Convergence Analysis of One-Step Schemes in the Method of Lines

J.M. Sanz-Serna
Dpto. de Ecuaciones Funcionales, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain

J.G. Verwer
Centre for Mathematics and Computer Science P.O. Box 4079, AB Amsterdam, The Netherlands

We present an expository account of some fundamental results concerning the analysis of one-step schemes for semidiscretizations of evolutionary problems in partial differential equations. In the paper the emphasis lies in the interplay between the stability and convergence properties of the fully discrete scheme and those of the ordinary differential equations solver. Much attention is paid to the phenomenon of order reduction.

1. Introduction

It is well known that many discretizations of time dependent problems in partial differential equations (PDEs) can be derived by means of the following two stage procedure. First, the space variables are discretized so as to convert the PDE into a system of ordinary differential equations (ODEs) with the time as independent variable. Then, the discretization in time of this ODE system yields the sought, fully discrete scheme. In the literature this two-stage procedure is often referred to as the *method of lines* (MOL). The purpose of the present contribution is to give an expository account of some fundamental results concerning the stability and convergence analysis of one-step MOL schemes. In our exposition, which is largely based on our earlier papers [2, 3 (Ch. 10), 11, 12, 14, 17, 18] the emphasis lies in the interplay between the *stability* and *convergence* properties of the *fully discrete scheme* and those of the *ODE solver*.

The contents of the paper is as follows. Section 2 deals with preliminaries. Here we introduce the PDE considered and collect some basic material which is needed later in the paper. Section 3 deals with stability aspects. In this section we briefly mention the standard PDE analysis, the (shortcomings) of the spectral condition property, and the notion of contractivity which in the last decade has attracted much attention [3]. In Section 4 we discuss consistency and convergence properties of the fully discrete MOL schemes. In particular, attention is paid to the order of convergence under simultaneous refinement of the grids in time and in space. This leads us to the somewhat awkward phenomenon of order reduction, i.e., in many cases it is found that under simultaneous refinement the order of convergence in time of the fully discrete scheme is less than the order of convergence of the ODE solver. A numerical example is given to illustrate the order reduction phenomenon.

0096-3003/89/$03.50

2. Preliminaries

2.1 Partial differential problem

We consider linear problems of the form

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

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

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

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 $\leqslant q-1$, acts on the boundary Γ and serves to introduce the boundary conditions (2.1b). 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.

Most of the following considerations may be extended, with varying degrees of difficulty, to problems more general than (2.1), including nonlinear cases. Nevertheless, the class (2.1) is wide enough to describe a large number of interesting practical situations and also to display some major difficulties to be expected in the time-integration of evolutionary PDEs. We therefore focus our attention in this paper on problems of the form (2.1), and briefly comment on other models when appropriate. It is assumed that $A_\Omega, A_\Gamma, f_\Omega, f_\Gamma$ and $u(x,0)$ are such that (2.1) possesses a unique solution u.

2.2 Space discretization

The discretization in space of the problem (2.10), by means of finite differences, results in a Cauchy problem

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

which is assumed to be uniquely solvable. 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). Finite-element discretizations can be catered for with minor modifications (see [11]) and will not be treated here.

Note that the dimension m of U_h depends on h. Throughout the paper, $\|\cdot\|$ denotes a chosen norm for m-dimensional vectors and the corresponding operator norm for $m \times m$ matrices.

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 [11]). Then (2.2) is said to be a *convergent* semidiscretization of (2.1) if, as $h \to 0$,

$$\max_{0 \leqslant t \leqslant T} \|u_h(t) - U_h(t)\| = o(1),$$

provided that $\|u_h(0) - U_h(0)\| = o(1)$. Convergence of order $\hat{p}$ is defined in the obvious way, i.e. replacing $o(1)$ by $O(h^{\hat{p}})$ in both occurrences of the symbol $o(1)$. For simplicity we assume hereafter that

$$U_h(0) = u_h(0) = [u(x_1,0), \cdots, u(x_m,0)]^T,$$

i.e., in the semidiscretization there is no error involved in approximating the initial function.

The vector $u_h(t)-U_h(t)$ is referred to as the *global error* of the semidiscretization. Also of interest later is the *truncation error* of (2.2) defined by

