Computing 36, 17–34 (1986)

Computing
© by Springer-Verlag 1986

A Study of B-Convergence of Runge-Kutta Methods

K. Burrage, Auckland, **W. H. Hundsdorfer**, and **J. G. Verwer**, Amsterdam

Received August 5, 1985

Abstract — Zusammenfassung

A Study of B-Convergence of Runge-Kutta Methods. This paper deals with the convergence analysis of implicit Runge-Kutta methods as applied to stiff, semilinear systems of the form $\dot{U}(t)=QU(t)+g(t, U(t))$. A criterion is developed which determines whether the order of optimal B-convergence is at least equal to the stage order or one order higher. This criterion is studied for a number of interesting classes of methods.

AMS Subject Classification: 65L05.

C.R. number: 5.17.

Key words: Numerical analysis, implicit Runge-Kutta methods, stiff problems, B-convergence.

Eine Untersuchung über B-Konvergenz von Runge-Kutta Verfahren. Dieser Aufsatz befaßt sich mit der Analyse der Konvergenz von impliziten Runge-Kutta Verfahren für steife, semi-lineare Systeme der Form $\dot{U}(t)=QU(t)+g(t, U(t))$. Ein Kriterium wird entwickelt, welches entscheidet, ob die Ordnung der optimalen B-Konvergenz mindestens gleich der Stufenordnung oder um eine Ordnung höher ist. Dieses Kriterium wird untersucht für eine Zahl von interessanten Klassen von Verfahren.

1. Introduction

Consider the stiff system of ordinary differential equations

$$\dot{U}(t)=f(t, U(t)), 0\leq t\leq T, U(0)=u_0, \tag{1.1}$$

with $f: \mathbb{R}\times\mathbb{R}^m\to\mathbb{R}^m$ satisfying the one-sided Lipschitz condition (with one-sided Lipschitz constant ν)

$$\langle f(t,\tilde{u})-f(t,u), \tilde{u}-u\rangle\leq\nu|\tilde{u}-u|^2, \ \forall t\in\mathbb{R}, \ \forall\tilde{u}, u\in\mathbb{R}^m, \tag{1.2}$$

for the inner product $\langle\cdot,\cdot\rangle$ in $\mathbb{R}^m$ ($|\cdot|$ being the related norm). For the numerical integration of (1.1) we consider the Runge-Kutta method given by

$$\begin{aligned} u_{n+1}&=u_n+\tau\sum_{i=1}^{s} b_i f(t_n+c_i\tau, y_i^{(n)}),\\ y_i^{(n)}&=u_n+\tau\sum_{j=1}^{s} a_{ij} f(t_n+c_j\tau, y_j^{(n)}), \ i=1(1)s, \end{aligned} \tag{1.3}$$

where τ is the stepsize $t_{n+1}-t_n$ and u_n approximates the exact solution $U(t)$ of (1.1) at $t=t_n$.

For a long time the interesting phenomenon of stiffness has been related solely to the stability of the Runge-Kutta method. However, it is now known that stiffness has a significant impact on the accuracy as well. Even if the solution U is smooth (no layers) and the scheme (1.3) is stable, the accuracy of the approximation is often worse than expected when the order of consistency of (1.3) is taken into account. This fundamental point was perceived first by Prothero and Robinson [15] in their analysis of the scalar test-equation $\dot{U}(t)=\lambda U(t)+\dot{g}(t)-\lambda g(t)$. Frank, Schneid and Ueberhuber [7, 8, 9] extended the ideas of Prothero and Robinson to the general nonlinear problem (1.1) in the B-convergence theory.

Let τ be constant, $t_N=N\tau=T$ and $\varepsilon_N=U(t_N)-u_N$, that is, the global error at $t=T$. The main object of the B-convergence theory is the derivation of bounds for ε_N of the form

$$|\varepsilon_N|\leq C\tau^p, \quad \forall\tau\in(0,\bar{\tau}], \tag{1.4}$$

where the stepsize bound $\bar{\tau}$ is a constant determined only by ν and C is a constant determined only by ν, T and by bounds for certain derivatives $d^i U(t)/dt^i$. Hence no other quantities, which might be disproportionately large due to stiffness (e.g., the (two-sided) Lipschitz constant), are allowed to be present in the bound (*optimal B-convergence*). Such bounds are often in better accord with the true error behaviour ([7, 9, 18], [6], Section 7.5) than the classical error bounds.

We are now ready to discuss the main goal of our paper. Let $\hat{u}_{n+1}$ be the Runge-Kutta result from the transition $U(t_n)\to\hat{u}_{n+1}$ and $l_{n+1}=U(t_{n+1})-\hat{u}_{n+1}$ the local truncation error. In their analysis Frank, Schneid and Ueberhuber [7, 8, 9] essentially bound this local error, as in (1.4) but with p replaced by $p+1$ (B-consistency), which is then transferred to (1.4) by stability arguments (see e.g. [6], Chapter 7 for the useful notion of C-stability). For many of the implicit schemes they are thus able to prove optimal B-convergence of order $p=\tilde{p}$, where $\tilde{p}$ is the minimal order of all stages in (1.3) (the stage order). It is known, however, that the approach of first bounding all local errors and then adding via the stability argument not necessarily leads to the best possible result [6, 8]. An example is provided by the implicit midpoint rule for which $p=\tilde{p}+1=2$. This was proved by Kraaijevanger [13] and earlier, but in a more complicated way, by Stetter [17]. We have strong numerical evidence ([6], Section 7.5 and [18]) that for many other interesting schemes $p=\tilde{p}+1$ uniformly on the problem class (1.1)–(1.2). In this paper we analyse this discrepancy between the local and global order reduction for stiff problems of the semi-linear form

$$\dot{U}(t)=QU(t)+g(t,U(t)), \tag{1.5}$$

where the constant $m\times m$ matrix Q and the vector function $g:\mathbb{R}\times\mathbb{R}^m\to\mathbb{R}^m$ satisfy

$$\langle Qu,u\rangle\leq\beta|u|^2, \quad \forall u\in\mathbb{R}^m, \tag{1.6a}$$

$$|g(t,\tilde{u})-g(t,u)|\leq\alpha|\tilde{u}-u|, \quad \forall\tilde{u},u\in\mathbb{R}^m \text{ and } t\in\mathbb{R} \tag{1.6b}$$

thus tacitly assuming that the stiffness is contained in the constant coefficient linear part of the problem.

For various A-stable methods we prove that for this semilinear problem we have $p=\tilde{p}+1$. One of the outcomes of our investigations is that there are Runge-Kutta methods with $p=s+1$, whereas it is known that for the Gauss methods with $s\geq 2$ we only have $p=s$ (see [5]).

2. Recursion Schemes for the Global Error

In order to write the Runge-Kutta scheme (1.3) in a more compact way we introduce some notation. The $s\times s$ and $m\times m$-identity matrices will be denoted by I_s, I_m, respectively, or, if no confusion can arise, simply by I. The vector e stands for the vector in $\mathbb{R}^s$ with all components equal to one. Further we put $\mathbf{A}=A\otimes I_m$, $\mathbf{b}^T=b^T\otimes I_m$, $\mathbf{e}=e\otimes I_m$, $\mathbf{I}=I_s\otimes I_m$ where A is the $s\times s$-matrix with entries a_{ij}, $b^T=(b_1, b_2, \ldots, b_s)^T$, and $\otimes$ is the Kronecker product. On the space $\mathbb{R}^{sm}$ we shall deal with the norm $\|y\|=(\Sigma_{i=1}^s|y_i|^2)^{1/2}$ for $y=(y_1, y_2, \ldots, y_s)^T\in\mathbb{R}^{sm}$, $|\cdot|$ being the inner-product norm on $\mathbb{R}^m$. Also the corresponding operator norms on $L(\mathbb{R}^m)$ and $L(\mathbb{R}^{sm})$ (spaces of linear operators) will be denoted by $|\cdot|$, $\|\cdot\|$, respectively.

