Numer. Math. 50, 405–418 (1986)

Numerische Mathematik

Convergence and Order Reduction of Runge-Kutta Schemes Applied to Evolutionary Problems in Partial Differential Equations

J.M. Sanz-Serna[1], J.G. Verwer[2], and W.H. Hundsdorfer[2]

1 Departamento de Ecuaciones Funcionales, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain
[2] Centre for Mathematics and Computer Science, Kruislaan 413, NL-1098 SJ Amsterdam, The Netherlands

Summary. We address the question of convergence of fully discrete Runge-Kutta approximations. We prove that, under certain conditions, the order in time of the fully discrete scheme equals the conventional order of the Runge-Kutta formula being used. However, these conditions, which are necessary for the result to hold, are not natural. As a result, in many problems the order in time will be strictly smaller than the conventional one, a phenomenon called order reduction. This phenomenon is extensively discussed, both analytically and numerically. As distinct from earlier contributions we here treat explicit Runge-Kutta schemes. Although our results are valid for both parabolic and hyperbolic problems, the examples we present are therefore taken from the hyperbolic field, as it is in this area that explicit discretizations are most appealing.

Subject Classifications: AMS(MOS): 65X02, 65M10, 65M20; CR: G1.7.

1. Introduction

In many cases of practical interest evolutionary problems in partial differential equations (PDEs) are solved numerically by schemes which can be derived and implemented along the ideas of the well-known method of lines (MOL) approach. In this technique the numerical treatment of the PDE problem is thought of as consisting of two parts, viz. the discretization in space and the integration in time. In the space discretization the PDE is converted into a time continuous system of ordinary differential equations (ODEs) by finite difference or finite element techniques. This ODE system is then integrated in time by one of the many available integration schemes, e.g., a Runge-Kutta (RK) or a linear multistep scheme. To mention an example, which we discuss later in this paper, the classical 4-th order, 4 stage, explicit RK formula is sometimes used to integrate in time hyperbolic problems arising in fluid dynamics [8, 14].

In this paper we address the question of convergence of fully discrete RK approximations to the PDE solution. We prove, that under certain conditions,

the order in time of the fully discrete scheme equals the conventional order of the RK formula being used. However, these conditions, which are necessary for the result to hold, are not natural. As a result, in many problems the order in time will be strictly smaller than the conventional one, a phenomenon called order reduction.

In the MOL literature the phenomenon of order reduction has got very little attention. In fact, we are only aware of a few papers on this topic. The contributions [1] and [12] deal with *implicit* RK schemes. When applied to stiff systems of ODEs, not necessarily semi-discrete PDEs, these schemes also suffer from reduction of the order. This is the central issue of the B-convergence theory developed in [5]. In fact, the MOL paper [12] heavily relies on results from the B-convergence theory, whereas [1] is completely independent of it and concentrates on discretizations of ODEs in Banach space. As distinct from these contributions we here treat *explicit* RK schemes. Although our results are valid for both parabolic and hyperbolic problems, the examples we present are therefore taken from the hyperbolic field, as it is in this area that explicit discretizations are most appealing.

The contents of the paper is as follows. In Sect. 2 we collect preliminaries on the (linear) PDE problem, the space discretization, and the RK method. In Sect. 3 we examine the full *local* error. Here we present a detailed discussion of the order reduction phenomenon and explain that it will be present unless certain boundary conditions are fulfilled. It is emphasized, however, that these conditions are not natural to the problem but arise as constraints by the use of the Runge-Kutta method. Sect. 4 deals with the behaviour of the full *global* error. Following [1, 2, 12], we here discuss a special technique for transferring estimates of the local errors to the global one. This technique shows that the decrease in global order, although present, is not as marked as the standard convergence analysis would predict. Section 5 is devoted to a numerical illustration which nicely supports the theory. Then, in Sect. 6, we present a simple means for avoiding the reduction by transforming the given problem. Sect. 7 contains some final remarks and concludes the paper.

2. Preliminaries

2.1. Partial Differential Problem

We consider linear problems of the form

$$u_t = A_\Omega u + f_\Omega(t), \quad x \in \Omega, \quad 0 \leqq t \leqq T < \infty, \tag{2.1 a}$$

$$A_\Gamma u = f_\Gamma(t), \quad x \in \Gamma, \quad 0 \leqq t \leqq T, \tag{2.1 b}$$

$$u(x, 0) \text{ given}, \quad x \in \Omega, \tag{2.1 c}$$

where Ω is a spatial domain in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$, with boundary Γ and A_Ω denotes a linear, q-th order differential operator in Ω which differentiates the (possibly vector valued) unknown function u with respect to the spatial variables. The

linear differential operator A_Γ possesses order $\leqq q-1$, acts on the boundary Γ and serves to introduce the boundary conditions (2.1 b). Note that the inhomogeneous terms f_Ω, f_Γ and the coefficients of A_Ω, A_Γ may depend on x. This dependence is not however reflected in the notation.

2.2. *Space Discretization*

The discretization in space of the problem (2.1), by means of finite-elements or finite-differences, results in a Cauchy problem

$$\dot{U}_h = A_h U_h + f_h(t), \qquad 0 \leqq t \leqq T, \qquad U_h(0) \text{ given.} \tag{2.2}$$

Here h is the parameter of a grid in $\Omega \cup \Gamma$ and $U_h = U_h(t)$ is an m-dimensional real vector consisting of approximations to u at grid points. The time-independent matrix A_h originates from A_Ω, A_Γ and the vector $f_h(t)$ arises from the inhomogeneous terms of (2.1).