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

2.3 An illustration

The following example might be helpful in order to become familiar with the preceding notation. We consider the simple hyperbolic problem

$$u_t=-u_x+f_\Omega(x,t), \quad 0\leqslant x\leqslant 1, \quad 0\leqslant t\leqslant T, \tag{2.4a}$$

$$u(0,t)=f_\Gamma(t), \quad 0\leqslant t\leqslant T, \tag{2.4b}$$

$$u(x,0) \text{ given}, \quad 0\leqslant x\leqslant 1, \tag{2.4c}$$

which is of the form (2.1) with $q=1$. Let m be a positive integer. A uniform grid $x_j=j/m$ $(0\leqslant j\leqslant m)$ is introduced in the x-interval [0,1] and (2.4a,b) is discretized in space by first order, backward differences to yield the semidiscretization

$$\begin{bmatrix}\dot{U}_1\\ \dot{U}_2\\ \\ \dot{U}_m\end{bmatrix}=\begin{bmatrix}-1/h & & \\ 1/h & -1/h & \\ & & \\ & 1/h & -1/h\end{bmatrix}\begin{bmatrix}U_1\\ U_2\\ \\ U_m\end{bmatrix}+\begin{bmatrix}f_\Omega(x_1,t)+h^{-1}f_\Gamma(t)\\ f_\Omega(x_2,t)\\ \\ f_\Omega(1,t)\end{bmatrix}, \tag{2.5}$$

where $U_h(t)=[U_1(t),\cdots,U_m(t)]^T$, with $U_j(t)$ an approximation to $u(jh,t)$, $j=1(1)m$.

Note that (2.5) represents a family of ODEs depending on the parameter h and that the dimension $m=1/h$ of the system and the spectral radius $1/h$ of the matrix A_h grows as $h\to 0$. For smooth solutions u the semidiscretization (2.5) can be proved to be L^p-convergent of the first order for $p=1,2,\infty$, i.e.,

$$\max_{0\leqslant t\leqslant T}\|u_h(t)-U_h(t)\|_p=O(h), \quad h\to 0.$$

Here, $\|\cdot\|_p$ is the L^p-norm for grid functions, i.e., for $p=1,2$

$$\|w\|_p^p=\sum_{j=1}^{m} h|w(x_j)|^p, \quad x_j=jh,$$

with the obvious modification for $p=\infty$. The convergence is proved in two steps. i) Prove that for the truncation error (2.3), with components

$$\alpha_j(t)=h^{-1}(u(x_{j-1},t)-u(x_j,t))+u_x(x_j,t),$$

a bound

$$\max_{0\leqslant t\leqslant T}\|\alpha_h(t)\|_p=O(h) \tag{2.6}$$

holds. This is trivially achieved by means of Taylor expansions. ii) Note that the global error u_h-U_h of the semidiscretization satisfies, by using (2.2) and (2.3), the differential equation $d(u_h-U_h)/dt=A_h(u_h-U_h)-\alpha_h(t)$. Next, use the variation of constant

formula or the energy method (as in [18]) to bound $\|u_h(t)-U_h(t)\|_p$ in terms of $\|\alpha_h(t)\|_p$.

The fact that we have chosen the model hyperbolic problem (2.4) is dictated by the simplicity in presentation. Further, this problem also proves to be useful for us in two later instances. Other examples of convergence proofs of semidiscretizations can be found in [18].

2.4 The ODE solver and full convergence

In order to get a fully discrete scheme, the problem (2.2) is discretized in time by a convergent, p-th order, *one-step* ODE solver with step τ independent of t. We suppose that τ satisfies

$$\tau \leqslant \lambda h^q \tag{2.7}$$

with $\lambda \in (0, \infty]$ a fixed constant and q the order in space of (2.1). We denote by U^n the corresponding fully discrete solution at time $t_n = n\tau$ (the dependence on h is suppressed in the notation U^n and in other notations introduced later).

Our task is to study the behaviour of the *global error* $e^n = u_h(t_n) - U^n$ of the fully discrete solution and more precisely to bound

$$\max_{0 \leqslant t_n \leqslant T} \|e^n\| = \max_{0 \leqslant t_n \leqslant T} \|u_h(t_n) - U^n\| \tag{2.8}$$

under an appropriate choice of λ in (2.7). Again we assume that there is no error involved in approximating the initial condition, i.e., $U^0 = U_h(0) = u_h(0)$. A minimal requirement that the (full) discretization should satisfy is that of *convergence,* i.e., that as $h, \tau \to 0$, subject to (2.7), the quantity in (2.8) tends to zero.

An important point here is that the convergence of the semidiscrete approximation U_h to u, together with the use of a convergent ODE solver, is not sufficient to guarantee the convergence of the fully discrete approximations U^n. For example, time-stepping in the convergent semidiscretization (2.5) with the convergent forward Euler rule results in a well known explicit method for (2.4), which does *not* converge if $\lambda > 1$. (The CFL condition [10,15] is violated).

Let us write, for fixed $n\tau = t_n$,

$$\|u_h(t_n) - U^n\| \leqslant \|u_h(t_n) - U_h(t_n)\| + \|U_h(t_n) - U^n\|. \tag{2.9}$$

The convergence of the semidiscrete approximation implies that the first term in the right hand-side of (2.9) tends to zero as $h \to 0$. For a convergent ODE solver, $\|U_h(t_n) - U^n\|$ tends to zero as $\tau \to 0$ for fixed h. However, the system (2.2) to which the ODE solver is applied changes with h. Therefore, in order to achieve the convergence of the fully-discrete scheme we must demand that, as h varies, the convergence of the ODE solver be *uniform* in the members of the family (2.2).

