5. Conclusion

In this paper the class of iterations discussed by Osborne and Michaelson has been examined critically with a view to determining the rate of convergence, and modifications have been suggested which ensure that the iteration is of third-order. The modified iteration (3.2) has already been tested and proved effective. The results obtained have been consistently better than those obtained using the iteration (1.3).

The iteration (4.15) has been tested on a family of matrices all of which depended non-linearly on the eigenvalue parameter. These eigenvalue problems were also solved using the iteration (1.3). In terms of the number of iterations required there was nothing to choose between the two methods (ten runs using equation (4.15) required a total of 44 iterations, the same problems with the same starting values required a total of 43 iterations using equation (1.3)). Thus equation (4.15) provided the more efficient procedure for the problems considered.

References

Osborne, M. R., and Michaelson, S. (1964). "The numerical solution of eigenvalue problems in which the eigenvalue parameter appears nonlinearly, with an application to differential equations," *The Computer Journal*, Vol. 7, pp. 58–65.

Wilkinson, J. H. (1961). "Rigorous Error Bounds for Computer Eigensystems," *The Computer Journal*, Vol. 4, pp. 230–241.

Wilkinson, J. H. (1958). "The Calculation of the Eigenvectors of Codiagonal Matrices," *The Computer Journal*, Vol. 1, pp. 90–96.

An error analysis of finite-difference methods for the numerical solution of ordinary differential equations

By M. R. Osborne*

A method is given for the calculation of strict, a-posteriori error bounds for the numerical solution by finite-difference methods of ordinary linear differential equations. The suggested procedure is illustrated by some numerical results for a particular differential equation.

1. Introduction

This paper is concerned with the derivation of strict, a-posteriori error bounds for the solution of ordinary differential equations by finite-difference methods. The basic idea of the method of error analysis is due to J. H. Wilkinson who has applied it to obtain strict error bounds for the solutions of sets of linear algebraic equations (see Wilkinson (1963) and the references quoted there). He argues as follows. Let the set of linear equations to be solved be

$$\boldsymbol{A x} = \boldsymbol{b} \tag{1.1}$$

In any process of calculation rounding errors almost always occur so that the process of numerical solution leads not to $\boldsymbol{x}$ but to a vector $\boldsymbol{z}$ satisfying the system of equations

$$(A + \delta A)\boldsymbol{z} = \boldsymbol{b} + \delta\boldsymbol{b}. \tag{1.2}$$

By a careful analysis, bounds can be found for the magnitudes of the elements of δA and $\delta\boldsymbol{b}$. By suitably combining equations (1.1) and (1.2) and taking norms the result is obtained that

$$||\boldsymbol{x} - \boldsymbol{z}|| \leqslant \frac{||(A + \delta A)^{-1}||}{1 - ||(A + \delta A)^{-1}||\,||\delta A||} \{||\delta\boldsymbol{b}|| + ||\delta A||\,||\boldsymbol{z}||\}. \tag{1.3}$$

At this stage the quantities on the right-hand side of equation (1.3) can be estimated with the exception of $||(A + \delta A)^{-1}||$ but, as equation (1.2) is the equation which is actually solved, there remains the possibility of obtaining at least an upper bound for the norm of the inverse of $(A + \delta A)$.

An essential feature of the argument is the representation of the vector $\boldsymbol{z}$ as the solution of a set of linear equations. This permits the original problem to be treated as a perturbation of the one actually solved once bounds have been obtained for δA and $\delta\boldsymbol{b}$.

In general an error analysis of Wilkinson type does not seem to be applicable to the numerical solution of differential equations by finite-difference methods. This is because the process of solution requires the inversion of an operator of a different kind (the operator associated with a finite system of linear or non-linear algebraic equations) to the operator in the original problem.