In what follows, we are interested in the behaviour of (2.2) as $h \to 0$. A crucial consideration is that, as the grid is refined, both the dimension m of (2.2) and the size of the entries of A_h will grow (these entries contain negative powers of the grid-spacing). As a result the problem (2.2) becomes increasingly stiffer for $h \to 0$. We assume that, for $h \to 0$, the entries of A_h grow like h^{-q}, with q the order in space of (2.1).

We denote by $u_h(t)$ the restriction of $u(x, t)$ to the spatial grid (or other suitable representation of u in that grid [10]) and by $\alpha_h(t)$ the space truncation error defined by

$$\alpha_h(t) = A_h u_h(t) + f_h(t) - \dot{u}_h(t). \tag{2.3}$$

We assume that (2.2) is consistent with (2.1) in the sense that, as $h \to 0$, $\max_{0 \leqq t \leqq T} \|\alpha_h(t)\| \to 0$. Throughout this paper, $\|\cdot\|$ denotes a chosen norm for m-dimensional vectors and the corresponding operator norm for $m \times m$ matrices. The space truncation error will enter the analysis in Sect. 3.

2.3. *The Runge-Kutta Scheme*

In order to numerically advance in time the solution of (2.2), we employ an explicit Runge-Kutta method. For our purpose it is convenient to describe this ODE method as it applies to a linear system of ODEs of the form

$$\dot{w} = M w + g(t), \tag{2.4}$$

with M a constant matrix. If w^n denotes the approximation to $w(n\tau)$ generated by the method with stepsize τ, the step $w^n \to w^{n+1}$ is performed by first computing recursively intermediate approximations $Y_1, Y_2, \ldots, Y_s$ through

$$Y_i = w^n + \tau \sum_{j=1}^{i-1} a_{ij} [M Y_j + g(t_n + c_j \tau)], \tag{2.5}$$

and then setting

$$w^{n+1}=w^n+\tau\sum_{i=1}^{s} b_i[MY_i+g(t_n+c_i\tau)]. \tag{2.6}$$

Here a_{ij}, b_i, c_i, $i=1,\ldots,s$, $j=1,\ldots,i-1$, are coefficients associated with the particular RK method being used and s is the number of stages. We denote by p the (classical) order of the method and assume that $\sum_{i=1}^{s} b_i=1$, $\sum_{j=1}^{i-1} a_{ij}=c_j$, $j=1,\ldots,s$. We also set $a_{s+1,j}=b_j$, $j=1,\ldots,s$ and $c_{s+1}=1$. The local accuracy of (2.5)–(2.6) will now be investigated in a manner related to that common in the B-convergence theory [5, 4] and slightly different from that based on Butcher trees.

We first consider a perturbed step $w^n\to w^{n+1}$

$$\tilde{Y}_i=\tilde{w}^n+\tau\sum_{j=1}^{i-1} a_{ij}[M\tilde{Y}_j+g(t_n+c_j\tau)]+r_i, \tag{2.7}$$

$$\tilde{w}^{n+1}=\tilde{w}^n+\tau\sum_{i=1}^{s} b_i[M\tilde{Y}_i+g(t_n+c_j\tau)]+r_{s+1}, \tag{2.8}$$

where the residuals r_i, $i=1,\ldots,s+1$, measure to what extent the perturbed values $\tilde{w}^{n+1}$, $\tilde{w}^n$, $\tilde{Y}_i$ fail to satisfy the equations (2.5)–(2.6). If we now subtract (2.5)–(2.6) from (2.7)–(2.8), we obtain a set of relations satisfied by the differences $\tilde{w}^n-w^n$, $\tilde{w}^{n+1}-w^{n+1}$, $\tilde{Y}_i-\tilde{Y}_i$, $i=1,\ldots,s$. A straightforward recursive elimination of the intermediate differences $\tilde{Y}_i-Y_i$, $i=1,\ldots,s$, leads to an expression for $\tilde{w}^{n+1}-w^{n+1}$ in terms of the residuals, i.e.,

$$\tilde{w}^{n+1}-w^{n+1}=P(\tau M)(\tilde{w}^n-w^n)+\sum_{i=1}^{s+1} Q_i(\tau M)\,r_i, \tag{2.9}$$

where P, Q_i, $i=1,\ldots,s+1$, are polynomials. The degree of P is $\leqq s$ and Q_i has degree $\leqq s+1-i$. The coefficients of P, Q_i can readily be expressed as functions of the coefficients a_{ij}, b_i, c_i of the method, but those expressions play no role here. Note that P is the usual stability polynomial.

We next consider the particular case of (2.7), (2.8) given by

$$\tilde{w}^{n+1}=w(t_{n+1}),\qquad \tilde{w}^n=w(t_n),\qquad \tilde{Y}_i=w(t_n+c_i\tau),\qquad i=1,\ldots,s,$$

i.e., all the values are taken from the theoretical solution $w(t)$. In this case, and assuming that w is smooth, we can write, for $i=1,\ldots,s,s+1$.

$$\begin{aligned} r_i&=w(t_n+c_i\tau)-w(t_n)-\tau\sum_{j=1}^{i-1} a_{ij}[Mw(t_n+c_j\tau)+g(t_n+c_j\tau)]\\ &=w(t_n+c_i\tau)-w(t_n)-\tau\sum_{j=1}^{i-1} a_{ij}\dot{w}(t_n+c_j\tau)\\ &=d_{i2}\tau^2\ddot{w}(t_n)+\ldots+d_{ip}\tau^p w^{(p)}(t_n)+R_i, \end{aligned} \tag{2.10}$$

where again, d_{ij} are scalar functions of the coefficients of the method, whose expression is not needed here. Note that $r_1=0$, since $c_1=0$. In (2.10) the

remainder R_i is $O(\tau^{p+1})$ and the constant in the $O(\tau^{p+1})$ term depends only on the RK method and on $w^{(p+1)}$. Substitution of (2.10) in (2.9) leads to the error relation, where we have taken into account that $r_1=0$,

$$w(t_{n+1})-w^{n+1}=P(\tau M)[w(t_n)-w^n] + \sum_{i=2}^{s+1} Q_i(\tau M)\sum_{j=2}^{p} d_{ij}\tau^j w^{(j)}(t_n)+\sum_{i=2}^{s+1} Q_i(\tau M)\,R_i. \tag{2.11}$$

