# Chebyshev collocation methods for ordinary differential equations

## By K. Wright

This paper describes some methods for the solution of ordinary differential equations based on the Lanczos "selected points" or collocation principle. Some properties of the methods are examined theoretically, and some examples of their use given.

In this paper I consider some methods for the solution of ordinary differential equations in terms of finite Chebyshev series. This form is often convenient, particularly if the solution is required for some other calculation where its rapid and accurate evaluation is important.

With the solution expressed in this form or as an explicit polynomial, there are two approaches to the problem of choosing coefficients to make the function satisfy the differential equation approximately. The first is to equate coefficients either of powers of $x$ or of the polynomials. The second is to satisfy the differential equation at certain values of the independent variable. Both these approaches have been considered by Lanczos (1957), but most work since then has followed the first method (Clenshaw, 1957; Fox, 1962). Recently, however, Clenshaw and Norton (1963) have given a method using the second alternative. Here I give some variations of the second approach, and discuss some of their properties.

Although the methods are described in terms of Chebyshev series, they are not essentially dependent upon this; similar methods could be constructed to give explicit polynomial solutions.

The principle of these methods is that the solution is represented by a finite Chebyshev series with unknown coefficients; this expression is substituted into the differential equation and the coefficients are determined so that the differential equation is satisfied at certain points within the range under consideration. The number of points is chosen so that, along with the initial or boundary conditions, there are enough equations to find the unknown coefficients. The positions of the points in the range are chosen to make small the residual obtained when the approximate solution is substituted into the differential equation.

The range in which the solution is required is assumed finite, and, for convenience, a linear transformation of the independent variable is made to make the new range $(-1, 1)$.

Suppose that the independent variable of the differential equation is $x$, and it is solved exactly at $n$ points $x_i, i = 1, 2, \ldots, n$. The residual will then be of the form

$$E = \prod_{i=1}^{n} (x - x_i)\psi(x), \qquad (1)$$

where $\psi(x)$ is an unknown function which will depend

upon the particular differential equation. The term $\Pi(x - x_i)$ is, however, independent of the differential equation, and this term is minimized if it is made a multiple of a Chebyshev polynomial, or explicitly

$$x_i = \cos\left\{\frac{(2i + 1)\pi}{2n}\right\}. \qquad (2)$$

Lanczos (1957), p. 504, calls this choice of points the "selected points" principle or the method of collocation. It applies in a similar manner to sets of equations, the function $\psi(x)$ then being a vector.

## Linear differential equations

This principle is easily applied to linear differential equations. It is most conveniently illustrated with a single first-order differential equation, since this gives the essential points without unnecessary complications.

Consider the equation

$$y' = f(x)y + g(x), \qquad (3)$$

where the solution is required in the range $(-1, 1)$, and the initial condition is imposed at one of the end points of the range. Suppose that $y, y'$ are represented by the Chebyshev series

$$\left. \begin{aligned} y &= \sum_{0}^{n}{}' a_r T_r(x) \\ y' &= \sum_{0}^{n-1}{}' b_r T_r(x) \end{aligned} \right\}, \qquad (4)$$

where the $\Sigma'$ denotes that $a_0, b_0$ are taken with weight $\frac{1}{2}$ as is usual with Chebyshev series. The $a_r$ and $b_r$ are connected by the relations

$$b_{i-1} - b_{i+1} = 2ia_i, \qquad i = 1, \ldots n \qquad (5)$$

where $b_n, b_{n+1}$ are taken to be zero; this result is obtained by integrating the series for $y'$ (Clenshaw, 1957). If the expressions for $y$ and $y'$ are substituted in the differential equation and the resulting relation satisfied at the $n$ points $x_i$, the zeros of $T_n(x)$, these equations with the initial condition and the relations (5) provide enough equations to find the $2n + 1$ unknowns $a_r, b_r$. The initial conditions are of a very simple form, for

$$\left. \begin{aligned} y(-1) &= \tfrac{1}{2}a_0 - a_1 + a_2 \ldots \\ y(+1) &= \tfrac{1}{2}a_0 + a_1 + a_2 \ldots \end{aligned} \right\}, \qquad (6)$$

which follow from the definition.

The set of linear equations produced in this way could be solved directly by any standard method, but it is more economical to eliminate the $b_i$ coefficients first. The equations (5) can be used in succession to eliminate $b_0, b_1, \ldots$ in order, by altering the coefficients of the later $b_i$ and the $a_i$ in the set of equations. Thus the set is reduced to order $n + 1$. The equations are then solved by a standard method.