This difficulty can be avoided in the case of finite-difference approximation to ordinary linear differential equations. Here it can be shown that the solution of the differential equation is also the solution of a linear difference equation, and bounds can be given for the difference between the coefficients in this equation and those in the equation produced by finite-difference approximation. The exact difference equation leads to

* *Edinburgh University Computer Unit*, 7, *Buccleuch Place, Edinburgh* 8.

a set of linear equations which can be treated as a perturbation of the set obtained by finite-difference approximation. Thus, in this case, a theory of Wilkinson type can be developed.

Another possibility for making an error analysis is to interpolate the solution of the finite-difference equations and then to compare this interpolation with the solution of the differential equation. In this case the method of interpolation provides another possible source of error. An analysis of this kind has been given in Schröder (1960).

It is readily shown that any solution of an ordinary linear differential equation exactly satisfies a certain difference equation. Consider the nth order ordinary differential equation

$$L_n(y) = 0. \tag{1.4}$$

This equation has a fundamental set of linearly independent solutions $U_1(x), \ldots, U_n(x)$, and any solution to (1.4) can be expressed as a linear combination of them. Let $y(x)$ be such a solution, then the linear dependence implies

$$\begin{vmatrix} y(x_1) & y(x_2) & \ldots & y(x_{n+1}) \\ U_1(x_1) & U_1(x_2) & \ldots & U_1(x_{n+1}) \\ \cdots & \cdots & \cdots & \cdots \\ U_n(x_1) & U_n(x_2) & \ldots & U_n(x_{n+1}) \end{vmatrix} = 0 \tag{1.5}$$

for arbitrary distinct points $x_1, x_2 \ldots . x_{n+1}$. Equation (1.5) gives the difference equation in the form

$$\sum_{i=1}^{n+1} a_i y(x_i) = 0 \tag{1.6}$$

where the a_i are the signed cofactors of the $y(x_i)$.

Consider now the inhomogeneous equation

$$L_n(y) = f(x). \tag{1.7}$$

Using the method of variation of parameters, a solution of this equation can be obtained in the form

$$y(x) = \int^x f(t)\left\{\sum_{j=1}^{n} b_j(t)U_j(x)\right\} dt. \tag{1.8}$$

If this expression for $y(x)$ is inserted in equation (1.6), the result is

$$\sum_{i=1}^{n+1} a_i y(x_i) = \sum_{i=1}^{n+1} a_i \int_{x_1}^{x_i} f(t)\left\{\sum_{j=1}^{n} b_j(t)U_j(x_i)\right\} dt. \tag{1.9}$$

The right-hand side has the form

$$\int_{x_1}^{x_{(n+1)}} K(t)f(t)dt$$

where $K(t)$ is determined solely by the differential operator L_n. Equation (1.9) is the desired difference equation.

In this paper the errors introduced by rounding are ignored (so that all arithmetic is assumed exact), and only perturbations of the finite difference equations to account for truncation error are considered. In taking norms the maximum norm for vectors is used together with its subordinate matrix norm (maximum row sum of the modulus of the matrix).

In the following sections only second-order differential equations are considered. However, there does not appear to be any difficulty of principle in extending the results obtained here to differential equations of higher order. The results which form the basis of the method are derived in the next section. The manner in which these are applied to yield strict bounds for the error in the numerical solution is described in Section 3. Two methods for obtaining upper bounds for the norm of the inverse of the matrix of the finite-difference equations are given in Section 4, and a numerical example is given in Section 5.

2. Derivation of the basic results

This section is devoted to a derivation of the following result.

