ERROR ANALYSIS FOR POLYNOMIAL AND FOURIER SERIES EVALUATION

by B. N. SUR, *Mathematical Sciences Group, Systems Engineering Division, National Aeronautical Laboratory, Bangalore* 570017

(*Received* 12 *September* 1977)

Floating-point forward error analysis is developed for the (1) evaluation of a real polynomial at a real argument by Horner's scheme, (2) evaluation of Fourier series by Clenshaw's algorithm. The computable error bounds derived for the two cases are of the same order as those derived by Newbery for floating-point backward error analysis.

1. INTRODUCTION

Newbery (1973, 1974) has derived error bounds for polynomial (real) evaluation at a real argument by Horner's scheme and Fourier-series evaluation by Clenshaw's algorithm by backward analysis. In this paper it is shown that forward error analysis may be carried out for the above two cases and the error bounds derived are of the same order as those derived by Newbery (1973, 1974). The purpose of this paper is to show for the two cases that forward error analysis can be as good as backward analysis.

2. FORWARD ERROR ANALYSIS FOR POLYNOMIAL EVALUATION

The Horner's scheme of evaluating the real polynomial

$$P(x) = \sum_{r=0}^{N} a_r \, x^{N-r}$$

at a real argument α is given by

$$\left.\begin{aligned} q_0 &= a_0 \\ q_n &= a_n + \alpha \, q_{n-1}, \; n = 1, 2, \ldots N \\ q_N &= P(\alpha). \end{aligned}\right\} \qquad \ldots(1)$$

Without loss of generality we can assume $|\alpha| \leqslant 1$, for if it is not, then one can transform the polynomial $P(x)$ into another polynomial $H(y)$ by suitable transformation so that $|y| \leqslant 1$. In the subsequent analysis we assume that computation is carried under normalized floating-point arithmetic. Since each floating-point

arithmetic operation in (1) is subject to maximum relative error of magnitude ϵ, we will get a sequence of computed numbers q_n^* given by

$$\left.\begin{aligned} q_0^* &= a_0 \\ q_n^* &= a_n + \alpha\, q_{n-1}^*\, \delta_n,\ n = 1, 2, \ldots, N \end{aligned}\right\} \qquad \ldots(2)$$

where

$$\delta_n = a_n\, \epsilon_1 + \alpha\, q_{n-1}^*\, \sigma_1$$

$$\sigma_1 = \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2$$

$$| \epsilon_1 | \leqslant \epsilon$$

$$| \epsilon_2 | \leqslant \epsilon$$

and

$$| \sigma_1 | \leqslant 2\epsilon + \epsilon^2 = \sigma \text{ (say)}.$$

The error accumulated in computing q_n is given by

$$r_n = q_n^* - q_n = \alpha(1 + \sigma_1)\, r_{n-1} + \xi_n \qquad \ldots(3)$$

where

$$\xi_n = a_n\, \epsilon_1 + \alpha\, \sigma_1\, q_{n-1}.$$

The equation (3) is a linear difference equation in $\{r_n\}$ of order one with initial conditions $r_0 = 0$. The solution of (3) is given by

$$r_n = \sum_{r=1}^{n} \alpha^{n-r}\, (1 + \sigma_1)^{n-r}\, \xi_r. \qquad \ldots(4)$$

Normally the positive integer N is much smaller than $1/\sigma$ so that we can assume $N\sigma < 0.1$.

Therefore we have, as suggested by Wilkinson (1963),

$$(1 + \sigma_1)^{n-r} \leqslant (1 + \sigma)^{n-r} \leqslant (1 + \sigma)^N \leqslant (1 + 1{\cdot}06\, N\sigma).$$

The evaluation error E is given by

$$| E | = | q_N^* - q_N | = | r_N | \leqslant (1 + 1.06\, N\sigma) \left[\epsilon \sum_{n=1}^{N} | a_n |\ | \alpha |^{N-n} + \sigma \sum_{n=1}^{N} | \alpha |^{N-n+1} | q_{n-1} | \right]. \qquad \ldots(5)$$

Since

$$\sum_{n=1}^{N} | q_{n-1} | \; | \alpha |^{N-n+1} \leqslant | \alpha | \tilde{P}' (| \alpha |) \leqslant N \tilde{P} (| \alpha |)$$

where

$$\tilde{P}(x) = \sum_{r=0}^{N} | a_r | x^{N-r}$$

it follows from (5)

$$| E | \leqslant (1 + 1.06 N\sigma) \tilde{P} (| \alpha |) (\epsilon + N\sigma). \qquad ...(6)$$

The error bound derived by Newbery (1974) by backward analysis is given by

$$| E | \leqslant \tilde{P} (| \alpha |) (\epsilon + N\sigma)/(1 - N\sigma). \qquad ...(7)$$

It may be noted that error bounds in (6) and (7) are of the same order.

