Applied Numerical Mathematics 5 (1989) 117–132
North-Holland

STABILITY AND CONVERGENCE AT THE PDE/STIFF ODE INTERFACE

J.M. SANZ-SERNA
Departamento de Ecuaciones Funcionales, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain

J.G. VERWER
Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

This is an expository paper showing the interplay between the analysis of numerical methods for evolutionary partial differential equations and some developments in the stiff ordinary differential equation literature. The notions of contractivity, one-sided Lipschitz conditions, logarithmic norms, B-convergence and order reduction are of particular importance.

1. Introduction

The last decade has witnessed a large development in the analysis of numerical methods for linear and nonlinear stiff systems of ordinary differential equations (ODEs). The simple scalar test equation $u' = \lambda u$ has been supplemented by more general model systems, like (following Dahlquist [9] and Butcher [7]) dissipative systems. We know at present the behaviour of large classes of one-step or multistep ODE methods when applied to such more demanding tests. The notions of logarithmic norm (Dahlquist [8]) and one-sided Lipschitz condition have assumed an ever-increasing role etc. The developments have not been confined to the issue of stability. For one-step schemes the structure of the local error is now much better understood than before, mainly due to the B-convergence theory of Frank, Schneid and Ueberhuber [11–13] (which extends earlier work by Prothero and Robinson, see [10]). For one-step methods, Dekker and Verwer [10] have gathered together most of the results we are referring to.

The purpose of this expository paper is to show the relevance of this stiff ODE material for the field of analysis of numerical methods in partial differential equations (PDEs). We have in our mind several possible interactions. People with an ODE background may wish to apply their material to concrete PDE cases or may like to know the sort of ODE result that would be more beneficial to the PDE research. People with a PDE background should know that the recent stiff ODE literature can help them considerably.

The paper is confined to one-step (two-level) discretizations. Furthermore, and in order to keep within reasonable bounds, the exposition is centered around *contractive, linear non-autonomous problems*. The final Section 6 discusses the extensions to more general situations. Sections 2, 3 and 4 are devoted, respectively, to the PDEs to be solved, to their discretizations in space and to their discretizations in time. Our treatment emphasizes the parallelism between these three realms. Section 5 examines certain aspects of the local error, notably the *order reduction* phenomenon, which renders possible for the order of convergence in time of a PDE scheme to be strictly lower than the classical order of the ODE method used for the integration in time.

2. Initial value problems in PDEs

In this section we introduce the PDE problems to be considered in the rest of the paper. With those readers without a large PDE background in our mind, the discussion is slightly lengthier than it would be necessary otherwise.

2.1. A simple example

We begin by presenting one of the simplest examples of a time-dependent PDE problem, namely

$$u_t = u_{xx}, \qquad 0 \leqslant x \leqslant 1, \quad t \geqslant 0, \tag{2.1a}$$

$$u(0, t) = u(1, t) = 0, \quad t \geqslant 0, \tag{2.1b}$$

$$u(x, 0) = u_0(x), \quad 0 \leqslant x \leqslant 1, \tag{2.1c}$$

where the initial datum u_0 is in $L^2(0, 1)$, i.e. is a real, square-integrable function in $0 \leqslant x \leqslant 1$. The method of separation of variables leads to the solution

$$u(x, t) = \sum_{n=1}^{\infty} a_n \exp(-n^2\pi^2 t) \sin(n\pi x), \tag{2.2}$$

where the a_n are the (sin) Fourier coefficients of u_0, i.e.

$$u_0(x) = \sum_{n=1}^{\infty} a_n \sin(n\pi x). \tag{2.3}$$

From these expressions, and recalling that the L^2-norm of a function equals the square root of the sum of the squares of its Fourier coefficients, it follows easily that, if we consider a second initial datum v_0 and denote by $v = v(x, t)$ the corresponding solution, then, for each two nonnegative times t and s with $t > s$,

$$\| u(\cdot, t) - v(\cdot, t) \|_{L^2_{(0,1)}} \leqslant \| u(\cdot, s) - v(\cdot, s) \|_{L^2_{(0,1)}}. \tag{2.4}$$

Here a symbol like $u(\cdot, t)$ represents the corresponding function of x obtained by fixing the time at the value t. Thus, problem (2.1) is such that: (i) To each initial datum there corresponds a unique solution (2.2). (ii) According to (2.4) solutions u and v stemming from two different initial data become closer to each other in the evolution $s \to t$, a behaviour called *contractivity*. Of importance is the fact that contractivity guarantees that small changes in the datum lead to small changes in the solution.

However, not everything is plain sailing in the considerations above. In fact, it is well known that the series in (2.3) does not necessarily converge in the pointwise sense and the same must be true for that in (2.2), which reduces to (2.3) when $t = 0$. Therefore the right-hand side of (2.2) does not generally define a continuous function of x and t and the solutions we have been referring to are only *generalized* solutions. (See [22, Section 3.1; 21,26,27] for further discussion of this important point. Recall that generalized solutions may possess physical importance.) To have genuine solutions, i.e. solutions for which u_t and u_{xx} exist and for which the relations in (2.1) hold, it is necessary to impose additional conditions on u_0. These conditions are of two sorts: (i) u_0 should possess continuous derivatives. (ii) u_0 should satisfy certain *compatibility*

conditions with the boundary information. Since compatibility conditions play an important role later in the paper, it is appropriate to comment on them. We first note that if the solution u of (2.1) is smooth, it satisfies, not only the relations (2.1b), but also

$$(\partial^2/\partial x^2)u(0, t) = (\partial^2/\partial x^2)u(1, t) = 0, \quad t \geqslant 0, \tag{2.5a}$$

$$(\partial^4/\partial x^4)u(0, t) = (\partial^4/\partial x^4)u(1, t) = 0, \quad t \geqslant 0, \tag{2.5b}$$

$$\vdots$$

a hierarchy of conditions at the boundary that follows by noticing that from (2.1b) we can write

$$(\partial/\partial t)u(0, t) = (\partial/\partial t)u(1, t) = 0, \qquad (\partial^2/\partial t^2)u(0, t) = (\partial^2/\partial t^2)u(1, t) = 0, \ldots$$

while, from differentiation of (2.1a), $(\partial^2/\partial t^2)u = (\partial^4/\partial x^4)u, \ldots$. On taking into account (2.1c) we can conclude that for the solution u to be smooth it is necessary that the initial datum u_0 satisfies successively

$$u_0(0) = u_0(1) = 0,$$

$$(\partial^2/\partial x^2)u_0(0) = (\partial^2/\partial x^2)u_0(1) = 0,$$

$$(\partial^4/\partial x^4)u_0(0) = (\partial^4/\partial x^4)u_0(1) = 0,$$

$$\vdots$$

In general, the smoothness of the solution (2.2) increases as the number of fulfilled compatibility conditions and the number of continuous derivatives of u_0 increase.

