BIT 23 (1983), 84–91

ON THE EXISTENCE OF SOLUTIONS TO THE ALGEBRAIC EQUATIONS IN IMPLICIT RUNGE-KUTTA METHODS

M. CROUZEIX*, W. H. HUNDSDORFER** and M. N. SPIJKER**

* *U.E.R. Mathématiques et Informatique, Université de Rennes, Campus Rennes-Beaulieu, 35042 Rennes CEDEX, France*

** *Institute of Applied Mathematics and Computer Science, University of Leiden, Wassenaarseweg 80, Leiden, The Netherlands*

Abstract.

This paper deals with the systems of algebraic equations arising in the application of B-stable Runge-Kutta methods. It is shown that under natural assumptions such systems do not always have a solution. In addition, general sufficient conditions are presented under which such systems do have unique solutions.

1. Introduction.

We shall deal with the initial-value problem

$$(1.1)\qquad \frac{d}{dt}U(t) = f(t, U(t)) \qquad (t \geqq 0), \qquad U(0) = u_0,$$

where u_0 is a given vector in the s-dimensional complex vector space $\mathbb{C}^s$ and $f\colon \mathbb{R} \times \mathbb{C}^s \to \mathbb{C}^s$ is a given continuous function such that

$$(1.2)\qquad \operatorname{Re}\langle f(t, \tilde{\xi}) - f(t, \xi), \tilde{\xi} - \xi\rangle \leqq 0 \qquad (\text{for all } t \in \mathbb{R} \text{ and } \tilde{\xi}, \xi \in \mathbb{C}^s).$$

Here $\langle \cdot, \cdot \rangle$ stands for an arbitrary inner-product on $\mathbb{C}^s$. The corresponding norm will be denoted by $|\cdot|$.

We consider Runge-Kutta methods for the numerical solution of (1.1), written in the form

$$(1.3a)\qquad u_n = u_{n-1} + h \sum_{j=1}^{m} b_j f(t_{n-1} + c_j h, y_j),$$

$$(1.3b)\qquad y_j = u_{n-1} + h \sum_{k=1}^{m} a_{jk} f(t_{n-1} + c_k h, y_k) \qquad (1 \leqq j \leqq m),$$

Received November 2, 1981. Revised July 7, 1982.

where b_j, a_{jk} are real parameters defining the method, $c_j = a_{j1} + a_{j2} + \ldots + a_{jm}$, $n \geqq 1$, $h > 0$, $t_n = nh$ and $u_n \simeq U(t_n)$.

We define the $m \times m$ matrix Q by

$$Q = BA + A^T B - BEB$$

where $A = (a_{jk})$, $B = \operatorname{diag}(b_1, b_2, \ldots, b_m)$ and E is the $m \times m$ matrix all of whose entries equal 1. In [1, 2] the condition

$$(1.4) \qquad b_j > 0 \quad (1 \leqq j \leqq m),\ Q \text{ is positive semi-definite}$$

was proved to imply that $|\tilde{u}_n - u_n| \leqq |\tilde{u}_{n-1} - u_{n-1}|$ $(n \geqq 1)$ for any two sequences $\{\tilde{u}_n\}$, $\{u_n\}$ computed by the Runge-Kutta method (this property was called B-stability or BN-stability).

For some time it has been conjectured that the assumptions (1.2), (1.4) also imply that the system of algebraic equations (1.3b) has a unique solution $y = (y_1, y_2, \ldots, y_m)^T \in \mathbb{C}^{sm}$. In section 2 we shall show that this conjecture is false.

In section 3 we shall prove a theorem stating that (1.3b) has a unique solution under a slightly modified version of condition (1.4), namely the condition that

(1.5) there exists a diagonal matrix D such that D and $DA + A^T D$ are positive definite

In addition, we prove in section 3 a theorem implying that (1.3b) has a unique solution under assumption (1.4) provided f satisfies a condition which is slightly stronger than (1.2). Uniqueness will be proved under the assumption that

$$(1.6) \qquad \operatorname{Re}\langle f(t, \tilde{\xi}) - f(t, \xi), \tilde{\xi} - \xi \rangle < 0 \qquad (\text{for all } t \in \mathbb{R} \text{ and } \tilde{\xi} \neq \xi \in \mathbb{C}^s),$$

and existence under the assumption

$$(1.7) \qquad \lim_{|\xi| \to \infty} \operatorname{Re}\langle f(t, \xi + \eta), \xi \rangle = -\infty \qquad (\text{for all } t \in \mathbb{R} \text{ and } \eta \in \mathbb{C}^s).$$

