

Chebyshev Methods for Ordinary Differential Equations

By L. FOX

The solution of linear ordinary differential systems, with polynomial coefficients, can be approximated by a finite Chebyshev series. The computation can be performed so that the solution satisfies exactly a perturbed differential system, the perturbations being computed multiples of one or more Chebyshev polynomials. An upper bound to the errors in the solution can often be estimated by approximate solution of the differential system satisfied by the error. Various devices are used to make the errors smaller, including a priori integration of the given differential equation and proper choice of the resulting constants of integration. The paper discusses various aspects of these topics, for both initial-value and boundary-value problems, and also suggests a method for automatic computation.

Introduction

1. Two methods have been proposed for solving ordinary differential equations which take advantage of the special properties of Chebyshev polynomials. For best convenience the equations are linear with polynomial coefficients, though they can often be adapted to other linear equations and, by iterative techniques, to non-linear equations. Here we consider only the favourable case, since our aim is to compare and contrast the two methods and to suggest some possible improvements in both.

Lanczos (1938) approximates to the solution by a polynomial, and determines the coefficients in that polynomial which satisfies the original differential equation perturbed by a small term or terms which are calculated as part of the process. The error satisfies a similar differential equation and its approximate solution can sometimes be obtained to provide an upper bound for the error. In other cases he increases the degree of the polynomial and compares results.

2. Clenshaw (1957) prefers to calculate directly the coefficients of the Chebyshev series, effectively approximating to the solution of an infinite set of equations for the coefficients by assuming that terms beyond a certain point are negligible and concentrating on the resulting finite set. He measures the accuracy of the result by estimating the number of figures to which his computed coefficients agree with those of the infinite Chebyshev series, which gives a bound to the rounding error, and adding the remainder term assessed by examining the rate of convergence of the series. If necessary he also repeats the computation with more coefficients included in the calculation, and often prefers to do this to guarantee more closely the maximum error of a finite series obtained by truncation to a smaller number of terms.

In this paper we compare with that of Lanczos a modification of the Clenshaw process, for obtaining a

finite approximation without truncating a higher-order approximation, and for this purpose we perform the arithmetic so accurately that the coefficients obtained can be regarded as exact solutions of the finite process used to find them. The arithmetic of the two methods is quite different, but we show that, in the absence of rounding error, and with careful attention in the Clenshaw process to the details of the reduction to a finite set of equations, the two methods give identical results. In the following we drop the term "modification" applied to our version of Clenshaw's method, since its persistent use might imply a completely different numerical technique. The main difference is in the method of estimating the error, and we refer to this briefly in the penultimate section of the paper.

3. We also consider delayed procedures, in which the methods are applied to integrated forms of the differential equation, showing that a certain device of Clenshaw is effectively equivalent to integration but also needs care for best accuracy. The delayed process is often advantageous, producing significantly better approximations for the same degree of polynomial, and the error analysis also is often simplified. Finally we suggest a process which may have extra advantages for automatic computation, and comment on some remarks of Clenshaw (1957), on the connection between his process and the number of convergent Chebyshev expansions which satisfy the differential equation.

The Lanczos Method

4. The Lanczos method has been described in the literature, and a particularly valuable account is contained in his book *Applied Analysis* (Lanczos, 1957). We give here the relevant details for the sake of completeness, for subsequent comparison with Clenshaw's method, and also to suggest a possible re-organization of the computation. An example will provide the details more conveniently.

The point $x = 0$ corresponds to $\cos \theta = -1$, $\sin \theta = 0$, and here

$$\tau_1 = (-1)^{n-1} \frac{\tau}{4(n^2 - 1)}$$

$$\text{so that } A = (-1)^n \frac{\tau}{4(n^2 - 1)} \tag{14}$$

We can now assert that the maximum error is hardly likely to exceed

$$\varepsilon_{max} = \frac{\tau}{4} \left\{ \frac{1}{n^2 - 1} + \frac{1}{2(n+1)} + \frac{1}{2(n-1)} \right\}, \tag{15}$$

which for $n = 4$ has the value $\tau/12 \approx 0.00030$. The point $x = 1$ is special, since here $\tau_1 = g_1 = g_2$, with a

maximum value of $\frac{\tau}{4} \left\{ \frac{1}{2(n-1)} + \frac{1}{2(n+1)} \right\}$. At this

point we would expect to find $\varepsilon_{max} \leq \frac{\tau}{2(n^2 - 1)} \approx 0.00012$

in this case. These general and special predictions are confirmed by the following errors, the differences between the true $y = (1+x)^{-1/2}$ and the approximation (7):

$$\begin{matrix} x & 0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1.0 \\ 10^5 \varepsilon & +2 & +16 & +23 & -17 & +6 & -4 & -7 & -2 & -7 & -8. \end{matrix} \tag{16}$$

8. This analysis breaks down if the coefficient of x^r in the equation for the error, vanishes at any point in the range, though it will still be valid for values of x sufficiently remote from this. Sometimes, also, a single τ -term will not suffice, since one or more equations are still unsatisfied. These points are illustrated by the approximate solution in the range (0, 1) of the equation

$$x^2 y'' - y = 0, \quad y(1) = 1, \tag{17}$$

whose exact solution is e^{1-x} . The assumption (2) gives the equations

$$\left. \begin{aligned} a_0 &= 0 \\ a_1 &= 0 \\ \tau a_r - a_{r-1} &= 0, \quad r = 1, 2, \dots, n-1 \\ na_n &= 0 \end{aligned} \right\}, \tag{18}$$

of which the last equation refers to the coefficient of $x^n = 1$. We should therefore need to add a term τT_n^* to the differential equation, but a_1 can then be determined from two equations of the set (18) (in Lanczos' phraseology a_1 is "overdetermined") so that we take $\tau_1 T_{n-1} + \tau_0 T_n^*$ as perturbations, the extra equation for a_1 , that is the second of (18), and the boundary condition giving two linear equations for the determination of τ_1 and τ_0 . In passing we note the desirability of keeping as large as possible the order of the perturbing Chebyshev polynomials, so that we prefer this perturbation, for example, rather than any other combination such as $\tau_1 T_{n-1} + \tau_0 T_2^*$.

We then find

$$\begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & \dots \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \dots \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{\tau_0 C_n^{(0)}}{\tau_0 C_n^{(2)}} + \tau_1 C_{n-1}^{(0)} & \dots & \dots \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{\tau_0 C_n^{(2)}}{\tau_0 C_n^{(4)}} + \tau_1 C_{n-1}^{(2)} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{1}{4} & \dots & \dots & \dots & \dots & \tau_1 C_{n-1}^{(n-1)} \end{bmatrix}, \tag{19}$$

with the extra equation

$$-a_1 = \tau_0 C_n^{(1)} + \tau_1 C_{n-1}, \tag{20}$$

and the boundary condition

$$\sum_{r=0}^n a_r = 1. \tag{21}$$

With $n = 4$ we calculate $\tau_0 = -27/2907 = -0.0093 \dots$, $\tau_1 = 32/2907 = 0.0110 \dots$, and

$$y = \frac{1}{2907} (59 - 2464x + 14656x^2 - 13440x^3 + 4096x^4). \tag{22}$$

9. The error satisfies the differential system

$$x^2 z'' - z = -\tau_0 T_n^*(x) - \tau_1 T_{n-1}^*(x), \quad z(1) = 0, \tag{23}$$

and it is easy to see that, since $|z|$ is a maximum at $x = 0$ or $x = 1$ or whenever $z' = 0$, we have the certainty that $|z|$ cannot exceed the maximum of $|\tau_0 T_n^*(x)| + |\tau_1 T_{n-1}^*(x)|$. With τ_0 and τ_1 of opposite sign, as here, the maximum is at $x = 0$ where $z = (-1)^n (\tau_0 - \tau_1)$, in this case -0.0203 , the value of $-y(0)$.

10. When the maximum error or its close estimate can be obtained by these processes, in terms of the parameters of type τ , it is clear that we do not need to calculate more than one or two of the individual a_r , until the value of n has been decided, quantities like $\sum a_r$ being obtained by summing the columns of the matrix. The computation of the τ process, in this matrix form, is therefore easily arranged and not excessive.

