IMA Journal of Numerical Analysis (1991) **11**, 449–455

A New Class of Results for the Algebraic Equations of Implicit Runge–Kutta Processes

J. M. Sanz-Serna

Departamento de Matemática Aplicada y Computación, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain

D. F. Griffiths

Department of Mathematics and Computer Science, University of Dundee, Dundee DD1 4HN, Scotland, UK

Dedicated to Professor A. R. Mitchell on the occasion of his 70th birthday

[Received 23 April 1990 and in revised form 13 December 1990]

We are concerned with the solvability of the discrete equations arising in the use of Runge–Kutta methods. Under suitable assumptions on the RK tableau, we show that in the neighbourhood of a sink (asymptotically stable equilibrium) a unique solution exists for arbitrarily large stepsizes. Furthermore in the neighbourhood of a slowly varying integral curve $z = z(t)$ that attracts neighbouring integral curves of the ODE system, a unique solution of the algebraic equations exists, provided that the stepsize is suitably restricted. This restriction does not depend on the stiffness of the ODEs being integrated.

1. Introduction

We are concerned with the algebraic equations

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j), \quad 1 \leqslant i \leqslant s, \tag{1.1}$$

that result from the application of an implicit Runge–Kutta process with real coefficients (a_{ij}) and real abscissae (c_i) to the m-dimensional system

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y), \quad t \geqslant t_0. \tag{1.2}$$

Throughout the paper the steplength h is supposed to be greater than 0.

The available results on the existence and uniqueness of a solution $Y_1, ..., Y_s$ to (1.1) are of two types:

(i) Results that hold for h *sufficiently* small, see for example Hairer *et al.* (1987: p. 201). These are based on the contraction mapping/implicit function theorem and work under mild assumptions on f, typically local Lipschitz continuity. The results of this group may be regarded as not powerful enough, since implicit methods are resorted to in order to operate with steplengths h that are *large*. (Here 'small' and 'large' are measured with respect to the reciprocal of the Lipschitz constant.)

(ii) Results where the steplength h is not restricted by the size of the classical Lipschitz constant. The first theorems in this direction were given by Crouzeix, Dekker, Hundsdorfer, and Spijker in papers of the early 1980s. Their work is surveyed in Dekker & Verwer (1984: Ch. 5). More recent references are Di Lena & Peluso (1985), Hundsdorfer & Spijker (1987), Liu & Kraaijevanger (1988), and Spijker (1985). This sort of theorem is powerful in the case of dissipative problems in that, for each $h > 0$, a unique solution $Y_1, \ldots, Y_s$ is shown to exist, under appropriate assumptions on (a_{ij}). However, such a strong conclusion is only possible at the price of imposing on f the very restrictive requirement of *global* dissipativity: a real constant ν should exist, such that in an appropriate inner product $\langle \bullet, \bullet \rangle$ in $\mathbb{R}^m$, with associated norm $\|\bullet\|$,

$$\langle f(t, z_1) - f(t, z_2), z_1 - z_2 \rangle \leq \nu \|z_1 - z_2\|^2, \quad \text{for all } z_1, z_2 \in \mathbb{R}^m, \ t \geq t_0. \tag{1.3}$$

The one-sided Lipschitz constant ν is required to be either <0 or ≤ 0, depending on the specific theorem. The requirement (1.3) is so strong that it is satisfied by very few functions f arising in real applications. Correspondingly theorems in this group have conclusions that are stronger than actually needed in practice. While it is true that in stiff problems we wish to take steps that are significantly larger than the reciprocal of the Lipschitz constant, accuracy considerations preclude using values of h that are arbitrarily large. Also, the insistence on *global* uniqueness is not necessary in practice. The solution should be unique with $Y_1, \ldots, Y_s$ near y_0. Spurious solutions away from y_0 usually exist. They cause no problem provided that the iterative nonlinear solver and the initial guess used to deal with (1.1) are judiciously chosen. (It should be mentioned that it is possible to treat, with theorems in this second group, the case where ν in (1.3) is positive, see Dekker & Verwer, (1984: Thms 5.3.9, 5.3.12).)

The following example may clarify these issues. Consider the backward Euler rule applied to the logistic equation

$$y' = -y + y^2 \tag{1.4}$$

arguably the simplest nonlinear model (cf. Iserles, 1989). Solutions of (1.4) with initial value $y_0 > 1$ only exist in a bounded time interval $(0, t_{\max})$, with $t_{\max} \equiv t_{\max}(y_0)$. Solutions with $y_0 < 1$ are attracted, eventually exponentially, towards the equilibrium $y = 0$. Condition (1.3) is not satisfied and the results of the second group do not apply (or at least not directly). On the other hand, if y_0 is very close to 0 we would like to take a step h substantially larger than unity (reciprocal of the local Lipschitz constant), a situation not covered by results of the first group. It is only in this y_0 close to 0 situation that the application to (1.4) of a stiff solver is practically meaningful. In other regimes an explicit method would certainly be more advantageous.

