ERROR ANALYSIS OF AN ALGORITHM FOR SUMMING CERTAIN FINITE SERIES

DAVID ELLIOTT

(Received 18 August 1966, revised 12 February 1967)

1. Introduction

An algorithm for summing the series $P_N = \sum_{n=0}^{N} a_n p_n$, where the coefficients a_n are assumed known, and the quantities p_n satisfy a linear three term recurrence relation, has been given by Clenshaw [1]. If we suppose that the p_n satisfy the recurrence relation

$$p_n + \alpha_n p_{n-1} + \beta_n p_{n-2} = 0, \tag{1.1}$$

where α_n and β_n are, in general, functions of n, then P_N may be found by constructing a sequence $\{b_n\}$ for $n = N(-1)0$, where the b_n satisfy the inhomogeneous recurrence relation

$$b_n + \alpha_{n+1} b_{n+1} + \beta_{n+2} b_{n+2} = a_n, \tag{1.2}$$

with the conditions,

$$b_{N+1} = b_{N+2} = 0. \tag{1.3}$$

The sum P_N is then given by

$$P_N = p_0 b_0 + (p_1 + \alpha_1 p_0) b_1. \tag{1.4}$$

This result can be readily verified by multiplying each side of equation (1.2) by p_n, summing from $n = 0$ to N, and making use of equations (1.1) and (1.3).

This algorithm is very convenient for use on a digital computer since it is easy to program, and may in fact be looked upon as a generalisation of the well known "nested-multiplication" process for summing power series. It has been frequently used for summing Chebyshev series. In this paper, we propose to discuss this algorithm in far more detail than appears to have been given previously in the literature. In Section 3 we shall give a fairly complete error analysis. Partial error analyses have already been given in particular cases. Clenshaw [1] has given a partial analysis for a series of shifted Chebyshev polynomials; Elliott [2] has generalised this result to series of ultraspherical polynomials, and Smith [3] has further extended

this result to the case when the p_n are general orthogonal polynomials. However, all these analyses are incomplete since they consider only the effect of the error in each a_n, and neglect the errors which might arise in computing the b_n from equation (1.2), and P_N from equation (1.4). In the next section we shall first discuss the algorithm in some detail.

2. Further discussion of the algorithm

Before proceeding with the error analysis, we shall first obtain an explicit expression for the quantities b_n, as defined by equations (1.2) and (1.3), in terms of functions which we shall consider to be known. It is well known that equation (1.1) possesses two linearly independent solutions. One of these solutions is the given p_n; let a second solution of the recurrence relation be denoted by q_n. Since these two solutions are linearly independent, we define the Casorati determinant W_n (see, for example, Milne-Thomson [4]) by

$$W_n = p_{n-1}q_n - p_n q_{n-1}. \tag{2.1}$$

We have $W_n \neq 0$, and furthermore

$$W_n = \beta_n W_{n-1} \qquad \text{for } n = 1, 2, 3, \cdots. \tag{2.2}$$

An explicit solution for the quantities b_n can now be given in terms of a_n, p_n and q_n. That this is not a completely trivial problem may be seen from the fact that the homogeneous equation

$$b_n + \alpha_{n+1}b_{n+1} + \beta_{n+2}b_{n+2} = 0 \tag{2.3}$$

is not, in general, the same as equation (1.1) satisfied by p_n and q_n. However it can be verified by substitution that two linearly independent solutions of equation (2.3) are given by

$$\frac{p_{n-1}}{\prod_{i=1}^{n}\beta_i} \quad \text{and} \quad \frac{q_{n-1}}{\prod_{i=1}^{n}\beta_i},$$

where we have assumed (as in the cases to be considered later) that none of the β_i is zero. With this result, it is then readily shown by standard techniques that

$$b_n = \frac{1}{W_n}\sum_{m=n}^{N}(p_{n-1}q_m - p_m q_{n-1})a_m, \qquad \text{for } n = 0(1)N. \tag{2.4}$$

It is now convenient to introduce the sum Q_N, which is defined by

$$Q_N = \sum_{n=0}^{N} a_n q_n. \tag{2.5}$$

https://doi.org/10.1017/S1446788700005267 Published online by Cambridge University Press