11. Similar techniques can be used for second-order equations. For example the function $y = e^{x^2}$ satisfies the system

$$y'' - 2(1 + 2x^2)y = 0, \quad y(0) = 1, \quad y'(0) = 0, \tag{24}$$

and we seek an even solution with the approximation

$$y = a_0 + a_2 x^2 + a_4 x^4 + \dots + a_{2n} x^{2n}, \tag{25}$$

which already satisfies the second given condition. Substitution in (24) gives the equations

$$\left. \begin{aligned} 2a_2 - 2a_0 &= 0 \\ 2r(2r-1)a_{2r} - 2a_{2r-2} - 4a_{2r-4} &= 0, \\ & \quad r = 2, 3, \dots, n \\ -2a_{2n} - 4a_{2n-2} &= 0 \\ & \quad -4a_{2n} = 0 \end{aligned} \right\}, \tag{26}$$

the last equation corresponding to the coefficient of

Table 1

a_0	1	1	13	43	771	4641	$\tau_0 C_{2n}^{(2)} + \tau_2 C_{2n}^{(2)}$
	-4	8	16	32	64	128	
a_2	1	1	32	87	2877		$\tau_0 C_{2n}^{(4)} - \tau_2 C_{2n}^{(4)}$
	-4	8	16	32	64		
...	1	1	1	1	57	147	...
	-4	8	16	32	64	32	...
...	1	1	1	1	91		...
	-4	8	16	32	64	16	...
...	1	1	1	1	1	1	...
	-4	8	16	32	64	8	...
a_{2n}	1	1	1	1	1	1	$\tau_2 C_{2n}^{(2n/2)}$
	-4	8	16	32	64	4	

x^{2n+2} . The first equation represents a single over-determination, and we must also satisfy the first given condition, so that we must add the terms $\tau_2 T_{2n+2}(x) - \tau_0 T_{2n}(x)$ on the right of the differential equation.

We then find equation 27 (see Table 1).

Taking $2n = 10$ we calculate $\tau_0 = -0.00066 \dots$

$$\tau_2 = -0.00003 \dots, \text{ and} \tag{28}$$

$$y = 1 + 1.0003161x^2 + 0.4974742x^4 + 0.1745248x^6 + 0.0304575x^8 + 0.0155269x^{10},$$

in which the coefficients, for simplicity, are given as rounded versions, to seven decimals, of their exact fractional values.

12. The error analysis for this case is more difficult, but we can get a good estimate by solving iteratively the equation

$$z'' - 2(1 + 2x^2)z' - \tau_0 T_{2n} - \tau_2 T_{2n+2} \tag{29}$$

in the form

$$z'' = 2(1 + 2x^2)z' - \tau_0 T_{2n} \tag{30}$$

and neglecting the τ_2 term, since τ_2 is here very much smaller than τ_0 . We may note, however, that the inclusion of τ_2 increases only slightly the complexity of the analysis.

With $z_0 = 0$ we find, by double integration and satisfaction of $z(0) = 0$, the next term

$$z_1 = \frac{1}{4} \tau_0 \left[\frac{2T_{2n}}{4n^2 - 1} - \frac{T_{2n+2}}{(2n + 1)(2n + 2)} - \frac{T_{2n-2}}{(2n - 1)(2n - 2)} + C \right],$$

$$C = (-1)^n \frac{1}{n^2 - 1}. \tag{31}$$

For the next iterate the part involving the double integral of $2(1 + 2x^2)$ multiplying the T terms will be of order

n^{-4} , and can reasonably be neglected. The C term will, however, make significant contribution here and in later terms. The first four contributions to the total are found to be

$$C = C(x^2 - \frac{1}{3}x^4) + C(\frac{1}{6}x^4 - \frac{7}{45}x^6 + \frac{1}{42}x^8),$$

$$C(\frac{x^6}{90} + \frac{11x^8}{630} - \frac{211x^{10}}{56700} + \frac{x^{12}}{462}). \tag{32}$$

and they are decreasing quite rapidly. The total contribution is, of course, Ce^{x^2} , but such knowledge is not generally available and is in any case unnecessary.

The error is probably a maximum at $x = 1$, and for its estimate we have

$$z_{max} = \frac{1}{4} \tau_0 \left[-\frac{3}{(4n^2 - 1)(n^2 - 1)} + \frac{(-1)^{n+1}}{n^2 - 1} (1 + 1.33 + 0.35 + 0.03) \right], \tag{33}$$

giving, for $n = 5$, the result

$$z_{max} = 0.028\tau_0 = -0.000018. \tag{34}$$

The accuracy of this prediction is verified by the following table of errors:

x	0.0	0.1	0.2	0.3	0.4	0.5
$10^6 z$	0	-3	-9	-13	-11	-7
		0.6	0.7	0.8	0.9	1.0
		-5	-9	-15	-15	-18.

13. The approximation (28) is clearly not the best possible result, its error being of one sign, and the reason for this becomes apparent from inspection of (31). It is clear that, instead of satisfying exactly the boundary condition $y(0) = 1$, we would do better to satisfy

$$y(0) = a_0 = 1 - \frac{(-1)^n \tau_0}{4(n^2 - 1)}, \tag{36}$$

so that C is zero in (31). We then find the result

$$y = 0.99999931 + 1.0003092x^2 + 0.4974708x^4 + 0.1745236x^6 - 0.0304573x^8 + 0.0155268x^{10}, \quad (37)$$

with the following smaller errors:

x	0.0	0.1	0.2	0.3	0.4	0.5		
$10^6 \epsilon$	-7	-4	-2	-5	-3	+2		
	0.6	0.7	0.8	0.9	1.0			
	+5	+2	-2	0	-1			

Lanczos (1957) has made similar comments with other error analysis for first-order equations, and a similar device can improve somewhat the approximation (7) for the problem of Section 6.

The Clenshaw Method

14. Clenshaw finds directly the coefficients of the Chebyshev series, or at least approximations to them, by assuming such an expansion for y , with coefficients $a_r^{(0)}$, and similar expansions for the derivatives, with coefficients $a_r^{(s)}$ for the s th derivative. In this way he avoids the necessity for differentiating Chebyshev polynomials, and can relate the coefficients of type $a^{(s)}$ to those of type $a^{(s-1)}$ by integration formulae. The product of powers of x with Chebyshev polynomials is expressible in terms of Chebyshev polynomials, so that it is possible to substitute in the differential equation, equate to zero successive coefficients of the polynomials, and solve the resulting infinite set of algebraic equations by truncation to a finite set.

The finite set is solved conveniently by a recurrence process, almost equivalent to standard "back substitution," since the matrix of coefficients is triangular or almost triangular.

15. For example, in the problem of Section 4 we take

$$y = \frac{1}{2}a_0^{(0)}T_0^* + a_1^{(0)}T_1^* + a_2^{(0)}T_2^* + \dots \quad (39)$$

$$y' = \frac{1}{2}a_0^{(1)}T_0^* + a_1^{(1)}T_1^* + a_2^{(1)}T_2^* + \dots,$$

substitute in the differential equation (1), and determine equations for the coefficients, given by

$$\frac{3}{2}a_0^{(1)} + \frac{1}{2}a_1^{(1)} + \frac{1}{2}a_0^{(0)} = 0 \quad (40)$$

$$3a_r^{(1)} + \frac{1}{2}a_{r-1}^{(1)} + \frac{1}{2}a_{r+1}^{(1)} + a_r^{(0)} = 0, \quad r = 1, 2, \dots,$$

In addition we have the general integral relation

$$a_{r-1}^{(1)} = 4ra_r^{(0)} + a_{r-1}^{(0)}, \quad (41)$$

and if we use this in (40), to give

$$(2r + 1)a_r^{(0)} = -3a_{r-1}^{(0)} + a_{r-2}^{(0)}, \quad (42)$$

we can apply (42) and (41), in turn, to obtain $a_r^{(0)}$ and $a_{r-1}^{(1)}$ in terms of coefficients of higher order in the two series.

Starting at some point $r = n$, with the assumption $a_n^{(0)} = 1, a_{n-1}^{(1)} = 0$, and all terms of higher order assumed negligible, we can calculate all the coefficients and finally determine an appropriate multiplying factor so that the

Table 2

r	$a_r^{(0)}$	$a_r^{(1)}$
0	145488/35	-52992/35
1	-12528/35	-13488/35
2	1616/35	-576/7
3	-48/7	16
4	1	0

initial condition is satisfied. Starting with $n = 4$, for example, we find the results of Table 2. The boundary condition gives

$$\frac{1}{2}a_0^{(0)} - a_1^{(0)} + a_2^{(0)} - a_3^{(0)} + a_4^{(0)} - \dots = 1, \quad (43)$$

so that we multiply the results of Table 2 by 35/87163, and have the solution

$$y = \frac{1}{87163} \{ 72744T_0^* - 12528T_1^* + 1616T_2^* - 240T_3^* + 35T_4^* \}$$

$$- 0.83457T_0^* - 0.14373T_1^* + 0.01854T_2^* - 0.00275T_3^* + 0.00040T_4^*, \quad (44)$$

rounded to five decimals.