In the case where $w_n=w(t_n)$ the difference $w(t_{n+1})-w^{n+1}$ is by definition the local error l^{n+1}. We have assumed the method to be of order p, so that $l^{n+1}=O(\tau^{p+1})$. Therefore in the right hand-side of (2.11) all terms involving powers τ^k, $k\leqq p$, must cancel and this leaves us finally with an expression

$$l^{n+1}=\sum_{l,j}\mu_{lj}\tau^{l+j}M^l w^{(j)}(t_n)+\sum_{i=2}^{s+1} Q_i(\tau M)\,R_i, \tag{2.12}$$

where, once more, μ_{lj} are scalar functions of the coefficients of the RK method and the indices l,j satisfy $1\leqq l\leqq s-1$, $2\leqq j\leqq p$, $p+1\leqq l+j$.

Example 2.1. We shall illustrate the foregoing derivation for the classical 4-stage, 4-th order scheme with the parameters

$$\begin{array}{c|cccc} c_1 & a_{11} & & & \\ & & & & \\ & & & & \\ c_4 & a_{41} & & & a_{44} \\ \hline & b_1 & \ldots & & b_4 \end{array} \quad = \quad \begin{array}{c|cccc} 0 & 0 & & & \\ 1/2 & 1/2 & 0 & & \\ 1/2 & 0 & 1/2 & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \tag{2.13}$$

The stability polynomial P arising first in equation (2.9) is the (4, 0)-Padé approximation to e^z,

$$P(z)=1+z+1/2\,z^2+1/6\,z^3+1/24\,z^2 \tag{2.14}$$

and the polynomials $Q_1,\ldots,Q_5$ arising in (2.9) are given by

$$\begin{aligned} Q_1(z)&=\frac{1}{6}z+\frac{1}{6}z^2+\frac{1}{12}z^3+\frac{1}{24}z^4, \quad Q_2(z)=\frac{1}{3}z+\frac{1}{6}z^2+\frac{1}{12}z^3,\\ Q_3(z)&=\frac{1}{3}z+\frac{1}{6}z^2, \quad Q_4(z)=\frac{1}{6}z, \quad Q_5(z)=1. \end{aligned} \tag{2.15}$$

The expansions of the residuals r_i introduced in (2.10) are

$$\begin{pmatrix} r_1\\ r_2\\ r_3\\ r_4\\ r_5 \end{pmatrix}=\begin{pmatrix} 0 & 0 & 0\\ 1/8 & 1/48 & 1/384\\ -1/8 & -2/48 & -3/384\\ 0 & 2/48 & 8/384\\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \tau^2 w^{(2)}(t_n)\\ \tau^3 w^{(3)}(t_n)\\ \tau^4 w^{(4)}(t_n) \end{pmatrix}+\begin{pmatrix} 0\\ R_2\\ R_3\\ R_4\\ R_5 \end{pmatrix} \tag{2.16}$$

The local error l^{n+1} given by (2.12) is found to be

$$l^{n+1}=\left(\frac{1}{576}Mw^{(4)}+\frac{-1}{288}M^2w^{(3)}+\frac{1}{96}M^3w^{(2)}\right)\tau^5$$
$$+\left(\frac{-1}{1152}M^2w^{(4)}+\frac{1}{576}M^3w^{(3)}\right)\tau^6$$
$$+\frac{1}{4608}M^3w^{(4)}\tau^7+\sum_{i=2}^{5}Q_i(\tau M)R_i, \tag{2.17}$$

where all derivatives are evaluated at $t=t_n$. The form of (2.17) will be used later in the paper. □

3. Behaviour of the Full Local Error

In this section we examine the behaviour of the full local error, i.e., the local error associated with the true PDE solution u_h instead of the local error associated with the intermediate ODE solution U_h (cf. [13]). The subsequent analysis is carried out under the following hypotheses.

(H1) The restriction $u_h(t)$ of the PDE solution possesses $p+1$ derivatives $u_h^{(j)}(t)$. Furthermore, $\|u_h^{(j)}(t)\|$, $j=0, 1, \ldots, p+1$, can be bounded uniformly in t and h.

(H2) The space and time grid refinements are carried out subject to a restriction

$$\tau\leqq\lambda h^q, \tag{3.1}$$

where λ is a fixed positive constant and q the order in space of (2.1).

(H3) For grid refinements satisfying (3.1), the expression $\tau\|A_h\|$ can be bounded independently of τ and h. (The bounds can nevertheless depend on λ.)

The local error (at t_{n+1}) of the fully discrete solution as an approximation to the PDE solution is defined by

$$l_h^{n+1}=u_h(t_{n+1})-\mathscr{R}u_h(t_n), \tag{3.2}$$

where $\mathscr{R}u_h(t_n)$ represents the result of a RK step for the system (2.2) starting from $u_h(t_n)$. Our task in this section is to derive bounds for $\|l_h^{n+1}\|$ of the form

$$C(\tau^k+\tau\max_{0\leqq t\leqq T}\|\alpha_h(t)\|), \tag{3.3}$$

where C denotes a constant independent of t_n, τ and h and k is a positive number. We will see that in order that the bound (3.3) be uniform in h, the exponent k must sometimes be taken smaller than $p+1$, the value one naively expects from the behaviour of the RK method as applied to ODEs.