Since a simple differential equation and a simple method are being considered, it is straightforward to solve, in closed form, the corresponding nonlinear equation

$$Y = y_0 + h(-Y + Y^2). \tag{1.5}$$

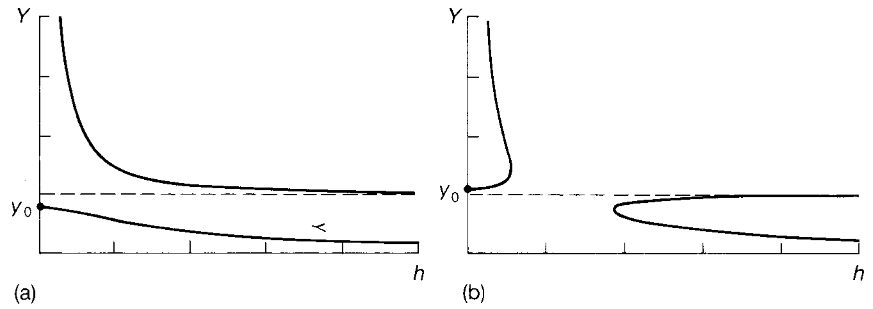

FIG. 1. Solutions of equation (1.5) with (a) $0<y_0<1$ and (b) $y_0>1$.

The result when $y_0 \in (0\,,1)$ is plotted in Fig. 1a, while Fig. 1b depicts the case $y_0>1$. Note that in Fig. 1a a solution does exist for arbitrarily large h, so that contraction mapping results are over pessimistic. On the other hand, in Fig. 1b, the branch emanating from y_0 can only be continued over a finite h interval $0 \leqslant h \leqslant h_{\max}$. This matches the fact that the true solution only exists over a finite time interval. However, while the true solution ceases to exist as it becomes infinite, the numerical solution has a turning point at $h = h_{\max}$. Finally, observe in both figures that when a numerical solution exists, it is generally nonunique; real solutions occur in pairs as correspond to a second-degree equation for Y.

In this paper we give existence and uniqueness results that are not of the types (i)–(ii) above. In Section 2 we examine the situation where (1.2) is autonomous and has an equilibrium that, without loss of generality, may be assumed to be $y=0$. If this equilibrium is a hyperbolic sink, so that neighbouring solutions of (1.2) are attracted towards 0 at an exponential rate, we shall prove that, under appropriate conditions on (a_{ij}), the system (1.1) possesses a solution for arbitrary $h>0$. Furthermore this solution is locally unique. In Section 3 we look at the case where (1.2) is nonautonomous and possesses a slowly varying solution $y=z(t)$ defined for all $t \geqslant t_0$ and such that $z(t)$ exponentially attracts neighbouring solutions. Under suitable assumptions on the RK method, we prove that, near $z(t)$, (1.1) possesses a unique solution, provided that $0<h<h_{\max}$. Here $h_{\max}$ is only determined by the size of the derivatives of z and not by the stiffness of (1.2). In other words, in the situations considered, a large Lipschitz constant does not unduly restrict the solvability of the algebraic equations (1.1) (cf. the B-convergence theory of Frank, Schneid, and Ueberhuber (see Dekker & Verwer, 1984: Ch. 7)).

It should be pointed out that we have not attempted to present our results in all possible generality or under minimal hypotheses. Our aim has been to illustrate the underlying problems and ideas.

There has been considerable interest recently in the investigation of the behaviour of numerical methods for evolutionary problems in cases where the time step is large (Griffiths & Mitchell, 1988; Iserles, 1989; Iserles *et al.*, 1990; Sleeman *et al.*, 1988; Stuart, 1989; Sweby *et al.*, 1991). The application of implicit methods in these situations requires, as a first step, the study of the solvability of the nonlinear equations along lines similar to those delineated in this paper.

2. The case of a sink

In the remainder of the paper we consider the following assumptions on the RK method.

(RK1) Each internal stage is consistent, i.e.

$$\sum_{j=1}^{s} a_{ij} = c_i, \quad 1 \leqslant i \leqslant s.$$

(RK2) The abscissae c_i, $1 \leqslant i \leqslant s$, are nonnegative.