16. We note two interesting points. First, this result is identical with the polynomial approximation (7), and the multiplying factor is equal to $\tau/9$ of that process. In fact, if we deliberately truncate the assumptions (39) at T_n^* for y , and T_{n-1}^* for y' , so that y' is exactly the derivative of y , the set of equations (40) is finite, the last two of them having the form

$$(2n - 1)a_n^{(0)} - 3a_{n-1}^{(1)} = 0$$

$$(2n + 1)a_n^{(0)} = 0.$$

By adding the constant τ to the right of the last equation, which is equivalent to adding τT_n^* to the right of the differential equation, and proceeding with the recurrence as before, we find all the results of Table 2 multiplied by $1/9$, and the multiplying factor is then exactly the τ of the polynomial method.

17. Second, we observe the good convergence of the Chebyshev series, with coefficients much smaller than those of the polynomial. We must, however, remember that the coefficients in (44) are only approximations to those of the infinite series, so that although the trend of the coefficients indicates that the next two terms of that series, to five decimals, are something like $-0.00006T_5^* + 0.000017T_6^*$, the maximum discrepancy between our approximation (44) and the true solution cannot be estimated at about 7×10^{-5} .

As we have seen the error is as much as 23×10^{-5} , and this can be reduced only by using a larger value of n to find more accurate approximations to the Chebyshev coefficients. (We refer in Section 45 to Clenshaw's method of assessing the number of figures in the approximate coefficients which agree with those of the infinite series.)

For example, if we take $n = 5$ we find

$$\begin{aligned} \tau &= -231/339323 \doteq -0.00068, \text{ and} \\ y &= 0.834621T_0^* - 0.143733T_1^* + 0.018519T_2^* \\ &\quad - 0.002652T_3^* + 0.000413T_4^* - 0.000062T_5^*, \end{aligned} \quad (46)$$

with significant changes, for example, in the fourth decimal of a_3 . The analysis of Section 7 would give maximum error of 0.00004, and an error at $x = 1$ of little more than 0.00001. These figures are confirmed, as good upper estimates, by the following table of errors:

x	0.0	0.1	0.2	0.3	0.4	0.5
$10^5 \epsilon$	0	± 1	± 3	± 2	-1	-2
	0.6	0.7	0.8	0.9	1.0	
	-1	± 1	± 2	± 1	0	

18. The Clenshaw method will need to combine solutions, as in the Lanczos process, whenever any equation is left unsatisfied, and the same problem will affect both methods in the same way. For example, the problem of Section 8, with the assumption (39) for the solution, gives for the coefficients the equations

$$\left. \begin{aligned} \frac{1}{16}a_2^{(1)} + \frac{1}{4}a_1^{(1)} + \frac{3}{16}a_0^{(1)} - \frac{1}{2}a_0^{(0)} &= 0 \\ \frac{1}{16}a_3^{(1)} + \frac{1}{4}a_2^{(1)} + \frac{7}{16}a_1^{(1)} + \frac{1}{4}a_0^{(1)} - a_1^{(0)} &= 0 \end{aligned} \right\} \quad (48)$$

$$\left. \begin{aligned} \frac{1}{16}a_{r+4}^{(1)} + \frac{1}{4}a_{r+3}^{(1)} + \frac{3}{8}a_{r+2}^{(1)} - \frac{1}{4}a_{r+1}^{(1)} + \frac{1}{16}a_r^{(1)} \\ - a_{r+2}^{(0)} = 0, \quad r = 0, 1, \dots, j \end{aligned} \right\}$$

the integrating equations

$$a_{r-1}^{(1)} = 4r a_r^{(0)} + a_{r+1}^{(1)} \quad (49)$$

and the boundary condition

$$\frac{1}{2}a_0^{(0)} + a_1^{(0)} + a_2^{(0)} + \dots = 1. \quad (50)$$

Again it is convenient to express $a_1^{(0)}$ in the third of (48) in terms of $a_{r+2}^{(0)}$ and $a_{r+1}^{(0)}$ from (49), giving the general equation

$$\begin{aligned} \frac{1}{16}a_{r+4}^{(1)} + \frac{1}{4}a_{r+3}^{(1)} + \frac{7}{16}a_{r+2}^{(1)} + \frac{1}{4}a_{r+1}^{(1)} + \frac{1}{4}(r+1)a_{r+1}^{(0)} \\ - a_{r+2}^{(0)} = 0, \quad r = 0, 1, \dots \end{aligned} \quad (51)$$

Similar treatment of the second of (48) gives

$$\frac{1}{16}a_3^{(1)} + \frac{1}{2}a_2^{(1)} + \frac{7}{16}a_1^{(1)} = 0, \quad (52)$$

and we note that the term in $a_1^{(0)}$ has disappeared. We can calculate all the coefficients by recurrence, from $r = n$ down to $r = 1$, using in turn (51) and (49). The coefficient $a_0^{(0)}$ then comes from the first of (48), and two equations, (50) and (52), remain unsatisfied. We therefore need two independent trial solutions, and a linear combination of them will yield the final solution.

19. The nature of the starting conditions is of some importance. We might, for example, start first with

Table 3

r	$a_r^{(0)}$	$a_r^{(1)}$	r	$a_r^{(0)}$	$a_r^{(1)}$
0	58	128	0	$\frac{1625}{24}$	$\frac{425}{3}$
1	44	32	1	$\frac{587}{12}$	$\frac{128}{3}$
2	2	-48	2	$\frac{10}{3}$	-54
3	-4	16	3	$\frac{55}{12}$	16
4	1	0	4	1	1

$a_n^{(0)} = 1$, $a_n^{(1)} = 0$, and all higher coefficients assumed zero, and then perhaps with $a_n^{(0)} = 1$, $a_n^{(1)} = 1$, with all higher coefficients neglected. We then find, for $n = 4$, the results of Table 3.

For the factors A and B multiplying the respective solutions we compute the values $A = -1056/30915$, $B = 1296/30915$, and then

$$\begin{aligned} y &= \frac{1}{30915}(13251T_0^* + 16932T_1^* + 2208T_2^* \\ &\quad - 1716T_3^* + 240T_4^*) \\ &= 0.4286T_0^* + 0.5477T_1^* + 0.0714T_2^* \\ &\quad - 0.0555T_3^* + 0.0078T_4^*, \end{aligned} \quad (53)$$

which has an error at $x = 0$ of amount 0.0156.

20. The τ -method applied to Clenshaw's process, starting with a finite approximation, that is truncating (39) after T_n^* and the corresponding expression for the derivative after T_{n-1}^* , would lead to a finite set of equations for the coefficients, of which the last three are

$$\left. \begin{aligned} \frac{7}{16}a_{n-1}^{(1)} + \frac{1}{4}a_{n-2}^{(1)} + \frac{1}{4}(n-2)a_{n-2}^{(0)} - a_{n-1}^{(0)} &= 0 \\ \frac{1}{4}a_{n-1}^{(1)} + \frac{1}{4}(n-1)a_{n-1}^{(0)} - a_n^{(0)} &= 0 \\ \frac{1}{4}na_{n-1}^{(0)} &= 0 \end{aligned} \right\} \quad (54)$$

Since we need two τ -terms, and the finite expansion gives a term in x^{n-1} in the differential equation, we take the form $\tau_1 T_{n-1}^* + \tau_0 T_n^*$, which means that τ_0 and τ_1 are inserted respectively on the right of the last two equations of (54). The two trial solutions then correspond exactly to the starting conditions

$$a_n^{(0)} = 1, a_n^{(1)} = 0, a_{n-1}^{(0)} = 1, a_{n-1}^{(1)} = 0; \quad (55)$$

and if they are multiplied respectively by A and B , to satisfy the remaining equations, the resulting finite solution satisfies exactly the system

$$\begin{aligned} x^2 y' - y - \tau_1 T_{n-1}^*(x) + \tau_0 T_n^*(x), \quad y(1) = 1, \\ \tau_1 = \frac{nA}{4}, \quad \tau_0 = \frac{(n-1)B}{4}, \end{aligned} \quad (56)$$

and is identical with the corresponding result for the Lanczos polynomial method.

