MATHEMATICS OF COMPUTATION, VOLUME 28, NUMBER 127, JULY 1974, PAGES 789-793

Error Analysis for Polynomial Evaluation*

By A. C. R. Newbery

Abstract. A floating-point error analysis is given for the evaluation of a real polynomial at a real argument by Horner's scheme. A computable error bound is derived. It is observed that when a polynomial has coefficients of constant sign or of strictly alternating sign, one cannot expect better accuracy by reformulating the problem in terms of Chebyshev polynomials.

Given that the real polynomial $P(x) \equiv \sum_0^n p_r x^r$ is to be evaluated at a real argument α under conditions of normalized floating-point arithmetic, we wish to study the bounds for accumulated round-off effects. Two algorithms for evaluation will be compared: (i) the standard Horner scheme and (ii) Clenshaw's algorithm [1] applied to the Chebyshev form.

First, we note that for practical purposes there is no loss of generality in assuming that $|\alpha| \leqq 1$. For instance, on a binary machine one could define $H(x) \equiv P(2^k x)$, and, in the absence of overflow/underflow, the coefficients of H are exactly determined in terms of those of P. The problem of evaluating $P(\alpha)$ can be replaced by that of evaluating $H(x)$ at $x = 2^{-k}\alpha$, where $|2^{-k}\alpha| \leqq 1$. The two computations give rise to identically the same sequence of significands; only the exponents may differ. Bauer [2] made an analogous observation with respect to the scaling of linear equation problems. In order to have a fair basis for comparison with Clenshaw's algorithm, we shall assume the problem has been normalized so that $|\alpha| \leqq 1$.

According to the Horner scheme, we have

$$(1)\quad P(\alpha) = q_0, \quad \text{where } q_n = p_n, \text{ and } q_r = p_r + \alpha q_{r+1}, \qquad r = n-1, n-2, \ldots, 0.$$

Computationally, since the result of each arithmetical operation in (1) is subject to a relative error in the range $\pm\varepsilon$, we shall generate a sequence $\{q_r^*\}$ given by

$$(2)\quad \begin{aligned} &q_n^* = p_n, \\ &q_r^* = p_r + \alpha q_{r+1}^* + \delta_r, \ r = n-1, n-2, \ldots, 0, \\ &\text{where } \delta_r \text{ denotes the difference between the floating-point and} \\ &\text{true evaluation of } p_r + \alpha q_{r+1}^*. \end{aligned}$$

It can be verified that

Received May 31, 1973.

AMS (MOS) subject classifications (1970). Primary 65G05.

Key words and phrases. Error analysis, polynomials.

*Much of the work on this paper was done while the author was on leave at the Oxford University Computing Laboratory, Oxford, England.

$$|\delta_r| \leqq \varepsilon|p_r| + \sigma|\alpha q_{r+1}^*|, \quad \text{where } \sigma = \varepsilon(2+\varepsilon). \tag{3}$$

Proceeding as in [3], [4], we can write (1), (2) in matrix form:

$$A\bar{q} = \bar{p}, \qquad A\bar{q}^* = \bar{p} + \bar{\delta}, \text{ where}$$

$$A = \begin{bmatrix} 1 & & & \\ -\alpha & 1 & & \\ & \ddots & \ddots & \\ & & -\alpha & 1 \end{bmatrix} \text{ and } A^{-1} = \begin{bmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha^2 & \alpha & 1 & & \\ \vdots & & & \ddots & \\ \alpha^n & \alpha^{n-1} & \alpha^2 & \alpha & 1 \end{bmatrix}. \tag{4}$$

The evaluation error E is given by

$$E = |q_0^* - q_0| = \left|\sum_0^n \delta_r \alpha^r\right| \leqq \sum_0^n |\delta_r \alpha^r|. \tag{5}$$

Combining this with (3) and noting that $q_{n+1}^* = 0$, we find

$$E \leqq \varepsilon \sum_0^n |p_r \alpha^r| + \sigma \sum_1^n |q_r^* \alpha^r|. \tag{6}$$

In order to find a bound for $\sum_1^n |q_r^* \alpha^r|$, we define a vector $\bar{g}_r = \{\alpha^n, \alpha^{n-1}, \ldots, \alpha^r, 0, 0, \ldots, 0\}$, and we note from (4) that the successive row-vectors of A^{-1} are $\bar{g}_n/\alpha^n$, $\bar{g}_{n-1}/\alpha^{n-1}, \ldots, \bar{g}_0/\alpha^0$. Hence

$$|q_r^* \alpha^r| = |\bar{g}_r \cdot (\bar{p} + \bar{\delta})| \leqq \sum_{k=r}^n (|p_k| + |\delta_k|)|\alpha|^k.$$

