cs 55

A STOPPING CRITERION FOR POLYNOMIAL ROOT FINDING

BY

DUANE A. ADAMS

TECHNICAL REPORT NO. CS 55
FEBRUARY 10, 1967

Supported in part by the
Office of Naval Research and
National Science Foundation

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

1. Introduction

When locating the zeros of a polynomial, it is usually difficult to know just when to terminate the iteration process. It is desirable to terminate the process when the zero is known to within roundoff accuracy. Various ad hoc methods have been used as stopping criteria; however, such methods do not take into account particular properties of the polynomial being evaluated. Such properties might include the condition of the polynomial, multiple zeros, or clusters of zeros. In this paper a stopping criterion is presented which requires that the value of the polynomial be less than a calculated bound for the roundoff error.

Bounds for the roundoff error can be obtained by using the methods of range arithmetic [1] or interval arithmetic [4], but such methods require a large amount of computation. The algorithm described here produces similar bounds, and offers the advantage of being easily calculated, Kahan and Farkas [3] have used this algorithm to bound the roundoff error for a real polynomial evaluated at a real point, but they offer no motivation as to why the algorithm works. In this paper Kahan's bounds for a real polynomial evaluated at a real point are summarized, and then the analysis is extended to a real polynomial evaluated at a complex point. The use of this bound as a stopping criterion is discussed.

2. Summary of Results for a Real Polynomial Evaluated at a Real Point

This section contains a brief summary of Kahan's[1] results on bounding the roundoff error for a real polynomial evaluated at a real point. Consider the polynomial

$$P(Z) = a_o Z^n + a_1 Z^{n-1} + \ldots + a_n .$$

The Horner recurrence is given by

$$\begin{aligned} b_o &= a_o \\ b_k &= x \cdot b_{k-1} + a_k \; ; \; k = 1,\ldots n . \end{aligned} \tag{1}$$

The last term of this recurrence, b_n , is $P(x)$.

Associated with (1) we have a second recurrence, given by

$$\begin{aligned} e_o &= |a_o|\pi/(\pi + a) \\ e_k &= |x|e_{k-1} + |b_k| \; ; \; k = 1,\ldots n . \end{aligned} \tag{2}$$

In (2), b_k represents the calculated quantity and π and σ are the maximum absolute rounding errors which take place in any single product or sum, respectively. We have

$$\pi \leq \tfrac{1}{2} \beta^{1-t} \quad , \quad \sigma \leq \tfrac{1}{2} \beta^{1-t} \quad ,$$

where β represents the base in which the machine floating point arithmetic is performed and t represents the number of digits in the mantissa. For the Burroughs B5500, an octal machine with a 39 bit mantissa, we have $\beta = 8$ and $t = 13$.

---

1. Lectures presented by Professor Kahan at Stanford University, Spring 1966.

Kahan shows that a bound for the roundoff error is given by

$$|P(x) - b_n| < (a + \pi)e_n - |b_n|\pi \quad . \tag{3}$$

He also shows that a suitable stopping criterion for having found a zero of P to within the bounds given for the roundoff error is

$$|b_n| \leq 2E \quad ,$$

$$E = (a + \pi)e_n - |b_n|\pi \quad . \tag{4}$$

The reason for having the factor of 2 in (4) is to guarantee that there is a computer representable number which satisfies (4). Note that the above criterion does not tell us how close we are to a zero, but only that we are in some interval about the zero where roundoff error may be dominating our calculations.

3. Rounding Error Bounds for a Real Polynomial Evaluated at a Complex Point

Now suppose that

$$P(Z) = a_o Z^n + a_1 Z^{n-1} + \ldots + a_n$$

is a polynomial with real coefficients a_i , but that we wish to evaluate the polynomial at a point $Z = x + i \cdot y$, By taking a quadratic factor out of the polynomial and then equating coefficients with the original polynomial, we can obtain the well known recurrence for evaluating this polynomial at a point in the complex plane which involves only real arithmetic and a total of 2.n multiplications. Thus if we write

