MATHEMATICS OF COMPUTATION, VOLUME 28, NUMBER 126, APRIL 1974, PAGES 449–464

The Application of Implicit Runge-Kutta and Collocation Methods to Boundary-Value Problems*

By Richard Weiss

Abstract. The solution of a nonlinear system of first order differential equations with nonlinear boundary conditions by implicit Runge-Kutta methods based on interpolatory quadrature formulae is examined. An equivalence between implicit Runge-Kutta and collocation schemes is established. It is shown that the difference equations obtained have a unique solution in a neighbourhood of an isolated solution of the continuous problem, that this solution can be computed by Newton iteration and that it converges to the isolated solution. The order of convergence is equal to the degree of precision of the related quadrature formula plus one. The efficient implementation of the methods is discussed and numerical examples are given.

1. Introduction. We investigate the application of certain implicit Runge-Kutta methods (cf. Butcher [2]) to the numerical solution of nonlinear boundary-value problems of the form

$$y'(t) - f(t, y(t)) = 0, \qquad a \leqq t \leqq b, \tag{1.1a}$$

$$g(y(a), y(b)) = 0. \tag{1.1b}$$

Here, y, f and g are vector valued functions of dimension N. It is clear that most two point boundary-value problems can be reduced to (1.1a, b).

The schemes will be used to obtain approximations to $y(t)$ on grids π_I;

$$\pi_I = \Big\{ t_0, t_1, \cdots, t_I : a = t_0 < t_1 < \cdots < t_I = b; \quad t_i = t_{i-1} + h_{i-1}, h = \max_i h_i \leqq \lambda \min_i h_i \Big\}, \tag{1.2}$$

where λ, the ratio between the largest and the smallest grid spacing, is uniformly bounded for all families of grids to be considered. High order accuracy on the grid π_I will be obtained by introducing appropriately spaced intermediate points on each interval of π_I. Using interpolatory quadrature based on these intermediate points yields implicit Runge-Kutta methods. Identical schemes are obtained if the intermediate points are used for collocation with piecewise polynomials.

We shall now introduce the Runge-Kutta schemes. Collocation will be discussed in Section 2.

Received February 23, 1973.

AMS (MOS) subject classifications (1970). Primary 65L10.

Key words and phrases. Implicit Runge-Kutta method, collocation method, boundary-value problem.

* This research was supported in part by the Atomic Energy Commission, Grant AT (04-3)-767.

Let

$$0 \leqq u_1 < u_2 < \cdots < u_n = 1 \tag{1.3}$$

be a fixed set of points and define

$$\omega(t) = (t - u_1)(t - u_2) \cdots (t - u_n), \tag{1.4}$$

$$L_k(t) = \omega(t)/((t - u_k)\omega'(u_k)), \qquad k = 1, \cdots, n, \tag{1.5}$$

$$w_{jk} = \int_0^{u_j} L_k(s)\, ds, \qquad k = 1, \cdots, n;\ j = 1, \cdots, n. \tag{1.6}$$

This leads to the set of quadrature rules

$$\int_0^{u_j} \varphi(s)\, ds \approx \sum_{k=1}^{n} w_{jk}\varphi(u_k), \qquad j = 1, \cdots, n. \tag{1.7}$$

Now introduce a subgrid of (1.2), viz.

$$t_{ij} = t_i + u_j h_i, \qquad j = 1, \cdots, n;\ i = 0, \cdots, I - 1, \tag{1.8}$$

and, for $t_{ij} \in (t_i, t_{i+1}]$, rewrite (1.1a) as

$$y(t_{ij}) - y(t_i) - \int_0^{u_j h_i} f(t_i + s, y(t_i + s))\, ds = 0. \tag{1.9}$$

The use of (1.7) to approximate the integral term in (1.9) then leads to the numerical method

$$h_i N_h Y_{ij} \equiv Y_{ij} - Y_{i-1,n} - h_i \sum_{k=1}^{n} w_{jk} f(t_{ik}, Y_{ik}) = 0, \tag{1.10a}$$

$$j = r, \cdots, n;\ i = 0, \cdots, I - 1,$$

$$g(Y_{-1,n}, Y_{I-1,n}) = 0, \tag{1.10b}$$

where

$$r = 1 \quad \text{if } u_1 > 0,$$

$$= 2 \quad \text{if } u_1 = 0,$$

Y_{ij} represents an approximation to $y(t_{ij})$ and $Y_{-1,n}$ is an approximation to $y(a)$. If $u_1 = 0$, then $Y_{i,1} = Y_{i-1,n}$, $i = 1, \cdots, I - 1$. Equations (1.10a, b) represent the desired finite difference scheme for (1.1). We have $(n + 1 - r)I + 1$ relations for the $(n + 1 - r)I + 1$ unknowns $Y_{-1,n}$, Y_{ij}. Only the approximations to $y(t_i)$, $Y_i = Y_{i-1,n}$, $i = 0, \cdots, I$, are of interest. The values Y_{ij}, $j = r, \cdots, n - 1$, are auxiliary quantities.

We also consider the case when u_n in (1.3) is <1. Then (1.10) has to be replaced by

$$\left.\begin{aligned} &Y_{ij} - Y_i - h_i \sum_{k=1}^{n} w_{jk} f(t_{ik}, Y_{ik}) = 0, \quad j = r, \cdots, n \\ &Y_{i+1} - Y_i - h_i \sum_{k=1}^{n} \bar{w}_k f(t_{ik}, Y_{ik}) = 0 \end{aligned}\right\}, \quad i = 0, \cdots, I - 1, \tag{1.11a}$$

$$g(Y_0, Y_I) = 0, \tag{1.11b}$$

where

$$\bar{w}_k = \int_0^1 L_k(s)\, ds, \qquad k = 1, \cdots, n.$$

If $u_1 = 0$, then $Y_{i1} = Y_i$, $i = 0, \cdots, I-1$. Equation (1.11) yields $I(n+2-r)+1$ relations for the $I(n+2-r)+1$ unknowns Y_i, Y_{ij}, $j = r, \cdots, n$; $i = 0, \cdots, I-1$, and Y_I.