Such a uniformity cannot be concluded from the standard convergence results for ODE solvers as the underlying error bounds there typically involve the factor $\exp(L_h t_n)$, with L_h the classical Lipschitz constant for A_h. This Lipschitz constant is of no use here due to the *negative powers* of h in A_h. This observation makes clear that for proving convergence of fully discrete MOL approximations it is necessary to derive error bounds which are basically independent of h or, using ODE terminology, independent of *stiffness.*

The derivation of stiffness independent error bounds has recently attracted much

attention in the field of implicit RK methods for ODEs (*B*-convergence theory, FRANK, SCHNEID & UEBERHUBER [4]). In [17] it has been shown that the results of this theory are also of use for the MOL convergence analysis. It should be noted however that this analysis is not based on the use of the error splitting (2.9), but compares u_h and U^n directly without employing the intermediate U_h.

3. STABILITY ASPECTS

3.1 The standard PDE analysis

The application to (2.2) of a one-step method usually results in a recursion

$$U^{n+1} = R(\tau A_h)U^n + F^n, \quad 0 \leqslant (n+1)\tau \leqslant T, \tag{3.1}$$

where $R(\cdot)$ is the stability function associated with the method and F^n is an m-dimensional vector originating from the inhomogeneous term $f_h(t)$. The standard PDE analysis relates (3.1) to (2.1) without resorting to (2.2) and requires the introduction of the *full truncation error* of (3.1) given by

$$\beta^{n+1} = u_h(t_{n+1}) - R(\tau A_h)u_h(t_n) - F^n. \tag{3.2}$$

Note that this is a residual associated with the PDE solution and is therefore different from the local error of (3.1) considered as a time-discretization of the ODE system (2.2). From (3.1)-(3.2) we find the following recursion for $e^n = u_h(t_n) - U^n$ which is the full global error,

$$e^{n+1} = R(\tau A_h)e^n + \beta^{n+1}, \tag{3.3}$$

whence (recall that $e^0 = 0$)

$$e^n = \sum_{i=1}^{n} R(\tau A_h)^{n-i}\beta^i. \tag{3.4}$$

From this expression (which is the discrete counterpart of the variation of constant formula we mentioned before) we conclude that, together, the following two conditions guarantee convergence [10]: (LS) (Stability) As τ, h vary subject to (2.7), the norms $\|R(\tau A_h)^j\|$, $0 \leqslant j\tau \leqslant T$, can be bounded by a constant S independently of h, τ and j.
(Cons.) (Consistency) As τ, h vary subject to (2.7),

$$\max_{0 \leqslant t_n \leqslant T} \|\beta^{n+1}\| = o(\tau).$$

In fact, it is enough to write

$$\|e^n\| \leqslant nS \max_n \|\beta^n\| \leqslant TS\tau^{-1} \max_n \|\beta^n\|. \tag{3.5}$$

The stability requirement LS is the *Lax-stability condition* and, under very general hypotheses, is also *necessary* for (full) convergence [8,9,10,13,15].

A somewhat more demanding stability property related to the concept of *strong stability* (KREISS [6]) is given by
(SS) As τ, h vary subject to (2.7)

$$\|R(\tau A_h)\| \leqslant 1 + C\tau, \tag{3.6}$$

where C is a constant independent of τ, h.

This requirement is stronger than (LS), because if (3.6) holds

$$\|R(\tau A_h)^j\| \leqslant \|R(\tau A_h)\|^j \leqslant (1+C\tau)^j \leqslant \exp(Cj\tau) \leqslant \exp(CT)$$

so that (LS) holds with S = exp (CT). Also note that if (SS) is satisfied there is no need to consider the representation (3.4), since in this case (3.3) leads directly to

$$\|e^{n+1}\| \leqslant (1+C\tau)\|e^n\| + \|\beta^{n+1}\|, \tag{3.7}$$

a recursion for $\|e^n\|$ which can be easily used to prove convergence.

3.2 Contractivity and C-stability - MOL analysis

The condition (SS) and the recursion (3.7) have often been used in convergence proofs of one-step MOL schemes (see [3], Ch. 10; there the term *C-stability* is used). In fact, a particular case of (3.6) is the condition $\|R(\tau A_h)\| \leqslant 1$ which implies that for any two solutions V^n, W^n of (3.1), stemming from different initial functions V^0, W^0, there holds

$$\|V^{n+1} - W^{n+1}\| \leqslant \|V^n - W^n\|. \tag{3.8}$$

In the field of stiff ODEs this behaviour is called *contractivity*.