If $x_{i-1} = x_i - h \leqslant \xi \leqslant x_i + h = x_{i+1}$, y satisfies the differential equation

$$y^{(2)} + py^{(1)} + qy = f, \tag{2.1}$$

and

$$K_i = \frac{h^2}{8}\{Q_i + 16P_i/h\} < 1 \tag{2.2}$$

where

$$\left.\begin{aligned} Q_i &= \max.\,|q(x)| \\ P_i &= \max.\,|p(x)| \end{aligned}\right\} x_{i-1} \leqslant x \leqslant x_{i+1},$$

then

$$y(\xi) = \alpha^i_{i-1}y_{i-1} + \alpha^i_i y_i + \alpha^i_{i+1}y_{i+1} + \frac{\beta_i}{2h}(y_{i+1} - y_{i-1}) + \gamma_i \tag{2.3}$$

where

$$|\alpha^i_{i-1}| + |\alpha^i_i| + |\alpha^i_{i+1}| \leqslant (1 + 3h^2Q_i/8)/(1 - K_i),$$

$$|\beta_i| \leqslant h^2P_i/8\,(1 - K_i),$$

$$|\gamma_i| \leqslant h^2F_i/8\,(1 - K_i),$$

$$F_i = \max.\,|f(x)|, \quad x_{i-1} \leqslant x \leqslant x_{i+1}.$$

Also

$$y^{(1)}(\xi) = \sigma^i_{i-1}y_{i-1} + \sigma^i_i y_i + \sigma^i_{i+1}y_{i+1} + \frac{\delta_i}{2h}(y_{i+1} - y_{i-1}) + \nu_i \tag{2.4}$$

where

$$|\sigma^i_{i-1}| + |\sigma^i_i| + |\sigma^i_{i+1}| \leqslant 6hQ_i/(1 - K_i),$$

$$\delta_i \leqslant (1 + 2hP_i)/(1 - K_i),$$

$$|\nu_i| \leqslant 2hF_i/(1 - K_i).$$

Two preliminary formulae are required for the derivation of equations (2.3 and 2.4). First the use of linear interpolation plus remainder gives

$$y(\xi) = \frac{x_i - \xi}{x_i - x_{i-1}} y_{i-1} + \frac{\xi - x_{i-1}}{x_i - x_{i-1}} y_i + y^{(2)}(\xi_1) \frac{(x_i - \xi)(x_{i-1} - \xi)}{2}$$

$$x_{i-1} \leqslant \xi, \xi_1 \leqslant x_i,$$

$$= \frac{x_{i+1} - \xi}{x_{i+1} - x_i} y_i + \frac{\xi - x_i}{x_{i+1} - x_i} y_{i+1} + y^{(2)}(\xi_2) \frac{(x_{i+1} - \xi)(x_i - \xi)}{2}$$

$$x_i \leqslant \xi, \xi_2 < x_{i+1}, \quad (2.5)$$

also

$$y^{(1)}(\xi) = \frac{1}{2h}(y_{i+1} - y_{i-1}) + \int_\eta^\xi y^{(2)}(t)dt$$

where η is defined by $y_{i+1} - y_{i-1} = 2h\, y^{(1)}(\eta)$

$$= \frac{1}{2h}(y_{i+1} - y_{i-1}) + (\xi - \eta) y^{(2)}(\xi_3) \quad (2.6)$$

by an application of the mean-value theorem to the integral term.

Equations (2.3) and (2.4) can now be obtained by iteration. First the second derivatives occurring in equations (2.5) and (2.6) are expressed in terms of y and $y^{(1)}$ using equation (2.1). Equations (2.5) and (2.6) are now applied to transform the resulting expressions into ones in which mean values of y and $y^{(1)}$ are replaced by mesh values of y and mean values of $y^{(2)}$. The mean values of $y^{(2)}$ now have smaller coefficients provided h is small enough. The whole procedure can now be repeated iteratively and leads to equations (2.3) and (2.4) in the limit. To prove convergence, and to derive the inequalities on the coefficients, a dominant process is constructed by replacing all quantities in the iteration by upper bounds for their absolute values in $x_{i-1} \leqslant \xi \leqslant x_{i+1}$. The maximum value of the modulus of $y^{(2)}(\xi)$ in this range is denoted by $||y^{(2)}||_i$.