The purpose of this paper is to investigate the convergence properties and computational aspects of (1.10) and (1.11). In Section 2, we shall present the alternative derivation of the schemes using collocation. The stability of the methods for linear equations (1.1) will be established in Section 3. The results of Section 3 will be used in Section 4 for the treatment of the nonlinear case where it will be shown that, for sufficiently small h, (1.10) (or (1.11)) has a unique solution in a neighbourhood of an isolated solution of (1.1), that this solution can be computed by Newton iteration and that the finite difference approximations converge to the isolated solution. The order of convergence is at least n. The results of Sections 3 and 4 are derived using the theory of Keller ([7], [8]). In Section 5, we refine the error estimates and show that the order of convergence is, in fact, equal to $p+1$, where p is the degree of precision of the quadrature formula

$$\int_0^1 \varphi(s)\, ds \approx \sum_{k=1}^{n} \varphi(u_k) \int_0^1 L_k(s)\, ds.$$

This implies that convergence of order up to $2n$ can be obtained for suitably chosen points (1.3). We conclude this section by showing that the use of Lobatto points is computationally most efficient. Convergence results similar to those derived in Section 5 have been established by Axelsson [1] for the initial value problem for ordinary differential equations, by de Hoog and Weiss [5] for Volterra integral equations of the second kind and by de Hoog [4] for certain integro-differential equations. Finally, in Section 6, efficient ways of solving the linear systems arising in the implementation of Newton's method are discussed and numerical examples are given.

Although we only treat two point boundary-value problems, the schemes and the analyses can be extended to multipoint boundary conditions as considered in Keller [8, Appendix B]. Also, for linear equations, we can include the case of piecewise continuous coefficients and data if we proceed as in Keller [7].

2. Collocation. The alternative derivation of our schemes via collocation proceeds as follows: Let $L(\pi_I)$ be the family of N-dimensional vector valued functions $v(t) \in C[a, b]$ such that v satisfies (1.1b) and each component of v is a polynomial of degree n on $[t_i, t_{i+1}]$, $i = 0, \cdots, I-1$. We now collocate at $t = t_{ij}$, i.e., we require that $v(t)$ satisfies (1.1a) for $t = t_{ij}$, $j = 1, \cdots, n$; $i = 0, \cdots, I-1$. If $u_1 = 0$, then $v'(t_{i1})$ is to be taken as the right derivative, and similarly, if $u_n = 1$, then $v'(t_{in})$ is the left derivative.

It is convenient to represent v as

$$(2.1) \qquad v(t) = p_i(t) = \sum_{k=0}^{n} c_{ik}(t - t_i)^k, \qquad t_i \leqq t \leqq t_{i+1};\ i = 0, \cdots, I-1.$$

Then the above conditions become

$$p_i'(t_{ij}) - f(t_{ij}, p_i(t_{ij})) = 0, \qquad j = 1, \cdots, n;\ i = 0, \cdots, I-1, \tag{2.2a}$$

$$p_i(t_i) - p_{i-1}(t_i) = 0, \qquad i = 1, \cdots, I-1, \tag{2.2b}$$

$$g(p_0(a), p_{I-1}(b)) = 0. \tag{2.2c}$$

This is a set of $(n+1)I$ relations for the $(n+1)I$ unknowns c_{ik}, $k = 0, \cdots, n$; $i = 0, \cdots, I-1$.

We shall now establish an equivalence between the schemes (2.2) and (1.10).

THEOREM 2.1. *Let $p_i(t)$, $i = 0, \cdots, I-1$, satisfy* (2.1), (2.2) *and define*

$$p_{-1,n} = p_0(a), \qquad p_{ij} = p_i(t_{ij}), \qquad j = 1, \cdots, n;\ i = 0, \cdots, I-1.$$

Then $p_{-1,n}$, p_{ij} satisfy (1.10). *Conversely, let $Y_{-1,n}$, Y_{ij} be a solution of* (1.10) *and denote by $\tilde{p}_i(t)$ the unique polynomial of degree n satisfying*

$$\tilde{p}_i(t_i) = Y_i, \qquad \tilde{p}_i(t_{ij}) = Y_{ij}, \qquad j = 1, \cdots, n,$$

if $u_1 > 0$, or

$$\tilde{p}_i(t_i) = Y_i, \qquad \tilde{p}_i'(t_i) = f(t_i, Y_i), \qquad \tilde{p}_i(t_{ij}) = Y_{ij}, \qquad j = 2, \cdots, n,$$

if $u_1 = 0$. Then $\tilde{p}_i(t)$, $i = 0, \cdots, I-1$, satisfy (2.2).

Proof. Since the $\{w_{jk}\}$ are weights for interpolatory quadrature of degree n we have that

$$h_i \sum_{k=1}^{n} w_{jk} P(t_{ik}) = \int_{t_i}^{t_{ij}} P(s)\, ds$$

for all $P(t)$ which are polynomials of degree $\leqq n-1$ on $[t_i, t_{i+1}]$. From (2.1),

$$p_i(t_{ij}) - p_i(t_i) = \int_{t_i}^{t_{ij}} p_i'(s)\, ds = h_i \sum_{k=1}^{n} w_{jk} p_i'(t_{ik})$$

$$= h_i \sum_{k=1}^{n} w_{jk} f(t_{ik}, p_i(t_{ik})),$$

provided (2.2a) is satisfied. Thus, from (2.2b), the p_{ij} satisfy (1.10).