A similar method could be used with $y$ in the form of an explicit polynomial of degree $n$; here the same result would be obtained expressed in a different form, apart from differences in rounding error. The Chebyshev series form is convenient, however, since a good estimate of the accuracy can be obtained from the size of the last few terms of the series, and a better one is obtained by comparing the two sets of coefficients obtained when different values of $n$ are used. These estimates are justified by the rapid convergence properties of Chebyshev series (Lanczos, 1957, p. 453).

The method outlined involves more work than the methods of Clenshaw (1957) and Lanczos (1957) Ch. VIII for equations of simple form, but it does apply just as easily to equations with complicated functions $f(x)$ and $g(x)$. The methods based on equating coefficients, on the other hand, require these functions to be explicit polynomials or finite Chebyshev series, and they become rapidly more complicated as the order of the polynomials increases.

*Systems of equations* can be treated in just the same way; here each unknown function must be represented by a Chebyshev series, so that larger sets of linear algebraic equations have to be solved. The method is not dependent upon the position of the associated conditions, and boundary-value problems are no more difficult than initial-value problems. *Higher-order equations* could be reduced to sets of first-order equations, but it is usually better to treat these directly, as the number of algebraic equations can be reduced using relations analogous to equations (5). Conditions involving higher derivatives can be treated using

$$\left.\begin{array}{l} y'(-1) = \Sigma'(-1)^{r+1}r^2 a_r \\ y'(+1) = \Sigma' r^2 a_r \end{array}\right\} \tag{7}$$

and similar expressions for higher derivatives; mixed boundary conditions are also not much more difficult.

## Non-linear differential equations

The "selected points" principle can also be applied to non-linear differential equations; here, however, it produces a set of non-linear algebraic equations for the unknown Chebyshev coefficients. These equations could be solved by a general iterative method, such as Newton's method, but one would not expect this to be very efficient.

I intend to discuss two main methods, and briefly mention a third, all of which take account of the origin of the algebraic equations in differential equations. The first is derived from the Picard existence theorem, the second uses a linearized form of the differential equation, and requires the solution of a linear differential equation at each stage of the iteration.

## The Picard method

For the sake of clarity I again outline the method for the single first-order equation

$$y' = f(x, y) \tag{8}$$

in the range $(-1, 1)$, with a condition at $x = \pm 1$. For this equation the Picard iteration takes the form

$$y'^{(n+1)} = f(x, y^{(n)}),$$

$$y^{(n+1)} = y(x_0) + \int_{x_0}^{x} f(t, y^{(n)})\, dt, \tag{9}$$

where $y^{(n)}$ are the successive iterates. The conditions on the convergence of this sequence are given by Ince (1955), p. 62, and Tricomi (1960), p. 11. It should be noted that these conditions can be generalized only for sets of equations of initial-value type. More consideration is needed for boundary-value problems.

The application of Chebyshev series and the selected points method is straightforward, but it is simplified by an additional property of the special points. The Chebyshev polynomials satisfy the orthogonal summation relations

$$\sum_{i}^{m} T_r(x_i)T_s(x_i) = \begin{cases} 2n & \text{if } r = s = 0 \\ n\delta_{rs} & \text{otherwise and } r, s < m, \end{cases} \tag{10}$$

where $x_i$ are the zeros of $T_m(x)$. This enables a Chebyshev series for a function $f(x)$ to be found from the function values $f(x_i) = f_i$ by summation instead of by the solution of a set of linear equations. If the coefficients in the series are $b_r$ we have

$$b_r = \frac{2}{m}\sum_{i=1}^{m} f_i T_r(x_i). \tag{11}$$

In the iteration the functions $y^{(n)}$ are represented by Chebyshev series

$$y^{(n)} = \sum_{r=0}^{m'} a_r T_r(x), \tag{12}$$

and some simple starting function, usually a constant, is taken for $y^{(0)}$. The right-hand side of equation (8) is evaluated at the points $x_i$, using the recurrence method given by Clenshaw (1955) for the $y$ values. A Chebyshev series for $y'^{(n+1)}$ is then constructed using the formulae (11), and $y^{(n+1)}$ is found by integrating this series using equations (5) and the initial condition. If the iteration converges the differential equation is satisfied at the points $x_i$.

The method of Clenshaw and Norton (1963) appears to be identical to this except that they use collocation over the different set of points given by $x_i = \cos\dfrac{i\pi}{m}$.

Since the Chebyshev polynomials satisfy an orthogonality relation over these points also this is just as convenient computationally. The use of the zeros seems better justified on account of the selected points principle, but there is probably little difference in practice.

This method again generalizes immediately to sets of first-order equations, and also to higher-order equations where some economy can be made in the amount of work involved.

