feel that the limitation to polynomial equations is unrealistic, since many problems seem to involve trigonometric functions. But, if we regard $\sin\theta$ and $\cos\theta$ as the variables, rather than θ, and link them via $\sin^2\theta + \cos^2\theta = 1$, we have a system of polynomial equations. This technique also avoids having to deal with infinite values such as $\tan\pi/2$, since this becomes the quotient of two variables, and we can then clear denominators to arrive at a polynomial equation in which all values are finite.

Most of the input equations we have to treat in robotics applications will either be linear equations or equations of the form

$$(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = k,$$

which expresses the fact that (x_1, y_1, z_1) and (x_2, y_2, z_2) are part of the same rigid body. Computer-aided design tends to restrict itself to polyhedra (i.e. linear equations) and spheres or cylinders (i.e. quadratic equations). When more general objects are treated, they are usually treated as a compositum of polynomially-defined patches, i.e. as a system of polynomial equations and inequalities.

Notation. We will use $\mathbf{Z}$ to stand for the integers, $\mathbf{Q}$ to stand for the rational numbers, $\mathbf{R}$ to stand for the real numbers and $\mathbf{C}$ to stand for the complex numbers. In addition we will use $\mathbf{A}$ to stand for the algebraic numbers (see section I.4 below). The notation $R[x]$ will mean "polynomials in x, with coefficients coming from R", and $F(x)$ will mean "rational functions in x, with coefficients coming from F". Additional notation will be introduced as necessary, for example $\delta(p)$ and $\|p\|$ in section I.5.

We will use a little complexity theory to measure the cost of our algorithms. We assume that the reader is familiar with the O notation for the asymptotic order of growth of a function (normally running time). By **operation**, we mean an elementary operation acting on fixed-size data and taking bounded time: whether this is bit operations or machine word operations only affects the constant implied by O. The phrase **arithmetic operation** will mean an operation on integers, of potentially unbounded length. We recall that addition of two numbers less than 2^n takes $O(n)$ operations, and multiplication takes $O(n^2)$ operations, or $O(n \log n \log\log n)$ by "fast" methods [Aho, Hopcroft & Ullman, 1974].

No change has been made to the body of this text for the Bath reprint, though a few foot-notes have been added. A few additional developments ought to be reported: the author and J. Heintz (to appear in *J. Symbolic Computation*) have shown that there is a doubly-exponential lower bound for the space complexity of cylindrical algebraic calculation, and the author and M. Mignotte (submitted to *SIAM J. Computing*) have further investigated the optimality of root estimates (see Proposition I.4.4).

I. One Polynomial Equation

In this section, we discuss the, apparently trivial, case of one polynomial equation, and show that there is more to it than meets the eye.

1. Low Degree Equations.

A linear equation is certainly trivial. A quadratic equation $ax^2 + bx + c = 0$ has, as we know, solutions

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Assuming, as we generally do, that the coefficients are real, we know that this equation has real roots if, and only if, $b^2 - 4ac \geq 0$

For cubic equations, the situation is more complex. The reduced cubic equation $x^3 + bx + c$ has roots of the form

$$\sqrt[3]{\frac{-c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} + \sqrt[3]{\frac{-c}{2} - \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}},$$

but this would appear to have two cube roots and two square roots, thus giving a choice of 36 possibilities, while a cubic, as we know, has three roots. Some of this complexity is easily removed by remarking that the square roots are really the same, and that we must make a consistent choice of signs for them. Furthermore, changing the choice of sign merely interchanges the two cube roots, so does not affect the value of the formula. The two cube roots are, in fact, related, since their product should be $-b/3$. Hence we can re-express our formula as

$$\sqrt[3]{\frac{-c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} - \frac{b}{3\sqrt[3]{\frac{-c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}}},$$

provided that we choose the sign of the square root to avoid $0/0$ (unless both b and c are zero). This gives us one root, and, by taking different values of the cube root, we obtain the other two roots of the cubic. For the general cubic $px^3 + qx^2 + rx + s$, this formula becomes the frightening

$$\frac{-q}{3p} + \sqrt[3]{-\frac{s}{2p} + \frac{qr}{6p^2} - \frac{q^3}{27p^3} + \sqrt{\left(\frac{s}{2p} - \frac{qr}{6p^2} + \frac{q^3}{27p^3}\right)^2 + \left(\frac{r}{3p} - \frac{q^2}{9p^2}\right)^3}} - \frac{\frac{r}{p} - \frac{q^2}{3p^2}}{3\sqrt[3]{-\frac{s}{2p} + \frac{qr}{6p^2} - \frac{q^3}{27p^3} + \sqrt{\left(\frac{s}{2p} - \frac{qr}{6p^2} + \frac{q^3}{27p^3}\right)^2 + \left(\frac{r}{3p} - \frac{q^2}{9p^2}\right)^3}}}$$

There is one puzzling phenomemon. A cubic has either one or three real roots (excluding the case of co-incident roots), while this formula would appear to have two cases:

(a) $c^2/4 + b^3/27$ is positive, so that the square root is real, and hence there is one real solution coming out of the real cube root;

(b) $c^2/4 + b^3/27$ is negative, so that the square root is imaginary, and we are combining the cube roots of imaginary numbers.

Oddly enough, case (b) corresponds to the case of three real roots, when, as if by magic, the imaginary parts all cancel out, whereas in case (a), where there is one evident real root, the other two roots are complex conjugates. The moral of this is that we may need complex numbers in order to compute purely real numbers via this formula.

For the reduced quartic equation $x^4 + bx^2 + cx + d$, there is the following formula: the roots are

$$\begin{aligned}
&\frac{1}{2}\left(\sqrt{\alpha} + \sqrt{\beta} + \sqrt{\gamma}\right)\\
&\frac{1}{2}\left(\sqrt{\alpha} - \sqrt{\beta} - \sqrt{\gamma}\right)\\
&\frac{1}{2}\left(-\sqrt{\alpha} + \sqrt{\beta} - \sqrt{\gamma}\right)\\
&\frac{1}{2}\left(-\sqrt{\alpha} - \sqrt{\beta} + \sqrt{\gamma}\right)
\end{aligned}$$

where α, β and γ are the three roots of

$$y^3 + 2by^2 + (b^2 - 3d)y - c^2,$$

the *resolvent* of the original equation, and we have chosen the signs of square roots so that $\sqrt{\alpha}\sqrt{\beta}\sqrt{\gamma} = c$. The reader is left to consider the case of the general quartic, and the generalisation of the various complications that were mentioned for the cubic (see van der Waerden [1960] Section 59).

2. General Equations.

The previous section should have convinced the reader that explicit solutions in terms of n-th roots (otherwise known as *radicals*) are perhaps not as simple, or as useful, as they may seem. There is a further objection.

Theorem [Abel]. *The general equation of degree greater than 4 is not soluble by radicals.*

We cannot give an account of this theorem here, which belongs to that branch of mathematics generally known as "Galois Theory". It suffices to remark that the usual proofs give little indication as to when particular equations might be soluble in terms of radicals, such as $x^n = a$, which is always soluble. Recently some progress has been made on efficient algorithms for discovering whether an equation is soluble in radicals — see Landau & Miller [1983] and Landau [1984]. These methods do not, though, readily give the expression in terms of radicals.

Hence we are led to consider roots of polynomials which do not necessarily have solutions in radicals, or where we may not know the solution in radicals even if it exists. A natural solution to this problem would be to abandon symbolic working, and to work with numerical approximations, i.e. to replace $\sqrt{2}$ by $1.4142\ldots$.

Unfortunately the roots of polynomials may well be very ill-conditioned functions of the coefficients of the polynomials. The classic case of this is when the polynomial has repeated, or nearly repeated roots. The case of repeated roots can be solved by the device of *square-free decomposition*.

Proposition. *If $p(x)$ is a polynomial, then $p/\gcd(p, p')$ has the same roots as p, but each root occurs only once.*

Proof. Write $p(x) = \prod_{i=1}^{n}(x - \alpha_i)^{n_i}$, where the α_i are the distinct roots of p. Then

$$p'(x) = \sum_{i=1}^{n} n_i(x - \alpha_i)^{n_i - 1} \prod_{\substack{j=1 \\ i \neq j}}^{n} (x - \alpha_j)^{n_j}.$$

$x - \alpha_i$ divides p' exactly $n_i - 1$ times, since it divides one summand $n_i - 1$ times and the rest n_i times. Hence the greatest common divisor is $\prod_{i=1}^{n}(x - \alpha_i)^{n_i - 1}$, and dividing this out gives the required result.

Assuming that we are prepared to define the concept of "greatest common divisor" when our polynomials might have floating-point coefficients, we can then reduce our problems to finding roots of polynomials with only simple roots. Unfortunately, this does not deal with "nearly repeated" roots. One might hope that these were rare, but in fact "nearly repeated" covers a very wide area.

The following example, due to Wilkinson [1959], illustrates this. Let p have roots at -1 , -2 , ..., -20, so that $p = (x+1)(x+2)\ldots(x+20) = x^{20} + 210x^{19} + \cdots + 20!$. Consider now the polynomial $p(x) + 2^{-23}x^{19}$. One might expect this to have twenty real roots close to the original ones, but in fact it has ten real roots,

at approximately -1, -2, $\ldots -7$, -8.007, -8.917 and -20.847, and five pairs of complex conjugate roots, $-10.095 \pm 0.6435i$, $-11.794 \pm 1.652i$, $-13.992 \pm 2.519i$, $-16.731 \pm 2.813i$ and $-19.502 \pm 1.940i$. It must be emphasised that these *are* the roots of the perturbed polynomial, and that the difference is caused by a change in the roots, not by any numeric solution process. In section 6, we will demonstrate that there are only 10 real roots. Hence a purely numeric approach to the manipulation of roots of polynomials seems doomed to error.

We will spend the rest of this chapter outlining a semi-algebraic approach, which will combine the best, we hope, of algebraic accuracy and numeric information.

One concept from the above discussion will be needed later. We can write any polynomial p as $\prod p_i^i$, where the p_i are relatively prime square-free polynomials. We saw above that $\gcd(p,p') = \prod p_i^{i-1}$ and that $p/\gcd(p,p') = \prod p_i$. Writing q for this second polynomial, we now see that $\gcd(\gcd(p,p'),q) = \prod_{i>1} p_i$, and so $q/\gcd(\gcd(p,p'),q) = p_1$. By applying the same process to $\gcd(p,p')$, we can deduce p_2, and so on. Hence the entire square-free decomposition of p can be computed via gcd calculations.

3. The Resultant.

In this section we introduce a piece of mathematical technology which will be useful throughout this course, the *resultant*. Throughout this section p and q will be assumed to be two polynomials with coefficients coming from some integral domain R. We will assume that $p(x) = a_m x^m + \cdots + a_0$ and that $q(x) = b_n x^n + \cdots + b_0$, and that $m > 0$, $n > 0$.

Definition. Sylvester's matrix *of p and q is the $m+n$ by $m+n$ matrix*

$$\begin{pmatrix}
a_m & a_{m-1} & \ldots & a_0 & 0 & 0 & \ldots & 0 \\
0 & a_m & a_{m-1} & \ldots & a_0 & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ldots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \ddots & \ddots & \ldots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & a_m & a_{m-1} & \ldots & a_0 \\
b_n & b_{n-1} & \ldots & b_0 & 0 & 0 & \ldots & 0 \\
0 & b_n & b_{n-1} & \ldots & b_0 & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ldots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \ddots & \ddots & \ldots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & b_n & b_{n-1} & \ldots & b_0
\end{pmatrix}$$

where there are n rows of the a_i followed by m rows of the b_i.

Definition. *The* **resultant** *of p and q, denoted by* $\operatorname{res}(p,q)$ *is the determinant of Sylvester's matrix. If we wish to emphasise that we are regarding p and q as polynomials in x, we will write* $\operatorname{res}_x(p,q)$.

Proposition 1. $\operatorname{res}(p,q) = (-1)^{mn} \operatorname{res}(q,p)$.

There is a strong connection between Sylvester's matrix and greatest common divisors. In fact, the operation of Gaussian elimination in Sylvester's matrix is equivalent to the calculations one normally performs when calculating a greatest common divisor of two polynomials by repeated subtraction. We shall content ourselves with the following remark in this area.

Proposition 2. $\operatorname{res}(p,q)$ *is zero if, and only if, p and q have a non-constant common divisor.*

Proof. Suppose first that p and q have a non-trivial common divisor g. Write $\hat{p} = p/g$, $\hat{q} = -q/g$, so that $\hat{q}p + \hat{p}q = 0$ and $\hat{q}$ has degree less than n, while $\hat{p}$ has degree less than m. Suppose, in fact, that

$\hat{p} = c_{m-1}x^{m-1} + \cdots + c_0$, while $\hat{q} = d_{n-1}x^{m-1} + \cdots + d_0$. Now consider the matrix product

$$(d_{n-1} \quad \ldots \quad d_0 \quad c_{m-1} \quad \ldots \quad c_0) \begin{pmatrix} a_m & a_{m-1} & \ldots & a_0 & 0 & 0 & \ldots & 0 \\ 0 & a_m & a_{m-1} & \ldots & a_0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ldots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ldots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & a_m & a_{m-1} & \ldots & a_0 \\ b_n & b_{n-1} & \ldots & b_0 & 0 & 0 & \ldots & 0 \\ 0 & b_n & b_{n-1} & \ldots & b_0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ldots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ldots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & b_n & b_{n-1} & \ldots & b_0 \end{pmatrix}$$

This is a $m+n$ column vector, whose elements are the coefficients of the polynomial $\hat{q}p+\hat{p}q$, treated as having degree $m+n-1$, even though its actual degree might be smaller. But this polynomial is zero, so we have found a non-zero vector which Sylvester's matrix annihilates, and hence this matrix must have determinant 0.

Conversely, if this matrix has determinant 0, there is a non-trivial vector which it annihilates. Call this vector $(d_{n-1}, \ldots, d_0, c_{m-1}, \ldots, c_0)$, and construct polynomials $\hat{p}$ and $\hat{q}$ as above. Then $\hat{q}p + \hat{p}q = 0$, and so q divides $\hat{q}p$. But if p and q have no common factor, this would imply that q divides $\hat{q}$, which is impossible since $\hat{q}$ is of lower degree than q.

Proposition 3. *If the α_i are the roots of p and the β_j are the roots of q, then*

$$\mathrm{res}(p,q) = a_m^n b_n^m \prod_{i=1}^{m} \prod_{j=1}^{n} (\alpha_i - \beta_j). \qquad (*)$$

Proof. It is sufficient to prove this result in the special case when R is $S[a_m, b_n, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n]$, an extension of an integral domain S by $m+n+2$ new indeterminates, since any particular case can be obtained by substituting special values for these indeterminates. Another way of saying this is that we intend to prove the proposition in the most general case possible, when there are no relationships between the items appearing on the right-hand side of $(*)$.

Each of the a_i can be expressed as polynomials in the α_i multiplied by a_m, and similarly for the b_j. For example, $a_0 = (-1)^m a_m \prod \alpha_i$. Hence we can express $\mathrm{res}(p,q)$, which is a sum of $\pm$ products of the a_i and b_j by Cramer's rule, as $a_m^n b_n^m$ times a sum of $\pm$ products of the α_i and β_j.

How many α_i and β_j appear in each summand? The surprising answer is that a total of mn appear in every summand. mn certainly appear in the summand $a_m^n b_0^m$, since $b_0 = b_n \prod_{j=1}^{n} \beta_j$. Henceit is sufficient to show that ecah term in Cramer's rule contributes the same number as this term does. As we walk along a row in Sylvester's matrix, each entry contains one more α_i or β_j than the term on its left (it does not matter how many α_i or β_j we say appear in the zero entries, so we can make this statement about them as well). Hence increasing the indices along the rows increases the number of α_i and β_j, and vice versa. But all sums appearing in Cramer's rule have precisely the same sum of the row indices, viz. $(m+n)(m+n+1)/2$. Hence all terms have the same number of α_i and β_j.

So both the left-hand and right-hand sides of $(*)$ are of the form $a_m^n b_n^m$ times a polynomial in the α_i and β_j, all of whose terms contains mn occurrences of α_i and β_j. But the resultant is divisible by $\alpha_i - \beta_j$ for every i and j, since it is zero when $\alpha_i = \beta_j$, i.e. when p and q have a common factor of $x - \alpha_i$, by the previous proposition. Hence the left-hand side of $(*)$ is divisible by the right-hand side. Since they have the same degree, they are therefore equal up to a constant factor, and this factor has in fact got to be unity, since both contain the term $a_m^n b_0^m = a_m^n b_n^m (-1)^{mn} \left(\prod_{j=1}^{n} \beta_j \right)^n$.

Corollary 1. $\mathrm{res}(p,q) = a_m^n \prod_{i=1}^{m} q(\alpha_i)$.

Corollary 2. $\mathrm{res}(p,q) = (-1)^{mn} b_n^m \prod_{j=1}^{n} p(\beta_j)$.

Sylvester's matrix is not normally the best way of calculating resultants. Collins [1971] presents a modular algorithm for calculating resultants over the integers or over polynomial domains, whose running time for polynomials over the integers of degree n and coefficients with at most d digits is $O(n^3d + n^2d^2)$.

We should also remark that resultants may, in practice, be quite large objects. If the coefficients of p and q have d and e digits respectively (more precisely if $\sqrt{\sum a_i^2}$, otherwise known as $\|p\|_2$, has d digits and $\sqrt{\sum b_j^2}$ has e digits), then by Hadamard's bound [1893; Mignotte, 1982] on determinants, the resultant will have no more than $nd + me$ digits. The same bound applies if the coefficients of p and q are polynomials in another variable, with maximal degree d and e respectively. In practice, resultants do tend to be of this order of magnitude, which means that they can be troublesome to work with.

Definition. *The* **discriminant** *of a polynomial* p, $\operatorname{disc}(p)$, *with leading coefficient* a_n *and roots* $\alpha_1, \ldots, \alpha_n$ *is defined to be* $a_n^{2n-2}\prod_{1\le i<j\le n}(\alpha_i - \alpha_j)^2$.

The discriminant of $ax^2 + bx + c$ is $b^2 - 4ac$, which accords with the usual definition. Furthermore, for the cubic $x^3 + bx + c$, the discriminant is $4b^3 + 27c^2$, which is essentially the quantity appearing under the square root sign in the formula of section 1. The discriminant has some useful geometric properties, which we will meet in chapter II. There is an alternative form of the product expression for the discriminant, which we will need later.

Proposition 4.

$$\prod_{1\le i<j\le n}(\alpha_i - \alpha_j) = (-1)^{n(n-1)/2}\begin{vmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \ldots & \alpha_n^{n-1} \end{vmatrix}.$$

Proof. We will actually prove the equivalent result, that the determinant on the right-hand side is equal to $\prod_{1\le i<j\le n}(\alpha_j - \alpha_i)$.

The proof is similar to that of proposition 3, and will be conducted in the settting where the α_i are indeterminates. Both sides of this equation are polynomials in the α_i, of degree precisely $n(n-1)/2$. If we subtract the i-th column from the j-th column in the determinant (which will not change the determinant's value) we obtain the column $(0, \alpha_j - \alpha_i, \alpha_j^2 - \alpha_i^2, \ldots, \alpha_j^{n-1} - \alpha_i^{n-1})$. Each of these entries, and hence the determinant, is divisible by $\alpha_j - \alpha_i$. Therefore the two sides of the identity to be proved divide one another, and hence are equal up to a constant factor. To determine this factor, we observe that taking all the α_j terms from the product $\prod_{1\le i<j\le n}(\alpha_j - \alpha_i)$ gives us $\alpha_1^0\alpha_2^1\ldots\alpha_n^{n-1}$, which is exactly the term of the determinant coming from expanding the leading diagonal. Hence the constant factor is 1, and the two are equal.

Clearly the discriminant is zero if, and only if, the polynomial has a repeated root. $\operatorname{res}(p, p')$ has the same property, which might lead one to conjecture the following result.

Proposition 5. $a_n \operatorname{disc}(p) = \operatorname{res}(p, p')$.

Proof.

$$\begin{aligned}
\operatorname{res}(p,p') = a_n^{n-1}\prod_{i=1}^{n} p'(\alpha_i) &= a_n^{n-1}\prod_{i=1}^{n}\left(a_n\prod_{j=1}^{n}(x-\alpha_j)\right)'(\alpha_i) \\
&= a_n^{2n-1}\prod_{i=1}^{n}\left(\sum_{k=1}^{n}\prod_{\substack{j=1\\ j\ne k}}^{n}(x-\alpha_j)\right)(\alpha_i) \\
&= a_n^{2n-1}\prod_{i=1}^{n}\left(\prod_{\substack{j=1\\ j\ne i}}^{n}(\alpha_i-\alpha_j)\right) \\
&= a_n\operatorname{disc}(p).
\end{aligned}$$

Now the first column of Sylvester's matrix for $\operatorname{res}(p, p')$ contains a_n and na_n only, and so the determinant is divisible by a_n. A particular consequence of this result is that the discriminant is a polynomial function of the coefficients of p, and is an integer if the coefficients of p are integers.

Proposition 6. $\operatorname{disc}(pq) = \operatorname{disc}(p)\operatorname{disc}(q)\operatorname{res}(p,q)^2$.

Proof. Every root of pq is a root of p or of q. Those terms in the definition of $\operatorname{disc}(pq)$ which come from two roots of p are accoutned for by the $\operatorname{disc}(p)$ term, and similarly for q, while the $\operatorname{res}(p,q)^2$ term accounts for the hybrid pairs. Note that the resultant needs to be squared, since every pair of roots occurs twice in a discriminant.

This proposition explains the often-observed fact that discriminants tend to factor, and to have non-trivial square-free decompositions. Note that the existence of a squared factor is independent of the field over which pq factors as p and q, so the discriminant of an apprently irreducible (but not absolutely irreducible) polynomial will tend to have squared factors.

Proposition 7. $\operatorname{res}(pq,r) = \operatorname{res}(p,r)\operatorname{res}(q,r)$.

Proof. Similar to the previous proposition, by using proposition 3 to express the resultant as a product of differences of roots.

4. Algebraic Numbers.

Definition. *A number is said to be an* **algebraic number** *if it is a root of a polynomial with integer coefficients.*

Familiar algebraic numbers are all the integers, since n is the root of $x - n = 0$. Similarly all the rational numbers are algebraic numbers, since p/q is the root of $px - q = 0$. All radicals of rational numbers are algebraic numbers, since $\sqrt[n]{p/q}$ is a root of $px^n - q = 0$. The case of more general radicals will be discussed later, after Corollary 5.

We could equally well have defined algebraic numbers as roots of polynomials with rational number coefficients, since clearing denominators will reduce such a polynomial to one with integer coefficients. The next definition, though, can not readily be so phrased.

Definition. *A number is said to be an* **algebraic integer** *if it is a root of a monic polynomial with integer coefficients, i.e. one whose leading coefficient is unity.*

For example, $\frac{1}{2}\left(1+\sqrt{5}\right)$ is an algebraic integer, since it is a root of $x^2 - x - 1$.

Proposition 1. *Every algebraic number can be expressed as an algebraic integer divided by an ordinary integer.*

Proof. Let α be an algebraic number, and suppose that it is a root of the polynomial $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$. If we multiply this polynomial by ${a_n}^{n-1}$, we obtain

$$(a_n x)^n + a_{n-1}(a_n x)^{n-1} + \cdots + a_1 {a_n}^{n-2}(a_n x) + a_0 {a_n}^{n-1}.$$

This polynomial has the same roots as the original one. But it can be re-written as $y^n + a_{n-1}y^{n-1} + \cdots + a_1 {a_n}^{n-2} y + a_0 {a_n}^{n-1}$, and this polynomial is monic, with integer coefficients, so that its roots are algebraic integers. But they are merely a_n times the roots of the original polynomial, so that the roots of the original polynomial, if multiplied by a_n, become algebraic integers.

Most computer algebra systems prefer to deal with algebraic integers, rather than algebraic numbers. This is because such systems are fundamentally polynomial-based, and the product of two polynomials with integer coefficients and with algebraic integers amongst the "variables" is still a polynomial with integer coefficients. This is not true for general algebraic numbers, as can be seen by considering α as a root of $2x^2 - 1$, when α^2 simplifies to $\frac{1}{2}$.