2. Construction of a counterexample.

Let d_1, d_2, d_3 be column vectors in $\mathbb{R}^3$ with $d_j^T d_k = 0$ $(1 \leqq j < k \leqq 3)$, $d_j^T d_j = 1$ $(1 \leqq j \leqq 3)$, $d_3 = (1/\sqrt{3})(1, 1, 1)^T$, and let $\sigma \neq 0$, $\rho \geqq \frac{3}{2}$ be given real numbers. We define the real 3×3 matrix $S = (s_{jk})$ by

$$Sd_1 = \sigma d_2, \qquad Sd_2 = -\sigma d_1, \quad Sd_3 = \rho d_3.$$

We note that

$$(2.1a) \qquad v^T S v = \rho [v^T d_3]^2 \qquad (\text{for all } v \in \mathbb{R}^3),$$

(2.1b) $$I - i(\sigma^{-1})S \text{ is singular,}$$

(2.1c) $$S \text{ is regular,}$$

(2.1d) there exist numbers b_j $(1 \leqq j \leqq 3)$ satisfying $b_j > 0$ $(1 \leqq j \leqq 3)$,

$$\sum_{j=1}^{3} b_j = 1 \text{ and } \sum_{n=1}^{3} s_{jn} b_n \neq \sum_{n=1}^{3} s_{kn} b_n \qquad (1 \leqq j < k \leqq 3).$$

The statements (2.1a)–(2.1c) follow from an easy calculation. Using (2.1c) it can be seen that (2.1d) holds; we can take $b_j = (1+t+t^2)^{-1} t^{j-1}$ $(1 \leqq j \leqq 3)$ where t is such that $t > 0$ and $\sum_{n=1}^{3} (s_{jn} - s_{kn}) t^{n-1} \neq 0$ $(1 \leqq j < k \leqq 3)$. Let $B = \operatorname{diag}(b_1, b_2, b_3)$ where the parameters b_j are chosen according to (2.1d). We define the matrix $A = (a_{jk})$ by

(2.2) $$A = SB.$$

From (2.1a), (2.1d), (2.2) it follows that the Runge-Kutta method (1.3) (with $m = 3$ and a_{jk}, b_j as indicated) satisfies (1.4), and

(2.3) $$c_j \neq c_k \qquad (1 \leqq j < k \leqq 3).$$

In view of (2.1b), (2.1c) and (2.2), there exists a vector $z = (z_1, z_2, z_3)^T \in \mathbb{C}^3$ such that the equation

(2.4) $$(I - iA(\sigma B)^{-1})y = Az$$

has no solution $y \in \mathbb{C}^3$.

We choose $s = 1$, $n = 1$, $u_0 = 0$, $h = 1$, and

$$f(t, \xi) = i g_0(t)\xi + g_1(t) \qquad (\text{for } t \in \mathbb{R},\ \xi \in \mathbb{C}).$$

Here $g_0 : \mathbb{R} \to \mathbb{R}$ and $g_1 : \mathbb{R} \to \mathbb{C}$ are continuous functions satisfying

$$g_0(c_j) = (\sigma b_j)^{-1}, \qquad g_1(c_j) = z_j \qquad (1 \leqq j \leqq 3)$$

(note that such g_0, g_1 exist by (2.3)).

With the definitions above, (1.3b) reduces to (2.4). Consequently the equation (1.3b) has no solution $y = (y_1, y_2, y_3)^T \in \mathbb{C}^3$, whereas (1.2) and (1.4) are fulfilled.

Example. By choosing

$$d_1 = \frac{1}{\sqrt{2}}(1, -1, 0)^T, \qquad d_2 = \frac{1}{\sqrt{6}}(1, 1, -2)^T, \qquad \rho = \tfrac{3}{2}, \qquad \sigma = 2\sqrt{2},$$

$$b_1 = b_2 = \tfrac{1}{4}, \qquad b_3 = \tfrac{1}{2}$$

we obtain by this construction a fourth order Runge-Kutta method with

$$A = \begin{bmatrix} \frac{1}{8} & \frac{1}{8}-\frac{1}{6}\sqrt{6} & \frac{1}{4}+\frac{1}{3}\sqrt{6} \\ \frac{1}{8}+\frac{1}{6}\sqrt{6} & \frac{1}{8} & \frac{1}{4}-\frac{1}{3}\sqrt{6} \\ \frac{1}{8}-\frac{1}{6}\sqrt{6} & \frac{1}{8}+\frac{1}{6}\sqrt{6} & \frac{1}{4} \end{bmatrix}.$$

This method thus fulfils condition (1.4) whereas for some continuous function f satisfying (1.2) the corresponding system of equations (1.3b) has no solution.

