Mathematics and Algorithms for Computer Algebra

Part 1 © 1992 Dr Francis J. Wright – CBPF, Rio de Janeiro

July 9, 2003

6: Matrices, polynomials and equations

Matrix computation is important in its own right, but it is also an important tool in understanding polynomial remainder sequences, because polynomial division can be expressed in terms of Gaussian elimination in a matrix. One of the techniques for isolating roots of a univariate polynomial uses a "negative" remainder sequence.

1 Bareiss' Gaussian elimination algorithm

The determinant of a matrix over a ring R also belongs to R, because a determinant can be defined in terms only of ring operations. The normal method of computing a determinant is to reduce the matrix to (upper) triangular form by performing elementary (row) operations, which is called Gaussian elimination. However, in its standard form, Gaussian elimination requires division, which is not a ring operation. If R were an integral domain then one could compute in its field of fractions, but this is unsatisfactory because it requires expensive gcd computations, and if R is not an integral domain then it does not have a field of fractions. An analogue of pseudo-division is required, in which one effectively multiplies by a coefficient that ensures that each division can then be performed, but it is desirable to keep such coefficients as small as possible. Bareiss' Gaussian elimination algorithm achieves this.

For matrix elements that have bounded complexity, as with conventional floating-point approximations to real numbers that use a fixed amount of memory, Gaussian elimination is as fast as any algorithm for computing determinants. However, as we have already seen for power computations, the complexity of the data can have a large affect on the overall complexity of an algorithm. For a matrix of integers or univariate polynomials, Bareiss' algorithm is probably the fastest, but for multivariate polynomials other methods, such as cofactor expansion, may be faster overall.

The standard Gaussian elimination algorithm, with division but without pivoting, applied to an $n \times n$ matrix M with elements $m_{ij} = m_{ij}^{(0)}$ in some field, is the following:

```
for $k := 1$ to $n-1$ do
    {triangularize column $k$:}
    for $i := k+1$ to $n$ do
        {zero its subdiagonal elements using the elementary}
        {row operation $R_i^{(k)} := R_i^{(k-1)} - \text{constant} \times R_k^{(k-1)}$:}
        for $j := k$ to $n$ do
            $m_{ij}^{(k)} := m_{ij}^{(k-1)} - \left(m_{ik}^{(k-1)} \Big/ m_{kk}^{(k-1)}\right) m_{kj}^{(k-1)}$;
```

It is also generally necessary to pivot, which means that if $m_{kk}^{(k-1)} = 0$ then a row with index $i > k$ is exchanged with row k and account kept of the consequent change of sign of the determinant. (When computing exactly this simple pivoting algorithm is sufficient, although when computing approximately more sophistication is required to avoid instability.)

Now note that the assignment that effects the elimination can also be written

$$m_{ij}^{(k)} = \left(m_{kk}^{(k-1)} m_{ij}^{(k-1)} - m_{ik}^{(k-1)} m_{kj}^{(k-1)}\right) \Big/ m_{kk}^{(k-1)} = D_{ij}^{(k)} \Big/ m_{kk}^{(k-1)},$$

where $D_{ij}^{(k)}$ denotes the 2×2 determinant

$$D_{ij}^{(k)} = \begin{vmatrix} m_{kk}^{(k-1)} & m_{kj}^{(k-1)} \\ m_{ik}^{(k-1)} & m_{ij}^{(k-1)} \end{vmatrix}.$$

The crucial observation by Bareiss is that, whilst in general $D_{ij}^{(k)}$ is not exactly divisible by $m_{kk}^{(k-1)}$, it is always exactly divisible by $m_{k-1,k-1}^{(k-2)}$ for $k > 1$. So, defining $m_{00}^{(-1)} = 1$, if $m_{ij}^{(k-1)} \in R$ where R is a ring then $D_{ij}^{(k)} / m_{k-1,k-1}^{(k-2)} \in R$. Hence, the optimal factor by which to multiply each $D_{ij}^{(k)}$ before dividing it by $m_{kk}^{(k-1)}$ is effectively $m_{kk}^{(k-1)} / m_{k-1,k-1}^{(k-2)}$. Therefore, Bareiss' elimination algorithm is the same as the Gaussian algorithm given above, but with the initialization

$$m_{00}^{(-1)} := 1$$

and the main assignment

$$m_{ij}^{(k)} := D_{ij}^{(k)} / m_{k-1,k-1}^{(k-2)}.$$

Pivoting is still required in general to avoid 0/0.

The Bareiss elimination procedure computes the determinant of the matrix as follows. For each value of k, row i of the matrix is multiplied by the factor $m_{kk}^{(k-1)}/m_{k-1,k-1}^{(k-2)}$ for $k+1 \leq i \leq n$, which has the following effect on the determinant of the matrix:

$$|M^{(k)}| = |M^{(k-1)}| \left(m_{kk}^{(k-1)} \Big/ m_{k-1,k-1}^{(k-2)} \right)^{n-k}.$$

Applying this relation successively gives

$$\begin{aligned}
|M^{(1)}| &= |M^{(0)}| \left(m_{11}^{(0)} \Big/ m_{00}^{(-1)} \right)^{n-1} = |M^{(0)}| \left(m_{11}^{(0)} \right)^{n-1}, \\
|M^{(2)}| &= |M^{(1)}| \left(m_{22}^{(1)} \Big/ m_{11}^{(0)} \right)^{n-2} = |M^{(0)}| m_{11}^{(0)} (m_{22}^{(1)})^{n-2}, \\
&\vdots \\
|M^{(n-1)}| &= |M^{(0)}|\, m_{11}^{(0)} m_{22}^{(1)} \cdots m_{n-1,n-1}^{(n-2)}.
\end{aligned}$$

But at the end of the elimination process $M^{(n-1)}$ is triangular, so that

$$|M^{(n-1)}| = m_{11}^{(0)} m_{22}^{(1)} \cdots m_{n-1,n-1}^{(n-2)} m_{n,n}^{(n-1)},$$

and hence by comparison with the last of the previous sequence of equations

$$|M| = |M^{(0)}| = m_{n,n}^{(n-1)}.$$

The reason that $D_{ij}^{(k)}/m_{k-1,k-1}^{(k-2)} \in R$ is based on a generalization of an identity due to Sylvester. If the terms of the matrix sequence M^k are defined by

$$m_{ij}^{(k)} = \begin{vmatrix} m_{kk}^{(k-1)} & m_{kj}^{(k-1)} \\ m_{ik}^{(k-1)} & m_{ij}^{(k-1)} \end{vmatrix} \Bigg/ m_{k-1,k-1}^{(k-2)}$$

then it can be shown that

$$m_{ij}^{(k)} = \left| \begin{array}{cccc|c} m_{11} & m_{12} & \cdots & m_{1k} & m_{1j} \\ m_{21} & m_{22} & \cdots & m_{2k} & m_{2j} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{k1} & m_{k2} & \cdots & m_{kk} & m_{kj} \\ \hline m_{i1} & m_{i2} & \cdots & m_{ik} & m_{ij} \end{array} \right|,$$

which shows that $m_{ij}^{(k)} \in R$ because it is defined as a determinant over R. In this notation, $|M| = m_{nn}^{(n-1)}$.

