Neural, Parallel, and Scientific Computations 24 (2016) 269-276

ON THE CONVERGENCE OF A FINITE DIFFERENCE SCHEME FOR A SECOND ORDER DIFFERENTIAL EQUATION CONTAINING NONLINEARLY A FIRST DERIVATIVE

S. MCKEE, JOSÉ A. CUMINATO, AND R. K. MOHANTY

Department of Mathematics and Statistics, University of Strathclyde, Livingstone
Tower, 26 Richmond Street, Glasgow G1 1XH, UK, s.mckee@strath.ac.uk
Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo,
São Carlos, SP, Brazil, jacuminato@gmail.com
Department of Applied Mathematics, South Asian University, Akbar Bhawan,
Chanakyapuri, New Delhi - 110021, India, rmohanty@sau.ac.in

ABSTRACT. This note is concerned with the convergence of a finite difference scheme to the solution of a second order ordinary differential equation with the right-hand-side nonlinearly dependent on the first derivative. By defining stability as the linear growth of the elements of the inverse of a certain matrix and combining this with consistency, convergence is demonstrated. This stability concept is then interpreted in terms of a root condition.

Key Words: Finite Difference, Numerical Methods for ODE's, Multistep Methods

Subject Classification: 65L06, 65L12, 65L20

1. Introduction

Convergence of finite difference methods (linear multistep methods) to $y' = f(t,y)$ was first studied by [2]. Since that famous paper, the subject has been treated by a number of authors, most notably [4] whose analysis is quite general in that it includes the methods of [6], [7] and [5].

However, it would appear that no one has considered the problem of convergence for second order ordinary differential equations whose right-hand-side function contains y and y'. This note seeks to remedy this.

Consider

$$y'' = f(t, y(t), y'(t)) \tag{1.1}$$

subject to

$$y(0) = \tilde{y}_0, \quad y(h) = \tilde{y}_1$$

and the associated difference scheme

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f(t_n, y_n, (y_n - y_{n-1})/h) \tag{1.2}$$

Received January 20, 2016

where $y_n \simeq y(t_n)$ is defined on the grid

$$t_n = nh, \quad n = 0, 1, \ldots, N.$$

Here $h = 1/N$ is the (constant) mesh spacing.

It will be assumed that the function f is Lipschitz continuous in its second and third variable, i.e. there exists L_1, L_2 such that

$$|f(t, y, z) - f(t, y^*, z^*)| < L_1|y - y^*| + L_2|z - z^*|. \tag{1.3}$$

2. **Consistency**

Re-write (1.2) as

$$\begin{aligned} y'_n - y_n/h + y_{n-1}/h &= 0 && (2.1)\\ y_{n+1} - 2y_n + y_{n-1} &= h^2 f(t_n, y_n, y'_n) && (2.2) \end{aligned}$$

Consider the totality of (2.1) and (2.2):

$$\left(\begin{array}{cccccccccccc}
1 & & & & & & & & & & & \\
0 & 1 & & & & & & & & & & \\
\frac{1}{h} & -\frac{1}{h} & 1 & & & & & & \text{\Huge 0} & & & \\
1 & -2 & 0 & 1 & & & & & & & & \\
0 & \frac{1}{h} & 0 & -\frac{1}{h} & 1 & & & & & & & \\
0 & 1 & 0 & -2 & 0 & 1 & & & & & & \\
0 & 0 & 0 & \frac{1}{h} & 0 & -\frac{1}{h} & 1 & & & & & \\
0 & 0 & 0 & 1 & 0 & -2 & 0 & 1 & & & & \\
0 & 0 & 0 & 0 & 0 & \frac{1}{h} & 0 & -\frac{1}{h} & 1 & & & \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & -2 & 0 & 1 & & \\
\vdots & & & & & & & & & & & \ddots
\end{array}\right)
\left(\begin{array}{c}
y_0 \\ y_1 \\ y'_1 \\ y_2 \\ y'_2 \\ y_3 \\ y'_3 \\ y_4 \\ y'_4 \\ y_5 \\ \vdots
\end{array}\right) =$$