2.2. *Abstract formulation*

It is useful to recast problem (2.1) in the following abstract form. We set $X = L^2(0, 1)$ and denote by D the subspace of X consisting of functions w for which (i) w'' exists and is square-integrable and (ii) the homogeneous boundary conditions $w(0) = w(1) = 0$ hold. Furthermore, we introduce the (linear) operator A in X, with domain D, that maps each w belonging to D into its second derivative w''. With this notation, (2.1) can obviously be rewritten in the compact form.

$$\mathrm{d}u/\,\mathrm{d}t = Au, \quad t \geqslant 0, \tag{2.6a}$$

$$u(0) = u_0. \tag{2.6b}$$

The fact that the formulae (2.6) have the appearance of an initial value problem for a linear system of ODEs should not hide the following features which make the problem at hand essentially different from any system of ODEs.

(i) The eigenvalues of the operator A (i.e. of the two-point boundary value problem $w'' = \lambda w$, $w(0) = w(1) = 0$) are given by $-n^2\pi^2$, $n = 1, 2, 3, \ldots$, and therefore are negative but with arbitrary large magnitude. In this sense, (2.6) possesses *infinite stiffness*.

(ii) As a consequence, A cannot satisfy in the L^2-norm a (classical) Lipschitz condition

$$\| Aw_1 - Aw_2 \| \leqslant L \| w_1 - w_2 \|, \tag{2.7}$$

(just take for $w_1 - w_2$ the nth normalized eigenfunction, then the left-hand side equals $-n^2\pi^2$ while the right-hand side equals L, so that the inequality cannot hold).

In functional analysis jargon, A is a densely defined *unbounded* operator. Typically, parabolic problems lead to the infinite stiffness situation, while hyperbolic problems often possess purely imaginary eigenvalues of arbitrarily large magnitude.

In principle, it is possible to discretize (2.6) in time by means of any of the standard ODE methods. However, we should note in this connection that the convergence of such a discretization cannot be immediately guaranteed: the classical theory of ODE methods (e.g. [15,16]) relies heavily on the use of a Lipschitz condition like (2.7), something which is not available here. See [3,4] for examples of treatments of time discretizations (without space discretization) of PDEs like (2.6).

2.3. *Well-posed contractive problems*

In the remainder of the paper, we let Ω be a bounded domain in $\mathbb{R}^d$ and let X be a Banach space composed of functions defined in Ω and taking values in $\mathbb{R}^s$. We denote by A a time-independent linear differential operator which differentiates the functions of X with respect to the d spatial variables (the coefficients of A may depend on the space variables). With this notation (2.6) represents now a system of s partial differential equations for the s components of u. It is assumed that appropriate *homogeneous* boundary conditions have been incorporated by suitably restricting the domain of A, just as we did for the heat equation example. This abstract formulation can include both parabolic and hyperbolic problems. We suppose that A is such that:

(H1) To each u_0 in X there corresponds a unique (possibly generalized) solution of (2.6).
(H2) For solutions of (2.6) the following contractivity property holds

$$\| u(t) - v(t) \|_X \leqslant \| u(s) - v(s) \|_X, \quad t > s > 0. \tag{2.8}$$

A necessary and sufficient condition for these requirements on A to hold is given in the Hille–Yoshida–Philips theorem (see e.g. Aubin [1, Chapter 14] or Kato [17, Chapter 9]). Under some auxiliary technical hypotheses, the condition is

$$\text{for each } \tau > 0,\ I - \tau A \text{ is invertible and } \| (I - \tau A)^{-1} \| \leqslant 1. \tag{2.9}$$

This requirement has an interesting numerical analysis interpretation: $(I - \tau A)^{-1}$ is the operator which maps each element w in X into the result of a step, starting from w, of the backward Euler rule applied to (2.6) with step length τ.

When the norm in X derives from an inner product $\langle \cdot, \cdot \rangle$, it is possible to substitute (2.9) by the following *dissipativity* condition:

$$\text{for each } x \text{ in the domain of } A,\ \langle x, Ax \rangle \leqslant 0. \tag{2.10}$$

Dissipativity conditions are a particular instance of *one-sided* Lipschitz conditions, see e.g. [10] (recall that here classical two-sided Lipschitz conditions like (2.7) do not hold). For conditions analogous to (2.10) and valid when X is not an inner product space, see e.g. [30].

It is not difficult to show [17] that (2.9) or (2.10) imply in particular that the spectrum of A does not intersect the positive half-plane $\mathrm{Re}(\lambda) > 0$. However this *spectral requirement is not, in general, sufficient* to guarantee the contractivity of (2.6). An exception is given by the situation where X is an inner product space and the operator A is normal: in this case (2.9) and (2.10) are equivalent to the condition that the spectrum of A (which is real) lies in $\lambda \leqslant 0$.