However, given that

$$m_{00}^{(-1)} = 1, \quad M^{(0)} = M,$$

$$m_{ij}^{(k)} = \left| \begin{matrix} m_{kk}^{(k-1)} & m_{kj}^{(k-1)} \\ m_{ik}^{(k-1)} & m_{ij}^{(k-1)} \end{matrix} \right| \bigg/ m_{k-1,k-1}^{(k-2)},$$

the most direct way to prove that $m_{ij}^{(k)} \in R$ is by induction on k. The base case is that for $k = 1$

$$m_{ij}^{(1)} = (m_{11}m_{ij} - m_{i1}m_{1j})/1 \ \in R.$$

Inserting (with the help of REDUCE) the values

$$\begin{aligned}
m_{k,k}^{(k-1)} &= \left(m_{k-1,k-1}^{(k-2)} m_{k,k}^{(k-2)} - m_{k-1,k}^{(k-2)} m_{k,k-1}^{(k-2)}\right) \Big/ m_{k-2,k-2}^{(k-3)} \\
m_{k,j}^{(k-1)} &= \left(m_{k-1,k-1}^{(k-2)} m_{k,j}^{(k-2)} - m_{k-1,j}^{(k-2)} m_{k,k-1}^{(k-2)}\right) \Big/ m_{k-2,k-2}^{(k-3)} \\
m_{i,k}^{(k-1)} &= \left(m_{k-1,k-1}^{(k-2)} m_{i,k}^{(k-2)} - m_{k-1,k}^{(k-2)} m_{i,k-1}^{(k-2)}\right) \Big/ m_{k-2,k-2}^{(k-3)} \\
m_{i,j}^{(k-1)} &= \left(m_{k-1,k-1}^{(k-2)} m_{i,j}^{(k-2)} - m_{k-1,j}^{(k-2)} m_{i,k-1}^{(k-2)}\right) \Big/ m_{k-2,k-2}^{(k-3)}
\end{aligned}$$

into

$$m_{i,j}^{(k)} = \left(m_{i,j}^{(k-1)} m_{k,k}^{(k-1)} - m_{i,k}^{(k-1)} m_{k,j}^{(k-1)}\right) \Big/ m_{k-1,k-1}^{(k-2)}$$

gives

$$\begin{aligned}
m_{i,j}^{(k)} = \Big(& m_{k-1,k-1}^{(k-2)} m_{i,j}^{(k-2)} m_{k,k}^{(k-2)} - m_{k-1,k-1}^{(k-2)} m_{i,k}^{(k-2)} m_{k,j}^{(k-2)} - \\
& m_{k-1,j}^{(k-2)} m_{i,k-1}^{(k-2)} m_{k,k}^{(k-2)} + m_{k-1,j}^{(k-2)} m_{i,k}^{(k-2)} m_{k,k-1}^{(k-2)} + \\
& m_{k-1,k}^{(k-2)} m_{i,k-1}^{(k-2)} m_{k,j}^{(k-2)} - m_{k-1,k}^{(k-2)} m_{i,j}^{(k-2)} m_{k,k-1}^{(k-2)}\Big) \Big/ \left(m_{k-2,k-2}^{(k-3)}\right)^2.
\end{aligned}$$

The denominator $m_{k-1,k-1}^{(k-2)}$ has exactly divided out of this result. Hence, if we make the induction hypothesis that $m_{i,j}^{(k-1)} \in R$ then the denominator $m_{k-2,k-2}^{(k-3)}$ must in fact exactly divide out of each formula for $m_{i,j}^{(k-1)}$, and hence because $m_{i,j}^{(k)}$ is quadratic in $m_{i,j}^{(k-1)}$ its full denominator must in fact exactly divide out, so that $m_{i,j}^{(k)} \in R$. Then by induction this holds for all k.

2 Polynomial division and matrices

The aim of this section is to sketch some important theory that underlies a large part of computer algebra. Much of this is classical pure mathematics that has been know for the last hundred years, and now finds important practical application. My presentation is based on the first half of the paper "Generalized Polynomial Remainder Sequences" by R. Loos in Buchberger, Collins & Loos. I primarily want to convey the flavour of the subject, so I have expanded the discussion of the concepts and omitted several of the proofs. My primary motivation is to understand the subresultant pseudo-remainder sequence, and the section finishes with this topic.

The process of polynomial division is very similar to that of Gaussian elimination in a matrix, and this analogy leads to a useful tool for the theoretical analysis of polynomial division.

Initially, let F be a field and let the polynomials $a, b \in F[x]$ have degrees $m = \deg a$, $n = \deg b$, $m \geq n$. Then we already know that there exist unique polynomials $q = \text{quot}(a, b)$, $r = \text{rem}(a, b)$ such that either $r = 0$ or $\deg r < \deg b$. The division algorithm, expressed in terms of operations on complete polynomials rather than on individual terms, is the following:

```
input: $a = \sum_{i=0}^{m} a_i x^i$, $b = \sum_{j=0}^{n} b_j x^j$
$a^{(0)} := a$;
for $k := 1$ to $m - n + 1$ do
begin
    $q^{(k)} := \text{lc}\, a^{(k-1)} / b_n$;
    $a^{(k)} := a^{(k-1)} - x^{m-n+1-k} q^{(k)} b$
end;
output: $q = \sum_{i=0}^{m-n} q^{(m-n+1-i)} x^i$, $r = a^{(m-n+1)}$
```

A polynomial can be represented densely as a row vector of its coefficients, and a sequence of polynomials can be represented as the sequence of rows comprising a matrix, as follows.

Definition 1 *Let $P_i = \sum_{j=0}^{d_i} p_{ij} x^j, 1 \leq i \leq k$ be a sequence of k polynomials P_i over an integral domain D, where* $\deg P_i = d_i$. *Then the $k \times l$* associated matrix *of $\{P_i\}$ is*

$$\text{mat}(P_1, P_2, \ldots, P_k) = (a_{i,l-j}),$$

where

$$l = 1 + \max_{1 \leq i \leq k} (d_i) \quad \textit{and} \quad a_{i,l-j} = 0 \textit{ if } d_i < j < l.$$

Thus the associated matrix is made just wide enough to hold the polynomial of highest degree, and polynomials of lower degree are padded with leading zeros as necessary. [The mat notation used here is identical to that used by REDUCE for matrix input in terms of a sequence of rows.]

Then the sequence of polynomials involved in the division process can be represented by the $(m-n+2) \times (m+1)$ associated matrix $M = \mathrm{mat}(x^{m-n}b, x^{m-n-1}b, \ldots, b, a)$, namely

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

The division algorithm for the polynomials a, b is identical to Gaussian elimination in the matrix M, which in this special case only changes the last row, because the first $m-n+1$ rows are already in upper triangular form. Moreover, no pivoting is required, because the pivot is always $b_n \neq 0$, the leading coefficient of b. After the Gaussian elimination, the matrix M has the upper triangular form

$$M' = \begin{pmatrix} b_n & b_{n-1} & \cdots & b_0 & 0 & \cdots & 0 \\ 0 & b_n & b_{n-1} & \cdots & b_0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \ddots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & b_n & b_{n-1} & \cdots & b_0 \\ 0 & \cdots & 0 & 0 & r_{n-1} & \cdots & r_0 \end{pmatrix},$$

i.e. $M' = \mathrm{mat}(x^{m-n}b, x^{m-n-1}b, \ldots, b, \mathrm{rem}(a, b))$.

The associated matrix constructs a matrix from any sequence of polynomials; the following definition constructs one polynomial from any matrix.