Taking moduli and inserting upper bounds in equations (2.5) and (2.6) gives

$$|y(\xi)| < |y_{i-1}| + |y_i| + |y_{i+1}| + \frac{h^2}{8} ||y^{(2)}||_i, \quad (2.7)$$

and

$$|y^{(1)}(\xi)| \leqslant \frac{1}{2h} |y_{i+1} - y_{i-1}| + 2h\, ||y^{(2)}||_i. \quad (2.8)$$

Clearly it is sufficient to begin the iteration with $y^{(2)}(\xi)$. Equations (2.1), (2.7) and (2.8) can be combined to give

$$||y^{(2)}||_i \leqslant F_i + \frac{P_i}{2h} |y_{i+1} - y_{i-1}| + Q_i \{|y_{i-1}| + |y_i| + |y_{i+1}|\} + K_i ||y^{(2)}||_i. \quad (2.9)$$

The inequality (2.9) can be solved for $||y^{(2)}||_i$ provided $K_i < 1$ (h small enough), and this is equivalent to inserting upper bounds into the iterative procedure and summing the result. Thus the procedure implied in solving equation (2.9) for $||y^{(2)}||_i$ is the required dominant. The solution is

$$||y^{(2)}||_i \leqslant \frac{F_i}{1 - K_i} + \frac{P_i}{2h} \frac{|y_{i+1} - y_{i-1}|}{1 - K_i} + \frac{Q_i}{1 - K_i} \{|y_{i+1}| + |y_i| + |y_{i-1}|\}. \quad (2.10)$$

The inequalities for the coefficients in equations (2.3) and (2.4) are obtained by taking moduli in equations (2.5) and (2.6), and using equation (2.10).

3. Procedure for the error estimate

The standard finite-difference approximation to equation (2.1) together with an expression for its truncation error can be written

$$\delta^2 y_i + p_i h \mu \delta y_i + h^2 q_i y_i = h^2 f_i + \frac{h^4}{12} \{y^{(4)}(\xi_1) + 2 p_i y^{(3)}(\xi_2)\}. \quad (3.1)$$

Also by repeated differentiation of the differential equation

$$y^{(4)}(\xi_1) + 2\, p_i y^{(3)}(\xi_2) = A_i y(\xi_1) + B_i y^{(1)}(\xi_1) + C_i y(\xi_2) + D_i y^{(1)}(\xi_2) \quad (3.2)$$

where bounds for A_i, B_i, C_i, and D_i can readily be computed in terms of bounds for p, q, f and their first and second derivatives.

Formulae (2.3) and (2.4) can now be applied to write the right-hand side of equation (3.2) in terms of y_{i-1}, y_i and y_{i+1}. Substituting the result in equation (3.1) leads to a difference equation which is satisfied exactly by the desired solution of the differential equation, and which has been obtained by perturbing the difference equation obtained by finite-difference approximation. A method similar to that described above can be used to construct perturbed forms of the finite-difference approximations to the initial or boundary conditions which are satisfied exactly by the desired solution of the differential equation.

Let the matrix form of the finite-difference equations be

$$\boldsymbol{Mz} = \boldsymbol{b}, \quad (3.3)$$

then the matrix form of the perturbed equations satisfied by the values of the exact solution at the mesh points can be written

$$(\boldsymbol{M} + \boldsymbol{R} + \boldsymbol{S})\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{c} \quad (3.4)$$

where $\boldsymbol{R}$ is the matrix constructed from the α_i and σ_i, $\boldsymbol{S}$ the matrix constructed from the β_i and δ_i (so that it has $1/h$ as a scalar multiplier), and $\boldsymbol{c}$ the vector constructed from γ_i and ν_i.