It is obvious that the negative of an algebraic number is an algebraic number, since we need merely substitute $-x$ for x in the defining polynomial. Similarly, the reciprocal of a non-zero algebraic number is an algebraic number. Sums and products are dealt with by the following propositions.

Proposition 2 [Loos, 1982b]**.** *If α and β are algebraic numbers defined as roots of the polynomials p and q, then $\alpha + \beta$ is an algebraic number defined by the polynomial $r(x) = \mathrm{res}_y(p(x-y), q(y))$.*

Proof. Let the α_i be all the roots of p, so that $p(z) = a_m \prod_{i=1}^{m}(z - \alpha_i)$, and the β_j be all the roots of q. Then

$$\begin{aligned}\mathrm{res}_y(p(x-y), q(y)) &= (-1)^{mn} b_n^m \prod_{j=1}^{n} p(x - \beta_j) \\ &= (-1)^{mn} a_m^n b_n^m \prod_{j=1}^{n}\prod_{i=1}^{m}(x - \beta_j - \alpha_i)\end{aligned}.$$

Hence the roots of this polynomial are all the $\alpha_i + \beta_j$, in particular $\alpha + \beta$.

Corollary 1. *If α and β are algebraic integers, so is $\alpha + \beta$.*

For example, $\sqrt{2} + \sqrt{3}$ is an algebraic integer, satisfying the polynomial $\mathrm{res}_y((x-y)^2 - 2, y^2 - 3) =$

$$\begin{vmatrix} 1 & -2x & x^2-2 & 0 \\ 0 & 1 & -2x & x^2-2 \\ 1 & 0 & -3 & 0 \\ 0 & 1 & 0 & -3 \end{vmatrix} = x^4 - 10x^2 + 1,$$

from which we deduce that $\sqrt{2} + \sqrt{3} = \sqrt{5 + 2\sqrt{6}}$.

Proposition 3 [Loos, 1982b]**.** *If α and β are algebraic numbers defined as roots of the polynomials p and q, then $\alpha\beta$ is an algebraic number defined by the polynomial $r(x) = \mathrm{res}_y(y^m p(x/y), q(y))$.*

Corollary 2. *If α and β are algebraic integers, so is $\alpha\beta$.*

Of course, there is no guarantee that this process will produce the most simple result for a given input. For example, $\sqrt{2}\sqrt{3}$ is an algebraic integer defined by $\mathrm{res}_y(y^2((x/y)^2 - 2), y^2 - 3)$, which evaluates to $x^4 + 12x^2 + 36$, but this can equally well be written as $(x^2 + 6)^2$, and this can be discovered by the technique of square-free decomposition described above. Similarly, applying the previous proposition to $\sqrt{2}^2$ gives the polynomial $x^4 + 8x^2 + 16$, which reduces to $(x-2)^2(x+2)^2$.

There is an alternative proceure, which is perhaps simpler to apply by hand, and which is guaranteed to give the minimal polynomial that can be deduced from the information given. Given some number x that is a combination (sums and products) of several algebraic numbers α_i, this process forms x^2, x^3, ...

Each of these will be a combination of the α_i, and can be regarded as lying in the $\mathbf{Q}$-vector space whose basis is $\{1, \alpha_1, \ldots, \alpha_n, \alpha_1^2, \alpha_1\alpha_2, \ldots, \alpha_n^2, \ldots\}$, where the basis is finite since each α_i appears only to powers less than the degree of the equation which defines it. After computing each x^i, we check to see if there is a non-trivial linear combination of $1, x, x^2, \ldots, x^i$ with rational coefficients which is zero. If so, this determines a polynomial equation for x, which has to be minimal, since otherwise there would be an equation of lower degree, which would have been found sooner.

Applying this to $x = \sqrt{2} + \sqrt{3}$, we see that $x^2 = 5 + 2\sqrt{2}\sqrt{3}$, and there is no such combination. $x^3 = 11\sqrt{2} + 9\sqrt{3}$, and again there is no combination. $x^4 = 49 + 20\sqrt{2}\sqrt{3}$, and now there is the combination $x^4 - 10x^2 + 1$. If we apply this technique to $x = \sqrt{2} + \sqrt{2}$, we must first make the distinction as to whether we know that these are the same roots or not. If we do, then $x^2 = 8$, and this gives us our equation. If we do not, then write x as $\sqrt{2} + \widehat{\sqrt{2}}$ to make the difference obvious. Then $x^2 = 4 + 2\sqrt{2}\widehat{\sqrt{2}}$, and now there is no equation. But $x^3 = 8\sqrt{2} + 8\widehat{\sqrt{2}}$, and we have the equation $x^3 = 8x$, which is the obvious equation multiplied by x, a factor corresponding to the choice of opposite signs for the two square roots, so that $x = 0$.

Corollary 3. *The algebraic integers form a ring.*

Corollary 4. *The algebraic numbers form a field, known as* $\mathbf{A}$.

Proposition 4. *If $p(x)$ is a polynomial with algebraic number coefficients, then the roots of p are algebraic numbers.*

Proof. By induction on k, the number of *distinct** algebraic numbers occurring as, or amongst, the coefficients of p. Clearly the proposition is true when $k = 0$. In the general case, let α be an algebraic number occurring as, or in the expression for, a coefficient of p, so that the coefficients of p are expressed in terms of rational numbers, α and $k-1$ other algebraic numbers (say $\beta_2, \ldots, \beta_k$). Let $q(y)$ be the polynomial defining α as an algebraic number, and suppose that this is of degree n, with roots $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$. Consider $r(x) = \mathrm{res}_y(p(x)_{\alpha:=y}, q(y))$, where by $p(x)_{\alpha:=y}$ we mean the result of replacing every occurrence of α in p by y. This is a polynomial in x whose coefficients involve rational numbers and $\beta_2, \ldots, \beta_k$ only. Furthermore, by Corollary 2 in the previous section, it is

$$(-1)^{mn} b_n^m \prod_{j=1}^{n} p(x)_{\alpha:=\alpha_j}.$$

In particular, $p(x)$ is the factor in the product with $j = 1$, so this polynomial has all the roots that p has.

Corollary 5. *The roots of a monic polynomial with algebraic integer coefficients are algebraic integers.*

This helps to settle the question mentioned above, as to whether more complex radicals were algebraic integers. Of course, some expressions that might seem to involve division are in fact algebraic integers, such as $\sqrt{\frac{1}{2}(1+\sqrt{5})}$, but we know now that expressions involving integers, addition, subtraction and multiplication and the extraction of radicals are always algebraic integers.

We now know how to reduce a complex expression, such as "$\alpha + \beta$ where α satisfies $\alpha^3 + (1-\sqrt{5})\alpha^2 = \sqrt[3]{7}$ and β satisfies $\sqrt{1+\sqrt{2}}\beta^2 + \beta = 1$" into a polynomial with integer coefficients, which may be assumed to have no repeated roots.

* More formally, we are inducting on the number of levels in the algebraic extension $\mathbf{Q}(\beta_2)\ldots(\beta_k)(\alpha)$ defining the coefficients of p.

5. On the Roots of Polynomials.

In the next section, we will need to know various facts about the roots of polynomials: how big they can be, how far apart they can be, etc. We will therefore spend this section in collecting various facts about the roots of polynomials for later use. Throughout this section, we will be interested in the polynomial $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$, where the a_i are arbitrary numbers (though we will generally be interested in the case when the a_i are integers), and a_n is non-zero (we may as well assume that a_n is positive). n is therefore the degree of p, which we will also write $\delta(p)$. For any integer k, let

$$\|p\|_k = \sqrt[k]{\sum_{i=0}^{n} |a_i|^k}.$$

This is a decreasing function of k, and we will abuse notation so as to write

$$\|p\|_\infty = \max_{0 \le i \le n} |a_i|.$$

These are related by the inequalities

$$(n+1)\|p\|_\infty \ge \|p\|_1 \ge \|p\|_2 \ge \ldots \ge \|p\|_\infty,$$

where the first inequality is strict unless all the coefficients are equal in absolute value, and the subsequent inequalities are strict unless there is only one non-zero coefficient.

The two quantities we will primarily be concerned with are the maximum root of a polynomial, and the minimum distance between them. So let the roots of p be $\alpha_1, \ldots, \alpha_n$, and define

$$\mathrm{rb}(p) = \max_{1 \le i \le n} |\alpha_i|,$$

$$\mathrm{sep}(p) = \min_{1 \le i < j \le n} |\alpha_i - \alpha_j|.$$

$\mathrm{sep}(p)$ is zero if, and only if, p has a repeated factor.

Proposition 1 [Cauchy, 1829, p. 122]**.** $\mathrm{rb}(p) \le 1 + \|p\|_\infty / a_n$.

Proof. Write H for $\max_{0 \le i \le n-1} |a_i|$. Let α be a root of p, and suppose that $|\alpha| > 1$, else the Proposition is trivially true. Then $a_n\alpha^n = -a_{n-1}\alpha^{n-1} - \cdots - a_0$, and so

$$\begin{aligned} a_n|\alpha|^n &\le \left|a_{n-1}\alpha^{n-1} + a_{n-2}\alpha^{n-2} + \cdots + a_0\right| \\ &\le |a_{n-1}||\alpha|^{n-1} + |a_{n-2}||\alpha|^{n-2} + \cdots + |a_0| \\ &\le H\left(|\alpha|^{n-1} + |\alpha|^{n-2} + \cdots + 1\right) \\ &= \frac{H|\alpha|^n}{|\alpha| - 1}. \end{aligned}$$

Hence $(|\alpha| - 1)a_n \le H$, and the result follows.

Corollary 1. *If the polynomial p does not take the value 0 at $x = 0$, then every root of p has absolute value at least $|a_0|/(|a_0| + \|p\|_\infty)$.*

Proof. Apply the proposition to the polynomial $a_0x^n + \cdots + a_n$, which has the same $\|\|_\infty$ and whose roots are the reciprocals of the roots of p.

Proposition 2 [Cauchy, 1829, p. 122]**.** $\mathrm{rb}(p) \le \max\left(\frac{n|a_{n-1}|}{a_n}, \sqrt{\frac{n|a_{n-2}|}{a_n}}, \ldots, \sqrt[n-1]{\frac{n|a_1|}{a_n}}, \sqrt[n]{\frac{n|a_0|}{a_n}}\right)$.

Proof. As above, $a_n|\alpha|^n \le |a_{n-1}||\alpha|^{n-1} + \cdots + |a_0|$. Let k be the index for which the right-hand summand is maximal, i.e. $|a_k||\alpha|^k \ge |a_i||\alpha|^i$ for all i. Then $a_n|\alpha|^n \le k|a_k||\alpha|^k$, which means that $|\alpha| \le \sqrt[n-k]{n|a_k|/a_0}$.

A very similar inequality was stated by Knuth [1969, 4.6.2 exercise 20].

Proposition 3. $\mathrm{rb}(p) \leq 2 \max\left(\frac{|a_{n-1}|}{a_n}, \sqrt{\frac{|a_{n-2}|}{a_n}}, \ldots, \sqrt[n-1]{\frac{|a_1|}{a_n}}, \sqrt[n]{\frac{|a_0|}{a_n}}\right)$.

Proof. Write the right-hand side of the inequality to be proved as $2B$. Then, as above,

$$a_n|\alpha|^n \leq |a_{n-1}||\alpha|^{n-1} + \cdots + |a_0|.$$

Hence

$$\begin{aligned} 1 &\leq \frac{|a_{n-1}|}{a_n|\alpha|} + \frac{|a_{n-2}|}{a_n|\alpha|^2} + \cdots + \frac{|a_0|}{a_n|\alpha|^n} \\ &\leq \frac{B}{|\alpha|} + \frac{B^2}{|\alpha|^2} + \cdots + \frac{B^n}{|\alpha|^n} \\ &\leq \frac{B/|\alpha|}{1 - B/|\alpha|}. \end{aligned}$$

This reduces to $|\alpha| \leq 2B$, as required.

Many other inequalities on similar lines have been proved over the years: see, e.g., Knuth [*loc. cit.* exercise 19]. We note that the previous two results are close to being optimal, since, for example, we know that there is one root of p larger (in absolute value) than $\sqrt[n]{|a_0|}/a_n$, and similarly that there is one root larger than $|a_{n-1}|/na_n$. This can be made more formal, as the following result shows.

Proposition 4. *There is always a root of a polynomial whose absolute value is at least $B/2n$, where B is the bound of proposition 3.*

Proof. Without loss of generality, we can assume that our polynomial is monic. B then reduces to $2 \max\left(|a_{n-1}|, \sqrt{|a_{n-2}|}, \ldots, \sqrt[n-1]{|a_1|}, \sqrt[n]{|a_0|}\right)$. Let this maximum be attained at the k-th element, so that $B = 2\sqrt[k]{|a_{n-k}|}$. Now a_{n-k} is a sum of $\binom{n}{k}$ products of roots, viz.

$$a_{n-k} = (-1)^k \sum_{\substack{1 \leq i_j \leq n \\ i_j \text{distinct}}} \alpha_{i_1} \ldots \alpha_{i_k}.$$

Hence there is some choice of indices such that $|\alpha_{i_1} \ldots \alpha_{i_k}| \geq |a_{n-k}|/\binom{n}{k}$. Therefore the largest, in absolute value, of these α_{i_j} is greater than

$$\sqrt[k]{|a_{n-k}|/\binom{n}{k}} = B/2\sqrt[k]{\binom{n}{k}} \leq B/2n.$$

We will need one more result, which we do not prove here, but refer the reader to Mignotte [1982] for a proof.

Proposition 5 [Landau, 1905]**.** *The product of all the roots of p greater than 1 in absolute value is bounded by $\|p\|_2/|a_n|$.*

We now pass to consideration of the minimum separation between roots.

Proposition 6 [Mahler, 1964]**.** $\mathrm{sep}(p) > \sqrt{3|\operatorname{disc}(p)|}n^{-(n+2)/2}\|p\|_2^{1-n}$.

Corollary 2. *If p is square-free and has integer coefficients, then its discriminant is a non-zero integer, so that* $\mathrm{sep}(p) > \sqrt{3}n^{-(n+2)/2}\|p\|_2^{1-n}$.

As noted in section 3, these bounds can be very large. For example, the discriminant of Wilkinson's polynomial $W(x) = (x-1)\ldots(x-20)$ is about $2.74 * 10^{276}$. Despite this, proposition 6 gives about 10^{-244} as a bound for the root separation, even though the true value is, of course, 1. The corollary gives about 10^{-382}. What is worse, the corollary is dependent on scale, so that, for the perturbed Wilkinson's polynomial $W(x) + 2^{-23}x^{19}$, the proposition gives about the same bound as before, while the corollary is not applicable

to this polynomial, but must be applied to $2^{23}W(x) + x^{19}$, when it gives a bound of about 10^{-443}. The following bound, due to Collins & Horowitz [1974], is, however, even worse than the corollary, though it has the same asymptotic order for $\log \operatorname{sep}(p)$, viz. $O(n \log(n\|p\|_2))$, and this is generally the quantity of interest. We will refer to $\log \operatorname{sep}(p)$ as the **accuracy** required to evaluate the roots of a polynomial. It is a measure of the number of digits in isolating intervals, and the number of bisections required to make an interval smaller than $\operatorname{sep}(p)$.

Proposition 7. *If p is square-free and has integer coefficients, then* $\operatorname{sep}(p) > \frac{1}{2}e^{-n/2}n^{-3n/2}\|p\|_2^{-n}$.

Fortunately, sep is rarely of interest in actual algorithms, as opposed to bounds on their complexity. One technique which can often improve the above estimates for sep is to transform the polynomial. The transformation $x \longrightarrow x - a_{n-1}/na_n$ is known as a *Tschirnhausen transformation*, and it annuls the coefficient of x^{n-1}, which is also the negative of the sum of the roots. It does not affect the difference between any pair of roots, and hence the separation and the discriminant are unchanged. This transformation may not be integral, but the corollary above is still applicable, since the discriminant is integral. Applying this transformation to Wilkinson's polynomial, and then applying the corollary, yields a bound of 10^{-240}, which is a distinct improvement, though still far from the truth. If we are interested in separating *all* the real roots of a polynomial, the following result is more useful.

Proposition 8 [Davenport, 1985]**.** *Let $\alpha_1, \ldots \alpha_{k+1}$ be the $k+1$ real roots of p in descending order, with $k \geq 1$. Then*

$$\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}| \geq 3^{k/2}\sqrt{|\operatorname{disc}(p)|}\|p\|_2^{-n+1}n^{-k-n/2}.$$

Proof. Without loss of generality, we can assume that p is monic, since multiplying by a constant will change $\operatorname{disc}(p)$ as much as $\|p\|_2^{2n-2}$. We need only consider the case of p square-free, since otherwise the right-hand side of the inequality to be proved is zero. Let us order the roots of p, which we shall call β_i, in decreasing order of modulus, so that

$$|\beta_1| \geq |\beta_2| \geq \cdots |\beta_M| > 1 \geq |\beta_{M+1}| \geq \cdots |\beta_n|,$$

where M can have any value from 0 to n.

By proposition 3 of section 3, we can write $\sqrt{\operatorname{disc}(p)}$ (up to sign) as

$$P = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ \beta_1 & \beta_2 & \ldots & \beta_n \\ \beta_1^2 & \beta_2^2 & \ldots & \beta_n^2 \\ \vdots & \vdots & & \vdots \\ \beta_1^{n-1} & \beta_2^{n-1} & \ldots & \beta_n^{n-1} \end{vmatrix}.$$

Perform the following column operations on P, which do not change its value.

1) Take a α_i of greatest absolute value, which can be chosen, by the condition imposed on the ordering of the α_i, to be one of α_1 and α_{k+1}. Subtract from its column the column corresponding to the α_j which occurs subtracted from α_i (or vice versa) in $\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|$. Delete this α_i from the list.
2) Take a remaining α_i of greatest absolute value, which can be chosen, by the condition imposed on the ordering of the α_i, to be one of the α_i at the end of the chain: one of α_1 and α_k in the case that step (1) deleted α_{k+1}, or one of α_2 and α_{k+1} if α_1 was deleted. Subtract from its column the column corresponding to the α_j which occurs subtracted from α_i (or vice versa) in $\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|$. Delete this α_i from the list.
3) ....

Eventually,we will have performed k column operations, and have accounted for all the subtractions in $\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|$. The new determinant for P has columns that look like $(1, \beta_i, \beta_i^2, \ldots, \beta_i^{n-1})$ if this column was not affected by the column operations, or $(0, \beta_i - \beta_j, \beta_i^2 - \beta_j^2, \ldots, \beta_i^{n-1} - \beta_j^{n-1})$ if this column had the j-th column subtracted from it. Note that the ordering of the column operations means that $i < j$. Such a column has a common factor of $\beta_i - \beta_j$, which by construction is one of the factors of $\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|$. Write this column as

$$(\beta_i - \beta_j)(0, 1, \gamma_i^{(2)}, \ldots, \gamma_i^{(n-1)}),$$

where

$$\gamma_i^{(l)} = \frac{\beta_i^l - \beta_j^l}{\beta_i - \beta_j} = \beta_i^{l-1} + \beta_i^{l-2}\beta_j + \cdots + \beta_j^{l-1}.$$

So we have shown that

$$\frac{\sqrt{|\operatorname{disc}(p)|}}{\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|} = \begin{vmatrix} 1 & 0 & \dots & 1 \\ \beta_1 & 1 & \dots & \beta_n \\ \beta_1^2 & \gamma_2^{(2)} & \dots & \beta_n^2 \\ \vdots & \vdots & & \vdots \\ \beta_1^{n-1} & \gamma_2^{(n-1)} & \dots & \beta_n^{n-1} \end{vmatrix}, \tag{1}$$

where the precise format of the determinant on the right depends on which β_j are α_i. All that remains is to estimate the determinant on the right-hand side. This will be done via a column-oriented version of Hadamard's inequality [1893; Mignotte, 1982], viz. that

$$\left| \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} \right| \le \prod_{i=1}^{n} \sqrt{\sum_{j=1}^{n} |a_{ji}|^2}.$$

There are four kinds of column that can occur in (1), and each will contribute differently to the product in Hadamard's inequality, and hence to our final bound.

Type 1) The column is $(1, \beta_i, \dots \beta_i^{n-1})$ and $i \le M$. Then this column contains n numbers, each of absolute value at most $|\beta_i|^{n-1}$, and hence the contribution to the product is less than $\sqrt{n}|\beta_i|^{n-1}$.

Type 2) The column is $(0, 1, \gamma_i^{(2)}, \dots, \gamma_i^{(n-1)})$ and $i \le M$. The absolute value of the l-th element of this column is at most $(l-1)|\beta_i|^{l-1}$, since $\gamma_i^{(l)}$ is the sum of $l-1$ items, each a power product of β_i and β_j, where $|\beta_j| \le |\beta_i|$, so the total contribution of the column is bounded by

$$\sqrt{\sum_{l=1}^{n} \left((l-1)|\beta_i|^{l-1}\right)^2} \le |\beta_i|^{n-1} \sqrt{\sum_{l=1}^{n} (l-1)^2} < |\beta_i|^{n-1} \sqrt{\frac{n^3}{3}}.$$

Type 3) The column is $(1, \beta_i, \dots \beta_i^{n-1})$ and $i > M$. Then this column contains n numbers, each at most 1 in absolute value, and hence the contribution to the product is less than or equal to $\sqrt{n}$.

Type 4) The column is $(0, 1, \gamma_i^{(2)}, \dots, \gamma_i^{(n-1)})$ and $i > M$. The absolute value of the l-th element of this column is at most $(l-1)$, so the total contribution of the column is bounded by

$$\sqrt{\sum_{l=1}^{n} (l-1)^2} < \sqrt{\frac{n^3}{3}}.$$

Hence every column contributes a $\sqrt{n}$. Those of types 1 and 2 contribute an additional $|\beta_i|^{n-1}$, and there are a total of M of these columns. Columns of types 2 and 4 contribute an additional $n/\sqrt{3}$, and there are k such columns. Multiplying these contributions together, we get

$$n^{n/2} \left(\frac{n}{\sqrt{3}}\right)^k \prod_{i=1}^{M} |\beta_i|^{n-1}.$$

By proposition 5 (Landau's inequality), $\prod_{i=1}^{M} |\beta_i| \le \|p\|_2$. This is the only place $\|p\|_2$ enters into the proof: an observation that will be use in the next section. Combining these results with (1) shows that

$$\frac{\sqrt{|\operatorname{disc}(p)|}}{\prod_{i=1}^{k} |\alpha_i - \alpha_{i+1}|} \le n^{k+n/2} 3^{-k/2} \|p\|_2^{n-1}.$$

Re-arranging this gives the inequality stated, and completes the proof.

Mahler's inequality corresponds to the case $k = 1$ of this result. We can view this result as saying that there is only a certain amount of "closeness", which can either be concentrated on one pair of roots, or spread between several pairs.

6. Approximations to Roots.

Section 4 showed how to reduce any complex expression to a simple polynomial with integer coefficients which has all the roots of the original expression (it may, of course, have other roots as well). However, it is not always sufficient to know an equation which a number satisfies. When we write $\sqrt{2}$, we mean more that "a number whose square is 2": we mean "that positive number whose square is 2", and in more general circumstances, where we have high degree polynomials, we will wish to know, to arbitrarily high accuracy, which root we are talking about. We will confine ourselves here to *real* roots of polynomials with integer (or rational) coefficients: complex zeros are treated by Collins [1977].

We can assume that the polynomial is square-free, as outlined in section 2. This means that every zero is simple, and that the polynomial changes sign in the neighbourhood of a zero.

Definition. *An* **isolating interval** *for a polynomial is an interval of two rational numbers, between which there is precisely one zero of the polynomial. We can assume that* 0 *does not belong to the isolating interval.*

We need not worry about cases where we find an exact zero of the polynomial by evaluating at a rational: in this case we know the zero exactly, and no longer have an algebraic number to deal with. Once we have an isolating interval (a, b), we can make it as precise as we please, merely by investigating the sign of the polynomial at $\frac{1}{2}(a+b)$, and letting this value replace whichever of a and b at which the polynomial has the same sign.

