MATHEMATICS OF COMPUTATION, VOLUME 27, NUMBER 123, JULY, 1973

Error Analysis for Fourier Series Evaluation

By A. C. R. Newbery

Abstract. A floating-point error analysis is given for the standard recursive method of evaluating trigonometric polynomials. It is shown that, by introducing a phase-shift, one can hold the error growth down to an essentially linear function of the degree. Explicit computable error bounds are derived and numerically verified.

Given the problem of evaluating the Fourier series

$$F(\theta) = \sum_{0}^{N-1} C_r \cos r\theta + \sum_{0}^{N-1} S_r \sin r\theta,$$

the most efficient, known method is Clenshaw's algorithm, which is also known as the "Goertzel-Watt" algorithm [1], [2]. The numerical properties of the floating-point algorithm were analyzed by Gentleman [2], and the principal conclusion was that the cumulative effect of rounding errors could become very severe whenever θ was small modulo π. By using the phase-shift $\phi = \pi/2 - \theta$, one can always replace the Fourier series $F(\theta)$ by the equivalent series $G(\phi) = F(\pi/2 - \phi)$, and one can therefore always avoid evaluating a Fourier series at a small argument. If we arrange to perform the phase-shift whenever θ is in the range $(-\pi/4, \pi/4)$ modulo π, we can guarantee that all evaluations occur with arguments in the range $[\pi/4, 3\pi/4]$ modulo π. Under these conditions, the behavior of the Clenshaw algorithm will be shown to be quite good. The transformation consists in determining the coefficients C'_r, S'_r of $G(\phi)$, and these are related to C_r, S_r by

$$\begin{bmatrix} C'_r \\ S'_r \end{bmatrix} = T_r \begin{bmatrix} C_r \\ S_r \end{bmatrix}, \tag{1}$$

where

$$T_r = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix},$$

according as $r = 0, 1, 2, 3$ modulo 4.

The transformation, therefore, does not involve any arithmetic; it simply involves some sign-reversals and some swapping of terms between one part of the series and the other.

Assuming that the phase-shift, if necessary, has been performed, we now examine the error-sensitivity of Clenshaw's algorithm, on the assumption that θ is in $[\pi/4, 3\pi/4]$ modulo π. The algorithm is defined by

Received March 2, 1972.
AMS (MOS) subject classifications (1970). Primary 42-04, 65G05.
Key words and phrases. Error analysis, Fourier series.

$$
\begin{aligned}
&\text{(i)} \quad u_N = u_{N+1} = 0,\\
&\text{(ii)} \quad u_r = f_r + 2u_{r+1}\cos\theta - u_{r+2}, \qquad r = N-1, N-2, \cdots, 1,\\
&\text{(iii)} \quad \sum_0^{N-1} C_r \cos r\theta = f_0 + u_1\cos\theta - u_2, \quad \text{with } f_r \text{ interpreted as } C_r,\\
&\text{(iv)} \quad \sum_0^{N-1} S_r \sin r\theta = u_1 \sin\theta,
\end{aligned} \tag{2}
$$

with f_r interpreted as S_r in the recursion (ii) defining $\{u_r\}$.

Clearly, the recursion defining $\{u_r\}$ is the most sensitive part of the computation. Written in matrix form, this recursion is equivalent to $M\bar{u} = \bar{f}$, where $\bar{u} = \{u_{N-1}, u_{N-2}, \cdots, u_1\}^T$, $\bar{f} = \{f_{N-1}, f_{N-2}, \cdots, f_1\}^T$ and M is a matrix with units on the diagonal and sub-subdiagonal, $-2\cos\theta$ on the subdiagonal and zero elsewhere. The inverse of M, quoted in [2], is $[m'_{ij}]$, where

$$
\begin{aligned}
m'_{ij} &= 0, && j > i,\\
&= \sin(i - j + 1)\theta/\sin\theta, && j \leqq i.
\end{aligned} \tag{3}
$$

We now perform a backward error analysis for the recursion (2)(ii) defining $\{u_r\}$. Let $\{u_r^*\}$ be the sequence actually developed, in view of the fact that the right sides of (2)(ii) are subject to rounding errors δ_r. Hence, we shall have

$$\bar{u}^* = M^{-1}(\bar{f} + \bar{\delta}) = \bar{u} + M^{-1}\bar{\delta}. \tag{4}$$

Although we do not ordinarily know δ_r, we can compute bounds Δ_r such that $|\delta_r| \leqq \Delta_r$. The value we determine will depend on what assumptions are made about the precision and sequencing of the arithmetic. In order to bound the error in the computed values of (2)(iii), (iv), it is only necessary to find bounds for $|u_1^* - u_1|$, $|u_2^* - u_2|$. From (3), (4), we see that

$$|u_1^* - u_1| = \left|\sum_{r=1}^{N-1} \frac{\sin(N-r)\theta}{\sin\theta}\,\delta_{N-r}\right| \leqq ||\bar{\Delta}||_1\,|\operatorname{cosec}\theta| \leqq 2^{1/2}\,||\bar{\Delta}||_1. \tag{5}$$