The concepts of contractivity and C-stability have two merits: i) They can be extended in a natural way to nonlinear problems. When they hold, they imply, together with full consistency, the convergence of the fully discrete approximations. ii) It is possible to give general results for the contractivity and C-stability of Runge-Kutta methods. For instance, the backward Euler method is contractive in any norm when applied to any dissipative system of ODEs.

The investigation of the concepts of contractivity (B-stability) and C-stability has been dominant in the recent studies of stability in ODE-solvers. The points i) and ii) above are adequately covered in the monograph [3] and the interested reader is referred to this work for the study of these issues.

3.3 The spectral condition

Consider the *stability region* $\mathbb{S} = \{z \in \mathbb{C} : |R(z)| \leqslant 1\}$ of the method. A stability requirement that easily comes to mind in MOL applications is the demand that the constant λ in (2.7) should be chosen to guarantee that, as τ and h vary, the *eigenvalues* $\tau\lambda_{ih}$ of $\tau A_h (i = 1(1)m)$ should lie in $\mathbb{S}$ or even in the interior of $\mathbb{S}(|R(z)| < 1)$. The demand that $\tau\lambda_{ih} (i = 1(1)m)$ lies in the interior of $\mathbb{S}$ seems particularly appealing in that it guarantees that if, for *fixed* τ and h, j increases without bound that then $\|R(\tau A_h)^j\| \to 0$. Hence, any error such as round-off will be then eventually damped.

However, this *spectral condition* approach may be dangerous, since in general it provides little information on the behaviour of $\|R(\tau A_h)^j\|$ as $\tau, h \to 0$, nor does it exclude excessive growth of $\|R(\tau A_h)^j\|$ for finite values of j with τ, h fixed. A classical example showing such a disastrous behaviour is given by the semi-discrete hyperbolic problem (2.5) when integrated in time by the forward Euler method [10, p. 152], [3, p. 272]. The choice $\lambda = 2$ in (2.7) satisfies the spectral condition, whereas the Lax stability condition (LS) or the condition (SS) (in the common L^p norms) dictates the choice $\lambda = 1$ which means $\tau/h \leqslant 1$. In practice, computations with $1 \leqslant \tau/h \leqslant 2$ easily lead to inacceptably large errors.

In spite of the previous remarks, there are cases where conditions on the stability

region of the ODE method are sufficient to guarantee stability in the senses (LS) or (SS):

(i) If $\|\cdot\|$ is an inner product norm and A_h is *normal* with respect to this inner product, then the condition $\tau\lambda_{ih} \in \mathcal{S}, i = 1, \cdots, m$, implies $\|R(\tau A_h)^j\| \leq 1$, $j = 1,2, \cdots$. This is a consequence of the fact that $R(\tau A_h)^j$ is normal and therefore $\|R(\tau A_h)^j\| = \rho(R(\tau A_h)^j) = (\rho(R(\tau A_h))^j$, where $\rho(\cdot)$ denotes the spectral radius.

(ii) If $\|\cdot\|$ is an inner product norm and the solver is A-stable then $\|R(\tau A_h)^j\| \leq 1$, $j = 1,2, \cdots$ for arbitrary τ. Here is assumed that $<A_h v_h, v_h> \leq 0$ for any grid function v_h. This result follows from a deep theorem by VON NEUMANN [5] [7] and does not require the normality of A_h.

The result by von Neumann has been recently used by SPIJKER [16] to derive an interesting sufficient condition for contractivity.

4. CONSISTENCY ASPECTS

4.1 The structure of the (full) local error

After our review of the behaviour of the powers $R(\tau A_h)^j$ we now turn our attention towards the local errors β^{n+1}, the other factor that according to (3.4) determines the global error. Our aim is to investigate the behaviour of β^{n+1} in terms of the smoothness in time of the PDE solution $u_h(t)$ and the space truncation error $\alpha_h(t)$ introduced in (2.3).

We now assume that the ODE solver used for the system (2.2) is an s-stage, p-th order *explicit Runge-Kutta* method given by the array

$$\begin{array}{c|cccc} c_1 & & & & \\ c_2 & a_{21} & & & \\ & & & & \\ c_s & a_{s1} & & a_{ss-1} & \\ \hline & b_1 & & b_{s-1} & b_s \end{array} \tag{4.1}$$

As usual we let $\Sigma_{i=1}^{s} b_i = 1$, $\Sigma_{j=1}^{i-1} a_{ij} = c_i (1 \leq i \leq s)$ and set $a_{s+1,j} = b_j (1 \leq j \leq s)$ and $c_{s+1} = 1$.

It is emphasized that the main conclusions of the following analysis are also valid for *implicit* Runge-Kutta methods, but the technical details are somehow different and also slightly more complicated (see [14,17]). For the sake of presentation in this expository contribution we therefore concentrate on the explicit methods.