Table 4

r	$a_r^{(0)}$	$a_r^{(1)}$	r	$a_r^{(0)}$	$a_r^{(1)}$
0	58	128	0	$\frac{35}{-2}$	-8
1	44	32	1	-5	-32
2	2	-48	2	-4	12
3	-4	16	3	1	0
4	1	0			

We find the results of Table 4, with $n = 4$. Satisfaction of the two remaining equations gives $A = 32/2907$, $B = -36/2907$, and

$$\begin{aligned}
 y &= \frac{1}{2907}(1243T_0^* + 1588T_1^* + 208T_2^* \\
 &\quad - 164T_3^* + 32T_4^*) \\
 &= 0.4276T_0^* + 0.5463T_1^* + 0.0716T_2^* \\
 &\quad - 0.0564T_3^* + 0.0110T_4^*. \quad (57)
 \end{aligned}$$

We see from (56) that $\tau_1 = 32/2907$, $\tau_0 = -27/2907$, as before, and the error analysis is also identical in all respects with the polynomial method of Section 8.

21. The analysis of the previous result (53) is more complicated, but we can see that the expression for the derivative, obtained from Table 3 and the multiplying factors, cannot be exactly the derivative of (53). It is in fact the derivative of $y + \frac{1}{20}BT_5^*$; we have effectively started at $n = 5$, with an extra term, and our solution satisfies exactly the system

$$\begin{aligned}
 x^2y' - y &= \left(A + \frac{5}{4}B\right)T_5^* + \frac{1}{16}BT_6^* - \frac{1}{20}Bx^2\frac{d}{dx}T_5^*, \\
 y(1) &= 1. \quad (58)
 \end{aligned}$$

The maximum error is now estimated less conveniently. 22. If we carry out the τ -process with an extra term, properly as in Section 20, we find

$$\begin{aligned}
 y &= 0.4277T_0^* + 0.5465T_1^* + 0.0713T_2^* \\
 &\quad - 0.0554T_3^* + 0.0077T_4^* + 0.0021T_5^*, \quad (59)
 \end{aligned}$$

which gives rise to a term $0.0161T_5^* + 0.0026T_6^*$ on the right-hand side of the differential equation, and has a somewhat smaller maximum error than the approximation (53).

23. For second-order equations the possible choice in starting the Clenshaw process can be settled in a similar manner, and generally with advantage, by borrowing the idea of the τ -method. For the example of Section 11, for instance, we assume the expansions

$$\left. \begin{aligned}
 y &= \frac{1}{2}a_0^{(0)}T_0 + a_2^{(0)}T_2 + a_4^{(0)}T_4 + \dots \\
 y' &= a_1^{(0)}T_1 + a_3^{(0)}T_3 + \dots \\
 y'' &= \frac{1}{2}a_0^{(2)}T_0 + a_2^{(2)}T_2 + \dots \end{aligned} \right\}, \quad (60)$$

and produce for solution the equations

$$\left. \begin{aligned}
 \frac{1}{2}a_0^{(2)} - 2a_0^{(0)} - a_2^{(0)} - a_4^{(0)} &= 0 \\
 a_1^{(2)} - 5a_1^{(0)} - a_3^{(0)} &= 0 \\
 a_2^{(2)} - 4a_2^{(0)} - a_{r-2}^{(0)} - a_{r-2}^{(0)} &= 0, \quad r = 2, 3, \dots \end{aligned} \right\}, \quad (61)$$

together with the integrating equations

$$\left. \begin{aligned}
 a_{r-1}^{(2)} &= a_{r+1}^{(2)} + 2ra_r^{(1)} \\
 a_{r-1}^{(0)} &= a_{r+1}^{(0)} + 2ra_r^{(0)} \end{aligned} \right\}, \quad (62)$$

and the remaining boundary condition

$$y(0) = \frac{1}{2}a_0^{(0)} - a_2^{(0)} + a_4^{(0)} - \dots = 1. \quad (63)$$

Inspection reveals an overdeterminacy in the first of (61) and, with the boundary condition still to be satisfied, two trial solutions and a linear combination thereof are necessary. Clenshaw (1957) starts the first solution with $a_{12}^{(0)} = 1$, $a_{12}^{(2)} = 0$ and all higher coefficients assumed negligible. For the second he states that the obvious $a_{12}^{(0)} = 0$, $a_{12}^{(2)} = 1$ gives a result too similar to the first. Making their combination uncertain through ill-conditioning, and therefore takes $a_{12}^{(0)} = 1$, $a_{12}^{(2)} = 10$. His final result (corrected here for an unnecessary rounding error in his determination of the equations for the constants) is

$$\begin{aligned}
 y &= 1.75338750T_0 + 0.85039158T_2 + 0.10520868T_4 \\
 &\quad + 0.00872210T_6 + 0.00054344T_8 \\
 &\quad + 0.00002712T_{10} - 0.00000118T_{12}, \quad (64)
 \end{aligned}$$

with a maximum error at $x = 1$ of about 2.3×10^{-7} .

24. The τ -method would have no starting problem. For a polynomial of degree $2n$ we would truncate (60) after T_{2n} , T_{2n-1} and T_{2n-2} respectively, the last three equations of the set (61) would be

$$\left. \begin{aligned}
 a_{2n-2}^{(2)} - 4a_{2n-2}^{(0)} - a_{2n-4}^{(0)} - a_{2n}^{(0)} - 0 \\
 \quad \text{(coefficient of } T_{2n-2}) \\
 - 4a_{2n}^{(0)} - a_{2n-2}^{(0)} - \tau_0 \\
 \quad \text{(coefficient of } T_{2n}) \\
 - a_{2n}^{(0)} - \tau_2 \\
 \quad \text{(coefficient of } T_{2n-2}) \end{aligned} \right\}, \quad (65)$$

and we have added $\tau_0 T_{2n} + \tau_2 T_{2n+2}$ to the right-hand side of the differential equation. With $n = 5$ we find the Chebyshev rearrangement of the result (28) of the polynomial method, and with $n = 6$ we find

$$\begin{aligned}
 y &= 1.75338727T_0 + 0.85039147T_2 + 0.10520867T_4 \\
 &\quad + 0.00872210T_6 + 0.00054344T_8 \\
 &\quad + 0.00002704T_{10} + 0.00000124T_{12}, \quad (66)
 \end{aligned}$$

with $\tau_0 \doteq 3.2 \times 10^{-5}$, $\tau_2 \doteq 1.2 \times 10^{-6}$, and the coefficients correctly rounded to eight decimals. The error at $x = 1$ is 6.0×10^{-7} , confirming incidentally the estimate (33) which gives exactly this result.

25. Clenshaw's result (64) is somewhat more accurate than (66), but when we examine his process more closely we find that it is effectively equivalent to adding an

extra term. Indeed, a term in T_{14} in the function is necessary to make it possible for the second derivative to have a term in T_{12} . We find that Clenshaw has effectively added to the right-hand side of the differential equation the terms

$$\frac{B}{364}(5T_{16} - 384T_{14} + 5T_{12}) \div AT_{14}, \tag{67}$$

with $B \doteq 3 \cdot 1 \times 10^{-6}$, $A \doteq 2 \cdot 0 \times 10^{-6}$,

has satisfied the boundary condition $y(0) = 1 \div \frac{5}{364}B$, and gave $y = \frac{5}{364}BT_{14}(x)$ as the solution, y being a polynomial of degree fourteen. The inclusion of the extra term with the τ process involves a perturbation of amount $\tau_0 T_{14} + \tau_2 T_{16}$ on the right of the differential equation, and we find $\tau_0 \doteq 1 \cdot 3 \times 10^{-6}$, $\tau_2 \doteq 0 \cdot 4 \times 10^{-7}$, and a result $y = 1 \cdot 753387666T_0 + 0 \cdot 850391659T_2 \div 0 \cdot 105208694T_4 \div 0 \cdot 008722105T_6 \div 0 \cdot 000543438T_8 \div 0 \cdot 000027114T_{10} \div 0 \cdot 000001126T_{12} + 0 \cdot 0000000044T_{14}$, (68)

the estimate (33) of maximum error giving the value $-1 \cdot 8 \times 10^{-8}$. This method of including the extra term, in fact, gives a full extra correct figure.

The device of Section 13, of course, could be used also in the τ -application of Clenshaw's method to reduce the error still further.

The Integrated Equations

26. The τ -method, with either process, ensures that the differential equation is satisfied within small and known amounts. Error analysis of the kinds indicated in Sections 7 and 12 may then sometimes give the error in the approximate solution. This error may involve several T terms, as for example in (31), through the single or double integration of the T terms in the perturbed differential equation. The inclusion of an arbitrary constant, also as in (31), may be a further substantial factor whose elimination, as in (36), involves extra work.

We may therefore get better results by applying the τ process to integrated forms of the original differential equation, so that boundary conditions are already satisfied and the minimum number of T terms is necessary in the perturbation.