This follows from the fact that when θ is in $[\pi/4, 3\pi/4]$,

$$|\sin(N-r)\theta/\sin\theta| \leqq |1/\sin\theta| \leqq 2^{1/2} \quad \text{and} \quad ||\Delta||_1 = \sum_1^{N-1} |\Delta_r|.$$

We now study the recursion

$$u_r^* = fl[fl(f_r - u_{r+2}^*) + fl(2u_{r+1}^*\cos\theta)]$$

which is equivalent to

$$u_r^* = f_r - u_{r+2}^* + 2u_{r+1}^*\cos\theta + \delta_r,$$

where the quantity δ_r is not explicitly added; it is there to accommodate the difference between the indicated floating operations and the true mathematical values. If we are working in single precision b-digit binary arithmetic, the relative error ζ in each

operation will be bounded in magnitude by $\epsilon = 2^{-b}$. We shall therefore obtain

$$u_r^* = (1 + \zeta_1)[(f_r - u_{r+2}^*)(1 + \zeta_2) + 2u_{r+1}^* \cos\theta(1 + \zeta_3)],$$

$$\delta_r = (f_r - u_{r+2}^*)[(1 + \zeta_1)(1 + \zeta_2) - 1]$$

$$+ 2u_{r+1}^* \cos\theta[(1 + \zeta_1)(1 + \zeta_3) - 1], \quad \text{and} \tag{6}$$

$$|\delta_r| \leqq \Delta_r = K(|f_r| + |u_{r+2}^*| + 2\,|u_{r+1}^* \cos\theta|),$$

$$\text{where } K = (1 + \epsilon)^2 - 1 = 2\epsilon + \epsilon^2.$$

Using the same argument as [2, Eq. (6)], we deduce from (6) that

$$||\bar{\Delta}|| \leqq K[||\bar{f}|| + (1 + 2\,|\cos\theta|)\,||\bar{u}^*||]. \tag{7}$$

Using any vector norm and subordinate matrix norm, it follows from (4) that

$$||\bar{u}^*|| = ||M^{-1}(\bar{f} + \bar{\delta})|| \leqq ||M^{-1}||\,(||\bar{f}|| + ||\bar{\delta}||) \leqq ||M^{-1}||\,(||\bar{f}|| + ||\bar{\Delta}||). \tag{8}$$

Combining this with (7), we obtain

$$||\bar{\Delta}|| \leqq K[||\bar{f}|| + (1 + 2\,|\cos\theta|)\,||M^{-1}||\,(||\bar{f}|| + ||\bar{\Delta}||)].$$

Hence,

$$||\bar{\Delta}||\,[1 - K(1 + 2\,|\cos\theta|)\,||M^{-1}||] \leqq K\,||\bar{f}||\,[1 + (1 + 2\,|\cos\theta|)\,||M^{-1}||].$$

We now interpret this using the L_1 norm, noting that $||M^{-1}||_1 \leqq (N - 1)/|\sin\theta|$. On substituting into (5), we conclude that

$$|u_1^* - u_1| \leqq \frac{K\,||\bar{f}||_1\,|\operatorname{cosec}\theta|\,[1 + (1 + 2\,|\cos\theta|)(N - 1)\,|\operatorname{cosec}\theta|]}{1 - K(1 + 2\,|\cos\theta|)(N - 1)\,|\operatorname{cosec}\theta|}, \tag{9}$$

provided the denominator is positive.

The above proviso is not a severe constraint. The shortest word length commonly used in scientific computation is 21 binary digits. In these conditions, it would take a series of over a quarter-million terms to make the denominator vanish. When θ is in the range $[\pi/4, 3\pi/4]$ modulo π, we may note that the quantity $(1 + 2\,|\cos\theta|)\,|\operatorname{cosec}\theta|$ takes its maximum value of $2 + 2^{1/2}$ when $\theta = \pi/4$ modulo π. This enables us to rewrite (9) in a form that is independent of θ, provided that θ is in the stated range. Thus

$$|u_1^* - u_1| \leqq \frac{K\,||\bar{f}||_1\,2^{1/2}[1 + (2 + 2^{1/2})(N - 1)]}{1 - K(2 + 2^{1/2})(N - 1)}. \tag{10}$$

In all "reasonable" situations, the denominator of (10) will be very close to unity, and the error bound can be taken as a linear function of N. The same bounds given by (9), (10) for $|u_1^* - u_1|$ will also serve for $|u_2^* - u_2|$, the reasoning being the same, but the vectors and matrices occurring are of order one less, so that $N - 1$ could be replaced by $N - 2$.

