MATHEMATICS OF COMPUTATION, VOLUME 25, NUMBER 113, JANUARY, 1971

Error Analysis of the Algorithm for Shifting the Zeros of a Polynomial by Synthetic Division

By G. W. Stewart III*

Abstract. An analysis is given of the role of rounding errors in the synthetic division algorithm for computing the coefficients of the polynomial $g(z) = f(z + s)$ from the coefficients of the polynomial f. It is shown that if $|z + s| \cong |z| + |s|$ then the value of the computed polynomial $g^*(z)$ differs from $g(z)$ by no more than a bound on the error made in computing $f(z + s)$ with rounding error. It may be concluded that well-conditioned zeros of f lying near s will not be seriously disturbed by the shift.

1. Introduction. Let the polynomial

$$f(z) = a_0 + a_1 z + \cdots + a_n z^n$$

have the zeros $r_1, r_2, \cdots, r_n$. Then the polynomial

$$g(z) = f(z + s) = b_0 + b_1 z + \cdots + b_n z^n$$

has zeros $r_1 - s, r_2 - s, \cdots, r_n - s$. The coefficients of g may be evaluated by repeated synthetic division:

$$\begin{aligned} b_{n-i}^{(i-1)} &= a_{n-i}, && (i = 0, 1, \cdots, n), \\ b_n^{(k)} &= b_n^{(k-1)}, && (k = 0, 1, \cdots, n), \\ b_{n-i}^{(k+1)} &= b_{n-i}^{(k)} + s b_{n-i+1}^{(k)}, && (i = 1, 2, \cdots, k+1;\ k = 0, 1, \cdots, n-1). \end{aligned} \tag{1.1}$$

The coefficients of g are then given by $b_i = b_i^{(n)}$. This scheme is a rearrangement of the usual synthetic division algorithm; however, the two are computationally equivalent. The purpose of this note is to analyze the effects of rounding error on the algorithm defined by (1.1).

2. The Principal Result. We shall assume that all calculations are carried out in complex floating-point arithmetic. Specifically, let $fl(a \circ b)$ denote the result of executing the binary operation $\circ$ in floating-point arithmetic. Then, we shall assume that there is a number η such that

$$fl(a \pm b) = a\alpha \pm b\beta,$$

and

$$fl(ab) = ab\gamma,$$

Received August 19, 1968, revised April 7, 1970.

AMS 1970 *subject classifications.* Primary 65G05, 65H05.

Key words and phrases. Rounding error, shifting algorithm, synthetic division, zeros of polynomials.

* Oak Ridge Graduate Fellow from the University of Tennessee under appointment from the Oak Ridge Associated Universities. Oak Ridge National Laboratory is operated by Union Carbide Corporation for the U. S. Atomic Energy Commission.

where

$$|\alpha - 1|, \quad |\beta - 1|, \quad |\gamma - 1| \leqq \eta - 1. \tag{2.1}$$

For brevity, the following notational convention will be observed. A lower case Greek letter, say η, will denote a number presumed to be near unity, and

$$\hat{\eta} = \eta - 1.$$

In this notation, the bounds (2.1) become

$$|\hat{\alpha}|, \quad |\hat{\beta}|, \quad |\hat{\gamma}| \leqq \hat{\eta},$$

The results of this note are closely connected with the problem of evaluating the polynomial f in the presence of rounding error. This is usually done by synthetic division, and the value of $f(s)$ is given by $b_0^{(n)}$ in (1.1). The following theorem is well known [1, p. 50].

THEOREM 2.1. *Let $fl(f(s))$ denote the computed value of $b_0^{(n)}$. Then*

$$fl(f(s)) = a_0\alpha_0 + a_1\alpha_1 s + \cdots + a_n\alpha_n s^n,$$

where

$$|\hat{\alpha}_n| \leqq \eta^{2n} - 1$$

and

$$|\hat{\alpha}_i| \leqq \eta^{2i+1} - 1, \qquad (i = 0, 1, \cdots, n - 1).$$

Thus, the computed value of $f(s)$ is the exact value of a polynomial whose coefficients differ from those of f by small relative amounts.

COROLLARY 2.2. *Let*

$$f_a(z) = |a_0| + |a_1|x + \cdots + |a_n|z^n.$$

Then

$$|fl(f(s)) - f(s)| \leqq (\eta^{2n} - 1)f_a(|z|).$$

For the shifting algorithm, a nice result would be an analogue of Theorem 2.1 stating that the polynomial g computed with rounding error from (1.1) is the polynomial that would be obtained by applying (1.1) without rounding error to a polynomial slightly perturbed from f. This is not true in general. For example, if the algorithm is applied in four-digit decimal arithmetic to the polynomial

$$f(z) = z^2 - 427.8z - 1.610,$$

with shift $s = 416.9$, the result is the polynomial

$$g(z) = z^2 + 406.0z - 4546.$$

In fact, g is the polynomial obtained by applying (1.1) with $s = 416.9$ exactly to the polynomial

$$h(z) = z^2 - 427.8z - 2.210,$$

whose low order coefficient differs considerably from that of f.