$$P(Z) = (Z^2 + pZ + q)(b_o Z^{n-2} + b_1 Z^{n-3} + \ldots + b_{n2}) + R(Z - x) + S$$

we obtain for the recurrence

$$
\begin{aligned}
b_o &= a_o \\
b1 &= a_1 - p\cdot b_o \\
b_k &= a_k - p\cdot b_{k-1} - q\cdot b_{k-2} \;;\; k = 2,\ldots n-1 \\
b_n &= a_n + x\cdot b_{n\,1} - q\cdot b_{n-2}
\end{aligned} \tag{5}
$$

where

$$p = -2x \;,\; q = x^2 + y^2 \;,\; b_{n-1} = R \text{ and } b_n = S \;.$$

Note that

$$Rl(P(x + iy)) = b_n \quad \text{and} \quad Im(P(x + iy)) = y\cdot b_{n-1} \;.$$

The coefficients a_i which appear in the machine may not be identical to the coefficients of the original problem because of the error in converting from decimal to binary. We shall not be concerned with this error, but rather with the errors which accumulate in attempting to evaluate the polynomial represented in the machine.

The elements of the recurrence (5), as represented within the machine, are given by

$$
\begin{aligned}
b_o &= a_o \\
b_1 &= (a_1 - \bar{p}\cdot b_o(1 + \pi_{11}))/(1 + \sigma_{11}) \\
b_k &= ((a_k - \bar{p}\cdot b_{k-1}(1 + \pi_{1k}))/(1 + \sigma_{1k}) - \bar{q}\cdot b_{k2}(1+\pi_{2k}))/(1+\sigma_{2k}) \\
&\qquad k = 2,\ldots n-1 \\
b_n &= ((a_n + x\cdot b_{n-1}(1 + \pi_{1n}))/(1+ \sigma_{1n}) - \bar{q}\cdot b_{n-2}(1 + \pi_{2n}))/(1 + \sigma_{2n}) \;,
\end{aligned} \tag{6}
$$

We can bound each of the quantities σ_{ij} and π_{ij} on the basis of the floating point arithmetic of the computer being used, that is,

$$|a_{ij}| \leq \tfrac{1}{2}\beta^{1-t} \quad , \quad |\pi_{ij}| \leq \tfrac{1}{2}\beta^{1-t} \quad .$$

In (6), $\bar{p}$ and $\bar{q}$ represent the true values, but will actually have rounding errors associated with their calculations. For the sake of simplifying the analysis slightly, let us assume that $\bar{q}$ is calculated in double precision and then rounded to single precision. Then we may write

$$p = \bar{p}/(1+\pi_p) \text{ and } q = \bar{q}/(1+\sigma_q) = (x^2 + y^2)/(1 + \sigma_q) \quad ,$$

where

$$|\pi_p| \leq \tfrac{1}{2}\beta^{1-t} \quad , \quad |\sigma_q| \leq \tfrac{1}{2}\beta^{1-t} \quad .$$

In practice this double precision calculation is not necessary,

Solving for the a_i in (6) we find

$$\begin{aligned}
a_o &= b_o \\
a_1 &= b_1(1+\sigma_{11}) + p\cdot b_o(1+\pi_p)(1+\pi_{11}) \\
a_k &= b_k(1+\sigma_{1k})(1+\sigma_{2k}) + q\cdot b_{k-2}(1+\sigma_q)(1+\pi_{2k})(1+\sigma_{1k}) \\
&\quad + p\cdot b_{k-1}(1+\pi_p)(1+\pi_{1k}) \qquad ; \quad k = 2,\ldots n-1 \\
a_n &= b_n(1+\sigma_{1n})(1+\sigma_{2n}) + q\cdot b_{n-2}(1+\sigma_q)(1+\pi_{2n})(1+\sigma_{1n}) \\
&\quad - x\cdot b_{n-1}(1+\pi_{1n}) \quad .
\end{aligned} \tag{7}$$

Note that in (7), and for the rest of this analysis, the letters a_i, b_i, p, and q shall represent the numbers within the machine, and any deviation from the true values is represented by the error bounds,,