The vector $\boldsymbol{y}$–$\boldsymbol{z}$ satisfies

$$(\boldsymbol{M} + \boldsymbol{R} + \boldsymbol{S})(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{c} - (\boldsymbol{R} + \boldsymbol{S})\boldsymbol{z}, \quad (3.5)$$

whence

$$(\boldsymbol{I} + \boldsymbol{M}^{-1}(\boldsymbol{R} + \boldsymbol{S}))(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{M}^{-1}(\boldsymbol{c} - (\boldsymbol{R} + \boldsymbol{S})\boldsymbol{z}). \tag{3.6}$$

Taking norms in equation (3.6) gives the result

$$||\boldsymbol{y} - \boldsymbol{z}|| \leqslant \frac{||\boldsymbol{M}^{-1}||}{1 - ||\boldsymbol{M}^{-1}||\,||\boldsymbol{R} + \boldsymbol{S}||} \{||\boldsymbol{c}|| + ||\boldsymbol{R}||\,||\boldsymbol{z}|| + ||\boldsymbol{S}\boldsymbol{z}||\} \tag{3.7}$$

provided $||\boldsymbol{M}^{-1}||\,||\boldsymbol{R} + \boldsymbol{S}|| < 1.$

Equation (3.7) is in a form suitable for the calculation of error bounds except for the term $||\boldsymbol{Sz}||$. This can be replaced by $||\boldsymbol{S}||\,||\boldsymbol{z}||$; however this is in general a poor upper bound. A better bound can be found by a more careful analysis of $\boldsymbol{S}$. This is easiest when the boundary conditions on equation (2.1) have the form $y(a) = y(b) = 0$. In this case the ith row of $\boldsymbol{Sz}$ is (using equations (3.1), (3.2), (2.3), and (2.4))

$$\frac{h^3}{24}\{A_i\beta_i(\xi_1) + B_i\delta_i(\xi_1) + C_i\beta_i(\xi_2) + D_i\delta_i(\xi_2)\} (z_{i+1} - z_{i-1}) = G_i(z_{i+1} - z_{i-1}). \tag{3.8}$$

If the mesh is chosen so that z_0 corresponds to y(a), and z_{n+1} to y(b), then $\boldsymbol{Sz}$ may be written (using the boundary conditions)

$$\boldsymbol{Sz} = \begin{bmatrix} G_1 & & & \\ & G_2 & & \\ & & \ddots & \\ & & & G_n \end{bmatrix} \begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & -1 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \boldsymbol{GHz}. \tag{3.9}$$

Now

$$||\boldsymbol{Sz}|| \leqslant ||\boldsymbol{G}||\,||\boldsymbol{Hz}|| \tag{3.10}$$

where all quantities on the right-hand side of this equation can be computed. Equation (3.10) provides a much smaller bound for $||\boldsymbol{Sz}||$ in general. This is exemplified in Table 5.1.

4. Bounds for the norm of the inverse matrix

To carry out the error analysis described in Section 3 it is necessary to know a bound for $||\boldsymbol{M}^{-1}||$. In this section two possible methods for finding such a bound are noted. One method is appropriate for initial-value problems, while the second is more suitable for boundary-value problems. The required bound for $||\boldsymbol{M}^{-1}||$ is sought as a bound for $||\boldsymbol{z}||/||\boldsymbol{b}||$ over all non zero vectors $\boldsymbol{b}$.

Consider first the solution of equation (2.1) with given initial conditions. If the standard finite-difference approximations are made, then $\boldsymbol{M}$ is the lower-triangular matrix

$$\begin{bmatrix} 1 & & & & \\ -1 + \frac{h^2q_0}{2} & 1 & & & \\ 1 - hp_1/2 & -2 + h^2q_1 & 1 + hp_1/2 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ & & 1 - hp_{(n-1)}/2 & -2 + h^2q_{(n-1)} & 1 + hp_{(n-1)}/2. \end{bmatrix} \tag{4.1}$$

In this case equation (3.3) can readily be solved by a process of forward substitution. A bound for the $||\boldsymbol{M}^{-1}||$ is found by solving a related problem whose solutions dominate those of equation (3.3).