The condition about 0 is convenient, since the reciprocal of a number near zero can be of either sign. If we exclude this case, then the sum, difference, product and quotient of two intervals are also intervals, and we shall make heavy use of this fact later.

This poses the problem: given a polynomial, can we find isolating intervals for all its real roots? One obvious method, which Collins & Loos [1982] attribute to Kronecker, is to take that portion of the real line between $-\mathrm{rb}(p)$ and $\mathrm{rb}(p)$ and divide it into intervals of size $\mathrm{sep}(p)$, and look for a change of sign in p between the limits of each segment. Unfortunately, the root separation is exponential in n, so this algorithm requires an exponential amount of time in general.

To do better, we must have some method of working out where the real zeros are, e.g. a technique for knowing how many there are in any interval (a, b) (presumably, if there are zeros exactly at a or b, we will detect the fact in any case). There are several such methods: probably the best known is that of *Sturm sequences.*

Definition. *The* **Sturm sequence** *for a polynomial $p(x)$ with real coefficients is a sequence of polynomials $\{f_i\}$, defined by $f_0 = p$, $f_1 = p'$ and, in general, $f_i = -\mathrm{remainder}(f_{i-2}, f_{i-1})$.*

The f_i have decreasing degrees, so the sequence is finite, and, if p is square-free, the sequence will terminate with a constant, since, up to sign, we are merely performing the Euclidean algorithm on p and p'. We will write f_N for this last element of the sequence.

Definition. *The* **sign variation** *for a Sturm sequence at a point a, $V(a)$, is defined as the number of times the sequence of non-zero elements of $f_0(a), f_1(a), \ldots, f_N(a)$ changes sign.*

We note that the sign variation of a Sturm sequence is unchanged if we multiply any of the elements by a positive constant. Hence we will often find it convenient to ignore denominators that arise in the calculation of the remainders when we divide by non-monic polynomials.

The use of Sturm sequences comes from the following result, which we state and prove only in the limited case of interest to us, though it is true in much greater generality.

Proposition 1. *If p is non-zero at a and b, the number of zeros of p in the interval (a, b) is $V(a) - V(b)$.*

Proof. Since a polynomial can only change sign by going through a zero, $V(x)$ is invariant except at the zeros of any of the f_i. Furthermore, f_N has no zeros, and so we need only consider two cases:

(a) f_i has a zero at x_0, where $0 < i < N$. In this case, we wish to prove that V is unchanged as we pass this zero. It is impossible for two consecutive f_i to be zero simultaneously, for this would imply that they had a common divisor, and hence that all subsequent f_i had a common divisor. Hence, even if two f_i are zero at x_0, their effects on V will be independent. So consider just one of them. $f_{i+1} = -\mathrm{remainder}(f_{i-1}, f_i)$, so that $f_{i-1} = q_i f_i - f_{i+1}$ for some quotient polynomial q_i. Hence $f_{i-1}(x_0) = -f_{i+1}(x_0)$, and these two

have opposite signs at, and therefore on both sides of, x_0. So, even though f_i changes sign, the sign variation from the triple f_{i-1}, f_i, f_{i+1} is 1 on both sides of x_0.

(b) f_0 has a zero at x_0. By the argument of case (a), we can ignore any other f_i which are also zero at this point. We wish to prove that V decreases by 1 as we pass this zero of f_0. Since p is square-free, we know that f_1 is non-zero at x_0. If it is negative, then $p(x_0 - \epsilon)$ is positive and $p(x_0 - \epsilon)$ is negative. Hence there was a sign variation between f_0 and f_1 at $x_0 - \epsilon$, but not at $x_0 + \epsilon$, so V has decreased by 1. If $f_1(x_0)$ is positive, then f_0 changes sign from negative to positive at x_0, and again V decreases by 1.

This would seem to be an excellent solution to our problem, since we can first evaluate the Sturm sequence at ∞ and $-\infty$ (which consists merely in examining the sign of the leading coefficients) to determine how many real zeros there are, and we can then perform binary division on the interval $(-b, b)$, where b is a bound for the roots of the polynomial, until they are isolated.

Proposition 2. *We can isolate the roots of a square-free polynomial p with integer coefficients in at most $O(n \log(n\|p\|_2))$ evaluations of the Sturm sequence.*

Proof. Let the α_i be the real roots of p, in increasing order. We know that these are contained in $(-b, b)$, where we will take b from proposition 1 of the previous section, viz. $1 + \|p\|_\infty / a_n$. If there are $k+1$ real roots, we will need to construct k separation points. The denominator of a point separating α_i from α_{i+1} will be at most $1/2|\alpha_i - \alpha_{i+1}|$. This will require $\lceil \log_2(2b/|\alpha_i - \alpha_{i+1}|) \rceil$ operations, hence the total number of sub-divisions is at most

$$k + k \log_2 2b + \sum_{i=1}^{k} - \log_2 |\alpha_i - \alpha_{i+1}|$$

(where the opening $k+$ bounds the amount of rounding that took place), which is bounded by

$$k + k \log_2 2b + (n-1) \log_2 \|p\|_2 + \left(k + \frac{n}{2}\right) \log_2 n - \frac{1}{2} \log_2 |\operatorname{disc}(p)| - \frac{k}{2} \log_2 3$$

by proposition 8 of the previous section. $\operatorname{disc}(p)$ is a non-zero integer, and b is bounded by $1 + \|p\|_\infty$, which in turn is bounded by $1 + \|p\|_2$, and hence by $2\|p\|_2$. Our bound then simplifies to

$$k \left(1 + \log_2(4/\sqrt{3})\right) + (k+n) \log_2 \|p\|_2 + \left(k + \frac{n}{2}\right) \log_2 n.$$

Since $k \leq n$, this bound can be written $O(n \log(n\|p\|_2))$. We still have to ensure that our isolating intervals do not include 0, but this can be done with one evaluation at 0 and a bound on the distance of roots of a polynomial from 0, wich is provided by corollary 1 of the previous section.

We can assume that the Sturm sequence is computed only once, which computation will take $O(n^2)$ *arithmetic* operations if done in the obvious way. Each computation of V requires evaluating polynomials of degree n, $n-1$, ..., 0, i.e. $O(n^2)$ *arithmetic* operations. Hence the cost of finding the isolating intervals is at most $O(n^3 log(n\|p\|_2))$ *arithmetic* operations.

In order to understand the practical draw-backs of this scheme, let us look at an example, that of Wilkinson's polynomial, $W(x) = (x+1)\ldots(x+20)$. The leading coefficients of its Sturm sequence are all positive, which means that $V(\infty) = 0$ and $V(-\infty) = 20$, so the polynomial has 20 real zeros. The largest coefficients in $W(x)$ and $W'(x)$ are both the x^2 coefficients, $1.38 * 10^{19}$ and $3.86 * 10^{19}$ respectively. The largest coefficient encountered in the Sturm sequence (assuming that we throw away denominators, and divide polynomials by their content) is the x^2 coefficient of f_4, which is about $7.78 * 10^{20}$. The situation is very different for the perturbed polynomial, which we can write as $\hat{W} = 2^{23} W + x^{19}$. Here the largest coefficients in $\hat{W}$ and $\hat{W}'$ are $1.08 * 10^{26}$ and $3.24 * 10^{26}$ respectively. The leading coefficients, instead of all being positive, have signs $+,+,+,+,+,+,+,+,-,-,+,+,+,-,-,+,+,+,-,-,-$, so that $V(\infty) = 5$ and $V(-\infty) = 15$, thus proving that, as we asserted in section I.2, $\hat{W}$ has ten real roots. However, the largest coefficient encountered in the Sturm sequence is now the constant term of f_{19}, which is $3.14 * 10^{320}$.

Conventional subresultant theory [Loos, 1982a] tells us that the coefficients of the i-th polynomial in this sequence may have coefficients as large as $n^i \|p\|_2^{2i-1}$ (which would give 10^{987} in the case described at the end of the previous paragraph, so we see that these estimates are not outrageous). Since we may need to

evaluate this polynomials at rational numbers whose denominators (and therefore presumably numerators) are of the order of $\operatorname{sep}(p)$, this may take as many as $O(n^4 \log^2(n\|p\|_2))$ operations using Horner's rule and classical operations. This contrasts unfavourably with the figure of n arithmetic operations that was assumed above. Combining the present figure with the analysis for the number of polynomial evaluations given above, we get a total time of $O(n^6 \log^3(n\|p\|_2))$. "Fast" operations would reduce this to $O(n^5 \log^3(n\|p\|_2)\epsilon)$, where ϵ stands for a variety of log log terms. The computation of the Sturm sequence, even by "classical" methods, will take only $O(n^4 \log^2(n\|p\|_2))$ operations, so this is not the limiting factor.

In this context, Schwartz & Sharir [1983] suggested an interesting variant. The Sturm sequence is defined via relations of the form $f_i = -\text{remainder}(f_{i-2}, f_{i-1})$, which can be re-written as $f_{i-2} = q_i f_{i-1} - f_i$. If we stored the q_i, then this sequence of rules would generate the values of the entire Sturm sequence from the values of f_N and the q_i, since $f_{N-1} = q_N f_N$. The total degree of these polynomials is n, the degree of p, so that we have an evaluation in $O(n)$ *arithmetic operations*, which one might presume to be faster when bit operations are counted.

There are several snags in converting Schwartz & Sharir's idea to a practical algorithm. The q_i computed in the sub-resultant scheme are not the q_i needed for recurrence-relation evaluation of $f_i(x)$, since various numeric factors have been introduced along the way: either multiplied in to avoid fractions, or divided out to reduce the size of the numbers. Davenport [1985] has analysed the process for various possible values of d, the number of digits in x (if x is a fraction we mean the greater of the number in the numerator and denominator). There are three obvious orders of magnitude for d: $O(1)$, which corresponds to the case of a small number of widely-separated roots; $O(\log(n\|p\|_2))$, which corresponds to a large number of roots, but "averagely" spaced; and $O(\log \operatorname{sep} p) = O(n \log(n\|p\|_2))$, which corresponds to the worst case of roots as close together as possible. The results are shown in table 1, for the naïve method of evaluating a polynomial as $\sum a_i x^i$ (which is not as stupid as it seems in our case, since we have up to n polynomials to evaluate at the same x), for Horner's rule $a_0 + x(a_1 + x(\ldots))$ and for a modified version of the recurrence-relation method.

Table 1:	Classical $O(k^2)$ multiplication		
d	Naïve Method	Horner's Method	Recurrence Method
1	$n^4 \log(n\|p\|_2)$	$n^3 \log(n\|p\|_2)$	$n^3 \log^2(n\|p\|_2)$
$\log(n\|p\|_2)$	$n^4 \log^2(n\|p\|_2)$	$n^3 \log^2(n\|p\|_2)$	$n^3 \log^2(n\|p\|_2)$
$n \log(n\|p\|_2)$	$n^5 \log^2(n\|p\|_2)$	$n^5 \log^2(n\|p\|_2)$	$n^4 \log^2(n\|p\|_2)$

On the last line, the constants involved in the two $O(n^5 \log^2(n\|p\|_2))$ methods are about the same. Conversely, on the previous line there are two $O(n^3 \log^2(n\|p\|_2))$ methods, and here the constants involved for the recurrence-relation method are about fifteen times greater than those for Horner's rule. It is fairly easy to see that, when d is small, the recurrence-relation method is not advantageous, since it involves multiplying the, potentially large, coefficients of the Sturm sequence and the q_i. The naïve method seems to have little to recommend it.

If we combine this data with proposition 2, we have shown the following result.

Proposition 3. *Using classical arithmetic, we can separate all the real roots of a square-free polynomial p in time $O(n^5 \log^3(n\|p\|_2))$.*

This result in fact seems to have little connection with reality. Heindel [1971] reported that the running time of his Sturm-sequence based method for isolating real roots seemed to be dominated by the cost of calculating the Sturm sequence, and the data presented by Collins & Loos [1976] indeed seem to show a $O(n^4)$ behaviour.

Of course, all these algorithms could benefit from the use of "fast" algorithms, e.g. $O(k \log k \log\log k)$ methods for multiplying k-bit integers. The analyses become distinctly tedious in this case, but the final results are shown in table 2. The recurrence-relation method is now definitely the fastest, but the practical impact of this should not be over-rated.

Table 2:	Fast $O(k \log k \log\log k)$ multiplication		
d	Naïve Method	Horner's Method	Recurrence Method
1	$n^3 \log(n\|\|p\|\|_2) \log n\, \epsilon$	$n^3 \log(n\|\|p\|\|_2)$	$n^2 \log(n\|\|p\|\|_2) \log n\, \epsilon$
$\log(n\|\|p\|\|_2)$	$n^3 \log(n\|\|p\|\|_2) \log n\, \epsilon$	$n^3 \log(n\|\|p\|\|_2)\epsilon$	$n^2 \log(n\|\|p\|\|_2) \log n\, \epsilon$
$n \log(n\|\|p\|\|_2)$	$n^4 \log(n\|\|p\|\|_2) \log n\, \epsilon$	$n^4 \log(n\|\|p\|\|_2) \log n\, \epsilon$	$n^3 \log(n\|\|p\|\|_2) \log n\, \epsilon$

ϵ stands for a variety of $\log\log$ terms

Proposition 3′. *We can separate all the real roots of a square-free polynomial p in time*

$$O(n^4 \log^2(n\|p\|_2) \log n\epsilon).$$

If we wish to isolate the distinct roots of a polynomial that may not be square-free, the obvious starting point is to find the square-free part (see section 2), and isolate its roots. Unfortunately, the coefficients of the square-free part may be larger (up to 2^n times larger — see Mignotte [1981]) than those of the original polynomial. Hence it might seem that we have to replace $\log\|p\|_2$ by $n \log\|p\|_2$ everywhere. Fortunately we can do somewhat better, since the only places that the coefficent length is used is in the length of the coefficients of the Sturm sequence and in proposition 6 and 8 of the previous section. The former we can do little about, but for the latter we can remark that we only used this via Landau's inequality, and the product of the roots larger than 1 can not increase. Hence these results can be applied with the $\|p\|_2$ for the original polynomial. The exact details are rather messy, but lead to two further tables, and propositions 4 and 4′ below.

Table 3:	Classical multiplication, arbitrary polynomials		
d	Naïve Method	Horner's Method	Recurrence Method
1	$n^5 \log(\|\|p\|\|_2)$	$n^4 \log(\|\|p\|\|_2)$	$n^5 \log^2(\|\|p\|\|_2)$
$\log(n\|\|p\|\|_2)$	$n^5 \log(\|\|p\|\|_2) \log(n\|\|p\|\|_2)$	$n^4 \log(\|\|p\|\|_2) \log(n\|\|p\|\|_2)$	$n^5 \log^2(\|\|p\|\|_2)$
$n \log(n\|\|p\|\|_2)$	$n^6 \log(\|\|p\|\|_2) \log(n\|\|p\|\|_2)$	$n^5 \log^2(n\|\|p\|\|_2)$	$n^5 \log(\|\|p\|\|_2) \log(n\|\|p\|\|_2)$

Table 4:	Fast multiplication, arbitrary polynomials		
d	Naïve Method	Horner's Method	Recurrence Method
1	$n^4 \log(\|\|p\|\|_2) \log n\, \epsilon$	$n^4 \log(\|\|p\|\|_2)$	$n^3 \log(\|\|p\|\|_2) \log n\, \epsilon$
$\log(n\|\|p\|\|_2)$	$n^4 \log(\|\|p\|\|_2) \log n\, \epsilon$	$n^4 \log(\|\|p\|\|_2)\epsilon$	$n^3 \log(\|\|p\|\|_2) \log n\, \epsilon$
$n \log(n\|\|p\|\|_2)$	$n^4 \log(n\|\|p\|\|_2) \log n\, \epsilon$	$n^4 \log(n\|\|p\|\|_2) \log n\, \epsilon$	$n^3 \log(n\|\|p\|\|_2) \log n\, \epsilon$

ϵ stands for a variety of $\log\log$ terms

Proposition 4. *Using classical arithmetic, we can separate all the distinct real roots of a polynomial p in time $O(n^6 \log^2(n\|p\|_2) \log n)$.*

Proposition 4′. *We can separate all the distinct real roots of a polynomial p in time*

$$O(n^4 \log^2(n\|p\|_2) \log n\epsilon).$$

There are other possible methods for isolating the real roots. Collins & Loos [1976] present one, based on the elementary remark that there is at most one root of p between two roots of p'. Hence, if we can isolate the roots of p', we can inductively isolate the roots of p. Note, however, that p' need not be square-free, even if p is. The theoretical running time of this algorithm is very bad: it is proved to be $O(n^{10} + n^7 \log^3\|p\|_2)$ by Collins & Loos [1976], but this can be improved to $O(n^9 + n^6 \log^3\|p\|_2)$ by applying proposition 8 of the previous section. In practice, though, it would seem from the figures quoted by Collins & Loos to be a $O(n^3)$ algorithm.

Proposition 5 (Descartes' Rule of Signs). *The number of positive real roots (counting multiplicity) of p is equal to, or an even positive integer less than, the number of variations of sign in the coefficients of p.*

Proof. Without loss of generality, we can consider p to be monic, and not to have 0 as a root (for then we could divide by x, which changes no coefficient). Let q be p divided by $(x - \alpha_i)$ for all positive real roots α_i of p. Then it is sufficient to prove that q has even variation, and that multiplying a polynomial by $x - \alpha_i$ increases the variation by an odd integer.

Now the leading coefficient of q is 1, so $q(x)$ is positive for large x. If the trailing coefficient of q were negative, then q would be negative at the origin, and hence have a positive real root. Hence the sequence of signs begins and ends $+$, and so must have an even number of variations. When we multiply a polynomial $r(x)$ by $(x - \alpha_i)$, the leading coefficient keeps the same sign, but the trailing coefficient changes sign, and hence the parity of the number of sign variations changes. We need only prove, therefore, that the number of variations can not decrease between $r(x)$ and $r(x)(x - \alpha_i)$. Suppose that we have a variation $+ -$ in r, say that the k-th coefficient, b_k, is positive and the $(k-1)$-st is negative. Then the k-th coefficient of $r(x)(x-\alpha_i)$ is $b_{k-1} - \alpha_i b_k$, and hence is certainly negative. Similarly a $- +$ variation in r leads to a positive coefficient in $r(x)(x-\alpha_i)$. (We have not considered the case of interposed zero coefficients, but these do not affect the argument.) Hence every variation in r gives rise to a variation in $r(x)(x-\alpha_i)$, so the total number of variations has not decreased. This proves the result.

Collins & Akritas [1976] have presented a further real root separation algorithm, based on Descartes' rule of signs. They use repeated bisection to separate the positive roots of p, and also to ensure that the "excess" even number of Descartes' rule is eventually reduced to 0. Their algorithm's theoretical computing time is given as $O(n^6 \log^2 \|p\|_2)$: again a factor of n can be removed by appealing to proposition 8 of the previous section. The practical running time of this algorithm seems to be $O(n^3)$, with a smaller constant than that of the previous paragraph.

All the analyses above that we have given relate to polynomials over the integers. The algorithms, though, are applicable to polynomials with real coefficients, in particular algebraic numbers. The bounds for the absolute value of the root are still applicable, in terms of the absolute values of the coefficients. The bounds for the separation (propositions 6 and 8) are not directly applicable unless we compute the discriminants, since these may be algebraic numbers less than 1 in absolute value. Fortunately, the algorithms do not need these bounds — they are only required for complexity analysis.

Here there is a major difference between the algorithms in terms of the manipulations to be performed on the coefficients: Sturm's method requires division, while the Collins & Loos method is based on differentiation and the Collins & Akritas method is based on bisection, substituting $2x$ or $2x-1$ for x. Thus, they may prove faster when applied to polynomials with algebraic number coefficients, though, to the best of my knowledge, this area has not been explored.

Clearly the whole question of practical real root isolation requires further study.

7. Algebraic Numbers (II).

We can now represent an algebraic number α as a triple $\langle p, l, r\rangle$, where p is a polynomial which α satisfies, and l and r are fractions (probably with binary denominators in practice) such that $l < \alpha < r$. The coefficients of p will certainly be real, and should either be integers or (simpler) real algebraic numbers. Given an algebraic number α, we will refer to the corresponding components of its description by p_α, l_α and r_α.

We now have a variety of questions to ask, of which the following is a sample: "If $\gamma = \alpha + \beta$, what are p_γ, l_γ and r_γ?". Proposition 2 of section 4 has largely answered the first part of this question, saying that $p_\gamma(x) = \mathrm{res}_y(p_\alpha(x-y), p_\beta(y))$. This resultant need not be square-free, so we might as well at least compute the square-free part of it. Let us note here, as a question be to considered later, that we could always perform a complete factorisation of this resultant, and select the factor which has $\alpha + \beta$ as a root. Whether or not we do a complete factorisation, we now have a square-free polynomial, which has γ as a simple root, and whose degree is at most $\delta(p_\alpha)\delta(p_\beta)$.

Certainly γ lies in the interval $(l_\alpha + l_\beta, r_\alpha + r_\beta)$, but this might not be an isolating interval. It may contain 0, but this by itself is easily rectified: we evaluate $p_\gamma(0)$ to decide which side of 0 γ lies, and then, if we wish, choose a point just away from 0 as a new left (or right) marker. Corollary 1 of section 5 can guide our choice of point. More fundamentally, though, there may be more than one root of p_γ in our interval, or at least we may not know that there is not.

Let us first make two preliminary remarks. If $p(l)$ and $p(r)$ have the same sign, then there is certainly more than one root, and we know that we will have to work on isolating the roots of p. The transformation $x \to 1/(r-x) - 1/(r-l)$ converts the interval (l, r) into the interval $(0, \infty)$, and Descartes' rule of signs (proposition 5 of the previous section) is now applicable. If the variation in sign of the coefficients is one, we know that there is only one root. If it is even, then we know that there are an even number of roots, but in fact this test is a re-phrasing of the one mentioned earlier. If the variation is odd, but greater than one, then we are no wiser. However, it can be shown that this implies that p has other zeros whose real parts lie in this interval, and this is unlikely to happen if p was constructed from isolating intervals for α and β which did not have other roots with the same real parts.

In general, though, we need to perform a real-root isolation process on p, and, as we saw in the previous section, this can be expensive. Once we know that there is more than one root of p in (l, r), we have to refine our values of l_α etc. in order to narrow (l, r). Since the interval (l_α, r_α) is an isolating interval, this refinement can be done merely by evaluating p_α at $(l_\alpha + r_\alpha)/2$, and observing in which half-interval the sign of p_α changes. We note here that it is probably worth always refining the larger interval in the case of $\gamma = \alpha + \beta$, and that different rules will apply for the different operations.

Such an isolation process will be expensive, if only because of the high degree that p_γ may have. As an example of the process we have just talked about, suppose that $\alpha = \langle x^2 - 3, 0, 2\rangle$ ("the positive square root of 3") and that $\beta = \langle x^2 - 2, -2, 0\rangle$ ("the negative square root of 2"). γ is then approximately $1.732 - 1.414 \approx .318$. The polynomial for γ is then $x^4 - 10x^2 + 1$, while the interval (l, r) in which we think γ lies is $(-2, 2)$. This contains two roots of the polynomial, and we have to narrow our intervals until $\alpha = \langle x^2 - 3, 1.5, 2\rangle$ and $\beta = \langle x^2 - 3, -1.5, -1\rangle$ before we have isolated γ as $\langle x^4 - 10x^2 + 1, 0, 1\rangle$. If we wanted our intervals not to include 0, we could either re-write this as $\langle x^4 - 10x^2 + 1, 1/11, 1\rangle$, relying on corollary 1 of section 5, or do further bisections and narrow our interval to $\langle x^4 - 10x^2 + 1, 0.25, 0.5\rangle$.

This poses a fundamental question: is this really necessary? Might we not be better off working with α and β, rather than with γ? The answer for addition (and similarly multiplication etc.) would seem to be that the "primitive element" approach of representing everything by a root of a polynomial will be very expensive. Furthermore, any cancellation, such as $\sqrt{2} - \sqrt{2}$, will be quite hard to spot. When it comes to the question of roots of polynomials with algebraic number coefficients, opinion is evenly divided. Schwartz & Sharir [1983] recommend a "recursive" representation, in which algebraic numbers are allowed in the coefficients of defining polynomials for other algebraic numbers. Arnon *et al.* [1984] recommend the opposite. Again, further research is called for.