$$h^2 \left(\begin{array}{ccccccccccc}
0 & & & & & & & & & & \\
0 & 0 & & & & & & & & & \\
0 & 0 & 0 & & & & & & \Large 0 & & \\
0 & 0 & 1 & 0 & & & & & & & \\
0 & 0 & 0 & 0 & 0 & & & & & & \\
0 & 0 & 0 & 0 & 1 & 0 & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \\
\vdots & & & & & & & & & & \ddots
\end{array}\right)
\begin{pmatrix} f_0 \\ f_1 \\ f_1 \\ f_2 \\ f_2 \\ f_3 \\ f_3 \\ f_4 \\ f_4 \\ f_5 \\ \vdots \end{pmatrix}
+
\begin{pmatrix} \tilde{y}_0 \\ \tilde{y}_1 \\ 0 \\ 0 \\ \vdots \\ \\ \\ \\ \\ \\ \\ \end{pmatrix}$$

or

$$A_h \mathbf{x_h} = h^2 B_h \mathbf{F_h} + \mathbf{g_h} \tag{2.3}$$

where

$$\begin{aligned}
\mathbf{x_h} &= (y_0, y_1, y_1', y_2, y_2', \cdots, y_{N-1}, y_{N-1}', y_N)^T, & (2.4)\\
\mathbf{F_h} &= (f(t_0, y_0, y_0'), f(t_1, y_1, y_1'), f(t_1, y_1, y_1'), f(t_2, y_2, y_2'), f(t_2, y_2, y_2'), \\
& \quad \ldots, f(t_{N-1}, y_{N-1}, y_{N-1}'), f(t_{N-1}, y_{N-1}, y_{N-1}'), f(t_N, y_N, y_N'))^T, & (2.5)\\
\mathbf{g_h} &= (\tilde{y}_0, \tilde{y}_1, 0, \ldots, 0)^T,
\end{aligned}$$

are $2N \times 1$ vectors.

Note A_h and B_h are $2N \times 2N$ matrices with elements

$$(A_h)_{ij}, (B_h)_{ij}, \ i, j = 0, 1, \ldots, 2N-1.$$

Define the vectors

$$\mathbf{x} = (y(t_0), y(t_1), y'(t_1), y(t_2), y'(t_2), \ldots, y(t_{N-1}), y'(t_{N-1}), y(t_N))^T \tag{2.6}$$

and

$$\begin{aligned}
\mathbf{F} = \Big(& f(t_0, y(t_0), y'(t_0)), f(t_1, y(t_1), y'(t_1)), f(t_1, y(t_1), y'(t_1)), f(t_2, y(t_2), y'(t_2)), \\
& f(t_2, y(t_2), y'(t_2)), \ldots, f(t_{N-1}, y(t_{N-1}), y'(t_{N-1})), f(t_{N-1}, y(t_{N-1}), y'(t_{N-1})), \\
& f(t_N, y(t_N), y'(t_N))\Big)^T. \qquad (2.7)
\end{aligned}$$

The local truncation errors associated with (2.1) and (2.2) are, respectively, $O(h^2)$ and $O(h^3)$. We may write the totality of local truncation errors as

$$\boldsymbol{\theta_h} = \Big((\boldsymbol{\theta_h})_0, (\boldsymbol{\theta_h})_1, (\boldsymbol{\theta_h})_2, \ldots, (\boldsymbol{\theta_h})_{2N-2}, (\boldsymbol{\theta_h})_{2N-1}\Big)^T,$$

where

$$(\boldsymbol{\theta_h})_0 = O(h^2) \ \text{ and } \ (\boldsymbol{\theta_h})_1 = O(h^2) \tag{2.8}$$

and

$$(\boldsymbol{\theta_h})_n = \begin{cases} O(h^2), & n \geq 2, \quad n \text{ even} \\ O(h^3), & n \geq 3, \quad n \text{ odd}. \end{cases} \tag{2.9}$$

Thus consistency may be expressed as follows:

$$A_h\mathbf{x} - h^2B_h\mathbf{F} - \mathbf{g} = \boldsymbol{\theta_h}, \tag{2.10}$$

where, here,

$$\mathbf{g} = (y(t_0), y(t_1), 0, \ldots, 0)^T.$$

3. The inverse matrix

The proof of convergence will depend upon the behaviour of A_h^{-1}, and in particular the behaviour of its elements as $N \to \infty$.

Although the results of [3] may be invoked it is a simple matter in this case to compute the elements directly whereupon we observe that there exists an M, independently of N, such that

$$(1/N) \max_{0\leq i\leq j\leq 2N-1} |(A_h^{-1})_{ij}| \leq M.$$

Furthermore, direct computation shows that

$$(A_h^{-1})_{ij} = 0, \quad j = 2, 4, \ldots, 2N-2, \, i \neq j. \tag{3.1}$$

These can be formally established by induction.