2.4. Non-autonomous problems

So far the problems considered have been autonomous, since A in (2.6) has been assumed to be independent of t. The class of linear, autonomous problems is too narrow to display some important aspects of our subject. We therefore introduce the slightly more general problem,

$$du/dt = Au + f(t), \quad t \geqslant 0, \tag{2.11a}$$

$$u(0) = u_0, \tag{2.11b}$$

where f is a function of t taking values in X. We assume that A satisfies the hypotheses (H1)–(H2) mentioned in Section 2.3 and that f is smooth. In this case (see e.g. [1, Chapter 14]) the problem (2.11) possesses for each u_0 a unique (possibly generalized) solution. Furthermore, if $u(t)$ and $v(t)$ are two solutions stemming from the two initial data u_0 and v_0, the contractivity property (2.8) holds: this follows immediately from (H2) and the fact that the difference of two solutions of (2.11a) is a solution of the homogeneous equation (2.6a).

A simple example on non-autonomous problem (2.11) is given by

$$u_t = u_{xx} + f(x, t), \qquad 0 \leqslant x \leqslant 1, \quad t \geqslant 0, \tag{2.12}$$

along with the boundary conditions (2.1b) and initial condition (2.1c). On differentiating (2.12) we find that, in this case, a smooth solution satisfies

$$(\partial^2/\partial x^2)u(0, t) + f(0, t) = (\partial^2/\partial x^2)u(1, t) + f(1, t) = 0, \qquad t \geqslant 0, \tag{2.13a}$$

$$(\partial^4/\partial x^4)u(0, t) + (\partial^2/\partial x^2)f(0, t) + (\partial/\partial t)f(0, t)$$
$$= (\partial^4/\partial x^4)u(1, t) + (\partial^2/\partial x^2)f(1, t) + (\partial/\partial t)f(1, t) = 0, \quad t \geqslant 0, \tag{2.13b}$$

$$\vdots$$

rather than (2.5). The relations (2.13) induce appropriate compatibility conditions that u_0 and f should fulfill if the solution is to be smooth.

3. Space discretization

The discretization in space of our PDE problem (2.11), by means of finite differences, results in an initial value ODE problem

$$dU_h/dt = A_h U_h + f_h(t), \quad t > 0, \tag{3.1a}$$

$$U_h(0) = U_{0h}, \tag{3.1b}$$

where h is the parameter of a grid in the closure of Ω having, say, m points; $U_h = U_h(t)$ is an array with m components consisting of approximations to u at the grid points (note that each component of U_h is in turn an s-dimensional real vector). The $(m \times s)$-dimensional, real square matrix A_h, the inhomogeneous term $f_h(t)$ and the initial condition U_{0h} result from discretization of A, $f(t)$ and u_0 respectively. Note that the dimension of U_h increases with decreasing h. Finite element and spectral space discretizations can be catered for with very minor modifications (see [25]) and will not be treated here. We assume that a norm $\|\cdot\|$ for $(m \times s)$-dimensional real vectors has been chosen which is a discrete analogue of the norm employed in the space of functions X.

As an illustration of the foregoing notation we consider the central difference discretization of the single ($s = 1$) equation (2.12) on an equidistant grid $x_i = ih$, $i = 1(1)m$, $h = 1/(m+1)$. The matrix A_h takes the well-known form

$$A_h = h^{-2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{bmatrix}, \tag{3.2}$$

$f_h(t)$ is the grid restriction of $f(x, t)$ and a suitable discrete L^2-norm is given by

$$\| V \| = \left(\sum_{i=1}^{m} h | V_i |^2 \right)^{1/2}.$$

The eigenvalues of A_h are

$$h^{-2}(-2 + 2\cos(n\pi/(m+1)), \quad n = 1(1)m.$$

Like those $-n^2\pi^2$ of the PDE, they are negative, but, of course, now there are only a finite number of them. The norm of A_h is given by the smallest eigenvalue and therefore is less than $4h^{-2}$. Thus A_h satisfies a classical Lipschitz condition (2.7) with constant $L_h = 4h^{-2}$. Of importance is the fact that this constant detoriates as h is decreased, something which could have been anticipated by observing that in the limit $h \to 0$ the matrices A_h approximate the PDE operator, for which (2.7) does not hold.

3.1. Contractive semidiscretizations

Leaving the heat example and returning to the general problem (2.11) and its semidiscretization (3.1), we observe that it is very natural to require that (3.1) should also satisfy a contractivity condition, thus imitating a property of the problem being discretized. More precisely, we say that (3.1) is contractive if, for two solutions U_h and V_h of (3.1a), stemming from two different initial conditions, and for each nonnegative times t and s with $t > s$,

$$\| U_h(t) - V_h(t) \| \leqslant \| U_h(s) - V_h(s) \|. \tag{3.3}$$

Note that, being linear, the problem (3.1a) possess a unique solution for all positive t, provided that f_h is continuous. The Hille–Yoshida–Philips theorem applies, in particular, to matrices and, therefore, the conditions (2.9) and (2.10) (with A_h instead of A) still characterize the contractive behaviour (3.3). It is again true that the *spectral* requirement that A_h has all its eigenvalues in $\mathrm{Re}(\lambda) \leqslant 0$ is necessary for (3.3) to hold, but (unless A_h is normal and we use an inner-product norm) it is not sufficient. Note in passing that this shows the contractivity of the heat equation semidiscretization studied above.

A further characterization of the contractivity of (3.1), not available for the PDE itself, uses the notion of logarithmic norm $\mu[A_h]$ of A_h defined by

$$\mu[A_h] = \lim_{\tau \to 0^+} \frac{\| I + \tau A_h \| - 1}{\tau}.$$

This number, which depends on the matrix norm being employed, was introduced by Dahlquist

in his thesis [8] and independently by Lozinskij (see [10]) and has the important property of being the smallest constant C for which

$$\| \exp(\xi A_h) \| \leqslant \exp(C\xi), \quad \xi > 0. \tag{3.4}$$

Upon noticing that, in (3.3),

$$U_h(t) - V_h(t) = \exp((t-s)A_h)(U_h(s) - V_h(s)),$$

we conclude that the condition

$$\mu[A_h] \leqslant 0 \tag{3.5}$$

is necessary and sufficient for (3.3) to hold. The condition (3.5), unlike (2.10), is valid in norms which not necessarily stem from an inner product. Also (3.5) can be checked in practice more easily than (2.9), since closed-form expressions exist for $\mu[A_h]$ in the most commonly employed norms [10].