We note that if our problem had been one of evaluating Q_N instead of P_N, then we would have calculated precisely the same quantities b_n. The value of Q_N is given by

$$Q_N = q_0 b_0 + (q_1 + \alpha_1 q_0) b_1,$$

and of course depends on the specific values taken by q_0 and q_1, which then uniquely define all the quantities q_n. Thus we see that the quantities b_0 and b_1, may be expressed in terms of P_N and Q_N; in fact,

$$(2.6) \qquad \begin{cases} b_0 = \dfrac{1}{W_1}\{(q_1+\alpha_1 q_0)P_N - (p_1+\alpha_1 p_0)Q_N\}, \\ b_1 = -\dfrac{1}{W_1}\{q_0 P_N - p_0 Q_N\}. \end{cases}$$

This result is of interest in so far as we can see the conditions under which the algorithm should *not* be used. Suppose that $|q_n| \gg |p_n|$ for most values of n in $0(1)N$, then in general we shall have $|Q_N| \gg |P_N|$. The sum P_N will then be obtained as the difference of two large numbers and a considerable loss of accuracy will occur.

An excellent example of such behaviour is given if we use the algorithm to sum the series

$$P_{12} = J_0(1) + 2\sum_{n=1}^{6} J_{2n}(1),$$

where $J_n(x)$ denotes the Bessel function of the first kind. The value of P_{12} is 1, correct to 10 decimal places. Suppose we calculate the quantities b_n using floating point decimal arithmetic to ten significant figures, using only single precision operations. Now we have

$$b_n = a_n + 2nb_{n+1} - b_{n+2},$$

with

$$b_{13} = b_{14} = 0$$

and

$$a_n = \begin{cases} 2 & (n \text{ even}, \neq 0), \\ 1 & (n = 0), \\ 0 & (n \text{ odd}). \end{cases}$$

The quantities b_n for $n = 4(1)12$ may be evaluated exactly (being integers with 10 or less digits), and rounded values have to be taken for the remaining b_n. We find in particular that $b_4 = 0{\cdot}32797\ 06418 \times 10^{10}$ (exact), together with $b_0 = -0{\cdot}73772\ 45906 \times 10^{11}$ and $b_1 = 0{\cdot}12828\ 18767 \times 10^{12}$. If now, in equation (1.4) we use values of $J_0(1)$ and $J_1(1)$ correctly rounded to ten decimal places, we find that $P_{12} = -30$ instead of 1, i.e. we have no correct significant digits. It is of course well known that two linearly in-

https://doi.org/10.1017/S1446788700005267 Published online by Cambridge University Press

dependent solutions of the recurrence relation satisfied by $J_n(1)$ are $J_n(1)$ and $Y_n(1)$ (the Bessel function of the second kind), and that for $n \gg 1$, we have $J_n(1) \sim 1/2^n n!$ and $Y_n \sim -2^n(n-1)!/\pi$. This is an example where $|q_n| \gg |p_n|$ for most values of n in the range $n = 0(1)12$, over which the series is summed.

The question then arises as to how such series might be summed. It is suggested that we first compute the values of p_n from equation (1.1) by making use of Miller's recurrence algorithm. This algorithm, and the errors which arise in its use, have recently been discussed in some detail by Olver [5]. With the values of p_n then determined, the series may be summed directly by accumulating the products $a_n p_n$ for $n = 0(1)N$. We shall not discuss the summation of such series any further in this paper. In the next section we shall consider a complete error analysis of the algorithm of Section 1.

3. Analysis of errors

We shall now consider the total effect of round-off errors on the computed value of P_N, due to the possible round-off errors arising at each step of the algorithm. The possible sources of error in P_N are:

(i) the round-off error in each a_n,
(ii) errors in the computed values of α_n and β_n,
(iii) errors in computing the b_n from equation (1.2), and
(iv) errors in computing P_N from equation (1.4).

Previous error analyses [1], [2] and [3] have only considered the error in P_N due to the errors in the a_n. Smith [3] correctly noted that if there is an error ϕ_n in each a_n then the error in P_N, assuming that no other errors occur, is obviously $\sum_{n=0}^{N} \phi_n p_n$. Clenshaw [1] and Elliott [2] obtained this result in their particular cases, in a far more elaborate manner.

