On the Convergence of Numerical Solutions to Ordinary Differential Equations*

By J. C. Butcher

Numerical methods for the solution of the initial value problem in ordinary differential equations fall mainly into two categories: multi-step methods and Runge-Kutta methods. For these and for some closely related methods, the convergence of the numerical solution to the exact solution as the step size tends to zero, has been studied by a number of authors [1, 2, 3]. It is the aim of the present paper to make a similar study for a fairly general class of method which includes both main classes of method as special cases. Also, it is applicable to methods which combine features common to both multi-step and Runge-Kutta methods such as the methods of Urabe [4], Gragg and Stetter [5] and Gear [6].

Although the standard treatments of convergence theory can be simply modified to include these new methods, there is some advantage in having a theory which includes them in a completely natural way. It is hoped also that some previously untried but useful methods may be suggested by the formalism of this paper.

The initial value problem we suppose can be written in the form

$$(1) \qquad \frac{d\mathbf{y}}{dx} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{y}(x_0) = \boldsymbol{\mathfrak{n}},$$

where $\mathbf{y}$ is a point in the (real) Euclidean M-space R_M and $\mathbf{f}(\mathbf{y})$ is a mapping of R_M onto itself satisfying the Lipschitz condition

$$(2) \qquad |\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z})| \leqq L\,|\mathbf{y} - \mathbf{z}|,$$

for any pair of points $\mathbf{y}$, $\mathbf{z} \in R_M$. L is a constant and $|\mathbf{v}|$ for $\mathbf{v} \in R_M$ denotes a norm. Although the particular norm used is irrelevant for most purposes, a number of details in the results of this paper take a simpler form if the norm used is defined by

$$(3) \qquad |\mathbf{v}| = \max\{|v^1|, |v^2|, \cdots, |v^M|\},$$

$v^1, v^2, \cdots, v^M$ denoting the components of $\mathbf{v}$. Accordingly, we adopt (3) as the definition of $|\mathbf{v}|$.

It will be necessary to consider sets of points $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N \in R_M$ and we shall regard such a set as corresponding to the point $\mathbf{V} = \mathbf{v}_1 \oplus \mathbf{v}_2 \oplus \cdots \oplus \mathbf{v}_N \in R_{MN}$. The norm of $\mathbf{V} \in R_{MN}$ will be defined in a similar way to (3) and a similar notation $|\mathbf{V}|$ will be used. Clearly

$$(4) \qquad |\mathbf{V}| = \max\{|\mathbf{v}_1|, |\mathbf{v}_2|, \cdots, |\mathbf{v}_N|\}.$$

We will have to make use of mappings from R_{MN} to R_{MN} such as $\mathbf{V} \to \mathbf{W} = \mathbf{w}_1 \oplus \mathbf{w}_2 \oplus \cdots \oplus \mathbf{w}_N$, where

Received April 22, 1965.

* Work supported by University of Canterbury, New Zealand, and by U. S. Atomic Energy Commission.

$$\mathbf{w}_i = \sum_{j=1}^{N} a_{ij}\mathbf{v}_j, \qquad i = 1, 2, \cdots, N, \tag{5}$$

and $a_{11}, a_{12}, \cdots, a_{NN}$ are elements of a matrix A. For this mapping we shall use the notation

$$\mathbf{W} = [A]\mathbf{V}, \tag{6}$$

so that $[A]$ is a linear operator on R_{MN} to R_{MN}. $| A |$ will denote the norm $\max_i \sum_{j=1}^{N} | a_{ij} |$ so that

$$|[A]\mathbf{V}| \leqq |A| \cdot |\mathbf{V}|. \tag{7}$$

Another type of mapping that will arise is that given by $\mathbf{V} \to \mathbf{W}$, where

$$\mathbf{w}_i = \mathbf{f}(\mathbf{v}_i), \qquad i = 1, 2, \cdots, N \tag{8}$$

and $\mathbf{f}$ is the function occurring in the statement of the initial value problem (1). We shall write

$$\mathbf{W} = \mathbf{F}(\mathbf{V}) \tag{9}$$

to denote this mapping and we see that $\mathbf{F}$ satisfies a Lipschitz condition with the same constant L as for $\mathbf{f}$.