3.2. *Convergence of the semidiscrete solution*

For the convergence analysis of this subsection, we suppose that (2.11) possesses a smooth genuine solution $u(x, t)$ and denote by $u_h(t)$ its restriction to the grid (generalized solutions could also be considered in convergence analysis, see [20,22,26,27]). We restrict our attention to a bounded time interval $0 \leqslant t \leqslant T$ and say that (3.1) is *convergent* if

$$\max_{0 \leqslant t \leqslant T} \| u_h(t) - U_h(t) \| = o(1) \quad \text{as } h \to 0,$$

provided that $\| u_h(0) - U_h(0) \| = o(1)$. Convergence of order p^* is defined in the obvious way, i.e. replacing o(1) by $O(h^{p^*})$ in both occurrences of the symbol o(1). For simplicity we assume hereafter that $U_h(0)$ is taken to be $u_h(0)$, so that there is no error in approximating the initial function.

The vector $u_h(t) - U_h(t)$ is referred to as the *global error* in the semidiscretization. Also of interest is the *truncation error* of (3.1) defined by

$$\alpha_h(t) = A_h(t)u_h(t) + f_h(t) - (\mathrm{d}/\mathrm{d}t)u_h(t), \tag{3.6}$$

which, in practical settings, can be easily bounded by means of a simple Taylor expansion (recall that $u_h(t)$ is, by assumption, smooth). The semidiscretization is *consistent* if

$$\max_{0 \leqslant t \leqslant T} \| \alpha_h(t) \| = o(1) \quad \text{as } h \to 0,$$

with consistency of order p^* defined again in the obvious way. (Our heat equation semidiscretization is, of course, consistent of the second order.)

So far the notion of contractivity of (3.1) has been introduced as a desirable property that the semidiscretization should satisfy in order to imitate the corresponding PDE property. The main result of this section is that, for a (p^*-) consistent semidiscretization, contractivity implies (p^*-) convergence. To prove this, subtract (3.6) from (3.1a) to arrive at

$$(\mathrm{d}/\mathrm{d}t)(u_h(t) - U_h(t)) = A_h(u_h(t) - U_h(t)) - \alpha_h(t)$$

and use the variation of constant formula

$$u_h(t) - U_h(t) = \mathrm{e}^{tA_h}(u_h(0) - U_h(0)) - \int_0^t \mathrm{e}^{\xi A_h}\alpha_h(t-\xi)\,\mathrm{d}\xi.$$

As noted before, contractivity is equivalent to $\|\exp(\xi A_h)\| \leqslant 1$ and convergence thus follc easily. Note that contractivity is *not* necessary for convergence: the previous proof also hc under the less demanding hypothesis that $\mu[A_h]$ can be bounded above by a constant *indep dent of* h, so that $\|\exp(\xi A_h)\|$ can be bounded independently of h (cf. (3.4)). Examples convergence proofs of semidiscretizations along the previous lines can be seen in [33].

4. Time discretization

In order to get numerical approximations to the solution u of (2.11), the semidiscretizat (3.1) must be integrated in time. We suppose that this is done by means of a converge pth-order *one-step* ODE solver, with a constant time step τ, leading to a recursion

$$U^{n+1} = R(\tau A_h)U^n + F_n, \quad n = 0, 1, 2, \ldots, \qquad U^0 \text{ given}, \tag{4}$$

where $R(\cdot)$ is the so-called stability function associated with the method and F_n is a vec originating from the nonhomogeneous term of (3.1).

4.1. Contractive time stepping

In the remainder of the paper, it is always assumed that the semidiscretization (3.1) onsistent and contractive (and hence convergent). Once more it is natural to demand that ılly discrete solution U^n also exhibits a contractive behaviour, or more precisely that, if V^n i second sequence, satisfying the recursion (4.1),

$$\|U^{n+1} - V^{n+1}\| \leqslant \|U^n - V^n\|, \quad n = 0, 1, 2, \ldots. \tag{4}$$

Clearly a necessary and sufficient condition for (4.2) to hold is that

$$\|R(\tau A_h)\| \leqslant 1. \tag{4}$$

The question thus arises of how to choose the ODE method and the value of τ, so that ti stepping in the contractive system (3.1) leads to contractive fully discrete solutions. It is probal here that the recent literature in stiff ODEs is helpful to the PDE researcher:

(C1) The implicit Euler rule performs contractively when applied to any contractive OI problem, regardless of the value of τ and of the norm employed. This follows trivially from c remark after formula (2.9). A direct proof can be seen in [10, pp. 46–47].

(C2) Spijker [30] has shown that if an ODE method behaves contractively for any OI problem, for any norm and any τ, then its order p cannot exceed 1.

(C3) If the matrix A_h and the value τ_0 are such that $\|(I + \tau_0 A_h)\| \leqslant 1$ with τ_0 maximal (i the explicit Euler rule with step τ_0 behaves contractively), then Spijker [31] shows that a one-st method applied to (3.1) behaves contractively for any step size $\tau \leqslant r\tau_0$, with r the so-call *contractivity radius* of the method (see [31]). The upper bound on τ is optimal, in the sense th matrices A_h exist such that violation of the bound results in lack of contractivity. Here the noı can be arbitrary; Spijker provides an interesting application to a convection-diffusion proble studied in the maximum norm.

(C4) The negative result in (C2) is, to some extent, counterbalanced by the fact that the exist implicit Runge–Kutta (RK) methods of arbitrarily high order that perform contractive

for any value of τ and any contractive ODE problem, *provided* that the norm considered derives from an *inner product*. In fact the literature on contractivity of RK methods is very well developed, starting with the paper by Butcher [7] which was in turn motivated by work by Dahlquist [9] on multistep methods. To survey all the contributions by Burrage, Butcher, Crouzeix, Dahlquist, Hairer, Hundsdorfer, Jeltsch, Spijker, Wanner and others is out of the scope of this article and the reader is referred to [10, Chapter 4].