To establish the converse, note that

$$\tilde{p}_i(t_{ij}) - \tilde{p}_i(t_i) = Y_{ij} - Y_i = h_i \sum_{k=1}^{n} w_{jk} \tilde{p}_i'(t_{ik}).$$

It follows from (1.10) that

$$h_i \sum_{k=1}^{n} w_{jk} \tilde{p}_i'(t_{ik}) = h_i \sum_{k=1}^{n} w_{jk} f(t_{ik}, \tilde{p}_i(t_{ik})). \tag{2.3}$$

Since the matrix $W = (w_{jk})$, $j = r, \cdots, n$; $k = r, \cdots, n$, is nonsingular, it follows that

$$\tilde{p}_i'(t_{ij}) = f(t_{ij}, \tilde{p}_i(t_{ij})), \qquad j = 1, \cdots, n. \qquad \square$$

In the same way, we obtain for the case $u_n < 1$:

THEOREM 2.2. *Let $p_i(t)$, $i = 0, \cdots, I-1$ satisfy* (2.1), (2.2) *and define*

$$p_i = p_i(t_i),\ ''p_{ij} = p_i(t_{ij}),\ ''j = r, \cdots, n;\ i = 0, \cdots, I-1;\ p_I = p_{I-1}(t_I);$$

then p_i, p_{ij} *satisfy* (1.11). *Conversely, let* Y_i, Y_{ij} *be a solution of* (1.11) *and let* $\tilde{p}_i(t)$, $i = 0, \cdots, I-1$, *be defined as in Theorem* 2.1. *Then* $\tilde{p}_i(t)$ *satisfy* (2.2).

The equivalence of implicit Runge-Kutta and collocation schemes has been observed for initial-value problems by Wright [12] and Hulme [6].

Collocation by piecewise polynomials as a tool for solving boundary-value problems has been studied extensively. (For a bibliography, see Russel and Shampine [10].) For $u_1 = 0$ and $u_n = 1$, our collocation procedure coincides with that of Russel and Shampine [10]. For the remaining cases, our procedure is different since then $v(t)$ defined by (2.2) is not an element of $C^1[a, b]$ in general but only of $C[a, b]$. We have shown that for first order systems each such collocation scheme is identical with an appropriate difference scheme.

The theory of Russel and Shampine [10] has recently been extended by deBoor and Swartz [3]. For the case of a scalar equation (1.1a) with linear boundary conditions (1.1b), their results coincide with ours.

Osborne [9] has considered a class of collocation procedures for linear scalar differential equations which include the methods presented here and selected schemes which have a minimal local discretization error. His theory is closely related to the results of Section 5.

3. Stability for Linear Equations. In the remainder of the paper the analysis will be presented only for the case $u_1 > 0$, $u_n = 1$. However, with slight notational modifications, all results extend to the remaining cases.

We now investigate the stability of (1.10) for linear equations (1.1):

$$y'(t) - A(t)y(t) - \alpha(t) = 0, \qquad a \leqq t \leqq b, \tag{3.1a}$$

$$B_a y(a) + B_b y(b) - \beta = 0. \tag{3.1b}$$

THEOREM 3.1. *Let* $A(t) \in C[a, b]$ *and* B_a, B_b *be such that* (3.1) *has a unique solution for all* $\alpha(t) \in C[a, b]$ *and all* β. *Then there exist constants* h_0, C_1 *and* C_2 *such that the difference equations*

$$v_{ij} - v_{i-1,n} - h_i \sum_{k=1}^{n} w_{jk} A(t_{ik}) v_{ik} = h_i \gamma_{ij}, \tag{3.2a}$$

$$j = 1, \cdots, n; \; i = 0, \cdots, I-1,$$

$$B_a v_{-1,n} + B_b v_{I-1,n} = \delta, \tag{3.2b}$$

have a unique solution for $h \leqq h_0$ *and*

$$\max\left\{||v_{-1,n}||, \max_{0 \leqq i \leqq I-1} \max_{1 \leqq j \leqq n} ||v_{ij}||\right\} \leqq C_1 \left(\max_{0 \leqq i \leqq I-1} \max_{1 \leqq j \leqq n} ||\gamma_{ij}|| \right) + C_2 \, ||\delta||.^{**} \tag{3.3}$$

Proof. Introduce the block vectors

$$\left.\begin{aligned} v_i^T &= (v_{i1}^T, \cdots, v_{in}^T) \\ \bar{v}_i^T &= (v_{i-1,n}^T, \cdots, v_{i-1,n}^T) \\ \gamma_i^T &= (\gamma_{i1}^T, \cdots, \gamma_{in}^T) \end{aligned}\right\}, \qquad i = 0, \cdots, I-1, \tag{3.4}$$

** Unless otherwise specified, $||\cdot||$ will denote the maximum norm in R^N or the induced operator norm.

and the appropriate block unit matrix $\bar{J}$ and block matrices $\bar{A}_i$, $i = 0, \cdots, I-1$, so that (3.2a) can be written as

$$(\bar{J} - h_i \bar{A}_i)v_i = \bar{v}_i + h_i\gamma_i, \qquad i = 0, \cdots, I-1. \tag{3.5}$$

Since $A(t) \in C[a, b]$, there exists a constant C_3 such that

$$\max_{0 \leq i \leq I-1} ||\bar{A}_i|| \leqq C_3,$$

where $||\cdot||$ is the operator norm induced by the maximum norm on R^{nN}. Hence, if $h \leqq h_1 < 1/C_3$, it follows from Banach's lemma that

$$v_i = (\bar{J} + h_i \bar{A}_i + h_i^2 \bar{R}_i)(\bar{v}_i + h_i\gamma_i), \qquad i = 0, \cdots, I-1, \tag{3.6}$$

where

$$\max_{0 \leq i \leq I-1} ||\bar{R}_i|| \leqq C_4, \qquad C_4 = \text{const.}$$

The last N equations of (3.6) take the form

$$\begin{aligned} v_{in} = v_{i-1,n} + h_i \sum_{k=1}^{n} w_{nk} A(t_{ik})v_{i-1,n} + h_i^2 R_i v_{i-1,n} \\ + h_i\gamma_{in} + h_i^2 S_i\gamma_i, \qquad i = 0, \cdots, I-1, \end{aligned} \tag{3.7}$$

where the linear operators R_i, S_i are uniformly bounded in i and h. Since, due to consistency, $\sum_{k=1}^{n} w_{nk} = 1$, Eq. (3.7) can be written as

$$\begin{aligned} h_i L_h v_{in} &\equiv v_{in} - v_{i-1,n} - h_i A(t_i + h_i/2)v_{i-1,n} - h_i[Q_i + h_i R_i]v_{i-1,n} \\ &= h_i\gamma_{in} + h_i^2 S_i\gamma_i, \qquad i = 0, \cdots, I-1, \end{aligned} \tag{3.8}$$

where $||Q_i|| \leqq C_5\theta(h/2)$, $C_5 = \text{const}$ and $\theta(h)$ is the modulus of continuity of $A(t)$. Equation (3.8) combined with (3.2b) are $(I+1)$ equations for the $(I+1)$ unknowns v_{in}, $i = -1, \cdots, I-1$. Multiplying the ith equation of (3.8) by $(J - h_i A(t_i + h_i/2)/2)$, where J is the (N, N) unit matrix, leads to a new difference operator

$$\hat{L}_h = \bar{L}_h + \tilde{L}_h \tag{3.9}$$

where $\bar{L}_h$ is the difference operator obtained by applying the centered Euler scheme to (3.1a), viz.