We are now in a position to formulate the general method with which the rest of this paper is concerned. It consists of the performance of a sequence of steps numbered 1, 2, 3, $\cdots$ such that at the start of step n, N points in R_M are given. We denote these by $\mathbf{y}_1^{(n-1)}, \mathbf{y}_2^{(n-1)}, \cdots, \mathbf{y}_N^{(n-1)}$ and write $\mathbf{Y}^{(n-1)} = \mathbf{y}_1^{(n-1)} \oplus \mathbf{y}_2^{(n-1)} \oplus \cdots \oplus \mathbf{y}_N^{(n-1)}$. At the end of the step $\mathbf{Y}^{(n)} = \mathbf{y}_1^{(n)} \oplus \mathbf{y}_2^{(n)} \oplus \cdots \oplus \mathbf{y}_N^{(n)}$ is given by

$$\mathbf{y}_i^{(n)} = \sum_{j=1}^{N} a_{ij}\mathbf{y}_j^{(n-1)} + h \sum_{j=1}^{N} \{b_{ij}\mathbf{f}(\mathbf{y}_j^{(n)}) + c_{ij}\mathbf{f}(\mathbf{y}_j^{(n-1)})\}, \tag{10}$$

which can be written as

$$\mathbf{Y}^{(n)} = [A]\mathbf{Y}^{(n-1)} + h[B]\mathbf{F}(\mathbf{Y}^{(n)}) + h[C]\mathbf{F}(\mathbf{Y}^{(n-1)}), \tag{11}$$

where the matrices A, B, C with elements a_{ij}, b_{ij}, c_{ij} $(i, j = 1, 2, \cdots, N)$ characterize the method. We interpret $\mathbf{y}_1^{(n-1)}, \mathbf{y}_2^{(n-1)}, \cdots, \mathbf{y}_N^{(n-1)}$ as approximations to $\mathbf{y}(x)$ for a set of N values of x and $\mathbf{y}_1^{(n)}, \mathbf{y}_2^{(n)}, \cdots, \mathbf{y}_N^{(n)}$ as approximations when the values of x are each increased by h (the step size). For simplicity with no loss of generality we shall assume $h > 0$ and that the method is used to find $\mathbf{y}(x)$ only when $x > x_0$.

The method defined by A, B, C will be denoted by (A, B, C) and in the particular case when C is the zero matrix by (A, B). There is no loss of generality in considering only methods of this last form since (A, B, C) is equivalent to $(\bar{A}, \bar{B})$, where

$$\bar{A} = \begin{bmatrix} A & O \\ I & O \end{bmatrix}, \tag{12}$$

$$\bar{B} = \begin{bmatrix} B & C \\ O & O \end{bmatrix} \tag{13}$$

and O, I are the $N \times N$ zero matrix and unit matrix respectively.

Before proceeding, it must be remarked that (11) is of the form

$$\mathbf{Y}^{(n)} = \mathbf{G}(\mathbf{Y}^{(n)}) \tag{14}$$

and in general does not define $\mathbf{Y}^{(n)}$ explicitly. However, if $\mathbf{Y} = \mathbf{y}_1 \oplus \mathbf{y}_2 \oplus \cdots \oplus \mathbf{y}_N$ and $\mathbf{Z} = \mathbf{z}_1 \oplus \mathbf{z}_2 \oplus \cdots \oplus \mathbf{z}_N$ are any two points in R_{MN}, then

$$(15)\quad |\mathbf{G}(\mathbf{Y}) - \mathbf{G}(\mathbf{Z})| = h|[B]\{\mathbf{F}(\mathbf{Y}) - \mathbf{F}(\mathbf{Z})\}| \leqq hL|B|\cdot|\mathbf{Y} - \mathbf{Z}|$$

so that if

$$(16)\qquad h < 1/(L|B|)$$

then $\mathbf{Y} \rightarrow \mathbf{G}(\mathbf{Y})$ is a contraction mapping. Thus if h is sufficiently small, $\mathbf{Y}^{(n)}$ is defined uniquely by (11) and may be evaluated iteratively. For a computer realization of the procedure for evaluating $\mathbf{Y}^{(n)}$, it is more convenient to use an iteration process based on the equation

$$(17)\qquad \mathbf{Y}^{(n)} = \overline{\mathbf{G}}(\mathbf{Y}^{(n)}),$$

where $\overline{\mathbf{G}}(\mathbf{Y}) = \bar{\mathbf{g}}_1(\mathbf{Y}) \oplus \bar{\mathbf{g}}_2(\mathbf{Y}) \oplus \cdots \oplus \bar{\mathbf{g}}_N(\mathbf{Y})$ is related to $\mathbf{G}(\mathbf{Y}) = \mathbf{g}_1(\mathbf{Y}) \oplus \mathbf{g}_2(\mathbf{Y}) \oplus \cdots \oplus \mathbf{g}_N(\mathbf{Y})$ by