II. Equations in Two Variables

As part of our general program of solving systems of equations in many variables, we now treat the case of two variables, conventionally x and y. The coefficients will be assumed to come from some field k, which the reader can think of as the rational numbers $\mathbf{Q}$. This is significantly easier to talk about (and to draw!) than the case of equations in arbitrarily many variables, so we will treat this case specially before doing the general induction. We will introduce a variety of notation and definitions which will be useful more generally. In particular, section 5 defines the concept of a *cylindrical algebraic decomposition*.

1. Two Polynomial Equations.

Suppose that p and q are two polynomials in $k[x,y]$, and we are interested in **common zeros** of them, i.e. values $\hat{x}$ and $\hat{y}$ of x and y such that $p(\hat{x},\hat{y}) = q(\hat{x},\hat{y}) = 0$. We may as well suppose that p and q are relatively prime, i.e. that there is no non-trivial polynomial that divides both of them, for all zeros of this common divisor g would be common zeros of p and q. In fact, the problem reduces to describing the zeros of g (which as a polynomial in two variables has an infinite number of zeros) and finding the common zeros of p/g and q/g.

Proposition 1. *The taking of resultants commutes with evaluation provided that the degrees of the polynomials involved do not change, i.e.*

$$(\mathrm{res}_y(p(x,y),q(x,y)))(x_1) = \mathrm{res}_y(p(x_1,y),q(x_1,y))$$

(where the left-hand side is a polynomial in x evaluated at the value x_1).

Proof. Since the degrees do not change, the resultants are both determinants of matrices of the same size, in one of which x has been evaluated. The result then follows from the same proposition about determinants of matrices with polynomial entries, and that is true since the determinant is a sum of products of the entries.

In order to ensure the applicability of this result, we will now make the further assumption that p and q are **y-monic**, i.e. that when regarded as polynomials in y with coefficients from $k[x]$, their leading coefficients are 1. Note that not every polynomial can be written this way: for example xy^2+1 is not y-monic, and trying to write it so gives us $y^2 + 1/x$, which is no longer a polynomial. This difficulty will prove one of the major stumbling blocks in our development, and we wish to bypass it for the moment.

Proposition 2. *The $\hat{x}$ which occur as x-components of common zeros of the relatively prime y-monic polynomials p and q are precisely the roots of $r(x) = \mathrm{res}_y(p,q)$.*

Proof. Let $(\hat{x},\hat{y})$ be a common zero. Define the polynomials $\hat{p}(y)$ and $\hat{q}(y)$ to be $p(\hat{x},y)$ and $q(\hat{x},y)$ respectively. These polynomials no longer have rational coefficients, but they are certainly polynomials in $\mathbf{C}[y]$. These polynomials have a common factor, viz. at least $(y-\hat{y})$, and hence (proposition I.3.2) their resultant $\mathrm{res}_y(\hat{p},\hat{q})$ is zero. But this is equal to $\mathrm{res}_y(p,q)(\hat{x})$ by proposition 1. Hence $\hat{x}$ is a root of $r(x)$.

Conversely, let $\hat{x}$ be a zero of $r(x)$. Defining $\hat{p}$ and $\hat{q}$ as before, we see that $\mathrm{res}_y(\hat{p},\hat{q}) = 0$, so that $\hat{p}$ and $\hat{q}$ have a non-trivial common divisor. If $\hat{y}$ is a root of this common divisor, then $(\hat{x},\hat{y})$ is a common zero of p and q.

Corollary 1. *Under the hypotheses stated, the $\hat{x}$ are algebraic numbers.*

Corollary 2. *Under the hypotheses stated, the $\hat{y}$ are algebraic numbers.*

The second corollary can be proved either by interchanging the roles of x and y (if p and q are also x-monic), or by observing that the $\hat{y}$ are the roots of $\gcd(p(\hat{x},y),q(\hat{x},y))$, and this is a polynomial with algebraic number coefficients, to which we can apply proposition I.4.4.

Corollary 3. *The $\hat{x}$ which occur as x-components of common real zeros of the relatively prime y-monic polynomials p and q are among the real roots of $r(x) = \mathrm{res}_y(p,q)$.*

It is not true that all real roots of $r(x)$ are the x-components of common real zeros of p and q, since the corresponding y values might not be real. As an example of this, consider two circles: $p = x^2+y^2-1$, $q = (x-4)^2+y^2-1$. In this case $r(x) = 64(x-2)^2$, and we have the value $\hat{x}=2$. But the corresponding y values are $\pm i\sqrt{3}$, and are not real. We note that $\hat{x}=2$ was a double root of the resultant, and that there were two corresponding y values. This is not a coincidence, as the following result shows.

Proposition 3. *The multiplicity of $\hat{x}$ as a common root of $r(x) = \mathrm{res}_y(p,q)$ is greater than or equal to the degree of* $\gcd(p(\hat{x},y), q(\hat{x},y))$.

Proof. Write $\hat{p}(y)$ for $p(\hat{x},y)$, and similarly for q. If p and q came from $R[x,y]$, $\hat{p}$ and $\hat{q}$ come from $R[y]$. Let k be the degree of $\hat{g}$, the gcd of $\hat{p}$ and $\hat{q}$. Then $\hat{S}$, Sylvester's matrix for $\hat{p}$ and $\hat{q}$, is singular, as in proposition I.3.2. Furthermore, it in fact has nullity* k, since each of the vectors corresponding to $(\hat{q}/\hat{g}, -\hat{p}/\hat{g})$, $(x\hat{q}/\hat{g}, -x\hat{p}/\hat{g})$, $(x^2\hat{q}/\hat{g}, -x^2\hat{p}/\hat{g})$, ..., $(x^{k-1}\hat{q}/\hat{g}, -x^{k-1}\hat{p}/\hat{g})$ is annihilated by $\hat{S}$, and these k vectors are linearly independent. So there is a sequence of invertible linear operations, defined over R, which transform $\hat{S}$ into a form where the last k rows are all zero.

Now S, Sylvester's matrix for p and q, can be written as $\hat{S} + (x-\hat{x})S_1$, where S_1 is Sylvester's matrix for $(p-\hat{p})/(x-\hat{x})$ and $(q-\hat{q})/(x-\hat{x})$. If we apply the same sequence of linear operations to S that we discovered for $\hat{S}$, we obtain a matrix where the last k rows are all divisible by $x-\hat{x}$, since all the terms from $\hat{S}$ have been annihilated. Hence the determinant of this matrix is divisible by $(x-\hat{x})^k$. But this determinant is the result of p and q, to within a factor coming solely from R, and hence $\hat{x}$ is at least a k-fold root of the resultant.

Having proved this inequality, it is natural to ask whether there is always equality between the degree of the gcd and the multiplicity of the root of the resultant. After all, the two were equal in the example we considered. The answer is no. For example, let p and q be polynomials of degree n in y whose resultant is not divisible by $(x-\hat{x})$. Then the gcd of $(x-\hat{x})^2 p$ and q is q, or degree n, but the resultant of $(x-\hat{x})^2 p$ and q is divisible by $(x-\hat{x})^{2n}$. As another example, $\mathrm{res}_y(y-x^3, y-x^2) = x^2(x-1)$, but the gcd of two linear polynomials has to be linear. What is happening here is that S_1 is itself divisible by x, which gives an extra factor of x in the resultant. However, it is true to say that equality in the previous proposition is the "normal" case.

2. Vanishing Leading Coefficients.

Throughout the previous section, we have assumed that p and q were y-monic, i.e. that their leading coefficients were 1 (or, at least, that they did not depend on x). Let us write $\mathrm{lc}_y(p)$ for the leading coefficient of p, regarded as a polynomial in y, and $\mathrm{red}_y(p)$ for the *reductum* of p, i.e. p minus its y-leading term. In this section, we will explore what happens when this restriction is lifted. It should be noted that this section is particular to the case of two variables, and the more general treatment will be somewhat different. We have four types of common zeros $(\hat{x},\hat{y})$ to consider.

1 $lc_y(p)(\hat{x}) \neq 0$, $lc_y(q)(\hat{x}) \neq 0$. In this case, the work of the previous section carries through unaltered. All the results of that section are still true for these zeros, since the only assumptions made were that the leading coefficients did not vanish at any of the zeros.

2 $lc_y(p)(\hat{x}) \neq 0$, $lc_y(q)(\hat{x}) = 0$. If q vanishes identically at $\hat{x}$, then every coefficient must be divisible by $x-\hat{x}$, and so the resultant certainly vanishes at $\hat{x}$. Otherwise, $\mathrm{red}_y(q)$ is a non-zero polynomial, even when $\hat{x}$ is substituted. We can assume that it does not have degree 0, for then there can be no common zeros.

In this case, any zero is a common zero of p and $\mathrm{red}_y(q)$, as well as of $\mathrm{lc}_y(q)$. Hence

$$\gcd(\mathrm{lc}_y(q), \mathrm{res}_y(p, \mathrm{red}_y(q))) \neq 1,$$

and the x-parts of these zeros are the roots of this gcd. But every term in the determinant of Sylvester's matrix for p and q either contains, and so is divisible by, $\mathrm{lc}_y(q)$, or corresponds to a term in $\mathrm{res}_y(p, \mathrm{red}_y(q)))$, but multiplied by a power of $\mathrm{lc}_y(p)$. This power is, in fact, $\delta_y(q) - \delta_y(\mathrm{red}_y(q))$, since this is the number of rows by which Sylvester's matrix has shrunk. Call this k. Hence

$$\mathrm{lc}_y(p)^k \gcd(\mathrm{lc}_y(q), \mathrm{res}_y(p, \mathrm{red}_y(q))) = \gcd(\mathrm{lc}_y(q), \mathrm{res}_y(p,q)),$$

and so these x-values are, in any case, roots of the resultant. So these roots will, in fact, be found by the same process as in the previous section, and there was no need to make any special case.

---

* We only need, and only prove, that the nullity is at least k. But the techniques of I.3.2 easily show that the nullity must be exactly k.

3 $lc_y(p)(\hat{x}) = 0$, $lc_y(q)(\hat{x}) \neq 0$. This case is clearly identical.
4 $lc_y(p)(\hat{x}) = 0$, $lc_y(q)(\hat{x}) = 0$. The same reasoning can be applied twice, and we see that all the roots for which both leading coefficients vanish are roots of the resultant. This is a result that could have been predicted anyway, since every term of the determinant is divisible by one or other leading coefficient.

Hence we have proved the following result, which is a weaker form of proposition 2 of the previous section.

Proposition. *The $\hat{x}$ which occur as x-components of common zeros of the relatively prime polynomials p and q are among the roots of $r(x) = \mathrm{res}_y(p, q)$.*

We can not say that the two are precisely the same any more. Consider $p = (x-1)y + 1$ and $q = (x-1)y + 2$. Since both leading coefficients vanish when $x = 1$, the resultant certainly vanishes here. But neither polynomial has a root there, so there is certainly no common root. (It would be possible to argue that there is a common root "at infinity". This would take us into projective geometry, which is very interesting, though not directly relevant. In projective geometry many of the questions relating to vanishing leading coefficients become somewhat simpler.) Hence the restriction in the previous section to y-monic polynomials was necessary as far as the precise statement of the results was concerned, though the conclusions we draw from them remain valid.

3. Expressing the Roots.

Proposition 1. *Let δ_x and δ_y stand for degrees in x and y. Then $\delta_x \,\mathrm{res}_y(p, q) \leq \delta_x(p)\delta_y(q) + \delta_x(q)\delta_y(p)$.*

Proof. $\mathrm{res}_y(p, q)$ is a determinant, where $\delta_y(q)$ rows contain the coefficients of p, which have degree at most $\delta_x(p)$. These rows will contribute at most $\delta_x(p)\delta_y(q)$ to the degree of a product of elements, one from each row. The $\delta_y(p)$ rows whose elements are coefficients of q will contribute at most $\delta_x(p)\delta_y(q)$, and since the determinant is a sum of such elements, its degree is at most $\delta_x(p)\delta_y(q) + \delta_x(q)\delta_y(p)$.

Corollary. *All the $\hat{x}$ are roots of a polynomial of degree at most $\delta_x(p)\delta_y(q) + \delta_x(q)\delta_y(p)$.*

This formula is symmetric in x and y, and therefore (if p and q are x-monic) this result is also true of $\hat{y}$. In fact a much stronger result is true.

Proposition 2. *All the $\hat{x}$ and $\hat{y}$ can be expressed in terms of the roots of a polynomial of degree at most $\delta_x(p)\delta_y(q) + \delta_x(q)\delta_y(p)$.*

Proof. The result is trivial if the resultant has no repeated roots, for then each $\hat{y}$ is determined by a linear polynomial in the corresponding $\hat{x}$. If we do have repeated roots, then the situation is more complex. Let the resultant have degree n, and factorise as $\prod p_i^{n_i}$, where the p_i are square-free polynomials of degree d_i. Then $\sum n_i d_i = n$. We will treat each p_i separately, and prove that its $\hat{x}$ and $\hat{y}$ can be expressed in terms of the roots of a polynomial of degree at most $n_i d_i$, and then the result will follow by multiplying these polynomials together. By proposition 3 of section 1, $\hat{y}$ is a root of a polynomial $q(z)$ of degree at most n_i, whose coefficients involve $\hat{x}$, roots of a polynomial of degree d_i. We may as well assume that q is square-free (but see the "remark" following the proof). Proposition I.4.4 assures us that we can express the $\hat{y}$ in terms of a polynomial $r(z) = \mathrm{res}_w(q(z)_{\hat{x}:=w}, p_i(w))$ of total degree $\delta_z(q)\delta_x p_i(x) \leq n_i d_i$

Regrettably†, there is no guarantee that $\hat{x}$ will be expressible in terms of this polynomial. Indeed, suppose that $p_i(x) = x^2 - 2$, so that $\hat{x} = \sqrt{2}$ or $-\sqrt{2}$, while $q(z) = z^2 - 3$, so that $\hat{y} = \sqrt{3}$ or $-\sqrt{3}$. Since q does not depend on $\hat{x}$, the resultant is merely $q(z)^2$, and we have information only about the $\hat{y}$. We intend to prove the following, which will complete the proof of this proposition.

For all but finitely many values of the integer λ, $r_\lambda(z) = \mathrm{res}_w(q(z - \lambda w)_{\hat{x}:=w}, p(w))$ is a square-free polynomial, in terms of whose roots we can express all the roots of p and q.

Let α_i be all the roots of p, and for each root α_i of p, let β_{ij} be the corresponding roots of $q(z)$, whose coefficients may depend on α_i. Now the roots of $r_\lambda(z)$ are all the $\beta_{ij} - \lambda\alpha_i$, by corollary I.3.2. We wish

† Algebraists will recognise the rest of this proof as a "low-brow" version of van der Waerden's proof of the *Theorem on the Primitive Element*, section 43, pp. 138–139 (but section 40, pp. 126–127 in the english translation). The complications arise because we do not necessarily have minimal polynomials for α or β.

to show that $r_\lambda(z)$ is square-free, i.e. that its roots are distinct. So, when is $\beta_{ij} - \lambda\alpha_i = \beta_{i'j'} - \lambda\alpha_{i'}$? If $i \neq i'$, then $\alpha_i \neq \alpha_{i'}$ since p is square-free, and hence we have a linear equation for λ, with at most one integer solution for each set $(i, j, i' \neq i, j')$, i.e. finitely many solutions. If $i = i'$, then the equation reduces to $\beta_{ij} = \beta_{ij'}$, and this is impossible since q was assumed square-free. Hence there are only finitely many values of λ for which $r_\lambda(z)$ is not square-free. Choose λ as a fixed integer away from any of these values, and write $r(z)$ for $r_\lambda(z)$. Let γ be a root of $r(z)$ so that $\gamma = \beta_{ij} - \lambda\alpha_i$ for some (i, j). The only common root of $p(z)$ and $q(\gamma - \lambda z)$ is α_i, so the gcd of these polynomials will give a defining equation for α_i in terms of γ. Once we know γ and α_i, we have $\beta_{ij} = \gamma + \lambda\alpha_i$.

Remark. Having just finished such a lengthy proof, it is somewhat embarrassing to have to admit that a complication (which seems to have escaped some previous writers on the subject) was glossed over in the proof. We said "We may as well assume that q is square-free" as if this were a triviality. Now q is a polynomial in z, whose coefficients involve the roots of $p(x)$. When we say that q is square-free, we mean more than that it should be square-free as an element of $k[x, z]$: we need it to be square-free even when the dependencies of its coefficients on the algebraic numbers $\hat{x}$ have been taken into account. A trivial example will make this clear. Suppose that p is $x - 1$, and q is $(z - 1)(z - x)$. q is square-free as an element of $k[x, z]$, but not when we take the value of x, i.e. $x = 1$ into account. Before we describe how to do this in general, let us first remark that, if q does not involve x, there is no problem. We will also assume that q has been made square-free as an element of $k[x, z]$.

Since q is square-free as an element of $k[x, z]$, its discriminant $d(x) = \mathrm{res}_z(q,\, dq/dz)/\,\mathrm{lc}(z)$ is non-zero. $d(x)$ is a polynomial in x, which take on the value zero if, and only if, q is not square-free for that particular value of x. There are three possible cases to consider.

1) $\gcd(d(x), p(x)) = 1$. Then d is non-zero at all roots of p, and q is indeed square-free.
2) $p(x)$ divides $d(x)$. Then q is square-free at none of the roots of p. If we perform a gcd calculation on q and dq/dz, we will end up with a polynomial which is divisible by $p(x)$. Hence this polynomial should be counted as 0, and we have found a gcd r of q and dq/dz. We should now replace q by q/r and repeat the process (we can not have case (2) again, but we might arrive in case (3)). Of course, we could take advantage of the fact that we have found a partial factorisation of q, if we wished to keep our polynomials more factored.

   As an example of this case, consider $p(x) = x^2 - 2$, $q(y) = (y^2 - 2)(y - x)$. Since x is a square root of 2, $y - x$ is a factor of $y^2 - 2$: how can we discover this fact? Note that we can not just plough ahead with the algorithm given in the proof of proposition 2. $r_\lambda(z) = \mathrm{res}_w(q(z)_{x:=w}, p(w))$ is $(y^2 - 2\lambda^2 + 4\lambda - 2)(y^2 - 2\lambda^2 - 4\lambda - 2)^2$, and so is never square-free. $\mathrm{res}_y(q(y),\, dq(y)/dy) = -8(x^2 - 2)^2$, and this is certainly divisible by p, so that we know that q is not really square-free. This is the last element in Euclid's algorithm as applied to q and dq/dy, and is really 0. The previous element was $-yx^2 - 6y + 8x$, which can be re-written as $-8(y - x)$. Hence $y - x$ is a repeated factor of q.
3) Neither of the above, i.e. $\gcd(d(x), p(x))$ is a polynomial between 1 and p. As a particular consequence of this, we have found a factor of p, and so this case can not happen if p is irreducible. Write $p_1 = \gcd(d, p)$ and $p_2 = p/p_1$. Then q is square-free at the roots of p_2, but not at those of p_1. We have to split our analysis in the proof of proposition 2 into two cases, and consider $\mathrm{res}(q, p_2)$ and $\mathrm{res}(\hat{q}, p_1)$, where $\hat{q}$ is the square-free part of q (when x is a root of p_1). This splitting does not change our analysis of the total degrees involved. Of course, p_1 may split further, so that a practical implementation would have a recursive element here.

   As an example of this phenomenon, consider the, slighty contrived, case of $p = (x - 1)(x - 2)$ and $q = (y - 1)(y - x)$. $r_\lambda(z) = \mathrm{res}_w(q(z)_{x:=w}, p(w))$ is $(y - \lambda - 1)^2(y - 2\lambda - 1)(y - 2\lambda - 2)$, and again is never square-free. $\mathrm{disc}_y(q(y),\, dq(y)/dy) = -(x - 1)^2$, so that we have to split p into its factors $x - 1$ and $x - 2$. Corresponding to the factor $x - 2$, q is square-free and has two roots (actually 1 and 2), while corresponding to the factor $x - 1$, $q = (y - 1)^2$, and has one root 1.

Let us conclude this section by making some remarks on the practical implementation of these remarks on the solutions of two polynomials. For simplicity, let us assume that p and q have the same degree n in x and y, and that their coefficients have at most d digits. We can find $\mathrm{res}_y(p, q)$, and this is a polynomial of degree $2n^2$, whose coefficients have $O(n(d + \log n))$ digits. A modular method due to Collins [1971] can compute this in $O(n^5 d + n^4 d^2)$ operations. We should clearly perform a square-free decomposition of this polynomial, in order to understand the way in which y depends on x. We will also need to discover

how many real roots there are, and, by the results of section I.6, this will take at most $O(n^{15}(d+\log n)^3)$ operations using Sturm sequences and classical arithmetic, $O(n^{10}\log n\,(d+\log n)^2)$ using Sturm sequences and fast arithmetic, and possibly as few as $O(n^6)$ if we use one of the other algorithms presented there. For each of these (potentially n^2) real roots, we have to discover if the corresponding y values are real. y can be expressed either as the root of a polynomial with coefficients involving x, and this polynomial will have degree at most n. Or we can express it as a root of a polynomial having degree at most $2n^2$, but with integer coefficients. In either case, the real root isolation is likely to prove very expensive.

4. One Polynomial Equation.

Now let us consider a very special case of the previous sections: that of one square-free polynomial p and its y-derivative dp/dy. Then the resultant we considered before is now just the y-discriminant of p, $\mathrm{disc}_y(p)$. Can we attach any special meaning to this?

Firstly, we note that, as a corollary of proposition 1 of section 1, the discriminant commutes with evaluation, i.e. that

$$(\mathrm{disc}_y(p(x,y)))(x_1) = \mathrm{disc}_y(p(x_1,y)).$$

Hence roots of the discriminant will tell us about those special x-values at which the discriminant vanishes. But the discrminant of a univariate polynomial is zero if, and only if, that polynomial has a repeated root. So the roots of the discriminant will tell us about those x-values at which the polynomial $p(x,y)$ has a repeated root. In particular, we learn about all self-crossings.

Secondly, we can look at the discriminant as a resultant: From section 1, we know that the roots of it are the x-values at which both p and dp/dy are zero. The slope of the tangent to the curve of y as a function f of x given by $p(x,y)=0$ is exactly $1/\,dp/dy$. Hence, we are finding the points on the curve at which the slope of the tangent is infinite, i.e. the tangent is vertical. These points are, in effect, also multiple points, since what happens is that two distinct branches of the function f meet at this point.

Proposition. *Between roots of the discriminant, the number of real zeros of $p(x,y)$ (regarded as a function of y alone) is independent of x.*

We will not prove this result here, since the proof is analytic rather than algebraic. In outline, though, the proof says that, as x varies, *one* real root can not become complex, since complex roots come in conjugate pairs. Hence there must have been two co-incident real roots, i.e. a root of the discriminant. A complete proof on these lines is given by Schwartz & Sharir [1983, pp. 321–322].

As an example of this, consider the polynomial

$$y^4 - 2y^3 + y^2 - 3x^2y + 2x^4,$$

which is drawn in figure 1 (taken from Arnon *et al.* [1984]).

The discriminant of this polynomial is

$$x^6(2048x^6 - 4608x^4 + 37x^2 + 12),$$

which has five real roots, whose approximate values are 0, ± 1.4969 and ± 0.2365. The last four correspond to the points with vertical tangents, while the root of 0 corresponds to the self-crossings. Hence we can characterise the various x-values for which this polynomial has different numbers of real roots, and show that the structure really is as we have drawn it, with two self-crossings etc. Furthermore, and this is our real use of it, we can divide two-dimensional space into different regions, on each of which the polynomial has a constant sign, positive, negative or zero.

5. Cellular Algebraic Decompositions.

The time has come for some general definitions to provide a framework into which the previous sections can be placed. These definitions will be stated in terms of n-dimensional real space $\mathbf{R}^n$. The definitions are taken largely from Arnon *et al.* [1984].

Definition. *A nonempty connected subset of* $\mathbf{R}^n$ *is termed a* **region**, *or an* n*-region if we wish to make the dimension of the containing space clear.*

Note that the dimension of the region itself may well be less than n. A point of $\mathbf{R}^n$, for example, is a 0-dimensional n-region.