In the subsequent analysis, we shall denote by $\bar{u}$ the value of a quantity u as it is computed during the course of the algorithm. We note that $\bar{u}$ is not necessarily always the correctly rounded value of u. The calculations may be performed in either fixed or floating point arithmetic, and to any number base. In particular if we assume that we are working in fixed point arithmetic to t decimal places, then $\bar{u}$ is the correctly rounded value of u if $|u-\bar{u}| \leqq \frac{1}{2} \times 10^{-t}$. On the other hand, if we are working with floating point numbers, with $\bar{u} = a \cdot 10^b$ where $0 \cdot 1 \leqq |a| < 1$, b an integer and a given to t decimal places, then $\bar{u}$ is the correctly rounded value of u if $|u-\bar{u}| \leqq \frac{1}{2} |u| \times 10^{1-t}$. In both cases we can put an upper bound on the round-off error; the modification for binary arithmetic is straightforward.

Let us now consider the algorithm again, but in terms of quantities

https://doi.org/10.1017/S1446788700005267 Published online by Cambridge University Press

that are actually computed. First, from equation (1.2), we compute $\bar{b}_n$ say, where

$$\bar{b}_n = (\bar{a}_n - \bar{\alpha}_{n+1}\bar{b}_{n+1} - \bar{\beta}_{n+2}\bar{b}_{n+2}) + r_n. \tag{3.1}$$

The quantity r_n is introduced as the round-off error which arises on rounding off the quantity in the brackets ( ). It will depend upon the way in which the computation of $(\bar{a}_n - \bar{\alpha}_{n+1}\bar{b}_{n+1} - \bar{\beta}_{n+2}\bar{b}_{n+2})$ is performed, and will, for example, be less if the sum of products is accumulated to double, rather than single length. Since equation (3.1) is not suitable for analysis as it stands, we rewrite it as

$$\bar{b}_n = \bar{a}_n - \alpha_{n+1}\bar{b}_{n+1} - \beta_{n+2}\bar{b}_{n+2} + \varepsilon_n + r_n, \tag{3.2}$$

where the quantity ε_n is defined by

$$\varepsilon_n = (\alpha_{n+1} - \bar{\alpha}_{n+1})\bar{b}_{n+1} + (\beta_{n+2} - \bar{\beta}_{n+2})\bar{b}_{n+2}. \tag{3.3}$$

Now we may write,

$$\varepsilon_n = (\alpha_{n+1} - \bar{\alpha}_{n+1})b_{n+1} + (\beta_{n+2} - \bar{\beta}_{n+2})b_{n+2}, \tag{3.4}$$

approximately, on neglecting second order small quantities such as $(\alpha_{n+1} - \bar{\alpha}_{n+1})(b_{n+1} - \bar{b}_{n+1})$ etc.

Let ϕ_n and ψ_n denote the errors in a_n and b_n respectively, i.e. we define

$$\begin{cases} \bar{a}_n = a_n + \phi_n \text{ and } \bar{b}_n = b_n + \psi_n \text{ for } n = 0(1)N, \\ \text{where } \phi_n = \psi_n = 0 \text{ for } n = N+1, N+2. \end{cases} \tag{3.5}$$

On subtracting equation (1.2) from (3.2), we find that the errors ψ_n are completely defined by

$$\begin{cases} \psi_n + \alpha_{n+1}\psi_{n+1} + \beta_{n+2}\psi_{n+2} = \phi_n + \varepsilon_n + r_n, \\ \text{where } \psi_{N+1} = \psi_{N+2} = 0. \end{cases} \tag{3.6}$$

Thus the quantities ψ_n satisfy a set of equations similar to that of the b_n, although with different right hand sides. On comparing with equation (1.4), we have immediately that

$$p_0\psi_0 + (p_1 + \alpha_1 p_0)\psi_1 = \sum_{m=0}^{N} (\phi_m + \varepsilon_m + r_m)p_m. \tag{3.7}$$

We shall make use of this result when we consider the errors in the evaluation of the sum P_N.