$$(18)\qquad \begin{aligned}
\bar{\mathbf{g}}_1(\mathbf{Y}) &= \mathbf{g}_1(\mathbf{y}_1 \oplus \mathbf{y}_2 \oplus \cdots \oplus \mathbf{y}_N),\\
\bar{\mathbf{g}}_2(\mathbf{Y}) &= \mathbf{g}_2(\bar{\mathbf{g}}_1(\mathbf{Y}) \oplus \mathbf{y}_2 \oplus \cdots \oplus \mathbf{y}_N),\\
\bar{\mathbf{g}}_3(\mathbf{Y}) &= \mathbf{g}_3(\bar{\mathbf{g}}_1(\mathbf{Y}) \oplus \bar{\mathbf{g}}_2(\mathbf{Y}) \oplus \cdots \oplus \mathbf{y}_N),\\
&\vdots\\
\bar{\mathbf{g}}_N(\mathbf{Y}) &= \mathbf{g}_N(\bar{\mathbf{g}}_1(\mathbf{Y}) \oplus \bar{\mathbf{g}}_2(\mathbf{Y}) \oplus \cdots \oplus \bar{\mathbf{g}}_{N-1}(\mathbf{Y}) \oplus \mathbf{y}_N).
\end{aligned}$$

With the norm defined by (3), it is trivial to prove that $\mathbf{Y} \rightarrow \overline{\mathbf{G}}(\mathbf{Y})$ is a contraction mapping if the same is true for $\mathbf{Y} \rightarrow \mathbf{G}(\mathbf{Y})$, so that (16) is sufficient for either type of procedure.

To illustrate the variety of methods that can be written in the form (A, B) we note that the multi-step method given by

$$(19)\quad \mathbf{y}_n = q_1\mathbf{y}_{n-1} + \cdots + q_k\mathbf{y}_{n-k} + h(r_0\mathbf{f}(\mathbf{y}_n) + r_1\mathbf{f}(\mathbf{y}_{n-1}) + \cdots + r_k\mathbf{f}(\mathbf{y}_{n-k})),$$

where $\mathbf{y}_n$ denotes the numerical solution at the point $x_0 + nh$, is equivalent to (A, B) with $N = K + 1$ and

$$(20)\qquad A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ 0 & q_k & q_{k-1} & \cdots & q_1 \end{bmatrix},$$

$$(21)\qquad B = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & 0\\ r_k & r_{k-1} & r_{k-2} & \cdots & r_0 \end{bmatrix}.$$

On the other hand an $N - 1$ stage Runge-Kutta process takes the form (A, B) with

$$A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{22}$$

$$B = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ b_{21} & 0 & 0 & \cdots & 0 \\ b_{31} & b_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ b_{N1} & b_{N2} & b_{N3} & \cdots & 0 \end{bmatrix}. \tag{23}$$

In the example of the classical fourth order process we have

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 0 \end{bmatrix}. \tag{24}$$

A final example we consider is neither a linear multi-step nor a Runge-Kutta method. It has the form (A, B) where

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{25}$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{3}{4} & -\frac{1}{4} & 0 & 0 & 0 \\ -2 & 1 & 2 & 0 & 0 \\ \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} & 0 \end{bmatrix}. \tag{26}$$

As it happens, this method yields values of $\mathbf{y}_5^{(n)}$ which differ from $\mathbf{y}(x_0 + nh)$ by about the same amount as for the classical Runge-Kutta method if it is started by the formulae $\mathbf{y}_5^{(0)} = \mathfrak{n}$, $\mathbf{y}_3^{(0)} = \mathfrak{n} - \frac{1}{2}h\mathbf{f}(\mathfrak{n})$. It has the advantage over 4th order Runge-Kutta methods in that it requires only three derivative calculations per step.

We shall not be concerned in this paper with methods of obtaining the starting vector $\mathbf{Y}^{(0)}$ but we shall suppose this is done in such a way that in the limits as $h \to 0$, $\mathbf{y}_i^{(0)} \to \mathfrak{n}$ for $i = 1, 2, \cdots, N$. We now define convergence as follows:

1. (Definition). (A, B) *is said to be convergent if for any initial value problem* (1) *satisfying* (2), *the following statement can be made: If* (A, B) *is used to compute* $\mathbf{Y}^{(\nu)}$ *with step size* $h = (x - x_0)/\nu$, *where* $\mathbf{Y}^{(0)}$ *is given in such a way that* $| \mathbf{Y}^{(0)} - \mathfrak{n} \oplus \mathfrak{n} \oplus \cdots \oplus \mathfrak{n} | \to 0$ *as* $\nu \to \infty$ *then* $| \mathbf{Y}^{(\nu)} - \mathbf{y}(x) \oplus \mathbf{y}(x) \oplus \cdots \oplus \mathbf{y}(x) | \to 0$ *as* $\nu \to \infty$.