We begin by defining the *residual* r_i associated with the i-th stage ($i = 1, \cdots, s+1$) of the step $t_n \to t_{n+1}$,

$$r_i = u_h(t_n + c_i\tau) - u_h(t_n) - \tau \sum_{j=1}^{i-1} a_{ij}[A_h u_h(t_n + c_j\tau) + f_h(t_n + c_j\tau)].$$

Note that this residual is defined for the PDE solution u_h as in Section 3.1 and that, by definition, $r_1 = 0$. Using (2.3) we can write

$$r_i = u_h(t_n + c_i\tau) - u_h(t_n) - \tau \sum_{j=1}^{i-1} a_{ij}[\dot{u}_h(t_n + c_j\tau) + \alpha_h(t_n + c_j\tau)],$$

and if we assume that $u_h(t)$ possesses $p+1$ derivatives we can Taylor expand $u_h(t_n + c_j\tau)$,

$\dot{u}_h(t_n+c_j\tau)$ to arrive at an expresssion

$$r_i = d_{i2}\tau^2\ddot{u}_h(t_n) + \cdots + d_{ip}\tau^p u_h^{(p)}(t_n) + R_i. \tag{4.2}$$

Here d_{ij} are coefficients which only depend on the array (4.1) and R_i is the sum of the remainder in the Taylor expansions plus the term $\tau\Sigma a_{ij}\alpha_h(t_n+c_j\tau)$, which is the space error contribution.

The intermediate residuals are used for deriving an expression for β^{n+1} which is more amenable for the task we have set ourselves than the expression given in (3.2). This is done as follows. We write down the ordinary Runge-Kutta equation (cf. (4.1))

$$Y_i = U^n + \tau\sum_{j=1}^{i-1} a_{ij}[A_h Y_j + f_h(t_n+c_j\tau)] \quad (1\leqslant i\leqslant s+1)$$

where $U^{n+1} = Y_{s+1}$, and the *perturbed* equations

$$u_h(t_n+c_i\tau) = u_h(t_n) + \tau\sum_{j=1}^{i-1} a_{ij}[A_h u_h(t_n+c_j\tau) + f_h(t_n+c_j\tau)] + r_i \quad (1\leqslant i\leqslant s+1).$$

Then we subtract these two formulas to obtain a set of relations satisfied by the full global errors $e^n = u_h(t_n) - U^n, e^{n+1}$ and the intermediate errors $u_h(t_n+c_i\tau) - Y_i$ $(1\leqslant i\leqslant s)$. A straightforward recursive elimination of the latter errors leads to the recursion (3.3) for the full global error, but with β^{n+1} now in the form

$$\beta^{n+1} = \sum_{i=1}^{s+1} Q_i(\tau A_h) r_i, \tag{4.3}$$

where Q_i is a polynomial of degree $\leqslant s+1-i$ whose coefficients depend on (4.1). Note that the behaviour of these polynomials accounts for the *internal* stability of the *RK*-scheme, i.e., for the effect on U^{n+1} of perturbations in the stages of the step $t_n \to t_{n+1}$. Substitution of (4.2) in (4.3) finally leads to the *full local error expression*

$$\beta^{n+1} = \sum_{l,j} \mu_{lj}\tau^{l+j} A_h^l u_h^{(j)}(t_n) + \sum_{i=2}^{s+1} Q_i(\tau A_h) R_i, \tag{4.4}$$

where μ_{lj} are scalars which only depend on the parameters in (4.1) and the summation l,j extends to $1\leqslant l\leqslant s-1$, $2\leqslant j\leqslant p$, $p+1\leqslant l+j$.

An important point to notice is that in (4.4) we find not only derivatives $u_h^{(j)}(t_n)$ (that are expected to behave nicely as $h\to 0$), but also powers A_h^l that will increase as $h\to 0$ due to the *negative* powers of h contained in A_h^l. Thus the analysis of (4.4) can be expected to be delicate and, indeed, we will see below that the negative powers of h cause difficulties.

4.2 Behaviour of the full local error - Local order reduction

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 $\|u_h^{(j)}(t)\|$ can be bounded uniformly in t and h. (H2) The space-time grid refinement is carried out subject to a condition (2.7) with $\lambda<\infty$ and for this refinement the expression $\tau\|A_h\|$ can be bounded independently of τ and h.

Hypothesis (H2) is natural here since we are discussing explicit methods. Recall that the order in space of (2.1) is q and that therefore the entries of A_h are expected to

increase like h^{-q}.

Our task in this subsection is to derive for $\|\beta^{n+1}\|$ bounds of the type

$$C(\tau^k + \tau \max_{0 \leq t \leq T} \|\alpha_h(t)\|), \tag{4.5}$$

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 (4.5) be uniform in h, *the exponent k must usually be taken smaller than* $p+1$, the value one naively expects from the behaviour of the RK method applied to ODEs.