Definition 2 *Let $\mathcal{M}$ be a $k \times l$ matrix, $k \leq l$, over an integral domain D. The* determinant polynomial *of $\mathcal{M}$ is*

$$\mathrm{detpol}(\mathcal{M}) = |\mathcal{M}_{(k)}|x^{l-k} + \cdots + |\mathcal{M}_{(l)}|,$$

where $\mathcal{M}_{(j)}$ denotes the submatrix of $\mathcal{M}$ consisting of the first $k-1$ columns followed by the j^{th} column, for $k \leq j \leq l$. Clearly, if $\mathcal{M}$ is square then $\mathrm{detpol}(\mathcal{M}) = |\mathcal{M}|$, *and* $\deg \mathrm{detpol}(\mathcal{M}) \leq l - k$ *or* $\mathrm{detpol}(\mathcal{M}) = 0$. *If the* *matrix $\mathcal{M}$ is composed of the k rows $\mathcal{M}_i, 1 \leq i \leq k$, then it is convenient to extend the notation and also write*

$$\operatorname{detpol}(\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k) = \operatorname{detpol}(\mathcal{M}).$$

[See below for an explicit example.] Then, because Gaussian elimination does not change a determinant, and because M' is upper triangular,

$$\operatorname{detpol}(M) = \operatorname{detpol}(M') = b_n^{m-n+1} \operatorname{rem}(a, b).$$

2.1 Pseudo-division and Bareiss elimination

Now suppose that $a, b \in D[x]$ where the integral domain D is not a field. Then polynomial division is in general not possible and neither is Gaussian elimination in the matrix M. However, polynomial pseudo-division is possible and so is Bareiss elimination in M, and not surprisingly the two are related. The essential problem is to compute

$$\operatorname{detpol}(M) = \operatorname{detpol}(x^{m-n}b, x^{m-n-1}b, \ldots, b, a),$$

which is well defined over a ring.

Immediately after step $k-1$ of the Bareiss elimination, the k^{th} row of M excluding the last row, i.e. for $1 < k \leq m-n+1$, is b_n^{k-1} times what it was originally, and M has the following form:

$$\begin{array}{c c}
\underline{1}\rfloor & \begin{array}{cccccccc} & & & \text{col } k & & & & \end{array} \\
\begin{array}{c} \\ \\ \\ \\ \text{row } k \\ \\ \\ \\ \end{array} &
\left(\begin{array}{cccccccc}
b_n & b_{n-1} & \cdots & \cdots & \cdots & 0 & 0 & 0 \\
0 & b_n^2 & b_n b_{n-1} & \cdots & \cdots & 0 & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & b_n^{k-1} & b_n^{k-2} b_{n-1} & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & b_n^k & b_n^{k-1} b_{n-1} & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & b_n^k & b_n^{k-1} b_{n-1} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & a_{m+1-k}^{(k-1)} & \cdots & \cdots & \cdots & a_0^{(k-1)}
\end{array}\right)
\end{array}$$

The 1 at the top left represents the value $m_{00}^{(-1)} = 1$ necessary to initialize the Bareiss elimination.

As indicated explicitly in the above matrix, the cumulative effect of the first $k-1$ steps of Bareiss elimination on all but the last row, i.e. for

$1 < k \leq m - n + 1$, is

$$R_i^{(k-1)} = \begin{cases} b_n^{i-1} R_i^{(0)} \text{ for } 1 < i \leq k \leq m + n + 1, \\ b_n^{k-1} R_i^{(0)} \text{ for } 1 < k < i \leq m + n + 1. \end{cases}$$

To see why this is so, recall that the Bareiss elimination formula in terms of rows is

$$R_i^{(k)} = \left(m_{kk}^{(k-1)} R_i^{(k-1)} - m_{ik}^{(k-1)} R_k^{(k-1)} \right) \Big/ m_{k-1,k-1}^{(k-2)},$$

which simplifies for all but the last row, because of the triangular structure, to

$$R_i^{(k)} = \left(m_{kk}^{(k-1)} \Big/ m_{k-1,k-1}^{(k-2)} \right) R_i^{(k-1)} \text{ for } 1 \leq k < i \leq m - n + 1.$$

The factor $m_{kk}^{(k-1)}/m_{k-1,k-1}^{(k-2)} = b_n/1 = b_n$ for $k = 1$ because $m_{00}^{(-1)} = 1$, and hence every row with $1 < i \leq m - n + 1$ is multiplied by b_n. Then $m_{kk}^{(k-1)}/m_{k-1,k-1}^{(k-2)} = b_n^2/b_n = b_n$ for $k = 2$, and so every row with $2 < i \leq m - n + 1$ is again multiplied by b_n. This procedure continues, and $m_{kk}^{(k-1)}/m_{k-1,k-1}^{(k-2)} = b_n^k/b_n^{k-1} = b_n$ for general k, so that the elimination formula simplifies still further to

$$R_i^{(k)} = b_n R_i^{(k-1)} \text{ for } 1 \leq k < i \leq m - n + 1.$$

This leads to the intermediate state shown above, and at the end of the Bareiss elimination each row with $1 \leq i \leq m - n + 1$ has been multiplied by b_n^{i-1}.

For the last row, the k^{th} step for $1 \leq k \leq m - n + 1$ of the elimination formula in terms of rows simplifies to

$$\begin{aligned} R_{m+n+2}^{(k)} &= \left(b_n^k R_{m+n+2}^{(k-1)} - a_{m+1-k}^{(k-1)} R_k^{(k-1)} \right) \Big/ b_n^{k-1} \\ &= \left(b_n^k R_{m+n+2}^{(k-1)} - a_{m+1-k}^{(k-1)} b_n^{k-1} R_k^{(0)} \right) \Big/ b_n^{k-1} \\ &= b_n R_{m+n+2}^{(k-1)} - a_{m+1-k}^{(k-1)} R_k^{(0)}, \end{aligned}$$

which involves no division at all.

For the polynomial division process, the transformations of the b rows of M are irrelevant, and each transformation of the a line, in polynomial notation, has the form

$$a^{(k)} := b_n a^{(k-1)} - x^{m-n+1-k} \operatorname{lc} a^{(k-1)} b.$$

This is precisely the pseudo-division algorithm presented earlier, which can therefore be regarded as a specialization of Bareiss elimination involving no division at all. The pseudo-remainder is $\operatorname{prem}(a, b) = a^{(m-n+1)}$ and the pseudo-quotient is $\operatorname{pquot}(a, b) = \sum_{i=0}^{m-n} \operatorname{lc} a^{(i)} x^i$. Moreover, the pseudo-division process clearly effectively multiplies a by b_n at each of the $m-n+1$ steps. Hence

$$\begin{aligned}\operatorname{detpol}(M) &= \operatorname{detpol}(x^{m-n}b, x^{m-n-1}b, \ldots, b, a) \\ &= \operatorname{prem}(a, b) = \operatorname{rem}(b_n^{m-n+1}a, b),\end{aligned}$$

because from the previous analysis of Bareiss elimination, a determinant is given by the final value of the element in the bottom row, which in this case is a coefficient of a^{m-n+1}, i.e. the determinant polynomial is independent of the b rows.

In fact, if it happens that $\operatorname{lc} a^{(k-1)} = 0$ then the straightforward pseudo-division algorithm introduces an unnecessary factor of b_n, which can be avoided by replacing the k^{th} step by

```
if $\operatorname{lc} a^{(k-1)} = 0$ then $a^{(k)} := a^{(k-1)}$ else
$a^{(k)} := b_n a^{(k-1)} - x^{m-n+1-k} \operatorname{lc} a^{(k-1)} b$;
```

This algorithm is called *sparse pseudo-division*, and the *sparse pseudo-remainder* is denoted $\operatorname{sprem}(a, b) = a^{(m-n+1)}$. It is related to the straightforward pseudo-remainder by

$$b_n^e \operatorname{sprem}(a, b) = \operatorname{prem}(a, b) \text{ for some } e \geq 0,$$

where e is the number of times that the sparsity condition $\operatorname{lc} a^{(k-1)} = 0$ is satisfied during the sparse pseudo-division algorithm. The essence of subresultant pseudodivision is to take advantage of similar but more subtle sparsity effects.

