J G VERWER

Convergence and order reduction of diagonally implicit Runge–Kutta schemes in the method of lines

1. INTRODUCTION

The method of lines (MOL) idea is simple in concept: for a given time dependent partial differential equation (PDE) discretize the space variables so that the equation is converted into a continuous time system of ordinary differential equations (ODEs). This ODE system is then numerically integrated by an integration scheme, often one which can handle *stiffness*. Various known numerical schemes for PDEs can be viewed in this way. This contribution is devoted to an analysis for the *full error* of implicit Runge-Kutta MOL schemes. We will particularly concern ourselves with a class consisting of four known *diagonally implicit methods* although much of this paper will apply to other schemes as well. However, within the class of general implicit methods there is a significant computational advantage in diagonally implicit RK (DIRK) methods, especially for PDEs. With the exception of special circumstances, other types of implicit RK methods are in fact of rather limited practical value here.

An overview of the paper reads as follows. In §2 we discuss the type of evolution problems our analysis applies to. The third paragraph is devoted to preliminaries on the discretization. Here we present the four DIRK schemes and we anticipate on the *convergence analysis* which is presented for these schemes in detail in §4. This analysis is centered around the semi-discrete approximation, i.e., the ODE system. That means that the stability concept we use is borrowed from the field of nonlinear, stiff ODEs [7]. Our error analysis is reminiscent of the analysis developed in the B-convergence theory by Frank, Schneid & Ueberhuber [9,10]. The central theme of this theory is that of *order reduction*. We examine this unwanted phenomenon in detail for a 3-rd and 4-th order DIRK scheme in the MOL framework. An interesting feature of these DIRK schemes is that the reduction for the global error is less than for the local error, although it still may be considerable when it occurs. To illustrate that the results of our analysis have real practical significance we have performed a number of numerical experiments which are presented in §5. There we also summarize some conclusions on the merits of higher order DIRK schemes in the method of lines.

2. PRELIMINARIES ON THE PROBLEM CLASS

We consider a real abstract Cauchy problem

$$u_t = \mathfrak{F}(x,t,u), \quad 0 < t \leqslant T, \quad u(x,0) = u^0(x), \tag{2.1}$$

where $\mathfrak{F}$ represents a partial differential operator which differentiates the unknown function $u(x,t)$ w.r. to its space variable x in the space domain in $\mathbb{R}, \mathbb{R}^2$ or $\mathbb{R}^3$. $\mathfrak{F}$ should not differentiate w.r. to the time variable t. The function $u(x,t)$ may be a vector function. Boundary conditions are supposed to be included in

the definition of $\mathfrak{F}$.

To the problem (2.1) we associate a real Cauchy problem for an ODE system,

$$\dot{U} = F(t,U),\ 0 < t \leq T,\ U(0) = U^0,\ F(t,\cdot):\mathbb{R}^m \to \mathbb{R}^m, \tag{2.2}$$

which is defined by a discretization of the space variable in (2.1). For the moment it is not necessary to discuss in detail how the semi discrete, continuous time approximation (2.2) arises from (2.1). Nor is it necessary, for the time being, to be specific about the partial differential equation. The reason is that our convergence analysis is centered around the ODE system (2.2). This is most convenient for the analysis and allows for the general treatment we aim at. We merely assume that U and F represent the values of grid functions on a space grid covering the space domain of (2.1). Further, we let h refer to the grid spacing, i.e., to the grid distances which may vary over the grid. In what follows, $h \to 0$ means that the grid is refined arbitrary far in a suitable manner. Note that the dimension m of problem (2.2) depends on h. The formulation (2.2) of the semi-discrete problem indicates that we concentrate on finite difference space discretizations. However, finite element or spectral methods could also be considered.

Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^m$ (we shall use the same symbol for the subordinate matrix norm) and $\mu[\cdot]$ the corresponding logarithmic matrix norm. Let $F'(t,\cdot)$ be the Jacobian matrix of $F(t,\cdot)$. Our analysis applies to problems (2.1)- (2.2) for which $\mu[F'(\cdot,\zeta)]$, $\zeta \in \mathbb{R}^m$ can be bounded from above by a constant, $\mu_{\max}$ say, which is *independent of the grid spacing*, i.e., $\mu_{\max}$ should satisfy

$$\mu_{\max} \geq \max_{\zeta \in \mathbb{R}^m} \mu[F'(\cdot,\zeta)] = \max_{\zeta \in \mathbb{R}^m} \lim_{\Delta \downarrow 0} \frac{\|I + \Delta F'(\cdot,\zeta)\| - 1}{\Delta} \tag{2.3}$$

uniformly in h. We let ζ lie in the whole of $\mathbb{R}^m$ for convenience of presentation. In actual applications it suffices to take ζ in a tube around the exact solution. For inner product norms $\|\zeta\| = (<\zeta,\zeta>)^{1/2}$ condition (2.3) can be reformulated as the one- sided Lipschitz condition (see [7], §1.5)

$$<F(\cdot,\tilde{\zeta}) - F(\cdot,\zeta), \tilde{\zeta} - \zeta> \leq \mu_{\max} \|\tilde{\zeta} - \zeta\|^2,\ \forall \tilde{\zeta}, \zeta \in \mathbb{R}^m. \tag{2.4}$$

Hypothesis (2.3), or (2.4), implies that any two solutions $\tilde{U}$, U of (2.2) satisfy the *exponential stability estimate* (a result due to Dahlquist [6])

$$\|\tilde{U}(t) - U(t)\| \leq e^{\mu_{\max} t} \|\tilde{U}(0) - U(0)\|,\ \forall t \in [o,T], \tag{2.5}$$

uniformly in h. Hence, in view of this well-posedness inequality, conditions (2.3)-(2.4) are natural. We wish to remark, however, that given a certain pair of problems (2.1)-(2.2), it may be far from trivial to select a specific norm for which (2.3) or (2.4) can proved to be valid.

Example 2.1. To illustrate the foregoing we mention two equations which were analysed in [18]. The first is the scalar, nonlinear parabolic equation

$$u_t = f(t,x,u,\ \frac{\partial}{\partial x}(d(x,t)\frac{\partial u}{\partial x})),\ t > 0,\ x \in (0,1), \tag{2.6}$$

$$u(0,t) = b_0(t),\ u(1,t) = b_1(t),\ t > 0,$$

where f and d satisfy the familiar conditions of uniform ellipticity. The second is the nonlinear Schrödinger equation

$$v_t + w_{xx} + (v^2+w^2)w = 0,\ t > 0,\ x \in (x_L,x_R), \tag{2.7}$$

$$w_t - v_{xx} - (v^2+w^2)v = 0,\ t > 0,\ x \in (x_L,x_R),$$

$$v_x(x,t) = w_x(x,t) = 0,\ x = x_L,x_R,\ t > 0.$$

Applying 3-point finite differences on nonequidistant grids ODE systems result which can be proved to satisfy (2.3), the parabolic problem in the l^∞ – norm and the Schrödinger problem in the l^2 – norm. □

In this paper we avoid questions concerning existence, uniqueness and smoothness of exact and numerical solutions. Hence, we suppose throughout that the two Cauchy problems at hand possess unique solutions $u(x,t)$ and $U(t)$, respectively. In addition, it is supposed that the true PDE solution is as smooth as the numerical analysis requires.

3. PRELIMINARIES ON THE FULL DISCRETIZATION

For the time integration of the ODE system (2.2) we define the implicit Runge-Kutta step $U^n \rightarrow U^{n+1}$ given by

$$U^{n+1} = U^n + \tau \sum_{i=1}^{s} b_i F(t_n + c_i\tau, Y_i),\ n = 0,1,..., \tag{3.1}$$

$$Y_i = U^n + \tau \sum_{j=1}^{s} a_{ij} F(t_n + c_j\tau, Y_j),\ i = 1(1)s,$$

where $t_0 = 0$ and U^{n+1} is the approximation to $U(t_{n+1}), t_{n+1} = t_n + \tau$. Throughout, we adopt the usual convention $c_i = a_{i1} + ... + a_{is}$, all i, and $b_1 + ... + b_s = 1$. Consequently, it is supposed that the *order of consistency* p of the integration formula of (3.1) is at least one.