$$h_i \bar{L}_h v_{in} \equiv v_{in} - v_{i-1,n} - \tfrac{1}{2} h_i A(t_i + h_i/2)(v_{in} + v_{i-1,n})$$

and $\tilde{L}_h$ is a linear perturbation satisfying

$$||\tilde{L}_h|| \leqq C_6\theta(h/2), \qquad C_6 = \text{const.} \tag{3.10}$$

Keller [7] has established stability for the centered Euler scheme. Thus, from (3.9) and (3.10), the application of the Banach lemma in the standard way guarantees stability for (3.8), (3.2b), viz. there exist constants h_2, C_7, C_8 such that (3.8), (3.2b) is uniquely solvable for $0 < h \leqq h_2$ and

$$\max_{-1 \leq i \leq I-1} ||v_{in}|| \leqq C_7 \max_{0 \leq i \leq I-1} \max_{1 \leq j \leq n} ||\gamma_{ij}|| + C_8||\delta||. \tag{3.11}$$

Clearly, (3.11) and (3.6) imply (3.3). □

4. Solution of the Nonlinear Difference Equations and Convergence. A solution of (1.1) will be called isolated if the linear system

$$w'(t) - A(t)w(t) = 0, \tag{4.1a}$$

$$B_a w(a) + B_b w(b) = 0 \tag{4.1b}$$

where

$$A(t) = f_y(t, y(t)), \qquad B_a = g_{y(a)}(y(a), y(b)), \qquad B_b = g_{y(b)}(y(a), y(b))$$

has only the trivial solution.

In the sequel, we shall use the notation

$$S_\rho[y(t)] \equiv \{z \mid z \in R^N, \ ||z - y(t)|| \leqq \rho\} \subset R^N \tag{4.2}$$

and

$$S_\rho\{y(t_{ij})\} \equiv \{v_{-1,n}; v_{ij}, j = 1, \cdots, n, i = 0, \cdots, I-1 \mid v_{-1,n} \in S_\rho[y(a)], v_{ij} \in S_\rho[y(t_{ij})]\}. \tag{4.3}$$

The main result of this section is summarized in

THEOREM 4.1. *Let* (1.1) *have an isolated solution* $y(t) \in C^{n+1}[a, b]$ *and let* $f(t, z) \in C^{n+1}\{[a, b] \times S_\rho[y(t)]\}$, $g(v, w) \in C^2\{S_\rho[y(a)] \times S_\rho[y(b)]\}$ *for some* $\rho > 0$. *Then there exist constants* ρ_0 *and* h_0 *such that, for* $0 < h \leqq h_0$,

(i) *Eqs.* (1.10) *have a unique solution* $\{Y_{ij}\} \in S_{\rho_0}\{y(t_{ij})\}$,

(ii) *the solution can be computed by Newton's method which converges quadratically for any initial iterate* $\{Y_{ij}^{(0)}\} \in S_{\rho_1}\{y(t_{ij})\}$ *provided* ρ_1 *and* ρ_1/h *are sufficiently small,*

(iii) $||Y_{-1,n} - y(a)|| \leqq D_1 h^n$, $||Y_{ij} - y(t_{ij})|| \leqq D_1 h^n$, $j = 1, \cdots, n$; $i = 0, \cdots, I-1$, $D_1 = \text{const.}$

Proof. This theorem is the analogue of a result proved by Keller for the centered Euler scheme [8, Main Theorem]. The techniques employed in [8] are not restricted to this scheme but are a general tool for the study of finite difference methods for nonlinear boundary-value problems.

We shall therefore not present a detailed proof of Theorem 4.1, but shall proceed only until the connection with the theory of [8] becomes obvious.

Consider the linear system (3.2) with $A(t)$, B_a and B_b given by (4.1). Introduce the (nN, N) matrices

$$\hat{C}_i = \begin{bmatrix} \hat{J}_i \\ \vdots \\ \hat{J}_i \end{bmatrix}$$

where $\hat{J}_i = J/h_i$, the (nN, nN) matrices

$$\bar{C}_i = \left[\mathbf{O} \vdots \hat{C}_i\right], \qquad \bar{D}_i = h_i^{-1} J - \bar{A}_i, \qquad i = 0, \cdots, I-1,$$

where $\bar{J}$, $\bar{A}_i$ are given by (3.5) and the $(N + nNI)$-dimensional vectors

$$V^T = (v_{-1,n}^T, v_0^T, \cdots, v_{I-1}^T), \qquad \Gamma^T = (\delta^T, \gamma_1^T, \cdots, \gamma_{I-1}^T)$$

with v_i, γ_i defined by (3.4). Then, by (3.5), the system (3.2) can be written as

$$\mathfrak{L} V = \Gamma \tag{4.4}$$

with the $(N + nNI, N + nNI)$ matrix

$$\mathfrak{L} = \begin{bmatrix} B_a & & & & O & B_b \\ -\hat{C}_0 & \bar{D}_0 & & & & \\ & -\bar{C}_1 & \bar{D}_1 & O & & \\ & & \ddots & \ddots & & \\ & O & & -\bar{C}_{I-1} & \bar{D}_{I-1} & \end{bmatrix}.$$

From Theorem 3.1, $\mathfrak{L}$ is nonsingular for $h \leqq h_0$ and

$$||\mathfrak{L}^{-1}|| \leqq \max\{C_1, C_2\} = C_0. \tag{4.5}$$

Here $||\cdot||$ is the operator norm induced by the maximum norm on $R^{(N+nNI)}$.