2.2 Polynomial remainder sequences and subresultants

The general notion of a polynomial remainder sequence (PRS) can conveniently be expressed by using a generalization of the notion of associates in a ring (with unity), because the Euclidean remainder, pseudo-remainder and sparse pseudo-remainder differ only in their content, i.e. by an element of the coefficient domain.

Definition 3 *Let F be the quotient field of an integral domain D, then $F[x] > D[x]$. Two polynomials a, b over D or F are called* similar*, denoted $a \sim b$, if and only if there exist $u, v \in D$, $uv \neq 0$, such that $ua = vb$.*

The polynomial $a \sim 0$ if and only if $a = 0$, and $\sim$ is an equivalence relation. In the special case that u, v are units, a, b are associates.

Definition 4 *A sequence of polynomials $p_1, p_2, \ldots, p_r \in D[x]$ is a* polynomial remainder sequence (PRS) *of the initial polynomials $p_1, p_2 \neq 0$ if, for $r \geq 2$,*

$$p_{i+2} \sim \operatorname{prem}(p_i, p_{i+1}) \neq 0 \textit{ for } 1 \leq i \leq r-2, \quad \operatorname{prem}(p_{r-1}, p_r) = 0.$$

Hence, there exist $e_i, f_i \in D$, not both zero, and $q_i(x) \in D[x]$ such that

$$e_i p_i = q_i p_{i+1} + f_i p_{i+2},$$

which conversely implies that $p_{i+2} \sim \operatorname{prem}(p_i, p_{i+1})$.

In a PRS, each polynomial occurs first as divisor and then as dividend, and so a matrix representation is required that treats the two polynomials in a division more symmetrically than that discussed above. This motivates the following:

Definition 5 *Let $a, b \in D[x]$, $\deg a = m > 0$ and $\deg b = n > 0$. For $0 \leq k < \min(m, n)$ let*

$$M_k = \operatorname{mat}(x^{n-k-1}a(x), x^{n-k-2}a(x), \ldots, a(x),\ x^{m-k-1}b(x), \ldots, b(x)).$$

Then the k^{th} subresultant *of a and b is $S_k = \operatorname{sres}_k(a, b) = \operatorname{detpol}(M_k)$.*

The matrix M_k has $m + n - 2k$ rows and $m + n - k$ columns and therefore, from the definition of determinant polynomial,

$$\deg S_k \leq (m + n - 2k) - (m + n - k) = k.$$

The subresultant S_0, which is a polynomial of degree 0 with respect to x, is precisely the *resultant* $\operatorname{Res}_x(a, b)$ as defined previously. The matrix M_k is the Sylvester matrix of a and b with the first (or equivalently the last) k of each of the a and b rows omitted (which therefore removes $2k$ rows but only k columns), hence the name *subresultant*.

For example, suppose $m = 4, n = 3$; then for $0 \leq k < \min(m, n) = 3$, i.e. $k = 0, 1, 2$:

$$
\begin{aligned}
M_0 &= \begin{pmatrix}
a_4 & a_3 & a_2 & a_1 & a_0 & 0 & 0 \\
0 & a_4 & a_3 & a_2 & a_1 & a_0 & 0 \\
0 & 0 & a_4 & a_3 & a_2 & a_1 & a_0 \\
b_3 & b_2 & b_1 & b_0 & 0 & 0 & 0 \\
0 & b_3 & b_2 & b_1 & b_0 & 0 & 0 \\
0 & 0 & b_3 & b_2 & b_1 & b_0 & 0 \\
0 & 0 & 0 & b_3 & b_2 & b_1 & b_0
\end{pmatrix} = \operatorname{Res}(a, b), \\
M_1 &= \left(\begin{array}{ccccc|c}
a_4 & a_3 & a_2 & a_1 & a_0 & 0 \\
0 & a_4 & a_3 & a_2 & a_1 & a_0 \\
b_3 & b_2 & b_1 & b_0 & 0 & 0 \\
0 & b_3 & b_2 & b_1 & b_0 & 0 \\
0 & 0 & b_3 & b_2 & b_1 & b_0
\end{array}\right), \\
M_2 &= \left(\begin{array}{ccc|c|c}
a_4 & a_3 & a_2 & a_1 & a_0 \\
b_3 & b_2 & b_1 & b_0 & 0 \\
0 & b_3 & b_2 & b_1 & b_0
\end{array}\right).
\end{aligned}
$$

The coefficients of S_k are the determinants of the square submatrices of M_k whose last columns are any of those shown with a vertical line to their right, including the last column of M_k, e.g.

$$
S_2 = \begin{vmatrix} a_4 & a_3 & a_2 \\ b_3 & b_2 & b_1 \\ 0 & b_3 & b_2 \end{vmatrix} x^2 + \begin{vmatrix} a_4 & a_3 & a_1 \\ b_3 & b_2 & b_0 \\ 0 & b_3 & b_1 \end{vmatrix} x + \begin{vmatrix} a_4 & a_3 & a_0 \\ b_3 & b_2 & 0 \\ 0 & b_3 & b_0 \end{vmatrix}.
$$

A sequence of the form $a, b, \operatorname{sres}_{n-1}(a, b), \operatorname{sres}_{n-2}(a, b), \ldots, \operatorname{sres}_0(a, b) = \operatorname{Res}(a, b)$, where $n = \deg b \leq \deg a$, is called a *subresultant chain.* It is closely related to the PRS generated by a, b, but is easier to analyse. In fact, for any PRS with $n_i = \deg p_i$ and $\operatorname{prem}(p_{r-1}, p_r) = 0$,

$$
p_i \sim S_{n_{i-1}-1} \sim S_{n_i} \quad \text{for} \quad 3 \leq i \leq r
$$

and

$$
S_k = 0 \quad \text{for} \quad n_i < k < n_{i-1} - 1, \quad 3 \leq i \leq r.
$$

2.3 Analysis of subresultant chains

We saw earlier that sparsity of polynomials being divided can reduce the size of the factor necessary in pseudo-division. In order to make use of this observation it is necessary to study the way that sparsity propagates through a subresultant chain, and to trace in detail the effect of the similarity coefficients e_i, f_i on a PRS.

It is convenient to use Kronecker's method of "indeterminate coefficients", in which polynomial coefficients are regarded temporarily not as elements of some number domain but as "indeterminates", which later can be given particular numerical values, including zero. Then we temporarily regard $a = \sum_{i=0}^{n+1} a_i x^i$, $b = \sum_{j=0}^{n} b_j x^j$ as polynomials over $\mathbb{Z}$ in x with indeterminate coefficients. The particular relationship between the degrees of a and b can be changed by allowing some leading coefficients to take the value zero later.