## Linearization method

For the first-order equation (8) the iteration for the linearization method takes the form

$$y'^{(n+1)} = f(x, y^{(n)}) + (y^{(n+1)} - y^{(n)}) \frac{\partial f(x, y^{(n)})}{\partial y}, \quad (13)$$

which is obtained from the Taylor expansion of $f(x, y)$ in terms of $y - y^{(n)}$, where $y$ is the true solution. Second-order terms are omitted and $y$ put equal to $y^{(n+1)}$. If $|y - y^{(n)}|$ is small one would expect this type of method to converge for either initial or boundary-value problems, though it is not obvious how small $|y - y^{(n)}|$ needs to be. Kalaba (1959) has shown that in certain cases the iteration will certainly converge.

The application of the selected points principle need hardly be described; each iteration involves just one solution of a linear equation by the method given above. However, the values of $y^{(n)}$ at the points $x_i$ must be found as in the Picard method, and the function $\partial f/\partial y$ must be calculable.

This method also generalizes to simultaneous equations and higher-order equations, but the method becomes rapidly more complex as the order of the system increases, since all the first-order partial derivatives of the functions are needed.

## Other possibilities

There are other iterative schemes for non-linear differential equations to which Chebyshev series can easily be applied. A whole set of such schemes which contains the Picard method as a particular case is suggested by Nicolovius (1961). These schemes use the Taylor's series with the remainder included in the form of an integral; the different methods are obtained by retaining different numbers of terms.

For equation (8) we have

$$y(x) = \sum_{r=0}^{p-1} \frac{(x - x_0)^r}{r!} \frac{d^r y(x_0)}{dx^r} + \frac{1}{(p-1)!}$$

$$\int_{x_0}^{x} \frac{d^p y(t)}{dt^p} (x - t)^{p-1} dt, \quad (14)$$

which for $p = 1$ gives the Picard iteration.

Since the higher derivatives $\frac{d^r y(x)}{dx^r}$ are rather complicated functions of the partial derivatives of $f(x, y)$, these methods become rapidly more complex as $p$ increases,

unless $f(x, y)$ is particularly simple. Consequently I have only considered the method with $p = 2$ in addition to the Picard method. This gives iteration of the form

$$y^{(r+1)}(x) = y(x_0) + (x - x_0) \left(\frac{dy}{dx}\right)_{x=x_0}$$

$$+ \int_{x_0}^{x} \left\{ \frac{\partial f}{\partial x} + f(t, y^{(r)}) \frac{\partial f}{\partial y} \right\} (x - t) dt. \quad (15)$$

Here just two series integrations are required at each step of the iteration, so that more work is involved than in the Picard method. The derivatives $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ are also needed and this again might be inconvenient. The applications of Chebyshev series and the selected points principle are the same as those for the Picard process.

## Practical results

Some indication has already been given of the amount of work involved in these methods, at least for a single iteration of the processes. From this it appears that the Picard method is the simplest and most easily applied to large sets of equations. It is, however, obviously necessary to consider how many iterations are needed for convergence, and indeed whether the different methods all converge in comparable cases.

I give below a comparison of the three methods outlined for three simple first-order non-linear differential equations with known solutions. The equations are:

(i) $y' = 1 + y^2$, $\quad y(0) = 1$, interval $(0, 1)$,
solution $y = \tan x$.

(ii) $y' = \sqrt{(1 - y^2)}$, $y(0) = 0$, interval $(0, 1)$,
solution $y = \sin x$.

(iii) $y' = e^{-y}$, $\quad\quad y(0) = 0$, interval $(0, 1)$,
solution $y = \ln(1 + x)$.

In all cases the initial function was taken as a constant, and identical tests for convergence of the methods were made. The number of terms in the series was initially taken as 6 but was automatically increased, so that the last term should not be greater than $2^{-20}$ times the first term, and the iteration was continued until corresponding coefficients in two successive iterates agreed to 20 binary digits. These tests were chosen just for simplicity and are not expected to be adequate in all circumstances. (If the solution function is nearly odd or even, the criterion for increasing the number of terms is clearly unsatisfactory.)

All the methods converged and the results are given in Table 1. They were obtained using the Oxford University Mercury computer programmed in Autocode; the linearization method used in addition Mercury Autocode matrix facilities which involve slow auxiliary transfers. Since these are incidental to the process the linearization is better than is indicated by the times given.

I also give three examples comparing the Picard and linearization methods for boundary-value problems.