Now consider (1.10) and write it in vector form

$$\Phi(Y) = 0, \tag{4.6}$$

where

$$Y^T = (Y_{-1,n}^T, Y_{01}^T, \cdots, Y_{0n}^T, \cdots, Y_{I-1,1}^T, \cdots, Y_{I-1,n}^T)$$

and

$$\Phi(Y) = \begin{bmatrix} g(Y_{-1,n}, Y_{I-1,n}) \\ N_h\ Y_{01} \\ \vdots \\ N_h\ Y_{0n} \\ \vdots \\ N_h\ Y_{I-1,1} \\ \vdots \\ N_h\ Y_{I-1,n} \end{bmatrix}. \tag{4.7}$$

With (4.4) and (4.6), we have reduced Eqs. (3.2) and (1.10) to the form used in [8] for the treatment of the centered Euler scheme. Due to (4.5), we may proceed as in [8, Section 3] and parts (i) and (ii) of the theorem follow by slightly changing some of the details of the analysis of the centered Euler scheme.

To establish (iii), note that from the Taylor series expansion the local discretization error $\tau_{ij} = -N_h y(t_{ij})$ satisfies

$$||\tau_{ij}|| \leqq D_2 h^n, \qquad j = 1, \cdots, n;\ i = 0, \cdots, I-1, \qquad D_2 = \text{const.} \tag{4.8}$$

Part (iii) now follows from the arguments of [8, Section 4] with appropriate modifications resulting from (4.8). □

5. High Order Convergence. In this section, estimates for $e_i = Y_i - y(t_i)$, $i = 0, \cdots, I$, will be derived which are sharper than the bounds provided by Theorem 4.1.

We shall say that $\omega(t) \in \mathcal{P}_0$ if $\int_0^1 \omega(s)ds \neq 0$ and that $\omega(t) \in \mathcal{P}_\nu$, $\nu > 0$, if

$$\int_0^1 s^r \omega(s)\, ds = 0, \qquad r = 0, \cdots, \nu - 1,$$

$$\int_0^1 s^\nu \omega(s)\, ds \neq 0.$$

Clearly, $n + \nu - 1$ is the degree of precision of the quadrature formula

$$\int_0^1 \varphi(s)\, ds \approx \sum_{k=1}^{n} w_{nk}\varphi(u_k)$$

if $\omega(t) \in \mathcal{P}_\nu$.

The following lemma which provides an estimate of the local discretization error of (1.10) will be required further on.

LEMMA 5.1. *Let $\omega(t) \in \mathcal{P}_\nu$ and consider the initial-value problem*

$$x'(t) - \psi(t, x(t)) = 0, \qquad 0 \leqq t \leqq h, \; x(0) = x_0, \tag{5.1}$$

where $\psi \in C^{n+\nu+1}\{[0, h] \times R^N\}$ and ψ is uniformly Lipschitz continuous with respect to x for $0 \leqq t \leqq h$ and $||x|| < \infty$. Then the implicit Runge-Kutta scheme

$$X_j - x_0 - h \sum_{k=1}^{n} w_{jk}\psi(u_k h, X_k) = 0, \qquad j = 1, \cdots, n,$$

has a unique solution if h is sufficiently small and $||X_n - x(h)|| \leqq D_3 h^{n+\nu+1}$, $D_3 =$ const.

This lemma is a generalization of a result given in Axelsson [1, Section 3] and can be proved by the technique used there. Alternatively, it may be established by the arguments used in de Hoog and Weiss [5] and Weiss [11] for the treatment of implicit Runge-Kutta methods for Volterra integral equations of the second kind (cf. [5, Theorem 4.1] or [11, Theorem 2.1]).

We shall also require

LEMMA 5.2. *Let*

$$u = hF_h(u) + v, \qquad u, v \in R^M, \tag{5.2}$$

be a family of nonlinear equations depending on the real parameter h, with F_h satisfying

$$F_h \in C^3[R^M], \qquad ||F_h^{(\nu)}(u)|| \leqq L_\nu, \tag{5.3}$$
$$\nu = 0, \cdots, 3, \; u \in R^M, \; 0 \leqq h \leqq \bar{h}, \qquad L_\nu = \text{const.}^{***}$$

Then, for $0 \leqq h \leqq \tilde{h} = \min(\bar{h}, 1/2L_1)$, (5.2) has a unique inverse $u = u_h(v) \in C^3[R^M]$ which can be represented as

$$u_h(v) = v + hF_h(v) + h^2 R_h(v) \tag{5.4}$$

where

*** Here $||\cdot||$ is the maximum norm on R^M or the induced operator norm.

$$||R_h'(v)|| \leqq L_4, \qquad ||R_h'(v_1) - R_h'(v_2)|| \leqq L_5||v_1 - v_2||, \tag{5.5}$$
$$v, v_1, v_2 \in R^M, \qquad L_4, L_5 = \text{const.}$$

Proof. The existence, uniqueness and differentiability of u_h follow from the contraction mapping principle and the implicit function theorem.

Using v as a starting iterate for the functional iteration, we have

$$u_h(v) = v + hr_h(v) \tag{5.6}$$

where

$$r_h(v) = \lim_{m\to\infty} F_h(v + hF_h(v + (\cdots)\overbrace{\cdots}^{m})).$$

Clearly,

$$||r_h(v)||, ||r_h'(v)|| \leqq L_6, \qquad v \in R^M, \qquad L_6 = \text{const.} \tag{5.7}$$

Equating (5.2) and (5.4) leads to

$$R_h'(v) = [F_h'(u_h(v))u_h'(v) - F_h'(v)]/h.$$

From (5.2),

$$u_h'(v) = hF_h'(u_h(v))u_h'(v) + J,$$

where J is the (M, M) unit matrix. Hence,

$$u_h'(v) = (J - hF_h'(u_h(v)))^{-1},$$

and, using (5.6),

$$\begin{aligned} R_h'(v) &= \left[F_h'(u_h(v)) - F_h'(v) + h(F_h'(u_h(v)))^2 \sum_{m=0}^{\infty} (hF_h'(u_h(v)))^m\right] \Big/ h \\ &= \int_0^1 F_h''(v + shr_h(v))\, ds\, r_h(v) + F_h'(u_h(v))^2 \sum_{m=0}^{\infty} (hF_h'(u_h(v)))^m. \end{aligned} \tag{5.8}$$