(RK3) An $s \times s$ diagonal matrix D with positive diagonal elements exists such that

$$\psi_{\mathrm{D}}(A) = \min_{\xi \neq 0} \frac{\langle A\xi, \xi\rangle_{\mathrm{D}}}{\langle \xi, \xi\rangle_{\mathrm{D}}} > 0,$$

where $\langle \bullet, \bullet \rangle_{\mathrm{D}}$ is the inner product in $\mathbb{R}^s$ defined by

$$\langle \xi, \eta \rangle = \sum_{i=1}^{s} d_i \xi_i \eta_i.$$

The condition (RK2) is useful when (1.2) is not autonomous and f is not defined for $t < t_0$. Condition (RK3) holds for many stiff solvers and has been often used in the literature (see Dekker & Verwer, 1984: Sections 5.5–5.10). Under this condition, A is invertible and $\Psi_{\mathrm{D}}(A^{-1}) > 0$ (Dekker & Verwer, 1984: Corollary 5.1.4). We point out the suggestion in Dekker & Verwer (1984: Remark 5.4.5) that (RK3) may be a stronger property than required.

In this section our assumptions on f are as follows:

(F1) f does not depend on t and is defined and continuously differentiable in a neighbourhood of $0 \in \mathbb{R}^m$.

(F2) 0 is an equilibrium of (1.2), i.e. $f(0) = 0$.

(F3) 0 is a hyperbolic sink of (1.2), that is, the Jacobian J of f at 0 has no eigenvalue with real part $\geqslant 0$ (Guckenheimer, & Holmes, 1983).

Under these conditions solutions of (1.2) that start near 0 are defined for all $t \geqslant t_0$ and approach 0 at an exponential rate. For the discrete equations we have

THEOREM 1. If RK3 and F1–F3 above hold, there exist neighbourhoods Ω_1,Ω_2 of 0 in $\mathbb{R}^m$ so that, for each $h > 0$ and each $y_0 \in \Omega_1$, system (1.1) has a unique solution $Y_1,\ldots, Y_s$ with $Y_i \in \Omega_2$, $1 \leqslant i \leqslant s$.

Proof. The hypothesis F3 on J implies (Hirsch & Smale, 1974: Ch. 7) that an inner product $\langle \bullet, \bullet \rangle$ in $\mathbb{R}^m$ and a negative number ν can be found such that

$$\langle z, Jz \rangle \leqslant \nu \, \|z\|^2, \quad z \in \mathbb{R}^m.$$

Since f is continuously differentiable near 0,

$$\langle f(z_1) - f(z_2), z_1 - z_2 \rangle = \langle z_1 - z_2, J(z_1 - z_2) \rangle + o(\|z_1 - z_2\|^2), \qquad z_1, z_2 \to 0,$$

and therefore a neighbourhood Ω_2 of 0 in $\mathbb{R}^m$ exists such that

$$\langle f(z_1) - f(z_2), z_1 - z_2 \rangle \leqslant 0, \quad z_1\, z_2 \in \Omega_2. \tag{2.1}$$

The uniqueness of solutions with $Y_i \in \Omega_2$ can now be proved in a standard way as in Dekker & Verwer (1984: Thm 5.3.9).

To prove the existence we employ an argument based on topological degree (Ortega & Rheinboldt, 1970). We first rewrite (1.1) in the compact format

$$Y = e \otimes y_0 + h(A \otimes I)F(Y), \tag{2.2}$$

where

$$Y = [Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}}, \dots, Y_s^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{ms}, \qquad e = [1, 1, \dots, 1]^{\mathrm{T}} \in \mathbb{R}^{s},$$

$\otimes$ denotes Kronecker product, A is the $s \times s$ coefficient matrix, I is the $m \times m$ identity, and $F(Y) = [f(Y_1)^{\mathrm{T}}, f(Y_2)^{\mathrm{T}}, \dots, f(Y_s)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{ms}$. Also we introduce in $\mathbb{R}^{ms}$ the inner product

$$[X, \tilde{X}]_{\mathrm{D}} = \sum_{j=1}^{s} d_j \langle X_j, \tilde{X}_j \rangle, \quad X, \tilde{X} \in \mathbb{R}^{ms}, X_j, \tilde{X}_j \in \mathbb{R}^{m},$$

and denote by $|||\bullet|||_{\mathrm{D}}$ the associated norm. (We use throughout the notation of Dekker & Verwer (1984).) After these preliminaries, (2.2) leads to

$$(A^{-1} \otimes I)Y - hF(Y) = (A^{-1} \otimes I)(e \otimes y_0),$$

$$[(A^{-1} \otimes I)Y, Y] - h[F(Y), Y] = [(A^{-1} \otimes I)(e \otimes y_0), Y]. \tag{2.3}$$