In order to derive an expression for l_h^{n+1}, we consider in (2.7)-(2.8) the perturbed step $\tilde{w}^{n+1}=u_h(t_{n+1})$, $\tilde{w}^n=u_h(t_n)$, $\tilde{Y}_j=u_h(t_n+c_j\tau)$, $j=1, \ldots, s$. The re-

siduals r_i now take the form (cf. (2.3), (2.10))

$$\begin{aligned} r_i &= u_h(t_n + c_i\tau) - u_h(t_n) - \tau \sum_{j=1}^{i-1} a_{ij}[A_h u_h(t_n + c_j\tau) + f_h(t_n + c_j\tau)] \\ &= u_h(t_n + c_i\tau) - u_h(t_n) - \tau \sum_{j=1}^{i-1} a_{ij}[\dot{u}_h(t_n + c_j\tau) + \alpha_h(t_n + c_j\tau)] \\ &= d_{i2}\tau^2 \ddot{u}_h(t_n) + \ldots + d_{ip}\tau^p u_h^{(p)}(t_n) + R_i, \end{aligned} \tag{3.4}$$

where d_{ij} are the coefficients found in (2.10) and R_i contains not only the remainder in the Taylor expansion of $u_h^{(j)}(t_n + c_j\tau)$, but also the term $\tau \sum a_{ij}\alpha_h(t_n + c_j\tau)$. From these considerations and the hypothesis (H1) it is clear that the norms $\|R_i\|$ satisfy a bound of the form (3.3) with $k = p+1$. On proceeding now in a manner similar to that in the previous section, we find

$$l_h^{n+1} = \sum_{l,j} \mu_{lj}\tau^{l+j} A_h^l u_h^{(j)}(t_n) + \sum_{i=2}^{s+1} Q_i(\tau A_h) R_i, \tag{3.5}$$

where the summation l, j extends to $1 \leqq l \leqq s-1$, $2 \leqq j \leqq p$, $p+1 \leqq l+j$. We now proceed to bound l_h^{n+1}.

Lemma 3.1. *Under the hypotheses above the norms* $\|Q_i(\tau A_h)\|$, $i = 1, \ldots, s+1$, *can be bounded independently of* τ, h.

Proof. This follows directly from (H3), since $\|\tau^j A_n^j\| \leqq \|\tau A_h\|^j$. □

After this lemma, it is clear that the second term in the right hand side of (3.5) can be bounded in the form (3.3) with $k = p+1$. In estimating the first sum at least two different settings may be considered.

(S1) If the further assumption is made that the norms $\|A_h^l u_h^{(j)}(t_n)\|$ are bounded uniformly in t_n and h, then $\|l_h^{n+1}\|$ is bounded by (3.3) with $k = p+1$.

(S2) If no relation is assumed between the powers of A_h and the derivatives of $u_h(t)$, then to bound a term like $\tau^{l+j} A_h^l u_h^{(j)}$ uniformly in h, one must write

$$\|q^{l+j} A_h^l u_h^{(j)}\| = \tau^j \|(\tau A_h)^l u_h^{(j)}\| \leqq \tau^j \|\tau A_h\|^l \|u_h^{(j)}\|$$

and employ (H1) and (H3). The price to be paid is that now the order in τ is j rather than $p+1$, and in general the local error (3.5) contains terms with $j = 2$. (See in (2.17) the term (1/96) $\tau^5 A_h^3 u_h^{(2)}$ that one gets for the classical RK_4 scheme). In this way only an $O(\tau^2)$ bound is obtained, regardless of the value of the classical order p. Note that *this order reduction is not induced by lack of smoothness in* $u(x, t)$, but rather by the presence of powers of A_h in the expression for the local error, as these powers will contain negative powers of h.

In the above it was tacitly assumed that for the l and j considered the coefficient μ_{lj} of $\tau^{l+j} A_h^l u_h^{(j)}$ in (3.5) is not equal to zero. Trivially, if $\mu_{lj} = 0$, *this* term does not cause reduction. In the standard schemes of order p with p stages ($p = 2(1)4$) the coefficient $\mu_{p-1,2}$ associated with the term with highest order reduction cannot be zero. Schemes can be constructed with zero μ_{lj} coefficients. However, only at the price of introducing additional stages for a given order p.

Between the extreme settings (S1)-(S2) one can conceive situations (S3) where one knows that $\|A_h^l u_h^{(j)}\| \cdot h^\gamma = O(1)$ for a certain $\gamma < ql$. Then (H2) shows that $\|\tau^{l+j} A_h^l u_h^{(j)}\|$ behaves like $O(\tau^{j+l-\gamma/q})$, which is a more favourable estimate than the $O(\tau^j)$ stemming from (S2).

The following example should be helpful in illustrating the relevance of distinguishing the three situations (S1)-(S3).