360

**Table 1**

| METHOD | | EQN. (i) | EQN. (ii) | EQN. (iii) |
|---|---|---|---|---|
| Picard | No. of Iterations | 17 | 8 | 12 |
| | Time (sec.) | 35 | 10 | 15 |
| | Final no. of terms | 12 | 7 | 9 |
| Linearization | No. of Iterations | 11 | 5 | 6 |
| | Time (sec.) | 90 | 35 | 48 |
| | Final no. of terms | 11 | 7 | 8 |
| Nicolovius $(p = 2)$ | No. of Iterations | 33 | 6 | 11 |
| | Time (sec.) | 87 | 10 | 24 |
| | Final no. of terms | 12 | 7 | 9 |

**Table 2**

| METHOD | EQN. (iv) | EQN. (v) | EQN. (vi) |
|---|---|---|---|
| Picard | 9 | 7 | 26 |
| Linearization | 4 | 4 | 5 |

Here the number of terms in the series was fixed throughout the iteration. The equations are

(iv) $y'' + \dfrac{1}{y} + \dfrac{y'^2}{y} = 0$;   $y(0) = 1, y(1) = 2$;
    using 11 terms    [from Milne (1953), p. 104].

(v) $y'' + \sin(y') + 1 = 0$;   $y(0) = 0, y(1) = 1$;
    using 8 terms    [from Fox (1957), p. 86].

(vi) $y'' = 1 \cdot 5y^2$;   $y(0) = 4, y(1) = 1$;
    using 8 terms    [from Collatz (1960)].

In all cases the iteration was terminated when all coefficients in the series agreed to $0 \cdot 00001$. Both methods converged for all the equations and Table 2 gives the number of iterations taken.

For these equations both methods are adequate, though for equation (vi) Picard is rather slowly convergent.

**Use of the methods in a step-by-step form for initial-value problems**

So far I have described methods in a form which attempts to find the solution of a differential equation as a Chebyshev series valid in the whole of the required range. This is not the only possible way of using them; for initial-value problems they can also be used as step-by-step methods with very little added complication. The equation can be solved in a sub-range with one end-point at the initial point, then the solution can be used to provide starting values for the next range, and this can be repeated as often as necessary.

It is difficult to find a general rule for choosing the interval size which applies to all methods. For the linear equation matrix method, however, I give a rule which is convenient in practice and should approximately minimize the total time involved in the calculation. Naturally, many assumptions have to be made to establish this rule.

The time for solving the problem is usually dominated by the solution of the sets of simultaneous equations; if there are $n$ unknowns this time is proportional to $n^3$. So the time for the solution in a single interval using $n$ terms may be taken to be

$$T_1 = \lambda n^3. \tag{16}$$

I now consider whether there is any gain in time if this interval is halved and a smaller number of terms used in each of the half-intervals so that the accuracy is approximately the same. This last requirement may be represented by requiring that the last coefficients of the series be approximately equal--assuming that similar behaviour occurs in both intervals. Suppose that the series coefficients in the original interval are $a_o, \ldots a_n$, and in a half-interval are $b_o, \ldots b_m$, then we require

$$b_m = a_n \tag{17}$$

approximately.

Another relationship between the $a_i$ and $b_i$ can be obtained by considering the infinite Chebyshev expansions in the corresponding intervals, which should agree approximately with the finite series. The Chebyshev coefficients $c_s$ of a function $f(x)$ in the range $(a, b)$ are given by

$$c_s = \int_{-1}^{1} \frac{f(\alpha\xi + \beta)T_s(\xi)}{\sqrt{(1 - \xi^2)}} d\xi, \tag{18}$$

where $\alpha = (b - a)/2$, $\beta = (b + a)/2$.

Expanding $f(\alpha\xi + \beta)$ by Taylor's theorem, we find

$$f(\alpha\xi + \beta) = f(\beta) + \cdots \frac{(\alpha\xi)^{s-1}}{(s-1)!} \frac{d^{s-1}}{du^{s-1}} f(\beta)$$

$$+ \frac{(\alpha\xi)^s}{s!} \frac{d^s f(\eta)}{du^s},$$

where $a < \eta < b$. Now using the orthogonality properties of the Chebyshev polynomials we get

$$c_s = \frac{\alpha^s}{s!} \int_{-1}^{1} \frac{\frac{d^s f(\eta)}{dx^s} T_s(\xi)\xi^s}{\sqrt{(1 - \xi^2)}} d\xi.$$