By collecting terms and introducing the forward differences operator Δ the equation determining z_{i+1} becomes

$$\left(1 + \frac{hp_i}{2}\right)\Delta z_i - \left(1 - \frac{hp_i}{2}\right)\Delta z_{i-1} + h^2q_iz_i = b_i. \tag{4.2}$$

Taking moduli in this equation, and inserting a bound for $|b_i|$, gives

$$\left(1 - \frac{h|p_i|}{2}\right)|\Delta z_i| \leqslant \left(1 + \frac{h|p_i|}{2}\right)|\Delta z_{i-1}| + h^2|q_i||z_i| + ||\boldsymbol{b}||. \tag{4.3}$$

$$\left.\begin{aligned} \text{Also} \quad & |z_0| \leqslant ||\boldsymbol{b}||, \\ & |\Delta z_0| \leqslant ||\boldsymbol{b}||\left(1 + \frac{h^2q_0}{2}\right), \\ \text{and} \quad & |z_i| \leqslant |z_0| + \sum_{j=0}^{i-1}|\Delta z_j|. \end{aligned}\right\} \tag{4.4}$$

The required dominant is now readily found for, by equation (4.4), it is sufficient for it to have forward differences which bound the moduli of the forward differences of z_i provided its initial value and initial forward difference bound $|z_0|$ and $|\Delta z_0|$. The bounds for the forward differences of z_i can be found from equation (4.3) provided $h|p_i|/2 < 1$ for $i = 1, 2, \ldots, n-1$. A suitable dominant for $\frac{|z_i|}{||\boldsymbol{b}||}$ for arbitrary $\boldsymbol{b}$ is Z_i^*, where Z_i^* satisfies the difference equation

$$\left(1 - h\frac{|p_i|}{2}\right)Z_{i+1}^* - (2 + h^2|q_i|)Z_i^* + \left(1 + h\frac{|p_i|}{2}\right)Z_{i-1}^* = 1, \quad i = 1, 2, \ldots, n-1, \tag{4.5}$$

and the initial conditions

$$\left.\begin{aligned} Z_0^* &= 1, \\ \Delta Z_0^* &= 1 + \frac{h^2|q_0|}{2}. \end{aligned}\right\} \tag{4.6}$$

As the Z_i^* are monotonically increasing with i it follows that

$$||M^{-1}|| \leqslant Z_n^*. \quad (4.7)$$

An explicit bound can be found by solving the difference equation with constant coefficients

$$\left(1-\frac{hP}{2}\right)W_{i+1}-(2+h^2Q)W_i+\left(1+\frac{hP}{2}\right)W_{i-1}=1, \quad (4.8)$$

with

$$W_0 = 1, \quad \Delta W_0 = 1 + \frac{h^2 Q}{2} \quad (4.9)$$

and P and Q are numbers such that

$$|q_i| \leqslant Q, \quad i = 0, 1, \ldots, n-1$$

and

$$|p_i| \leqslant P, \quad i = 1, 2, \ldots, n-1.$$

It is necessary that $\frac{hP}{2} < 1$. In this case

$$||M^{-1}|| \leqslant Z_n^* \leqslant W_n. \quad (4.10)$$

If boundary conditions are imposed on equation (2.1) then the matrix M is no longer triangular. The set of linear equations can be solved by factorizing M into an upper and a lower-triangular matrix, and carrying out a forward and back substitution. Specifically let

$$M = (I + L)\,(D + U) \quad (4.11)$$

where D is a diagonal matrix, and L and U are matrices (respectively lower and upper-triangular) with zeros on the diagonal. Then the solution of (3.3) is obtained by solving the equations

$$(I + L)r = b, \quad (4.12a)$$

and

$$(D + U)z = r. \quad (4.12b)$$

To obtain a bound for $||M^{-1}||$ the set of equations

$$(I - |L|)(|D| - |U|)Z^* = e \quad (4.13)$$

is solved where e is the vector each of whose elements is 1, and where the modulus signs indicate that the elements of the matrices concerned are to be replaced by their moduli. The desired result is that

$$||M^{-1}|| \leqslant ||Z^*||. \quad (4.14)$$

The proof is straightforward. From equation (4.12a)

$$\left.\begin{aligned} y_i &= b_i - \sum_{i=1}^{j-1} l_{ij} y_j, \\ \text{whence} \quad |y_i| &\leqslant ||b|| + \sum_{j=1}^{i-1} |l_{ij}|\,|y_j|. \end{aligned}\right\} \quad (4.15)$$