3. Sufficient conditions for existence and uniqueness of solutions.

THEOREM 1. *Let $f: \mathbb{R} \times \mathbb{C}^s \to \mathbb{C}^s$ be continuous and satisfy* (1.2). *Let D be a diagonal matrix such that D and $DA + A^T D$ are positive definite. Then the system* (1.3b) *has a unique solution $y = (y_1, y_2, \ldots, y_m)^T \in \mathbb{C}^{sm}$.*

PROOF. Let $n \geqq 1$, $h > 0$ and $u_{n-1} \in \mathbb{C}^s$ be given. In the subsequent we shall deal with the inner product $[\cdot, \cdot]$ and norm $\|\cdot\|$ on $\mathbb{C}^{sm}$ defined (as in [5, pp. 12–13]) by

$$[x, y] = \sum_{j=1}^{m} d_j \langle x_j, y_j \rangle, \qquad \|x\| = [x, x]^{\frac{1}{2}} \text{ (for } x, y \in \mathbb{C}^{sm}),$$

where $D = \operatorname{diag}(d_1, d_2, \ldots, d_m)$, $d_j > 0$, $x = (x_1, x_2, \ldots, x_m)^T$ and $y = (y_1, y_2, \ldots, y_m)^T$.

We define $\boldsymbol{A} = A \otimes I_s$ where I_s is the $s \times s$ identity matrix and $\otimes$ stands for the Kronecker product. Further we define the function $\boldsymbol{F}: \mathbb{C}^{sm} \to \mathbb{C}^{sm}$ by

$$\begin{aligned} \boldsymbol{F}(x) = h(f(t_{n-1}+c_1 h, x_1+u_{n-1}), f(t_{n-1}+c_2 h, x_2+u_{n-1}), \ldots, \\ f(t_{n-1}+c_m h, x_m+u_{n-1}))^T \end{aligned}$$

$$\text{(for } x = (x_1, x_2, \ldots, x_m)^T \in \mathbb{C}^{sm}).$$

Writing $y_j = x_j + u_{n-1}$ $(1 \leqq j \leqq m)$, we transform the system (1.3b) into the equivalent equation

$$x - \boldsymbol{A}\boldsymbol{F}(x) = 0. \tag{3.1}$$

Uniqueness. From lemma (2.2) in [4], it is found that

$$\operatorname{Re}[\boldsymbol{A}w, w] > 0 \qquad \text{(for all } w \in \mathbb{C}^{sm} \text{ with } w \neq 0).$$

This implies $\boldsymbol{A}$ is regular and there exists a constant $\beta > 0$ such that

$$\operatorname{Re}[\boldsymbol{A}^{-1}w, w] \geqq \beta\|w\|^2 \qquad \text{(for all } w \in \mathbb{C}^{sm}). \tag{3.2}$$

Assuming

$$0 = \tilde{x} - AF(\tilde{x}) = x - AF(x),$$

we obtain (from (3.2), (1.2))

$$\beta\|\tilde{x}-x\|^2 \leqq \mathrm{Re}\,[A^{-1}(\tilde{x}-x), \tilde{x}-x] = \mathrm{Re}\,[F(\tilde{x})-F(x), \tilde{x}-x] \leqq 0.$$

Hence $\tilde{x} = x$.

Existence. We define

$$G_0(x) = A^{-1}x - F(x) \qquad (\text{for } x \in \mathbb{C}^{sm}).$$

We have from (3.2)

$$\mathrm{Re}\,[G_0(x)-G_0(0), x] \geqq \beta\|x\|^2 \qquad (\text{for all } x \in \mathbb{C}^{sm}),$$

and therefore $\mathrm{Re}\,[G_0(x), x] \geqq \|x\|(\beta\|x\| - \|G_0(0)\|)$. This implies $\mathrm{Re}\,[G_0(x), x] \geqq 0$ for all x with $\|x\| \geqq \|G_0(0)\|/\beta$. By a classical result (see e.g. [8, p. 163] or [7, p. 74]) it follows that there exists an

$$x \in \mathbb{C}^{sm} \text{ with } \|x\| \leqq \|G_0(0)\|/\beta \text{ such that } G_0(x) = 0.$$

Clearly x is a solution to (3.1). ■

We now give a theorem with a weaker requirement on A and a slightly stronger requirement on f.