For a given stepsize $\tau>0$ and $f:\mathbb{R}\times\mathbb{R}^m\to\mathbb{R}^m$ we define the function $F:\mathbb{R}\times\mathbb{R}^{sm}\to\mathbb{R}^{sm}$ by

$$F(t,y)=(f(t+c_1\tau, y_1),\ f(t+c_2\tau, y_2), \ldots, f(t+c_s\tau, y_s))^T$$

for $t\in\mathbb{R}$ and $y=(y_1, y_2, \ldots, y_s)^T\in\mathbb{R}^{sm}$.

With these notations the Runge-Kutta scheme (1.3) can be written as

$$u_{n+1}=u_n+\tau\,\mathbf{b}^T F(t_n, y_n), \tag{2.1 a}$$

$$y_n=\mathbf{e}\,u_n+\tau\,\mathbf{A}\,F(t_n, y_n), \tag{2.1 b}$$

where $y_n=(y_1^{(n)}, y_2^{(n)}, \ldots, y_s^{(n)})^T\in\mathbb{R}^{sm}$.

Let $Y_n=(Y_1^{(n)}, Y_2^{(n)}, \ldots, Y_s^{(n)})^T\in\mathbb{R}^{sm}$, $Y_i^{(n)}=U(t_n+c_i\tau)$ with U the solution of (1.1). Following [9] we define the residual errors $\rho_n\in\mathbb{R}^m$ and $r_n=(r_1^{(n)}, r_2^{(n)}, \ldots, r_s^{(n)})^T\in\mathbb{R}^{sm}$ such that

$$U(t_{n+1})=U(t_n)+\tau\,\mathbf{b}^T F(t_n, Y_n)+\rho_n, \tag{2.2 a}$$

$$Y_n=\mathbf{e}\,U(t_n)+\tau\mathbf{A}F(t_n, Y_n)+r_n. \tag{2.2 b}$$

In the next sections we shall use the following order conditions on the Runge-Kutta method,

$$B(p): b^T c^{j-1}=\frac{1}{j}\ (1\leq j\leq p),$$

$$C(q): A c^{j-1}=\frac{1}{j}\,c^j\ (1\leq j\leq q),$$

with $c^j=(c_1^j, c_2^j, \ldots, c_s^j)^T\in\mathbb{R}^s$. For a given $q\in\mathbb{N}$ we define the vector $k=(k_1, k_2, \ldots, k_s)^T\in\mathbb{R}^s$ by

$$k=\frac{1}{q!}\left(\frac{1}{q+1}\,c^{q+1}-A\,c^q\right). \tag{2.3}$$

From (2.2) it easily follows by a Taylor series expansion that these order conditions $B(p)$, $C(q)$ are equivalent to saying that

$$\rho_n = \tau^{p+1} \frac{1}{p!} \left(\frac{1}{p+1} - b^T c^p \right) U^{(p+1)}(t_n) + \mathcal{O}(\tau^{p+2}) \quad (\tau \downarrow 0),$$

$$r_i^{(n)} = \tau^{q+1} k_i U^{(q+1)}(t_n) + \mathcal{O}(\tau^{q+2}) \quad (\tau \downarrow 0),$$

where in the order terms only higher derivatives of U are involved (see also [9]). We note that the *stage order* of the Runge-Kutta method equals q iff both $B(q)$ and $C(q)$ hold.

Subtraction of (2.1) from (2.2) yields the following recursion scheme for the errors $\varepsilon_n = U(t_n) - u_n$ and $\delta_n = (\delta_1^{(n)}, \delta_2^{(n)}, \ldots, \delta_s^{(n)})^T = Y_n - y_n$,

$$\varepsilon_{n+1} = \varepsilon_n + \mathbf{b}^T \mathbf{Z}_n \delta_n + \rho_n, \tag{2.4a}$$

$$\delta_n = \mathbf{e} \varepsilon_n + \mathbf{A} \mathbf{Z}_n \delta_n + r_n \tag{2.4b}$$

where $\mathbf{Z}_n \in L(\mathbb{R}^{sm})$ is the block diagonal matrix with blocks $Z_i^{(n)} \in L(\mathbb{R}^m)$ on the diagonal, defined by

$$Z_i^{(n)} = \tau \int_0^1 f'(t_n + c_i \tau, y_i^{(n)} + \theta(Y_i^{(n)} - y_i^{(n)}))\, d\theta \quad (1 \le i \le s),$$

with $f'(t,u)$ the Jacobian matrix $\frac{\partial}{\partial u} f(t,u)\,(t \in \mathbb{R}, u \in \mathbb{R}^m)$.

Assuming $\mathbf{I} - \mathbf{A}\mathbf{Z}_n$ to be regular we obtain from (2.4) the recursion

$$\varepsilon_{n+1} = [I + \mathbf{b}^T \mathbf{Z}_n (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} \mathbf{e}]\, \varepsilon_n + \mathbf{b}^T \mathbf{Z}_n (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} r_n + \rho_n. \tag{2.5}$$

Besides this recursion we also use a perturbed version. For given vectors $v_n \in \mathbb{R}^m$, $w_n \in \mathbb{R}^{sm}$ we define

$$\hat{\varepsilon}_n = \varepsilon_n + v_n, \quad \hat{\delta}_n = \delta_n + (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} w_n. \tag{2.6}$$

Inserting this into (2.4) we arrive at

$$\hat{\varepsilon}_{n+1} = [I + \mathbf{b}^T \mathbf{Z}_n (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} \mathbf{e}]\, \hat{\varepsilon}_n + \mathbf{b}^T \mathbf{Z}_n (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} \hat{r}_n + \hat{\rho}_n, \tag{2.7}$$

where

$$\hat{\rho}_n = \rho_n + v_{n+1} - v_n - \mathbf{b}^T \mathbf{Z}_n (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} w_n, \tag{2.8a}$$

$$\hat{r}_n = r_n - \mathbf{e}\, v_n + w_n. \tag{2.8b}$$

In the proof of our convergence results we shall sometimes use (2.7) instead of (2.5). This generalizes an idea used by Kraaijevanger [13], who considered (2.7) with $w_n = 0$ in his study on the implicit midpoint rule.

3. B-Convergence for Semi-Linear Problems

3.1. The Convergence Results

In this section we present some convergence results for the semi-linear problem (1.5) satisfying (1.6) with given constants $\alpha, \beta \in \mathbb{R}$. We assume that the function g is continuously differentiable. If the order condition $C(q)$ holds we shall tacitly assume that the solution U of (1.5) is $q+2$ times continuously differentiable. The formulas given in Section 2 can be applied with $f(t,u)=Qu+g(t,u)$ $(t\in\mathbb{R}, u\in\mathbb{R}^m)$. This function satisfies the one-sided Lipschitz condition (1.2) with constant $\nu=\alpha+\beta$.

The stability function of the Runge-Kutta method (1.3) will be denoted by R,

$$R(z)=1+b^T z(I-Az)^{-1}e \quad (z\in\mathbb{C}).$$

In order for the method to be stable uniformly on the class of (nonlinear) problems (1.5) satisfying (1.6) we do not need B-stability. It will be assumed that the method is A-stable,

$$|R(z)|\leq 1 \quad \text{for all} \quad z\in\mathbb{C}^-=\{\zeta\in\mathbb{C}: \operatorname{Re}\zeta\leq 0\}.$$

Similarly we shall not need the BSI- and BS-stability concepts of Frank, Schneid and Ueberhuber [8], but only their linear, scalar counterparts.