Now the hypothesis (H1) implies that in (4.4) the terms R_i satisfy a bound of the form (4.5) with $k = p+1$. On the other hand, (H2) implies that $\|Q_i(\tau A_h)\|$ are bounded uniformly in τ, h and therefore the second summation in (4.4) can be bounded in the form (4.5) with an *optimal* $k = p+1$. In estimating the first sum at least two different settings may be considered (see [14] for a third setting).

(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 $\|\beta^{n+1}\|$ is bounded by (4.5) 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 $\tau^{l+j} A_h^l u_h^{(j)}(t_n)$ *uniformly* in h, one must write

$$\|\tau^{l+j} A_h^l u_h^{(j)}(t_n)\| = \tau^j \|(\tau A_h)^l u_h^{(j)}(t_n)\| \leq \tau^j \|\tau A_h\|^l \|u_h^{(j)}(t_n)\|$$

and employ (H1) and (H2). The price to be paid is that for such a term the order in τ is j rather than $l+j \geq p+1$. In general the local error (4.4) contains terms with $j=2$ so that in this way *only an* $O(\tau^2)$ *bound is obtained regardless of the value of the classical order p.* We emphasize 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 local error (4.4).

4.3 Behaviour of the (full) global error - Global order reduction

Once β^{n+1} has been bounded in the form (4.5) (possibly with a reduced k, i.e., $k < p+1$) a stability assumption like (LS) or (SS) mentioned in Section 3.1 immediately leads to the global error bound

$$\|e^n\| \leq C(\tau^{k-1} + \max_{0 \leq t \leq T} \|\alpha_h(t)\|), \tag{4.6}$$

by applying the standard arguments given there (cf. (3.5) or (3.7)). An important point we wish to make now is that if $k < p+1$, these standard arguments of transferring the local errors to the global error (first bounding and then adding via stability) can be unduly *pessimistic* [1,2,14,18].

Consider (3.4) and (4.4). As already concluded in Section 4.2 the second summation in (4.4) can be bounded in the form (4.5) with an optimal $k = p+1$, implying that this part of the local error causes no order problem and can be dealt with in the standard way. So we now confine ourselves to the first sum in (4.4) and treat only *one* of the terms $\mu_{lj} \tau^{l+j} A_h^l u_h^{(j)}$ that may suffer from reduction. The other terms can be dealt with in the same fashion.

According to (3.4) the term considered contributes to the global error e_h^n by an amount

$$a_h^n = \mu_{lj} \tau^{l+j} \sum_{i=1}^{n} R(\tau A_h)^{n-i} A_h^l u_h^{(j)}(t_{i-1}). \tag{4.7}$$

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

$$\|(I-R(\tau A_h))^{-1}\tau A_h\| \leqslant \mathcal{K}, \tag{4.8}$$

with $\mathcal{K}$ independent of τ, h. (The feasibility of this condition is discussed in [14]). Then (4.7) can be rewritten as

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

Further we write

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

The following result now follows easily: Suppose that as h, τ vary (4.8) holds and $\|R(\tau A_h)\| \leqslant 1$. Then the *global error contribution* a_h^n possesses a bound of the form

$$C\mu_{lj}\tau^{l+j-1}(\max_{t,h}\|A_h^{l-1}u_h^{(j+1)}\| + \max_{t,h}\|A_h^{l-1}u_h^{(j)}\|). \tag{4.9}$$

The crucial observation 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. This is of importance since the reduction emanates from the negative powers of h contained in A_h.

If we collect all bounds (4.9) for l, j from their range of summation

$$(1 \leqslant l \leqslant s-1, 2 \leqslant j \leqslant p, l+j-1 \geqslant p),$$

take into account the second summation in (4.4) and the hypotheses (H1), (H2) of Section 4.2, we finally arrive at a global error bound (cf. (4.6))

$$\|e^n\| \leqslant C(\tau^g + \max_{0 \leqslant t \leqslant T}\|\alpha_h(t)\|), \quad k-1 < g \leqslant p. \tag{4.10}$$

The particular value of g depends on the *problem*, i.e., on the order in space q and on the possible growth with h of the grid functions occurring in (4.9). In the *worst case*, where no relation is assumed between the powers of A_h and the derivatives of u_h (setting (S2)) the order in τ of the individual bounds (4.9) can be put equal to j, so that if $p > 1$ we have $g \geqslant 2$ in (4.10). Hence in the worst case setting it is possible that the drop in global order is *one unit less* than that in local order (see Section (4.2)). Obviously, in the setting (S 1) we have $g = p$ and the special derivation of this subsection is not necessary.

In the next subsection we shall discuss a particular example with the aim of illustrating the analysis, but also to show that the (minimal) order $g = 2$ in (4.10) really may occur.