We also note that A_h may be written as

$$A_h = \begin{pmatrix} I & 0 \\ d & A_N \end{pmatrix} \tag{3.2}$$

where I is the 2×2 unit matrix and d and A_N are clear. Thus

$$A_h^{-1} = \begin{pmatrix} I & 0 \\ -A_N^{-1}d & A_N^{-1} \end{pmatrix}. \tag{3.3}$$

4. Convergence

Subtract (2.3) from (2.10) to obtain

$$A_h(\mathbf{x} - \mathbf{x_h}) = h^2B_h(\mathbf{F} - \mathbf{F_h}) + \boldsymbol{\theta_h}$$

or

$$\mathbf{x} - \mathbf{x_h} = h^2A_h^{-1}B_h(\mathbf{F} - \mathbf{F_h}) + A_h^{-1}\boldsymbol{\theta_h}. \tag{4.1}$$

i.e. when $i = j$.

We first note that

$$B_h(\mathbf{F} - \mathbf{F_h}) \;=\; (0, 0, 0, f(t_1, y(t_1), y'(t_1)) - f(t_1, y_1, y_1'), 0, f(t_2, y(t_2), y'(t_2))$$

$$-f(t_2, y_2, y_2'), 0, f(t_3, y(t_3), y'(t_3)) - f(t_3, y_3, y_3'), \ldots, 0,$$
$$f(t_{N-1}, y(t_{n-1}, y'(t_{N-1})) - f(t_{N-1}, y_{N-1}, y_{N-1}'))^T.$$

Thus, taking moduli and using the triangle inequality in (4.1) results in

$$(4.2)|(\mathbf{x} - \mathbf{x_h})_{2i+1}| \;\; \leq \;\; h^2 \max_{0\leq i\leq j\leq 2N-1} |(A_h^{-1})_{ij}| \sum_{j=1}^{i} |f(t_j, y(t_j), y'(t_j)) - f(t_j, y_j, y_j')| + \max_{0\leq k\leq 2N-1} |(A_h^{-1}\boldsymbol{\theta_h})_k|$$

and

$$(4.3)\, |(x - x_h)_{2i}| \;\; \leq \;\; h^2 \max_{0\leq i\leq j\leq 2N-1} |(A_h^{-1})_{ij}| \sum_{j=1}^{i-1} |f(t_j, y(t_j), y'(t_j)) - f(t_j, y_j, y_j')| + \max_{0\leq k\leq 2N-1} |(A_h^{-1}\boldsymbol{\theta_h})_k|.$$

By appealing to (4.3) we observe that

$$\begin{aligned} &\max_{0\leq k\leq 2N-1} |(A_h^{-1}\boldsymbol{\theta}_h)|_k| \\ \leq\;& \max\left\{ \max_{0\leq i\leq 2N-1} |(A_h^{-1})_{i0}|\,|(\boldsymbol{\theta}_h)_0|,\ \max_{0\leq i\leq 2N-1} |(A_h^{-1})_{i1}|\,|(\boldsymbol{\theta}_h)_1|, \right. \\ &\left. \max_{0\leq i\leq 2N-1} \left| \sum_{j=2}^{2N-1} (A_h^{-1})_{ij} (\boldsymbol{\theta}_h)_j \right| \right\} \\ =\;& \max\left\{ \max_{0\leq i\leq 2N-1} |(A_h^{-1})_{i0}|\,|(\boldsymbol{\theta}_h)_0|,\ \max_{0\leq i\leq 2N-1} |(A_h^{-1})_{i1}|\,|(\boldsymbol{\theta}_h)_1|, \right. \\ &\left. \max_{0\leq i\leq 2N-1} \left| \sum_{j=1}^{N-1} (A_h^{-1})_{i,2j+1} (\boldsymbol{\theta}_h)_{2j+1} \right| \right\} \end{aligned}$$

using (4.1).

Thus

$$\begin{aligned} &\max_{0\leq k\leq 2N-1} |(A_h^{-1}\boldsymbol{\theta}_h)_k| \\ \leq\;& \max_{0\leq i\leq j\leq 2N-1} |(A_h^{-1})_{ij}| \max\left\{ |(\boldsymbol{\theta}_h)_0|, |(\boldsymbol{\theta}_h)_1|,\ \sum_{j=1}^{N-1} |(\boldsymbol{\theta}_h)_{2j+1}| \right\}. \end{aligned}$$