Definition 3.1: The Runge-Kutta method (1.3) is called *ASI-stable* if the matrix $I-Az$ is regular for all $z\in\mathbb{C}^-$, and $(I-Az)^{-1}$ is uniformly bounded for $z\in\mathbb{C}^-$.

Definition 3.2: The Runge-Kutta method (1.3) is said to be *AS-stable* if $I-Az$ is regular for all $z\in\mathbb{C}^-$, and $b^T z(I-Az)^{-1}$ is uniformly bounded for $z\in\mathbb{C}^-$.

We note that the concept of AS-stability has been introduced by Crouzeix and Raviart [4]; they called a method $\bar{A}$-stable if it is A-stable and AS-stable.

Let $q\in\mathbb{N}$ be such that the order condition $C(q)$ holds, and let the vector $k\in\mathbb{R}^s$ be defined by (2.3). Defining the rational function ψ by

$$\psi(z)=[b^T(I-Az)^{-1}e]^{-1}[b^T(I-Az)^{-1}k] \quad (z\in\mathbb{C}) \tag{3.1}$$

we state the following result.

Theorem 3.3: *Let $\alpha,\beta\in\mathbb{R}$ be given. Assume the Runge-Kutta method* (1.3) *is A-stable, AS-stable and ASI-stable. Then we have for the class of problems* (1.5) *satisfying* (1.6) *the* (*optimal*) *B-convergence result*

$$|\varepsilon_N|\leq C\tau^p \quad (0<\tau\leq\bar{\tau})$$

with order

(*a*) $p=q$ *if* $B(q), C(q)$,

(*b*) $p=q+1$ *if* $B(q+1)$, $C(q)$ *and* ψ *is uniformly bounded on* $\mathbb{C}^-$.

In this theorem the constant $\bar{\tau}$ only depends on α,β and the coefficients of the method, and C only depends on α, β, T, the coefficients of the method and bounds for the derivatives of U.

For a large class of methods (Gauss, Radau IA and IIA) the result of part (a) has already been proved in [9], even for the more general problems (1.1) which satisfy (1.2). Since most A-stable methods which are used in practice are ASI- and AS-stable as well (see Section 4), part (a) is applicable to a larger class of methods than those considered in [9].

Part (b) of the theorem shows that the global order of a method can be higher than its stage order (which equals q if $B(q+1)$ and $C(q)$ holds). This result has been proved for the implicit midpoint rule in [13], [17] (for the problems (1.1) satisfying (1.2)). A surprising corollary of part (b) is that for $s\geq 2$ there are s-stage methods with a higher global order than the s-stage Gauss method. For instance, for the Radau IIA methods we have $p = s+1$ (see Section 4), whereas it has been shown in [5] that the global order for the Gauss methods on the semi-linear problems is only $p = s$.

The conditions on the methods we imposed in Theorem 3.3 will be analyzed in Section 4.

3.2. The Proof of Theorem 3.3

For proving Theorem 3.3 we shall first derive some technical results, and then proceed with the actual proof.

If $\phi : \mathbb{C} \to \mathbb{C}$ is a rational function and $Z \in L(\mathbb{R}^m)$, the operator $\phi(Z) \in L(\mathbb{R}^m)$ is defined by $\phi(Z) = [\phi_1(Z)]^{-1} \phi_2(Z)$ (provided $\phi_1(Z)$ is regular) where ϕ_1, ϕ_2 are polynomials without common factors such that $\phi(z) = \phi_1(z)^{-1} \phi_2(z)$ (whenever $z \in \mathbb{C}$, $\phi(z)$ is defined). If $\phi_1(Z)$ is regular we shall say that $\phi(Z)$ exists.

A proof of the following result, essentially due to J. von Neumann, can be found in [10].

Lemma 3.4: *Let* $\omega \in \mathbb{R}$ *and let* ϕ *be a rational function without poles in* $\{z \in \mathbb{C} : \mathrm{Re}\, z \leq \omega\}$. *Suppose* $Z \in L(\mathbb{R}^m)$, $\langle u, Zu \rangle \leq \omega |u|^2$ *(for all* $u \in \mathbb{R}^m$*). Then* $\phi(Z)$ *exists and*

$$|\phi(Z)| \leq \sup \{|\phi(z)| : z \in \mathbb{C},\ \mathrm{Re}\, z \leq \omega\}.$$

In the rest of this section we shall write Z for τQ, $\mathbf{Z} = I_s \otimes Z$ and $\mathbf{Z}_n$ will be as in Section 2. All constants $\bar{\tau}_i, \gamma_i$ appearing further on will only depend on α, β and the coefficients of method (1.3), and the constants C_i will only depend on α, β, T, the coefficients of the method and bounds for the derivatives of the solution U of (1.5).

Lemma 3.5: *Suppose the Runge-Kutta method is ASI-stable. Then there are positive constants* $\bar{\tau}_1, \gamma_1$ *such that* $\mathbf{I} - \mathbf{A}\mathbf{Z}_n$ *is regular and*

$$\|(\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1}\| \leq \gamma_1 \ \textit{for} \ 0 < \tau \leq \bar{\tau}_1.$$

Proof: We first prove the statement of the lemma with $\mathbf{Z}_n$ replaced by $\mathbf{Z}$. Let $V(z) = (v_{ij}(z)) = I - Az$ $(z \in \mathbb{C})$ and $W(z) = (w_{ij}(z)) = V(z)^{-1}$ (if $z \in \mathbb{C}$, $V(z)$ is regular). From our assumption it follows that there exists an $\omega > 0$ such that $V(z)$ is regular for $\mathrm{Re}\, z \leq \omega$, and all entries $w_{ij}(z)$ of $W(z)$ are uniformly bounded for $\mathrm{Re}\, z \leq \omega$.

Let $\bar{\tau}_0$ be such that $\bar{\tau}_0 \beta \leq \omega$. By applying Lemma 2.4.6 in [11] it can be seen that for any $\tau \in (0, \bar{\tau}_0]$ the matrix $V(\mathbf{Z}) \in L(\mathbb{R}^{sm})$ is regular and $V(\mathbf{Z})^{-1} = W(\mathbf{Z}) = (\mathbf{I} - \mathbf{A}\mathbf{Z})^{-1}$ is a block-matrix with blocks $w_{ij}(Z) \in L(\mathbb{R}^m)$ $(1 \leq i, j \leq s)$. From Lemma 3.4 it follows that there are $\gamma_{ij} > 0$ such that $|w_{ij}(Z)| \leq \gamma_{ij}$ $(0 < \tau \leq \bar{\tau}_0)$, and hence there is a $\gamma_0 > 0$ such that $\|W(\mathbf{Z})\| \leq \gamma_0$ $(0 < \tau \leq \bar{\tau}_0)$.