The determinant polynomial defining a subresultant can be related to a pseudo-remainder in a way similar to that used previously. By definition, for $0 \leq k < n$,

$$\begin{aligned} S_k &= \operatorname{sres}_k(a,b) = \operatorname{detpol}(M_k) \\ &= \operatorname{detpol}(x^{n-k-1}a, x^{n-k-2}a, \ldots, a,\ x^{n-k}b, x^{n-k-1}b, \ldots, b), \end{aligned}$$

because $m = \deg a = n+1$. The matrix M_k has $2n+1-2k$ rows, hence moving the first row to the last without otherwise changing the order of the rows requires $2(n-k)$ row exchanges, which is always even and so does not change the sign of the determinant. Applying the same argument to all of the a rows leads to

$$S_k = \operatorname{detpol}(x^{n-k}b, x^{n-k-1}b, \ldots, b,\ x^{n-k-1}a, x^{n-k-2}a, \ldots, a).$$

Multiplying each of the $n-k$ rows of a coefficients by b_n^2 gives

$$S_k = \operatorname{detpol}(x^{n-k}b, \ldots, b,\ x^{n-k-1}b_n^2 a, \ldots, b_n^2 a) b_n^{-2(n-k)}.$$

Gaussian elimination using appropriate b rows allows each a row to be reduced to $\operatorname{rem}(b_n^2 a, b) = \operatorname{prem}(a,b)$, thus

$$S_k = \operatorname{detpol}(x^{n-k}b, \ldots, b,\ x^{n-k-1}b_n^2 \operatorname{prem}(a,b), \ldots, b_n^2 \operatorname{prem}(a,b)) b_n^{-2(n-k)}.$$

But $\deg \operatorname{prem}(a,b) \leq n-1$, whereas the matrix was originally constructed using $\deg a = n+1$. Hence the $\operatorname{prem}(a,b)$ rows *all* have at least 2 leading zeros, and so each determinant comprising a coefficient of the determinant polynomial can be cofactor expanded about its first two columns.

For $0 \leq k \leq n-2$ this leads to $S_k =$

$$\operatorname{detpol}(x^{n-2-k}b, \ldots, b,\ x^{n-k-1}b_n^2 \operatorname{prem}(a,b), \ldots, b_n^2 \operatorname{prem}(a,b)) b_n^{-2(n-k)+2}$$

and hence, in a form that does not require division,

$$b_n^{2(n-k-1)} S_k = b_n^{2(n-k-1)} \operatorname{sres}_k(a,b) = \operatorname{sres}_k(b, \operatorname{prem}(a,b)), \quad 0 \le k \le n-2.$$

For $k = n-1$, $S_{n-1} = \operatorname{detpol}(xb, b, \operatorname{prem}(a,b)) b_n^{-2}$ and cofactor expansion about its first two columns leads to

$$S_{n-1} = \operatorname{sres}_{n-1}(a,b) = \operatorname{prem}(a,b).$$

(In fact, the matrix is upper triangular.)

It is convenient to extend the notation and define

$$S_{n+1} = a, \quad S_n = b,$$

hence

$$S_{n-1} = \operatorname{sres}_{n-1}(S_{n+1}, S_n) = \operatorname{prem}(S_{n+1}, S_n).$$

We will also need

Definition 6 *The k^{th}* principal subresultant coefficient*, for $0 \le k \le n+1$, is*

$$R_k = \begin{cases} \textit{coefficient of } x^k \textit{ in } S_k & \textit{for } 0 \le k \le n, \\ 1 & \textit{for } k = n+1. \end{cases}$$

[With indeterminate coefficients, $R_k = \operatorname{lc} S_k$ for $0 \le k \le n$, but this is not necessarily true if the coefficients are allowed to take numerical values.]

The following theorem, which is not new, relates subresultants to their predecessors (in decreasing subscript order) in the subresultant chain.

Theorem 1 (Habicht, 1948) *Let a, b be polynomials of degrees $n+1, n$ respectively with indeterminate coefficients, and let*

$$S_{n+1} = a, S_n = b, S_{n-1}, \ldots, S_0$$

be the subresultant chain of a, b. Then, for all j, $0 < j \le n$,

$$\begin{aligned} R_{j+1}^{2(j-r)} S_r &= \operatorname{sres}_r(S_{j+1}, S_j), \quad 0 \le r < j, \\ R_{j+1}^2 S_{j-1} &= \operatorname{prem}(S_{j+1}, S_j). \end{aligned}$$

Habicht's Theorem gives the links that exist within a completely general subresultant chain. In order to consider sparsity, we must now let the indeterminate coefficients take values in D, such that $\deg a = n_1$, $\deg b = n_2$. If

$n_1 > n_2$ we let $n_1 = n + 2$, $b_n = \cdots = b_{n_2+1} = 0$, whereas if $n_1 \leq n_2$ we let $n_2 = n$, $a_{n+1} = a_n = \cdots = a_{n_1+1} = 0$. Therefore,

$$n = \begin{cases} n_1 - 1 \text{ if } n_1 > n_2, \\ \quad n_2 \quad \text{if } n_1 \leq n_2. \end{cases}$$

With indeterminate coefficients it is not possible for any of the determinants defining the coefficients of a subresultant to vanish, whereas when the coefficients are allowed to take values in D such cancellation can occur. Therefore, $\deg S_j \leq j$, and if $\deg S_j = r < j$ then S_j is called *defective* of degree r; otherwise it is called *regular*. [This terminology dates back over one hundred years!] A subresultant chain or PRS is called *regular* if all its elements are regular, and defective otherwise. Moreover, the k^{th} principal subresultant coefficient R_k is defined to be the coefficient of x^k in S_k, which if it vanishes is not the same as the actual leading coefficient $\operatorname{lc} S_k$.

Allowing the coefficients to take values in D leads to what is essentially a special case of Habicht's Theorem:

Theorem 2 (Subresultant Theorem) *Let $S_{n+1}, S_n, \ldots, S_0$ be the subresultant chain in $D[x]$ of S_{n+1}, S_n. Let S_{j+1} be regular and S_j be defective of degree $r < j$ (with $\deg 0 = -1$). Then*

$$S_{j-1} = S_{j-2} = \cdots = S_{r+1} = 0, \quad -1 \leq r < j < n,$$

$$R_{j+1}^{j-r} S_r = \operatorname{lc}(S_j)^{j-r} S_j, \quad 0 \leq r \leq j < n, \tag{1}$$

$$(-1)^{j-r} R_{j+1}^{j-r+2} S_{r-1} = \operatorname{prem}(S_{j+1}, S_j), \quad 0 < r \leq j < n. \tag{2}$$

This theorem shows how a defective subresultant (one or more of whose leading terms happen to vanish) causes *gaps* in the subresultant chain, in which successive subresultants vanish completely, and related jumps in the degrees of the elements of the associated PRS.

2.4 Analysis of polynomial remainder sequences

The final step is to relate a PRS to the subresultant chain starting from the same pair of polynomials. Only a regular subresultant chain can actually be identical to a PRS, because a PRS cannot contain any zero elements (otherwise all subsequent elements would vanish). The *subresultant* PRS, introduced by George Collins in 1967, is a PRS that has (not surprisingly) a fairly simple relationship to the subresultant chain, as expressed in the following

Theorem 3 (Collins-Loos) *Let $A_1, A_2, \ldots, A_r$ be a PRS over D with $n_i = \deg A_i$, and $e_i, f_i \in D*$ (i.e. $e_i, f_i \neq 0$), such that*

$$e_i A_i = Q_i A_{i+1} + f_i A_{i+2} \quad (1 \leq i \leq r-1).$$

Then

$$A_i = S_{n_{i-1}-1} \quad \textit{for} \quad 1 \leq i \leq r$$

where by definition $n_0 = n_1 + 1$, if e_i, f_i are defined as follows.

Let $\delta_i = n_i - n_{i+1}$, $c_i = \operatorname{lc} A_i$, and

$$e_i = c_{i+1}^{\delta_i+1} \quad \textit{for} \quad 1 \leq i \leq r-1,$$

$$f_1 = 1, \quad f_i = -c_i(-R_{n_i})^{\delta_i} \quad \textit{for} \quad 2 \leq i \leq r-2,$$

where $R_{n_i} = \operatorname{lc} S_{n_i}$ for $i > 1$.