Finally, from equation (1.4), we actually compute a quantity $\bar{P}_N$ say, where

$$\bar{P}_N = [\bar{p}_0\bar{b}_0 + (\bar{p}_1 + \bar{\alpha}_1\bar{p}_0)\bar{b}_1] + s, \tag{3.8}$$

where the quantity s performs a similar role to that of r_n in equation (3.2),

https://doi.org/10.1017/S1446788700005267 Published online by Cambridge University Press

i.e., it is the round-off error introduced on evaluating the sum of products in the brackets [ ]. If we rewrite $\bar{P}_N$ as

$$\bar{P}_N = p_0\bar{b}_0+(p_1+\alpha_1 p_0)\bar{b}_1+\xi+s, \tag{3.9}$$

where ξ is defined by

$$\xi = (\bar{p}_0-p_0)\bar{b}_0+(\bar{p}_1-p_1)\bar{b}_1+[\bar{\alpha}_1(\bar{p}_0-p_0)+(\bar{\alpha}_1-\alpha_1)p_0]\bar{b}_1, \tag{3.10}$$

then on subtracting equation (1.4) from (3.9), we have

$$\bar{P}_N-P_N = p_0\psi_0+(p_1+\alpha_1 p_0)\psi_1+\xi+s. \tag{3.11}$$

If we now make use of equation (3.7), we have the required result,

$$\bar{P}_N-P_N = \sum_{m=0}^{N}(\phi_m+\varepsilon_m+\gamma_m)p_m+\xi+s. \tag{3.12}$$

Previous analyses ([1], [2] and [3]) have essentially given only the partial result that $\bar{P}_N-P_N = \sum_{m=0}^{N}\phi_m p_m$, and this is not necessarily the major contribution to the total error.

Our previous example on summing a series of Bessel functions is a good example of this. Here, the coefficients a_m are given exactly so that $\phi_m = 0$. Furthermore, $\varepsilon_m = 0$ since the coefficients α_m and β_m are also given exactly. Thus the error $\bar{P}_{12}-P_{12}$ is given by

$$\bar{P}_{12}-P_{12} = \sum_{m=0}^{12}\gamma_m p_m+\xi+s$$

in this example, and as we have already seen this is *not* negligible with respect to P_{12}.

The results derived in this and the previous section provide sufficient information for an estimate of the error in P_N to be made. In the next section we propose to give a complete analysis of the error which may arise when we sum a finite series of Chebyshev polynomials.

4. Errors in summing a Chebyshev series

One frequently needs to evaluate a sum of the form

$$P_N = \sum_{n=0}^{N} a_n T_n(x), \text{ for } -1 \leqq x \leqq 1, \tag{4.1}$$

where $T_n(x)$ is the Chebyshev polynomial of the first kind defined by

$$T_n(x) = \cos n\theta \text{ where } x = \cos\theta, \text{ for } n = 0, 1, 2, \cdots. \tag{4.2}$$

These polynomials satisfy the recurrence relation

$$T_n(x)-2xT_{n-1}(x)+T_{n-2}(x) = 0, \tag{4.3}$$

https://doi.org/10.1017/S1446788700005267 Published online by Cambridge University Press

a second solution of which are the Chebyshev polynomials of the second kind, $U_n(x)$ which are defined by

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}, \quad x = \cos\theta, \quad n = 0, 1, 2, \cdots. \tag{4.4}$$

It can be readily shown that, in the notation of Section 2,

$$\frac{p_{n-1}q_m - p_m q_{n-1}}{W_n} = U_{m-n}(x), \tag{4.5}$$

so that from equation (2.4), we have

$$b_n = \sum_{m=n}^{N} U_{m-n}(x)a_m. \tag{4.6}$$

In order to carry out the error analysis, we shall assume that the calculation is done in fixed point arithmetic to t decimal places. We shall further assume that $|r_n|, |s| \leqq \frac{1}{2}\times 10^{-t}$. Since, in general, the quantity x will not be represented exactly, let us assume that in the calculations it is replaced by $\bar{x}$, where $\bar{x}-x = \gamma$, say. Then since $\alpha_n = -2x$ we have $\alpha_n - \bar{\alpha}_n = 2\gamma$ for all n. Finally since $\beta_n = 1$, for all n, we shall assume that $\beta_n - \bar{\beta}_n = 0$. Equation (3.4) now gives

$$\varepsilon_n = 2\gamma b_{n+1}, \tag{4.7}$$

approximately. Since $p_0 = 1$ and $p_1 = x$ we shall assume that $\bar{p}_0 - p_0 = 0$ and $\bar{p}_1 - p_1 = \gamma$. Thus, from equation (3.10) we have

$$\xi = -\gamma b_1, \tag{4.8}$$

approximately, where we have replaced $\bar{b}_1$ by b_1. With these results, equation (3.12) gives

$$\bar{P}_N - P_N = \sum_{m=0}^{N} (\phi_m + r_m)T_m(x) + 2\gamma \sum_{m=0}^{N} b_{m+1}T_m(x) - \gamma b_1 + s. \tag{4.9}$$