In order to prove the actual statement of the lemma we note that $|Z_i^{(n)} - Z| \leq \tau\alpha$ since the function g has a Lipschitz constant α. Therefore $\|\mathbf{Z}_n - \mathbf{Z}\| \leq \tau\alpha$, $\|(\mathbf{I} - \mathbf{A}\mathbf{Z}_n) - (\mathbf{I} - \mathbf{A}\mathbf{Z})\| \leq \tau\alpha_1$ with $\alpha_1 = \alpha \|\mathbf{A}\|$. It follows that $\mathbf{I} - \mathbf{A}\mathbf{Z}_n$ is regular and $\|(\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1}\| \leq \gamma_0/(1 - \gamma_0 \alpha_1 \tau)$ provided $\alpha_1 \tau < \gamma_0^{-1}$. We thus can take $\bar{\tau}_1 > 0$ such that $\alpha_1 \bar{\tau}_1 < \gamma_0^{-1}$, and define $\gamma_1 = \gamma_0/(1 - \gamma_0 \alpha_1 \bar{\tau}_1)$. □

Lemma 3.6: *Suppose the Runge-Kutta method is A-stable and ASI-stable. Then there exist positive constants $\bar{\tau}_2, \gamma_2$ such that*

$$|I+\mathbf{b}^T \mathbf{Z}_n(\mathbf{I}-\mathbf{A}\mathbf{Z}_n)^{-1}\mathbf{e}| \leq 1+\gamma_2\tau \text{ for } 0<\tau\leq\bar{\tau}_2.$$

Proof: As in the proof of Lemma 3.5 we can show that

$$|I+\mathbf{b}^T \mathbf{Z}(\mathbf{I}-\mathbf{A}\mathbf{Z})^{-1}\mathbf{e}| \leq 1+\gamma_0'\tau \quad (\text{for } 0<\tau\leq\bar{\tau}_0')$$

for certain $\gamma_0', \bar{\tau}_0'$ which only depend on β. Further we have

$$\mathbf{Z}_n(\mathbf{I}-\mathbf{A}\mathbf{Z}_n)^{-1} = \mathbf{Z}(\mathbf{I}-\mathbf{A}\mathbf{Z})^{-1} + (\mathbf{I}-\mathbf{A}\mathbf{Z})^{-1}(\mathbf{Z}_n-\mathbf{Z})(\mathbf{I}-\mathbf{A}\mathbf{Z}_n)^{-1}, \tag{3.2}$$

which can easily be derived by noticing that $\mathbf{Z}(\mathbf{I}-\mathbf{A}\mathbf{Z})^{-1}=(\mathbf{I}-\mathbf{Z}\mathbf{A})^{-1}\mathbf{Z}$ and $\mathbf{A}\mathbf{Z}=\mathbf{Z}\mathbf{A}$. By using the bounds for $\|(\mathbf{I}-\mathbf{A}\mathbf{Z})^{-1}\|$, $\|\mathbf{Z}_n-\mathbf{Z}\|$, $\|(\mathbf{I}-\mathbf{A}\mathbf{Z}_n)^{-1}\|$ as given in the proof of Lemma 3.5 the proof now easily follows. □

In the same way one can prove the following result.

Lemma 3.7: *Suppose the Runge-Kutta method is AS-stable and ASI-stable. Then there are positive constants $\bar{\tau}_3, \gamma_3$ such that*

$$|\mathbf{b}^T \mathbf{Z}_n(\mathbf{I}-\mathbf{A}\mathbf{Z}_n)^{-1} r| \leq \gamma_3 \|r\| \text{ for all } r\in\mathbb{R}^{sm} \text{ and } 0<\tau\leq\bar{\tau}_3.$$

We shall now start with the actual proof of part (a) of Theorem 3.3, thus assuming that the method is A-, ASI- and AS-stable and satisfies $B(q), C(q)$. This proof is essentially the same as the B-convergence proof of Frank, Schneid and Ueberhuber [9] who considered a more restricted class of methods but the more general problem (1.1) satisfying (1.2).

Consider the recursion (2.5) for the global error. Application of the Lemmas 3.6, 3.7 gives

$$|\varepsilon_{n+1}| \leq (1+\gamma_2\tau)|\varepsilon_n| + \gamma_3\|r_n\| + |\rho_n| \quad (0<\tau\leq\bar{\tau})$$

with $\bar{\tau}=\min\{\bar{\tau}_2, \bar{\tau}_3\}$. Further we know there are constants $C_1, C_2>0$ such that $\|r_n\|\leq C_1\tau^{q+1}$, $|\rho_n|\leq C_2\tau^{q+1}$ (see Section 2). The order q result of part (a) now follows in a standard way.

Next we assume in addition that $B(q+1)$ holds and the function ψ (defined by (3.1)) is uniformly bounded on $\mathbb{C}^-$.

From Lemma 3.4 it can be seen that there are constants $\bar{\tau}_4, \gamma_4>0$ such that $\psi(Z)$ exists and $|\psi(Z)|\leq\gamma_4$ for $0<\tau\leq\bar{\tau}_4$ (which we assume in the following).

In order to prove part (b) of Theorem 3.3 we use the perturbed error scheme (2.7) with $v_n=\psi(Z)\tau^{q+1}U^{(q+1)}(t_n)$ and $w_n=\mathbf{e}v_n-\mathbf{k}\tau^{q+1}U^{(q+1)}(t_n)$ where $\mathbf{k}=k\otimes I_m$, is defined by (2.3). With these choices we have

$$\hat{r}_n = r_n - \mathbf{k}\tau^{q+1}U^{(q+1)}(t_n),$$

$$\hat{\rho}_n = \mathbf{b}^T\mathbf{Z}_n(\mathbf{I}-\mathbf{A}\mathbf{Z}_n)^{-1}[\mathbf{k}-\mathbf{e}\psi(Z)]\tau^{q+1}U^{(q+1)}(t_n) + \\ +\psi(Z)\tau^{q+1}[U^{(q+1)}(t_{n+1})-U^{(q+1)}(t_n)]+\rho_n.$$

We have $\|\hat{r}_n\|\leq C_5\tau^{q+2}$ $(0<\tau\leq\bar{\tau}_5)$ for certain $C_5, \bar{\tau}_5>0$. By using formula (3.2), the relation $b^T\zeta(I-A\zeta)^{-1}$ $[k-e\psi(\zeta)]=0$ (for all $\zeta\in\mathbb{C}$) and Lemma 3.5 it can be seen that there are constants $C_6, \bar{\tau}_6>0$ such that $|\hat{\rho}_n|\leq C_6\tau^{q+2}$ $(0<\tau\leq\bar{\tau}_6)$.

The proof of part (b) can now proceed as the proof of part (a). We get

$$|\hat{\varepsilon}_N| \leq C_7 \tau^{q+1} \quad (0 < \tau \leq \bar{\tau}_7)$$

for certain $C_7, \bar{\tau}_7 > 0$, and since

$$|\hat{\varepsilon}_N - \varepsilon_N| \leq \gamma_4 \tau^{q+1} |U^{(q+1)}(t_n)|$$

the order $q+1$ result follows.

3.3. Remarks on Extensions of Theorem 3.3

Remark 3.8: The conclusions of Theorem 3.3 also hold if the function g only satisfies a local Lipschitz condition

$$|g(t,\tilde{u}) - g(t,u)| \leq \alpha |\tilde{u} - u| \quad (\text{for } (t,\tilde{u}),(t,u) \in \mathfrak{D})$$

where $\mathfrak{D} \subset \mathbb{R}^{m+1}$ is an open set containing $\{(t, U(t)) : 0 \leq t \leq T\}$, instead of the global condition (1.6).

This can be shown in a standard way, by considering a function $\bar{g} : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ which coincides with g on $\mathfrak{D}$ and satisfies a global Lipschitz condition, and proving convergence for the problem

$$\dot{U}(t) = QU(t) + \bar{g}(t, U(t)), \; U(0) = u_0.$$

Remark 3.9: For convenience we have considered thus far only constant stepsizes. Convergence results for the semi-linear problems can also be given for variable stepsizes τ_n where

$$\tau_n = t_{n+1} - t_n, \quad 0 < \tau_n \leq \tau \tag{3.3}$$

with $t_n \in [0, T]$, $t_0 = 0$, $t_N = T$.