Definition. *Given an* n*-region* A, *the* **cylinder over** a, *written* $Z(A)$ *is the set* $A \times \mathbf{R}$.

This is, in fact, an $(n+1)$-region.

Definition. *Suppose we are given functions* $f_1, \ldots, f_k$ *from a region* A *into* $\mathbf{R}$, *with* $f_i < f_{i+i}$ *throughout* A. *Then we say that* $f_1, \ldots, f_k$ *determine a* **stack** *over* A, *which is a partitioning of* $Z(A)$ *into the* $2k+1$ *regions* $\{(a,x) : a \in A, x < f_1(a)\}$, $\{(a,x) : a \in A, x = f_1(a)\}$, $\{(a,x) : a \in A, f_1(a) < x < f_2(a)\}$, $\ldots$, $\{(a,x) : a \in A, x = f_k(a)\}$, $\{(a,x) : a \in A, f_k(a) < x\}$.

The case $k = 0$ is legal, when the stack is just the whole cylinder. it is often convenient to add the "functions" $f_0(a) = -\infty$ and $f_{k+1}(a) = \infty$, so that the stack consists of $(k+1)$ regions where x is sandwiched between two functions, and k where x is defined precisely by a function.

Proposition 1. *By the constraints on the* f_i, *the objects defined by the equalities and inequalities really are regions. Furthermore they are disjoint, and partition* $Z(A)$.

Definition. *A* **cylindrical decomposition** *of* $\mathbf{R}^n$ *is defined recursively as a set of stacks: one over each region comprising a cylindical decomposition of* $\mathbf{R}^{n-1}$. *A cylindrical decomposition of* $\mathbf{R}^0$ *is one point.*

Hence a cylindrical decomposition of $\mathbf{R}^1$ is a set of points splitting the line up into segments. Each point and each line segment is an element of the decomposition. A cylindrical decomposition of $\mathbf{R}^2$ is obtained from this by erecting cylinders above each point or line segment, and then splitting it up into a stack, and so on.

With these definitions in mind, we can express the conclusion of the previous section better. The discriminant divides $\mathbf{R}^1$ into a cylindrical decomposition, with six line segments and five points. Above the left-hand line segment $(-\infty, \approx -1.4969)$, we have a cylinder consisting of just one region, and on this region the polynomial is positive. Above the point ≈ -1.4969 we have a stack consisting of two semi-infinite line segments and one point $(x \approx -1.4969, y \approx 1.759)$. The polynomial is zero at the point, and positive on the two line segments. Above the segment $(\approx -1.4969, \approx -.2365)$, we have two semi-infinite regions, one finite but two-dimensional region, and the two one-dimensional regions which separate the finite region fron the semi-infinite ones. The polynomial is positive on the semi-infinite regions, zero on the one-dimensional ones, and negative on the two-dimensional interior region. We have 18 two-dimensional regions, 27 one-dimensional regions and 10 zero-dimensional regions in the decomposition.

Definition. *A decomposition is said to be* **algebraic** *if the defining functions are all algebraic functions, i.e. solutions of polynomial equations.*

The importance of algebraic decompositions is that they are computable, whereas there are many unsolved problems in the theory of transcendental functions that might make a more general decomposition uncomputable. The importance of cylindrical decompositions is that there is an algorithm [Collins, 1975] to compute them, based essentially on induction in the dimension. The decomposition we have computed is both cylindrical and algebraic: such decompositions form the main thrust of the rest of this course, and will often be abbreviated as *c.a.d.*

Definition. *A* **sample point** *for a region in a c.a.d. is any point in that region. A* **sample set** *is a set of sample points: one for each region in the c.a.d.*

It will be in our interest to choose the simplest possible sample points for the various regions of a c.a.d. For example, we could choose $(-2, 0)$ for the large region at the left, whereas the 0-dimensional region $(\approx -1.4969, \approx 1.759)$ mentioned earlier contains only one point, which has to be its sample point. This point is

$$(\alpha = \langle 2048x^6 - 4608x^4 + 37x^2 - 12, -2, -1\rangle, \frac{-1024\alpha^4 + 2904\alpha^2 + 89}{827}),$$

which could also be written as

$$(\langle 2048x^6 - 4608x^4 + 37x^2 - 12, -2, -1\rangle, \langle 32x^4 - 96x^3 + 77x^2 - 12x - 1, 1.5, 2\rangle).$$

Proposition 2. *An r-dimensional n-region B in a c.a.d. has $(n - r)$ algebraic co-ordinates, and r that can be chosen to be rational.*

Proof. By induction on n. This n-region B arose from an $(n-1)$-region A in one of two ways: either that region has dimension r, and we took a region $\{(a, x) : a \in A, x = f_i(a)\}$ (for some i) as B; or A has dimension $(r-1)$, and we took a region $\{(a, x) : a \in A, f_i(a) < x < f_{i+1}(a)\}$ (for some i) as B. In the former case, we had r rational co-ordinates, and the co-ordinate we are adding is constrained to be the solution of an algebraic equation, while in the second case we had $(r-1)$ rational co-ordinates, and the new value of x as the n-th co-ordinate can be chosen to be any rational number between the values of f_i and f_{i+1} at the sample point of A.

Of course, some of the algebraic co-ordinates *may* turn out to be rational, but that is a rare stroke of luck. In any case, the point is that we have no choice over these co-ordinates.

Definition. *Let F be a set of polynomials in n variables, and C a decomposition of $\mathbf{R}^n$. We say that C is* **sign-invariant** *for F if each element of F has the same sign (positive, negative or zero) throughout each component of C.*

Then we have computed a sign-invariant c.a.d. for our polynomial. There is a stronger notion that is sometimes useful, but which requires an auxiliary definition.

Definition. *Let f be a function of n variables, and p a point in n-space. We say that f has* **order** *k at p if k is the least integer such that one of the partial derivates of f of order k does not vanish at p. If there is no such k, we say that the order is ∞.*

Order 0 means that $f \neq 0$ at p.

Definition. *Let F be a set of polynomials in n variables, and C a decomposition of $\mathbf{R}^n$. We say that C is* **order-invariant** *for F if each element of F has the same order throughout each component of C.*

Proposition 3. *An order-invariant decomposition is also sign-invariant.*

Proof. The only way in which this could happen would be for a polynomial of F to be both positive and negative on a region of C. But regions are connected and polynomials are continuous functions, so the polynomial would have to take on the value 0 between positive and negative values, and then the decomposition would not be order-invariant.

The converse is not true. As an example of this, take the function xy, and the decomposition consisting of the two axes (as one component) and the four quadrants that they enclose. This decomposition is sign-invariant, since the polynomial is zero on the first component, and positive or negative throughout each of the others. But it is not order-invariant, since the polynomial has order 2 at the origin (xy, x and y all vanish there), and 1 elsewhere on the first component.

6. C.A.D. for One Polynomial

Let us now consider how to calculate a cellular algebraic decomposition for the case of one polynomial, i.e. $F = \{f\}$. We have almost all the machinery at our disposal, but not quite all. Let us suppose that f has integral coefficients, and also initially that it is square-free. Write n for the maximum degree of f in any of its variables, and $\|f\|_2$ for its norm (generalising slightly the definition of norm given in section I.5: now $\|f\|_2 = \sqrt{\sum \sum a_{ij}^2}$ where the a_{ij} are the coefficients of f). The interesting variable is really d, the length (i.e. logarithm to a suitable base) of $\|f\|_2$.

There are a couple of trivial cases to deal with first. If f has degree 0 in y, i.e. is independent of y, then a c.a.d. of $\mathbf{R}^2$ sign-invariant for f is just the result of multipliying each element of a c.a.d. of $\mathbf{R}^1$ invariant for f by $\mathbf{R}^1$. If f has degree 1, then its discriminant is 1, and there we get no information from this. But there is one value of y above each value of x where the leading coefficient does not vanish, no values where the leading coefficient does vanish and the y^0 coefficient does not vanish, and a complete $\mathbf{R}^1$ where the both coefficients vanish, and f imposes no constraints at all on y.

Proposition 1. *A c.a.d. of $\mathbf{R}^2$ sign-invariant for some non-trivial $f \in \mathbf{Q}[x, y]$ can be obtained by taking a decomposition of $\mathbf{R}^1$ induced by $\mathrm{disc}_y(f)$, and constructing stacks over each region of this corresponding to all the real branches of f.*

Proof. This is "geometrically obvious", since the discriminant gives us all the "critical points" of f. The proposition of section 4 says that the number of roots of p is invariant between roots of the discriminant, i.e. on each of the one-dimensional regions of the decomposition of $\mathbf{R}^1$. Above the zero-dimensional regions $x = \alpha$, either f vanishes identically (so that $x - \alpha$ divides f), in which case we have a stack consisting of the whole cylinder $Z(x = \alpha)$, or it does not, when we can locate the roots precisely. These roots will give us zero-dimensional cells on which f vanishes, and between them f will be sign-invariant.

This then gives us essentially a three-phase algorithm, which is a paradigm to which we will return later.

1) Compute the discriminant, i.e. project our problem into fewer dimensions.
2) Solve the problem there, which in our case means that we compute a c.a.d. of $\mathbf{R}^1$.
3) Extend this c.a.d. to a c.a.d. of $\mathbf{R}^2$.

Step 1 can be solved in time $O(n^5 d + n^4 d^2)$, and leads us to a discriminant of degree $2n^2$ and coefficient length $O(n(d + \log n))$. As stated in section 3, the real root isolation, i.e. the computation of the c.a.d. of $\mathbf{R}^1$, will take $O(n^{15}(d + \log n)^3)$ operations classically, or $O(n^{10} \log n\ (d + \log n)^3)$ operations using fast arithmetic.

By proposition 2 of section 3, all the common roots of f and df/dy can be expressed in terms of a polynomial of degree at most $2n^2$, and so it might seem that all we need to do is to find them in order to find the zero-dimensional cells of our decomposition of $\mathbf{R}^2$. To see why this is not true, consider the 1-region $(x \approx -.2365)$, which is formed from the small negative root of the discriminant. Above this there lies a true critical point of the discriminant, whose y-value is $\approx .30025$. But there are two other real points of the curve with this x-value, and y-values $\approx .05$ and ≈ 1.35. These two points also enter into the cylindrical algebraic decomposition, even though they might not seem necessary. To clarify notation, we will call such points **apparent critical points**, to distinguish them from the **true critical points** that are common zeros of f and df/dy, and were considered in section 3.

To compute the cylindrical algebraic decomposition, we need to discover the structure of the stack lying over each 1-region. This means isolating all the real roots of f lying above a sample point of this 1-region, such that the roots determine sample points of 2-regions with the same dimension as the underlying 1-region, and points between the roots determine sample points of 2-regions with dimension one more than that of the 1-region.

Let us first deal with lifting a one-dimensional 1-region. This is defined by a rational sample point p_i, and let us suppose that the numerator and denominator of p_i have at most e_i digits. Substituting this into f we obtain a univariate polynomial of degree at most n and coefficients with $O(ne_i + d)$ digits. We can separate its roots in time $O(n^5(ne_i + d)^3)$ (ignoring $\log n$ terms). Summing this over all the one-dimensional regions, we obtain $O(n^5(n^2 d^3 + \cdots + n^3 \sum e_i^3)) \leq O(n^5(n^2 d^3 + \cdots + n^3(\sum e_i)^3))$. But the e_i are bounded by the separation (more accurately, since the e_i are numbers of digits, by the logarithm of the separation) of the roots of the discriminant, and we can use proposition I.5.8 to show that $\sum e_i$ is bounded by $O(n^2(\log(n^2) + n(d + \log n))) = O(n^3(d + \log n))$. Substituting this in gives us $O(n^{18}(d + \log n)^3)$, or

$O(n^{12}(d+\log n)^2 \log n)$ if we use fast arithmetic: a pretty frightening prospect. This is, of course, a worst case, and assumes that the discriminant had very closely spaced roots and that the root-separating process for the univariate polynomials that resulted also took worst-case time. Notice, as a matter of interest, that the major component in n^{18} or n^{12} arose from the $\log \|p\|_2$ components in the statement of proposition I.5.8.

This leaves us with the problem of extending the zero-dimensional 1-regions. Here we have to find the real roots of a univariate polynomial whose coefficients are algebraic numbers. One way of doing this is to reduce the problem, via the techniques of proposition I.4.4, to that of finding the roots of a polynomial with integer coefficients. This polynomial will have degree at most n^3, and integer coefficients with $O(n^2(\log n) + (n^2 d + n \log n))$ digits: call this $O(n^2 d)$ in keeping with our cavalier attitude to factors of $\log n$. Hence the accuracy required for root separation, from section I.5, will be $O(n^5 d)$, and this technique is the only known one for bounding the root separation. Finding the real roots of this polynomial will take $O(n^{24}d^3)$ operations, or $O(n^{16}d^2)$ with fast arithmetic, and the same remarks as before apply. We need only do this operation once: it will give us all the real roots of $f(\hat{x}, y)$ for all the $\hat{x}$ which are roots of the discriminant, whether they be true or apparent critical points. We still need to decide which of the real roots we have calculated belong to which of the $\hat{x}$ (if any: we could have found real roots here which corresponded to imaginary $\hat{x}$). While we could (and perhaps should) think of various ways of doing this in two dimensions, such techniques will not generalise.

The alternative technique is to use a root-isolating method for polynomials with algebraic coefficients. This is rather hard to analyse, but it would seem that the total cost, using the bounds on root separation from before, would be about the same.

In any case, it is clear that this is the most expensive step in the process. This is disappointing, because it would seem to be a conceptually unnecessary step, and merely an artefact of our cylindrical approach to decomposing — though no-one has proposed a better approach.

Proposition 2. *A c.a.d. of $\mathbf{R}^2$ sign-invariant for an arbitrary $f \in \mathbf{Q}[x,y]$ can be obtained by taking a decomposition of $\mathbf{R}^1$ induced by $\mathrm{disc}_y(g)$, or $\mathrm{lc}_y(g)$ if $\delta_y(g) \leq 1$, where g is the square-free part of f, and constructing stacks over each region of this corresponding to all the real branches of f. The computing time for this is bounded by $O(n^{16} \log n\,(d+\log n)^2)$, where n is the maximum degree of f in either variable.*

Proof. If the polynomial is not square-free, we have merely to make it square-free, since sign-invariant c.a.d.s are the same for both. The remark about $\mathrm{lc}_y(g)$ was made just before proposition 1, and the rest of the first sentence is a re-statement of proposition 1. The time complexity comes from the preceding analysis, where the last step dominates all the others (and the cost of a square-free decomposition of f). We have ignored any protential growth in the coefficients of g with respect to f: in principle this would add an extra factor of n to d, but in practice such growth is unlikely, and implies such special properties of the original polynomial that we will ignore it.

Proposition 3. *A c.a.d. of $\mathbf{R}^2$ order-invariant for a square-free $f \in \mathbf{Q}[x,y]$ can be obtained by taking a decomposition of $\mathbf{R}^1$ induced by $\mathrm{disc}_y(f)$ and $\mathrm{lc}_y(f)$, and constructing stacks over each region of this corresponding to all the real branches of f. The computing time for this is bounded by*

$$O(n^{16} \log n\,(d+\log n)^2),$$

where n is the maximum degree of f in either variable, and d is the length of the norm.

Proof. This is certainly stronger than the sign-invariant c.a.d. that proposition 2 constructed. How could it fail to be order-invariant? In general, the order of f changes only when f and all its derivatives up to a certain order vanish, and we have isolated all the points where f and df/dy both vanish. The only trouble is that this may be a complete 1-dimensional region in the case that f vanishes entirely along a strip perpendicular to the x-axis. But this can certainly only happen if the leading coefficient of f vanishes entirely. The computing-time calculations are the same.

We will have to postpone the discussion of order-invariant c.a.d.s for non square-free f, since this requires consideration of the c.a.d. induced by all the square-free factors of f, because the order of f can change where two of these meet.

The analysis of section 3 may well now seem redundant. Since we have to find the apparent critical points as well as the true one, why should we look for a simple method to find the true critical points? There

are two reasons for this. The first is that, if there are *no* true critical points, then the 1-region defined by this root of the discriminant is irrelevant, and can be discarded, and we can merge the two 1-dimensional 1-regions adjacent to it into one region. This simplification is based on true critical points only, and therefore requires the analysis of section 3. The second reason is that it may be cheaper to find the true critical points this way, and divide them out of the equation defining the totality of critical points before searching for the apparent ones. As far as I know, this question has not been experimentally analysed.

The reader may well feel that the analyses we have done are ludicrously pessimistic, and in some sense they are. But the problems of coefficients growth and root isolation that they describe do seem to be real. McCallum [1985] quotes the following example. Consider the decomposition of the plane induced by the polynomial

$$\begin{aligned} &x^6(y^2-2y+5)+x^5(-8y^4+26y^3-66y^2+56y-8)+x^4(16y^6-72y^5+201y^4-252y^3+156y-76)\\ &+x^3(-4y^6-56y^5+494y^4-788y^3+176y^2+340y-162)\\ &\qquad+x^2(-128y^6+242y^5+318y^4-754y^3+46y^2+456y-180)\\ &\qquad+x(144y^6-804y^5+1476y^4-936y^3-168y^2+396y-108)\\ &\qquad\qquad+229y^6-606y^5+363y^4+284y^3-405y^2+162y-27 \end{aligned}$$

which is actually the discriminant of

$$(y-1)z^4+xz^3+x(1-y)z^2+(y-x-1)z+y$$

with respect to z, after a factor of $y-1$ has been removed. Its discriminant is

$$4096x^8(331776x^{48}+\cdots-889053138892628678423391 39x^{15}-\cdots-6096743321322720854016).$$

Fortunately, this polynomial factors substantially, and can be written as $x^8p_4^2p_4p_6p_{10}^3$, where p_i stands for a polynomial of degree i (not necessarily the same one each time). p_{10} is

$$x^{10}+36x^9+594x^8+5400x^7+27621x^6+75681x^5+120933x^4+63693x^3+2916x^2-58320x-46656,$$

which has a root in $(\frac{-5}{4},\frac{-9}{8})$. Performing the extension of this root to the cylinder of $\mathbf{R}^2$ lying above it took over 6 hours on a VAX 11/780.

7. C.A.D. for Several Polynomials

Let us now generalise the previous section to calculate a cellular algebraic decomposition for the case of m polynomials, i.e. $F = \{f_1, \ldots, f_m\}$. Let us suppose that each f_i has integral coefficients, and also initially that it is square-free. We also suppose that the f_i are relatively prime. Write n for the maximum degree of any of the $f - i$ in either of its variables, and d for the maximum of the lengths of the $\|f_i\|_2$.

Clearly a sign-invariant c.a.d. for F has to be a sign-invariant c.a.d. for each of the f_i separately, and so proposition 2 of the previous section will need to be invoked. Write $\mathrm{disc}'_y(f)$ for either $\mathrm{disc}_y(f)$ or $\mathrm{lc}_y(f)$, depending on the degree of f.

Proposition 1. *A c.a.d. of $\mathbf{R}^2$ sign-invariant for such an F can be obtained by taking a decomposition of $\mathbf{R}^1$ induced by all the $\mathrm{disc}'_y(f_i)$ and $\mathrm{res}_y(f_i, f_j)$, and constructing stacks over each region of this corresponding to all the real branches of the f_i.*

Proof. Each $\mathrm{disc}'_y(f_i)$ induces enough division points in $\mathbf{R}^1$ to ensure that it is sign-invariant on the 2-regions produced from these points by looking at the real branches of f_i. The problem is that we have also got to divide these stacks according to the f_j. There will be no problem if these divisions are compatible, i.e. if a region that lies between two branches of f_i is further divided by an f_j running across the region. The problem comes if the two divisions are incompatible, i.e. if an f_j branch tries to cut an f_i branch. But, at such a cutting point, the corresponding resultant is zero, and so there will be a zero-dimensional region of $\mathbf{R}^1$, the stack above which will include the crossing point of f_i and f_j.

Again, we have a three-phase algorithm: project into $\mathbf{R}^1$, decompose $\mathbf{R}^1$, and extend this to $\mathbf{R}^2$. In the decomposition phase, we have to compute m discriminants, and $m(m-1)/2$ resultants. As in the previous section, this will take time $O(m^2(n^5 d + n^4 d^2))$. Each of these polynomials will have degree at most $2n^2$, and coefficients of $O(n(d + \log n))$. Isolating the real roots of each of them will take $O(m^2 n^{15}(d + \log n)^3)$ operations classically, or $O(m^2 n^{10}(d + \log n)^2 \log n)$ operations using fast arithmetic. From now on, we will just give the times for "fast" arithemtic: they will be sufficiently depressing.

However, it is not sufficient to isolate the roots separately: we must ensure that their defining intervals do not overlap. Clearly the correct way to do this is to isolate the roots separately, and then refine such intervals as overlap. The minimum separation between any two roots of any two of these polynomials is given by Mahler's inequality (I.5.6) applied to the product, and its logarithm is therefore $O(n^3(d + \log n))$. We notice that the distance to which we need to isolate any individual root is independent of m, which is perhaps a somewhat surprising result. In order to refine the intervals, we need to evaluate the polynomials, and even with fast arithmetic this will take $O(n^7(d + \log n) \log n)$ operations (ignoring the nested logarithmic terms etc.). There are (at most) $m^2 n^2$ roots, each of which needs to be isolated to the precision stated, and the total cost is $O(m^2 n^{12}(d + \log n)^2 \log n)$. We can obtain a different bound by applying proposition I.5.8 to the product of all the polynomials, thus bounding the total amount of "closeness" between all the roots, but this gives $O(m^4 n^{10}(d + \log n)^2 \log n)$: not all that different.

Now let us consider the extension phase. We will restrict ourselves to the problem of extending the zero-dimensional regions (which may well have algebraic sample points), since we saw in the previous section that this is likely to pose the greatest difficulties. Extending the roots *per se* is no more different than before, except that we have m^2 times as many roots, each of which has to be extended to m times as many polynomials. Hence this step costs $O(m^3 n^{16}(d + \log n)^2 \log n)$ operations (using fast methods). The last step is a further refinement, to ensure that the roots of f_i are distinguished from those of f_j. This means isolating each of the $m^3 n^3$ roots to a suitable precision. We are now talking about polynomials with algebraic number coefficients, and, as in the previous section, we can only bound the precision by taking a norm, getting a required logarithm for the inter-root separation of $O(n^5 d)$. Each evaluation will require $O(n^{11} d)$ operations, because our polynomial could have degree $2n^2$ in the algebraic number,which has also to be evaluated to at least the same precision. This gives a total cost of $O(m^3 n^{19} d^2)$.

This cost dominates all the other costs, and also the cost of refining our estimates of the roots of the defining polynomials for the decomposition of $\mathbf{R}^1$ to a precision of $O(n^5 d)$ if this was required during the extension phase.

Proposition 2. *Such a c.a.d. can be computed in time $O(m^3 n^{19}(d + \log n)^2 \log n)$ operations. It contains $O(m^3 n^3)$ cells, of which the defining polynomials for the algebraic co-ordinates are of degree at most $O(n^3)$ and have coefficient length at most $O(n^2 d)$.*

Proof. The timing has just been shown above (except that we have quoted a more accurate result here). The number of cells follows from the fact that each section of the $O(m^2n^2)$ division of $\mathbf{R}^1$ induces at most $2mn+1$ divisions above it. The defining polynomials for the 0-dimensional regions are the most troublesome, since we have to substitute an algebraic number into a polynomial, and then find the roots. To express these directly as polynomials with integer coefficients, we use proposition I.4.4. The statements about the degree and coefficient length of these defining polynomials were proved in the previous section.

In order to discuss general sets of polynomials, we need a further notion.

Definition. *A* **square-free basis** *for a set F of polynomials is a set G of polynomials, each of which is square-free and relatively prime to all the other elements of G, and such that every element of F is a product of powers of the elements of G.*

Such a basis can be computed by repeated square-free decomposition and the taking of greatest common divisors. We will not discuss the computing time of this operation, since it is certainly dominated by the enormous computing times we have just been talking about. Remark, though, that a square-free basis for a single polynomial of degree n may contain $O(\sqrt{n})$ polynomials, for consider $(x-1)(x-2)^2\ldots(x-k)^k$. However, the total degree of a square-free basis is at most that of the original set. Since m only enters into the previous computations via mn, as a measure of total degree, we can say that it will not hurt us to compute a square-free basis first, even if it is not strictly necessary.