Let

$$(I - |L|)y^* = ||b||\,e, \quad (4.16)$$

then from equations (4.15) and (4.16)

$$y_i^* \geqslant |y_i|. \quad (4.17)$$

Now from equation (4.12b)

$$z_i = \frac{1}{d_i}\left\{y_i - \sum_{j=i+1}^{n} u_{ij} z_j\right\}$$

whence, using equation (4.17),

$$|z_i| \leqslant \frac{1}{|d_i|}\left\{y_i^* + \sum_{j=i+1}^{n} |u_{ij}|\ |z_j|\right\}.$$

Now let

$$(|D| - |U|)W = y^*, \quad (4.18)$$

then

$$W_i \geqslant z_i \quad (4.19)$$

for any vector b. Also $\left(\frac{W_i}{||b||}\right) = Z^*$ so that (4.14) follows from (4.19).

The author was informed about this method for calculating a bound for $||M^{-1}||$ by Mr. Sidney Michaelson (unpublished). The above analysis has a close connection with a method due to Milne (Milne (1949)) which is also reported in Bodewig (1959). The method is particularly valuable for matrices obtained by making finite-difference approximations to boundary-value problems, because it gives the exact value of $||M^{-1}||$ for an important class of these problems.

If the above method is to give the exact value of $||M^{-1}||$ then

$$\pm (I + L)(D + U) = (I - |L|)(|D| - |U|). \quad (4.20)$$

In this case either M or $-M$ is monotonic in the sense of Collatz (1960), p. 42. In the same reference p. 178, Collatz gives an error calculation in which the fact that M is monotone is used essentially to find a bound for $||M^{-1}||$. His method also requires him to know bounds for the solution of the differential equation, and in his numerical example he is forced to substitute values from the solution of the finite-difference equations to obtain "approximate" bounds.

5. An example

In this section the results of the previous sections are exemplified by applying them to the Numerov difference approximation to the differential equation

$$y^{(2)} + (1 + x^2)y = -1 \quad (5.1)$$

subject to the boundary conditions

$$y(-1) = y(1) = 0. \quad (5.2)$$

The exact solution of equation (5.1) also satisfies

$$\left(1+\frac{h^2(1+x_{i-1}^2)}{12}\right)y_{i-1}-\left(2-\frac{5h^2(1+x_i^2)}{6}\right)y_i + \left(1+\frac{h^2(1+x_{i+1}^2)}{12}\right)y_{i+1} = -h^2 - \frac{h^6}{240}y^{(6)}(\xi_i), \quad (5.3)$$

where $x_{i-1} \leqslant \xi_i \leqslant x_{i+1}$. The Numerov difference approximation is obtained by ignoring the term $h^6 y^{(6)}(\xi_i)$.

By repeated differentiation of equation (5.1) it is found that

$$y^{(6)}(\xi) = A_1(\xi)y(\xi) + A_2(\xi)y^{(1)}(\xi) + A_3(\xi), \quad (5.4)$$

where $A_1(\xi) = (1 + \xi^2)^3 - 14(1 + \xi^2) - 16\xi^2$,

$$A_2(\xi) = 12\,\xi(1 + \xi^2),$$

$$A_3(\xi) = (1 + \xi^2)^2 - 12.$$

The exact difference equation can now be constructed by the procedure described in Section 3. The bounds for A_1, A_2, and A_3, in $-1 \leqslant x \leqslant 1$ are

$$|A_1| \leqslant 36,\ |A_2| \leqslant 24,\ |A_3| \leqslant 8;$$

also $(1 + x^2) \leqslant 2$ in this range. Using these bounds and equations (2.3) and (2.4) one finds