(C5) Since the norm of a matrix always exceeds its spectral radius, (4.3) implies that, for contractivity, all the eigenvalues of $R(\tau A_h)$ should be in modulus $\leqslant 1$. These eigenvalues are given by $R(\tau\lambda_h)$ with λ_h an eigenvalue of A_h, and, as a consequence, the *spectral condition*

$$\tau \text{ is such that the products } \tau\lambda_h,\ \lambda_h \text{ an eigenvalue of } A_h, \text{ belong to the region of absolute stability of the ODE method } \{z\colon |R(z)| \leqslant 1\} \tag{4.4}$$

is *necessary* for contractivity in any norm.

(C6) In the case of *inner-product norms and normal* A_h, the norm actually equals the spectral radius and (4.4) is also *sufficient* for contractivity. An interesting corollary of this result is that, for an A-stable method no restriction on τ is needed. As a further application of the sufficiency of (4.4), it is trivial to show that the explicit Euler time stepping applied to our heat equation semidiscretization (or in other words the standard explicit method for the heat equation) is contractive (in L^2) provided that $\tau/h^2 \leqslant \frac{1}{2}$. The implicit Euler and trapezoidal rules, being A-stable, behave L^2-contractively in our model problem, regardless of the value of τ.

(C7) It should be emphasized that in general the spectral condition (4.4) guarantees contractivity only under the stated hypotheses, namely that we work with an inner-product norm and that A_h is normal. As we will discuss later, attempts to use it outside of this setting may result in a catastrophic error propagation. However, it is possible to use a deep theorem due to von Neumann [19,23] to show that A-stable methods behave contractively for any step size when applied to contractive linear ODEs provided that the norm is of the inner-product type. Note that the normalilty of the matrix is not required, as distinct from (C6). Spijker [31, Theorem 6.1] provides a further application of von Neumann's result to contractivity studies. See also [14].

4.2. Convergence of the fully discrete solutions

We now position ourselves in the setting of Section 3.2, where (2.11) possesses a *smooth* solution u with grid restriction u_h and the interest is confined to a bounded time interval $0 \leqslant t \leqslant T$. We study the convergence of the fully discrete solutions, i.e. we wish to know whether

$$\max_{0 \leqslant n\tau \leqslant T} \| u_h(n\tau) - U^n \| = o(1),$$

as h and τ tend to zero subject perhaps to appropriate restrictions. For simplicity we assume hereafter that there is no error in approximating the initial condition so that $U^0 = u_h(0)$. The hypotheses made so far, namely that the semidiscretization is contractive and consistent (and hence convergent) and that the ODE method is convergent, do not guarantee by themselves such a fully discrete convergence if h and τ tend independently to zero. For example consider the explicit heat equation method mentioned above, where it is well known that the supplementary hypothesis $\tau/h^2 \leqslant \frac{1}{2}$ must be imposed to obtain convergence. More generally, consider the inequality

$$\| u_h(n\tau) - U^n \| \leqslant \| u_h(n\tau) - U_h(n\tau) \| + \| U_h(n\tau) - U^n \|. \tag{4.5}$$

The convergence of the semidiscretization implies that the first term in the right-hand side tends to 0 as $h \to 0$. For a convergent ODE solver $\| U_h(n\tau) - U^n \|$ tends to 0, as $\tau \to 0$, for *fixed* h. However, the system (3.1) 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 the convergence of the ODE solver be *uniform*, as h varies, in the family of problems (3.1). Such a uniformity cannot be established by means of the classical straightforward bounds for ODE solvers [15,16] as those bounds typically include factors $\exp(L_h n\tau)$, where L_h is the classical Lipschitz constant for A_h, and we know that L_h must grow with decreasing h. The idea of error bounds that hold uniformly for whole classes of stiff problems has been dominant in the recent ODE literature; see notably the B-convergence theory of Frank, Schneid and Ueberhuber [11–13].

A sufficient condition for the convergence of the fully discrete approximations will be presented next. We emphasize that we do not work with the splitting (4.5): any conceivable bound for $\| U_h(n\tau) - U^n \|$ would involve estimating the derivatives of the semidiscrete solution U_h and this is something we prefer to avoid [33]. In what follows, the time space grids are refined subject to a condition

$$\tau \leqslant r h^q \tag{4.6}$$

with $0 < q < \infty$, $0 < r \leqslant \infty$ ($r = \infty$ means, of course, no restriction).

We introduce the full truncation error of (4.1) defined by

$$\beta^{n+1} = u_h((n+1)\tau) - R(\tau A_h) u_h(n\tau) - F_n, \tag{4.7}$$

and say that the fully discrete method (4.1) is (fully) *consistent*, if, as τ and h tend to zero subject to (4.6),

$$B = \max_{0 \leqslant n\tau \leqslant T} \| \beta^{n+1} \| = o(\tau). \tag{4.8}$$

It is easy to prove that, if (i) a fully discrete method is consistent, and (ii) as τ and h vary subject to (4.6), the ODE solver with step τ is contractive on the problem (3.1), then the fully discrete method is convergent. To see this, subtract (4.7) from (4.1) to get

$$u_h((n+1)\tau) - U^{n+1} = R(\tau A_h)(u_h(n\tau) - U^n) + \beta^{n+1}, \quad n = 0, 1, \ldots, [T/\tau] - 1,$$

by contractivity and (4.8)

$$\| u_h((n+1)\tau) - U^{n+1} \| \leqslant \| u_h(n\tau) - U^n \| + B, \quad n = 0, 1, \ldots, [T/\tau] - 1,$$

and induction shows that

$$\| u_h(n\tau) - U^n \| \leqslant \| u_h(0) - U^0 \| + nB = nB \leqslant (T/\tau) B = o(1). \tag{4.9}$$

The requirement of contractivity, which we have just shown to be sufficient for the convergence of consistent fully discrete approximations, is not necessary. The minimal requirement leading to ıch a convergence is that of Lax stability

$$\sup \{ \| R(\tau A_h)^n \| : \tau, h \text{ subject to (4.6)}, 0 \leqslant n\tau \leqslant T \} < \infty$$

4.3) and see [20–22,26–28].

ry often in the literature the spectral condition (4.4) is used as a criterion for choosing the step. In cases where (4.4) does not imply contractivity (i.e. cases where the matrices are not

normal or we work with norms not deriving from inner products) this spectral condition is likely not to imply Lax stability and, therefore does not guarantee convergence. The numerical results reported in [10, pp. 273–274] are very illuminating in this connection: the spectral condition ensures that on any fixed grid any error will be eventually damped as t increases without bound, but makes possible for the errors to grow catastrophically prior to that damping. Useful references in this area are e.g. [18,24].