The dependence of the integral on $\alpha$ will vary with the particular function. Taking just the $\alpha^s$ term into consideration, halving $\alpha$ divides $c_s$ by $2^s$ and so we may take

$$b_s \approx 2^{-s} a_s. \tag{19}$$

Suppose also that the terms of the first series decrease

361

in a roughly constant ratio $k$, that is $a_s \approx k^{-s} a_o$. Condition (17) now gives

$$2^{-m} k^{-m} a_o = k^{-n} a_o,$$

and

$$m = \frac{n \log k}{\log (2k)}. \tag{20}$$

Now the time $T_2$ for the calculation using the smaller interval is given by

$$T_2 = 2\lambda m^3 = 2T_1 \left\{ \frac{\log k}{\log (2k)} \right\}^3,$$

and this will be smaller than $T_1$ if

$$2s^{1/3} \log k < \log (2k)$$

or

$$k < 14\cdot39. \tag{21}$$

This implies that time will certainly be gained by halving unless the series is rapidly convergent; with a computer which is normally able to keep only eight or nine decimals the last terms would be negligible after 6 or 7 terms with a $k$ value satisfying condition (21). So about this number of terms should be used, and if these do not give adequate representation of the series the interval should be reduced rather than the number of terms increased.

### Prediction with Chebyshev methods

Finite-difference predictor-corrector methods obtain high accuracy with few evaluations of the functions in the differential equations by using information from previous steps. In the step-by-step forms of the methods I have described the only use that is made of a previous interval is to provide starting values for the next. But here also more information is available and can be used, at least for the iterative methods, to obtain a better starting approximation for the iteration.

A starting approximation for a new interval can be constructed so that it has derivatives at the initial point which agree up to some order with those of the solution for the previous interval. First the derivatives of the known series can be evaluated by successive differentiation and evaluation at the end point. Then a series can be constructed in the second interval by repeated series integration, so that it has these derivatives at the beginning of the range. The number of derivatives used in this way can vary from one to the length of the whole series.

I applied this method using all possible derivatives to both the Picard and linearization methods. I also used just the first two derivatives with the Picard method.

I give a comparison of these two modifications of the Picard process with the original method for three sets of simultaneous equations

(vii) $y_1' = y_2$ , $y_1(1) = 0$
$y_2' = -\exp(-2y_1)$, $y_2(1) = 1$.
Solution $y_1 = \log x$; $y_2 = 1/x$.

(viii) $y_1' = y_2$ , $y_1(1) = 0$
$y_2' = y_3$ , $y_2(1) = 1$
$y_3' = 2 \exp(-3y_1)$ , $y_3(1) = 1$.
Solution $y_1 = \log x$; $y_2 = 1/x$; $y_3 = -1/x^2$.

(ix) $y_1' = y_2$ , $y_1(0) = 1$
$y_2' = -y_1$, $y_2(0) = 0$.
Solution $y_1 = \cos x$; $y_2 = -\sin x$.

In all the solutions the constant interval $h = 1$ was used, and the number of terms was also kept constant. The number of terms in the series is specified by $N$, and the calculations were terminated when successive iterates differed by less than $E$. Table 3 gives the number of iterations taken in the first 5 steps.

I also give the two versions of the linearization method for equations (vii) and (viii). The results are given in Table 4.

These modifications to the Picard and linearization processes do give some improvement in all cases, but it

### Table 3

| EQN. | PARAMETERS | PICARD (ORDINARY) | 1ST AND 2ND DERIVATIVES | ALL DERIVATIVES |
|---|---|---|---|---|
| (vii) | $N=6$ $E=0\cdot0001$ | 7,6,5,5,4. | 7,4,3,3,3. | 7,5,3,3,2. |
| | $N=6$ $E=0\cdot00001$ | 8,7,6,6,5. | 8,5,4,4,3. | 8,6,4,3,3. |
| (viii) | $N=6$ $E=0\cdot00001$ | 10,7,6,6,5. | 10,6,4,4,4. | 10,7,4,4,3. |
| | $N=8$ $E=0\cdot00001$ | 10,7,6,6,5. | 10,5,4,4,3. | 10,7,4,3,2. |
| (ix) | $N=6$ $E=0\cdot00001$ | 9,8,8,9,8. | 9,6,6,7,6. | 9,5,5,5,5. |
| | $N=8$ $E=0\cdot00001$ | 8,8,8,8,8. | 8,6,6,6,6. | 8,4,3,3,3. |

### Table 4

| EQN. | PARAMETERS | LINEARIZATION | USING ALL DERIVATIVES |
|---|---|---|---|
| (vii) | $N = 6$ $E = 0\cdot0001$ | 4,3,3,3,3. | 4,2,2,2,2. |
| | $N = 6$ $E = 0\cdot00001$ | 4,3,3,3,3. | 4,3,2,2,2. |
| (viii) | $N = 6$ $E = 0\cdot00001$ | 4,3,3,3,3. | 4,3,2,2,2. |

362

is not always very much. The second set of results for equation (ix) with the Picard method, however, show a considerable improvement. It is interesting to note the difference between the cases $N = 6$ and $N = 8$ for this equation; the poorer results for $N = 6$ appear to be caused by the series being only just long enough to represent the function.