By substituting the a_i of (7) into P and simplifying we find

$$P(z) = b_n + i\cdot y\cdot b_{n\,1} - x\cdot b_{n-1}\pi_n + \sum_{k=1}^{n} \sigma_k b_k Z^{n-k} + \sum_{k=0}^{n-2} b_k(\pi_{k+1}\cdot p/Z + \omega_{k+2}\,\bar{Z}/Z)\cdot Z^{n-k}$$

where

$$1 + \sigma_1 = (1 + \sigma_{11}) \quad ; \quad |\sigma_1| \le \tfrac{1}{2}\beta^{1-t}$$

$$1 + \pi_n = (1 + \pi_{1n}) \quad ; \quad |\pi_n| \le \tfrac{1}{2}\beta^{1-t}$$

$$1 + \sigma_k = (1 + \sigma_{2k})(1 + \sigma_{1k}) \quad ; \quad |\sigma_k| \le \beta^{1-t}$$

$$1 + \pi_k = (1 + \pi_{1k})(1 + \pi_p) \quad ; \quad |\pi_k| \le \beta^{1-t}$$

$$1 + \omega_k = (1 + \sigma_q)(1 + \pi_{2k})(1 + \sigma_{1k}) \quad ; \quad |\omega_k| \le \tfrac{3}{2}\beta^{1-t} .$$

Recalling that the calculated value of the polynomial as given by the recurrence is

$$b_n + i\cdot y\cdot b_{n\,1}$$

we have

$$\begin{aligned} |P(x + i\cdot y) - (b_n + i\cdot y\cdot b_{n-1})| \\ \le \pi|x|\cdot|b_{n-1}| + \sigma(|b_n| + |b_{n\,1}|\cdot|Z|) \\ + (\pi|p|/|Z| + \omega|\bar{Z}|/|Z|)|b_o|\cdot|Z^n| \\ + (\sigma + \pi|p|/|Z| + \omega|\bar{Z}|/|Z|)\sum_{k=1}^{n-2}|b_k Z^{n-k}| , \end{aligned}$$

where

$$\sigma = \max_i |\sigma_i| , \quad \pi = \max_i |\pi_i| , \quad \omega = \max_i |\omega_i| .$$

Now choose

$$\begin{aligned} e_o &= |b_o|(2\pi + \omega)/(2\pi + \omega + a> \\ e_k &= |Z|\, e_{k-1} + |b_k| \; ; \; k = 1,\ldots n . \end{aligned} \tag{8}$$

Hence

$$|b_o| = (2\pi + \omega + \sigma)\cdot e_o/(2\pi + \omega)$$

$$|b_k| = e_k - |Z|\cdot e_{k\,1} ,$$

and upon substituting into the above and simplifying we obtain

$$
\begin{aligned}
&|P(x + iy) - (b_n + iy \cdot b_{n-1})| \leq (2\pi + \omega + \sigma)e_n \\
&\quad - (2\pi + \omega)(|b_n| + |b_{n-1}||Z|) + \pi |x||b_{n-1}| ,
\end{aligned} \tag{9}
$$

where

$$
\pi \leq \beta^{1-t} \quad , \quad \sigma \leq \beta^{1-t} \quad , \quad \omega \leq \frac{3}{2} \beta^{1-t} \quad .
$$

The formula given in (9) is a generalization of the formula given in (3). To complete the parallelism between the real and the complex cases, we give the stopping criterion for having found a complex zero. A zero has been found to within roundoff accuracy when

$$
\begin{aligned}
&|b_n + i \cdot y \cdot b_{n-1}| < E , \\
&E = (2\pi + \omega + \sigma)e_n - (2\pi + \omega)(|b_n| + |b_{n-1}||Z|) + \pi|x||b_{n-1}| .
\end{aligned} \tag{10}
$$

The following section contains a discussion of the acceptability of this criterion.