5. The structure of the full truncation error: Order reduction

We have just seen in (4.9) that, for contractive time steppings, the error $u_h(n\tau) - U^n$ can be readily bounded, once bounds for the full truncation error β^{n+1} are available. The question remains of how to estimate this truncation error. Our aim is to derive, under reasonable hypotheses, bounds for $\| \beta^{n+1} \|$ of the form

$$C\Big(\tau^k + \tau \max_{0 \leqslant t \leqslant T} \| \alpha_h(t) \|\Big) \tag{5.1}$$

where k is a positive number and C denotes a constant, depending on T and on the smoothness in time of the PDE solution u, but independent of τ, h and n, $0 \leqslant n\tau \leqslant T$. Since our ODE method has been assumed to be of order p, we would naively expect that in (5.1) k can be taken equal to $p + 1$. However the fact is that often the exponent k can only be taken to be less than $p + 1$, so that the order of convergence in time of the fully discrete scheme is strictly less than the (classical) order of the ODE method used in the time stepping, a phenomenon called *order reduction*.

Examples of derivation of bounds (5.1) for commonly used, low-order ODE methods can be seen in [33]. In [3] Brenner, Crouzeix and Thomée consider the order reduction phenomenon mainly in the case where the time stepping is directly applied to the PDE (i.e. no space discretization). They consider *implicit* one-step methods. The implicit case has been further considered in [32] by Verwer; illustrative numerical experiments are given. Explicit Runge–Kutta schemes are dealt with in [29]. Lack of space prevents us from reporting all these contributions and we here limit ourselves to a *partial* presentation of the explicit case, which is nevertheless sufficient to show the flavour of this sort of research. There is a close connection with the B-convergence theory mentioned before and references [5,6,11–13] are relevant.

5.1. The structure of the full truncation error

In the remainder of the section, we restrict our attention to the case where the ODE method used for the system (3.1) is an σ-stage pth-order *explicit* Runge–Kutta method given by the array

$$\begin{array}{c|cccc} c_1 & & & & \\ c_2 & m_{21} & & & \\ \vdots & \vdots & \ddots & & \\ c_\sigma & m_{\sigma 1} & \cdots & m_{\sigma,\sigma-1} & \\ \hline & b_1 & \cdots & b_{\sigma-1} & b_\sigma \end{array} \tag{5.2}$$

As usual we let

$$\sum_{i=1}^{\sigma} b_i = 1, \quad \sum_{j=1}^{i-1} m_{ij} = c_i, \qquad 1 \leqslant i \leqslant \sigma$$

and set $m_{\sigma+1,j} = b_j$ $(1 \leqslant j \leqslant \sigma)$, $c_{\sigma+1} = 1$.

We begin by defining the *residual* associated with the ith stage $(i = 1, \ldots, \sigma + 1)$ of the step $n \to n + 1$

$$r_i = u_h((n + c_i)\tau) - u_h(n\tau) - \tau \sum_{j=1}^{i-1} m_{ij}\big[A_h u_h((n + c_j)\tau) + f_h((n + c_j)\tau)\big]. \tag{5.3}$$

Note that by definition $r_1 = 0$ and that the residuals are defined for the PDE solution u_h rather than for the solution U_h of the ODE problem (3.1) to which the RK scheme is applied. Upon using (3.6) we can write

$$r_i = u_h((n + c_i)\tau) - u_h(n\tau) - \tau \sum_{j=1}^{i-1} m_{ij}\big[(\mathrm{d}/\mathrm{d}t)u_h((n + c_j)\tau) + \alpha_h((n + c_j)\tau)\big]$$

and, if we assume that u_h possesses $p + 1$ derivatives, we can Taylor expand u_h and $(\mathrm{d}/\mathrm{d}t)u_h$ to arrive at an expression

$$r_i = d_{i2}\tau^2 u_h^{(2)}(n\tau) + \cdots + d_{ip}\tau^p u_h^{(p)}(n\tau) + R_i. \tag{5.4}$$

Here d_{ij} are coefficients which only depend on the array (5.2) and R_i is the sum of the remainder in the Taylor expansion plus the term $\tau \Sigma m_{ij}\alpha_h((n + c_j)\tau)$ which is the contribution of the space error.

We write down the Runge–Kutta equations (cf. (5.2))

$$Y_i = U^n + \tau \sum_{j=1}^{i-1} m_{ij}\big[A_h Y_j + f_h((n + c_j)\tau)\big], \quad 1 \leqslant i \leqslant \sigma + 1, \qquad U^{n+1} = Y_{\sigma+1},$$

and subtract from them the relations (5.3). In this way we obtain a set of equations linking the global errors

$$u_h((n + 1)\tau) - U^{n+1}, \qquad u_h(n\tau) - U^n,$$

the intermediate errors $u_h((n + c_i)\tau) - Y_i$ and the residuals r_i. Elimination of the intermediate errors yields an expression for the full truncation error

$$\beta^{n+1} = \sum_{i=1}^{\sigma+1} Q_i(\tau A_h) r_i, \tag{5.5}$$

where Q_i is a polynomial of degree $\leqslant \sigma + 1 - i$, whose coefficients depend only on (5.2). Note (in connection with the B-convergence theory) that the Q_i reflect the internal stability of the RK scheme, i.e. the effect on U^{n+1} of perturbations in the computations of the internal stages of the step $n \to n + 1$. Substitution of (5.4) in (5.5) finally leads to the full truncation error expression

$$\beta^{n+1} = \sum_{l,j} \mu_{lj}\tau^{l+j} A_h^l u_h^{(j)}(n\tau) + \sum_{i=2}^{\sigma+1} Q_i(\tau A_h) R_i, \tag{5.6}$$

where μ_{lj} are scalars which only depend on (5.2) and the summation l, j extends to $1 \leqslant l \leqslant \sigma - 1$, $2 \leqslant j \leqslant p$, $p + 1 \leqslant l + j$. The important point to notice is that in (5.6) we find not only derivatives of u_h (which can be supposed to behave nicely as $h \to 0$) but also powers of A_h. These are expected to have norms which increase with decreasing h, as commented in Section 3.