Using (5.8) together with (5.3), (5.6) and (5.7), it is now straightforward to establish (5.5). □

The main result of this section is

THEOREM 5.1. *Let $\omega(t) \in \mathcal{P}_\nu$ for $\nu \leqq n$ and $f(t, z) \in C^{n+\nu+1}\{[a, b] \times S_{\rho_0}[y(t)]\}$. Then*

$$||Y_i - y(t_i)|| \leqq D_3 h^{n+\nu}, \qquad i = 0, \cdots, I, \qquad D_3 = \text{const.}$$

Proof. Let the function $\mu(t, z) \in C^{n+\nu+1}\{[a, b] \times R^N\}$ satisfy $\mu(t, z) \equiv 1$, $(t, z) \in [a, b] \times S_{\rho_0}[y(t)]$ and $\mu(t, z) \equiv 0$, $(t, z) \notin [a, b] \times S_{\tilde{\rho}}$, $\tilde{\rho} > \rho_0$. Denote

$$\tilde{f}(t, z) = f(t, z)\mu(t, z).$$

If $f(t, y)$ in (1.1) is replaced by $\tilde{f}(t, y)$, then (1.1) has not been changed in $[a, b] \times S_{\rho_0}[y(t)]$. Also, the approximations Y_{ij} defined by (1.10) and Theorem 4.1 satisfy

$$Y_{ij} - Y_{i-1,n} - h_i \sum_{k=1}^{n} w_{jk}\tilde{f}(t_{ik}, Y_{ik}) = 0, \tag{5.9a}$$
$$j = 1, \cdots, n; \; i = 0, \cdots, I - 1,$$

$$g(Y_{-1,n},\ Y_{I-1,n}) = 0. \tag{5.9b}$$

The function $\tilde{f}(t, z)$ has compact support. Hence, from Lemma 5.2, there exists a constant h_3 such that, for $0 < h \leqq h_3$, (5.9a) can be written as

$$Y_{ij} - Y_{i-1,n} - h_i \sum_{k=1}^{n} w_{jk}\tilde{f}(t_{ik},\ Y_{i-1,n}) - h_i^2 G_h^{ij}(Y_{i-1,n}) = 0, \tag{5.10}$$

$$j = 1, \cdots, n;\ i = 0, \cdots, I-1.$$

In particular, for $j = n$,

$$Y_{i+1} - Y_i - h_i \sum_{k=1}^{n} w_{nk}\tilde{f}(t_{ik},\ Y_i) - h_i^2 G_h^{in}(Y_i) = 0, \tag{5.11}$$

$$i = 0, \cdots, I-1.$$

From Lemma 5.1,

$$y(t_{i+1}) - y(t_i) - h_i \sum_{k=1}^{n} w_{nk}\tilde{f}(t_{ik},\ y(t_i)) - h_i^2 G_h^{in}(y(t_i)) = \tau_i, \tag{5.12}$$

$$i = 0, \cdots, I-1,$$

where

$$||\tau_i|| \leqq D_4 h^{n+\nu+1}, \qquad i = 0, \cdots, I-1, \qquad D_4 = \text{const.}$$

Subtracting (5.12) from (5.11), applying Taylor's theorem and using Theorem 4.1(iii) and Lemma 5.2, we obtain

$$h_i L_h^* e_i \equiv e_{i+1} - e_i - h_i \sum_{k=1}^{n} w_{nk} A(t_{ik}) e_i - h_i^2 Q_i e_i = \sigma_i, \tag{5.13}$$

$$i = 0, \cdots, I-1,$$

where $A(t)$ is defined by (4.1), Q_i is a linear operator with

$$||Q_i|| \leqq D_5, \qquad i = 0, \cdots, I-1, \qquad D_5 = \text{const} \tag{5.14}$$

and

$$||\sigma_i|| \leqq D_6 h^{n+\nu+1}, \qquad i = 0, \cdots, I-1, \qquad D_6 = \text{const.}$$

Also, (1.1b), (5.9b) and the application of Taylor's theorem yield

$$B_a e_0 + B_b e_I = \eta$$

where B_a, B_b are given by (4.1b) and $||\eta|| \leqq D_7 h^{2n}$, $D_7 = \text{const}$. It follows from (5.14) by an argument similar to that used in the proof of Theorem 3.1 that the difference equations

$$L_h^* e_i = \alpha_i, \qquad i = 0, \cdots, I-1,$$

$$B_a e_0 + B_b e_I = \delta$$

are stable. This completes the proof. □

COROLLARY 5.1. *If $\nu > 0$ and $f(t, z)$ is as in Theorem 5.1, then*

$$||Y_{ij} - y(t_{ij})|| \leqq D_8 h^{n+1},\ j = 1, \cdots, n-1;\ i = 0, \cdots, I-1,\ D_8 = \text{const.}$$

Proof. Subtracting (1.9) from (5.9a), applying Taylor's theorem and using Theorem 4.1, we obtain

$$e_{ij} - h_i \sum_{j=1}^{n} w_{jk} A(t_{ik}, y(t_{ik}))e_{ik} = e_i + \kappa_{ij}, \tag{5.15}$$

$$j = 1, \cdots, n;\ i = 0, \cdots, I-1,$$

where $e_{ij} = Y_{ij} - y(t_{ij})$ and

$$||\kappa_{ij}|| \leqq D_9 h^{n+1}, \qquad j = 1, \cdots, n;\ i = 0, \cdots, I-1, \qquad D_9 = \text{const.}$$