Note that because the coefficients of a subresultant are defined as determinants over a ring, the subresultant is a polynomial over that ring. Hence this theorem proves that a subresultant PRS generates polynomials over the original coefficient ring, and involves only ring operations. The proof below shows how the similarity coefficients are derived from the properties of subresultants discussed above.

Proof This follows from the Subresultant Theorem by induction. By definition, $A_1 = S_{n_1} = S_{n_0-1}$ and $A_2 = S_{n_1-1}$, which establishes a base case. Take as the induction hypothesis that the theorem is true for i and $i+1$. Then equation (1) in the statement of the Subresultant Theorem gives, with $j = n_{i-1} - 1$, $r = n_i$, $j - r = n_{i-1} - n_i - 1 = \delta_{i-1} - 1$,

$$R_{n_{i-1}}^{\delta_{i-1}-1} S_{n_i} = \operatorname{lc}(S_{n_{i-1}-1})^{\delta_{i-1}-1} S_{n_{i-1}-1},$$

i.e.

$$R_{n_{i-1}}^{\delta_{i-1}-1} S_{n_i} = c_i^{\delta_{i-1}-1} A_i,$$

and taking leading coefficients of both sides gives

$$R_{n_{i-1}}^{\delta_{i-1}-1} R_{n_i} = c_i^{\delta_{i-1}}.$$

Dividing these two equations then gives

$$c_i S_{n_i} / R_{n_i} = A_i. \tag{3}$$

Similarly, equation (2) gives, with $j = n_i - 1$, $r = n_{i+1}$, $j - r = n_i - n_{i+1} - 1 = \delta_i - 1$,

$$(-R_{n_i})^{\delta_i+1} S_{n_{i+1}-1} = \mathrm{prem}(S_{n_i}, A_{i+1}). \tag{4}$$

Now by the definition of a PRS,

$$\begin{aligned} f_i A_{i+2} &= \mathrm{prem}(A_i, A_{i+1}) \\ &= c_i / R_{n_i} \, \mathrm{prem}(S_{n_i}, A_{i+1}) \text{ by (3)} \\ &= c_i / R_{n_i} (-R_{n_i})^{\delta_i+1} S_{n_{i+1}-1} \text{ by (4)} \\ &= -c_i (-R_{n_i})^{\delta_i} S_{n_{i+1}-1} \\ &= f_i S_{n_{i+1}-1} \text{ by definition of } f_i, \end{aligned}$$

and hence $A_{i+2} = S_{n_{i+1}-1}$. Therefore, by induction, the theorem is true generally. □

The relationship between a general PRS and its associated subresultant chain is more complicated, as expressed in the following

Theorem 4 (Fundamental PRS Theorem)
Let $A_1, A_2, \ldots, A_r$ be a PRS over D such that, for $e_i, f_i \in D*$ (i.e. $e_i, f_i \neq 0$)

$$e_i A_i = Q_i A_{i+1} + f_i A_{i+2} \quad (1 \leq i \leq r-2).$$

Let $n_i = \deg A_i$ and $c_i = \operatorname{lc} A_i$. The for any j, $1 < j < r$,

$$S_k = 0 \quad \textit{for} \quad 0 \leq k < n_r \ \textit{and}\ n_{j+1} < k < n_j - 1,$$

$$\left(\prod_{i=1}^{j-1} e_i^{n_{i+1}-n_j+1} \right) S_{n_j-1} =$$
$$\left(\prod_{i=1}^{j-1} (-1)^{(n_i-n_j+1)(n_{i+1}-n_j+1)} f_i^{n_{i+1}-n_j+1} c_{i+1}^{n_i-n_{i+2}} \right) c_j^{-n_j+n_{j+1}+1} A_{j+1},$$

$$\left(\prod_{i=1}^{j-1} e_i^{n_{i+1}-n_{j+1}} \right) S_{n_{j+1}} =$$
$$\left(\prod_{i=1}^{j-1} (-1)^{(n_i-n_{j+1})(n_{i+1}-n_{j+1})} f_i^{n_{i+1}-n_{j+1}} c_{i+1}^{n_i-n_{i+2}} \right) e_{j+1}^{n_j-n_{j+1}-1} A_{j+1}.$$

Beware that there are very subtle differences between the powers appearing in the last two equations.

Collins also discovered a very simple PRS called the *reduced* PRS, defined by

$$e_i = (\operatorname{lc} A_{i+1})^{n_i - n_{i+1} + 1} \quad \text{and} \quad f_1 = 1, \quad f_{i+1} = e_i \text{ for } 1 \leq r \leq r - 1.$$

The proof that this sequence defines a PRS requires use of the Fundamental PRS Theorem.

3 Sturm sequences and polynomial zero isolation

An equation can be rescaled, by multiplying it by any finite non-zero quantity that is independent of the unknowns, without changing the roots of the equation. Hence, a polynomial equation with rational coefficients can be multiplied by their common denominator to convert it into an equivalent equation with integer coefficients that has identical roots, and similarly a polynomial equation with integer coefficients can be multiplied by an arbitrary finite non-zero rational constant. Thus a polynomial equation over $\mathbb{Z}$ or $\mathbb{Q}$ can be trivially reduced to a primitive polynomial over $\mathbb{Z}$ or a monic polynomial over $\mathbb{Q}$ as is convenient, and I will refer to all polynomials having exact explicit numerical coefficients as being over $\mathbb{Q}$.

This section considers the problem of finding the *real* zeros (not complex zeros, although some of the ideas generalize to complex zeros) of a univariate polynomial $p(x)$ over $\mathbb{Q}$. In general, it is impossible to solve a polynomial equation to obtain results that are both exact and explicit, even if the coefficients are rationals, because the solutions are (algebraic) irrational numbers, which have no representation that is both exact and explicit. For example, even the trivial equation $x^2 - 2 = 0$ can be solved either as $x = \pm\sqrt{2}$, which is exact but implicit or symbolic, or as $x = \pm 1.414\ldots$, which is explicit but approximate.

An explicit exact solution must be expressed in terms of rational constants, and the best that can be achieved is to find *isolating intervals.* A set of isolating intervals for a univariate polynomial equation is a set of intervals with rational end-points such that every interval contains precisely one root of the equation and every root of the equation is contained in precisely one interval. Once a set of isolating intervals has been found, each interval can be made as small as desired by some suitably reliable technique that is essentially numerical, of which the simplest is interval bisection. The algebraic component of the calculation is finding the complete set of isolating intervals.

The "classical" technique for doing this uses a Sturm sequence, which is a particular kind of polynomial remainder sequence. Let $p(x)$ be a squarefree polynomial in x over $\mathbb{Q}$. Then its Sturm sequence is the sequence of polynomials $\{p_i(x)\}_{i=0}^k$ in x over $\mathbb{Q}$ defined by

$$\begin{aligned} p_0(x) &= p(x), \\ p_1(x) &= p'(x), \\ p_i(x) &= -\mathrm{rem}(p_{i-2}(x), p_{i-1}(x)), \quad 2 \le i \le k, \\ p_k(x) &= \text{non-zero constant,} \end{aligned}$$

where $'$ means (formal) derivative. This is precisely the Euclidean remainder sequence that was discussed at length in the context of computing gcds, except that the remainder is given a negative sign (*which is crucial*). I will discuss the theory in terms of polynomials over $\mathbb{Q}$, so "remainder" means remainder in a division over $\mathbb{Q}$, but in fact it will only be the signs of the polynomials in the sequence that are required, so in practice any kind of PRS can be used as long as care is taken to preserve the correct signs of the elements. Hence all of the previous discussion of PRSs is relevant to Sturm sequences also; the sign change is allowed for in the similarity coefficients.