The term $[F(Y), Y]$ is easily proved to be $\leqslant 0$ (see (2.1)), if $Y \in (\Omega_2)^s$. Results in Dekker & Verwer (1984, Section 5.3) show that

$$\psi_{\mathrm{D}}(A^{-1})\, |||Y|||_{\mathrm{D}}^2 \leqslant [(A^{-1} \otimes I)Y, Y],$$

$$[(A^{-1} \otimes I)(e \otimes y_0), Y] \leqslant \|A^{-1}\|_{\mathrm{D}}\, |||e \otimes y_0|||_{\mathrm{D}}\, |||Y|||_{\mathrm{D}}$$

$$= \|A^{-1}\|_{\mathrm{D}} \Big(\sum_{i=1}^{s} d_i\Big)^{\frac{1}{2}} \|y_0\|\; |||Y|||_{\mathrm{D}}$$

and hence (2.3) gives, whenever (2.2) holds with $Y \in (\Omega_2)^s$,

$$|||Y|||_{\mathrm{D}} \leqslant \beta \|y_0\|, \tag{2.4}$$

where

$$\beta = \|A^{-1}\|_{\mathrm{D}} \Big(\sum_{i=1}^{s} d_i\Big)^{\frac{1}{2}} / \Psi_{\mathrm{D}}(A^{-1}) \geqslant \Big(\sum_{i=1}^{s} d_i\Big)^{\frac{1}{2}}.$$

Now choose $r > 0$ such that the closed ball B_r centred at $0 \in \mathbb{R}^{ms}$ is contained in $(\Omega_2)^s$ and define Ω_1 to be the open ball centred at 0 with radius r/β. The estimate (2.4) implies that if $y_0 \in \Omega_1$ and Y satisfies (2.2), then Y cannot belong to the boundary of B_r. Thus the degree of the mapping

$$G_h(Y) = Y - e \otimes y_0 - h(A \otimes I)F(Y), \quad Y \in B_r$$

does not vary with h, $h \geqslant 0$. At $h = 0$ this degree is 1 because

$$|||e \otimes y_0|||_{\mathrm{D}} = \Big(\sum_{i=1}^{s} d_i\Big)^{\frac{1}{2}} \|y_0\| < \frac{r}{\beta} \Big(\sum_{i=1}^{s} d_i\Big)^{\frac{1}{2}} \leqslant r$$

so that $e \otimes y_0 \in$ interior of B_r. Hence the degree is 1 for each $h > 0$ and a solution exists. □

3. The case of a slowly varying solution

We now consider the following hypotheses on f:

(F4) f is defined and continuous in a tube

$$T_r = \{(t, \xi): t \geqslant t_0,\ \|\xi - z(t)\| < r\}, \quad r > 0$$

around a solution $z = z(t)$ of (1.2).

(F5) $\langle f(t, z_1) - f(t, z_2), z_1 - z_2 \rangle \leqslant v\|z_1 - z_2\|^2$, whenever $t \geqslant t_0$, $\|z_1 - z(t)\| < r$, $i = 1, 2$. Here v is a nonpositive constant.

In practice we are interested in the case where z is a slowly varying function and $v < 0$ so that z attracts neighbouring solutions. However, these hypotheses on z are not needed below.

THEOREM 2. *If* RK1–RK3 *and* F4–F5 *hold, there exists a neighbourhood* Ω_1, *of* $z(0)$ *in* $\mathbb{R}^m$, *a tube* T_{r_1} *around* $z = z(t)$, *and a constant* $h_0 > 0$ *such that, for* $h \in [0, h_0)$ *and* $y_0 \in \Omega_1$, *the system* (1.1) *has a unique solution* $Y_1, \ldots, Y_s$ *with* $(t_0 + c_i h, Y_i) \in T_{r_1}$, $1 \leqslant i \leqslant s$. *Furthermore, for a given* RK *method the upper bound* h_0 *on* h *only depends on the radius* r *in* F4, *on the norm* $\|\bullet\|$ *for which* F5 *holds, and on bounds of the norm of the* $(q+1)$th *derivatives of* z, *where* q *is the stage order of the method.*

Proof. The proof is an extension of that in the previous section and again we only present the existence part. The discrete equations are still of the form (2.2), where now

$$F(Y) = [f(t_0 + c_1 h, Y_1)^{\mathrm{T}}, \ldots, f(t_0 + c_s h, Y_s)^{\mathrm{T}}]^{\mathrm{T}}.$$