It is easily seen from Section 2 that the recursions (2.4) – (2.8) for the global error can now be used with $\mathbf{Z}_n = \mathbf{Q}_n + \mathbf{G}_n$ where $\mathbf{Q}_n = \tau_n (I_s \otimes Q)$ and $\mathbf{G}_n \in L(\mathbb{R}^{sm})$ is a block-diagonal matrix with blocks $G_i^{(n)} \in L(\mathbb{R}^m)$ on the diagonal satisfying $|G_i^{(n)}| \leq \tau_n \alpha$.

Examination of the proof of Theorem 3.3 shows that the conclusions of part (a) remain valid for the variable step sizes.

For part (b) the situation is more complicated. Consider the perturbed error scheme (2.7) with $v_n = \psi(\tau_n Q) \tau_n^{q+1} U^{(q+1)}(t_n)$ and $w_n = \mathbf{e} v_n - \mathbf{k} \tau_n^{q+1} U^{(q+1)}(t_n)$. Then (cf. (2.8)) $\hat{r}_n = r_n - \mathbf{k} \tau_n^{q+1} U^{(q+1)}(t_n)$,

$$\begin{aligned} \hat{\rho}_n = \mathbf{b}^T \mathbf{Z}_n (\mathbf{I} - \mathbf{A}\mathbf{Z}_n)^{-1} [\mathbf{k} - \mathbf{e}\psi(\tau_n Q)] \tau_n^{q+1} U^{(q+1)}(t_n) + \\ + \psi(\tau_n Q) \tau_n^{q+1} [U^{(q+1)}(t_{n+1}) - U^{(q+1)}(t_n)] + \rho_n + \\ + [\psi(\tau_{n+1} Q) - \psi(\tau_n Q)] \tau_{n+1}^{q+1} U^{(q+1)}(t_{n+1}) + \psi(\tau_n Q) [\tau_{n+1}^{q+1} - \tau_n^{q+1}] U^{(q+1)}(t_{n+1}). \end{aligned}$$

As in the proof of Theorem 3.3, part (b), the first three terms on the right hand side can be bounded in norm by $C_6 \tau^{q+2}$ provided $0 < \tau_n \leq \tau \leq \bar{\tau}_6$. Further we have

$$|\tau_{n+1}^{q+1} - \tau_n^{q+1}| \leq q \tau^q |\tau_{n+1} - \tau_n|,$$

$$|\psi(\tau_{n+1} Q) - \psi(\tau_n Q)| \leq \sup\{|\psi(\tau_{n+1} z) - \psi(\tau_n z)| : z \in \mathbb{C}, \operatorname{Re} z \leq \beta\} \text{ (see Lemma 3.4).}$$

By a tedious calculation, using the assumption that ψ is uniformly bounded on $\mathbb{C}^-$, it can be shown that there are constants $\gamma_8, \bar{\tau}_8 > 0$ such that

$$\sup\{|\psi(\tau_{n+1} z) - \psi(\tau_n z)| : z \in \mathbb{C}, \operatorname{Re} z \leq \beta\} \leq \gamma_8 \tau_{n+1}^{-1} |\tau_{n+1} - \tau_n|$$
$$(0 < \tau_n, \tau_{n+1} \leq \tau \leq \bar{\tau}_8).$$

Hence we have

$$|\hat{\rho}_n| \leq C_6 \tau^{q+2} + C_9 \tau^q |\tau_{n+1} - \tau_n| \quad (0 \leq \tau_n, \tau_{n+1} \leq \bar{\tau}_9)$$

for certain constants $C_9, \bar{\tau}_9 > 0$.

It follows that, under the assumptions of Theorem 3.3, part (b), we have the optimal B-convergence result

$$|\varepsilon_n| \leq C \tau^{q+1} + C' \tau^q \sum_{n=0}^{N-1} |\tau_{n+1} - \tau_n| \quad (\text{for } 0 < \tau_n \leq \tau \leq \bar{\tau}) \tag{3.4}$$

with $\bar{\tau}$ only depending on α, β and the coefficients of the Runge-Kutta method, and C, C' only depending on α, β, T, the coefficients of the method and bounds for the derivatives of U.

This upperbound for the global error shows that the order $q+1$ result of Theorem 3.3, part (b), can remain valid for variable stepsizes (cf. also [13]). For instance, if the number of changes in sign in the series $\{\tau_{n+1} : n = 0, 1, \ldots, N-1\}$ is less than a fixed number M (independent of N), then

$$\sum_{n=0}^{N-1} |\tau_{n+1} - \tau_n| \leq M \tau,$$

and thus we get from (3.4)

$$|\varepsilon_N| \leq (C + MC') \tau^{q+1} \quad (0 < \tau_n \leq \tau \leq \bar{\tau}).$$

4. Some Examples

In this section we will study the stability properties, introduced in the previous sections, for certain interesting families of Runge-Kutta methods. This will be accomplished by presenting some general results which are sufficient conditions for these stability properties.

Let $\sigma(A)$ denote the spectrum of a matrix A and define the following regions in the complex plane $\mathbb{C}$:

$$\begin{aligned} \mathbb{C}^- &= \{z \in \mathbb{C} : \operatorname{Re}(z) \leq 0\} \\ \mathbb{C}_0^+ &= \{z \in \mathbb{C} : z = 0 \text{ or } \operatorname{Re}(z) > 0\} \\ I_0 &= \{z \in \mathbb{C} : z \neq 0, \operatorname{Re}(z) = 0\} \end{aligned}$$

then the following results hold. We note that Lemma 4.2 is equivalent to a result that can be found in [4].

Lemma 4.1: *$I - Az$ regular on $\mathbb{C}^- \Leftrightarrow \sigma(A) \subset \mathbb{C}_0^+$.*

Lemma 4.2: $\sigma(A) \cap \mathbb{C}^- = \emptyset \Rightarrow AS-$*stability.*

Proof: A consequence of Lemma 4.1 is that $\sigma(A) \cap \mathbb{C}^- = \emptyset$ implies $I - Az$ is regular on $\mathbb{C}^-$ and furthermore that A can have no eigenvalues at zero. Hence the characteristic polynomial of $I - Az$ is of degree s which is greater than or equal to the degree of the polynomial of each of the s components in the numerator of $b^T z(I - Az)^{-1}$. □

Lemma 4.3: *If A is regular, or A has a simple eigenvalue at zero, then* $\sigma(A) \subset \mathbb{C}_0^+ \Rightarrow ASI$*-stability.*

Proof: If A has at most a single eigenvalue at zero then the characteristic polynomial of $I - Az$ is of degree $s-1$ which is greater than or equal to the degrees of the polynomials of the numerators of the s^2 rational functions, which are the elements of $(I - Az)^{-1}$. The proof concludes from Lemma 4.1 and Definition 3.2. □

The sufficient conditions expressed in Lemma 4.3 can actually be weakened to allow a multiplicity of eigenvalues at zero as long as the matrix A has a special structure. However, the maximum order of consistency of the family of methods with A having t eigenvalues equal to zero is $2s - t$ and there do not seem to be any methods in this class with practical significance. Hence for the rest of this paper we will usually assume that A has at most one eigenvalue at zero.

The properties of ASI-stability and AS-stability can be related very simply by the following Lemma (which is similar to a result given in [12]).