Applying a double summation, we conclude that

$$\sum_1^n |q_r^* \alpha^r| \leqq n(|p_n| + |\delta_n|)|\alpha|^n + (n-1)(|p_{n-1}| + |\delta_{n-1}|)|\alpha|^{n-1} + \cdots + (|p_1| + |\delta_1|)|\alpha|.$$

If we define the polynomials $\tilde{P}(x) \equiv \Sigma_0^n |p_r| x^r$ and $\tilde{D}(x) \equiv \sum_0^n |\delta_r| x^r$, then the foregoing inequality can be written in the form

$$\sum_1^n |q_r^* \alpha^r| \leqq |\alpha|(\tilde{P}'(|\alpha|) + \tilde{D}'(|\alpha|)) \tag{7}$$

Combining (5), (6), (7), we deduce that

$$E \leqq \tilde{D}(|\alpha|) \leqq \varepsilon \tilde{P}(|\alpha|) + \sigma|\alpha|(\tilde{P}'(|\alpha|) + \tilde{D}'(|\alpha|)). \tag{8}$$

From the second inequality above, we find that on evaluation at $|\alpha|$

$$\tilde{D} - \sigma|\alpha|\tilde{D}' \leqq \varepsilon \tilde{P} + \sigma|\alpha|\tilde{P}'. \tag{9}$$

For any nth degree polynomial P, it is clear that $|\alpha|\tilde{P}'(|\alpha|) \leqq n\tilde{P}(|\alpha|)$, and equality will hold only in the event that the highest-order coefficient is the only nonzero coefficient. Since the same is true of D, it follows from (9) that $(1 - n\sigma)\tilde{D} \leqq (\varepsilon + n\sigma\tilde{P})$. Combining this with the first inequality in (8), and assuming that

$n\sigma < 1$, we conclude that

$$E \leqq \tilde{P}(|\alpha|)(\varepsilon + n\sigma)/(1 - n\sigma). \tag{10}$$

It is clear from the expression that the bound grows quite steeply with $|\alpha|$, particularly when n is large, and particularly when the high-order coefficients of P are of relatively large magnitude. A simpler but cruder bound derivable from (10) is

$$E \leqq \|\bar{p}\|_1(\varepsilon + n\sigma)/(1 - n\sigma). \tag{11}$$

Now suppose the polynomial is written in Chebyshev form, so that $P(x) \equiv \sum_0^n p_r x^r \equiv \sum_0^n t_r T_r(x)$. Since this problem is equivalent to evaluating a cosine series, the error analysis can be extracted from [3], [4]. It is assumed that we use Clenshaw's algorithm with no phase shift. The algorithm states that

$$\begin{aligned} &P(\alpha) = t_0 + u_1\alpha - u_2, \\ &\text{where } u_{n+1} = u_{n+2} = 0, \\ &u_r = t_r + 2u_{r+1}\alpha - u_{r+2}, \qquad r = n, n-1, \ldots, 1. \end{aligned} \tag{12}$$

The error bound for u_1 turns out to be

$$|u_1^* - u_1| \leqq \frac{\sigma\|\bar{t}\|_1 n[1 + (1 + 2|\alpha|)n(n+1)/2]}{1 - \sigma(1 + 2|\alpha|)n(n+1)/2}. \tag{13}$$

A bound for $|u_2^* - u_2|$ is given by the same expression with $n - 1$ replacing n. These bounds are generally very conservative except when $|\alpha| \simeq 1$. If $|\alpha| \leqq 1/\sqrt{2}$, then the quantity $n(n+1)/2$ occurring twice in (13) can be replaced by $n\sqrt{2}$. Since the bound (13) is (for large $|\alpha|$) of order n^3, compared with (11) which is at most of order n, it might appear that the Chebyshev method was inferior to the Horner scheme. This can indeed be the case, but, more commonly, the advantage is reversed in consequence of $\|\bar{t}\|_1 \ll \|\bar{p}\|_1$.

The vectors $\bar{t}$, $\bar{p}$ are related by $\bar{p} = U\bar{t}$, where the matrices U, U^{-1} are given in [5]. It is evident that $\|U^{-1}\|_1 = 1$ and hence that $\|\bar{t}\|_1 \leqq \|\bar{p}\|_1$; moreover, since U^{-1} is column-stochastic, we shall have $\|\bar{t}\|_1 = \|\bar{p}\|_1$ whenever the elements of $\bar{p}$ are of uniform sign. Furthermore, if $\bar{p}$ has alternating signs, we can say that evaluating $P(x)$ is indistinguishable from evaluating a constant-sign polynomial at $-x$. Since the two special cases of constant signs and alternating signs figure prominently in Taylor expansions of elementary functions and elsewhere, it is worth noting that in these cases the conversion to Chebyshev form is *disadvantageous* from the error-bound point of view. The fact that the error bound is worsened does not permit any firm conclusion about the actual errors, since the errors and their bounds are rather loosely correlated. For example, by examining the structure of U^{-1} it becomes clear that when $\bar{p}$ is positive, so that $\|\bar{t}\|_1 = \|\bar{p}\|_1$, the t-sequence of coefficients is much more strongly damped than the p-sequence. It was shown in [4] that damping has a beneficial effect on computational stability, but no allowance is made for this in determining the error bound. The damping may indeed be sufficient to make the high-order terms of the Chebyshev series negligible, but, in this article, we are assuming that the polynomial is mathematically determined, and it is only the evaluation schemes that are under discussion. The empirical results given below