Example 3.1. We consider the simple model hyperbolic problem

$$u_t = -u_x + f_\Omega(x, t), \quad 0 \leqq x \leqq 1, \quad 0 \leqq t \leqq 1, \tag{3.6a}$$

$$u(0, t) = f_\Gamma(t), \quad 0 \leqq t \leqq 1, \tag{3.6b}$$

$$u(x, 0) = u_0(x), \quad 0 \leqq x \leqq 1, \tag{3.6c}$$

which is assumed to possess a smooth solution. (This requirement implies not only that u_0, f_Ω and f_Γ are smooth, but also that they satisfy certain compatibility conditions whose expressions are of no consequence here.) If m is a positive integer, a uniform grid $x_j = j/m (0 \leqq j \leqq m)$ is introduced in $[0, 1]$, and (3.6) is discretized in space as follows ($h = m^{-1}$)

$$\begin{bmatrix} \dot{U}_1 \\ \\ \\ \dot{U}_m \end{bmatrix} = \begin{bmatrix} -1/h & & & \\ 1/h & -1/h & & \\ & 1/h & -1/h & \\ & & 1/h & -1/h \end{bmatrix} \begin{bmatrix} U_1 \\ \\ \\ U_m \end{bmatrix} + \begin{bmatrix} f_\Omega(x_1, t) + h^{-1} f_\Gamma(t) \\ f_\Omega(x_2, t) \\ f_\Omega(x_3, t) \\ f_\Omega(1, t) \end{bmatrix}. \tag{3.7}$$

We work with the usual L^2-norm. When the matrix A_h acts on a vector v_h obtained by restricting to the grid a smooth function $v(x)$, $0 \leqq x \leqq 1$, the 2^{nd} 3^{rd} ..., m^{th} entries in $A_h v_h$ approximate values of v_x and therefore can be bounded independently of h. However the first entry in $A_h v_h$ will behave like h^{-1} leading to a $h^{-1/2}$ behaviour of $\|A_h v_h\|$, unless v satisfies the homogeneous boundary condition $v(0) = 0$. It follows that the term $\tau^{p+1} A_h u^{(p)}$ is $O(\tau^{p+1})$, uniformly in h, if $u_h^{(p)}$ is 0 at the boundary, a condition which is of course satisfied if the boundary term $f_\Gamma(t)$ is identically zero, but not in general. To sum up, if $f_\Gamma \equiv 0$ then the term $\tau^{p+1} A_h u_h^{(p)}$ that features in (3.5) if $s \geqq 2$, behaves like $O(\tau^{p+1})$ uniformly in h, but in other case it may behave only like $O(\tau^{p+1/2})$ (use the arguments in situation (S3) above, with $q = 1$ and $\gamma = 1/2$).

In a similar vein $A_h^2 v_h$ is bounded if $v(0) = 0$ *and* $v_x(0) = 0$. If in (3.6) $f_\Gamma \equiv 0$ *and* $f_\Omega(0, t) \equiv 0$, then both u and u_x will be zero at the boundary and as a consequence the same will be true for all their derivatives with respect to t. In this case the terms $\tau^{j+1} A_h^2 u_h^{(j)}$, $j = p-1, p$, which feature in (3.5) if $s \geqq 2$ are $O(\tau^{j+2})$ uniformly in h and consequently $O(\tau^{p+1})$.

However, in general, $\|A_h^2 v_h\|$ behaves like $h^{-3/2}$ and this results in a reduction to $O(\tau^{p-1/2})$ in the term $\tau^{p+1} A_h^2 u_h^{(p-1)}$, and a reduction to $O(\tau^{p+1/2})$ in $\tau^{p+2} A_h^2 u_h^{(p)}$. The general trend should now be clear: for a method with s stages the optimal exponent $k = p+1$ in (3.3) cannot be obtained unless the theoretical solution $u(x, t)$ satisfies $s-1$ boundary requirements

$$u(0, t) = 0, \quad u_x(0, t) = 0, \ldots, (\partial^{s-2}/\partial x^{s-2}) u(0, t) = 0$$

that render it possible for $A_h^l u_h^{(j)} (1 \leqq l \leqq s-1,\ 2 \leqq j \leqq p,\ p+1 \leqq l+j)$ to remain bounded uniformly in h. These $s-1$ boundary requirements for u will be satisfied if and only if f_Ω, f_Γ do not violate a set of $s-1$ constraints $f_\Gamma \equiv 0$, $f_\Omega(0,t)=0, \ldots, (\partial^{s-3}/\partial x^{s-3}) f_\Omega(0,t)=0$. We emphasize that such constraints are induced by the numerical method and are not related to the compatibility conditions that f_Γ, f_Ω, u_0 must satisfy in order that u be smooth. Perhaps it is useful to point out that for homogeneous problems (homogeneous boundary conditions and no forcing term), the above constraints are trivially satisfied and no order reduction occurs.

4. Behaviour of the Full Global Error

We now turn to the full global error defined by

$$e_h^n = u_h(t_n) - U^n \tag{4.1}$$

where U^n denotes the fully discrete solution at time t_n. For simplicity we assume $e_h^0 = 0$ our aim is to derive bounds of the form

$$\|e_h^{n+1}\| \leqq C(\tau^k + \max_{0 \leqq i \leqq T} \|\alpha_h(t)\|), \tag{4.2}$$

with C a constant independent of t_n, τ, h, and k a positive number that we would like to be p in view of the order of the RK method when applied to an ODE. Our first result is

Theorem 4.1. *Assume that* (H1)-(H3) *hold, that* $\|A_h^l u_h^{(j)}(t_n)\|$ *can be bounded uniformly in* h *and* t_n, *for* $2 \leqq j \leqq p$. $l = p+1-j$, *and that for each* h *and* τ, $\|P(\tau A_h)\| \leqq 1$. *Then the convergence estimate* (4.2) *holds with an optimal value* $k = p$.