Furthermore, by using the Lipschitz condition (2.3) we may write

$$|(\mathbf{x} - \mathbf{x}_h)_{2i+1}| \leq h^2 \max_{0\leq i\leq j\leq 2N-1} |(A_h^{-1})_{ij}| \sum_{j=1}^{i} \{L_1|y(t_j) - y_j| + L_2|y'(t_j) - y_j'|\}$$

$$(4.4) \quad + \max_{0\leq i\leq j\leq 2N-1} |(A_h^{-1})_{ij}| \max\{|(\boldsymbol{\theta}_h)_0|, |(\boldsymbol{\theta}_h)_1|,\ \sum_{j=1}^{N-1} |(\boldsymbol{\theta}_h)_{2j+1}|\}.$$

Let $L = \max\{L_1, L_2\}$ and recall that there exists an M, independent of N, such that

$$\max_{0 \le i \le j \le 2N-1} |(A_h^{-1})_{ij}| < MN.$$

Thus the inequality (5.4) becomes

$$\begin{aligned} |(\mathbf{x} - \mathbf{x}_h)_{2i+1}| \ \ &\le \ \ h^2 LMN \sum_{j=0}^{2i} |(\mathbf{x} - \mathbf{x}_h)_j| \\ &\qquad + MN \max\left\{ |(\boldsymbol{\theta}_h)_0|, |(\boldsymbol{\theta}_h)_1|, \sum_{j=1}^{N-1} |(\boldsymbol{\theta}_h)_{2j+1}| \right\} \end{aligned}$$

Note that $N = \frac{1}{h}$ and further that $(\boldsymbol{\theta}_h)_0 = O(h^2)$, $(\boldsymbol{\theta}_h)_1 = O(h^2)$ and $\sum_{j=1}^{N-1} |(\boldsymbol{\theta}_h)_{2j+1}| \le N \max_{1 \le j \le N-1} |(\boldsymbol{\theta}_h)_{2j+1}|$.

Thus we have

$$|(\mathbf{x} - \mathbf{x_h})_{2i+1}| \le \tilde{M} h \sum_{j=0}^{2i} |(\mathbf{x} - \mathbf{x_h})_j| + \delta$$

where $\delta = O(h)$, and $\tilde{M} = LM$ is independent of N.

By a similar argument we observe that

$$|(\mathbf{x} - \mathbf{x_h})_{2i}| \le \tilde{M} h \sum_{j=0}^{2i-1} |(\mathbf{x} - \mathbf{x_h})_j| + \delta.$$

Thus we have

$$|(\mathbf{x} - \mathbf{x_h})_k| \le \tilde{M} h \sum_{j=0}^{k-1} |(\mathbf{x} - \mathbf{x_h})_j| + \delta, \quad k = 1, 2, \ldots, 2N-1$$

and an application of (a mild generalization of) the standard discrete Gronwall lemma (see eg. [1]) results in

$$|(\mathbf{x} - \mathbf{x_h})_k| \le \delta \exp(\tilde{M} kh) \le \delta \exp(\tilde{M}(2N-1)h) < \delta \exp(2\tilde{M})$$

since $Nh = 1$ and so convergence of $O(h)$ has been demonstrated.

The restriction that the initial starting values be $O(h^2)$ is unnecessary; the assumption was employed to minimise the complexity of the argument. From (4.3) we note that d has only a small finite (i.e. independent of N) non-zero elements implying that $(A_N^{-1} d)_{ij}$ are independent of N, that is $(A_h^{-1})_{ij}$ $(i = 0, 1, \ldots, 2N-1, j = 0, 1)$ are independent of N allowing the last term in (5.4) to be replaced by

$$\begin{aligned} \max\Bigg\{ &\max_{0 \le i \le 2N-1} |(A_h^{-1})_{i0}||(\theta_h)_0|, \max_{0 \le i \le 2N-1} |(A_h^{-1})_{i1}||(\theta_h)_1|, \\ &\max_{0 \le i \le j \le 2N-1} |(A_h^{-1})_{ij}| \sum_{j=1}^{N-1} |(\theta_h)_{2j+1}| \Bigg\} \end{aligned}$$

5. **Associated root condition**

When [2] introduced linear multistep methods for solving $y'(t) = f(t, y(t))$ he characterized them by the polynomials $\rho(z)$ and $\sigma(z)$. He defined zero-stability to be the case when $\rho(z)$ has a single root at unity (required for consistency) and all the remaining roots strictly inside the unit circle or lying on the unit circle with multiplicity of one. He then proved that convergence was dependent on zero-stability and consistency.