Just as for linear multi-step processes it is convenient to introduce concepts of consistency and stability for (A, B). However, it is convenient first of all to consider A by itself.

2. (Definition). *A is consistent if $A\mathbf{s} = \mathbf{s}$, where $\mathbf{s}$ is the vector in R_N with every component equal to unity.*

3. (Definition). *A is stable*[1] *if there is a constant α such that for any positive integer n*

$$|A^n| \leqq \alpha. \tag{27}$$

The following results are consequences of these definitions.

4. *If all eigenvalues of A have magnitude less than 1 except for a simple eigenvalue at 1, A is stable.*

5. *If A is stable, no eigenvalue has magnitude greater than 1.*

6. *If A has minimal polynomial $P(z)$, then A is stable if and only if no zero of $P(z)$ exceeds 1 in magnitude and all roots of magnitude 1 are simple.*

7. *A is stable if and only if there is a nonsingular matrix T such that $|T^{-1}AT| \leqq 1$.*

8. *If A is consistent and has only non-negative elements, then A is stable.*

9. *$\bar{A}$ given by (12) is stable if and only if A is stable.*

10. *$\bar{A}$ given by (12) is consistent if and only if A is consistent.*

11. *A given by (20) is stable if and only if no zero of*

$$Q(z) = z^k - q_1 z^{k-1} - q_2 z^{k-2} - \cdots - q_k \tag{28}$$

exceeds 1 in magnitude and all zeros of magnitude 1 are simple.

12. *A given by (20) is consistent if and only if $Q(z)$ given by (28) has a zero equal to 1.*

13. *A given by (22) is stable.*

14. *A given by (22) is consistent.*

Proofs. **10, 12** and **14** are immediate consequences of the definition of consistency. **4, 5** and **11** are trivial consequences of **6. 13** is an example of **8** which follows from **7** with $T = I$. **9** is immediately seen from the obvious formula

$$\bar{A}^n = \begin{bmatrix} A^n & O \\ A^{n-1} & O \end{bmatrix} \tag{29}$$

so that $|\bar{A}^n| = \max(|A^n|, |A^{n-1}|)$.

It remains to prove **6** and **7**. Let the Jordan canonical form of A be $(\lambda_1 I_1 + \delta_1 J_1) \oplus (\lambda_2 I_2 + \delta_2 J_2) \oplus \cdots \oplus (\lambda_s I_s + \delta_s J_s)$, where the orders of the various blocks are $r_1, r_2, \cdots, r_s$ such that $r_1 + r_2 + \cdots + r_s = N$. I_i $(i = 1, 2, \cdots, s)$ is the $r_i \times r_i$ unit matrix and J_i is the $r_i \times r_i$ matrix with every element zero except those immediately below the main diagonal and these are unity. The λ_i correspond to the eigenvalues of A and the δ_i are arbitrary non-zero numbers. If for any i, $r_i = 1$, J_i consists of the 1×1 zero matrix and the term $\delta_i J_i$ is omitted in such a case. Consider the three statements

S_1 : $|\lambda_i| \leqq 1$ for $i = 1, 2, \cdots, s$ and for all i such that $|\lambda_i| = 1$, $r_i = 1$.

S_2 : T exists such that $|T^{-1}AT| \leqq 1$.

S_3 : A is stable.

From the relationship between the Jordan canonical form and the minimal equation we see that **6** asserts the equivalence of S_1 and S_3. Also **7** asserts the equivalence of S_2 and S_3. We will thus have proved **6** and **7** when we have shown

[1] In the theory of linear operators, the term "power-bounded" is used for this property.

that $S_1 \Rightarrow S_2$, $S_2 \Rightarrow S_3$, and $S_3 \Rightarrow S_1$. To deduce S_2 from S_1 we choose T so that $T^{-1}AT$ is the Jordan canonical form with $\delta_i = 1 - |\lambda_i|$ for every i for which $r_i > 1$. S_3 follows from S_2 since $|A^n| = |T(T^{-1}AT)^nT^{-1}| \leqq |T|\cdot|T^{-1}|$. Finally we deduce S_1 from S_3 by noting that $|(\lambda_iI_i + \delta_iJ_i)^n| \geqq |\lambda_i|^n$ for all i and that $|(\lambda_iI_i + \delta_iJ_i)^n| \geqq n|\lambda_i|^{n-1}|\delta_i|$ whenever $r_i > 1$.