Proposition 3. *An order-invariant c.a.d. for a set F of polynomials can be obtained by extending the decomposition of R_1 induced by $\mathrm{lc}_y(f_i)$ and $\mathrm{disc}_y(g_i)$, $\mathrm{res}_y(g_i,g_j)$ where the g_i belong to a square-free basis for F. Such a c.a.d. satisfies the previous proposition.*

Proof. This is obvious from the definition of a square-free basis. The c.a.d. is certainly sign-invariant, and the order of a polynomial p can only change at a root of dp/dy or when two components of p of different multiplicity meet, and this is taken care of by the square-free basis and the inclusion of the leading coefficients.

8. Complex Roots.

Let us just remark that a single polynomial in a complex variable is equivalent to two polynomials in two real variables. This means that the technology we have just given is, in principle, sufficient to tell us all about the behaviour of a complex-valued polynomial. There are some special remarks that can be made in this case.

For example, there are precisely n real roots, rather than the $2n^2$ roots that the general theory allows. The resultant of the real and imaginary parts will generally have a very special structure, especially if the original equation had real coefficients. Since all the complex roots of the original equation can be expressed with $y=0$ or with $x=0$, some special factors corresponding to the original equation can be removed.

In practice this is almost certainly a ridiculous way to do it, since the original problem has such definite structure. Pinkert [1976] gives an algorithm based on Sturm sequences for solving this problem directly. He does not analyse the running time there, but quotes a factor of n more than Collins & Loos do for Sturm sequences for real roots. It is likely that the same improvements as we made in section I.6 will apply to this case, but a detailed analysis has not been done.

III. Gröbner Bases

As part of our general program of handling systems of equations and inequalities in many variables, we now treat the case of equations in an arbitrary number of variables, known as $x_1, \ldots, x_n$ in this chapter. The coefficients will be assumed to come from some field k, which the reader can think of as the rational numbers $\mathbf{Q}$.

The question of interest will be the existence of common solutions to *all* the equations, we will not consider questions of reality (much) or of partial solutions. This question is therefore significantly easier to answer than the questions about c.a.d.s treated in the last chapter. The theory is quite complex, though, and several results are stated without proof.

1. Terminology.

Let F be a finite family of polynomials from $k[x_1, \ldots, x_n]$. The algebraist will say that these polynomials define an *ideal* in $k[x_1, \ldots, x_n]$. The material of this chapter can be treated on a very ideal-theoretic level, but we will attempt a fairly low-brow approach. We will define the word ideal, but it will be sufficient, after this section, to treat the concept as a black box.

Definition. *The* **ideal generated by** F, *denoted* (F), *is the set of all elements of* $k[x_1, \ldots, x_n]$ *which can be written as* $\sum a_i f_i$, *where the* f_i *are elements of* F *and the* a_i *are any members of* $k[x_1, \ldots, x_n]$.

In particular, each element of F belongs to (F). (F) is unchanged if we repeat elements of F, or multiply the elements of F by non-zero constants. We will write (F, f) for the ideal generated by the set $F \cup \{f\}$, and so on. Note that $(F, 0) = (F)$, and that the ideal $(F, 1)$ never has any solutions.

Proposition 1. *At any common zero of the elements of* F, *every element of* (F) *is zero.*

Proposition 2. *If* $(F) = (G)$, *then* F *and* G *have the same set of common zeros.*

Proof. At a common zero of all the elements of F, every element of (F) is zero. But, since $(F) = (G)$, this means that every element of (G), and hence every element of G, is zero. The converse argument shows that every element of F is zero at a common zero of the elements of G.

This proposition is of not much direct use, since the ideals are infinite sets, and we have no means of constructing them, let alone testing their equality. We should note that this proposition is only true in one direction. Even with one variable, the converse is false. For example, $(x) \neq (x^2)$, since x does not belong to (x^2). but x and x^2 have the same zeros, though with different multiplicity.

As this example shows, ideals provide a finer classification of sets of polynomials than the common zeros do. Nevertheless, the difference is not very important from the point of view of the elementary theory of the solution of equations (though it is extremely important from other points of view). The reader who wishes to think of (F) as "the set of common zeros of F" will not suffer. There is one anomaly, that can cause difficulties of terminology. If we add a new polynomial to F, the ideal increases (or, at least, does not decrease), while the set of common zeros decreases (or, at least, does not increase).

In one variable $(p_1, \ldots, p_n)$ is equal to (g) where g is the gcd of the p_i. Hence the use of () to denote both ideals and greatest common divisors is not ambiguous *for one variable*. In general there is a difference. For example, x and y have a gcd of 1, but $(x, y) \neq (1)$, since the latter has no zeros, which the former has a zero at $x = 0$, $y = 0$.

We can write elements of $k[x_1, \ldots, x_n]$ as $\sum a_{i_1 \ldots 1_n} x_1^{i_1} \ldots x_n^{i_n}$, where the products $x_1^{i_1} \ldots x_n^{i_n}$ are the *monomials* of $k[x_1, \ldots, x_n]$. We will assume that monomials have non-zero coefficients, so that 0 is a sum of no monomials, and other elements of k are coefficients of the monomial $x_1^0 \ldots x_n^0$. We will need to define an ordering on the monomials, which we do as follows.

Definition. *We say that the monomial* $x_1^{i_1} \ldots x_n^{i_n}$ **is more important than** $x_1^{j_1} \ldots x_n^{j_n}$, *written in symbols* $x_1^{i_1} \ldots x_n^{i_n} > x_1^{j_1} \ldots x_n^{j_n}$, *if there is an integer* k $(1 \leq k \leq n)$ *such that* $i_k > j_k$ *and that, for* $l < k$, $i_l = j_l$. *We define other relations such as* $<$ *and* $\geq$ *similarly.*

This means that, at the first exponent at which they differ, the exponent in $x_1^{i_1} \ldots x_n^{i_n}$ is greater than that in $x_1^{j_1} \ldots x_n^{j_n}$. This is known as **lexicographic** order on the monomials, since it corresponds to the order in which words are placed in a dictionary (if we regard a as greater than —). The theory to be

described in this chapter is also valid for other orderings, but in the interest of simplicity we will stick to this ordering. *The ordering depends on the numbering of the variables.* REDUCE* generally prints the monomials of polynomials in a lexicographic order, and there is a statement `ORDER` which lets one change the way the variables are numbered from the point of view of defining the lexical order.

Proposition 3. *If* $x_1^{j_1} \dots x_n^{j_n}$ *divides* $x_1^{i_1} \dots x_n^{i_n}$, *then* $x_1^{j_1} \dots x_n^{j_n} \leq x_1^{i_1} \dots x_n^{i_n}$.

With more than one variable, the converse is false. Indeed, this is one reaon why the theory of polynomial equations in more than one variable is so different. We will write $x_1^{j_1} \dots x_n^{j_n} \mid x_1^{i_1} \dots x_n^{i_n}$ to indicate that $x_1^{j_1} \dots x_n^{j_n}$ divides $x_1^{i_1} \dots x_n^{i_n}$.

Definition. *The* **head term** *of a polynomial is that monomial with non-zero coefficient that comes first in the importance order.*

2. Operations and Ideals.

Clearly $(F) = (F, f_1 + f_2)$, where the f_i are the elements of F. This is true in much greater generality.

Proposition 1. (F) *is unchanged if we add to* F *any linear combination of its elements (with coefficients from* $k[x_1, \dots, x_n]$*).*

Proof. Let $g = \sum a_i f_i$ be the linear combination. Every element of (F) is in (F, g), since a linear combination of elements of F is still valid as a linear combination of elements of a larger set. An element of (F, g) can be expressed as a linear combination of elements of F by making use of the defining relation for g, so the converse is also true, and the two ideals are equal.

More naïvely, adding this new element does not change the set of common zeros, since adding an element can only decrease the number of common zeros, and this new one is zero if all the old elements are. This does not prove that the ideals are equal, but it satisfies the elementary point of view.

Corollary 1. (F) *is unchanged if we delete from* F *any element that is a linear combination of the other elements (with coefficients from* $k[x_1, \dots, x_n]$*).*

Proposition 2. *Given any two polynomials* p *and* q *in a variable* y *with coefficients from an integral domain* R, *their resultant* $r = \mathrm{res}_y(p, q)$ *is a linear combination of* p *and* q *with coefficients from* R.

Proof. Clearly the proposition is trivial if the resultant is zero. Let S be the Sylvester's matrix of p and q, and consider the system of linear equations over R given in matrix form as $v.S = (0 \dots 01)$. Since the resultant is non-zero, we can solve this system over the field of fractions of R by multiplying by the inverse of S, and this gives a solution for v, whose elements are fractions with denominator at most the determinant of S. Regarding v as the coefficients of two polynomials f and g (as in proposition I.3.2), we see that $fp + gq = 1$. If we multiply this by r, all the fractions from f and g are cleared, since their denominator was at most r, the determinant of S, and we get a solution, in R, to the problem $\hat{f}p + \hat{g}q = r$.

Corollary 2. *If* $f, g \in (F)$, *then for any* i, $\mathrm{res}_{x_i}(f, g) \in (f)$.

Proposition 3. (F) *is unchanged if we replace an element* g *of* F *by* $g + h$, *where* h *is a linear combination of the remaining elements (with coefficients from* $k[x_1, \dots, x_n]$*).*

Proof. Write G for $F \backslash \{g\}$. Then we have to prove that $(G, g) = (G, g + h)$. But $(G, g) = (G, g, g + h)$ by proposition 1, and the corollary implies that $(G, g, g + h) = (G, g + h)$.

Of course, we can apply these transformations to F as much as we want. What do we want to achieve with them, and how do we know when we have finished? One particular use of them stems from the following concept.

---

* [Added in the Bath reprint.] REDUCE version 3.3 incorporates a Gröbner-base package due to R. Gebauer, A.C. Hearn and M. Möller. This package takes the ordered list of variables as one of the parameters to the `groebner` function.

Definition. *If one of the monomials of f, say $x_1^{i_1}\dots x_n^{i_n}$, is divisible by any of the head terms of the elements of F, say $x_1^{j_1}\dots x_n^{j_n}$ the head term of f_j, we can perform a* **reduction** *by replacing f by $\hat{f} = f - c(x_1^{i_1-j_1}\dots x_n^{i_n-j_n})f_j$, where c is the ratio of the coefficients of $x_1^{i_1}\dots x_n^{i_n}$ and $x_1^{j_1}\dots x_n^{j_n}$. We write $f \mapsto_F \hat{f}$.*

What we are doing is using the element f_j of F to eliminate the $x_1^{i_1}\dots x_n^{i_n}$ term in f, and replace it with other terms. These other terms will all be less important than the term we have eliminated, and hence it is natural to call this operation a reduction.

Corollary 3. *If $f \mapsto_F \hat{f}$, then $(F, f) = (F, \hat{f})$.*

This corollary corresponds to a very familiar concept in algebra. For example, if I give you the equations $x^2 + y^2 = 1$ and $x^2 = y$, you will naturally *reduce* the first by the second, to obtain $y^2 + y = 1$. In the symbolism we have introduced

$$x^2 + y^2 - 1 \mapsto_{\{x^2-y\}} y^2 + y + 1.$$

Note that $\mapsto_F$ is not a unique concept. For example, if $F = \{x\}$, then $x^2 + x \mapsto_F x^2$, and also $x^2 + x \mapsto_F x$, since we can eliminate either of the x^2 or x terms by subtracting an appropriate multiple of x. Clearly $x \mapsto_{\{x-1,x-2\}} 1$ and $x \mapsto_{\{x-1,x-2\}} 2$, so that in the presence of several polynomials there can be a great deal of choice about the reduction to perform.

3. Reduction and Gröbner-bases.

One of the reasons that reduction is not unique is that there could be many places to reduce a polynomial. Indeed, that is the reason why the first example in the previous section exhibited non-uniqueness. Hence we will typically be interested, not in single reductions, but in chains of reductions. We will need some notation for this. Write $f \mapsto_F^* f$ to mean that there exist $f_0 = f$, f_1, …, $f_n = \hat{f}$ such that

$$f = f_0 \mapsto_F f_1 \mapsto_F \cdots \mapsto_F f_n = \hat{f},$$

i.e. that f maps to $\hat{f}$ after a finite number of reductions by F. The word "finite" is, in fact, unnecessary.

Proposition 1. *There can be no infinite chain of reductions of a polynomial f by a fixed set F.*

Proof. This will proceed via a series of nested inductions.

There can be no infinite chain of reductions of a constant. In fact, a constant has either no reductions (if there is no constant in F), or one reduction, to 0, if there is a constant in F.

There can be no infinite chain of reductions of a polynomial in x_n alone, of degree k. The proof of this is by induction on k, and the previous assertion is the case $k = 0$. For any fixed k, there can not be a chain involving only the terms of degree less than k, for this would be a chain of reductions of a polynomial of degree less than k. Hence the term of degree k must be reduced at some stage. But, once it has been reduced, the remaining polynomial is of degree less than k, and so can only have finitely long chains, by the inductive hypothesis.

There can be no infinite chain of reductions of a polynomial in x_{n-1} and x_n alone, of degree k in x_{n-1}. The proof of this is by induction on k, and the previous assertion is the case $k = 0$. For any fixed k, there can not be a chain involving only the terms of degree less than k in x_{n-1}, whatever their complexity in x_n, for this would be a chain of reductions of a polynomial of degree less than k. Hence the terms of degree k must be reduced at some stage. Let the terms of degree k in x_{n-1} be $x_{n-1}^k(a_l x_n^l + \dots)$.

By induction on l, of precisely the same form as the proof of the previous assertion, the term $x_{n-1}^k x_n^l$ has to be reduced at some stage. But, once it has been reduced, the remaining polynomial is of degree less than (k, l), and so can only have finitely long chains, by the inductive hypotheses on k and l.

Similarly, by induction on the number of x_i, repeating the previous argument at each stage, we prove the final result.

We write $f \mapsto_F^\dagger \hat{f}$ to indicate both that $f \mapsto_F^* \hat{f}$ and that no further reductions of $\hat{f}$ are possible with F, i.e. $\hat{f}$ is **reduced** with respect to F. The previous result implies that there always is a $\hat{f}$ with this property, and the second example at the end of the previous section shows that there can be more than one.

Corollary 1 *(to proposition 2.3).* *If $f \mapsto^{\dagger}_{F} \hat{f}$, then $(F, f) = (F, \hat{f})$.*

Given a set of equations (F, f), we would naturally try to reduce it. Hence we would replace f by $\hat{f}$, where $f \mapsto^{\dagger}_{F} \hat{f}$, to obtain $(F, \hat{f})$. This process may change the head term of f, and it may now be possible (even if it was not possible to reduce with respect to f) to reduce some of the elements of F with respect to $\hat{f}$.

Proposition 2. *Given a finite set F of polynomials, after a finite number of operations of replacing each element f of F by $\hat{f}$ where $f \mapsto^{\dagger}_{F\setminus\{f\}} \hat{f}$, we arrive at a set where each element is reduced with respect to all the others.*

The proof of this proposition is similar to proposition 1, but rather more involved. Indeed, one really needs the general technology of *well-ordering* to express this proof clearly.

Such a set is said to be **auto-reduced**. By repeated use of the corollary above, it determines the same ideal as the previous one. Note that the reductions have to be done in series, not in parallel. Just because $f \mapsto^{\dagger}_{\{g\}} \hat{f}$ and $g \mapsto^{\dagger}_{\{f\}} \hat{g}$ does not mean that $(f, g) = (\hat{f}, \hat{g})$. The reduction of f may mean that the reductions of g can no longer be carried out.

One might hope that such a set, in which all the elements were reduced, had the property that $\mapsto^{\dagger}_{F}$ was uniquely defined. Alas, this is not so if there is more than one variable, as the following example shows. Let $F = \{xy^2 - 1, x^2y - 1\}$, which set certainly has this property, and consider $x^2y^2 \mapsto^{\dagger}_{F}$. We can reduce with respect to the first element, and get x, or reduce it with respect to the second, and get y. Both these are certainly reduced with respect to F, and are not equal.

Definition. *A set F is said to be a* **Gröbner bases** *(also called* **standard basis***) for the ideal (F) if $\mapsto^{\dagger}_{F}$ is a uniquely defined operation.*

Let us emphasise here that this definition is dependent on the definition of $\mapsto$, and hence on the definition of "head term", and so on the definition of $>$ for monomials. Changing the order of the variables, and so changing $>$, will affect whether or not a set is a Gröbner basis.

Corollary 2. *If F is a Gröbner basis, then $g \in (F)$ if, and only if, $g \mapsto^{\dagger}_{F} 0$.*

The following result is not directly used in what follows, but is useful in other applications.

Proposition 3. *For a fixed ordering, an ideal has a unique auto-reduced Gröbner basis, up to multiplication of the elements of the basis by non-zero members of k.*

4. Buchberger's Algorithm.

So far, we have defined a Gröbner basis, but not given any method of testing whether a set is a Gröbner basis or not, or any method of constructing Gröbner bases. Before introducing Buchberger's algorithm, which provides a method of constructing a Gröbner basis for any ideal (F), let us ask ourselves some questions.

How can a set fail to be a Gröbner basis? There must exist a f with $f \mapsto_F^\dagger \hat{g}$ and $f \mapsto_F^\dagger \hat{h}$, where $\hat{g} \neq \hat{h}$. Suppose the chains that lead to these two reduced values are

$$f = g_0 \mapsto_F g_1 \mapsto_F \cdots \mapsto_F g_n = \hat{g},$$

and

$$f = h_0 \mapsto_F h_1 \mapsto_F \cdots \mapsto_F h_m = \hat{h},$$

where n and m are not necessarily equal. If $g_1 = h_1$, then we have found a smaller (in the sense of $>$) polynomial than f, which has non-unique reductions. If there is some g_i such that $g_i \mapsto_F^\dagger \hat{h}$ is possible, then again this is a smaller example, and similarly with $h_j \mapsto_F^\dagger \hat{g}$. The most obvious minimal example is to take a monomial $x_1^{i_1} \ldots x_n^{i_n}$ which is a multiple of the had-terms of two different elements of F, and reduce it by both of them. In fact, it may as well be the least common multiple of their head-terms. The difference between these two reductions is a measure of how much we are failing to have unique reductions. This motivates the following definition.

Definition. *Given two polynomial f and g, with head terms $f_0 x_1^{i_1} \ldots x_n^{i_n}$ and $g_0 x_1^{j_1} \ldots x_n^{j_n}$, their* **S-polynomial**, *$S(f,g)$ is defined to be $g_0(x_1^{k_1-j_1} \ldots x_n^{k_n-j_n})f - f_0(x_1^{k_1-i_1} \ldots x_n^{k_n-i_n})g$, where $x_1^{k_1} \ldots x_n^{k_n}$ is the l.c.m. of the head terms of f and g, so that $k_l = \max(i_l, j_l)$.*

Up to constant multiples, this is the difference between reducing $x_1^{k_1} \ldots x_n^{k_n}$ via f and reducing it via g. $S(f,f) = 0$ for any element f. The following result, which we do not prove, gives a constructive test for Gröbner bases in terms of S-polynomials.

Corollary (*of proposition 2.1*). *$S(f,g) \in (f,g)$.*

Proposition 1 [Buchberger, 1970]. *A set F is a Gröbner basis for its ideal (F) if, and only if, $S(f,g) \mapsto_F^\dagger 0$ for all pairs (f,g) of elements of F.*

This proposition says much more than it appears to. Remember that $\mapsto_F^\dagger$ is not in general a well-defined concept, so that we are actually saying that, if there is any way of reducing all the S-polynomials to 0, then they must *always* reduce to 0.

Furthermore, this gives us some idea about what to do if the set is not a Gröbner basis. If it is not, then we have a definite "reason", viz. an S-polynomial that does not reduce to 0. This leads to Buchberger's algorithm for converting F into a Gröbner basis for (F).

while $\exists f_i, f_j \in F$ **such that** $S(f_i, f_j) \mapsto_F^\dagger g \neq 0$
do $F := F \cup \{g\}$

Clearly this will give a Gröbner basis if it terminates, for the condition of the previous proposition will be satisfied. Note that we can not write the algorithm in the form

for $i := 1 : |F|$ **do**
  for $j := i+1 : |F|$ **do**
    $S(f_i, f_j) \mapsto_F^\dagger g$
    if $g \neq 0$
      then $F := F \cup \{g\}$

since F, and its size, are always increasing. It is not sufficient to update $|F|$ on the fly, since the new element added may need to be considered along with some f_i that we thought we had finished with. A

better structure works with a set L of pairs of elements of F, and can be written as

$L := \emptyset$
$n := |F|$
for $i := 1 : n$ **do**
  for $j := i + 1 : n$ **do**
    $L := L \cup \{\{f_i, f_j\}\}$
while $L \neq \emptyset$ **do**
  choose $\{g_1, g_2\} \in L$
  $L := L \backslash \{\{g_1, g_2\}\}$
  $S(g_1, g_2) \mapsto_F^{\dagger} h$
  if $h \neq 0$ **then**
    for $i := 1 : n$ **do**
      $L := L \cup \{\{f_i, h\}\}$
    $n := n + 1$; $f_n := h$

Proposition 2 [Buchberger, 1970]**.** *This algorithm always terminates.*

We will not prove this here: see Buchberger [1970] or his more recent survey articles for a proof*. This is, as it stands, hardly an algorithm, since it is extremely under-determined. Not only is the choice of which pair of elements from F (i.e. which element of L) to take not specified, but there is also a choice of how to reduce the S-polynomial. Much has been written about efficient ways of making these choices, and also ways of determining *a priori* that some S-polynomials will reduce to 0, and so do not need to be constructed.

Also, it is possible, and indeed desirable, to keep the set at least partially auto-reduced on the way. Just how much reduction to do is another issue that must be faced in implementing this algorithm. The complexity of this algorithm is a little-understood area. The worst-case size of a Gröbner basis is doubly exponential in the number of variables, but this seems to be a pathological case. It seems that simple Gröbner bases can be found quite quickly via this method, but complex ones can indeed take a long time.

* [Added in the Bath reprint.] Since this was originally written, M. Giusti introduced the author to the following elegant proof. Consider the ideal $I(F)$, generated by the leading monomials of all the elements of F. Every time we add a new polynomial to F, $I(F)$ increases strictly, since the only way this could fail to happen is that the leading term of the new element would be expressible in terms of existing leading terms, and that would mean that this new element was not reduced. But the polynomials over a field are noetherian, i.e. an increasing chain of ideals has to be finite. Hence we can only add finitely many new elements to F.

5. Applications.

Given any Gröbner basis, or indeed any set, we can write it in **triangular** form, viz. as some polynomials in $k[x_1, \ldots, x_n]$ which involve x_1, followed by some that do not involve x_1 but do involve x_2, ..., some that involve only x_{n-1} and (possibly) x_n, follwed by at most one that involves only x_n. Of course, there may be none in any particular class. We will exclude the case where the auto-reduced Gröbner basis is $\{1\}$, i.e. the ideal is the whole of $k[x_1, \ldots, x_n]$ and there are no common solutions.

Definition. *We say that such a form is* **strongly triangular** *if each variable shows up (to some power) as a head term of some element of the basis.*

This excludes systems such as $x_1x_2 + 1$, $x_2^2 - 1$, since x_1 does not show up directly as a head term. This set is not, though, a Gröbner basis: the corresponding Gröbner basis is $x_1 + x_2$, $x_2^2 - 1$, which is strongly triangular.

Proposition. *A set of equations has finitely many common solutions if, and only if, its Gröbner basis is strongly triangular.*

This provides a test for this property, which we will need later. It may or may not be an efficient test, depending on the running time of the Gröbner basis algorithm.

In general, it is possible to say that a Gröbner basis is, in some sense, the "simplest" representation of an ideal. Once we know a Gröbner basis for an ideal, it is an easy task to reduce any polynomials to their smallest (in the sense of $>$) equivalent polynomials modulo that ideal.