We define

$$Z = [z(t_0 + c_1 h)^{\mathrm{T}}, \ldots, z(t_0 + c_s h)^{\mathrm{T}}]^{\mathrm{T}},$$

$$F(Z) = [f(t_0 + c_1 h, z(t_0 + c_1 h))^{\mathrm{T}}, \ldots, f(t_0 + c_s h, z(t_0 + c_s h))^{\mathrm{T}}]^{\mathrm{T}},$$

and

$$R = Z - e \otimes z(0) - h(A \otimes I)F(Z). \tag{3.1}$$

Since z solves (1.2), the components of $F(Z)$ are values of $z'(t)$. Hence, Taylor expansion of (3.1) leads, after taking RK1 into account, to $|||R|||_{\mathrm{D}} \leqslant Ch^{q+1}$, where q is the stage order of the RK method ($q \geqslant 1$) and C is essentially a bound for the $(q+1)$th derivative of z.

Now subtract (3.1) from (2.2), to obtain

$$Y - Z = e \otimes (y_0 - z(0)) + h(A \otimes I)[F(Y) - F(Z)] - R,$$

which leads, as in Section 2, to the a priori bound

$$|||Y - Z|||_{\mathrm{D}} \leqslant \{\|A^{-1}\|_{\mathrm{D}} / \Psi_{\mathrm{D}}(A^{-1})\}\{\, |||R|||_{\mathrm{D}} + |||e \otimes (y_0 - z(0))|||_{\mathrm{D}}\}, \tag{3.2}$$

an inequality that now plays the role of (2.4). If h is small and y_0 is close to $z(0)$, the right-hand side of (3.2) is small and therefore $(t_0 + c_1 h, Y_1), \ldots, (t_0 + c_s h, Y_s)$ cannot cross the boundary of a tube T_{r_1}. □

Remark. It is also possible to prove a similar theorem if v in F5 is positive. In such a case h should be less than $\min(h_0, \Psi_D(A^{-1})/v)$ where h_0 again depends on r and on the size of the norm of the $(q+1)$th derivative of z (see Dekker & Verwer, 1984: Thm 5.3.7).

References

DEKKER, K., & VERWER, J. G. 1984 *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations.* Amsterdam: North-Holland.

DI LENA, G., & PELUSO, R. I. 1985 On conditions for the existence and uniqueness of solutions to the algebraic equations in Runge–Kutta methods. *BIT* **25,** 223–232.

GRIFFITHS, D. F., & MITCHELL, A. R. 1988 Stable periodic bifurcations of an explicit discretization of a nonlinear partial differential equation in reaction diffusion. *IMAJ. Numer. Anal.* **8,** 435–454.

GUCKENHEIMER, J., & HOLMES, P. 1983 *Nonlinear oscillations, dynamical systems and bifurcations of vector fields.* New York: Springer.

HAIRER, E., NORSETT, S. P., & WANNER, G. 1987 *Solving Ordinary Differential Equations I, Nonstiff problems.* Berlin: Springer-Verlag.

HIRSCH, M. W., & SMALE, S. 1974 *Differential Equations, Dynamical Systems, and Linear Algebra.* New York: Academic Press.

HUNDSDORFER, W. H., & SPIJKER, M. N. 1987 On the algebraic equations in implicit Runge–Kutta methods. *SIAM J. Numer. Anal.* **24,** 583–594.

ISERLES, A. 1989 Nonlinear stability and asymptotics of ODE solvers. In: *International Conference on Numerical Mathematics* (R. P. Agarwal, Ed.). Basel: Birkhauser.

ISERLES, A., PEPLOW, A. T., & STUART, A. 1990 A unified approach to spurious solutions introduced by time discretisation Part I: Basic Theory. Report NA4, University of Cambridge.

LIU, M. Z., & KRAAIJEVANGER, J. F. B. M. 1988 On the solvability of the systems of equations arising in implicit Runge–Kutta methods. *BIT* **28,** 825–838.

ORTEGA, J. M., & RHEINBOLD, W. C. 1970 *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic Press.

SLEEMAN, B. D., GRIFFITHS, D. F., MITCHELL, A. R., & SMITH, P. D. 1988 Period doubling bifurcations in nonlinear difference equations. *SIAM J. Sci. and Stat. Comp.* **9,** 543–557.

SPIJKER, M. N. 1985 Feasibility and contractivity in implicit Runge–Kutta methods. *J. Comp. Appl. Math.* **12, 13,** 563–578.

STUART, A. 1989 Linear instability implies spurious periodic solutions. *IMA J. Numer. Anal.* **9,** 465–486.

SWEBY, P. K., YEE, H. C., & GRIFFITHS, D. F. 1991 On spurious steady state solutions of explicit Runge–Kutta schemes. NASA Technical Memorandum, in preparation.