27. Lanczos has already effectively observed this fact but, instead of integrating the original equation, suggested the incorporation of terms like $(d/dx)T(x)$ as perturbation of the original. This seems unnecessary and unduly complicated, involving the coefficients of the Chebyshev polynomials of the second kind, and it would appear, in fact, that the integration can always be performed. Clenshaw also sometimes carries out an integration, but in a disguised way which, as we shall see, needs care to give best results.

The Lanczos Integration

28. We consider first the problem presented by equation (1), and integrate the equation to produce

$$2(1 \div x)y = \int_0^x y dx = 2, \tag{69}$$

the constant on the right being adjusted to satisfy the initial condition $y(0) = 1$. The assumption (2) then gives the equations

$$\left. \begin{aligned} 2a_0 &= 2 \\ \left(2 \div \frac{1}{r}\right)a_{r-1} \div 2a_r &= 0, \quad r = 1, 2, \dots, n \\ \left(2 \div \frac{1}{n+1}\right)a_n &= 0 \end{aligned} \right\} \tag{70}$$

The coefficients on the left are multiples of those of the set (3) coming from the original differential equation, but for the non-trivial solution of (70) we now add the term $\tau T_{n+1}(x)$ to the right of (69), since the last of (70) refers to the coefficient of x^{n+1} . We can find all the coefficients in terms of τ , and its value then comes from the satisfaction of the first of (70), perturbed by $\tau C_{n+1}^{(0)}$, so that our approximation does not satisfy exactly the correct boundary condition.

With $n = 4$ we find $\tau = 126/725339 \doteq 17 \cdot 4 \times 10^{-5}$, and the solution

$$y = 0 \cdot 999913 - 0 \cdot 495614x \div 0 \cdot 336968x^2 - 0 \cdot 183528x^3 \div 0 \cdot 049411x^4. \tag{71}$$

The errors, considerably smaller than those of the previous method given in (16), are as follows:

x	0.0	0.1	0.2	0.3	0.4	0.5
$10^5 z$	-9	-8	-1	-6	+5	0
	0.6	0.7	0.8	0.9	1.0	
	-5	-5	0	-4	-4	

(72)

We note the expected increase in the number of changes of sign, and the fact that the error satisfies the equation

$$2(1 \div x)z = \int_0^x z dx = -\tau T_5'(x), \tag{73}$$

with a value of $-\tau/2 \doteq 9 \times 10^{-5}$ at $x = 0$. In fact the error is almost exactly $-\frac{1}{2(1-x)}\tau T_5'(x)$ everywhere.

29. Similar methods can be applied to second-order equations, and it seems a permanent feature of the polynomial method that the resulting equations for the coefficients are the same apart from constant factors. For the example of equation (24), with the assumption (25), a single integration gives

$$y' = 2 \int_0^x (1 + 2x^2)y dx = 0, \tag{74}$$

and, if the second term in (74) is called $g(x)$, a second integration gives

$$y' - \int_0^x g(x) dx = 1, \tag{75}$$

and the boundary conditions are satisfied.

The equations for the coefficients become

$$\begin{aligned} & a_0 = 1 \\ & a_2 = a_0 = 0 \\ & a_{2r} = \frac{1}{2r(2r-1)}(2a_{2r-2} + 4a_{2r-4}) = 0, \\ & r = 2, 3, \dots, n \end{aligned} \tag{76}$$

$$\begin{aligned} & - \frac{1}{(2n+2)(2n-1)}(2a_{2n} + 4a_{2n-2}) = 0 \\ & - \frac{1}{(2n+4)(2n-3)}4a_{2n} = 0 \end{aligned}$$

which may be compared with (26) for the original equation. By inserting terms on the right we can find all the coefficients without satisfying the first two of (76), so that two τ -terms are needed, and they must include a coefficient of x^{2n-4} , represented by the last of (76).

The choice depends on our requirements. If we must satisfy exactly the condition $y'(0) = 1$ we take $x^2(\tau_2 T_{2n-2} + \tau_0 T_{2n})$, which we note will tend to minimize the error in the neighbourhood of the origin. Otherwise we take $\tau_4 T_{2n-4} + \tau_2 T_{2n-2}$, and accept an error of $\tau_4 - \tau_2$ at the origin, with the probability of a more uniform error distribution. (The same choice, incidentally, existed in the previous example, and we took the second alternative.)

30. The first choice gives a result

$$y = 1 + 0.9999999x^2 + 0.4997978x^4 + 0.1682347x^6 + 0.0374490x^8 + 0.0127904x^{10}, \tag{77}$$

with $\tau_0 \doteq 4.4 \times 10^{-6}$, $\tau_2 \doteq 1.4 \times 10^{-7}$, and a maximum error of 10×10^{-6} at $x = 1$. This is about half the maximum error of the previous result (28).

The second choice gives

$$\begin{aligned} y = & 0.9999989 + 1.0000773x^2 + 0.4991135x^4 \\ & + 0.1703526x^6 + 0.0348576x^8 \\ & + 0.0138807x^{10}, \end{aligned} \tag{78}$$

with $\tau_4 \doteq -4 \times 10^{-8}$, $\tau_2 = -1.1 \times 10^{-6}$, and a very small error, distributed even more favourably than that obtained by the special device of Section 13. The errors are as follows, in units of the *seventh* decimal, showing that this process has produced a solution with a whole figure better than that of (28) and at least half a figure better than that of (37):

x	0.0	0.1	0.2	0.3	0.4	0.5
$10^7 \epsilon$	+11	-4	-9	-9	-3	-11
	0.6	0.7	0.8	0.9	1.0	
	-2	-11	+2	+8	+12.	

The Clenshaw Integration

31. The same integration processes can of course be used to produce directly the coefficients of a finite Chebyshev series, and there are several points of interest. First, the choice of perturbing terms involving only the T_r or T_r^* polynomials is more convenient than terms involving powers of x . Second, the equations for the coefficients coming from the integrated equations are usually more complicated than those from the original. For example, the coefficient of $T_r(x)$ on the left of (75) is

$$\begin{aligned} & \frac{1}{4r(r^2-1)} \{ -(r+1)a_{r-4}^{(0)} - 2(r+2)a_{r-2}^{(0)} \\ & + 2r(2r^2+1)a_r^{(0)} - 2(r-2)a_{r+2}^{(0)} \} \\ & - (r-1)a_{r+4}^{(0)}, \end{aligned} \tag{80}$$

which should be compared for complexity with the third of (61).

32. The extra complexity may be negligible, however, if the equation has a special form, and in this case a single integration corresponds to a device used by Clenshaw for removing coefficients of the highest derivative. In Clenshaw (1957) he examines the differential system

$$xy'' + y' + 16xy = 0, y(0) = 1, y'(0) = 0. \tag{81}$$

and obtains, for coefficients of expansions in T polynomials, the equations

$$\frac{1}{2}(a_{r-1}^{(2)} + a_{r-1}^{(2)}) + a_r^{(1)} + 8(a_{r-1}^{(0)} + a_{r-1}^{(0)}) = 0. \tag{82}$$

He then observes, by changing r to $r+2$ in (82), subtracting the two equations and using integrating relations like

$$a_{r-1}^{(s)} - a_{r-1}^{(s)} = 2ra_r^{(s-1)}, s = 2, \tag{83}$$

that the coefficients of type $a^{(2)}$ disappear, and he finds the equation

$$(r+1)(a_{r+2}^{(1)} + a_r^{(1)}) + 8(a_{r-1}^{(0)} - a_{r+3}^{(0)}) = 0. \tag{84}$$

Now this disappearance must be equivalent to an integration, and inspection shows that (84) will also be obtained by integrating (81) in the form

$$xy' + \int_{-1}^x 16xy = \text{constant}, \tag{85}$$

and using known Chebyshev expressions for products and integrals. The constant in (85) should be zero in virtue of the initial conditions, giving Clenshaw's equation (84).

A very important point, however, which becomes obvious only when we use the τ -method, is that we do not solve (85) exactly with our finite approximation, but include perturbing terms on the right. The choice of the "constant" should depend on this fact, and the correct choice will increase significantly the accuracy of our approximation.