It is difficult to come to any general conclusions from these results, particularly as equations (vii) and (viii) have similar solutions. It does seem possible that the amendments will produce considerable improvement in some cases, but more experience is needed to be sure of this.

Unlike finite-difference predictor-corrector methods this method does not need any special starting procedure, since the ordinary process is used in the first interval.

## Conclusions

The methods described provide generally applicable ways of solving differential equations giving the solution in Chebyshev series form. They should certainly be reliable for initial-value problems if a sufficiently small interval is used, but for boundary problems convergence is not certain. These methods often involve more work than conventional methods, but it does provide the solution in a form which is sometimes very useful.

The Picard method is often simpler and more convenient than the linearization method as it does not require the solution of a set of simultaneous equations; however, for certain types of equation considered in the Appendix linearization is much better. More experience is required to determine the relative merits of the various alternative forms of the methods.

## Acknowledgements

## References

CLENSHAW, C. W. (1955). "A Note on the Summation of Chebyshev Series," *M.T.A.C.*, Vol. 9, p. 118.

CLENSHAW, C. W. (1957). "The Numerical Solution of Linear Differential Equations in Chebyshev Series," *Proc. Camb. Phil. Soc.*, Vol. 53, p. 134.

CLENSHAW, C. W., and NORTON, H. J. (1963). "The solution of non-linear ordinary differential equations in Chebyshev series," *Comp. J.*, Vol. 6, p. 88.

COLLATZ, L. (1960). *The Numerical Treatment of Differential Equations*, Springer Verlag.

FOX, L. (1957). *The Numerical Solution of Two Point Boundary Problems*, Oxford University Press.

FOX, L. (1962). "Chebyshev Methods for Ordinary Differential Equations," *Comp. J.*, Vol. 4, p. 318.

FOX, L., and GOODWIN, E. T. (1949). "Some New Methods for Numerical Integration of Ordinary Differential Equations," *Proc. Camb. Phil. Soc.*, Vol. 45, p. 373.

HARTREE, D. R. (1958). *Numerical Analysis*, Oxford University Press.

INCE, E. L. (1955). *Ordinary Differential Equations*, Dover.

KALABA, R. (1959). "On Non-linear Differential Equations, the Maximum Operator and Monotone Convergence," *J. Math. Mech.*, Vol. 8, p. 519.

LANCZOS, C. (1957). *Applied Analysis*, Pitman.

MILNE, W. E. (1953). *The Numerical Solution of Differential Equations*, Wiley.

NICOLOVIUS, R. (1961). "Iterative Solution of Boundary Value Problems using Remainder Terms of the Taylor Expansion," P.I.C.C. Symposium on Differential and Integral Equations. Rome.

TRICOMI, F. G. (1960). *Differential Equations*, Blackie.

WRIGHT, K. (1962). "The Numerical Solution of Ordinary Non-linear Differential Equations," Thesis—Oxford.

# Appendix

### Stability properties and related phenomena

To investigate the stability properties of the methods when used as step-by-step processes I used the following two examples:

(x) $y' = -10y + 10x^2 - 8x - 1$, $\quad y(0) = 0$

(xi) $y' = -10y_1 + 6y_2$, $\quad y_1(1) = 0 \cdot 6666666$

$y'_2 = 13 \cdot 5y_1 - 10y_2$, $\quad y_2(1) = 1$ [from Fox and Goodwin (1949)].

Many ordinary methods suffer from induced instability when used on these examples, on account of the large negative exponential complementary functions which are not well represented. Instability takes place if such a complementary function is represented by the process as an increasing function instead of a decreasing function, and this happens with both Runge-Kutta and predictor-corrector type methods unless a very small interval is used, since they are both based on the truncated Taylor's series which is slowly convergent.

F

363

The method for linear equations and the Picard–Chebyshev method both gave interesting results when used to solve these two equations.

Firstly the linear equation method gave good results as long as the solution was well represented by the Chebyshev series, even though the complementary function would need many more terms. An even more important point is that there was no sign of instability when the method was used over a number of steps. For example (xi) intervals in $x$ of 1 and 2 were stable with only 6 terms in the series, and the accuracy remained good; for interval 2 the solution for $y_2$ at $x = 11$ was $0 \cdot 000045$, correct in all places.