The relationship between Sturm and Euclidean sequences shows why the sequence must terminate with a constant element: this will be a gcd of p and p', and because p has been assumed to be squarefree it will have no repeated factors and hence no multiple roots, so this gcd must be a constant. (The value of the constant will depend on exactly what PRS is used, but for a primitive PRS over $\mathbb{Q}$ it must be 1.)

By the effect of differentiating and the definition of remainder, each element of the Sturm sequence must have degree at least one less than the previous element, i.e. $\deg p_i(x) \le \deg p_{i-1}(x) - 1$, hence $0 = \deg p_k(x) \le \deg p(x) - k \le$. In other words, the number of elements of the sequence in addition to p is $\le \deg p$.

The Sturm sequence provides a way of determining how many real roots lie within some interval by counting sign variations. Let a be a real number that is not a zero of p. Then define the *variation* at a of p, written $V(p, a)$, to be the number of variations of sign in the elements of the Sturm sequence evaluated at a, i.e. the number of times that the numbers $p(a), p'(a), p_2(a), \ldots, p_k(a)$ change sign (ignoring any zeros). More formally,

$V(p,a)$ is the number of values of i such that $p_i(a)p_j(a) < 0$ and $p_l(a) = 0$ for $1 \le i < l < j \le k$, where the inequality is strict.

Theorem 5 (Sturm) *If a and b are two real numbers that are not zeros of $p(x)$, such that $a < b$, then the number of real zeros of $p(x)$ in the interval $(a, b]$ is $V(p,a) - V(p,b)$.*

Proof This proof is based on that given by Knuth in *Seminumerical Algorithms.* It amounts to showing that as x varies, only changes in the sign of $p(x)$ itself as it passes through zeros affect the number of sign variations in the Sturm sequence of p.

Recall that, whatever PRS is used, the $p_i(x)$ must satisfy a recurrence relation of the general form

$$\alpha_i p_i(x) = p_{i+1}(x)q_i(x) - \beta_i p_{i+2}(x), \quad 2 \le i \le k,$$

where $\alpha_i, \beta_i > 0$ to preserve the signs, and $p_k(x)$ is a non-zero constant.

From the definition of the PRS, $\{p_i(a)\}$ cannot contain two successive zeros, because then all subsequent elements would have to be zero also, but $p_k(a) \ne 0$. Moreover, if any $p_{i+1}(a) = 0$ then $p_i(a)$ and $p_{i+2}(a)$ must have opposite signs. Hence a zero can only appear in the "sign sequence" as "$+, 0, -$" or "$-, 0, +$".

Consider the changes in $V(p,x)$ as x increases. The polynomials $p_i(x)$ have finitely many zeros and $V(p,x)$ changes only when x encounters such zeros. As x passes through a zero of $p_i(x), i > 0$ then, by continuity and the above restrictions, the sign sequence around this element can only change from "$+, \pm, -$" through "$+, 0, -$" to "$+, \mp, -$" or from "$-, \pm, +$" through "$-, 0, +$" to "$-, \mp, +$", both of which give no change in the number of sign variations. If $p(x)$ itself passes through zero as x *increases* then the signs of $p(x), p'(x)$ can only change from "$+, -$" through "$0, -$" to "$-, -$" or from "$-, +$" through "$0, -$" to "$+, +$", both of which *decrease* the number of sign variations by 1. Hence the total change in the number of sign variations as x increases from a to b, namely $V(p,b) - V(p,a)$, is a *decrease* equal to the number of real zeros of p passed. □

In the limit $x \to \pm\infty$, $p(x) \to$ its leading term, hence $V(p,\infty)$ is the number of sign variations of the leading coefficients of the Sturm sequence, and $V(p,-\infty)$ is the number of sign variations of the leading coefficients after negating elements of odd degree. Hence the total number of zeros, or the number of positive or negative zeros, can be found.

Here is a simple example (which I originally produced using REDUCE interactively) of the Sturm sequence of a simple polynomial whose zeros are obvious, in which I first compute the remainders over $\mathbb{Q}$ and then reduce them to primitive remainders over $\mathbb{Z}$, to keep the coefficients as simple as possible. I then evaluate the Sturm sequence at points that I know are between zeros, count the sign variations, and show that these counts agree with Sturm's theorem.

$$\begin{aligned}
p_0 &= (x-1)(x-3)(x-5) = x^3 - 9x^2 + 23x - 15 \\
p_1 &= p_0' = 3x^2 - 18x + 23 \\
p_2 &= -\mathrm{rem}(p_0, p_1) = \tfrac{8}{3}(x-3) \to x - 3 \\
p_3 &= -\mathrm{rem}(p_1, p_2) = 4 \to 1
\end{aligned}$$

Hence the primitive Sturm sequence is:

$$S(x) = \{p_0, p_1, p_2, p_3\} = \{x^3 - 9x^2 + 23x - 15, 3x^2 - 18x + 23, x - 3, 1\}.$$

The Sturm sequence and its variations $V(x)$ at specified values of x, namely 0, 2, 4, 6, are:

$$\begin{aligned}
S(0) &= \{-15, 23, -3, 4\}, \quad V(0) = 3 \\
S(2) &= \{3, -1, -1, 4\}, \quad V(2) = 2 \\
&\Rightarrow V(0) - V(2) = 1 \text{ real roots in } (0, 2] \\
S(4) &= \{-3, -1, 1, 4\}, \quad V(4) = 1 \\
&\Rightarrow V(2) - V(4) = 1 \text{ real roots in } (2, 4] \\
&\Rightarrow V(0) - V(4) = 2 \text{ real roots in } (0, 4] \\
S(6) &= \{15, 23, 3, 4\}, \quad V(6) = 0 \\
&\Rightarrow V(4) - V(6) = 1 \text{ real roots in } (4, 6] \\
&\Rightarrow V(0) - V(6) = 3 \text{ real roots in } (0, 6]
\end{aligned}$$

We now know how to determine exactly how many real zeros lie in any given interval. If there is one zero in an interval then the interval is an isolating interval, if there are no zeros then the interval can be discarded, and if there are more than one zeros then the interval can be divided into smaller intervals that are then tested again. Normally, intervals containing more than one zero are bisected, because asymptotically this is optimal. But in order to be able to guarantee to find the complete set of isolating intervals in a finite number of steps we need a *finite* starting interval that is guaranteed to contain all the real roots.

4 Polynomial root bounds

A suitable starting interval for root isolation is given by an upper bound on the magnitude of all the real roots of a polynomial. Three such bounds are quoted without proof in DST. Let $p(x) = \sum_{r=0}^{n} a_r x^r$, $n > 0$, and let α be a zero, i.e. $p(\alpha) = 0$.

Proposition 6 (Cauchy, 1829)

$$|\alpha| < 1 + \max_{r=0}^{n-1} \left| \frac{a_r}{a_n} \right|$$

This formulation and the following proof are based on the paper by Mignotte in Buchberger, Collins & Loos.