5.2. Order reduction

The subsequent analysis is carried out under the following (reasonable) *hypotheses*:

(H1) The restriction $u_h(t)$ of the PDE solution u possesses $p + 1$ derivatives, which can be bounded uniformly in h and t $(0 \leqslant t \leqslant T)$.

(H2) The space-time grid refinement is carried out subject to a condition (4.6) with finite r (we are dealing with explicit methods) and, for this refinement, $\tau \| A_h \|$ can be bounded independently of τ and h. (In the heat equation model problem, (H2) clearly holds if we set in (4.6) $q = 2$, r arbitrary but finite.)

The hypothesis (H1) implies that in (5.6) the terms R_i, which originate from Taylor expanding u_h and the space discretization, satisfy a bound of the form (5.1) with an *optimal* $k = p + 1$. On the other hand (H2) implies that $\| Q_i(\tau A_h) \|$ can be bounded uniformly in τ and h, and therefore the second sum in (5.6) admits a bound (5.1) with $k = p + 1$.

Thus it remains to estimate the first sum in (5.6). We note again that each term is $O(\tau^{p+1})$, in agreement with the fact that the method has order p, but not uniformly in h.

(1) A *first* way of obtaining an h-uniform bound for a term like $\tau^{l+j} A_h^l u_h^{(j)}(n\tau)$ is to write

$$\| \tau^{l+j} A_h^l u_h^{(j)}(n\tau) \| = \tau^j \| (\tau A_h)^l u_h^{(j)}(n\tau) \| \leqslant \tau^j \| \tau A_h \|^l \| u_h^{(j)}(n\tau) \|$$

and employ (H1) and (H2). The price to be paid is that for such a term the order in τ is now j rather than the former $l + j \geqslant p + 1$. Generically (i.e. for most RK schemes) the truncation error (5.6) has $\mu_{lj} \neq 0$ for $l = p - 1$, $j = 2$ so that in this way we only obtain an $O(\tau^2)$ bound for the truncation error, *regardless of the classical order p of the method being used*. We emphasize that this order reduction, where the local error in time has only been shown to be $O(\tau^2)$, is not induced by lack of smoothness in the solution but rather by the presence of powers of A_h in the truncation error.

(2) The pessimistic conclusion we have just reached is in line with the results of the B-convergence theory, which gives prominence to the so-called stage order rather than to the classical order. Explicit methods possess a stage order equal to 1 regardless of the classical order. In actual fact the situation may not be so bad as predicted in (1), since it is often possible to estimate expressions like $A_h^l u_h^{(j)}(n\tau)$ in a *second*, more advantageous way that we present next. For simplicity we consider only the heat equation example (3.2). Let $v(x)$, $0 \leqslant x \leqslant 1$, be a smooth function and v_h its restriction to the grid. The 2nd,...,$(m-1)$th entry of $A_h v_h$ approximate values of v_{xx} and therefore can be bounded independently of h. However the first and last entry will behave like h^{-2} unless $v(0) = v(1) = 0$. Likewise, the 3rd,...,$(m-2)$th entries of $A_h^2 v_h$ approximate values of v_{xxxx} and are thus bounded. However, the 1st, 2nd, $(m-1)$th and mth entries will be bounded only if $v(0) = v(1) = v_{xx}(0) = v_{xx}(1) = 0$. The general trend should

now be clear. For $l = 1, 2, \ldots, \sigma - 1$ (the highest power of A_h which occurs in (5.6)) $\| A_h^l v_h \|$ is bounded in h if, at $x = 0, 1$,

$$\partial^{2k} v / \partial x^{2k} = 0, \quad k = 0, 1, \ldots, \sigma - 2.$$

Therefore it follows that an optimal exponent $k = p + 1$ in (5.1) can still be obtained, provided that the theoretical solution satisfies at the boundary the relations

$$(\partial^{2k}/\partial x^{2k}) u(0, t) = (\partial^{2k}/\partial x^{2k}) u(1, t) = 0, \qquad k = 0, 1, \ldots, \sigma - 2, \quad t \geqslant 0. \tag{5.7}$$

The crucial point is that these relations *do not in general* agree with the relations (2.13) that are necessary for the PDE solution u to be smooth. They do agree however if the problem is homogeneous, i.e. $f \equiv 0$, or if, by lucky coincidence, $f(0, t) = f(1, t) = 0$, $f_{xx}(0, t) = f_{xx}(1, t) = 0$, etc. We conclude that the homogeneous problem has *no order reduction* but some order reduction occurs in any other case, except for the exceptions we have just pointed out (i.e. f, $f_{xx}, \ldots$ vanish at the boundary). For an investigation of the exact amount of reduction and for a means to avoid reduction, see [29].

A final important point: the reductions we have been mentioning refer to the truncation error β; for the global error $u_h(n\tau) - U^n$ the reduction does take place, but often not so markedly as for the truncation error. The reason for this is that local errors at consecutive steps partly cancel. This cancellation can be taken into account in the analysis by a partial summation argument. It hus can be shown that in general the reduction in global order is one unit less than in local rder. Hence in the previous example of case (1) we also obtain an $O(\tau^2)$ bound for the global rror. For details, the interested reader is again directed to [29].