33. To see this we analyse closely the result given by Clenshaw, computed here to extra figures. Only one

recurrence is needed, and he starts at $r = 10$ with $a_{10}^{(0)} = 1$, $a_{10}^{(1)} = 0$, and higher coefficients assumed negligible. Alternate use of (84) and (83) (with $s = 1$) produces all the coefficients, and a single multiplying factor is used to satisfy the condition $y(0) = 1$. We find the solution

$$y = 0.0501358T_0 - 0.6652551T_2 + 0.2489761T_4 - 0.0332397T_6 + 0.0023019T_8 - 0.0000921T_{10}, \quad (86)$$

and this satisfies exactly the equation

$$xy' + \int_0^x 16xy = \tau T_{12}(x), \quad (87)$$

with $\tau = 8/10860.5 \doteq 0.00074$.

We want, however, to have a zero on the right of (87), at least at $x = 0$, so that we do better to satisfy exactly the equation

$$xy' + \int_0^x 16xy = \tau(T_{12} - T_{10}). \quad (88)$$

This is quite easy to do, involving no change in technique, and we find the result

$$y = 0.0501271T_0 - 0.6652266T_2 + 0.2489857T_4 - 0.0332534T_6 + 0.0023115T_8 - 0.0000956T_{10}, \quad (89)$$

with $\tau = -4/125461 \doteq -0.000032$. The errors of Clenshaw's result (86) and the improved solution (89), denoted respectively by I and II, are as follows:

x	0.0	0.1	0.2	0.3	0.4	0.5
$10^6(I)$	0	-10	-33	-52	-54	-40
$10^6(II)$	0	-3	-9	-9	-3	+4
	0.6	0.7	0.8	0.9	1.0	
	-23	-11	+1	+16	+23	
	-5	-1	-1	+5	+1.	

34. We can, of course, integrate again, but the Clenshaw process is now less attractive than that of Lanczos. With the polynomial assumption

$$y = a_0 + a_2x^2 + a_4x^4 + a_6x^6 + a_8x^8 + a_{10}x^{10}, \quad (91)$$

the result (85) of a single integration has the factor x^2 on the left-hand side so that, if we want to solve the singly integrated equation, we might put the factor $\tau x^2 T_{10}(x)$ on the right-hand side, or perhaps $\tau(T_{12} + T_{10})$.

If we defer the perturbation to the second integrated equation we first write (85) in the form

$$y' + \int_0^x 16xy dx = y' + g(x) = 0, \quad (92)$$

then integrate again to find

$$y + \int_0^x g(x) dx = 1, \quad (93)$$

the boundary condition $y(0) = 1$ being then satisfied.

We might satisfy (93) exactly at the origin, with the polynomial approximation, by adding $\tau x^2 T_{10}$ to its

right-hand side, or preferably we add τT_{12} and tolerate an error at the origin. The resulting equations are then

$$\left. \begin{aligned} a_0 - 1 + \tau & \\ a_2 + a_0 &= -72\tau \\ a_4 + a_2 &= +840\tau \\ a_6 + \frac{4}{3}a_4 &= +3584\tau \\ a_8 + \frac{1}{4}a_6 &= +6912\tau \\ a_{10} + \frac{1}{25}a_8 &= -6144\tau \\ \frac{1}{9}a_{10} &= +2048\tau \end{aligned} \right\} \quad (94)$$

and again it is interesting to note that these equations, as well as those coming from the first integration, are constant multiples of those obtained from the original differential equation. The solution is

$$\tau = -1/363397 \doteq -0.0000028, \text{ is} \\ y = 0.9999972 - 3.9997909x^2 + 3.9974793x^4 \\ - 1.7667950x^6 + 0.4226782x^8 - 0.0507214x^{10}, \quad (95)$$

and its error distribution is as follows:

x	0.0	0.1	0.2	0.3	0.4	0.5
10^6z	+3	-1	-2	-2	-1	+3
	0.6	0.7	0.8	0.9	1.0	
	+1	-3	0	-2	+3.	

35. The corresponding solution of the original differential equation (81), written in the perturbed form

$$y'' + \frac{1}{x}y' + 16y = \tau T_{10}(x), \quad (97)$$

is

$$y = 1 - 3.9996304x^2 + 3.9950109x^4 \\ - 1.7591352x^6 + 0.4139142x^8 - 0.0473045x^{10}, \quad (98)$$

with $\tau = -32/21647 \doteq -0.0015$, and its error distribution is as follows:

x	0.0	0.1	0.2	0.3	0.4	0.5
10^6z	0	-3	-8	-5	+10	+26
	0.6	0.7	0.8	0.9	1.0	
	-29	+17	+2	-1	-5.	

Inspection of the errors of (99), (90) and (96) shows a significant increase in accuracy with each integration. Moreover, because the Chebyshev perturbation terms are of successively increasing order, the error changes sign more often, and quantities like $\int g(x) dx$, in the equation for the error corresponding to (93), are likely to be very small. The error of y in the approximation (95), for example, is in fact almost identical with the perturbing term τT_{12} , and this approximation is therefore very close to the best possible polynomial solution of degree ten.

Alternative Elimination of $a_r^{(s)}$

36. The polynomial method solves only for the coefficients in the finite approximation to the required function whereas, without integration, the Chebyshev

method finds the coefficients also in the relevant derivatives. The latter is unnecessary if we extend the relations between $a_r^{(0)}$ and $a_r^{(1)}$, and find a set of algebraic equations for the $a_r^{(0)}$, which can be solved by Gauss elimination. An increase in the order of the approximation merely adds extra rows and columns to the matrix of coefficients, and the extra calculation is therefore reduced to a minimum. "Overdetermination" is manifest in a more obvious way, and we ensure that the redundant equations are those involving the τ coefficients, so that the latter can be omitted, if desired, or calculated directly.

37. The relevant relations, for the T polynomials, come from the integrating relations extended to the point where, in the finite approximation, successive terms vanish. We have

$$a_r^{(1)} = a_{r+2}^{(1)} + 2(r+1)a_{r+1}^{(0)} \\ = a_{r+4}^{(1)} + 2(r+3)a_{r+3}^{(0)} + 2(r+1)a_{r+1}^{(0)}, \dots \quad (100)$$

so that ultimately

$$a_r^{(1)} = 2\{(r+1)a_{r+1}^{(0)} + (r-3)a_{r+3}^{(0)} + \dots\}, \quad (101)$$

as far as a vanishing term. For the coefficients of the second derivative we find similarly the formula

$$a_r^{(2)} = 4(r+1)(r+2)a_{r+2}^{(0)} + 8(r+2)(r+4)a_{r+4}^{(0)} \\ + 12(r+3)(r+6)a_{r+6}^{(0)} + \dots \quad (102)$$

and the series terminates.

In the problem of Section 23, for example, we can express the equations (61) in terms of the $a_r^{(0)}$ coefficients only, incorporate the boundary condition (63), and produce the following simultaneous equations, of which the first represents the boundary condition:

$$\begin{array}{cccccccc} a_0^{(0)} & a_2^{(0)} & a_4^{(0)} & a_6^{(0)} & a_8^{(0)} & a_{10}^{(0)} & \dots & \dots \\ \frac{1}{2} & -1 & 1 & -1 & 1 & -1 & \dots & \dots \\ -2 & 3 & 32 & 108 & 256 & 500 & \dots & \dots \\ -1 & -4 & 47 & 192 & 480 & 960 & \dots & \dots \\ & & -1 & -4 & 119 & 384 & 840 & \dots \\ & & & -1 & -4 & 233 & 640 & \dots \\ & & & & -1 & -4 & 359 & \dots \\ & & & & & -1 & -4 & \dots \\ & & & & & & -1 & \dots \end{array} \quad (103)$$

Here, in any finite set drawn from (103), we have two more equations than unknowns, the last two equations referring to the coefficients of $T_{2n}(x)$ and $T_{2n+2}(x)$. We can clearly omit these equations from the set, and use them to determine the "error" of our approximation. For example, from the first six of the equations given explicitly in (103) we can calculate all the $a_r^{(0)}$, and the perturbing terms are just $-(a_8^{(0)} + 4a_{10}^{(0)})T_{10} - a_{10}^{(0)}T_{12}$.

The solution, of course, is identical with that of our standard τ -method, and is the Chebyshev rearrangement of the polynomial (28), with $-(a_8^{(0)} + 4a_{10}^{(0)})$ and $-a_{10}^{(0)}$ being identical with τ_0 and τ_2 , respectively, of that solution.

For solutions of larger or smaller order we merely add or subtract rows and columns from the matrix in

(103), omitting the last two equations in all cases, and the omitted equations will determine the error. The Gauss elimination process clearly facilitates the computation, and this method is probably the easiest, of those described in this paper, for automatic programming and computing.