IV. C.A.D. in Many Dimensions

As part of our general program of handling systems of equations and inequalities in many variables, we now treat the case of cylindrical algebraic decompositions for a system of polynomials in an arbitrary number of variables, known as $x_1, \ldots, x_r$ in this chapter. We will sometimes write x for x_r when the other variables are irrelevant, and $\underline{x}$ for the set of variables $x_1, \ldots, x_{r-1}$.

The basic theory comes from Collins [1975] as refined by Arnon *et al.* [1984] and McCallum [1985]. Some of the details are our own. While the general principles are a relatively straight-forward generalisation of chapter II, some of the underlying theory is quite complex and several results are stated without proof. We will work generally with order-invariant decompositions, since these seem to have a much better inductive structure than sign-invariant decompositions.

1. Initial Definitions.

The question of vanishing leading coefficients, and indeed vanishing polynomials in general, that troubled us somewhat in chapter II, will prove much more troublesome in the more general setting that we are now considering. We will first give a number of definitions of various sets which will be important in the algorithm, and which will enable us, in the next section, to state a constraint which means, more or less, that "vanishing of polynomials doesn't matter".

Definition. *Let A be a finite set of polynomials in a variable x. Define the* **coefficient set** *of A, $coeff(A)$, to be the set of all the coefficients of the elements of A. Define the* **discriminant set** *of A, $disc(A)$, to be the set of all discriminants of elements of A (of degree at least 2). Define the* **resultant set** *of A, $res(A)$, to be the set of all resultants of pairs of distinct elements of A (of degree at least 1).*

As a general convention, italic letters, as above, will denote **sets** of polynomials, while the corresponding roman letters will denote operations forming individual polynomials.

Definition. *If p is any polynomial in x, then the* **content** *of p, $\mathrm{cont}(p)$, is defined to be the gcd of all the coefficients of p. The* **primitive part** *of p, $\mathrm{pp}(p)$, is $p/\mathrm{cont}(p)$.*

A polynomial is said to be **primitive** if its content is 1. Clearly a primitive part is primitive. Non-primitive polynomials are a technical embarrassmment to us, since they vanish at roots of their contents, independent of x.

Definition. *If B is a set of square-free relatively prime primitive polynomials in x, then the* **projection** *of B, $P(B)$ is defined by*

$$P(B) = coeff(B) \cup disc(B) \cup res(B).$$

Definition. *If A is any set of polynomials in x, then the* **projection** *of A, $P(A)$ is defined by*

$$P(A) = cont(A) \cup P(B),$$

where $cont(A)$ is the set of contents of elements of A, and B is a relatively prime square-free basis for the primitive parts of the elements of A.

$P(A)$ is going to be, under favourable circumstances, the appropriate generalisation of the sets of polynomials that we used in chapter II to define c.a.d.s via projection into $\mathbf{R}^1$. Note that the definition of P is not precise if A is not already a relatively prime square-free basis. This looseness does not matter, since the only effect of using a finer basis is to factor more of the resultants and discriminants that emerge. In practice it seems to be advisable to factor the polynomials as much as possible, but it is sufficient to use the coarse square-free basis that is generated by repeated gcd calculations.

2. Well-ordered polynomials.

The following definition is taken from McCallum [1985], and is crucial to his improvements to the c.a.d. algorithm.

Definition. *A set A of polynomials in $x_1, \ldots, x_r$ is said to be* **well-ordered** *if the following two conditions hold:*

a) The primitive part of each element of A is identically zero (as a polynomial in x_r) at only finitely many $(r-1)$-tuples $(x_1, \ldots, x_{r-1})$.

b) $P(A)$ (as defined in the previous section) is a well-ordered set of polynomials in $x_1, \ldots, x_{r-1}$.

This definition is actually a slight mis-nomer. It is really not the polynomials, but the choice of axes implicit in the variables, which might not be well-oriented. Proposition 3 below says, essentially, that we can rotate the axes to avoid this.

We note that the Gröbner-basis methods of the previous chapter will let us tell whether or not a set of polynomials in $(r-1)$ variables, the coefficients of a polynomial in r variables, have finitely or infinitely many solutions. Of course, in many special cases, such as the existence of a constant coefficient, it is obvious that there are at most finitely many points at which it vanishes.

Proposition 1. *If $r \leq 3$, then every set of non-zero polynomials is well-ordered.*

Proof. If $r = 1$, then there are no tuples of the x-coordinates preceding x_r, and the definition is vacuously true. If $r = 2$, then suppose that a primitive non-zero p was identically zero (as a polynomial in x_2) at a particular value a of x_1. Then each coefficient (a polynomial in x_1) would have a root at a, and so would be divisible by $(x - a)$. But the polynomial was assumed primitive, i.e. without any gcd of the coefficients, and we have just constructed such a gcd.

If $r = 3$, then the situation is more complex. A primitive polynomial can certainly vanish identically at a particular pair (x_1, x_2). For example, $x_1 x_3 + x_2$ is primitive, and vanishes identically at $(0, 0)$. But a set of polynomials in two variables can only have finitely many common zeros unless they actually have a common factor (indeed, the number of zeros is bounded by $2n^2$, where n is the total degree, by the work of chapter II).

Proposition 2 [McCallum, 1985]**.** *If A is a well-oriented set of polynomials, then an algebraic decomposition of $\mathbf{R}^{r-1}$ which is order-invariant for $P(A)$ can be extended cylindrically to a decomposition of $\mathbf{R}^r$ which is order-invariant for A.*

We do not propose to prove this theorem here, which seems to rely on some fairly recent and deep results of Zariski in analytic geometry. Of course, this result *per se* does not help us when A is not well-oriented. One solution, which was the method adopted by Collins [1975] before the discovery of proposition 2, was to augment A with all its reducta with respect to x at every stage of the projection. This unfortunately greatly increases the size of the set $P(A)$, and has very bad consequences for the complexity, both theoretically and practically.

McCallum points out that we can add all the partial derivatives of all orders for any $f \in A$ for which there are an infinite set of points on which f vanishes. This check needs to be performed at every level of the induction, but in practice it will be extremely rare to have to add these extra polynomials often, because, as we shall see, the degrees mount considerably, thus decreasing the probability that the coefficients will all vanish simultaneously.

Proposition 3 [McCallum, 1985]**.** *Given any set A of polynomials, we can construct a linear transformation of space such that the transform of A is well-oriented.*

Again, we will not prove this theorem, though the proof is not very difficult. Indeed, it is possible to ensure even stronger properties (such as monic) of the transformed set. Note that such a transformation does not change the number of polynomials, or the total degree (though it may well change the degrees in individual variables). It is somewhat harder to say what it does to the coefficients, but the blow-up is certainly at worst polynomial. Based on this result, we shall restrict our attention to well-ordered sets of polynomials, even though non well-ordered sets may cause problems in practice.

3. Complexity of Projection.

In this section, we shall look at the complexity of the projection operation and the sizes of data that it induces. In chapter II, we used m to stand for the number of polynomials and n to stand for the degree, and then waved our hands somewhat to show that the increase in m caused by square-free decomposition and the search for common factors did not matter. Here we will formalise that hand-waving, using a conept due to McCallum [1985].

Definition. *A set of polynomials* **has the** (m,n) **property** *if it can be partitioned into at most m sets such that the product of the polynomials in each set has degree at most n in any variable.*

Proposition 1. *If A has the (m,n) property, then it has the (m',n') property for any $m' \geq m$ and $n' \geq n$.*

Proposition 2. *If A has the (m,n) property, then $cont(A)$ and $pp(A)$ (the set of primitive parts of A) both have the (m,n) property.*

Proposition 3. *If A has the $(m_1 m_2, n)$ property, then it also has the $(m_1, m_2 n)$ property.*

Proposition 4. *If A has the (m,n) property, then for any square-free basis B for A, $B \cup cont(A)$ also has the (m,n) property.*

Proof. Let A be $S_1 \cup \cdots S_m$, where the product of the elements of each S_i has degree at most n in any variable. Let T_1 be the set of elements of B which divide any element of S_1, T_2 be the set of elements of $B \backslash T_1$ which divide any element of S_2 and so on. Then the set $T_1, \ldots, T_m$ (some of which may be null) partition B. Since the product of the elements of each T_i divides the product of the elements of the corresponding S_i, we see that the total degree is at most n, so that this decomposition demonstrates that B has the (m,n) property. If we add the contents of the elements of S_i to T_i we obtain a decomposition of $B \cup cont(A)$ with the same property.

Proposition 5. *If A has the (m,n) property, then $P(A)$ has the (m',n') property, where $m' \leq \frac{1}{2}(m+1)^2$ and $n' \leq 2n^2$. Alternatively, $m' \leq m^2$ for $m \geq 2$.*

Proof. Recall that

$$P(A) = cont(A) \cup coeff(B) \cup disc(B) \cup res(B),$$

where B is any square-free basis for A. Let $B \cup cont(A)$ be (m,n) decomposed as the union of the T_i, as in the previous proof. For any particular i, let $T_i = \{f_1, f_2, \ldots, f_l\}$, and write $F = \prod_{j=1}^{l} f_j$. F has degree at most n in any variable, and so $\mathrm{res}(F, F/')$ has degree at most $2n^2$. But this resultant is divisible by the leading coefficient of F (and *a fortiori* by the leading coefficient of every f_j) by propositon I.3.5, and also by the discriminamt of F, which by proposition I.3.6 is divisible by every $\mathrm{disc}(f_j)$ and $\mathrm{res}(f_j, f_{j'})$. So all these polynomials form a single set whose product has degree at most $2n^2$.

If we take these m sets, we have accounted for all the elements of $cont(A)$, all the elements of $disc(B)$, the leading coefficients that made up $coeff(B)$ and the elements of $res(B)$ defined by pairs of polynomials coming from the same T_i. The set of non-leading coefficients clearly has the (mn,n), since a set of polynomials of total degree at most n has at most n non-leading coefficients. Hence it has the (m,n^2), and so the $(\lceil m/2 \rceil, 2n^2)$ property.

So we have only the additional resultants to account for. By repeated use of proposition I.3.7, expressing the resultant of products in terms of the product of resultants, we see that

$$\prod_{f_\alpha \in T_i,\, g_\beta \in T_j} \mathrm{res}(f_\alpha, g_\beta) = \mathrm{res}\Big(\prod_{f_\alpha \in T_i} f_\alpha, \prod_{g_\beta \in T_j} g_\beta\Big).$$

Hence the set of all these resultants (for a fixed choice of i and j) has the $(1, 2n^2)$ property, and hence the entire set of cross-resultants has the $(\frac{1}{2}m(m-1), 2n^2)$ property.

Adding up the three sets defined in the previous three paragraphs, we see that the whole of $P(A)$ has the $(\lceil \frac{1}{2}m(m+2) \rceil, 2n^2)$ property. The stated result then follows.

It is not intended that this method of multiplying polynomials together and then taking resultants actually be used: it is merely a device for proving the necessary degree bounds. Note that this result is actually somewhat better than that obtained by McCallum (his equation 6.1.1) since we replaced a (mn,n) partition by a $(\lceil m/2 \rceil, 2n^2)$ one.

Proposition 6. *Suppose no coefficient in A has more than d digits. Then no coefficient in $P(A)$ has more than $7rn^2 + 2nd$ digits.*

Proof. The elements of B and $cont(A)$ are factors of the elements of A. Hence, by a theorem due to Gelfond [1960], the lengths of their coefficients are bounded by $r \log n + rn + d$, and so the length of their sum is bounded by $2r \log n + rn + d$. Call this d'. The resultant of any two elements of B has coefficients which are therefore, by a generalization of Hadamard's inequality [Collins & Horowitz, 1974], bounded by $n \log n + 2nd'$. We can replace $\log n$ by n, and deduce that this is bounded by $7rn^2 + 2nd$.

This is certainly an unrealistic bound in practice. We would not expect any of the factors to have larger coefficients than the original, so that $d' = r \log n + d$, and then the bound would be $3nr \log n + 2nd$, with a substantially better behaviour as a function of n.

We shall normally be concerned with repeated projections, from r variables down to one. Hence let A_1 be A, and in general let $A_{k+1} = P(A_k)$, so that A_k is a set of polynomials in $r + 1 - k$ variables. Suppose that A_i has the (m_i, n_i) property, with maximum coefficient length d_i.

Proposition 7. *If A has the (m, n) property and maximum coefficient length d, then: $m_k \leq m^{2^{k-1}}$ if $m > 1$, otherwise $m_k \leq 2^{2^{k-2}}$; $n_k \leq \frac{1}{2}(2n)^{2^{k-1}}$; $d_k \leq r(2n)^{2^k} d$.*

Proof. The inequalities for n_k and m_k (general case) follow from proposition 5 by induction. The case when $m = 1$ gives $m_2 = 2$, and then the induction is straight-forward. For d_k, we note that the stated inequality is true when $k = 1$ (since $d \leq r(2n)^2 d$) and $k = 2$ (since $7rn^2 + 2nd \leq r(2n)^4 d$). To prove the inequality in general, we use induction.

$$\begin{aligned} d_{k+1} &\leq 7(r + 1 - k){n_k}^2 + 2n_k d_k \\ &\leq 7r\left(\frac{1}{2}(2n)^{2^{k-1}}\right)^2 + (2n)^{2^{k-1}} r(2n)^{2^k} d \\ &\leq r(2n)^{2^{k+1}} d. \end{aligned}$$

Again, the equation for d_k is very pessimistic, but the inherent nature of it, doubly exponential in n, is not changed by more optimistic assumptions. These equations are, again, somewhat better than McCallum's, reflecting the improvement in proposition 5. Essentially, his exponents are of the form $k2^k$ rather than 2^k. From now on, we will ignore the special case $m = 1$, leaving it to the reader to work out the special formulae for this case, or to replace m by $\min(m, 2)$.

When it comes to the total time for the projection phase, this is dominated by the time for the last projection, since here we have the most polynomials, of the largest degree and longest coefficients. As mentioned in section II.6, the time for taking the resultants and discriminants in this operation will be $m_{r-1}^2 \left(n_{r-1}^5 d_{r-1} + n_{r-1}^4 d_{r-1}^2\right)$, which simplifies to $\frac{3}{32} m^{2^{r-1}} (2n)^{2^{r+1}} r^2 d^2$. This time also dominates the time required for square-free and relatively prime basis calculations. In practice, it is generally cheaper to do these via factorisation, although the asymptotic time for this is somewhat worse.

4. The Extension Phase.

Between the projection phase and the extension phase , we ought to consider the base phase, that of decomposing $\mathbf{R}^1$. Here the methodology is exactly that of section II.7, and the time complexity (using fast operations) given there is $m_{r-1}^2 n_{r-1}^{12} \left(d_{r-1} + \log n_{r-1}\right)^2 \log n_{r-1}$. We will ignore the terms in logarithms, and substituting the values from propositions 3.7 gives $\frac{1}{4096} m^{2^{r-1}} (2n)^{2^{r+2}} r^2 d^2$.

We now have to consider the various extension phases, from $\mathbf{R}^i$ to $\mathbf{R}^{i+1}$. Let c_i be the number of regions in the decomposition of $\mathbf{R}^i$.

Proposition. $c_i \leq 2(2mn)^{2^r}$.

Proof. c_1 is the number of regions of $\mathbf{R}^1$, and is determined by the roots of a set of polynomials with the (m_r, n_r) property, which therefore have at most $m_r n_r$ roots between them. So $c_1 \leq 2m_r n_r + 1 \leq (2mn)^{2^{r-1}} + 1$. In general, above each region of $\mathbf{R}^{i-1}$ we erect $2N+1$ regions in $\mathbf{R}^i$, where N is the total number of roots of the polynomials in A_{r+1-i} above the sample point, which is therefore bounded by $m_{r+1-i}n_{r+1-i}$. Hence $c_i \leq \left((2mn)^{2^{r-i}} + 1\right) c_{i-1}$. So by induction

$$\begin{aligned}
c_i &\leq \prod_{j=r-i}^{r-1} \left((2mn)^{2^j} + 1\right) \\
&\leq \prod_{j=r-i}^{r-1} (2mn)^{2^j} \prod_{j=r-i}^{r-1} \left((2mn)^{-2^j} + 1\right) \\
&\leq (2mn)^{\sum_{j=r-i}^{r-1} 2^j} \prod_{j=r-i}^{r-1} \frac{1}{1-(2mn)^{-2^j}} \\
&\leq (2mn)^{2^r} \frac{1}{1-\sum_{j=r-i}^{r-1}(2mn)^{-2^j}} \\
&\leq (2mn)^{2^r} \frac{1}{1-mn},
\end{aligned}$$

and the result follows since we assume that $m > 1$.

For each region in the decomposition of $\mathbf{R}^i$, we have a sample point that has some algebraic co-ordinates and some rational co-ordinates. For the sake of simplicity in analysis, we will assume that all the co-ordinates for a given sample point are represented in terms of one algebraic number (the **primitive element** representation), whether or not this is done in practice (see section I.7). Under this assumption, the extension algorithm looks as follows, where we have numbered those steps that might require non-trivial computation.

for all $(\beta_1, \ldots, \beta_i)$ sample points of $\mathbf{R}^i$ **do**
  Assume the β_j are expressed in terms of α
  for all polynomials p in $i+1$ variables **do**
[1] Substitute the β_j into p
[2] Reduce the result by the defining equation for α
[3] Isolate real roots of this equation
  for each root β_{i+1} **do**
[4] Calculate $\bar{\alpha}$ as a primitive element defined by $\bar{p}$
[5] Express α in terms of $\bar{\alpha}$
[6–7] **for** $1 \leq j \leq i$ **do**
[6] Hence express β_j in terms of $\bar{\alpha}$
[7] Reduce these equations modulo $\bar{p}$
[8] Express β_{i+1} in terms of $\bar{\alpha}$
  Define the $(i+1)$-regions accordingly.
[9a] Refine the isolating intervals for the β_{i+1}
[9b] Separate the isolating intervals
  Define the intermediate regions by choosing simple rational β_{i+1}

Steps [9a] and [9b] are alternatives: we can either refine the intervals defining our points so much that we know that no other roots can be close to them, or we can ensure that the intervals do not overlap by refining only such intervals as do. This is discussed in section II.7 for the two-dimensional case. In practice, of course, one would use step [9b], but step [9a] is easier to analyse. It will turn out (proposition 6.8 below) that this is the dominating step for the theoretical analysis.

5. Data Sizes During Extension.

In this section, we will discuss the sizes of the data in the above algorithm, as we extend a c.a.d. of $\mathbf{R}^i$ to $\mathbf{R}^{i+1}$. Let $\hat{n}_i$ be the maximal degree of any a primitive element for a sample point of $\mathbf{R}^i$, and $\hat{d}_i$ be maximal coefficient length in any defining equation for such a primitive element. Furthermore, let $\check{d}_i$ be the maximal coefficient length (numerator or denomiator) in the equation determining any of the coefficients of a sample point in terms of the primitive element (the degree is automatically bounded by $\hat{n}_i$) and the lengths of all the isolating intervals involved.

For the case $i = 1$, we know from section II.7 that

$$\begin{aligned}
\hat{n}_1 &\leq n_r \leq \frac{1}{2}(2n)^{2^{r-1}} \\
\hat{d}_1 &\leq n_r + d_r \leq r(2n)^{2^r} d \\
\check{d}_1 &\leq O(n_r(\log n_r + d_r)) \leq O(r(2n)^{3\cdot 2^{r-1}} d),
\end{aligned}$$

where we have allowed for the increase in $\hat{d}_1$ since we wish to reduce our defining equations (in particular, to insist that they be square-free). In fact, the constant implied by the O is less than unity, since we have factors like $3^{n_r/2}$ on our side. As in section II.7, none of these quantities depends on m.

Proposition 1. $\hat{n}_i \leq (2n)^{2^r}$

Proof. Let $(\beta_1, \ldots, \beta_i)$ be a sample point for an i-region, and let α be an algebraic number in terms of which we can express these co-ordinates. As in chapter II, we do not actually require that the defining equation for α be irreducible, since, if it does factor for us, we can take advantage of the factorisation when we discover it. For each of the equations p in A_{r-i}, we have, in step [4], to construct an algebraic number $\bar{\alpha}$ in terms of which we can express all the existing coefficients and the β_{i+1} that arise. If ν is the degree of p, the degree of the polynomial defining $\bar{\alpha}$, i.e. $\hat{n}_{i+1}$, is at most $\nu\hat{n}_i$, which is bounded by $n_{r-i}\hat{n}_i$. Hence by induction

$$\hat{n}_i \leq \prod_{j=r-i}^{r-1} n_{j+1} \leq \prod_{j=r-i}^{r-1} (2n)^{2^j} = (2n)^{\sum_{j=r-i}^{r-1} 2^j} < (2n)^{2^r}.$$

This result is based on the same recurrence relation as McCallum [1985] obtains, which is better than that of Collins [1975]. McCallum attributes the improvement to his use of factorisation, but we have not, in fact, used any factorisation. The explanation is simple: $\bar{\alpha}$ is determined in terms of an equation of degree at most $n_{r-i}\hat{n}_i$ over the integers. *In general*, a primitive element for this and α, which is determined by an equation of degree $\hat{n}_i$, would be of degree $n_{r-i}\hat{n}_i^2$, and this is what Collins assumes. But $\bar{\alpha}$ is determined by an equation of degree n_{r-i} over the field containing α, and this is the situation we assumed in our primitive element theorem (proposition II.3.2).

Proposition 2. $\hat{d}_i, \check{d}_i \leq r(2n)^{i2^{r+3}} d$.

Proof. To compute a region in $(i+1)$-dimensional space, we have to decompose the stack above each region in i-dimensional space. Suppose such a region is defined by a sample point whose co-ordinates are expressed as polynomial functions (of degree at most $\hat{n}_i$ and coefficient length $\check{d}_i$) of some algebraic number α. Then substituting these co-ordinates [step 1] into a polynomial of degree n_{r-i} with coefficients of length d_{r-i} will give us a polynomial p, still of degree n_{r-i} in $x - i + 1$, but of degree $in_{r-i}\hat{n}_i$ in α, and with coefficients of length up to $in_{r-i}(\check{d}_i\hat{n}_i + d_{r-i})$. Since α is defined by a polynomial of degree at most $\hat{n}_i$ and with coefficients of length at most $\hat{d}_i$, we have [step 2] to reduce p with respect to the defining polynomial for α. This will increase the coefficient length to $in_{r-i}(\check{d}_i\hat{n}_i + d_{r-i}) + in_{r-i}\hat{n}_i\hat{d}_i$. Call this length e for the moment (but see the corollary below).

Now [step 4] we have to compute the algebraic number $\bar{\alpha}$ in terms of which we will express our sample points for the $(i+1)$-regions. Those regions whose dimension is one more than that of the i-region will have a new rational co-ordinate, and so they are defined in terms of α. Those $(i+1)$-regions whose dimension is the same as that of the i-region need a more complicated algebraic number. This number is computed by taking the norm of the reduced p and the defining equation for α. As in proposition II.3.2, we may need to

perform a substitution $x_{i+1} \mapsto x_{i+1} - \lambda\alpha$, but λ is at most $\hat{n}_{i+1}$, so the effect on the size of the coefficients is negligeable. The resulting polynomial will have degree $n_{r-i}\hat{n}_i$, as already stated in the proof of the previous proposition, and coefficient length $\hat{n}_i e + n_{r-i}\hat{d}_i + (n_{r-i} + \hat{n}_i)\log(n_{r-i} + \hat{n}_i)$. The last two terms in this equation are negligeable with respect to the first term, so we deduce a recurrence relation

$$\hat{d}_{i+1} \leq i n_{r-i}\hat{n}_i^2(\hat{d}_i + \check{d}_i) + i n_{r-i}\hat{n}_i d_{r-i}. \tag{1}$$

What of $\check{d}_{i+1}$? It is the maximum of the lengths appearing in the defining equations for the co-ordinates β_j in terms of $\bar{\alpha}$ and the lengths of the necessary isolating intervals. The coefficients of the polynomials defining α and β_{i+1} in terms of $\bar{\alpha}$ [steps 5 and 7] have the same magnitude as $\hat{d}_{i+1}$, since they are determined via the extended euclidean algorithm from the same data as define $\bar{\alpha}$. This definition of α must be substituted [step 6] into the polynomials defining β_j in terms of α. This gives us polynomials of degree $\hat{n}_i\hat{n}_{i+1}$ and coefficient length $\hat{n}_i\hat{d}_{i+1} + \hat{n}_i \log \hat{n}_i + \check{d}_i$. These polynomials must now [step 7] be reduced modulo $\bar{p}$, the defining equation for $\bar{\alpha}$, which can increase the coefficient length by $\hat{n}_i\hat{n}_{i+1}\hat{d}_{i+1}$.