Example 3.1. For future reference we already list the DIRK schemes we shall concentrate on later in the paper, viz., using Butcher's notation, *the implicit Euler rule*

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad p = 1 \tag{3.2}$$

the implicit midpoint rule

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad p = 2 \tag{3.3}$$

and *the 2-stage scheme*

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3},\ p = 3 \tag{3.4}$$

and *the 3-stage scheme*

$$\begin{array}{c|ccc} \gamma & \gamma & 0 & 0 \\ \frac{1}{2} & \frac{1}{2}-\gamma & \gamma & 0 \\ 1-\gamma & 2\gamma & 1-4\gamma & \gamma \\ \hline & \frac{1}{24(\frac{1}{2}-\gamma)^2} & 1-\frac{1}{12(\frac{1}{2}-\gamma)^2} & \frac{1}{24(\frac{1}{2}-\gamma)^2} \end{array} \qquad \gamma = \frac{1}{2} + \frac{1}{3}\sqrt{3}\cos(\frac{\pi}{18}),\ p = 4. \tag{3.5}$$

developed independently by Nørsett [15] and Crouzeix [5]. Observe that the order of consistency p ranges from 1 to 4. Later we will show that the 2-stage and 3-stage scheme may suffer from accuracy and order reduction. □

The RK result U^{n+1} is the *full approximation* to $u_h(t_{n+1}) = r_h u(x,t_{n+1})$. Here r_h stands for the natural restriction operator on the space grid. Hence $u_h(t)$ is a vector in $\mathbb{R}^m$. We want to study the full convergence of (3.1), i.e., the behaviour of the *full discretization error*

$$\epsilon^{n+1} = u_h(t_{n+1}) - U^{n+1} \tag{3.6}$$

as both $\tau \to 0$ and $h \to 0$. Unless otherwise stated, it is supposed that τ and h are *independent parameters.* Further, for ease of presentation we restrict ourselves to constant stepsizes τ, i.e., in the limit process we take $t_N = N\tau$ fixed and suppose that $\tau \to 0$, $N \to \infty$ in such a way that $N\tau = t_N$. As ϵ is a full error it does contain the error due to discretization of the space variables. According to the MOL approach we want to treat this part separately from the error due to discretization of the time variable. For this purpose we introduce the *space truncation error*

$$\alpha(t) = F(t,u_h(t)) - \dot{u}_h(t). \tag{3.7}$$

Here $\dot{u}_h(t) = du_h(t)/dt = r_h u_t(x,t)$, i.e., the restriction of the derivative u_t of the true PDE solution u to the space grid.

Our convergence analysis is aimed at deriving full error bounds at fixed times $t_N = N\tau$ of the form

$$\|\epsilon^N\| \leq C_1\tau^q + C_2 \max_{0\leq t\leq t_N} \|\alpha(t)\|, \quad \forall \tau \in (0,\bar{\tau}],\ 1 \leq q \leq p, \tag{3.8}$$

where C_1, C_2 and $\bar{\tau}$ are constants independent of τ and h. The term $C_1\tau^q$ emanates from the time integration. Clearly, the order q appearing in this bound must be smaller than or equal to p, the order of consistency of the RK formula. As C_1 and $\bar{\tau}$ are required to be independent of h (independent of the stiffness in the ODE terminology), it may very well happen that q is really smaller than p (order reduction). One can say that q is the *order uniform in h,* whereas p is the *order for fixed h.*

4. DERIVATION OF THE FULL ERROR BOUND

4.1 Convergence stability

Because our convergence analysis is centered around the semi-discrete problem, we can make fruitful use of stability results from the field of stiff ODEs [7]. Here the concept of C-stability [7,Ch.10] proves to be very useful for transferring the local errors (defined later on) to the full global error (in the definition below $\tilde{U}^n, \tilde{U}^{n+1}$ is a second numerical solution satisfying (3.1)).