THEOREM 2. *Let D be a positive definite diagonal matrix such that $DA + A^T D$ is positive semi-definite. Let $f\colon \mathbb{R} \times \mathbb{C}^s \to \mathbb{C}^s$ be continuous. Then the condition*

$$\mathrm{Re}\,\langle f(t,\tilde{\xi}) - f(t,\xi), \tilde{\xi}-\xi\rangle < 0 \qquad (\textit{for all } t \in \mathbb{R} \textit{ and } \tilde{\xi} \neq \xi \in \mathbb{C}^s)$$

implies that the system (1.3b) *has at most one solution, and*

$$\lim_{|\xi|\to\infty} \mathrm{Re}\,\langle f(t, \xi+\eta), \xi\rangle = -\infty \qquad (\textit{for all } t \in \mathbb{R}, \eta \in \mathbb{C}^s)$$

implies the existence of a solution to (1.3b).

PROOF.
Uniqueness. With the same notations as in the proof of theorem 1, we have

(3.3) $$\operatorname{Re}[Aw, w] \geqq 0 \qquad (\text{for all } w \in \mathbb{C}^{sm}),$$

and (in view of (1.6))

$$\operatorname{Re}[F(\tilde{x}) - F(x), \tilde{x} - x] < 0 \qquad (\text{for all } \tilde{x}, x \in \mathbb{C}^{sm} \text{ with } \tilde{x} \neq x).$$

Assuming that $\tilde{x}$, x are two different solutions to (3.1) we obtain

$$0 = \tilde{x} - AF(\tilde{x}) = x - AF(x),$$

and

$$0 \leqq \operatorname{Re}[A(F(\tilde{x}) - F(x)), F(\tilde{x}) - F(x)] < 0.$$

We thus have a contradiction.

Existence. We define for $r > 0$

$$\varphi(r) = \max\{\operatorname{Re}[F(x), x] : x \in \mathbb{C}^{sm} \text{ and } \|x\| = r\}.$$

Using (1.7) it can be proved that $\lim_{r\to\infty} \varphi(r) = -\infty$.

For $\tau > 0$ we define

$$G_\tau(x) = (A + \tau I)^{-1}x - F(x) \qquad (\text{for } x \in \mathbb{C}^{sm})$$

where I stands for the $sm \times sm$ identity matrix. We note that, in view of (3.3), $(A + \tau I)$ is regular. It follows that

$$\operatorname{Re}[G_\tau(x), x] \geqq -\operatorname{Re}[F(x), x] \geqq -\varphi(\|x\|).$$

Choosing $r > 0$ so large that $\varphi(r) \leqq 0$, we have

$$\operatorname{Re}[G_\tau(x), x] \geqq 0 \qquad (\text{for all } x \in \mathbb{C}^{sm} \text{ with } \|x\| = r).$$

Consequently, for each $\tau > 0$, there exists an $x(\tau) \in \mathbb{C}^{sm}$ with

$$G_\tau(x(\tau)) = 0, \qquad \|x(\tau)\| \leqq r.$$

It follows that there is a sequence $\tau_1, \tau_2, \tau_3, \ldots \downarrow 0$ such that the vectors $x(\tau_1)$, $x(\tau_2)$, $x(\tau_3)$, ... converge to some limit $x^* \in \mathbb{C}^{sm}$. Since $x(\tau_k) - [A + \tau_k I]F(x(\tau_k)) = 0$ $(k \geqq 1)$ we see that x^* is a solution to (3.1). ■

Clearly condition (1.4) implies that the assumption on A in theorem 2 is satisfied with $D = B$. We thus arrive at the following

COROLLARY. *Let the Runge-Kutta method satisfy condition* (1.4). *Then the system of equations* (1.3b) *has a unique solution whenever* $f\colon \mathbb{R} \times \mathbb{C}^s \to \mathbb{C}^s$ *is a continuous function satisfying* (1.6), (1.7).