As regards the root isolation, we must [step 9a] isolate the roots of each polynomial separately, and then refine them to a suitable accuracy. The required accuracy is given by Mahler's inequality applied to the product of the norms of two defining polynomials, as in chapter II. This product has degree $2\hat{n}_{i+1}$, and coefficient length $2\hat{d}_{i+1}\log \hat{n}_{i+1}$, and so the accuracy is $2\hat{n}_{i+1}(2\hat{d}_{i+1}\log \hat{n}_{i+1} + \log 2\hat{n}_{i+1})$. Combining these bounds, we have shown that

$$\check{d}_{i+1} \leq \hat{n}_i\hat{d}_{i+1} + \hat{n}_i \log \hat{n}_i + \check{d}_i + \hat{n}_i\hat{n}_{i+1}\hat{d}_{i+1} + 2\hat{n}_{i+1}(2\hat{d}_{i+1}\log \hat{n}_{i+1} + \log 2\hat{n}_{i+1}). \tag{2}$$

We can use the known values from proposition 2 and the previous section to simplify the recurrence inequalities (1) and (2), yielding

$$\begin{aligned} \hat{d}_{i+1} &\leq i\frac{1}{2}(2n)^{2^{r-i-1}}(2n)^{2^{r+1}}(\hat{d}_{i+1} + \check{d}_{i+1}) + i\frac{1}{2}(2n)^{2^{r-i-1}}(2n)^{2^r} r (2n)^{2^{r-i}} \\ &\leq i(2n)^{2^{r-i-1}+2^{r+1}}(\hat{d}_{i+1} + \check{d}_{i+1}) \end{aligned} \tag{1'}$$

and, since the term we retain dominates all the others,

$$\begin{aligned} \check{d}_{i+1} &\leq \hat{n}_i\hat{n}_{i+1}\hat{d}_{i+1} \\ &\leq 2(2n)^{2^{r+1}}\hat{d}_{i+1}. \end{aligned} \tag{2'}$$

Hence we have a chain of inequalities $\check{d}_{i+1} \geq \hat{d}_{i+1} \geq \check{d}_i \geq \hat{d}_i$, which we can use to simplify the equations further. They become $\hat{d}_{i+1} \leq (2n)^{3\cdot 2^r}\check{d}_i$ and $\check{d}_{i+1} \leq 2(2n)^{2^{r+1}}\hat{d}_{i+1}$, so that $\check{d}_{i+1} \leq 2(2n)^{7\cdot 2^r}\check{d}_i$. Since $\log 2 + 7 \cdot 2^r < 2^{r+3}$, and since the inequality to be proved is satisfied for $\check{d}_1$, the proposition follows by induction.

Corollary 1. *e, the length of the coefficients of the equation produced by step [2], is bounded by*

$$(2n)^{3\cdot 2^r}\check{d}_i \leq r(2n)^{(i+1)2^{r+3}}.$$

Proof. $e = i n_{r-i}(\check{d}_i\hat{n}_i + d_{r-i}) + i n_{r-i}\hat{n}_i\hat{d}_i$. Since $\check{d}_i > \hat{d}_i$ (with a sufficiently large margin to absorb the d_{r-i} term), we can simplify this to $e \leq 2i n_{r-i}\check{d}_i\hat{n}_i$. As in the deduction and simplification of (1') above, $e \leq (2n)^{3\cdot 2^r}\check{d}_i$. Applying the proposition itself, $e \leq r(2n)^{3\cdot 2^r + i2^{r+3}}d \leq r(2n)^{(i+1)2^{r+3}}$.

Corollary 2. *The required accuracy for the β_{i+1} [step 9a] is bounded by $(2n)^{3+r}\hat{n}_{i+1}\hat{d}_{i+1}$.*

Proof. Just before equation (2), we deduced that the accuracy was bounded by $2\hat{n}_{i+1}(2\hat{d}_{i+1}\log \hat{n}_{i+1} + \log 2\hat{n}_{i+1})$. This is dominated by $8\hat{n}_{i+1}\hat{d}_{i+1}\log \hat{n}_{i+1}$, and so by $(\log 2n)2^{3+r}\hat{n}_{i+1}\hat{d}_{i+1}$ (using proposition 1). We can bound this by $(2n)^{3+r}\hat{n}_{i+1}\hat{d}_{i+1}$.

6. Time for Extension.

Now we have to compute the time required for the extension operation, as defined in the algorithm of section 4. We use the convention that t_k denotes the number of operations required for one execution of step $[k]$ (i.e. excluding any effect due to the step being inside a loop).

Proposition 1. $t_1 = O((2n)^{(i+4)2^{r-i-1}} \check{d}_i^2 \hat{n}_i^2)$.

Proof. We know that the end-result of step 1 is a polynomial in x_{i+1} and α, whose maximum coefficients are bounded by $in_{r-i}(\check{d}_i\hat{n}_i + d_{r-i})$. Hence we can replace α by a number N greater than twice this, evaluate the polynomials in α to yield numbers, and then perform the computations. At the end, we re-express all numbers base N, and convert the results to polynomials in α. If we choose N to be a power of two, the conversion will be free [Davenport & Padget, 1985].

N will dominate all the coefficients involved in p, so the time for the Horner's rule calculation can be taken as $n_{r-i}^2 N^2$. We have a total of $1 + n_{r-i} + (n_{r-i}+1)^2 + \cdots + (n_{r-i}+1)^i \leq 2(n_{r-i}+1)^i$ such evaluations to perform. This gives a total running time of $2n_{r-i}^{i+2}N^2 \leq (2n)^{(i+4)2^{r-i-1}} \check{d}_i^2 \hat{n}_i^2$ by proposition 3.7, absorbing the $+1$, i and d_{r-i} terms into the factors of $\frac{1}{2}$ available.

Proposition 2. $t_2 = O(r^3(2n)^{3\cdot 2^{r-i-1}+3\cdot 2^r} \check{d}_i^2)$.

Proof. There are $n_{r-i}+1$ coefficients in the polynomial, each of which requires reduction. A reduction step involves dividing a polynomial of degree bounded by $\nu = in_{r-i}\hat{n}_i$ and coefficients bounded by $M = in_{r-i}(\check{d}_i\hat{n}_i + d_{r-i})$ by a polynomial of degree $\hat{n}_i$ and coefficients $\hat{d}_i$. The total time for such a division is

$$\frac{\hat{d}_i}{6}(3M\nu^2 + 6M\nu\hat{n}_i + \hat{d}_i(\nu^3 + 2\hat{n}_i^3)) \tag{3}$$

by considering a straight-forward operation in which,on each of $\nu - \hat{n}_i$ cycles, the dividend is multiplied by the leading coefficient and a suitable multiple subtracted. Substituting in the values of ν and M gives us $i^3 n_{r-i}^3 \hat{n}_i^3 \hat{d}_i(3\check{d}_i + \hat{d}_i)/6$ plus lesser terms, and the whole expression is dominated by $i^3 n_{r-i}^3 \hat{n}_i^3 \hat{d}_i \check{d}_i$. $\hat{d}_i \leq \check{d}_i$ and $i < r$, so we can reduce this to $r^3 n_{r-i}^3 \hat{n}_i^3 \check{d}_i^2$. We can now apply propositions 3.7 and 5.1 to deduce the final result.

Proposition 3. $t_3 = O(2^r \hat{n}_{i+1}^4 \hat{d}_{i+1}^2)$.

Proof. As remarked in chapter II, it is not clear how to bound the complexity of isolating roots of a polynomial with algebraic number coefficients, but it seems to be cheaper than isolating the roots of the norm of the polynomial. The norm has degree bounded by $\hat{n}_{i+1}$ and length of coefficients bounded by $\hat{d}_{i+1}$. By the best algorithm of section I.6, this will take time $O(2^r \hat{n}_{i+1}^4 \hat{d}_{i+1}^2)$, where we have ignored the logarithmic terms in comparison with $\hat{d}_{i+1}$, and have replaced $\log \hat{n}_{i+1}$ by 2^r, using proposition 5.1.

Proposition 4. $t_4 = O((2n)^{10\cdot 2^r} \check{d}_i^2)$.

Proof. This calculation is that of a resultant of two polynomials in two variables, whose degrees are bounded by $\hat{n}_i$ and whose coefficients are bounded by e. By the result of Collins [1971], this can be done in $O(\hat{n}_i^5 e + \hat{n}_i^4 e^2)$ operations. By Corollary 5.1, $e \leq (2n)^{3\cdot 2^r} \check{d}_i$. Hence e dominates $\hat{n}_i$, and the time is $O(\hat{n}_i^4 e^2)$. Substituting the values from proposition 5.2 and corollary 5.1 give the bound stated.

Corollary 1. *t_5 and t_8 are* $O((2n)^{10\cdot 2^r} \check{d}_i^2)$.

Proof. The expressions required in steps 5 and 8 are by-products of the resultant calculation in step 4, and some gcd calculations whose running time is no worse.

Proposition 5. $t_6 = O((2n)^{4\cdot 2^r} \hat{d}_{i+1}^2)$.

Proof. Step [6] involves substituting the definition of α, a polynomial of degree at most $\hat{n}_{i+1}$ and coefficient length at most $\hat{d}_{i+1}$ into the definition of β_j, which is a polynomial of degree at most $\hat{n}_i$ and coefficient length at most $\check{d}_i$. As in proposition 1, we can do this via numeric evaluation, replacing $\bar{\alpha}$ by $N > 2(\hat{n}_i\hat{d}_{i+1} + \hat{n}_i \log \hat{n}_i + \check{d}_i)$, which we can simplify to $N < 8\hat{n}_i\hat{d}_{i+1}$, where we have also allowed for rounding N up to a power of 2. Hence the total cost is that of Horner's rule, $\hat{n}_i N(\hat{n}_i N + \check{d}_i)$, and we can absorb the $\check{d}_i$ into the rounding up for N. This gives a total cost of $8^2 \hat{n}_i^4 \hat{d}_{i+1}^2$. Applying proposition 5.1 gives the stated result.

Proposition 6. $t_7 = O((2n)^{6\cdot 2^r}\hat{d}_{i+1}^2)$.

Proof. This operation is that of reducing a polynomial, whose running time was given in equation (3) above. Here, this equation gives

$$\frac{\hat{d}_{i+1}}{6}(3M\nu^2 + 6M\nu\hat{n}_{i+1} + \hat{d}_{i+1}(\nu^3 + 2\hat{n}_{i+1}^3)),$$

with $M = 2\hat{n}_i\hat{d}_{i+1}$ and $\nu = \hat{n}_i\hat{n}_{i+1}$. This expression is dominated by the terms in $\hat{d}_{i+1}^2\nu^3$, and their coefficient is less than unity. So we deduce a total bound of $\hat{d}_{i+1}^2\nu^3$, which simplifies to $\hat{d}_{i+1}^2\hat{n}_i^3\hat{n}_{i+1}^3$, Applying proposition 5.1 gives the stated result.

Corollary 2. $t_{6-7} = O(i(2n)^{6\cdot 2^r}\hat{d}_{i+1}^2)$.

Proof. t_7 clearly dominates t_6, and the whole loop is executed i times.

Proposition 7. $t_{9a} = O((2n)^{9+3r}\hat{n}_{i+1}^5\hat{d}_{i+1}^3)$.

Proof. We have to refine each root to an accuracy of s, which is given by Corollary 5.2. Each bisection involves evaluating a polynomial of degree n_{r-i} in x_{i+1} and $\hat{n}_{i+1}$ is $\bar{\alpha}$. The coefficients of this polynomial are negligeable compared with s. The evaluation of the x_{i+1} part will take $n_{r-i}^2 s^2$ operations per coefficient of $\bar{\alpha}$, or $\hat{n}_{i+1}n_{r-i}^2 s^2$ in all. We then have to evaluate this polynomial, of degree $\hat{n}_{i+1}$, at an interval value (bounded by s-digit numbers). To within a constant factor, this will take the same number of operations as evaluating the polynomial at a point value, viz $\hat{n}_{i+1}s(\hat{n}_{i+1}s + n_{r-i}s)$. The second term (resulting from the coefficients of the polynomial) is dominated by the first, so we have $O(\hat{n}_{i+1}^2 s^2)$. This time also dominates the first cost.

We may need to make up to s evaluations, so the total time is $O(\hat{n}_{i+1}^2 s^3)$. The result then follows by applying Corollary 5.2.

Proposition 8. *The total cost of the extension phase is dominated by* $m^{2^r}(2n)^{(24i+30)2^r}r^3d^3$.

Proof. t_{9a} dominates t_{6-7} since $\hat{d}_{i+1}$ dominates $\hat{n}_{i+1}$. t_{9a} can be re-written as $O((2n)^{9+3r+5\cdot 2^r}\hat{d}_{i+1}^3)$ by proposition 5.1, and then as $O(r^3(2n)^{9+3r+(5+24(i+1))2^r}d^3)$ by proposition 5.2. t_4, which by corollary 1 is the other term that arises in the inner loop, can be written $O(r^2(2n)^{(10+16i)2^r}d^2)$ by proposition 5.2, and now it is clear that t_{9a} dominates. This term also dominates t_1, t_2 and t_3, even without allowing for the fact that they are outside the innermost loop containing step 9a, and so are executed less frequently.

Step 9a is executed less than once for every two regions in the final decomposition of $\mathbf{R}^{i+1}$, since N roots result in $2N+1$ regions, and so the total time for this extension phase is bounded by

$$O(c_{i+1}t_{9a}) \leq (2mn)^{2^r}r^3(2n)^{9+3r+(5+24(i+1))2^r}d^3 \leq m^{2^r}(2n)^{(24i+30)2^r}r^3d^3$$

.

Proposition 9. *The cost of the cylindrical algebraic decomposition is dominated by* $m^{2^r}(2n)^{r2^{r+5}}d^3$.

Proof. . The extension phase dominates the other phases, and the last extension (with $i = r-1$) dominates all the others. hence we can replace $24i+30$ by $24r+6$ in the exponent from the previous proposition, and then replace this by $32r = r2^5$. In this latter replacement, we can afford to drop the r^3 factor, since $r < 2^{r-1}$.

This result is slightly better than that found by McCallum, whose final exponent was $r+7$ rather than our $r+5$. The analyses are sufficiently complex that it is hard to explain the difference precisely, but it stems from a combination of better bounds on root separation (I.5.8 again!) and the absence of factorisation anywhere in the complexity analysis (but see the next section).

7. Further Extensions and Problems.

The algorithm as outlined below will merely produce a sample point for each region. This is perfectly adequate for the application of this method to quantifier elimination, where we merely wish to know whether a formula is true on any, or on all, the regions of a particular stack.

However, we may wish to know more. The first question that springs to mind is "What are the regions?". In chapter II, we defined an algebraic decomposition as one in which the defininig equations for all the regions were algebraic, i.e. solutions of polynomial equations. The algorithm as given in section 4 is unable to do this because, as it happens, there are insufficient data available. For techinical reasons, we may need to have a finer set of polynomials in order to have available enoyugh data with which to construct the defining formulae for each region. Intuitively speaking, we know "what" each critical point and curve is, from the definition of P, but in order to construct defining formulae, we also need to know "why". Space does not permit us to go into the details, but we will give a brief summary here.

Definition. *Let A be a finite set of polynomials in a variable x. Define the* **derivative set** *of A, $der(A)$, to be the set of the greatest square-free divisor of the primitive part of every derivative (of positive order) with respect to x of all elements of A.*

Definition. *If B is a set of square-free relatively prime primitive polynomials in x, then the* **augmented projection** *of B, $AP(B)$ is defined by*

$$AP(B) = coeff(B) \cup disc(B) \cup res(B) \cup der(B).$$

Definition. *If A is any set of polynomials in x, then the* **augmented projection** *of A, $AP(A)$ is defined by*

$$AP(A) = cont(A) \cup AP(B),$$

where $cont(A)$ is the set of contents of elements of A, and B is a relatively prime square-free basis for the primitive parts of the elements of A.

Clearly P is a subset of AP. In order to compute defining equations for the various cells, it is necessary to replace P by AP throughout this chapter. The addition of $der(B)$ has the effect that the analogue of proposition 3.5 is no longer true. We can replace it by the following.

Proposition. *If A has the (m,n) property, then $AP(A)$ has the (m',n') property, where $m' \leq \frac{1}{2}(m+1)^2+mn$ and $n' \leq 2n^2$. Alternatively, $m' \leq m^2 + mn$ for $m \geq 2$.*

Proof. In view of the proof of proposition 3.5, it is suffiicnet to prove that $der(B)$, where B is a square-free basis for A, has the $(mn, 2n^2)$ property. But there are at most n possible derivatives of each polynomial, and the set of discriminants of greatest square-free divisors of primitive parts of k-th derivatives has the $(m, 2(n-k)^2)$ property, and so the $(m, 2n^2)$ property.

8. Practical Experience.

Practical experience with the c.a.d. algorithm is limited. There is essentially only one implementation, in SAC2 [Arnon *et al.*, 1984]. McCallum [1985] reports that he is able to solve several examples, such as the curve from section II.4 in 156 seconds (on a VAX 11/780) and the pair of equations $x^2 + y^2 + z^2 = 1$, $z^3 + xz + y = 0$ in about an hour. The pessimist would note that he then spent about eight hours determining which of the over 1000 regions were adjacent to which others.

He is unable to solve (in over 13 hours) the "random" equation

$$(y-1)z^4 + xz^3 + x(1-y)z^2 + (y-x-1)z + y,$$

or even to perform the first extension (see section II.6 for further details of this problem). While this might be surprising, there seems to be some peculiarity of the geometry of the plane induced by the discriminant. Its discriminant has a factor of multiplicity three, which the argument of proposition II.1.3 would lead one to suspect gave rise to three points in $7R^2$, but in fact there is only one.

Arnon & Smith [1983] have considered the problem of determining the constraints on the semi-axes (a, b) and the centre (c, d) of an ellipse such that it lies inside the unit circle $x^2 + y^2 = 1$. They have not solved this problem mechanically, but indicate some useful transformations.

The author has considered the, apparently trivial, problem of fitting a ladder of length three round a right-angled corridor. This problem is specified by nine equations in four variables. The projection phases take less than 10 minutes DEC 2060 time (if done correctly!), and yield 184 univariate polynomials, of total degree 801. These polynomials have 375 real roots.

We remarked earlier that we did not use factorisation in our theoretical complexity analysis. Nevertheless, it has an important practical aspect. It turned out, in the problem mentioned above, to be much cheaper (3 minutes as against 24), to factorise the polynomials in $\mathbf{R}^1$ rather than to calculate a relatively prime basis via cross-gcd computations.

References

Aho, A., Hopcroft, J.E. & Ullman, J.D., *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, Mass., 1974.

Arnon, D.S. & Smith, S.F., *Towards Mechanical Solution of the Kahan Ellipse Problem I.* Proc. EUROCAL '83 (Springer Lecture Notes in Computer Science 162) pp. 34–44.

Arnon, D.S., Collins, G.E. & McCallum,S., *Cylindrical Algebraic Decomposition.* SIAM J. Comp. **13**(1984) pp. 865–877, 878–889.

Buchberger, B., *Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystem.* Aequationes Math. **4**(1970) pp. 374–383.

Cauchy, A.-L., *Exercices de Mathématiques Quatrième Année.* De Bure Frères, Paris, 1829. Reprinted Œuvres, Sér. II, Tom. IX, Gauthier-Villars, Paris, 1891.

Collins, G.E., *The Calculation of Multivariate Polynomial Resultants.* J. ACM **18**(1971) pp. 515–532.

Collins, G.E., *Quantifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition.* Second GI Conf. Automata Theory and Formal Languages, Springer lecture Notes in Computer Science 33, 1975, pp. 134–183.

Collins, G.E., *Infallible Calculation of Polynomial Zeros to Specified Precision.* Mathematical Software III, Academic Press, New York, 1977, pp. 35–68.

Collins, G.E. & Akritas, A.G., *Polynomial Real Root Isolation Using Descartes' Rule of Signs.* Proc. SYMSAC 76 (ACM, New York), pp. 272–275.

Collins, G.E. & Horowitz, E., *The Minimum Root Separation of a Polynomial.* Math. Comp. **28**(1974) pp. 589–597.

Collins, G.E. & Loos, R.G.K., *Polynomial Real Root Isolation by Differentiation.* Proc. SYMSAC 76 (ACM, New York), pp. 15–25.

Collins, G.E. & Loos, R.G.K., *Real Zeros of Polynomials.* Computing Supplementum 4 (*ed.* B. Buchberger, G.E. Collins & R.G.K. Loos), Springer-Verlag, Wien-New York, 1982, pp. 83–94.

Davenport, J.H., *Real Zeros of Polynomials.* Manuscript, Sept. 1985. Submitted to B.I.T.

Davenport, J.H. & Padget, J.A., *On Numbers and Polynomials.* Computers and Computing (ed. P. Chenin, C. Dicrescenzo, F. Robert), Masson and Wiley, 1985, pp. 49–53.

Gelfond, A.O., *Transcendental and Algebraic Numbers*, Dover, New York, 1960.

Hadamard, J., *Résolution d'une Question Relative aux Déterminants.* Bull. des Sciences Math. (2) **17**(1983) pp. 240–248.

Heindel, L.E., *Integer Arithmetic Algorithms for Polynomial Real Zero Determination.* J. ACM **18**(1971) pp. 533–548.

Knuth, D.E., *The Art of Computer Programming II: Seminumerical Algorithms.* Addison-Wesley, Reading, Mass., 1969.

Landau, E., *Sur Quelques Théorèmes de M. Petrovic Relatif aux Zéros des Polynômes.* Bull. Soc. Math. France **33**(1905) pp. 251–261.

Landau, S., *Polynomial Time Algorithms for Galois Groups.* Proc. EUROSAM 84 (Springer Lecture Notes in Computer Science 174) pp. 225–236.

Landau, S., & Miller, G.L., *Solvability by Radicals is in Polynomial Time.* Proc. 15th. ACM Symposium on Theory of Computing, 1983, pp. 140–151.

Loos, R.G.K., *Generalized Polynomial Remainder Sequences.* Computing Supplementum 4 (*ed.* B. Buchberger, G.E. Collins & R.G.K. Loos), Springer-Verlag, Wien-New York, 1982, pp. 115–137.

Loos, R.G.K., *Computing in Algebraic Extensions.* Computing Supplementum 4 (*ed.* B. Buchberger, G.E. Collins & R.G.K. Loos), Springer-Verlag, Wien-New York, 1982, pp. 173–187.

McCallum, S., An Improved Projection Operation for Cylindrical Algebraic Decomposition. Computer Science Tech. Report 548, University of Wisconsin at Madison, Feb. 1985.

McLaughlin, H.W., *Approximation Theory & Graphics for Problem-Solving Environments.* IFIP WG 2.5 "The Mathematical Problem-Solving Environment", June 1985.

Mahler, K., *An Inequality for the Discriminant of a Polynomial.* Michigan Math. J. **11**(1964) pp. 257–262.

Mignotte, M., *An Inequality about Factors of Polynomials.* Math. Comp. **28**(1974) pp. 1153–1157.

Mignotte, M., *Some Useful Bounds.* Computing Supplementum 4 (*ed.* B. Buchberger, G.E. Collins & R.G.K. Loos), Springer-Verlag, Wien-New York, 1982, pp. 259–263.

Pinkert, J.R., An Exact Method for Finding the Roots of a Complex Polynomial. ACM TOMS **2**(1976) pp. 351–363.

Schwartz, J.T. & Sharir, M., *On the "Piano Movers" Problem. II. General Techniques for Computing Topological Properties of Real Algebraic Manifolds.* Advances in Applied Maths. **4**(1983) pp. 298–351.

van der Waerden, B.L., *Algebra* (Fifth Edition). Springer-Verlag, Berlin-Göttingen-Heidelberg, 1960.

Wilkinson, J.H., *The Evaluation of the Zeros of Ill-conditioned Polynomials.* Numerische Mathematik **1**(1959) pp. 152–166, 167–180.