Additional Comment on Clenshaw's Paper

38. In discussing the number of trial solutions, that is the number of τ terms needed for the solution of the algebraic equations, Clenshaw relates this to the number of convergent Chebyshev solutions of the differential equation. For example he states that "if the solution sought is the only solution of the differential equation with a convergent Chebyshev expansion, then a single trial solution obtained by recurrence will yield this required result when multiplied by a constant factor, ... normally given by the satisfaction of a boundary condition."

Presumably he has in mind the homogeneous case, but even here we cannot equate the number of trial solutions to the number of complementary solutions of the differential equation. The number expected by this consideration is likely to be increased according to the occurrence of various powers of x . For example, the system

$$x^2y' - y = 0, \quad y(1) = 1, \quad (104)$$

solved in Sections 8 and 18, has only the solution $y = e^{\frac{1}{2}x^2}$, and yet two τ -terms are needed for the recurrence solution of the algebraic equations.

39. The system

$$y' + x^2y = 0, \quad y(0) = 1, \quad -1 \leq x \leq 1 \quad (105)$$

has only the single solution $y = e^{-\frac{1}{3}x^3}$, and this needs as many as three τ -terms. This fact becomes quite obvious when we apply the method of Section 37, which gives for the coefficients the equations

$$\begin{array}{cccccccc} a_0^{(0)} & a_1^{(0)} & a_2^{(0)} & a_3^{(0)} & a_4^{(0)} & a_5^{(0)} & a_6^{(0)} & \dots \\ \frac{1}{2} & 0 & -1 & 0 & 1 & \dots & = & 1 \\ \frac{1}{4} & 1 & \frac{1}{4} & 3 & 0 & \dots & = & 0 \\ 0 & \frac{3}{4} & 4 & \frac{1}{4} & 8 & \dots & = & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & 6 & \frac{1}{4} & \dots & = & 0 \\ & \frac{1}{4} & 0 & \frac{1}{2} & 8 & \dots & = & 0 \\ & & \frac{1}{4} & 0 & \frac{1}{2} & \dots & = & 0 \\ & & & \frac{1}{4} & 0 & \dots & = & 0 \\ & & & & \frac{1}{4} & \dots & = & 0 \end{array} \quad (106)$$

Here truncation to a finite set leaves the last three equations unsatisfied, so that three τ -terms are involved. Successive solutions of low order, with the perturbing terms included, are as follows:

$$\begin{array}{l} y = T_0; \quad \frac{1}{4}T_0 + \frac{1}{4}T_2 \\ y = T_0 - \frac{1}{4}T_1; \quad -\frac{1}{16}T_1 + \frac{1}{4}T_2 - \frac{3}{16}T_3 \\ y = -\frac{1}{8}(6T_0 - 32T_1 + 6T_2); \\ \quad \quad \quad \frac{3}{32}T_0T_2 - \frac{3}{8}T_3 - \frac{1}{16}T_4 \end{array} \quad (107)$$

40. The integrated equations yield similar information. For example, we can solve (105) in the form

$$y + \int_0^x x^2 y dx = 1, \tag{108}$$

and the coefficient of $T_1(x)$ on the left of (108) is

$$a_r = \frac{1}{8r}(a_{r-1} + a_{r-3} - a_{r+1} - a_{r+3}), \tag{109}$$

reducing to

$$\frac{1}{2}a_0 + \frac{3}{2}a_1 - \frac{5}{6}a_3 = \frac{5}{6}a_3 \quad \text{for } r = 0. \tag{110}$$

The equations are then

$$\begin{matrix} a_0 & a_1 & a_2 & a_3 & a_4 & \dots & \dots \\ \frac{1}{2} & \frac{3}{2} & 0 & -\frac{5}{6} & 0 & \dots & = 1 \\ \frac{1}{8} & 1 & 0 & 0 & -\frac{1}{8} & \dots & = 0 \\ 0 & \frac{1}{8} & 1 & -\frac{1}{6} & 0 & \dots & = 0 \\ \frac{1}{24} & 0 & \frac{1}{24} & 1 & -\frac{1}{24} & \dots & = 0 \\ \frac{1}{2} & 0 & \frac{3}{2} & 0 & \frac{3}{2} & \dots & = 0 \\ & & \frac{4}{6} & 0 & \frac{4}{6} & \dots & = 0 \\ & & & \frac{4}{8} & 0 & \dots & = 0 \\ & & & & \frac{5}{6} & \dots & = 0 \end{matrix}, \tag{111}$$

and again the last three of any finite set are unnecessary except for determining the error. Stopping at a_2 , for example, we find the solution

$$y = 1.024T_0 - 0.256T_1 + 0.032T_2, \tag{112}$$

which satisfies exactly the equation

$$y + \int_0^x x^2 y dx = 1 + \frac{1}{1250}T_5 - \frac{1}{125}T_4 + \frac{13}{150}T_3, \tag{113}$$

and again it is worth remarking that the error of (112) differs, throughout the range, only very slightly from the perturbing terms in (113), even for such a low-order approximation.

The number of τ -terms needed is clearly connected with the manner in which the matrix of coefficients departs from a strict upper triangular form, and depends on the power of x in the various coefficients of the differential equation.

Boundary-Value Problems

41. So far all our examples, of second-order equations, have involved one-point boundary conditions. The treatment of boundary-value problems is very similar, though the integrated forms have a slight extra complication. To illustrate the technique we use the simple system

$y'' + y = x$ (114)

with various boundary conditions, and seek a polynomial approximation of degree four.

42. First we take the conditions

$$y'(0) = -1, \quad y(1) = 2, \tag{115}$$

substitute our approximation in (114), and determine equations for the coefficients a_r , given by

$$a_r + (r + 2)(r + 1)a_{r+2} = 0 \quad \text{for } r \neq 1, \quad = 1 \text{ for } r = 1, \tag{116}$$

and with $a_5 = a_6 = 0$.

We find that the equations can be satisfied exactly, and non-trivially, if we add the terms $\tau_3 T_3 + \tau_4 T_4$ on the right of the differential equation, and the parameters τ_3 and τ_4 are determined from the satisfaction of the boundary-condition equations

$$a_0 + a_1 + a_2 + a_3 + a_4 = \frac{1}{2}, \tag{117}$$

Calculation gives the five-decimal result

$$y = 4.96231 - x - 2.49104x^2 + 0.39012x^3 + 0.13861x^4, \tag{118}$$

which satisfies the system

$$y'' + y = x + 0.00108 \dots T_4 + 0.02085 \dots T_3, \tag{119}$$

and has the following errors:

x	0.0	0.2	0.4	0.6	0.8	1.0
$10^5 \epsilon$	332	330	255	126	32	0.

The error in the derivative is, of course, zero at $x = 0$. 43. With the same differential equation, but with boundary conditions

$$y(0) = 4.96563, \quad y(1) = 2, \tag{121}$$

which give the same analytical solution, to five decimals, as those of (115), the technique and equations are the same, except that the first of (117) is replaced by $a_0 = 4.96563$. We then find

$$y = 4.96563 - 1.00213x - 2.49271x^2 + 0.39053x^3 - 0.13869x^4, \tag{122}$$

which satisfies exactly equation (119), with almost the same values of τ_4 and τ_3 , and has the following errors:

x	0.0	0.2	0.4	0.6	0.8	1.0
$10^5 \epsilon$	0	+47	+32	-28	-47	0.

The error in the derivative at $x = 0$ is now 0.00213.

The Integrated Equations

44. As in the initial-value case we may find better solutions by integrating the original equation and by not insisting that either boundary condition is satisfied exactly.

A first integration of (114) gives

$$y' + \int_0^x y dx = \frac{1}{2}x^2 + a, \tag{124}$$

where a is a constant equal to the value of $y'(0)$, and if the second term in (124) is denoted by $g(x)$ a second integration gives

$$y + \int_0^x g(x) dx = \frac{1}{6}x^3 + ax + b, \tag{125}$$

where the constant b is in fact the value of $y(0)$.

Now in the initial-value case both a and b are known immediately. In the boundary-value case we can determine them so that, at this stage, the boundary values are satisfied. For example, with conditions (115) we have $a = -1$ in (125), and if we put $x = 1$ in (125) we find

$$b = a + \frac{1}{b} = 2 + \left(\frac{1}{3}a_0 + \frac{1}{6}a_1 + \frac{1}{15}a_2 + \frac{1}{20}a_3 + \frac{1}{30}a_4 + \dots\right). \tag{126}$$

With conditions (121) we know immediately $b = 4.96563$, and can now determine a from (126).