In this note we have demonstrated convergence subject to the method being consistent and the elements of A_h^{-1} being such that $|(A_h)_{ij}/N|$ are uniformly bounded with respect to N. It is, however, natural to ask if there exists an associated root condition. We shall now demonstrate that such root condition does in fact exist.

First let us characterize the matrix A_N in (3.2) by the two polynomials

$$g_1(A_N, z) = 1 - Nz + Nz^3 \quad \text{and} \quad g_2(A_N, z) = 1 - 2z^2 + z^4.$$

We introduce the functions

$$g_j^n(A_N, z) = \frac{1}{2}\sum_{\ell=1}^{2}(-1)^{2-(\ell-1)(n-1)} g_j(A_N, (-1)^{\ell-1}z), \quad n = 1, 2,$$

and the associated matrix

$$T(A_N, z) = \begin{pmatrix} g_1^1(A_N, z) & g_1^2(A_N, z) \\ g_2^2(A_N, z) & g_2^1(A_N, z) \end{pmatrix}.$$

We now define the vector

$$\mathbf{g}(A_N, z) = (g_1(A_N, z), g_2(A_N, z))^T.$$

It is not difficult to show that

$$T(A_N, z)\mathbf{g}(A_N^{-1}, z) = \mathbf{g}(I, z) = (1, 1)^T,$$

where we must interpret $\mathbf{g}(A_N^{-1}, z)$ as a truncated (vector) power series where all the terms of order $2N$ and above have been neglected.

Since $T(A_N, z)$ is a 2×2 matrix its inverse can be calculated in the normal way from the quotient of its adjoint and its determinant. Each element of its adjoint is simply sums and products of polynomials. Clearly how $\mathbf{g}(A_N^{-1}, z)$ behaves depends on the behaviour of $(\det T(A_N, z))^{-1}$.

In this case $\det |T(A_N, z)|$ is easily computed:

$$\det |T(A_N, z)| = \begin{vmatrix} 1 & -Nz + Nz^3 \\ 0 & 1 - 2z^2 + z^4 \end{vmatrix} = (1 - z)^2(1 + z)^2$$

providing the $O(N)$ growth of the elements of A_N^{-1} which have clearly been established by direct computation. Thus, the associated root condition, i.e. the zero-stability equivalent for the finite difference scheme applied to $y'' = f(t, y, y')$ is

$$\det |T(A_N, z)| = 0.$$

For the differential equation (1.1) this may be stated simply: the polynomial $\det |T(A_N, z)|$ must have a zero at one of multiplicity two (necessary for consistency) and all the other roots must either lie strictly within the unit circle, or if they lie on the unit circle their multiplicity may not be greater than two.

6. Concluding remarks

This note has analyzed a finite difference scheme which approximates a second order differential equation. It has been shown to be zero-stable (in the sense of Dahlquist) and convergent. The fact that this scheme is zero-stable means that consistency, itself, implies convergence. Although this work has treated a simple finite difference scheme, the ideas are quite general and may, in principle, be applied to demonstrate the convergence of other more sophisticated schemes.

Acknowledgements

The first author is grateful to the Institute for Mathematics and its Applications for a travel grant to visit ICMC in Brazil and also to Fapesp grant #2013/07375-0. The second author acknowledges the support of Fapesp grant #2013/07375-0 and CNPq grant #301787/2008-5. The third author would like to acknowledge the receipt of financial support from the Royal Society of Edinburgh under the INSA-RSE Bilateral Exchange Programme 2014.

REFERENCES

[1] R. Bellman, *Stability of Ordinary Differential Equations*, McGraw-Hill, 1953.

[2] G. Dahlquist, Convergence and stability in the numerical integration of ordinary differential equations, *Math. Scand.*, 4:33–53, 1956.

[3] S. McKee, Discretization methods and block isoclinal matrices, *IMA J. Numer. Anal.*, 3:467–491, 1983.

[4] J. C. Butcher, On the convergence of the numerical solutions to ordinary differential equations, *Maths. Comp.*, 20:1–10, 1972.

[5] C. W. Gear, Hybrid methods for initial values problems in ordinary differential equations, *SIAM J. Numer. Anal. Ser.B*, 2:69–86, 1965.

[6] M. Urabe, Theory of errors in numerical integration of ordinary differential equations, *J. Sci. Hiroshima Univ. Ser A-I*, 25:3–62, 1961.

[7] W. B. Gragg and H. J. Stetter, Generalized multistep predictor-corrector methods, *J. Assoc. Comp. Mach*, 11:188–209, 1964.