3. Forward Error Analysis for Fourier Series Evaluation

Let the Fourier series be expressed in the form

$$F(\theta) = \sum_{n=0}^{N} C_n \cos n\theta + \sum_{n=0}^{N} S_n \sin n\theta \qquad ...(8)$$

after the phase-shift transformation if necessary, so that θ is in the range $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$ modulo π. Newbery (1973) has shown that transformation involves no arithmetic but some sign-reversals.

Clenshaw's algorithm of evaluating the series (8) is given by

$$\left.\begin{aligned}
&(i) \quad u_{-1} = u_{-2} = 0 \\
&(ii) \quad u_n = a_n + 2u_{n-1} \cos\theta - u_{n-2}, n = 0, 1, 2, \ldots, N-1 \\
&(iii) \quad \sum_{n=0}^{N} C_n \cos n\theta = a_N + u_{N-1} \cos\theta - u_{N-2} \\
&\qquad \text{with } a_n \text{ interpreted as } C_{N-n} \\
&(iv) \quad \sum_{n=0}^{N} S_n \sin n\theta = u_{N-1} \sin\theta \\
&\qquad \text{with } a_n \text{ interpreted as } S_{N-n}.
\end{aligned}\right\} \qquad ...(9)$$

A sequence of numbers u_n^* $(n = 0, 1, 2, \ldots, N-1)$ is generated as follows :

$$u_n^* = a_n + 2u_{n-1}^* \cos\theta - u_{n-2}^* + a_n\sigma_1 - u_{n-2}^*\sigma_1$$

$$+ 2u_{n-1}^* \cos\theta\,\sigma_2 \qquad \ldots(10)$$

where

$$|\epsilon_1| \leqslant \epsilon, \quad |\epsilon_2| \leqslant \epsilon, \quad |\epsilon_3| \leqslant \epsilon,$$

$$\sigma_1 = \epsilon_1 + \epsilon_3 + \epsilon_1\epsilon_3, \quad \sigma_2 = \epsilon_2 + \epsilon_3 + \epsilon_2\epsilon_3$$

$$|\sigma_1| \leqslant \sigma, \quad |\sigma_2| \leqslant \sigma$$

and

$$\sigma = 2\epsilon + \epsilon^2.$$

Therefore, rounding error r_n at the nth step is given by

$$r_n = u_n^* - u_n = 2\cos\theta(1 + \sigma_2)\, r_{n-1} - (1 + \sigma_1)\, r_{n-2}$$

$$+ \delta_n \text{ for } n \geqslant 0 \qquad \ldots(11)$$

where

$$\delta_n = a_n\sigma_1 + 2\cos\theta\,\sigma_2 u_{n-1} - u_{n-2}\sigma_1.$$

We also note that

$$|\delta_n| \leqslant \sigma[\,|a_n| + |2\cos\theta|\,|u_{n-1}| + |u_{n-2}|\,]. \qquad \ldots(12)$$

Equation (11) is a linear difference equation in $\{r_n\}$ of order (2) with initial conditions $r_{-1} = r_{-2} = 0$. If we denote two independent solutions of the homogeneous linear difference equation obtained from (11) by $r_n^{(1)}$ and $r_n^{(2)}$, then a solution of (10) as suggested by Henrici (1964) is given by

$$r_n = \sum_{m=0}^{n} \frac{\begin{vmatrix} r_n^{(1)}, & r_n^{(2)} \\ r_{m-1}^{(1)}, & r_{m-1}^{(2)} \end{vmatrix}}{W_m} \text{ for } n = 0, 1, 2, \ldots, N-1, \qquad \ldots(13)$$

where

$$W_m = \begin{vmatrix} r_m^{(1)}, & r_m^{(2)} \\ r_{m-1}^{(1)}, & r_{m-1}^{(2)} \end{vmatrix}.$$

The zeros of the characteristic polynomial

$$t^2 - 2\cos\theta(1 + \sigma_2)\, t + (1 + \sigma_1)$$

of the homogeneous difference equation obtained from (11) are given by

$$t = \lambda \pm \sqrt{\lambda^2 - \mu}, \text{ where } \lambda = (1 + \sigma_2)\cos\theta$$

and

$$\mu = 1 + \sigma_1.$$

Since θ is in the range of $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$ modulo π and σ_1 and σ_2 are small quantities, therefore $\lambda^2 < \mu$.

So

$$r_n^{(1)} = \rho^n \cos n\phi \text{ and } r_n^{(2)} = \rho^n \sin n\phi,$$

where

$$\lambda = \rho\cos\phi \text{ and } \sqrt{\mu - \lambda^2} = \rho \sin\phi.$$

If we assume that $\delta_{-2} = \delta_{-1} = 0$ then from (13) it follows that (This assumption helps us to write the solution in compact form. It does not mean that the difference equation is defined for $n < 0$.)

$$r_n = \sum_{m=-2}^{n} \rho^{n-m} \frac{\sin(n - m + 1)\phi}{\sin\phi}\, \delta m \qquad \ldots(14)$$

satisfies the difference equation (11) as well as initial conditions $r_{-1} = r_{-2} = 0$.