Lemma 4.4: *If there exists a vector d such that $b^T = d^T A$ then ASI-stability $\Rightarrow AS$-stability.*

Proof: Consider the vector function ϕ defined by

$$\phi(z) = b^T z(I - Az)^{-1}, \ (z \in \mathbb{C}).$$

If there exists a vector d such that $b^T = d^T A$, then

$$\phi(z) = d^T (I - Az)^{-1} - d^T$$

which is uniformly bounded on $\mathbb{C}^-$, and hence AS-stable, if the method is ASI-stable. □

Before we analyse optimal B-convergence (for semi-linear problems) in greater depth, it is interesting to ask whether AS-stability and/or ASI-stability are necessary conditions for A-stability or for the uniform boundedness of ψ on $\mathbb{C}^-$. This question can be partially answered by the following method

$$\begin{array}{c|ccc}
1 & 1 & \theta_1 & -\theta_1 \\
\theta_2 & \theta_2 & 0 & 0 \\
\theta_2 & \theta_2 & 0 & 0 \\
\hline
 & 1 & 1 & -1
\end{array}$$

One can easily show that the method is not AS-stable and not ASI-stable. However

$$R(z) = \frac{1}{1-z}, \ \psi(z) = \frac{z}{2(1-z)},$$

so that the method is A-stable and ψ is uniformly bounded on $\mathbb{C}^-$. This method, however, is reducible and is equivalent to the implicit Euler method. The situation is further complicated if $\sigma(A)\notin\mathbb{C}_0^+$. In this case it is possible to construct a method which is A-stable, but neither AS-stable or ASI-stable at which the determinant of $I-zA$ vanishes at some point in $\mathbb{C}^-$. For example, consider the method given by

$$\begin{array}{c|cc} a & a & 0 \\ x+\frac{1}{2} & x & \frac{1}{2} \\ \hline & b_1 & b_2 \end{array}$$

with $a<0$ and $b_1=(x+\frac{1}{2}-a)^{-1}x$, $b_2=(x+\frac{1}{2}-a)^{-1}(\frac{1}{2}-a)$. In the solution of the linear test equation

$$y'=qy,\ \operatorname{Re}(q)<0$$

the above method has no solution for the $y_i^{(n)}$ if $hq=1/a$, but the stability function is given by

$$R(z)=\frac{1+z/2}{1-z/2}.$$

Hence in order to avoid such complications we will always assume that either $\sigma(A)\subset C_0^+$ or $\sigma(A)\cap\mathbb{C}^-=\emptyset$.

Let $\psi(z)=P(z)/Q(z)$, where P and Q are polynomials of degree at most $s-1$. In order to simplify the study of the boundedness of $\psi(z)$ (where $\psi(z)=(b^T(I-Az)^{-1}e)^{-1}(b^T(I-Az)^{-1}k)$) we will assume that $Q(z)$ is of degree $s-2$ or more and that $\sigma(A)\cap\mathbb{C}^-=\emptyset$. Furthermore, we will concentrate our study on Runge-Kutta methods that are A-stable. Hence $|R(z)|<1$ for all $z\in\mathbb{C}$ such that $\operatorname{Re}(z)<0$, so that $|R(z)|$ take its maximum value of 1 on the imaginary axis or as $z\to-\infty$.

Therefore, in the case that Q is of degree $s-1$, ψ will be bounded if $R(z)\neq 1$ on I_0, while if Q is of degree $s-2$ (that is $b^TA^{-1}e=0$) it is easily shown that ψ will be bounded on $\mathbb{C}^-$ if $b^TA^{-1}k=0$ and $R(z)\neq 1$ on I_0. Thus from Theorem 3.3 (part b) and the above discussion we see that if $C(q)$ and $B(q+1)$ hold the following conditions are sufficient for a Runge-Kutta method to be optimally B-convergent (for semi-linear problems) of order $q+1$.

I Degree of Q is $s-1$:
A-stability, $\sigma(A)\cap\mathbb{C}^-=\emptyset$, $R(z)\neq 1$ on $I_0, b^TA^{-1}e\neq 0$.

II Degree of Q is $s-2$:
A-stability, $\sigma(A)\cap\mathbb{C}^-=\emptyset$, $R(z)\neq 1$ on I_0, $b^TA^{-1}e=b^TA^{-1}k=0$.

We will now investigate the property of the boundedness of ψ on $\mathbb{C}^-$ for two interesting classes of methods; singly-implicit methods and methods of order $2s-2$ or more.

For any s-stage Runge-Kutta method let $R(z)=N(z)/D(z)$, where D is of degree s and N is of degree at most s, and define

$$E(y)=|N(iy)|^2-|D(iy)|^2,\ y\in\mathbb{R}.$$

Then Nørsett [14] has shown that for any Runge-Kutta method of order $2s-2$

$$E(y)=\theta y^{2s}, \text{ for all } y\in\mathbb{R}, \tag{4.1}$$

where θ depends only on the method. Thus if $\theta\neq 0$ then $E(y)\neq 0$ for $y\neq 0$, and hence $|R(z)|\neq 1$ on $I_0\cup\{\infty\}$. Using this fact we will give a characterization of almost all methods of order $2s-2$ satisfying Theorem 3.3 (b).

Butcher [3] has given an elegant characterization of all A-stable implicit Runge-Kutta methods of order $2s-2$ or more. For such methods we have

$$R(z)=\frac{w_0 N_0(z)+w_1 N_1(z)+w_2 N_2(z)}{w_0 D_0(z)+w_1 D_1(z)+w_2 D_2(z)} \tag{4.2}$$

where

$$N_0(z)=\sum_{j=0}^{s}\frac{(2s-j)!}{j!(s-j)!}z^j,\quad D_0(z)=N_0(-z)$$

$$N_1(z)=2\sum_{j=0}^{s-1}\frac{(2s-1-j)!}{j!(s-1-j)!}z^j,\quad D_1(z)=2s\sum_{j=0}^{s}\frac{(2s-1-j)!}{j!(s-j)!}(-z)^j$$

$$N_2(z)=2\sum_{j=0}^{s-2}\frac{(2s-2-j)!}{j!(s-2-j)!}z^j,\quad D_2(z)=2s(s-1)\sum_{j=0}^{s}\frac{(2s-2-j)!}{j!(s-j)!}(-z)^j.$$

Butcher has shown that a method whose stability function is given by (4.2) is A-stable iff

$$w_0+w_1+w_2=1,\quad w_2<2-1/s,\quad w_0\leq 1. \tag{4.3}$$

Finally, we note that θ (in (4.1)) is zero iff $|R(iy)|=1$ for all $y\in\mathbb{R}$, and this can only be true iff $R(z)=1$ or

$$R(z)R(-z)=1,\quad \text{for all } z\in\mathbb{C}. \tag{4.4}$$

Ignoring the trivial case one can show that (4.4) holds iff $w_0=1$ in which case $w_1=-w_2$ and

$$R(z)=\frac{N_0(z)+2sw_1 N_3(z)}{N_0(-z)+2sw_1 N_3(-z)}, \tag{4.5}$$

where

$$N_3(z)=\sum_{j=0}^{s-1}\frac{(2s-2-j)!}{j!(s-1-j)!}z^j.$$

Thus we have the following result:

Theorem 4.5: *Any Runge-Kutta method of order $2s-2$ whose stability function is given by* (4.2) *with*

$$w_0+w_1+w_2=1,\quad w_2<2-1/s,\quad w_0<1$$

and where $C(q)$ and $B(q+1)$ hold and $\sigma(A)\cap C^-=\emptyset$, is optimally B-convergent (for semi-linear problems) of order $q+1$.

Remark:

(i) By choosing $w_1=1$, $w_2=0$, $w_0=0$ or $w_2=1$, $w_1=0$, $w_0=0$ we see that the Runge-Kutta methods whose stability functions corresponds to the first two subdiagonals of the Padé table have the property that ψ is bounded on $\mathbb{C}^-$.