4. Remarks.

REMARK 1. If $m = 1$, then system (1.3b) is essentially the same as the system of algebraic equations arising if one uses a linear multistep method to compute the approximations u_n. For this case several authors have discussed the existence and uniqueness of solutions; see [4, 6, 9, 10, 11]. It should be noted that their results can also be used to obtain results on Runge-Kutta methods which are diagonally implicit, i.e. $a_{jk} = 0$ whenever $j < k$.

REMARK 2. It seems to us that many B-stable methods satisfy the hypothesis of theorem 1, and for these methods we have existence and uniqueness as soon as (1.2) is satisfied. However, the counterexample given in section 2 shows the existence of B-stable methods which do not satisfy the assumption of theorem 1. We also note that theorem 1 can be applied to some methods which are not B-stable.

REMARK 3. Let $\beta > 0$ be a given constant. It is easily verified that a function $f\colon \mathbb{R} \times \mathbb{C}^s \to \mathbb{C}^s$ satisfies both (1.6) and (1.7) whenever

$$(4.1) \quad \operatorname{Re}\langle f(t,\tilde{\xi}) - f(t,\xi), \tilde{\xi} - \xi\rangle \leqq -\beta|\tilde{\xi} - \xi|^2 \qquad (\text{for all } t \in \mathbb{R} \text{ and } \tilde{\xi}, \xi \in \mathbb{C}^s).$$

Condition (4.1) is equivalent to requiring that $f(t,\cdot) + \beta I$ is dissipative, or that $-f(t,\cdot)$ is uniformly (or strongly) monotone with monotonicity constant β (cf. [8, p. 141], [7, p. 61], [4, p. 63]).

REMARK 4. As in [1, 2] we might consider the following weaker version of requirement (1.4),

$$(4.2) \qquad b_j \geqq 0 \qquad (1 \leqq j \leqq m), \ Q \text{ is positive semi-definite.}$$

However, we would gain little by dealing with (4.2) instead of (1.4) since there are no Runge-Kutta methods of practical interest which satisfy (4.2) but not (1.4) (because such methods are equivalent to methods with fewer stages in which only strict inequalities occur – see [5] for more details). Moreover the next example shows that the conclusion in the corollary of section 3 is not necessarily valid for Runge-Kutta methods satisfying (4.2) but violating (1.4).

An example of a Runge-Kutta method which satisfies (4.2) and for which the system (1.3b) need not have a solution when (1.6), (1.7) hold, is given by

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

This method (with $m = 2$) is of no practical interest since the approximations u_n computed by it could have been calculated more easily from the Backward Euler Method (i.e. (1.3) with $m = 1, a_{11} = 1, b_1 = 1$).

REMARK 5. All conclusions in the sections 2 and 3 remain valid if we deal throughout with the space $\mathbb{R}^s$ instead of $\mathbb{C}^s$.

REFERENCES

1. K. Burrage and J. C. Butcher, *Stability criteria for implicit Runge-Kutta methods*, SIAM J. Numer. Anal. 16 (1979), 46–57.
2. M. Crouzeix, *Sur la B-stabilité des méthodes de Runge-Kutta*, Numer. Math. 32 (1979), 75–82.
3. M. Crouzeix and P. A. Raviart, Unpublished Lecture Notes, Université de Rennes 1980.
4. G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, Lecture Notes in Mathematics 506, Berlin, Springer-Verlag (1976).
5. G. Dahlquist and R. Jeltsch, *Generalized disks of contractivity for explicit and implicit Runge-Kutta methods*. Report TRITA-NA-7906, Dept. Comp. Sci., Roy. Inst. of Techn., Stockholm (1979).
6. C. A. Desoer and H. Haneda, *The measure of a matrix as a tool to analyse computer algorithms for circuit analysis*. IEEE Trans. Circuit Theory 19 (1972), 480–486.
7. H. Gajewski, K. Gröger and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Berlin, Akademie-Verlag (1974).
8. J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, New-York, Academic Press (1970).
9. I. W. Sandberg, *Theorems on the computation of transient response of nonlinear networks containing transistors and diodes*, Bell System Tech. J. 49 (1970), 1739–1776.
10. I. W. Sandberg and H. Shichman, *Numerical integration of systems of stiff nonlinear differential equations*, Bell System Tech. J. 47 (1968), 511–527.
11. J. Williams, *The problem of implicit formulas in numerical methods for stiff differential equations*, Numer. Anal. Report No. 40, Dept. of Math., University of Manchester, Manchester (1979).