4.4 An example

We consider the hyperbolic model problem (2.4) with its semidiscretization (2.5). It is supposed that the solution u of (2.4) is as smooth as the analysis requires. (This assumption 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). We shall work with the usual L^2-norm and L^∞-norm.

Let $v(x), 0 \leqslant x \leqslant 1$, be some smooth function. When the matrix A_h given by (2.5) acts

on the restriction v_h, the $2^{nd}, \cdots, m^{th}$ entries in $A_h v_h$ approximate values of v_x and can therefore be bounded independently of h. However, the first entry in $A_h v_h$ will behave like h^{-1} as $h \to 0$ unless v satisfies the *homogeneous* boundary condition $v(0)=0$. Likewise, the $3^{rd}, \cdots, m^{th}$ entries in $A_h^2 v_h$ approximate values of v_{xx} and are thus bounded. However, the first and second entries in $A_h^2 v_h$ will only be bounded if $v(0)=v_x(0)=0$. The general trend should now be clear. For $l-1=1,2, \cdots, s-2$ (= the highest power of A_h which occurs in the bounds (4.9)), $\|A_h^{l-1} v_h\|$ is bounded in h if

$$\frac{\partial^k v(0)}{\partial x^k}=0, k=0,1, \cdots, l-2.$$

In general, $\|A_h^{l-1} v_h\|_2$ behaves like $h^{(3/2-l)} (l \geq 2)$ while in L^∞ we have the behaviour $h^{(1-l)}$.

Next, since the highest power of A_h in the bounds (4.9) is $(s-2)$, it follows that the optimal exponent $g=p$ in (4.10) will be obtained if the theoretical solution $u(x,t)$ satisfies $s-2$ boundary requirements

$$u(0,t)=0,\ u_x(0,t)=0, \cdots, (\partial^{s-3}/\partial x^{s-3}) u(0,t)=0$$

that render it possible for $A_h^{l-1} u_h^{(j)}, A_h^{l-1} u_h^{(j+1)}$ $(1 \leq l \leq s-1, 2 \leq j \leq p,\ l+j \geq p+1)$ to remain bounded uniformly in h. These $s-2$ boundary requirements for u will be satisfied if and only if f_Ω, f_Γ do not violate a set of $s-2$ constraints

$$f_\Gamma \equiv 0,\quad f_\Omega(0,t)=0, \cdots,\quad (\partial^{s-4}/\partial x^{s-4}) f_\Omega(0,t)=0.$$

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. If one or more of the constraints are not satisfied the exponent g in (4.10) can be found by a simple inspection of the differentials featuring in (4.9) ($\mu_{ij} \neq 0$). The one with the largest reduction will determine g.

Finally, we have tacitly assumed that the number of stages s is greater than or equal to three. No reduction will take place with a 2-stage method, and, of course, with the Eul method.

For the time integration of the semidiscretization (2.5) we consider the classical stage, 4-th order scheme

0	0			
1/2	1/2	0		
1/2	0	1/2	0	
1	0	0	1	0
	1/6	1/3	1/3	1/6

From its local error expression [14]

$$\beta^{n+1}=\left(\frac{1}{576} A_h u_h^{(4)}+\frac{-1}{288} A_h^2 u_h^{(3)}+\frac{1}{96} A_h^3 u_h^{(2)}\right)\tau^5+$$

$$\left(\frac{-1}{1152} A_h^2 u_h^{(4)}+\frac{1}{576} A_h^3 u_h^{(3)}\right)\tau^6+\frac{1}{4608} A_h^3 u_h^{(4)} \tau^7+\sum_{i=2}^{5} Q_i(\tau A_h) R_i,$$

we deduce that none of the coefficients $\mu_{lj}(1\leqslant l\leqslant 3, 2\leqslant j\leqslant 4,\ l+j\geqslant 5)$ is zero so that all grid functions which feature in (4.9) may contribute to a reduction of the order. Obviously, the largest reduction will emanate from the two terms in (4.9) with $(l+j)$ minimal and $(l-1)$ maximal, which are here $\tau^4 A_h^2 u_h^{(2)}$ and $\tau^4 A_h^2 u_h^{(3)}$. Hence if the additional boundary requirements mentioned above ($v(0)=v_x(0)=0$, $v=u_h^{(2)}$, $v=u_h^{(3)}$) are not satisfied and τ/h is kept fixed (cf. (2.7)), we will have to face a reduction in global order from 4 to 2 if we measure in L^∞ and from 4 to 2.5 if we measure in L^2.

4.5 A numerical illustration

We have applied the above RK method to the semidiscretization (2.5) with the choice $u(x,0)=1+x$, $f_\Gamma(t)=1/(1+t)$, $f_\Omega(x,t)=(t-x)/(1+t)^2$ which yields the simple solution $u(x,t)=(1+x)/(1+t)$. Since this solution is linear in space, there is no error introduced by the space discretization ($\alpha_h\equiv 0$). The time derivatives of u are *not* zero at the boundary so that the reduction mentioned in the preceding example should occur.