Proof. For $1 \leqq l \leqq s-1$, $2 \leqq j \leqq p$, $p+1 \leqq l+j$ we can write

$$\|\tau^{l+j} A_h^l u_h^{(j)}\| \leqq \tau^{p+1} \|\tau A_h\|^{l+j-p-1} \|A_h^{p+1-j} u_h^{(j)}\| = O(\tau^{p+1}),$$

so that the local error in (3.5) possesses a bound (3.3) with $k = p+1$. This bound and the stability assumption $\|P(\tau A_h)\| \leqq 1$ lead, in the standard way, to (4.2) with $k = p$. □

Some remarks are in order: First, we have required assumptions on $\|A_h^l u_h^{(j)}\|$. We saw in the previous section that these requirements are *not* naturally fulfilled in the applications, except if the PDE problem is homogeneous. Secondly, the stability condition $\|P(\tau A_h)\| \leqq 1$ is satisfied if the norm under consideration derives from an inner product, the matrices A_h are normal and λ in (3.1) has been chosen so that the eigenvalues of τA_h lie in the stability region S of the RK method $S = \{z: |P(z)| \leqq 1\}$ [9]. For nonnormal matrices this condition on the eigenvalues is necessary but not sufficient. An interesting sufficient condition involving the stability region S has been given by Spijker, [11], Th. 6.1.

In the general case where $\|A_h^l u_h^{(j)}\|$ are not bounded the analysis in the previous section only guarantees a τ^2-bound for the local error, leading via stability to an exponent $k=1$ in (4.2). A finer study of the local error, along the lines of what we called (S3) may result in τ^{k+1}-bounds for the local error, with $2<k+1<p+1$ and lead to τ^k-estimates of the global error.

An important point we want to make now is that the standard approach of transferring the local errors to the global error via stability (first bounding and then adding) can be unduly pessimistic [12]. An alternative technique, essentially used in [1, 2, 12] will now be presented. We consider one of the terms $\mu_{lj}\tau^{l+j}A_h^l u_h^{(j)}$, $1\leqq l\leqq s-1$, $2\leqq j\leqq p$, $p+1\leqq l+j$, that may suffer from reduction. This term contributes to the global error e_h^n by an amount

$$a_h^n=\mu_{lj}\tau^{l+j}\sum_{i=1}^{n}P(\tau A_h)^{n-i}A_h^l u_h^{(j)}(t_{i-1}). \tag{4.3}$$

Assume that the matrix $(I-P(\tau A_h))^{-1}\tau A_h$ can be defined and satisfies a bound

$$\|(I-P(\tau A_h))^{-1}\tau A_h\|\leqq\mathscr{K}, \tag{4.4}$$

with $\mathscr{K}$ independent of τ, h. (The feasibility of this condition is discussed later.) Then in (4.3) we can write

$$\begin{aligned}a_h^n&=\mu_{lj}\tau^{l+j-1}[(I-P(\tau A_h))^{-1}\tau A_h](I-P(\tau A_h))\sum_{i=1}^{n}P(\tau A_h)^{n-i}A_h^{l-1}u_h^{(j)}(t_{i-1})\\&=\mu_{lj}\tau^{l+j-1}[(I-P(\tau A_h))^{-1}\tau A_h]\cdot\Big[A_h^{l-1}u_h^{(j)}(t_{n-1})-P(\tau A_h)^n A_h^{l-1}u_h^{(j)}(t_0)\\&\quad+\sum_{i=1}^{n-1}P(\tau A_h)^{n-i}A_h^{l-1}(u_h^{(j)}(t_{i-1})-u_h^{(j)}(t_i))\Big].\end{aligned}$$

The following result now follows easily:

Theorem 4.2. *Assume that* (H1)-(H3) *and* (4.4) *hold and that as* h,τ *vary* $\|P(\tau A_h)\|\leqq 1$. *Then the contribution to the global error of a term* $\mu_{lj}A_h^l u_l^{(j)}$, $1\leqq l\leqq s-1$, $2\leqq j\leqq p$, $p+1\leqq l+j$, *possesses a bound of the form*

$$C\tau^{l+j-1}(\max_{t,h}\|A_h^{l-1}u_h^{(j+1)}\|+\max_{t,h}\|A_h^{l-1}u_h^{(j)}\|). \tag{4.5}$$

Proof. It is enough to write

$$\|A_h^{l-1}(u_h^{(j)}(t_{i-1})-u_h^{(j)}(t_i))\|=\left\|\int_{t_{i-1}}^{t_i}A_h^{l-1}u_h^{(j+1)}(s)\,ds\right\|\leqq\tau\max_{t,h}\|A_h^{l-1}u_h^{(j+1)}. \quad\square$$

The advantage of the new approach is that we have got rid of one power of A_h, i.e., we are now dealing with A_h^{l-1} instead of the A_h^l we started with. In the *worst* case, where $j=2$ and no relation is assumed between A_h^{l-1} and the derivatives of u_h, the bound (4.5) is $O(\tau^2)$, as shown by (H3). Recall that in the standard approach we only proved an $O(\tau)$ bound for the global error in the worst setting (S2) (cf. Th. 4.1).

Before we close this section the feasibility of (4.4) should be discussed. The rational function $\phi(z)=(1-P(z))^{-1}z$ is finite if $P(z)\neq 1$. Now, by consistency, $P(z)=1+z+O(z^2)$, so that for $z=0$, $P(z)=1$. But nevertheless $\phi(0)$ is finite. Therefore, $(I-P(\tau A_h))^{-1}\tau A_h$ exists if τA_h has no nonzero eigenvalue on the boundary of the stability region, a requirement only marginally more demanding than the spectral *necessary* stability condition mentioned above. Furthermore, slight modifications of sufficient stability conditions guanrantee the existence of a uniform bound (4.4). Two instances are given in the next proposition.

Proposition 4.1. *Each of the following two conditions is sufficient for* (4.4) *to hold:*

(i) *The norm* $\|\cdot\|$ *is an inner product norm, the matrices* A_h *are normal and as* τ, h *vary the eigenvalues of* τA_h *remain in a closed set* F *contained in* $\{0\}\cup(S-\partial S)$, *where* ∂S *is the boundary of* S.