However, the following analogue of Corollary 2.2 holds.

COROLLARY 3.2. *Let g denote the polynomial obtained when the algorithm* (1.1) *is carried out with rounding error. Then*

$$|g(z) - f(z + s)| \leqq (\eta^{2n} - 1)f_a(|z| + |s|).$$

Thus when

$$|z| + |s| \cong |z + s|, \tag{2.2}$$

that is, when z lies in the direction of the shift or when z is small, the error in $g(z)$ has the same bound as the error made in evaluating $f(z + s)$ with rounding error. This is true of the example given above. On the other hand, $g(-416.9) = -2.210$, which differs considerably from $f(0)$. This is to be expected, since $z = -416.9$ does not satisfy (2.2).

In conjunction with Rouché's theorem, Corollary 3.2 suggests that the shifting algorithm will not perturb zeros near the shift by much more than they would be perturbed by the act of rounding the coefficients of f.

3. The Principal Theorem. From the Eqs. (1.1) which define the shifting algorithm it follows that the $b_i^{(k)}$ satisfy the matrix equation

$$\begin{bmatrix} b_n^{(k+1)} \\ b_{n-1}^{(k+1)} \\ \vdots \\ b_{n-k}^{(k+1)} \\ b_{n-k-1}^{(k+1)} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ s & 1 & & \\ & \ddots & \ddots & \\ & & s & 1 \\ & & & s & 1 \end{bmatrix} \begin{bmatrix} b_n^{(k)} \\ b_{n-1}^{(k)} \\ \vdots \\ b_{n-k}^{(k)} \\ a_{n-k-1} \end{bmatrix}. \tag{3.1}$$

From this, it is seen that the vector $b^{(k)} = (b_n^{(k)}, b_{n-1}^{(k)}, \cdots, b_{n-k}^{(k)})^T$ may be obtained by premultiplying the vector $a^{(k)} = (a_n, a_{n-1}, \cdots, a_{n-k})^T$ by a unit lower triangular matrix L_k of order $k + 1$. The idea of the following error analysis is to show that the vector $b^{(k)}$, calculated with rounding error, may be obtained by multiplying $a^{(k)}$ by a perturbed matrix $L_k + G_k$, where the elements of G_k are small.

Let the (i, j)-element of L_k be $l_{ij}^{(k)}$, $(i, j = 1, 2, \cdots, k + 1)$. Let $l_{k+2,k+2}^{(k)} = 1$, and for all other $(i, j) \notin \{1, 2, \cdots, k + 1\} \times \{1, 2, \cdots, k + 1\}$, let $l_{ij}^{(k)} = 0$ (n.b., these last defined $l_{ij}^{(k)}$ are not elements of L_k). Then, from (3.1), it follows that

$$l_{ij}^{(k+1)} = l_{ij}^{(k)} + s l_{i-1,j}^{(k)}, \qquad (i, j = 1, 2, \cdots, k + 2).$$

Since

$$L_1 = \begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix},$$

it follows by an easy induction that

$$l_{ij}^{(k)} = s^{i-j} C(k - j + 1, i - j), \qquad (i, j = 1, 2, \cdots, k + 1).$$

Here $C(m, n)$ denotes the binomial coefficient $m!/[n!(m - n)!]$ and is assumed to be zero for $n > m$.

Suppose now that the $b_i^{(k)}$ represent computed values. Then

$$b_{n-i}^{(k+1)} = b_{n-i}^{(k)}\epsilon_i^{(k)} + sb_{n-i+1}^{(k)}\delta_i^{(k)}, \tag{3.2}$$
$$(i = 1, 2, \cdots, k+1; k = 0, 1, \cdots, n-1),$$

where

$$|\hat{\epsilon}_i^{(k)}| \leqq \hat{\eta}, \qquad |\hat{\delta}_i^{(k)}| \leqq \eta^2 - 1.$$

Let $\epsilon_i^{(k)}$, $\delta_i^{(k)} = 1$ when i and k fall outside the bounds in (3.2).

THEOREM 3.1. *Let $b^{(k)}$ denote the computed vector. Then*

$$b^{(k)} = (L_k + G_k)a^{(k)},$$

where the (i, j)-element of G_k is $l_{ij}^{(k)}\hat{\gamma}_{ij}^{(k)}$ and

$$\hat{\gamma}_{11}^{(k)} = 0,$$
$$|\hat{\gamma}_{i1}^{(k)}| \leqq \eta^{k+i-1} - 1, \qquad (i = 2, 3, \cdots, k+1), \tag{3.3}$$
$$|\hat{\gamma}_{ij}^{(k)}| \leqq \eta^{k+i-2j+2} - 1, \qquad (j = 2, 3, \cdots, k+1; i = j, j+1, \cdots, k+1).$$

Proof. The proof is by induction. Throughout the proof, the symbols ϵ and δ will be used generically for the $\epsilon_i^{(k)}$ and $\delta_i^{(k)}$.