We now state two necessary conditions for convergence.

15. *If* (A, B) *is convergent,* A *is stable.*

16. *If* (A, B) *is convergent,* A *is consistent.*

Proofs. To prove **15** we suppose that (A, B) is convergent but A is not stable and we use (A, B) for the solution of the initial value problem defined by $M = 1$, $f^1 = 0$, $\eta^1 = 0$, $x_0 = 0$, $x = 1$. Let $\alpha_n = |A^n|$ and let $\mathbf{v}_n \in R_N$ be such that $|A^n\mathbf{v}_n| = \alpha_n$, $|\mathbf{v}_n| = 1$. Furthermore, let $\beta_n = \max(\alpha_1, \alpha_2, \cdots, \alpha_n)$ and define $\mathbf{w}_n = \beta_n^{-1}\mathbf{v}_n$ so that, since A is not stable, $|\mathbf{w}_n| \longrightarrow 0$. If we choose $\mathbf{Y}^{(0)}$ as $\mathbf{w}_\nu$, write $h = 1/\nu$ and perform the solution to the initial value problem using (A, B), we find $\mathbf{Y}^{(\nu)} = A^\nu\mathbf{w}_\nu$. Since the method is convergent and the true solution is $y^1(x) = 0$, we have $|A^\nu\mathbf{w}_\nu| \longrightarrow 0$ as $\nu \longrightarrow \infty$. But $|A^\nu\mathbf{w}_\nu| = \alpha_\nu/\beta_\nu$ which equals 1 for an infinite set of values of ν.

To prove **16**, we assume (A, B) is convergent and apply it to the solution of the initial value problem defined by $M = 1$, $f^1 = 0$, $\eta^1 = 1$, $x_0 = 0$, $x = 1$. We choose $\mathbf{Y}^{(0)} = \mathbf{s}$ independently of ν, so that convergence implies that $|A^\nu\mathbf{s} - \mathbf{s}| \longrightarrow 0$ as $\nu \longrightarrow \infty$. But

$$\begin{aligned} |A\mathbf{s} - \mathbf{s}| &\leqq |A^{\nu+1}\mathbf{s} - A\mathbf{s}| + |A^{\nu+1}\mathbf{s} - \mathbf{s}| \\ &\leqq |A|\cdot|A^\nu\mathbf{s} - \mathbf{s}| + |A^{\nu+1}\mathbf{s} - \mathbf{s}| \\ &\longrightarrow 0 \end{aligned}$$

so that $A\mathbf{s} = \mathbf{s}$.

Further definitions and theorems now follow.

17. (Definition). (A, B) *is semi-consistent if* A *is consistent and if there is a* $\mathbf{t} \in R_N$ *and a scalar* c *such that*

$$A\mathbf{t} + B\mathbf{s} = \mathbf{t} + c\mathbf{s}. \tag{30}$$

18. (Definition). (A, B) *is stable if* A *is stable.*

19. *If* (A, B) *is stable and semi-consistent, the value of* c *in* (30) *is unique.*

Proof. If (30) were also satisfied with $\mathbf{t}$, c replaced by $\mathbf{t}'$, c' where $c \neq c'$, we would have $A(\mathbf{t} - \mathbf{t}') = (\mathbf{t} - \mathbf{t}') + (c - c')\mathbf{s}$ so that $\mathbf{t} - \mathbf{t}'$ is a member of the null space of $(A - I)^2$ but not of $A - I$. Hence, the minimal equation of A contains a repeated unit root contrary to **6**.

It may be remarked that $\mathbf{t}$ in (30) is not unique but may be altered by the addition of any null vector (for example $\mathbf{s}$) of $A - I$.

20. *If* A *is consistent and the characteristic equation of* A *has only a simple root at* 1, *then* (A, B) *is semi-consistent.*

Proof. Let V be the range space of $A - I$ so that V is of dimension $N - 1$ and $\mathbf{s} \notin V$. Hence, an arbitrary vector of R_N can be written as a linear combination of $\mathbf{s}$ with a member of V. Write c as the component of $\mathbf{s}$ in $B\mathbf{s}$ and the result follows.

21. (Definition). (A, B) *is consistent if it is semi-consistent and the value of* c *in* (30) *is* 1.

22. *If (A, B) is semi-consistent with $c \neq 0$, $(A, (1/c)B)$ is consistent.*

The proof of this result is immediate. Before proceeding further we return to the examples (A, B) given by (12), (13), by (20), (21) and by (22), (23).