But this is clearly unnecessary. We are going to perturb (125) with τ -terms, and the boundary conditions will be affected by these amounts. For a polynomial approximation of degree four we shall need to use τ_0, τ_1 and $\tau_2 T_3^*$, and the equations for the coefficients relating to the constant and the term in x can be taken to be the "new" boundary conditions. With conditions (115), for example, we use

$$\left. \begin{aligned} a_1 = \dots = 1 + \tau_6 C_6^{(1)} - \tau_5 C_5^{(1)} \\ a_0 + a_1 + a_2 + a_3 + a_4 = 2 + \tau_6 + \tau_5 \end{aligned} \right\}, \tag{127}$$

while with conditions (121) the first of (127) is replaced by

$$a_0 = 4.96563 + \tau_6 - \tau_5. \tag{128}$$

The other equations come from powers of x from two to six, and in each case we find the six-decimal result

$$\begin{aligned} \tau_1 = 4.965566 - 0.996848x = 2.507378x^2 \\ + 0.398870x^3 + 0.139859x^4, \end{aligned} \tag{129}$$

with $\tau_6 = 2.3 \times 10^{-6}$, $\tau_5 = 6.6 \times 10^{-5}$, and with the following errors:

$$\begin{array}{cccccccc} x & 0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 & \\ 10^5 \epsilon & +6 & -1 & -5 & -5 & -1 & -7 & \end{array} \tag{130}$$

The error in the derivative at $x = 0$ is 315×10^{-5} , but the result (129) for the function is again virtually the best possible, the errors being almost identical with $\tau_5 T_5^*$.

Note on Clenshaw's Error Analysis

45. We remarked in the introduction that Clenshaw's original process seeks essentially to find the infinite Chebyshev series, and in fact he uses the recurrence process in such a way that he can estimate the number of figures to which his computed coefficients agree with those of the infinite series. The idea is originally due to Miller (British Association Mathematical Tables, 1952), who used it to compute the Bessel function $J_n(x)$ for fixed x and integer n from the recurrence relation

$$I_{n-1}(x) = \frac{2n}{x} I_n(x) - I_{n+1}(x), \tag{131}$$

and it was applied by Fox (1954) in similar calculations.

Starting with $I_{m+1} = 0$ and $I_m = 1$, for some large value m of n , we recur backwards down to $n = 0$. The result is a solution of (131), and we find an appropriate

multiplying factor by satisfying an extra condition, such as the known value of I_0 or the series

$$e^x = I_0 + 2I_1 + 2I_2 + \dots \tag{132}$$

Miller's analysis indicates, Fox states, and a private communication from Dr. F. W. J. Olver shows, that if we round off each value of the recurrence to the nearest integer, thereby *deliberately* keeping systematically erroneous end figures, then the number of correct significant figures, after applying the multiplying factor, is equal to the number of digits in the original rounded coefficients, or the number of correct digits in the check sum, whichever is less.

If this method is applied to the example of Section 15 we can decide that the maximum discrepancy between the computed coefficients and those of the infinite series is about 0.00020, giving a possible maximum *rounding* error of about 0.0010, and a possible maximum error in the fourth-order approximation of little more than this. Olver shows that by a slightly deeper analysis he can reduce this upper bound to agree quite closely with the actual error of 0.00023. This analysis has not been extended to the more complicated case when more than one trial solution is necessary, and Clenshaw's paper gives no guide to estimation when the equations are solved by other methods, such as iteration or Gauss elimination.

In any case the τ -method has certain desirable features, in the possibility of knowing something about the distribution of the error, the fact that it may be of different order at special points in the range, and in automatic computation of the kind illustrated in Sections 36 and 37.

Summary and Conclusion

46. We have demonstrated the close relation between the method of Lanczos, which produces for an approximate solution the coefficients of a polynomial, and that of Clenshaw, which produces the coefficients of the corresponding Chebyshev series, and have noted the advantages, in improving Clenshaw's process, of adapting the ideas of Lanczos involved in a *finite* approximation.

Whether we use the Lanczos or Clenshaw process for the original equation will depend largely on the nature of the problem. The equations for the coefficients of the polynomial are usually less complicated, and often have fewer terms, than those for the Chebyshev series. On the other hand the magnitude of the polynomial coefficients is usually greater than those of the Chebyshev series, for the same precision, and this gives an advantage to the Clenshaw method when high precision is required and when many terms are needed, the latter situation arising particularly when the function does not behave, at some points of the range, very much like a polynomial.

With the integrated forms, however, the equations for the Chebyshev coefficients will often be considerably more complicated, while those for the polynomial are effectively the same. Moreover, with the polynomial

approximation we can more easily divide throughout by any power of x common to numerator and denominator.

A tentative conclusion is that the production of the Chebyshev series should be reserved for those problems needing many terms and great accuracy, particularly when the error estimation is difficult or impossible. For in the latter case coefficients in the solutions of different orders have only small differences in the Chebyshev series compared with those in the polynomial, and the variation with order is more easily observed.

47. For most of the examples of this paper the process of integration has produced a more accurate result. The conditions under which this will apply need investigation, but we may remark tentatively that this is likely whenever, as a result of integration, the final expression for y in the new equation is multiplied by a function of x of reasonable size in the range. In particular the method may not succeed if this function vanishes, most probably at one end of the range.

For example, in the problem of Section 8, defined by

$$x^2y' - y = 0, \quad y(1) = 1, \quad (133)$$

we know immediately from the perturbed equation, with

two τ -terms, that the maximum error cannot exceed $|\tau_1| + |\tau_2|$. The integrated form is

$$x^2y = \int_0^x (1 + 2x)^y dx = 0, \quad (134)$$

and in order to be able to assess the error we must be able to divide (134) by x^2 . This means that, unlike the process of Section 8, we must satisfy an extra condition, $y = 0$, at the origin. Such a constraint may result in a poorer solution and, perhaps more important, we see that to satisfy

$$y = \frac{1}{x^2} \int_0^x (1 + 2x)^y dx = 0 \quad (135)$$

the perturbation on the right must involve terms T_n^* and T_{n-1}^* , whereas that of the original equation has the "better" perturbation involving T_n^* and T_{n+1}^* .

48. Each problem clearly merits some preliminary mathematical investigation, but the examples of this paper show that, at least in problems of reasonable size and complication, both this investigation and the corresponding error analysis are not prohibitive.

I am grateful for valuable discussion with Mr. C. W. Clenshaw, Dr. F. W. J. Olver and Dr. J. C. P. Miller on much of the material of this paper.

References

- BRITISH ASSOCIATION (1952). *Mathematical Tables X*. Cambridge University Press.
 CLENSHAW, C. W. (1957). "The Numerical Solution of Linear Differential Equations in Chebyshev Series," *Proc. Camb. Phil. Soc.*, Vol. 53, pp. 134-49.
 FOX, L. (1954). *A Short Table for Bessel Functions of Integer Order and Large Argument*, Royal Society Shorter Mathematical Tables 3, Cambridge University Press.
 LANCZOS, C. (1938). "Trigonometric Interpolation of Empirical and Analytical Functions," *Journal of Maths. and Physics*, Vol. 17, pp. 123-99.
 LANCZOS, C. (1957). *Applied Analysis*. London: Pitman.

Book Review

Electronic Digital Computers, by G. D. SMIRNOV (Translated by G. Segal). 1960, original Russian Edition 1958. 104 pages. (London: Pergamon Press Ltd., 42s. 0d.)

This book attempts to cover the entire field of digital computers—programming, logical design and construction—in 97 pages of text. The treatment is necessarily brief, but is in many cases ill-balanced: for example, fifteen pages are devoted to a discussion of systems of numeration and binary arithmetic, whilst arithmetic units are dealt with in four pages and input/output is dismissed in two. Sometimes a completely erroneous impression is conveyed, as in the section on control units, which is limited to a treatment of diode decoders.

It is not clear what type of reader the book is aimed at, for although the first chapter appears to be intended for the layman, later chapters discuss the design of logical elements using both valve and transistor techniques, which implies a knowledge of electronics. There is some interesting infor-

mation about current Russian techniques, though the detail is insufficient to satisfy the knowledgeable reader. One would like to know, for example, whether "peeping" is the only technique used for detection of program errors, and to have more detail about the magnetic-tape units, which are said to have a maximum tape speed of 2 metres/sec, and a packing density of about 300 words/metre.

The book is a photo-litho reproduction of typescript. The publishers explain that this form of reproduction is used "... in the interests of speedily making available the information contained in the publication". Although this is an admirable attitude to adopt for accounts of current research, the reviewer is doubtful of the need for speedy translation of general texts such as this. The translation is generally of a high standard, though there are a few puzzling things, as for example, "... magnetic tapes are made of combustible or incombustible film ..."

D. W. BARRON.