4. Use of Error Bound as Stopping Criterion

We may think of the E as given in (10) as defining a region about a zero in the complex plane, such that for the set of all machine representable points in this region the stopping criterion in (10) is satisfied. For each zero j of P let Ω_j denote the region defined by E . If indeed our error bound is a good one, then we will not be able to distinguish any of the points in Ω_j from the true zero Zj on the basis of calculated function values, for any non-zero quantities will only represent "noise". In general, E will define a larger region than the ideal one just described. We have made rather

extensive tests to see how the bound given in (10) compares with the actual roundoff errors. Included in our tests have been the polynomials given in Table 1 and Table 2 of Henrici [2].

The zeros of these polynomials were determined using the method suggested by Traub [5],[6]. The iteration process was terminated when (10) was satisfied, After all of the zeros of each polynomial had been located, they were then reevaluated in the original polynomial , in both single and double precision, and any zeros which did not satisfy (10) were purified. The roundoff error is then the difference between the evaluations in single and double precision.

Figure 1 shows a distribution of the ratios of roundoff error to the roundoff error bound when (10) was first satisfied for each zero, These calculations were performed on a Burroughs B5500, an octal machine, and hence the error bound contains an additional factor of 4 over that of a binary machine to account for the worst case where a rounding operation can cause a change in the exponent. From Figure 1 we see that in nearly 85% of our examples the roundoff error is bigger than 0.01 times the error bound, and this we feel is a reasonable bound for the error.

The distribution shown in Figure 1 tells us how the roundoff error compares with the error bound, but not how close we are to a zero of P. When (10) is satisfied we only know that we are within the region Ω_j. However, our analysis of the data shows that in the majority of the examples we have tested, we are sufficiently close to the zero when the stopping criterion is satisfied, that even one more iteration is unwarrented. In performing the extra iteration either no change occurs, there is a perturbation in the roundoff error but the answer is not

improved, or the answer is improved by 2 or 3 units in the last decimal.

In referring to the region Ω_j about each zero, we have not dealt with the case where Ω_j may be empty, If there is no machine representable number which satisfies the error bound, then the algorithm would search endlessly for such a value unless terminated after a certain number of steps, We have not been able to prove that there always exists a machine number which satisfies (10). On the other hand, we have not found an example where there is no such number, For the real case, Kahan has shown that by doubling the error bound, it is always satisfiable, For the complex case it can probably be shown that for some small multiple of the error bound, there is always a machine representable number which satisfies the bound, However, we have not shown this.

5. Conclusions

The stopping criterion given in (10) serves as a very adequate means of determining when a complex zero of a real polynomial has been obtained to within roundoff accuracy, The bound for the roundoff error used in (10) is easily calculated as the polynomial is evaluated by using the recurrence given in (8). Little is to be achieved by iterating beyond the stopping criterion, An open question at present is whether or not there always exists a machine representable number which satisfies (10).

Acknowledgements: I wish to thank Professor J. F. Traub for suggesting this problem to me and for his patience in reading the report and suggesting improvements, Also I wish to thank Professor W. Kahan for his many helpful suggestions,