Now the second sum on the right hand side of this equation may be summed explicitly, for

$$\begin{aligned}
\sum_{m=0}^{N} b_{m+1}T_m(x) &= \sum_{m=1}^{N} b_m T_{m-1}(x), \text{ since } b_{N+1} = 0, \\
&= \sum_{m=1}^{N} \left(\sum_{k=m}^{N} U_{k-m}(x)a_k \right) T_{m-1}(x), \text{ by equation (4.6)}, \\
&= \sum_{k=1}^{N} a_k \left(\sum_{m=1}^{k} U_{k-m}(x)T_{m-1}(x) \right), \\
&= \sum_{k=1}^{N} a_k \left(\frac{k+1}{2} \right) U_{k-1}(x).
\end{aligned}$$

https://doi.org/10.1017/S1446788700005267 Published online by Cambridge University Press

Thus equation (4.9) may be rewritten as

$$\bar{P}_N - P_N = \sum_{m=0}^{N} (\phi_m + r_m) T_m(x) + \gamma \sum_{m=1}^{N} m a_m U_{m-1}(x) + s. \tag{4.10}$$

At this point, we may obtain an upper bound for $|\bar{P}_N - P_N|$ by making use of the facts that for $-1 \leqq x \leqq 1$, $|T_m(x)| \leqq 1$ and $|U_{m-1}(x)| \leqq m$. If we further assume that both $|\phi_m|$ and $|\gamma|$ are less than $\frac{1}{2} \times 10^{-t}$, we have

$$|\bar{P}_N - P_N| \leqq \tfrac{1}{2} \times 10^{-t} \left\{ (2N+3) + \sum_{m=1}^{N} m^2 |a_m| \right\}. \tag{4.11}$$

At this point, we may recall Clenshaw's partial error analysis [1], which although given in terms of shifted Chebyshev polynomials would give for this problem $|\bar{P}_N - P_N| \leqq \frac{1}{2} \times 10^{-t} \times (2N+2)$. (Clenshaw introduced a factor of 2 to account for possible round-off errors in b_n). The additional term given in equation (4.11) is only likely to make a significant contribution to the error if we are summing a Chebyshev series where the coefficients a_n are slowly convergent. In numerical work a "slowly convergent" series would be one for which $a_n = 0(1/n^2)$. If we assume that there exists a constant A such that $|a_n| \leqq A/n^2$ for all $n \geqq 1$, then equation (4.11) gives the result that

$$|\bar{P}_N - P_N| \leqq \tfrac{1}{2} \times 10^{-t} \{(2+A)N + 3\}.$$

5. Conclusion

In this paper we have given a fairly complete discussion of the algorithm which may be used to sum a series $\sum_{n=0}^{N} a_n p_n$ where the p_n satisfy a three term recurrence relation. We have seen that the algorithm should not be used for series in which the second solution of the recurrence relation is considerably larger in modulus than $|p_n|$. For problems where the algorithm may be used, a complete discussion of the errors has been given, from which a bound may be obtained for the errors in the final sum.

6. Acknowledgement

This research has been sponsored in part by the United States Air Force Office of Scientific Research under Contract AF-AFOSR-660-64. The author wishes to thank Messrs. J. D. Donaldson and M. Rush for their comments.

References

[1] C. W. Clenshaw, 'A note on the summation of Chebyshev series', M.T.A.C. 9 (1955), 118–120.
[2] D. Elliott, 'On the expansion of functions in ultraspherical polynomials', *Journ. Aust. Math. Soc.* 1 (1960), 428–438.
[3] F. J. Smith, 'An algorithm for summing orthogonal polynomial series and their derivatives with applications to curve-fitting and interpolation', *Math. Comp.* 19 (1965), 33–36.
[4] L. M. Milne-Thomson, *The calculus of finite differences*, (Macmillan, London, 1933).
[5] F. W. J. Olver, 'Error analysis of Miller's recurrence algorithm', *Math. Comp.* 18 (1964), 65–74.

Mathematics Department
University of Tasmania
Hobart, Tasmania