The stability of this method for equations with large negative exponential complementary functions can be demonstrated theoretically by showing that the representations of these functions always decrease.

Consider a set of $s$ simultaneous linear equations with constant coefficients which may be written in matrix form

$$y' = Ay, \qquad (22)$$

where $A$ is an $(s \times s)$ non-singular matrix. $A$ may be taken non-singular for otherwise the number of independent equations could be reduced. The selected points method is carried out as described for equation (3) except that the series coefficients are now vectors $a_r$ and $b_r$.

For this type of equation the selected points method and the method equating coefficients given by Clenshaw (1957) are equivalent, for they both have error terms which are multiples of $T_n(x)$, and the vanishing of the error at the zeros of this function is sufficient to determine the coefficients uniquely.

Equating coefficients gives

$$b_r = Aa_r, \qquad r = 0, \ldots n - 1, \qquad (23)$$

then equations (5) give

$$\left.\begin{array}{l} 2ra_r = A(a_{r-1} - a_{r+1}), \quad r = 1, \ldots n - 2 \\[4pt] 2(n - 1)a_{n-1} = Aa_{n-2} \\[4pt] 2na_n = Aa_{n-1}. \end{array}\right\}$$

Now putting $B = 2A^{-1}$, which is permissible since $A$ is non-singular, we get

$$\left.\begin{array}{l} a_{n-1} = nBa_n \\[4pt] a_{n-2} = (n - 1)Ba_{n-1} = n(n - 1)B^2a_n \\[4pt] a_{n-3} = (n - 2)Ba_{n-2} + a_{n-1} \\[4pt] \qquad = \{n(n - 1)(n - 2)B^3 + nB\}\,a_n \end{array}\right\}$$

and, from the general equation, it follows by induction that $a_{n-i}$ is a polynomial in $B$ with positive coefficients multiplying $a_n$; if $i$ is even it is an even polynomial and if $i$ is odd an odd polynomial.

The initial conditions can be written in the form

$$\tfrac{1}{2}a_0 - a_1 + a_2 \ldots = \alpha$$

or

$$u - v = \alpha$$

where

$$\left.\begin{array}{l} u = \tfrac{1}{2}a_0 + a_2 + a_4 + \ldots \\[4pt] v = a_1 + a_3 + a_5 + \ldots \end{array}\right\} \qquad (24)$$

So if $n$ is even $u$ is an even polynomial in $B$ with positive coefficients multiplying $a_n$, and $v$ is odd. If $n$ is odd this is reversed. This gives $a_n$ in terms of $\alpha$.

Now $\alpha$ may be expressed as a linear combination of the latent vectors $\alpha_i$ of $A$ corresponding to latent roots $\lambda_i$, that is

$$\alpha = \sum_1^s \mu_i\alpha_i.$$

The $\alpha_i$ are also latent vectors of $B$ corresponding to the latent roots $2\lambda_i^{-1}$.

If the solutions of the differential equations for $\alpha = \alpha_i$ for negative $\lambda_i$ are decreasing, then the method will be stable in the sense I have indicated.

Consider therefore a single latent vector $\alpha$ corresponding to a negative $\lambda$. The problem now reduces to a scalar one, for all the vectors will be multiples of $\alpha$, and we may replace $a_n$ by $a_n\alpha$, etc., and $B$ by $2\lambda^{-1}$.

If $n$ is even equation (24) now gives

$$u = Ra_n, \quad R > 0,$$
$$v = Sa_n, \quad S < 0.$$

Since $u - v = 1$ then $(R - S)a_n > 0$ and $a_n > 0$. The value at the other end of the range is $(u + v)\alpha$, and

$$|u + v| \leqslant |u| + |v| = u - v$$

since $u > 0$, $v < 0$, so that the function values decrease.

Similarly if $n$ is odd we have

$$u = Ra_n, \quad R < 0,$$
$$v = Sa_n, \quad S > 0.$$

Again $u - v = 1$, so now $(R - S)a_n > 0$ and $a_n < 0$, and the modulus of the value at the other end of the range satisfies

$$|u + v| \leqslant |u| + |v| = u - v$$

and again the functions decrease.

So for both possibilities the Chebyshev representation of a negative exponential function decreases, and this occurs even though the representation of the function might be very poor. Consequently no instability should be caused by functions of this kind. Functions corresponding to complex latent roots are not covered by this theory, but in practice the method did not suffer from induced instability for any of the equations I tried.