(a) $||\boldsymbol{R}|| \leqslant \frac{h^6}{20} \frac{3 + 24h + 9h^2/4}{1 - h^2/4}$,

(b) $||\boldsymbol{G}|| \leqslant \frac{1}{2h} \frac{h^6}{10} \frac{1}{1 - h^2/4}$,

(c) $||\boldsymbol{S}|| \leqslant 2||\boldsymbol{G}||$,

(d) $||\boldsymbol{c}|| \leqslant \frac{h^6}{240} \frac{16h + h^2}{1 - h^2/4}$.

For this example it is found by calculation that the triangular factors satisfy equation (4.20) so that the second procedure of Section 4 actually gives the exact value for $||\boldsymbol{M}^{-1}||$. In fact because of the special form of the right-hand side in this example

$$||\boldsymbol{M}^{-1}|| = h^{-2}\,||z||. \quad (5.5)$$

Table 5.1 gives a list of the numerical values found for the quantities involved in evaluating $||\boldsymbol{y} - \boldsymbol{z}||$ for three different values of h. In Table 5.2 are given the calculated values of $||\boldsymbol{y} - \boldsymbol{z}||$ and, for comparison, the difference between the true solution and the solution of the difference equation at $x = 0$. The value of $y(0)$ is 0.932053718. It will be seen that our error analysis gives quite a good indication of the order of magnitude of the error in the solution of the difference equation.

Table 5.1. Values of quantities relevant to the error analysis

h	.2	.1	.05
$\|\|\boldsymbol{R}\|\|$	2.65×10^{-5}	2.72×10^{-7}	3.29×10^{-9}
$h\,\|\|\boldsymbol{G}\|\|$	3.24×10^{-6}	5.01×10^{-8}	7.81×10^{-10}
$\|\|\boldsymbol{S}\|\|$	3.24×10^{-5}	1.00×10^{-6}	3.12×10^{-8}
$\|\|\boldsymbol{c}\|\|$	8.73×10^{-7}	4.18×10^{-9}	5.23×10^{-11}
$\|\|z\|\|$	0.931945	0.932047	0.932053
$\|\|\boldsymbol{M}^{-1}\|\|$	23.2986	93.2047	372.821
$\frac{\|\|\boldsymbol{M}^{-1}\|\|\ \|\|\boldsymbol{Hz}\|\|}{h}$	68.2735	301.264	1253.46
$\frac{\|\|\boldsymbol{S}\|\|\ \|\|z\|\|}{\|\|\boldsymbol{G}\|\|\ \|\|\boldsymbol{Hz}\|\|}$	3.19	5.75	11.1

Table 5.2. Comparison between predicted and calculated errors

h	.2	.1	.05
$\|\|\boldsymbol{y} - \boldsymbol{z}\|\|$	8.18×10^{-4}	4.31×10^{-5}	2.14×10^{-6}
$y(0) - z(0)$	1.09×10^{-4}	6.7×10^{-6}	4×10^{-7}

References

Bodewig, E. (1959). *Matrix Calculus* (second edition), Amsterdam: North-Holland Publishing Company.
Collatz, L. (1960). *Numerical Treatment of Differential Equations*, Berlin: Springer.
Milne, W. E. (1949). *Numerical Calculus*, Princeton University Press.
Schröder, J. (1960). "Error estimates for boundary value problems using fixed-point theorems," included in *Boundary Problems in Differential Equations* (Editor R. E. Langer), University of Wisconsin Press.
Wilkinson, J. H. (1963). *Rounding Errors in Algebraic Processes*, H.M. Stationery Office.