If greater accuracy is required, it can be obtained by use of local double precision. If we replace (2)(ii) by

$$u_r^* = \mathrm{Rnd}(f_r + 2u_{r+1}^* \cos\theta - u_{r+2}^*),$$

where the quantity in parentheses is obtained by using higher-precision arithmetic on single-precision arguments in such a way that the computed value is the "mathematical" value correctly rounded to single precision, then we may say

$$\Delta_r = \epsilon\,|u_r^*|, \qquad ||\bar{\Delta}|| = \epsilon\,||\bar{u}^*|| = \epsilon\,||M^{-1}(\bar{f} + \bar{\delta})|| \leqq \epsilon\,||M^{-1}||\,(||\bar{f}|| + ||\bar{\Delta}||).$$

Hence

$$||\bar{\Delta}||_1 \leqq \epsilon\,||\bar{f}||_1\,(N-1)\,|\operatorname{cosec}\theta|/[1 - \epsilon(N-1)\,|\operatorname{cosec}\theta|]$$

and so

$$|u_1^* - u_1| \leqq \epsilon\,||\bar{f}||_1\,(N-1)\operatorname{cosec}^2\theta/[1 - \epsilon(N-1)\,|\operatorname{cosec}\theta|]. \tag{11}$$

This will ordinarily reduce the error bound by a factor between 2 and 5 as compared with (9).

One disadvantage of the proposed phase-shift is that it will convert a pure sine or cosine series into a mixed series, thereby approximately doubling the cost of evaluation in this case. In many cases however, the stability and error bounding will be considered to be worth the cost. An alternative method for achieving stability without additional cost was proposed by Reinsch (unpublished) and quoted at the end of [2].

The proposal, valid for $\theta \simeq 0$ modulo 2π, is to compute the following recursion, which is mathematically equivalent to (2):

$$\begin{aligned} d_r &= f_r + d_{r+1} - 2(1 - \cos\theta)u_{r+1}, \qquad r = N-1, N-2, \cdots, 1, \\ u_r &= u_{r+1} + d_r, \\ d_N &= u_N = 0. \end{aligned} \tag{12}$$

In our experience on general problems, the algorithm does indeed perform about as well as the phase-shifted Goertzel algorithm throughout the range $0 \leqq \theta < \pi/2$. However, no formal error bound is available for it, and it is capable of performing badly, as the following "loaded" example illustrates.

Example. Evaluate $\alpha \cos 100\theta$, where $\theta = 0$ and α is a floating number without many trailing zeros, e.g. $\alpha \simeq 3^{1/2}$.

The Reinsch algorithm will yield

$$d_{100} = d_{99} = \cdots = d_1 = \alpha, \quad \text{and}$$

$$u_{100} = \alpha, \quad u_{99} = \alpha + \alpha, \cdots, u_r = u_{r+1} + \alpha.$$

The "mathematical" value of u_r is $(101 - r)\alpha$, but computationally this quantity will result from $(100 - r)$ rounded additions of α to itself. Finally, the "true" result of α will be approximated by performing the subtraction $u_1 - u_2$, where $u_1 \simeq 100\alpha$ and $u_2 \simeq 99\alpha$. It is true that the unmodified Goertzel algorithm, using the recursion $u_r = 2u_{r+1} - u_{r+2}$, will perform even worse on this problem, but a phase-shift will translate the problem into an evaluation of $\alpha \cos 100\phi$, where $\phi = \pi/2$. Let $\{u_r'\}$ be the sequence of u's generated by the Goertzel algorithm on this problem; then $u_{100}' = \alpha$, $u_{99}' = 0$, $u_{98}' = -\alpha$, etc. Although this example is admittedly loaded, it does indicate that the Reinsch algorithm is capable of producing inferior results, even within the

argument range for which it is specifically designed. We have no evidence that such behavior is typical; we merely note that it *can* occur.

The case where a Chebyshev expansion is to be evaluated at an argument > 1 is equivalent to applying the recursion (2) with $\cos\theta > 1$. Equivalently, we have to replace the trigonometric functions in (2), (3) by their hyperbolic counterparts. Since we now have no periodicity, we cannot stabilize by means of a phase-shift. The Reinsch algorithm then appears to be the best approach, at least for arguments not greatly exceeding 1.

In order to validate the error bound (9), we defined three different Fourier series $\sum_1^{300} C_r \cos r\theta + \sum_1^{300} S_r \sin r\theta$ and evaluated each at 100 arguments randomly chosen. The "error" was taken to be the difference between the single-precision and double-precision evaluation, care being taken that the two runs operated on identical data, e.g. the double-precision value of $\cos\theta$ was taken to be the single-precision value with the less significant word set to zero. The three problems were:

(A) C_r, S_r uniform in $[-5, 5]$, $\sin\theta$ uniform in $[-.5, .5]$.