seem to justify the view that when a polynomial has coefficients of constant sign or of strictly alternating sign, there is no substantial difference in accuracy between Horner's scheme and Clenshaw's. The upper triangular matrix U contains in its $(k+1)$st column the coefficients of $T_k(x)$ with the constant term written in the first row. Corresponding to an nth degree polynomial, the order of U will be $n+1$ and $\|U\|_1$ will be the absolute sum of the coefficients of T_{n+1}. Equivalently,

$$\|U\|_1 = |T_{n+1}(i)| = (1/2)\left[(1+\sqrt{2})^{n+1} + (1-\sqrt{2})^{n+1}\right]. \tag{14}$$

Since we have $\|\bar{p}\|_1 \leqq \|U\|_1 \|\bar{t}\|_1$, with equality attained when $P(x) \equiv T_n(x)$, it can clearly happen that $\|\bar{p}\|_1 \gg \|\bar{t}\|_1$, and this explains the observation in [6] that the Horner scheme performs particularly badly on the evaluation of Chebyshev polynomials. In fact, although a comparison of (11) and (13) appears to favor the Horner method by a factor $O(n^2)$, the countervailing term $(1+\sqrt{2})^{n+1}$ in (14) may put the advantage overwhelmingly in favor of the Chebyshev-Clenshaw algorithm.

Some experiments were performed in order to test the two main hypotheses that seem to emerge from the above arguments, namely:

(A) The accuracy of the Horner scheme is highly sensitive to the magnitude of α. (The same is also true of the Clenshaw scheme, as was proved in [3].)

(B) When a polynomial has coefficients of constant sign or of strictly alternating sign, a translation into Chebyshev form will not bring any systematic improvement in accuracy of evaluation.

A polynomial $P(x) \equiv \sum_0^9 x^r/(10-r)$ was defined, with coefficients correct to six hexadecimal figures. Two ranges for the arguments α were defined, namely $I = [-.9, .9]$ and $I' = [-1, -.9] \cup [.9, 1]$. Within each range, 100 uniformly distributed values of α were chosen, and the mean and maximum of the evaluation errors were computed. The 'errors' were taken to be the difference between the results of single- and double-precision evaluation, where the double-precision coefficients of $P(x)$ were made identical to the single-precision coefficients. The polynomial was then translated into Chebyshev form and the experiments were repeated, using Clenshaw's algorithm on the same set of arguments. The fact that the translation cannot be performed with perfect accuracy should not be considered significant, because we never directly examine the discrepancies between the Clenshaw and Horner evaluations; these discrepancies would, of course, be partly attributable to translation errors, and they would not yield information on the merits of the two algorithms.

TABLE

	I $(\|\alpha\| \leqslant .9)$	I' $(.9 \leqslant \|\alpha\| \leqslant 1.)$
Horner	27	286
	497	745
Clenshaw	47	326
	263	556

The table above gives the results of four experiments corresponding to all combinations of the two methods and the two argument ranges. The two figures given at each grid point denote the mean and maximum errors measured in units of 10^{-8} for 100 argument values.

It seems that hypothesis (A) is well evidenced by these results. As for (B), the reader will have to form his own judgment, but it seems that neither method has a clear empirical advantage over the other. Our conclusion, therefore, is that although the Clenshaw algorithm is generally superior to Horner's, this superiority will generally not apply in a situation where the polynomial coefficients have uniform sign or strictly alternating sign.

Department of Computer Science
University of Kentucky
Lexington, Kentucky 40506

1. C. W. CLENSHAW, "A note on the summation of Chebyshev series," *MTAC*, v. 9, 1955, pp. 118-120. MR **17**, 194.

2. F. L. BAUER, "Optimally scaled matrices," *Numer. Math.*, v. 5, 1963, pp. 73-87. MR **28** #2629.

3. W. M. GENTLEMAN, "An error analysis of Goertzel's (Watt's) method for computing Fourier coefficients," *Comput. J.*, v 12, 1969/70, pp.160-165. MR **39** #5081.

4. A. C. R. NEWBERY, "Error analysis for Fourier series evaluation," *Math. Comp.*, v. 27, 1973, pp. 639-644.

5. C. LANCZOS, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1956. MR **18**, 823.

6. J. R. RICE, "On the conditioning of polynomial and rational forms," *Numer. Math.*, v. 7, 1965, pp. 426-435. MR **32** #6710.