For $k = 1$, define G_1 by

$$L_1 + G_1 = \begin{bmatrix} 1 & 0 \\ s\delta & \epsilon \end{bmatrix},$$

so that

$$\begin{bmatrix} b_n^{(1)} \\ b_{n-1}^{(1)} \end{bmatrix} = (L_1 + G_1)\begin{bmatrix} a_n \\ a_{n-1} \end{bmatrix}.$$

Moreover

$$G_1 = \begin{bmatrix} 0 & 0 \\ s\hat{\delta} & \hat{\epsilon} \end{bmatrix}.$$

Hence, the $\gamma_{ij}^{(1)}$ satisfy (3.3).

Assume that G_k is given and the $\gamma_{ij}^{(k)}$ satisfy (3.3). Consider the quantity

$$g_{ij}^{(k+1)} = l_{ij}^{(k)}\gamma_{ij}^{(k)}\epsilon_{i-1}^{(k)} + sl_{i-1,j}^{(k)}\gamma_{i-1,j}^{(k)}\delta_{i-1}^{(k)}, \qquad (i, j = 1, 2, \cdots, k+2).$$

Then, it is easily verified that the matrix $L_{k+1} + G_{k+1}$, whose (i, j)-element is $g_{ij}^{(k+1)}$, produces the vector $b^{(k+1)}$ when it premultiplies the vector $a^{(k+1)}$. Moreover, since

$$\arg(l_{ij}^{(k)}) = \arg(s^{i-j}) = \arg(sl_{i-1,j}^{(k)}),$$
$$g_{ij}^{(k+1)} = (l_{ij}^{(k)} + sl_{i-1,j}^{(k)})\gamma_{ij}^{(k+1)} = l_{ij}^{(k+1)}\gamma_{ij}^{(k+1)},$$

where

$$|\hat{\gamma}_{ij}^{(k+1)}| \leqq \max\{|\gamma_{ij}^{(k)}\epsilon - 1|, |\gamma_{i-1,j}^{(k)}\delta - 1|\}.$$

But, for $j = 1$,

$$|\gamma_{i1}^{(k)}\epsilon - 1| \leqq \eta^{k+i-1}\eta - 1 = \eta^{k+i} - 1,$$

and

$$|\gamma_{i-1,1}^{(k)}\delta - 1| \leqq \eta^{k+i-2}\eta^2 - 1 = \eta^{k+i} - 1.$$

For $j > 1$,

$$|\gamma_{ij}^{(k)}\epsilon - 1| \leqq \eta^{k+i-2j+2}\eta - 1 = \eta^{k+i-2j+3} - 1,$$

and

$$|\gamma_{i-1,j}^{(k)}\delta - 1| \leqq \eta^{k+i-2j+1}\eta^2 - 1 = \eta^{k+i-2j+3} - 1.$$

Hence, the $\gamma_{ij}^{(k+1)}$ satisfy (3.3). In particular

$$|\hat{\gamma}_{ij}^{(n)}| \leqq \eta^{2n} - 1.$$

This completes the proof of the theorem.

To establish Corollary 3.2, let $g(z)$ be the computed shifted polynomial. Then

$$\begin{aligned} g(z) &= \sum_{i=1}^{n+1} b_{n-i+1}z^{n-i+1} = \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} z^{n-i+1}l_{ij}^{(n)}\gamma_{ij}^{(n)}a_{n-j+1} \\ &= \sum_{j=1}^{n+1} a_{n-j+1}\sum_{i=j}^{n+1} z^{n-i+1}s^{i-j}C(n-j+1, i-j)\gamma_{ij}^{(n)} \\ &= \sum_{j=1}^{n+1} a_{n-j+1}(z+s)^{n-j+1} + \sum_{j=1}^{n+1} a_{n-j+1}\sum_{i=j}^{n+1} z^{n-i+1}s^{i-j}C(n-j+1, i-j)\hat{\gamma}_{ij}^{(n)} \\ &= f(z+s) + e(z), \end{aligned}$$

where

$$e(z) = \sum_{j=1}^{n+1} a_{n-j+1}\sum_{i=j}^{n+1} z^{n-i+1}s^{i-j}C(n-j+1, i-j)\hat{\gamma}_{ij}^{(n)}.$$

Hence

$$\begin{aligned} |e(z)| &\leqq (\eta^{2n} - 1)\sum_{j=1}^{n+1} |a_{n-j+1}|\sum_{i=j}^{n+1} |z|^{n-i+1}\,|s|^{i-j}\,C(n-j+1, i-j) \\ &= (\eta^{2n} - 1)\sum_{j=1}^{n+1} |a_{n-j+1}|\,(|z| + |s|)^{n-j+1} \\ &= (\eta^{2n} - 1)f_a(|z| + |s|). \end{aligned}$$

This proves Corollary 3.2 stated in the last section.

4. Acknowledgment. I am indebted to Professor A. S. Householder for his valuable comments and suggestions.

Computation Center
The University of Texas at Austin
Austin, Texas 78712

1. J. H. WILKINSON, *Rounding Errors in Algebraic Processing*, Prentice-Hall, Englewood Cliffs, N. J., 1963. MR **28** #4661