23. *(A, B, C) is semi-consistent (that is, $(\bar{A}, \bar{B})$ given by (12), (13) is semi-consistent) if and only if A is consistent and $\mathbf{t} \in R_N$ and c exist such that*

$$At + (B + C)\mathbf{s} = \mathbf{t} + c\mathbf{s}. \tag{31}$$

24. *If A given by (20) satisfies the conditions of* **11** *and* **12** *so that A is stable and consistent, and if B is given by (21), then (A, B) is semi-consistent with*

$$c = (r_0 + r_1 + \cdots + r_k)/(q_1 + 2q_2 + \cdots + kq_k).$$

25. *If A is given by (22) and B by (23), then (A, B) is stable and semi-consistent with $c = b_{N1} + b_{N2} + \cdots + b_{N,N-1}$.*

Proofs. **23** follows by noting that (31) is equivalent to

$$\bar{A}\bar{\mathbf{t}} + \bar{B}\bar{\mathbf{s}} = \bar{\mathbf{t}} + c\bar{\mathbf{s}}, \tag{32}$$

where $\bar{\mathbf{t}} = \mathbf{t} \oplus (\mathbf{t} - c\mathbf{s})$, $\bar{\mathbf{s}} = \mathbf{s} \oplus \mathbf{s}$. **24** can be verified immediately with $\mathbf{t}$ in (30) such that its component number i is $-c(k + 2 - i)$. The part of **25** not included in **13** and **14** is an example of **20.** It may be remarked at this point that the consistency and stability of (A, B) where A, B are given by (25), (26) follow in a similar way.

We now come to the two main theorems.

26. *If (A, B) is convergent, it is stable and consistent.*

Proof. In view of **15** and **16** we may assume A is stable and consistent if (A, B) is convergent. We need only prove that there is a $\mathbf{t} \in R_N$ such that

$$A\mathbf{t} + B\mathbf{s} = \mathbf{t} + \mathbf{s}. \tag{33}$$

As for the proofs of **15** and **16** we prove this result by considering a special example. We take $M = 1$, $f^1 = 1$, $\eta^1 = 0$, $x_0 = 0$, $x = 1$ and $\mathbf{Y}^{(0)} = 0$ independently of ν. With $h = 1/\nu$ we find

$$\mathbf{Y}^{(\nu)} = \frac{1}{\nu}(A^{\nu-1} + A^{\nu-2} + \cdots + I)B\mathbf{s} \tag{34}$$

and for convergence, this must tend to $\mathbf{s}$ as $\nu \to \infty$. Since A is stable, the range space and the null space of $A - I$ are disjoint so that we may write $B\mathbf{s} - \mathbf{s} = (I - A)\mathbf{t} + \mathbf{v}$ where $\mathbf{v}$ is in the null space of $A - I$. Substitute into (34) and we find

$$\mathbf{Y}^{(\nu)} - \mathbf{s} = \frac{1}{\nu}(I - A^{\nu})\mathbf{t} + \mathbf{v} \tag{35}$$

so that

$$|\mathbf{v}| \leqq |\mathbf{Y}^{(\nu)} - \mathbf{s}| + \frac{1}{\nu}(1 + |A^{\nu}|) \to 0$$

as $\nu \to \infty$. Hence $\mathbf{v} = 0$ so that (33) follows.

27. *If (A, B) is stable and consistent, it is convergent.*

Proof. Let $\mathbf{t}$ in (33) have components $t_1, t_2, \cdots, t_N$. We may assume by the remark following **19** that none of $t_1, t_2, \cdots, t_N$ is negative. We write

$$\mathbf{n}_i^{(n)} = \mathbf{y}(x_0 + h(n + t_i)) \tag{37}$$

for $i = 1, 2, \cdots, N$; $n = 0, 1, \cdots$ where $\mathbf{y}(x)$ denotes the true solution to the initial value problem (1). Also we write $\mathbf{H}^{(n)} = \mathbf{n}_1^{(n)} \oplus \mathbf{n}_2^{(n)} \oplus \cdots \oplus \mathbf{n}_N^{(n)}$ so that, by the continuity of $\mathbf{y}(x)$, convergence will be proved when we have shown that as $\nu \to \infty$ with $h = (x - x_0)/\nu$ and $|\mathbf{Y}^{(0)} - \mathbf{H}^{(0)}| \to 0$ then $|\mathbf{Y}^{(\nu)} - \mathbf{H}^{(\nu)}| \to 0$. It will be assumed that h is no more than some fixed h_0 satisfying (16).