Writing (5.15) in matrix vector notation similar to (3.5) and repeating the arguments following (3.5), we obtain

$$||e_{ij}|| \leqq D_{10}\Big(||e_i|| + \max_{1 \leqq j \leqq n} ||\kappa_{ij}||\Big),$$

$$j = 1, \cdots, n;\ i = 0, \cdots, I-1, \qquad D_{10} = \text{const.} \quad \square$$

For fixed n, it is desirable to choose $\{u_1, \cdots, u_n\}$ so that the order of convergence is as high as possible. For the cases ($u_1 = 0$, $u_n = 1$), ($u_1 > 0$, $u_n = 1$ or $u_1 = 0$, $u_n < 1$) and ($u_1 > 0$, $u_n < 1$) this leads to the Lobatto, Radau and Gauss points, respectively. The orders of convergence are $2n - 2$, $2n - 1$ and $2n$. When using Lobatto points, we have to solve a system of order $N(\mu I + 1)$ to obtain order 2μ convergence. For Radau points, a system of order $N(\mu I + 1)$ yields convergence of order $2\mu - 1$ and for Gauss points, a system of the same size yields convergence of order $2\mu - 2$. *Hence, Lobatto points are more efficient than Radau points, which again are more efficient than Gauss points.*

The Lobatto points for $n = 2, 3, 4$, are given below:

$n = 2$: $u_1 = 0, u_2 = 1$ (Trapezoidal rule),

$n = 3$: $u_1 = 0, u_2 = \frac{1}{2}, u_3 = 1$ (Simpson's rule),

$n = 4$: $u_1 = 0, u_2 = \frac{1}{2}(1 - 1/\sqrt{5}), u_3 = \frac{1}{2}(1 + 1/\sqrt{5}), u_4 = 1.$

6. Computational Aspects and Numerical Examples. In the case when (1.1) has separated endpoint boundary conditions, viz.

$$g(v, w) = \begin{bmatrix} g_1(v) \\ g_2(w) \end{bmatrix} \tag{6.1}$$

where $g_1(v)$ is a p-vector and $g_2(w)$ is a $q = (N - p)$-vector, it is advisable not to apply Newton's method to (4.7), but to rewrite (4.7) in the form

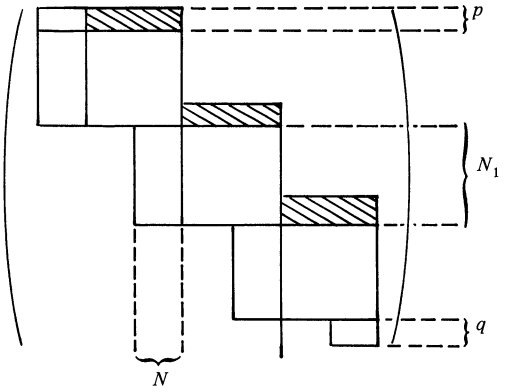
$$\Phi(Y) = \begin{bmatrix} g_1(Y_{-1,n}) \\ N_h Y_{01} \\ \vdots \\ N_h Y_{I-1,n} \\ g_2(Y_{I-1,n}) \end{bmatrix}.$$

Then the matrices $\mathfrak{A}^\nu$ in the implementation of Newton's method

$$\mathfrak{A}^{\nu}[Y^{\nu+1} - Y^{\nu}] = -\Phi(Y^{\nu}) \tag{6.2}$$

have a certain block-band structure. For $I = 3$, this structure is exhibited in the schematic representation

(6.3) $\mathscr{A}^{\nu} \equiv$



Here $N_1 = N(n + 1 - r)$ if $u_n = 1$ and $N_1 = N(n + 2 - r)$ otherwise. Only the white fields within the dark lines contain nonzero elements.

Equation (6.2) can be solved by the following procedures:

#1: *Gaussian elimination with partial pivoting.* If the elimination is performed with consideration of the zero-pattern in $\mathfrak{A}^{\nu}$, then only the shaded fields in (6.3) are filled and the amount of additional storage required is modest.

#2: *A "mixed" pivoting strategy with column interchanges while eliminating* Y_i, $i = 0, \cdots, I$, *and row interchanges during the elimination of the other unknowns.* Here the zero-pattern of $\mathfrak{A}^{\nu}$ is preserved. This procedure is slightly simpler to implement than #1.

The leading terms in the operational counts are

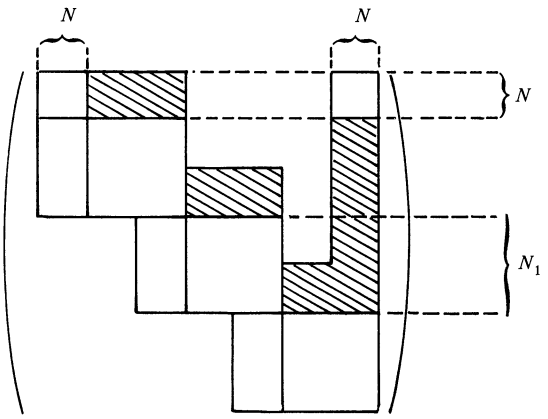
#1: $I\{N_1[N_1^2/3 + N_1(N + p)/2 + Np]\}$.

#2: $I\{p[p^2/3 + p(N_1 + q)/2 + N_1 q]$
$+ (N_1 - p)[(N_1 - p)^2/3 + (N_1 - p)(N + p)/2 + Np]\}$.

A simple calculation shows that #2 is faster than #1.

When the boundary conditions are not of the form (6.1), then the matrices $\mathfrak{A}^{\nu}$ in Newton's method for (4.7) have the following structure for $I = 3$:

(6.4) $\mathscr{A}^{\nu} \equiv$



Again, $\mathcal{A}^{\nu}$ can be inverted efficiently by Gaussian elimination with partial pivoting. The amount of fill-in introduced is indicated by the shaded areas in (6.4). The leading term in the operational count is $I\{N_1[N_1^2/3 + 3N_1N/2 + 2N^2]\}$.