(ii) The Radau IIA methods with $s \geq 2$ are optimally B-convergent with order $s+1$.

(iii) We have no general results about the order of optimal B-convergence for methods of order $2s-2$ or more satisfying (4.4). However, the family of Gauss methods of order $2s$ belong to this class and it is known (see [13], for example) that the implicit midpoint rule is optimally B-convergent of order 2, and recently Dekker et al. [5] have shown that all s-stage Gauss methods with $s \geq 2$ are not optimally B-convergent of order $s+1$. The proof of this result for the case that s is even is very simple.

Suppose that $C(q)$ and $B(q+1)$ hold and that A is nonsingular then it can easily be seen from Lemma 3.1 in [5] that if $b^T A^{-1} e=0$, the method cannot be optimally B-convergent of order $q+1$ if $b^T A^{-1} k \neq 0$. (We note that Dekker et al. [5] have a more general result.) Using this fact we will derive a general result about the order of optimal B-convergence for collocation methods satisfying $C(s)$ and $B(s+1)$. However, we will first derive a result which will be of help when studying the order of optimal B-convergence of singly-implicit methods.

Theorem 4.6: *For a Runge-Kutta method, with distinct c_j and a nonsingular matrix A, satisfying $C(s-1)$ and $B(s)$ we have*

$$b^T A^{-1} k = -\frac{1}{s!}(p(1)+(1-b^T A^{-1} e) e_1^T V^{-1} c^s)$$

where $p(x)=\prod_{j=1}^{s}(x-c_j)$ for $x \in \mathbb{R}$ and V is the $s \times s$ matrix whose (i,j) element is c_i^{j-1}.

Proof: Let $A=V\bar{A}V^{-1}$ and $\bar{b}^T=b^T V$. Then $B(s)$ is equivalent to

$$\bar{b}^T=(1, 1/2, \ldots, 1/s).$$

Since the stage order is $s-1$

$$k=\frac{1}{(s-1)!}\left(\frac{c^s}{s}-Ac^{s-1}\right)$$

and hence

$$b^T A^{-1} k=\frac{1}{(s-1)!}(\bar{b}^T \bar{A}^{-1} V^{-1} c^s/s-b^T c^{s-1})=\frac{1}{s!}(\bar{b}^T \bar{A}^{-1} V^{-1} c^s-1).$$

Burrage [1] has shown that for methods satisfying the condition stated in the theorem

$$\bar{b}^T \bar{A}^{-1}=e^T+(b^T A^{-1} e-1) e_1^T. \tag{4.6}$$

In addition one can easily show that

$$1-e^T V^{-1} c^s=p(1). \tag{4.7}$$

Hence

$$b^T A^{-1} k=-\frac{1}{s!}(p(1)+(1-b^T A^{-1} e) e_1^T V^{-1} c^s). \quad \square$$

Theorem 4.7: *For a Runge-Kutta method with distinct c_j and a nonsingular matrix A satisfying $C(s)$ and $B(s+1)$ we have*

$$b^T A^{-1} e = 1 \Leftrightarrow b^T A^{-1} k = 0 \Leftrightarrow p(1) = 0,$$

where $p(x) = \prod_{j=1}^{s} (x - c_j)$ *for* $x \in \mathbb{R}$.

Proof: Let V and $\bar{b}^T$ be as defined in Theorem 4.6 and let W be the $s \times s$ matrix whose (i, j) element is c_i^j/j. Then it is easily seen that $C(s)$ is equivalent to

$$A = W V^{-1}.$$

Since the stage order is s

$$k = \frac{1}{s!}\left(\frac{c^{s+1}}{s+1} - A c^s\right)$$

and hence

$$\begin{aligned} s!\, b^T A^{-1} k &= \bar{b}^T W^{-1} \frac{c^{s+1}}{s+1} - b^T c^s \\ &= -\frac{1}{s+1}(1 - \bar{b}^T W^{-1} c^{s+1}), \text{ since } B(s+1) \text{ holds} \\ &= -\frac{1}{s+1}(1 - e^T V^{-1} c^s) \\ &= -p(1)/(s+1), \text{ from (4.7).} \end{aligned}$$

But since the stage order is s we have from Theorem 4.6 that

$$(b^T A^{-1} e - 1)\, e_1^T V^{-1} c^s = p(1)$$

and the result is proved. □

Corollary 4.8: *Any Runge-Kutta method satisfying* $C(s)$ *and* $B(s+1)$ *with* A *nonsingular and* $b^T A^{-1} e = 0$ *cannot have optimal B-convergence order* $s+1$.

Corollary 4.9 (*see* [5]): *The even stage Gauss methods are not optimally B-convergent of order* $s+1$.

The methods studied so far have had an order of consistency of $2s-2$ or more and since such methods also have a high stage order they would appear to be attractive propositions for solving stiff differential equations. However, such high order methods cannot in general be implemented efficiently (in comparison with linear multistep methods for example). Nevertheless, there is one class of Runge-Kutta methods (characterized by the Runge-Kutta matrix having a one-point spectrum) which can be efficiently implemented. Such methods are called singly-implicit (SIRK's), and their order and stability properties have been studied by, for example, Nørsett [14] and Burrage [2]. A distinction is usually made between SIRKs and DIRKs (in which the Runge-Kutta matrix is lower triangular with constant value on the diagonal) since their order properties are very different. For example, the maximum stage order of any DIRK is 1 (while a SIRK can have a stage order of s) and this can be achieved in the case of DIRKs by the implicit midpoint rule or by the following two stage family.

$$\begin{array}{c|cc} \lambda & \lambda & \\ 1-\lambda & 1-2\lambda & \lambda \\ \hline & 1/2 & 1/2 \end{array}$$

If $\lambda \geq 1/4$ one can show that the conditions in part (b) of Theorem 3.3 are satisfied. (Note that $\lambda = 1/4$ is a special case since $b^T A^{-1} e = b^T A^{-1} k = 0$.) Thus from this viewpoint DIRKs do not appear to be very attractive for solving stiff problems if high accuracy is required. SIRKs on the other hand do not suffer from this drawback and Burrage [2] has constructed families of SIRKs satisfying either $C(s-1)$ and $B(s)$ or $C(s)$ and $B(s+1)$. These methods appear very attractive especially if they are optimally B-convergent with order s or $s+1$.

The stability function of an s-stage SIRK of order s is given by

$$R(z) = (-1)^s \frac{\sum_{k=0}^{s} \lambda^k L_s^{(s-k)}\left(\frac{1}{\lambda}\right) z^k}{(1-\lambda z)^s}, \quad L_s(x) = \sum_{j=0}^{s} (-1)^j \binom{s}{j} x^j/j! \tag{4.8}$$

Burrage [2] has studied the A-stability of such methods and gives the following ranges of λ which produce A-stable methods (note that since $\sigma(A) \cap \mathbb{C}^- = \emptyset$ these methods are ASI-stable and AS-stable).