Let $\mathbf{E}^{(n)} = \mathbf{e}_1^{(n)} \oplus \mathbf{e}_2^{(n)} \oplus \cdots \oplus \mathbf{e}_N^{(n)}$ be the truncation error in a single step defined by

$$\mathbf{E}^{(n)} = \mathbf{H}^{(n)} - [A]\mathbf{H}^{(n-1)} - h[B]\mathbf{F}(\mathbf{H}^{(n)}). \tag{38}$$

Our first task is to estimate $\mathbf{E}^{(n)}$. We have

$$\begin{aligned} &y^k(x_0 + h(n + t_i)) - y^k(x_0 + h(n - 1 + t_j)) \\ &\qquad = h(1 + t_i - t_j)f^k(\mathbf{y}(x_0 + h(n + \theta^k))) \end{aligned} \tag{39}$$

by the mean value theorem, where θ^k lies between $t_j - 1$ and t_i. Hence we have

$$\begin{aligned} &\mathbf{y}(x_0 + h(n + t_i)) - \mathbf{y}(x_0 + h(n - 1 + t_j)) \\ &\qquad - h(1 + t_i - t_j)\mathbf{f}(\mathbf{y}(x_0 + nh)) = \mathbf{u}, \end{aligned} \tag{40}$$

where

$$|\mathbf{u}| \leqq h^2Lm|1 + t_i - t_j| \max(t_i, |1 - t_j|) \tag{41}$$

and m is the maximum of the (continuous) function $|\mathbf{f}(\mathbf{y}(x))|$ for

$$x \in [x_0, x + h_0 \max(t_1, t_2, \cdots, t_N)].$$

Multiplying (40) by a_{ij} and summing over j we find

$$\begin{aligned} &\left| \mathbf{n}_i^{(n)} - \sum_{j=1}^{N} a_{ij}\, \mathbf{n}_j^{(n-1)} - h\left(\sum_{j=1}^{N} b_{ij}\right) \mathbf{f}(\mathbf{y}(x_0 + nh)) \right| \\ &\qquad = \left| \sum_{j=1}^{N} a_{ij}\{\mathbf{n}_i^{(n)} - \mathbf{n}_j^{(n-1)} - h(1 + t_i - t_j)\mathbf{f}(\mathbf{y}(x_0 + nh))\} \right| \\ &\qquad \leqq h^2Lm \sum_{j=1}^{N} \{|a_{ij}| \cdot |1 + t_i - t_j| \max(t_i, |1 - t_j|)\}. \end{aligned} \tag{42}$$

Similarly we have

$$|\mathbf{f}(\mathbf{n}_j^{(n)}) - \mathbf{f}(\mathbf{y}(x_0 + nh))| \leqq ht_jLm \tag{43}$$

so that

$$\left| h\sum_{j=1}^{N} b_{ij}\, \mathbf{f}(\mathbf{n}_j^{(n)}) - h\left(\sum_{j=1}^{N} b_{ij}\right) \mathbf{f}(\mathbf{y}(x_0 + nh)) \right| \leqq h^2Lm \sum_{j=1}^{N} |b_{ij}|\, t_j. \tag{44}$$

Combining (42) and (44) we find

$$|\mathbf{e}_i^{(n)}| \leqq h^2Lml_i \tag{45}$$

where l_i is given by

$$l_i = \sum_{j=1}^{N} \{|a_{ij}| \cdot |1 + t_i - t_j| \max(t_i, |1 - t_j|) + |b_{ij}| t_j\}. \tag{46}$$

We write for $\mathbf{l}$ for the vector in R_N whose typical component is l_i.

For the accumulated error we use the symbol $\mathbf{Z}^{(n)} = \mathbf{z}_1^{(n)} \oplus \mathbf{z}_2^{(n)} \oplus \cdots \oplus \mathbf{z}_N^{(n)}$ and define this quantity by $\mathbf{Z}^{(n)} = \mathbf{H}^{(n)} - \mathbf{Y}^{(n)}$. We also write $\mathbf{F}(\mathbf{H}^{(n)}) - \mathbf{F}(\mathbf{Y}^{(n)}) = \mathbf{W}^{(n)} = \mathbf{w}_1^{(n)} \oplus \mathbf{w}_2^{(n)} \oplus \cdots \oplus \mathbf{w}_N^{(n)}$, so that $| \mathbf{W}^{(n)} | \leqq L| \mathbf{Z}^{(n)} |$. Thus we may write

$$\mathbf{Z}^{(n)} - [A]\mathbf{Z}^{(n-1)} - h[B]\mathbf{W}^{(n)} = \mathbf{E}^{(n)} \tag{47}$$

so that

$$\begin{aligned} \mathbf{Z}^{(n)} &= [A^n]\mathbf{Z}^{(0)} + h([B]\mathbf{W}^{(n)} + [AB]\mathbf{W}^{(n-1)} + \cdots + [A^{n-1}B]\mathbf{W}^{(1)}) \\ &\quad + \mathbf{E}^{(n)} + [A]\mathbf{E}^{(n-1)} + \cdots + [A^{n-1}]\mathbf{E}^{(1)}. \end{aligned} \tag{48}$$