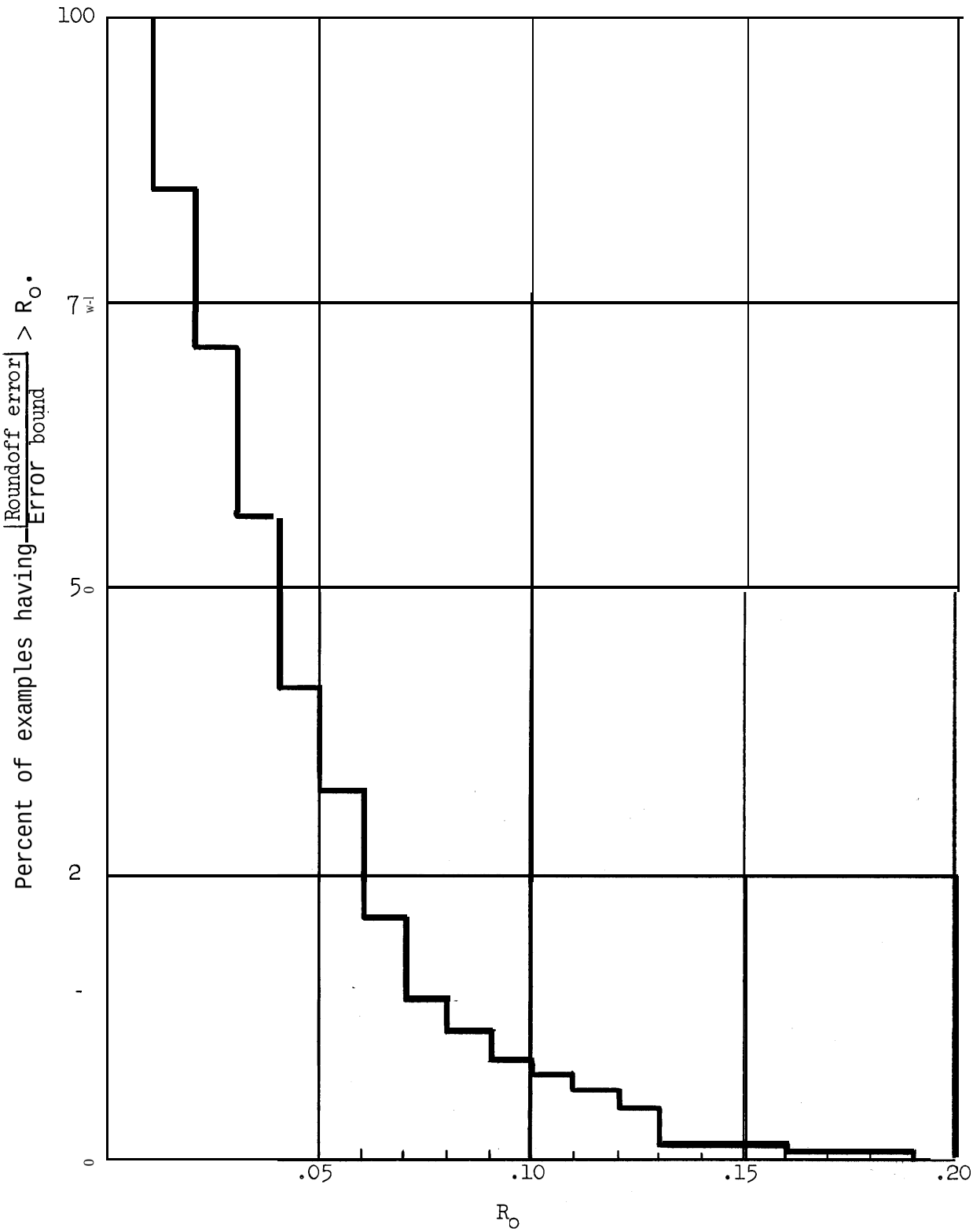

Fig. 1 Distribution of the ratio of roundoff error to error bound.

References

[1] Gibb, Allan. Algorithm 61 Procedures for Range Arithmetic. Comm. ACM 4(1961), 319-320.

[2] Henrici, P. and Watkins, Bruce O. Finding Zeros of a Polynomial By the Q-D Algorithm. Comm. ACM 8 (Sept 1965), 572-573. with corrections given by Thomas, Richard F. Jr,, Corrections to Numerical Data on Q-D Algorithm. Comm, ACM 9 (May 1966), 322.

[3] Kahan, W. and Farkas, I. Algorithm 168 and Algorithm 169. Comm. ACM 6 (Apr 1963), 165.

[4] Moore, R. E. Interval Arithmetic and Automatic Error Analysis in Digital Computing. Stanford University Applied Mathematics and Statistics Laboratories Technical Report No. 25. (Nov 1962) 134 pp.

[5] Traub, J. F. A Class of Globally Convergent Iteration Functions for the Solution of Polynomial Equations, Math. Comp. 20 (1966), 113-138.

[6] _____ The Calculation of Zeros of Polynomials and Analytic Function:. To Appear in Proceedings of a Symposium on Mathematical Aspects of Computer Science, American Mathematical Society, Providence, Rhode Island. Also available as Tech. Rept. 36, Computer Science Dept., Stanford University.