Proof If $|\alpha| \leq 1$ then the proposition is trivially true, so suppose $|\alpha| > 1$. If $p(x) = 0$ then

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

By the triangle inequality,

$$\begin{array}{rcl} |a_n||x|^n & \leq & |a_{n-1}||x|^{n-1} + \cdots + |a_0| \\ & \leq & \max_{r=0}^{n-1} |a_r|(|x|^{n-1} + \cdots + 1). \end{array}$$

Factoring out the maximum magnitude of the coefficients converts the right side into a geometric series, which can be summed to give

$$|x|^{n-1} + \cdots + 1 = \frac{|x|^n - 1}{|x| - 1} < \frac{|x|^n}{|x| - 1}.$$

Hence

$$|a_n||x|^n < \max_{r=0}^{n-1} |a_r| \frac{|x|^n}{|x| - 1}.$$

Cross-multiplying the denominator, which satisfies $|x| - 1 > 0$ by supposition, and re-organizing gives the required result. □

Because this bound depends on ratios of coefficients it is invariant under rescaling of $p(x)$ as a whole, which is an obvious condition to expect a bound to satisfy. But if the polynomial $q(y) = \sum_{r=0}^{n} b_r x^r$ is derived from $p(x)$ by rescaling each coefficient so that $b_r = a_r/\gamma^r$ then a zero at $x = \alpha$ of $p(x)$ becomes a zero at $y = \gamma\alpha$ of $q(y)$, i.e. the actual roots scale *linearly* with

γ. However, the above root bound rescales by γ^r for some r in the range $1 \leq r \leq n$ determined by the term for which the maximum used in the bound is attained. This scale dependence is clearly unsatisfactory.

The coefficient ratio $\frac{a_r}{a_n}$ rescales to $\frac{a_r/\gamma^r}{a_n/\gamma^n} = \gamma^{n-r}\frac{a_r}{a_n}$, and if this is to be linear in γ then each coefficient ratio must appear raised to the power $1/(n-r)$. This requirement is met by the next two bounds.

Proposition 7 (Cauchy, 1829)

$$|\alpha| \leq \max_{r=0}^{n-1} \left| \frac{na_r}{a_n} \right|^{\frac{1}{n-r}}$$

Proof As for the previous proof,

$$\begin{aligned} |a_n||x|^n &\leq |a_{n-1}||x|^{n-1} + \cdots + |a_0| \\ &\leq n \max_{r=0}^{n-1}(|a_r||x|^r). \end{aligned}$$

Then there exists an r such that

$$|x|^n \leq n \left| \frac{a_r}{a_n} \right| |x|^r$$

and hence

$$|x|^{n-r} \leq n \left| \frac{a_r}{a_n} \right|, \quad |x| \leq \left| \frac{na_r}{a_n} \right|^{\frac{1}{n-r}}.$$

Therefore the maximum value of the right side of the last inequality provides an upper bound on $|\alpha|$. □

Proposition 8 (Knuth, 1969)

$$|\alpha| \leq 2 \max_{r=0}^{n-1} \left| \frac{a_r}{a_n} \right|^{\frac{1}{n-r}}$$

Proof The bound is trivially satisfied if $x = 0$, so assume $x \neq 0$. As for the previous proof,

$$|x|^n \leq |a'_{n-1}||x|^{n-1} + \cdots + |a'_0|,$$

where $a'_r = a_r/a_n$, and hence, by dividing out $|x|^n$,

$$1 \leq |a'_{n-1}/x| + |a'_{n-2}/x^2| + \cdots + |a'_0/x^n| \leq t + t^2 + \cdots + t^n$$

where t has the smallest possible value such that for all r, $1 \leq r \leq n$,

$$t^r \geq |a'_{n-r}/x^r| \text{ which implies that } t \geq |a'_{n-r}|^{1/r}/|x|,$$

i.e.

$$t = \frac{1}{|x|} \max_{1 \leq r \leq n} |a'_{n-r}|^{1/r}.$$

Then

$$1 \leq t + t^2 + \cdots + t^n < t + t^2 + \cdots.$$

The infinite series sums formally to $t/(1-t)$, leading to the inequality $1 < t/(1-t) \Rightarrow t > 1/2$. Hence

$$\frac{1}{|x|} \max_{1 \leq r \leq n} |a'_{n-r}|^{1/r} > \frac{1}{2}$$

and the stated bound follows. [However, I am unhappy about the convergence aspect of this proof. Suggestions will be welcome!] □

This bound is the one recommended by Collins & Loos in Buchberger, Collins & Loos. Moreover, numerical experiments[1] suggest that Knuth's bound is best overall; in the tests it was often significantly better and never significantly worse than both of Cauchy's bounds.

The publication status of this bound is strange. It was published in the first edition of Knuth's *Seminumerical Algorithms* in 1969 on page 398 as exercise 20 to §4.6.2 on factorization of polynomials, together with a sketch solution on page 546. However, all reference to the bound appears to have been removed from the second edition published in 1981, despite the fact that the bound appears to be a good one. DST wrongly attribute this bound to Knuth's second edition in 1981. Knuth also gives another related but more complicated bound that he attributes to H. Zassenhaus, which I have never seen referenced elsewhere.

5 Strategy for polynomial root location

Given a univariate polynomial over $\mathbb{Z}$ or $\mathbb{Q}$ the first step is to perform a squarefree decomposition. Then each squarefree factor can be solved, and the multiplicity of the factor attached to each of its roots once they have

[1] Barbara C. Davies, *Computation of all the real roots of a real polynomial in REDUCE*, M.Sc. Thesis, Queen Mary College, University of London, 1989, Appendix 1.

been located. For each squarefree factor, a suitable root bound is computed, probably using Knuth's formula. It may then be convenient to round this up to the nearest power of 2 to facilitate interval bisection. Sturm's theorem can then be used to determine the number of real roots in the interval, which if necessary is bisected and each half re-examined, until a complete set of isolating intervals has been found. More precise algorithms are given by Collins & Loos, and by DST. If desired, each isolating interval can be reduced in width until each root is bracketed to within any chosen accuracy.

The reason why such exact algorithms may be necessary, rather than the more conventional algorithms using various numerical approximations, is that the roots of a polynomial may be very unstable to perturbation of the coefficients. This is particularly true of real roots, which can coalesce, turn complex and so disappear under perturbation. A classic example of this due to Wilkinson (1959) is quoted by DST. It is also possible for polynomial roots to be distinct but very close together, which makes them very hard to locate using numerical methods. There are formulae for the minimum root separation – see the paper by Mignotte in Buchberger, Collins & Loos for an example.

The methods described above, and related methods, provide the only methods I know of that can find *all real roots* with total reliability. Essentially the above Sturm sequence algorithm is implemented in REDUCE 3.4, together with extensions to find complex roots, as a package (called ROOTS) written by Stan Kameny. It is called automatically by the equation solver (SOLVE) if the rounded-real number domain is selected, although it does not return interval solutions to the user, but rather a specific value within each interval.

6 Exercises

The single assessed question in this set of exercises is the first.

1. (** **Assessed** **)
Let $p(x) = (x-1)(2x+1)(3x-1) = 6x^3 - 5x^2 - 2x + 1$.

(a) Compute the root bound for p given by all three formulae. Which is best?

(b) Compute a Sturm sequence for p (using whatever precise definition you prefer, e.g. Euclidean, primitive, subresultant, ...).

(c) Use a root bound together with Sturm's Theorem to compute isolating intervals for all the real roots of p.

2. Use Bareiss elimination to reduce the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 5 & 4 & 3 \\ 4 & 5 & 7 \end{pmatrix}$$

to upper triangular form, and hence using the formula given in the text find its determinant. Compute also the determinant of the triangular form, and by comparing the effect of the Bareiss elimination steps with those of Gaussian elimination relate this determinant to the determinant of the original matrix, and hence again compute the latter.

3. Compute the subresultant chain for the polynomials (with indeterminate coefficients) $a = a_2x^2 + a_1x + a_0$ and $b = b_1x + b_0$. Hence verify Habicht's Theorem. Repeat the computation with the explicit polynomials $a = 2x^2 + x$ and $b = 3x + 1$. Hence verify the Subresultant Theorem, and that the Subresultant PRS requires only integer (not rational) computations.