(ii) *The norm* $\|\cdot\|$ *is an inner product norm and a positive number* ρ *exists such that the disk* $\{z: |z+\rho|\leqq\rho\}$ *is contained in* $\{0\}\cup(S-\partial S)$ *and, as* τ, h *vary,* $\|\tau A_h+\rho I\|\leqq\rho$.

Proof. (i) The rational function $\phi(z)=(1-P(z))^{-1}z$ is bounded in F. If $|\phi(z)|\leqq\mathscr{K}$ in F, then

$$\|(I-P(\tau A_h))^{-1}\tau A_h\|=\max\{\phi(\mu): \mu\in\mathrm{Spec}(\tau A_h)\}\leqq\mathscr{K},$$

where we have used the spectral theorem and the fact that $\phi(\tau A_h)$ is normal.

(ii) This follows from a theorem due to von Neumann [7] (cf. [2, 6, 11]). □

5. Numerical Illustration

Example 5.1. A simple experiment will be presented first which clearly shows the order reduction phenomenon. We consider the simple semidiscretization of Example 3.1 together with the classical fourth order RK-scheme (2.13). The mesh-ratio parameter λ is taken to be 1, a choice that guarantees that $\|P(\tau A_h)\|\leqq 1$ and that (4.4) holds. (Use Th. 6.1 in [11] and Proposition 4.1, (ii)). Furthermore, we take $u_0(x)=1+x$, $f_\Gamma(t)=1/(1+t)$, $f_\Omega(t)=(t-x)/(1+t)^2$ so as to have the simple solution $u=(1+x)/(1+t)$. Since this is linear in space, $\alpha_h\equiv 0$, i.e., there is no error introduced by the space discretization.

The time derivatives of u are *not* zero at the boundary; and then the analysis in Example 3.1 shows that the term $\tau^5(1/96)A_h^3 u_h^{(2)}$ behaves only like $\tau^{2.5}$ uniformly in h, leading to a decrease in local order of 2.5 units. The other terms of the local error involve higher powers of τ or lower powers of A_h and therefore suffer from reductions which harm less than that of the $\tau^5 A_h^3 u_h^{(2)}$ term. The conventional bound for the global error would show a $O(\tau^{1.5})$ behaviour of the global error, uniformly in h. However, the use of Theorem 4.2, reveals that the global error possesses a better, $O(\tau^{2.5})$, bound. Moreover the exponent 2.5 cannot be increased because at t_1 the local and global errors

coincide and we know that the local is not better than $O(\tau^{2.5})$. Table 1 shows the L^2-errors at $t=1$.

Table 1

τ^{-1}	h^{-1}			
	10	20	40	80
10	$0.31_{10}-4$			
20	$0.12_{10}-5$	$0.49_{10}-5$		
40	$0.62_{10}-7$	$0.20_{10}-6$	$0.83_{10}-6$	
80	$0.35_{10}-8$	$0.10_{10}-7$	$0.34_{10}-7$	$0.14_{10}-6$

From the table we computed the observed order of convergence obtained. The notation (1/10, 1/10)2.66(1/20, 1/20) denotes that an order of 2.66 was observed when refining the grid from $\tau=1/10$, $h=1/10$ to $\tau=1/20$, $h=1/20$, i.e., $2.66=\log_{10}\xi/\log_{10}2$, where ξ denotes the ratio of the error at (1/10.1/10) to the error at (1/20, 1/20). The rows of Table 2 display the observed order in the *simultaneous* refinement of τ and h, where the effect of the reduction is clearly seen.

Table 2

(1/10, 1/10)	2.66	(1/20, 1/20)	2.56	(1/40, 1/40)	2.56	(1/80, 1/80)
(1/20, 1/10)	2.58	(1/40, 1/20)	2.55	(1/80, 1/40)		
(1/40, 1/10)	2.63	(1/80, 1/20)				

The rows of Table 3 provide the order observed when in Table 1, the attention is focused in successively having τ with h fixed along the row.

Table 3

(1/10, 1/10)	4.69	(1/20, 1/10)	4.27	(1/40, 1/10)	4.14	(1/80, 1/10)
(1/20, 1/20)	4.61	(1/40, 1/20)	4.32	(1/80, 1/20)		
(1/40, 1/40)	4.60	(1/80, 1/40)				

Thus, on a fixed spatial grid there is *no order reduction* visible. Of course, this is the behaviour one should expect as one is now solving a *fixed* system of ODEs. With our fourth order method, the order asymptotically behaves like $C\tau^4$ on each fixed grid. The issue at hand is that C depends on the choice of mesh and increases with decreasing h. This is very clearly borne out in the last row of Table 1.

6. Avoiding Order Reduction

In this section we suggest a simple means for avoiding the order reduction. Although the principle is quite general, we prefer to describe it in the context

of a concrete situation. We consider again the model problem (3.6) and the classical RK method, but now the simple discretization (3.7) is replaced by the 4-th order scheme

$$(1/6)[\dot{U}_{j-1}+4\dot{U}_j+\dot{U}_{j+1}]=(1/(2h))[U_{j-1}-U_{j+1}]$$
$$+(1/6)[f_\Omega(x_{j-1},t)+4f_\Omega(x_j,t)+f_\Omega(x_{j+1},t)], \quad j=1(1)m-1, \tag{6.1}$$

with

$$(1/6)[\dot{U}_{m-1}+2\dot{U}_m]=(1/(2h))[U_{m-1}-U_m]+(1/6)[f_\Omega(x_{m-1},t)+2f_\Omega(x_m,t)] \tag{6.2}$$

near the boundary $x=1$. Note that (6.1)-(6.2) is the result of the Product Approximation Galerkin technique based on piecewise linear test functions [3].