Table 1

s	λ
1	$\left[\frac{1}{2}, \infty\right]$
2	$\left[\frac{1}{4}, \infty\right)$
3	$\left[\frac{1}{3}, \frac{1}{2} + \frac{\sqrt{3}}{3} \cos\left(\frac{\pi}{18}\right)\right]$
4	$[0.39434, 1.28057]$
5	$[0.24651, 0.36180] \cup [0.42079, 0.47328]$

We now consider whether $R(z)$ can equal 1 for any $z \in I_0$ for the above ranges of λ. For $s = 1, 2$ and 3 one can easily show that $R(z) \neq 1$ on I_0 for the above ranges of λ. In the case with $s = 4$, $R(z) = 1$ on I_0 for some finite z iff there exists $z = ir$ such that

$$1 = r^2 (6\lambda^2 - 2\lambda + 1/6),$$

$$\frac{1}{2} - 4\lambda = r^2 \left(\frac{1}{24} - \frac{2}{3}\lambda + 3\lambda^2 - 4\lambda^3\right),$$

which is equivalent to requiring that λ be a zero of the polynomial

$$20x^3 - 8x^2 + x - 1/24.$$

Some computations show that there is no value for λ satisfying this polynomial which lies in range of values given in Table 1 for $s=4$.

Similarly, for the case $s=5$, $R(z)=1$ on I_0 for some finite z iff there exists $z=ir$ such that

$$1/2-5\lambda=r^2(1/24-5/6\lambda+5\lambda^2-10\lambda^3),$$
$$1-r^2(1/6-5/2\lambda+10\lambda^2)+r^4(1/120-5/24\lambda+5/3\lambda^2-5\lambda^3+5\lambda^4)=0.$$

This can be shown to be equivalent to the requirement that λ be a zero of the polynomial

$$792000x^6-504000x^5+116400x^4-11400x^3+300x^2+20x-1$$

and we have shown numerically that there is no zero of this polynomial which lies in the range of values given in Table 1 for $s=5$.

Hence all that remains is to check whether $\lim_{z\to-\infty} R(z)=1$ (or equivalently that $b^T A^{-1} e=0$).

From (4.8) it can be seen that

$$b^T A^{-1} e=0 \Leftrightarrow L_s(1/\lambda)=1.$$

Some numerical computations show that the only values for s and λ (assuming λ lies in the intervals given in Table 1) that must be considered are

$$\left.\begin{array}{l} s=2,\ \lambda=1/4 \\ s=3,\ \lambda=1/3 \\ s=4,\ \lambda=.39434 \\ s=5,\ \lambda=.42079. \end{array}\right\} \tag{4.9}$$

However, for SIRKs with order of consistency s whose stability function is given by (4.8) the degree of the denominator of ψ is at least $s-2$. Hence if $b^T A^{-1} e=0$, $R(z)\neq 1$ on I_0 and $b^T A^{-1} k=0$ then ψ will be uniformly bounded on $\mathbb{C}^-$ for all the A-stable methods. Since we are assuming that $C(s-1)$ and $B(s)$ hold we see from Theorem 4.6 that we can always make $b^T A^{-1} k=0$ if $b^T A^{-1} e=0$ (choose $c_1=0$, $c_s=1$, for example). Thus we can state

Theorem 4.10: *All A-stable SIRKs satisfying $C(s-1)$ and $B(s)$, for $s=1,\dots,5$, are optimally B-convergent with order s, except in the special case given by* (4.9). *Here, the values the abscissae must be chosen so that $p(1)+e_1^T V^{-1} c^s=0$, in order to have optimal B-convergence with order s.* □

For some very special choices of λ and the abscissae we can in fact obtain optimal B-convergence of order $s+1$ (the highest order possible). Burrage [2] has shown that if $L_{s+1}^{(1)}(1/\lambda)=0$ and $c_j/\lambda\ (j=1,\dots,s)$ are the zeros of $L_s(x)$ then $B(s+1)$ and $C(s)$ hold. The values of λ such that $L_{s+1}(1/\lambda)=0$ and λ lies in the range of values given in Table 1 are

$$s=1,\ \lambda=\frac{1}{2}$$

$$s=2,\ \lambda=(3+\sqrt{3})/6$$

$$s=3,\ \lambda=\frac{1}{2}+\frac{1}{3}\sqrt{3}\cos\left(\frac{\pi}{18}\right)$$

$$s=5,\ \lambda\approx 0.47328.$$

Furthermore, Wanner et al. [19] have shown that these are the only values of λ which give A-stability and a classical order of $s+1$. Thus these are the only values of λ which give optimal B-convergence of order $s+1$.

In conclusion the results of this section seem to confirm the use of SIRKs as appropriate methods for solving stiff differential equations and these theoretical results are backed up by some numerical work (see [6], for example) which illustrates the superiority of SIRKs over DIRKs for stiff problems when high accuracy is required.

References

[1] Burrage, K.: Stability and efficiency properties of implicit Runge-Kutta methods. Ph. D. Thesis, Dept. of Math., Univ. of Auckland, 1978.

[2] Burrage, K.: A special family of Runge-Kutta methods for solving stiff differential equations. BIT *18*, 22–41 (1978).

[3] Butcher, J. C.: On *A*-stable implicit Runge-Kutta methods. BIT *17*, 375–378 (1977).

[4] Crouzeix, M., Raviart, P. A.: Méthodes de Runge-Kutta. Unpublished lecture notes. Université de Rennes, 1980.

[5] Dekker, K., Kraaijevanger, J. F. B. M., Spijker, M. N.: The order of *B*-convergence of the Gaussian Runge-Kutta method. Computing (this issue).

[6] Dekker, K., Verwer, J. G.: Stability of Runge-Kutta methods for stiff nonlinear differential equations. Amsterdam: North-Holland 1984.

[7] Frank, R., Schneid, J., Ueberhuber, C. W.: The concept of *B*-convergence. SIAM J. Numer. Anal. *18*, 753–780 (1981).

[8] Frank, R., Schneid, J., Ueberhuber, C. W.: Stability properties of implicit Runge-Kutta methods. SIAM J. Numer. Anal. *22*, 497–514 (1985).

[9] Frank, R., Schneid, J., Ueberhuber, C. W.: Order results for implicit Runge-Kutta methods applied to stiff systems. SIAM J. Numer. Anal. *22*, 515–534 (1985).

[10] Hairer, E., Bader, G., Lubich, Ch.: On the stability of semi-implicit methods for ordinary differential equations. BIT *22*, 211–232 (1982).

[11] Hundsdorfer, W. H.: The numerical solution of nonlinear stiff initial value problems – an analysis of one-step methods. CWI Tract 12, Amsterdam 1985.

[12] Hundsdorfer, W. H., Spijker, M. N.: On the algebraic equations in implicit Runge-Kutta methods. SIAM J. Numerical Anal. (to appear).

[13] Kraaijevanger, J. F. B. M.: *B*-convergence of the implicit midpoint rule and the trapezoidal rule. BIT (to appear).

[14] Nørsett, S. P.: Semi-explicit Runge-Kutta methods. Report Math. and Comp. No. 6/74, Dept. of Math., Univ. of Trondheim, 1974.

[15] Nørsett, S. P.: *C*-polynomials for rational approximations to the exponential function. Numer. Math. *25*, 39–56 (1975).

[16] Prothero, A., Robinson, A.: On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. Math. Comp. *28*, 145–162 (1974).

[17] Stetter, H. J.: Zur *B*-Konvergenz der impliziten Trapez- und Mittelpunktregel, unpublished note.

[18] Verwer, J. G.: Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines. Proc. Dundee 1985, D. F. Griffiths (ed.), Pitman Publ. Co. (to appear).

[19] Wanner, G., Hairer, E., Nørsett, S. P.: Order stars and stability theorems. BIT *18*, 475–489 (1978).

K. Burrage
Dept. of Computer Science
University of Auckland
Auckland
New Zealand

W. H. Hundsdorfer
Centre for Mathematics and
Computer Science
Kruislaan 413
1098 SJ Amsterdam
The Netherlands

J. G. Verwer
Centre for Mathematics and
Computer Science
Kruislaan 413
1098 SJ Amsterdam
The Netherlands