The floating point numbers in the table below are the L^∞- errors at $t=1$ for certain values of τ,h. The fixed point numbers represent the observed orders of convergence upon either *simultaneous* halving of τ,h (the numbers in italics) or halving τ on a *fixed* grid. Recall that these computed orders are given by the expression $\log_2$ (error ratio).

h^{-1} / τ^{-1}	10		20		40		80
10	$.69_{10^{-4}}$						
	4.7	*2.1*					
20	$.26_{10^{-5}}$		$.16_{10^{-4}}$				
	4.2	*2.0*	4.7	*2.1*			
40	$.15_{10^{-6}}$		$.65_{10^{-6}}$		$.40_{10^{-5}}$		
	4.1	*2.2*	4.3	*2.0*	4.6	*2.0*	
80	$.85_{10^{-8}}$		$.32_{10^{-7}}$		$.16_{10^{-6}}$		$.97_{10^{-6}}$

For the simultaneous refinement the anticipated reduction from 4 to 2 is clearly seen. 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 the table.

An illustration of order reduction occurring in a RK-finite element scheme applied to problem (2.4) can be found in [14]. The interested reader should also consult [17] where examples with implicit RK methods applied to parabolic problems are discussed.

4.6 Some concluding remarks on order reduction

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 extension of the analysis to the nonlinear case is possible but becomes rather technical and offers no new insight.

As mentioned earlier, 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 hypotheses (H2) should be forsaken. The details of the analysis become

then quite different [1,17].

A simple means for avoiding order reduction has been suggested and tested in [14]. It is based on reformulating the PDE problem, prior to the space discretization, so that the additionally required boundary conditions are satisfied.

Finally, 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.

References

[1] P. Brenner, M. Crouzeix & V. Thomée, *Single step methods for inhomogeneous linear differential equations in Banach space,* R.A.I.R.O. Analyse numérique 16, 5-26, 1982.

[2] K. Burrage, W.H. Hundsdorfer & J.G. Verwer, *A study of B-convergence of Runge-Kutta methods,* Computing, to appear.

[3] K. Dekker, & J.G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations,* North-Holland, Amsterdam-New York-Oxford, 1984.

[4] R. Frank, J. Schneid & C.W. Ueberhuber, *Order results for implicit Runge-Kutta methods applied to stiff systems,* SIAM J. Numer.Anal. 22, 515-534, 1983.

[5] E. Hairer, G. Bader & CH. Lubich, *On the stability of semi-implicit methods for ordinary differential equations,* BIT 22, 211-232, 1982.

[6] H.O. Kreiss, *Ueber die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren,* BIT 2, 153-181, 1962.

[7] J. von Neumann, *Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes,* Mathematischen Nachrichten 4, 258-281, 1951.

[8] C. Palencia & J.M. Sanz-Serna, *Equivalence theorems for incomplete spaces: an appraisal.* IMA. J. Numer. Anal. 4, 109-115, 1984.

[9] C. Palencia & J.M. Sanz-Serna, *An extension of the Lax-Richtmeyer theory,* Numer. Math. 44, 279-283, 1984.

[10] R.D. Richtmyer & K.W. Morton, *Difference methods for initial value problems,* Interscience Publishers, New York-London-Sydney, 1967.

[11] J.M. Sanz-Serna, *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.).

[12] J.M. Sanz-Serna, *Stability and convergence in Numerical Analysis I: Linear problems, a simple, comprehensive account,* in Nonlinear differential equations and applications, J. Hale and P. Martinez-Amores (eds.), Pitman, Boston, pp. 64-113, 1985.

[13] J.M. Sanz-Serna & C. Palencia, *A general equivalence theorem in the theory of discretization methods,* Math. Comp. 45, 143-152, 1985.

[14] J.M. Sanz-Serna, J.G. Verwer & W.H. Hundsdorfer, *Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations,* Report NM-R8525, Centre for Math. and Comp. Sc., Amsterdam, 1985.

[15] J.M. Sanz-Serna & M.N. Spijker, *Regions of stability, equivalence theorems and the Courant-Friedrichs-Lewy condition,* to appear in Numer. Math.

[16] M.N. SPIJKER, *Stepsize restrictions for stability of one-step methods in the numerical solution of initial value problems,* Math. Comp. 45, 377-392, 1985.

[17] J.G. VERWER, *Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines,* Report NM-R8506, Centre for Math. and Comp. SC., Amsterdam, 1985 (to appear in Proc. Dundee Num. Anal. Conf. 1985).

[18] J.G. VERWER & J.M. SANZ-SERNA, *Convergence of method of lines approximations to partial differential equations,* Computing 33, 297-313, 1984.