From an analysis similar to that presented before an order reduction is to be feared, unless f_Ω, f_Γ satisfy the two constraints $f_\Gamma\equiv 0$, $f_\Omega(0,t)\equiv 0$ necessary for A_h, A_h^2 to act boundedly on the time derivatives of u_h. Now if $w(x,t)$ is a known function, then $v=u+w$ satisfies the transformed problem

$$v_t=-v_x+g_\Omega(x,t), \quad v(0,t)=g_\Gamma(t), \tag{6.3}$$

where

$$g_\Omega=f_\Omega+w_t+w_x, \quad g_\Gamma=f_\Gamma+w(0,\cdot) \tag{6.4}$$

are known functions. The idea is to choose w such the application of the numerical method to the problem (6.3) does not cause reduction (i.e., $g_\Gamma\equiv 0$, $g_\Omega(0,t)\equiv 0$), and then solve numerically for v and retrieve u from $u=v-w$. The finding of w is not difficult here. One may for instance choose $w(x,t)$ to be of the form $w(x,t)=\alpha(t)+x\beta(t)$ and then the conditions on g_Ω, g_Γ readily determine $\alpha(t)$ and $\beta(t)$.

The left half of Table 4 gives the L^2-errors for u when the integration is performed on (3.6) with f_Ω, f_Γ, u_0 chosen so that the solution is $u(x,t)=\cos(10t)\exp(-10x)$. The right half of the table corresponds to errors in u when the numerical integration is performed on the transformed problem (6.3). The results are in complete agreement with the theory.

Table 4

$\tau=h$	Error order		Error order	
1/10	$0.46_{10}-2$		$0.49_{10}-2$	
1/20	$0.52_{10}-3$	3.14	$0.21_{10}-4$	3.88
1/40	$0.76_{10}-4$	2.77	$0.21_{10}-4$	3.88
1/80	$0.13_{10}-4$	2.54	$0.14_{10}-5$	3.91

7. Concluding Remarks

The attention here has been restricted to linear problems. Order reduction also takes place for nonlinear problems and the mechanism involved there is essentially the one we have discussed. The extensions of the analysis to the nonlinear case is possible but becomes rather technical and offers no new insight.

For implicit RK schemes the main ideas of our analysis are still valid. However, the interest there is in situations where τ and h are not related and therefore our hypothesis (H2) and (H3) should be forsaken. The details of the analysis become then quite different [1, 12]. The technique for avoiding the order reduction outlined in Sect. 6, can also be used with implicit schemes. In fact we have employed it with success to retrieve the 3rd and 4th order of convergence of the diagonally implicit RK schemes discussed in [12].

It is fair to say that in practical problems the negative effects caused by order reduction are likely to be less important than those stemming from other sources, such as errors in space, instabilities at boundaries, curved boundaries, etc. However, the understanding of this phenomenon is essential in situations where one is interested in higher order methods.

Acknowledgements. The authors are indebted to Mrs. J. Blom for her help with the numerical computations. J.M.S. is also thankful to the "Centrum voor Wiskunde en Informatica" for the hospitality provided.

References

1. Brenner, P., Crouzeix, M., Thomée, V.: Single step methods for inhomogeneous linear differential equations in Banach space. RAIRO Anal. Numér. **16**, 5-26 (1982)
2. Burrage, K., Hundsdorfer, W.H., Verwer, J.G.: A study of *B*-convergence of Runge-Kutta methods. Computing **36**, 17-34 (1986)
3. Christie, I., Griffiths, D.F., Mitchell, A.R., Sanz-Serna, J.M.: Product approximation for nonlinear problems in the finite element method. IMA J. Numer. Anal. **1**, 253-266 (1981)
4. Dekker, K., Verwer, J.G.: Stability of Runge-Kutta methods for stiff nonlinear differential equations. Amsterdam, New York, Oxford: North-Holland 1984
5. 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 (1983)
6. Hairer, E., Bader, G., Lubich, C.: On the stability of semi-implicit methods for ordinary differential equations. BIT **22**, 211-232 (1982)
7. von Neumann, J.: Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. Math. Nachr. **4**, 258-281 (1951)
8. Praagman, N.: Numerical solution of the shallow water equations by a finite element method. Thesis, Technical University of Delft 1979
9. Sanz-Serna, J.M.: Convergent approximations to partial differential equations and stability concepts of methods for stiff systems of ordinary differential equations. Actas del VI CEDYA, Jaca, University of Zaragoza 1984 (available on request from J.M.S.)
10. Sanz-Serna, J.M.: Stability and convergence in Numerical Analysis I: Linear problems, a simple, comprehensive account. In: Nonlinear differential equations and applications. (J. Hale, P. Martinez-Amores, eds.), pp. 64-113. London: Pitman 1985
11. Spijker, M.N.: Stepsize restrictions for stability of one-step methods in the numerical solution of initial value problems. Math. Comput. **45**, 377-392 (1985)
12. Verwer, J.G.: Convergence and order reduction of diagonably implicit Runge-Kutta schemes in the method of lines. Report NM-R8506, Centre Math. Comput. Sci., Amsterdam 1985 (to appear in Proc. Dundee Num. Anal. Conf. 1985)
13. Verwer, J.G., Sanz-Serna, J.M.: Convergence of method of lines approximations to partial differential equations. Computing **33**, 297-313 (1984)
14. Wubs, F.W.: Stabilization of explicit methods for hyperbolic initial-value problems. Report NM-R8521, Centre Math. Comput. Sci., Amsterdam 1985

Received June 16, 1986/September 25, 1986