One would also expect that the results for linear equations would indicate similar properties for the linearization method for non-linear equations. This was verified in practice for one type of non-linear equation which has properties similar to equations (x) and (xi). The equations considered were obtained from one-dimensional Schrödinger equations by the Madelung

364

transformation (Hartree, 1958, p. 154). The solution of the transformed equation is slowly varying compared with the solution of the original equation, but with the Runge-Kutta and certain predictor-corrector methods I tried, a small interval was still needed for stability. The linearization method, on the other hand, was stable for large intervals.

The Picard–Chebyshev method, on the other hand, behaves in a much less satisfactory manner for this type of equation. In theory the Picard process converges in any interval for linear equations with constant coefficients, since it just produces the truncated Taylor's series of corresponding order, and this series is always convergent for exponential and trigonometric functions. The process does not always converge, however, if it is carried out using Chebyshev series with a fixed number of terms.

I tried both examples (x) and (xi) using a fixed number of terms which were sufficient to represent the solutions but not the complementary functions. Here the iteration behaved asymptotically, and the optimum accuracy obtained was greater when the number of terms was larger; also, when different initial conditions were used, the process diverged sooner. If the number of terms was sufficient to represent the complementary function well, either as a result of using a small interval or a large number of terms, then the process did converge.

This behaviour can again be explained theoretically by considering the form of the iteration. For this purpose it is convenient to consider the form of the equations for a general interval $(a, a + h)$ with initial conditions $y(a) = \alpha$ and the differential equations (22). When this equation is transformed to the interval $(-1, 1)$ the new differential equation will be of the form

$$y^{*\prime} = \frac{h}{2} A y^*, \quad y^* (-1) = \alpha. \tag{25}$$

Putting $B = \frac{h}{2} A$ and replacing $y^*$ by $y$ the iteration takes the form

$$y'^{(j+1)} = B y^{(j)}. \tag{26}$$

As before we get

$$2r c_r = B(a_{r-1} - a_{r+1}), \quad r = 1, \ldots n - 2$$

$$2(n - 1)c_{n-1} = B a_{n-2} \tag{27}$$

$$2n c_n = B a_{n-1}$$

and the initial conditions give

$$\tfrac{1}{2} c_0 = \alpha + c_1 - c_2 \ldots.$$

Here the $a_r$ represent the Chebyshev coefficients of $y^{(j)}$, and the $c_r$ those of $y^{(j+1)}$. In view of the initial conditions which are satisfied by all the $y^{(j)}$ only $n$ of the coefficients are linearly independent. So the new coefficients $c_r$ can be expressed just in terms of $a_1, \ldots, a_n$.

In equation (27) $a_0$ occurs only in the equation for $c_1$. Using the initial conditions on the $a_r$ we get

$$c_1 = \frac{B}{2} \{2(\alpha + a_1 - a_2 \ldots) - a_2\},$$

and the iteration may be expressed in partitioned matrix form

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ . \\ . \end{bmatrix} = \begin{bmatrix} B & -\tfrac{3}{2}B & B & -B \ldots \\ \tfrac{1}{4}B & & -\tfrac{1}{4}B & \\ & \tfrac{1}{6}B & & -\tfrac{1}{6}B \\ & & \tfrac{1}{8}B & \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \end{bmatrix} + \begin{bmatrix} B\alpha \\ 0 \\ 0 \\ . \\ . \end{bmatrix}$$

or $c = Pa + d$.

The iteration will converge if all the latent roots of $P$ are less than unity in modulus. For a six-term series, with $n = 5$, I found that $P$ has a latent root $\mu$ satisfying

$$\tfrac{3}{2}|\lambda| > |\mu| > |\lambda|,$$

where $\lambda$ is any latent root of $B$.

It is now interesting to see the effect of changing the interval $h$. If $v$ is a latent root of $A$ then the corresponding latent root of $B$ will be $\lambda = \frac{h}{2}v$, and the inequality becomes

$$\tfrac{3}{4}|hv| > |\mu| > \tfrac{1}{2}|hv|.$$

In example (xi) with $v = -19$ we therefore require $h < 2/19$ for convergence with this number of terms.

As the number of terms increases one would expect the latent roots to become smaller and eventually to be all less than unity in modulus.

The asymptotic behaviour of the iteration is explained by the choice of the initial vector. If the initial vector does not contain any component corresponding to the large negative latent root, then the process will at first converge, but the other components will inevitably be brought in by rounding error and grow rapidly as the iteration proceeds, thus causing divergence.

It is interesting that the convergence of the Picard–Chebyshev process is poor for just those equations where instability occurs with Runge-Kutta and other methods. They are both caused by the presence of the large negative exponential complementary functions.