(B) C_r, $S_r = e^{-r/30}(C'_r, S'_r)$, where C'_r, S'_r are uniform in $[-5, 5]$, so as to make a damped series; $\sin\theta$ uniform in $[-.5, .5]$.

(C) $C_{300} = 3^{1/2}$. All other coefficients zero, $\sin\theta$ uniform in $[-.25, .25]$.

In the following table, the columns refer to problems (A), (B), (C), respectively; the rows refer to the phase-shifted Goertzel (PSG), Reinsch (R) and unmodified Goertzel (UG) algorithms, respectively. For each θ value in each of the three tests, the error bound for the PSG algorithm was computed; firstly, (9) was used to bound the errors in u_1^*, u_2^* in terms of θ and the series coefficients, then (2)(iii) and (2)(iv) were used in order to place a bound on the extent to which errors could affect the computed evaluation. The resulting error bound E was used as a standard against which the observed errors were measured. For each series and each algorithm, the mean and maximum of the quotient $|\text{observed error}|/E$ were computed and are tabulated below.

	A (neutral)	B (damped)	C (highly undamped)
PSG	.000804	.000162	.0337
	.00234	.000719	.0668
R	.000355	.00105	.0483
	.00384	.00323	.150
UG	.0116	.00632	1.91
	.215	.0962	24.1

In case it may be thought that these results indicate that the bound (9) is too conservative, the following considerations should be borne in mind. Firstly, an error bound has always to be computed on the unlikely assumption that all the local errors conspire to maximize the total error; secondly, the bounds would have looked less conservative if a genuine binary machine had been used. It was necessary to consider six hexadecimal digits as uniformly equivalent to 21 binary, although some of the arithmetic had precision up to 24 binary digits.

In conclusion, it seems appropriate to ask whether the bound (9) is attainable, or whether it could possibly be replaced by a sharper bound. To determine this, we must look in turn at the various inequalities on which (9) depends. If the first inequality in (5) could ever be replaced by an equality, this would be equivalent to asserting that the quantities δ_r were proportional to $\sin(N-r)\theta$, *and* $|\sin(N-r)\theta| = 1$ for all r. While the first condition could occasionally be satisfied, the second clearly cannot, except in the case of a one-term 'series'. Something has therefore been given away by this inequality, but the give-away turns out to be quite small in general. The average value of $|\sin k\theta|$ as $k \to \infty$ is the average distance of points on the circumference of a unit circle from the diameter, namely $\pi/4$. By taking the average of $|\sin(N-r)\theta|$ to be unity, we are therefore not conceding substantially more than is necessary in the general case. We now examine all the inequalities on which (9) depends and write down the necessary conditions under which each individual inequality could be replaced by an equality. From (5) through (8), these conditions are seen to be the following:

(i) The sign of $\delta_{N-r} \sin(N-r)\theta$ is constant for all r.

(ii) The quantities u^*_{r+2}, u^*_{r+1} are in constant ratio, so that the vector $\bar{u}^*$ is 'geometric' in the sense that its elements form a geometric progression. Also $\bar{f}$ is parallel to $\bar{u}^*$.

(iii) $\bar{f}$ is parallel to $\bar{\delta}$ and $\bar{f}$ is parallel to $\bar{e}_1 = (1, 0, 0, \cdots)$.

This follows from the fact that $||M^{-1}||_1$ is the absolute sum of the *first* column of M^{-1}. These conditions are seen to be generally incompatible, since $\bar{f}$ has to be parallel to $\bar{u}^*$ and $\bar{e}_1$, and $\bar{e}_1$ is not a geometric vector. Hence, the bound (9) is not attainable, but a study of the inequalities will indicate what type of problem should generate errors that come closest to the bound. The inequalities differ greatly in the size of their contribution to the bound. Sometimes a magnitude, which could in fact be zero, is replaced by a small bound; sometimes it is replaced by a not-so-small bound. The replacement of $||M^{-1}(\bar{f} + \bar{\delta})||_1$ by $||M^{-1}||_1(||f||_1 + ||\Delta||_1)$ is the most important instance of the latter kind, and this substitution will only be realistic if $\bar{f} \simeq \alpha\bar{e}_1$ for some scalar α. From this, we deduce the following assertion:

"*The kind of Fourier series on which the (phase-shifted) Clenshaw algorithm will be least accurate is one in which the high-frequency terms have relatively large coefficients*".

This assertion is well evidenced by the table above. On examining the results for the PSG algorithm, which is the only one to which our analysis strictly applies, we see that the bound is indeed more nearly attained on the neutral series (A) than on the damped series (B); it is closest for the highly undamped series (C). Similar remarks apply to the unmodified Goertzel algorithm except that the theoretical bound can be (and actually is) exceeded.

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. 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.