Since N is small compared to $\frac{1}{\sigma}$, we can assume $N\sigma < 0.1$. Therefore, as before, we have

$$\begin{aligned} |\rho^{n-m}| &\leqslant |\rho|^{N-1-m} \leqslant |\rho|^{N-1} = (1 + \sigma_1)^{(N-1)/2} \\ &\leqslant 1 + 1.06\left(\frac{N-1}{2}\right)\sigma \\ &\leqslant [1 + 0.53\,(N-1)\sigma]. \end{aligned} \qquad \ldots(15)$$

From (14) and (15) it follows that

$$|u_{N-1}^* - u_{N-1}| = |r_{N-1}| \leqslant \{1 + 0.53\,(N-1)\sigma\} \times \left\{ |\operatorname{cosec}\phi| \sum_{m=0}^{N-1} |\delta_m| \right\}. \qquad \ldots(16)$$

From (12) it follows that

$$\sum_{m=0}^{N-1} | \delta_m | \leqslant \sigma [\| A \|_1 + \{2 | \cos \theta | + 1\} \| U \|_1] \qquad ...(17)$$

where

$$U = (u_0, u_1, u_2, ..., u_{N-1})^T$$

$$A = (a_0, a_1, a_2, ..., a_{N-1})^T.$$

The recurrence relation (9) can be expressed as

$$MU = A \qquad ...(18)$$

where M is a $(N \times N)$ matrix with units on diagonal and $-2 \cos \theta$ on the first sub-diagonal and units on the second sub-diagonal.

The inverse of M is (m_{ij}) where

$$m_{ij} = 0, j > 1$$

$$m_{ij} = \frac{\sin (i - j + 1) \theta}{\sin \theta}, \; j \leqslant i.$$

Therefore,

$$\| M^{-1} \|_1 \leqslant N | \operatorname{cosec} \theta |$$

and consequently

$$\| U \|_1 = \| M^{-1} A \|_1 \leqslant \| M^{-1} \|_1 \| A \|_1 \leqslant N | \operatorname{cosec} \theta | \| A \|_1. \qquad ...(19)$$

From (16), (17) and (19), it follows that

$$| u^*_{N-1} - u_{N-1} | \leqslant [1 + 0.53 (N - 1) \sigma] | \operatorname{cosec} \phi | \sigma \| A \|_1 \times [1 + \{2 | \cos \theta | + 1\} N | \operatorname{cosec} \theta |]. \qquad ...(20)$$

Since

$$\operatorname{cosec} \phi = \frac{\rho}{\sqrt{\mu - \lambda^2}} = \frac{1 + \sigma_1}{\sqrt{(1 + \sigma_1) - (1 + \sigma_2)^2 \cos^2 \theta}}$$

and σ_1, σ_2 are small quantities cosec ϕ will be nearly equal to cosec θ.

Replacing cosec ϕ by cosec θ in (20) we get

$$| u^*_{N-1} - u_{N-1} | \leqslant [1 + 0.53 (N - 1) \sigma] | \operatorname{cosec} \theta | \sigma \| A \|_1 \times [1 + N \{2 | \cos \theta | + 1\} | \operatorname{cosec} \theta |] \qquad ...(21)$$

Since θ lies in the range $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$ modulo π and $\cos\theta$, $\operatorname{cosec}\theta$ take their maximum value $\frac{1}{\sqrt{2}}$, $\sqrt{2}$ respectively when $\theta = \frac{\pi}{4}$ modulo π, we have from (21)

$$| u^*_{N-1} - u_{N-1} | \leqslant [1 + 0.53\,(N-1)\sigma]\, 2^{(1/2)}\, \sigma \,\| A \|_1 \times [1 + N(2 + 2^{1/2})] \qquad \ldots(22)$$

The error bound given by (22) is of the same order as that obtained by Newbery (1973, see inequality (10)).

Conclusion

Wilkinson (1963) has commented that backward analysis is often much simpler than forward analysis particularly in connection with floating-point computations. The author, however, feels that for linear algorithms, forward analysis is as simple as backward analysis for floating point computations.

Acknowledgement

The author is grateful to the Director, National Aeronautical Laboratory, Bangalore, for the permission to carry on this work. He gratefully acknowledges the help extended by Shri S. Janardhan in this work.

References

Henrici, P. (1964). Elements of Numerical Analysis. John Wiley and Sons, Inc., New York, pp. 119-45.

Newbery, A. C. R. (1973). Error analysis for Fourier series evaluations. *Math. Comp.*, **27**, 639-44.

———— (1974). Error analysis for polynomial evaluation. *Math. Comp.*, **28**, 789-93.

Wilkinson, J. H. (1963). Rounding errors in algebraic process. National Physical Laboratory — Notes on Applied Science No. 32, H.M. Stationary Office, London, 19.