We now choose constants α, β, γ such that $| A^n | \leqq \alpha$, $| A^nB | \leqq \beta$, $| A^n1 | \leqq \gamma$ for $n = 0, 1, 2, \cdots$ and use (45) with (48) to find

$$\begin{aligned} | \mathbf{Z}^{(n)} | &\leqq \alpha| \mathbf{Z}^{(0)} | + h\beta(| \mathbf{W}^{(n)} | + | \mathbf{W}^{(n-1)} | + \cdots + | \mathbf{W}^{(1)} |) + nh^2Lm\gamma \\ &\leqq \alpha| \mathbf{Z}^{(0)} | + hL\beta(| \mathbf{Z}^{(n)} | + | \mathbf{Z}^{(n-1)} | + \cdots + | \mathbf{Z}^{(1)} |) + nh^2Lm\gamma. \end{aligned} \tag{49}$$

Hence, it follows that $| \mathbf{Z}^{(n)} | \leqq \epsilon^{(n)}$, where $\epsilon^{(0)} = \alpha| \mathbf{Z}^{(0)} |$ and

$$\epsilon^{(n)} = \epsilon^{(0)} + hL\beta(\epsilon^{(n)} + \epsilon^{(n-1)} + \cdots + \epsilon^{(1)}) + nh^2Lm\gamma, \qquad n \geqq 1. \tag{50}$$

Thus

$$\epsilon^{(n)} - \epsilon^{(n-1)} = hL\beta\epsilon^{(n)} + h^2Lm\gamma, \qquad n \geqq 1, \tag{51}$$

so that

$$\begin{aligned} (\epsilon^{(n)} + hm\gamma/\beta) &= (1 - hL\beta)^{-1}(\epsilon^{(n-1)} + hm\gamma/\beta) \\ &= (1 - hL\beta)^{-n}(\epsilon^{(0)} + hm\gamma/\beta). \end{aligned} \tag{52}$$

If we suppose that $h \leqq h_0$ where h_0, besides satisfying (16) also satisfies $h_0L\beta < 1$, we have

$$(1 - hL\beta)^{-n} \leqq \exp\left(\frac{nhL\beta}{1 - hL\beta}\right) \tag{53}$$

so that, writing $n = \nu$ in (52) and using (53) we find

$$\begin{aligned} | \mathbf{Z}^{(\nu)} | \leqq \epsilon^{(\nu)} \leqq \alpha \, | \mathbf{Z}^{(0)} | \exp\left(\frac{(x - x_0)L\beta}{1 - hL\beta}\right) + \frac{(x - x_0)m\gamma}{\nu\beta} \\ \left\{\exp\left(\frac{(x - x_0)L\beta}{1 - hL\beta}\right) - 1\right\} \end{aligned} \tag{54}$$

and the right hand side tends to zero as $\nu \to \infty$.

Stanford Linear Accelerator Center
Stanford University
Stanford, California

1. G. Dahlquist, "Convergence and stability in the numerical integration of ordinary differential equations," *Math. Scand.*, v. 4, 1956, pp. 33–53. MR **18**, 338.

2. P. Henrici, *Discrete variable methods in ordinary differential equations*, Wiley, New York, 1962. MR **24** #B1772.

3. M. Urabe, "Theory of errors in numerical integration of ordinary differential equations," *J. Sci. Hiroshima Univ. Ser.* A-I, v. 25, 1961, pp. 3–62. MR **25** #759.

4. M. Urabe, H. Yanagiwara & Y. Shinohara, "Periodic solutions of van der Pol's equation with damping coefficient $\lambda = 2 \sim 10$," *J. Sci. Hiroshima Univ. Ser.* A, v. 23, 1960, pp. 325–366. MR **23** #1889.

5. W. B. Gragg & H. J. Stetter, "Generalized multistep predictor-corrector methods," *J. Assoc. Comp. Mach.*, v. 11, 1964, pp. 188–209. MR **28** #4680.

6. C. W. Gear, "Hybrid methods for initial value problems in ordinary differential equations," *J. SIAM Numer. Anal. Ser.* B, v. 2, 1965, pp. 69–86.