The scheme (1.10) with $n = 4$, Lobatto points (accuracy h^6) and algorithm #2 has been used to solve the following problems:

(6.5)(i) $$u''(t) = \exp(u(t)), \qquad u(0) = u(1) = 0.$$

The unique solution is

$$u(t) = 2 \ln\left\{c \sec\left(\frac{c(t - \frac{1}{2})}{2}\right)\right\} - \ln 2,$$

where $c \doteq 1.33605569490611$. As a first order system, (6.5) takes the form

$$y_1' - y_2 = 0,$$
$$y_2' - \exp(y_1) = 0,$$
$$y_1(0) = y_1(1) = 0.$$

A uniform grid was used. The starting values for the Newton iteration were obtained from

$$y_1^{(0)} = (t - 0.5)^2 - 0.25, \qquad y_2^{(0)} = 2t - 1.$$

The errors in u and u' for different gridspacings are shown in Table 1. Since u is symmetric about $t = \frac{1}{2}$, the values are only given for $t = 0$ and $t = \frac{1}{3}$. The iteration process was terminated when the norm of the difference between two successive iterates was $\leqq 10^{-14}$. For all h in Table 1 this was achieved in four iterations.

TABLE 1

	u	u'	
h	$t = \frac{1}{3}$	$t = 0$	$t = \frac{1}{3}$
1/3	2.66 E-9	−3.66 E-8	−9.06 E-9
1/6	5.07 E-11	−5.96 E-10	−1.47 E-10
1/12	8.30 E-13	−9.42 E-12	−2.32 E-12

(6.6)(ii) $$u''(t) + tu'(t) - u(t) = te^t - |t|(6 - 12t + 2t^2 - 3t^3),$$
$$u(-1) = e^{-1} - 2, \qquad u(1) = e.$$

The unique solution to (6.6) is

$$u(t) = e^t - (t^3 - t^4), \qquad t \geqq 0,$$
$$= e^t + (t^3 - t^4), \qquad 0 \leqq t.$$

Equation (6.6) was transformed to a first order system in the same way as (6.5). We chose this example to demonstrate that, for linear equations, the results of Section 5 are not affected by jump discontinuities in $A(t)$ and $g(t)$ or their derivatives, if the points of discontinuity are contained in π_I. We use uniform nets such that $t = 0$

is a gridpoint. The largest errors in u and u' occur at $t = 0$ and $t = 1$ respectively. The errors in u and u' are given in Tables 2 and 3.

Russel and Shampine [10] used collocation in the smooth Hermite space $H^{(3)}(\pi_I)$ to solve (6.6). The maximum of the absolute values of the errors obtained by this procedure is given in the last columns of Tables 2 and 3. It must be noted, however, that the operational count for our method in this example is about twice that required for the collocation procedure. So it is appropriate to compare our values for $2h$ with those of Russel and Shampine for h.

TABLE 2

	u			
h	$t = -\frac{1}{2}$	$t = 0$	$t = \frac{1}{2}$	$H^{(3)}$
1/2	−6.59 E-8	−9.81 E-8	−7.67 E-8	
1/4	−1.01 E-9	−1.50 E-9	−1.16 E-9	3.40 E-6
1/8	−1.57 E-11	−2.32 E-11	−1.80 E-11	2.17 E-7
1/16	−2.46 E-13	−3.62 E-13	−2.79 E-13	1.36 E-8

TABLE 3

		u'				
h	$t = -1$	$t = -\frac{1}{2}$	$t = 0$	$t = \frac{1}{2}$	$t = 1$	$H^{(3)}$
1/2	−2.88 E-7	−2.70 E-7	−1.80 E-7	3.24 E-9	3.13 E-7	
1/4	−4.45 E-9	−4.13 E-9	−2.67 E-9	2.76 E-10	5.34 E-9	1.00 E-5
1/8	−6.93 E-11	−6.42 E-11	−4.12 E-11	5.26 E-12	8.54 E-11	6.18 E-7
1/16	−1.04 E-12	−9.97 E-13	−6.67 E-13	8.97 E-14	1.34 E-12	3.85 E-8

All computations were done in double precision on the IBM 370/155 at the California Institute of Technology.

Acknowledgement. I wish to thank Professor H. B. Keller for many stimulating discussions during the work on this paper and for his valuable criticism of the first draft.

Applied Mathematics Department
California Institute of Technology
Pasadena, California 91109

1. O. AXELSSON, "A class of A-stable methods," *Nordisk Tidskr. Informationsbehandling (BIT)*, v. 9, 1969, pp. 185–199. MR **40** #8266.
2. J. C. BUTCHER, "Implicit Runge-Kutta processes," *Math. Comp.*, v. 18, 1964, pp. 50–64.
3. C. DE BOOR & B. SWARTZ, "Collocation at Gaussian points," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 582–606.
4. F. DE HOOG, "Implicit Runge-Kutta methods for Volterra integro-differential equations," *Nordisk Tidskr. Informationsbehandling (BIT)*. (To appear.)
5. F. DE HOOG & R. WEISS, "Implicit Runge-Kutta methods for second kind Volterra integral equations," *Numer. Math.* (To appear).
6. B. L. HULME, "Galerkin and related one-step methods for ordinary differential equations," *Math. Comp.*, v. 26, 1972, pp. 881–891.

7. H. B. KELLER, "Accurate difference methods for linear ordinary differential systems subject to linear constraints," *SIAM J. Numer. Anal.*, v. 6, 1969, pp. 8–30. MR **40** #6776.
8. H. B. KELLER, "Accurate difference methods for nonlinear two point boundary value problems," *SIAM J. Numer. Anal.* (To appear.)
9. M. R. OSBORNE, "Minimizing truncation error in finite difference approximations to ordinary differential equations," *Math. Comp.*, v. 21, 1967, pp. 133–145. MR **36** #6156.
10. R. D. RUSSEL & L. F. SHAMPINE, "A collocation method for boundary value problems," *Numer. Math.*, v. 19, 1972, pp. 1–28.
11. R. WEISS, *Numerical Procedures for Volterra Integral Equations*, Thesis, Computer Centre, The Australian National University, Canberra, 1972.
12. K. WRIGHT, "Some relationships between implicit Runge-Kutta, collocation and Lanczos τ methods, and their stability properties," *Nordisk Tidskr. Informationsbehandling (BIT)*, v. 10, 1970, pp. 217–227. MR **42** #1345.