6. Extensions

In this paper we have been concerned with *linear, contractive* PDE problems. This class of problems is far too small to include all the problems that arise in the applications. The class of *nonlinear, contractive* PDEs has received much attention in the literature: the contractivity of the solutions can be studied by examining the dissipativity properties of the PDE itself, just as we did in (2.11) for the linear case (see e.g. Barbu [2]). Fortunately much of the material in Sections 3 and 4 can be extended without difficulty to nonlinear contractive problems. In particular this is so for the "consistency + contractivity → convergence" results in Sections 3.2 and 4.2. Also the study of the contractivity properties of RK schemes, mentioned in Section 3.1 (C4), applies totally to the nonlinear case (in fact, this study was from the beginning carried out for nonlinear problems).

However even the class of *nonlinear* contractive problems is too narrow to include all the applications. Often (2.8) must be replaced by

$$\| u(t) - v(t) \|_X \leqslant e^{\omega(t-s)} \| u(s) - v(s) \|_X, \tag{6.1}$$

with ω a positive constant, a situation also studied in the stiff ODE literature (see e.g. [10,12]) and in which much of the present material would apply with suitable modifications. Unfortunately, (6.1) is not yet the most general conceivable well-posedness requirement in PDEs and wider classes of problems can be envisaged, such as those satisfying

$$\| u(t) - v(t) \|_X \leqslant C \| u_0 - v_0 \|_X, \quad 0 \leqslant t \leqslant T,$$

for arbitrary v_0 (cf. the linear class considered in [16, Chapter 3]) or even only for v_0 in a suitable ball around u_0. Such more general classes of PDEs will be investigated in a forthcoming paper by López-Marcos and Sanz-Serna.

References

[1] J.P. Aubin, *Applied Functional Analysis* (Wiley, New York, 1979).

[2] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces* (Noordhoff, Leiden, 1976).

[3] P. Brenner, M. Crouzeix and V. Thomée, Single step methods for inhomogeneous linear differential equations in Banach space, *R.A.I.R.O. Anal. Numér. 16* (1982) 5–26.

[4] P. Brenner and V. Thomée, On rational approximations of semigroups, *SIAM J. Numer. Anal. 16* (1979) 683–694.

[5] K. Burrage and W.H. Hundsdorfer, The order of B-convergence of algebraically stable Runge-Kutta methods, *BIT 27* (1987) 62–71.

[6] K. Burrage, W.H. Hundsdorfer and J.G. Verwer, A study of B-convergence of Runge-Kutta methods, *Computing 36* (1986) 17–34.

[7] J.C. Butcher, A stability property of implicit Runge-Kutta methods, *BIT 15* (1975) 358–361.

[8] G. Dahlquist, Stability and error bounds in the numerical integration of ordinary differential equations, *Trans. Roy. Inst. Technol. Stockholm 130* (1959).

[9] G. Dahlquist, Error analysis for a class of methods for stiff nonlinear initial value problems, in: G.A. Watson, ed., *Numerical Analysis*, Lecture Notes in Mathematics 506 (Springer, Berlin, 1976)

[10] K. Dekker and J.G. Verwer, *Stability of Runge-Kutta methods for Stiff Nonlinear Differential Equations* (North-Holland, Amsterdam, 1984).

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

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

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

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

[15] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations* (Wiley, New York, 1962).

[16] E. Isaacson and J.B. Keller, *Analysis of numerical Methods* (Wiley, New York, 1966).

[17] T. Kato, *Perturbation Theory for Linear Operators* (Springer, Berlin, 2nd ed., 1984).

[18] K.W. Morton, Stability of finite difference approximations to a diffusion-convection equation, *Internat. J. Numer. Methods Engrg. 15* (1980) 677–683.

[19] J. von Neumann, Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes, *Math. Nachr. 4* (1951) 258–281.

[20] C. Palencia and J.M. Sanz-Serna, Equivalence theorems for incomplete spaces: An appraisal. *IMA J. Numer. Anal. 4* (1984) 109–115.

[21] C. Palencia and J.M. Sanz-Serna, An extension of the Lax-Richtmeyer theory, *Numer. Math. 44* (1984) 279–283.

[22] R.D. Richtmeyer and K.W. Morton, *Difference Methods for Initial Value Problems* (Interscience, New York, 1967).

[23] F. Riesz and B. Sz-Nagy, *Leçons d'Analyse Fonctionelle* (Gauthier Villars, Paris, 1968).

[24] P.H. Sammon and P. Forsyth, Jr., Instability in Runge-Kutta schemes for simulation of oil recovery, *BIT 24* (1984) 373–379.

[25] J.M. Sanz-Serna, Convergent approximation to partial differential equations and stability concepts for stiff systems of ordinary differential equations, in: *Actas de VI CEDYA*, Jaca, Universidad de Zaragoza (1984) (available on request from J.M.S.).

[26] J.M. Sanz-Serna, Stability and convergence in numerical analysis 1: Linear problems—A simple comprehensive account, in: J.K. Hale and P. Martinez-Amores, eds., *Nonlinear Differential Equations and Applications* (Pitman, Boston, MA, 1985) 64–113.

[27] J.M. Sanz-Serna and C. Palencia, A general equivalence theorem in the theory of discretization methods, *Math. Comp. 45* (1985) 143–152.
[28] J.M. Sanz-Serna and J.G. Verwer, Convergence analysis of one-step schemes in the method of lines, Rept. NM-R8608, Centre for Mathematics and Computer Science, Amsterdam (1986).
[29] J.M. Sanz-Serna, J.G. Verwer and W.H. Hundsdorfer, Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations, *Numer. Math. 50* (1987) 405–418.
[30] M.N. Spijker, Contractivity in the numerical solution of initial value problems, *Numer. Math. 42* (1983) 271–290.
[31] M.N. Spijker, Stepsize restrictions for stability of one-step methods in the numerical solution of initial value problems, *Math. Comp. 45* (1985) 377–392.
[32] J.G. Verwer, Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines, in: D.F. Griffiths and G.A. Watson, eds., *Numerical Analysis*, Pitman Research Notes in Mathematics 140 (Pitman, Boston, MA, 1986) 220–237.
[33] J.G. Verwer and J.M. Sanz-Serna, Convergence of method of lines approximations to partial differential equations, *Computing 33* (1984) 297–313.