Definition 4.1. Let $\|\cdot\|$ be a norm on $\mathbb{R}^m$. The integration method is called *C-stable* for the Cauchy problem (2.2) with respect to this norm, if a positive real number $\tau_0 = \tau_0(h)$ and a real constant C_0, independent of τ and h exist, such that for each $\tau \in (0,\tau_0]$ and each $U^n, \tilde{U}^n \in \mathbb{R}^m$

$$\|\tilde{U}^{n+1} - U^{n+1}\| \leqslant (1 + C_0\tau)\|\tilde{U}^n - U^n\|. \quad \square \tag{4.1}$$

C-stability is an abbreviation for *convergence stability* and is linked with stability in the Lax-Richtmeyer sense [16] and, more closely with stability in the sense of Kreiss [13] (sometimes referred to as strong stability [16]). If $C_0 \leqslant 0$ and we think of U^n, as being a numerical solution, and of $\tilde{U}^n$ as being a perturbation of U^n, then (3.10) shows that the perturbation will not increase in time. The bound (3.10) then provides the definition of contractivity, also called *computing stability*, a concept which plays a major role in recent developments in ODES [7]. If $C_0 > 0$, we allow an increase in the difference $\tilde{U}^n - U^n$. In this case C-stability is mainly useful in the convergence analysis and not as a concept of computing stability. Notice that C-stability is a property for *nonlinear* problems. In general τ_0 may decrease with h. However, for the given DIRK schemes we have a fixed bound $\tau_0(h)$ for τ, under the hypothesis (2.3):

Theorem 4.1. Let hypothesis (2.3) be true for a given norm $\|\cdot\|$ on $\mathbb{R}^m$. Then (i) The implicit Euler method is C-stable for this norm (ii) The implicit midpoint rule and the 2-stage and 3-stage schemes (3.4) and (3.5) are C-stable if $\|\cdot\|$ is an inner product norm. Further, for all four schemes τ_0 and C_0 depend solely on $\mu_{\max}$. $\square$

The proof of this theorem can be found in the literature on nonlinear stiff ODEs (see the survey [7], §2.4 for (i), §7.4 for (ii)). The result for implicit Euler goes back to Desoer & Haneda [8]. The C-stability of implicit midpoint has been proved by various authors and is in fact known for a long time. The result for the 2-stage and 3-stage scheme is of a more recent date and can be concluded from the general Th.7.4.2 in [7]: *if an algebraically stable RK method is BSI-stable, then it is C-stable.* At this place we should like to mention that the proof of BSI-stability of the 2-stage and 3-stage DIRK scheme, given in [7], is largely due to Montijano [13].

It shall be clear now that for the DIRK schemes applied to the problem classes (2.1)-(2.2) satisfying (2.3), stability in the sense of Definition 4.1 is guaranteed. We now leave the subject of C-stability and shall proceed with the examination of a recurrence for the full error ϵ where, in the usual way, C-stability takes care of transferring full local errors to ϵ.

4.2. A recurrence for the full error

We consider the Runge-Kutta step $U^n \to U^{n+1}$ given by (3.1) and the perturbed fictitious step $u_h(t_n) \to u_h(t_{n+1})$ given by

$$u_h(t_{n+1}) = u_h(t_n) + \tau \sum_{i=1}^{s} b_i F(t_n + c_i\tau, u_h(t_n + c_i\tau)) + r_0, \tag{4.2}$$

$$u_h(t_n + c_i\tau) = u_h(t_n) + \tau \sum_{j=1}^{s} a_{ij} F(t_n + c_j\tau, u_h(t_n + c_j\tau)) + r_i, \quad i = 1(1)s.$$

The (specific) perturbations r_i are residuals depending exclusively on the true PDE solution u_h and on the space truncation error α. For, using (3.7),

$$r_0 = u_h(t_{n+1}) - u_h(t_n) - \tau \sum_{i=1}^{s} b_i \dot{u}_h(t_n + c_i\tau) - \tau \sum_{i=1}^{s} b_i \alpha(t_n + c_i\tau), \tag{4.3}$$

$$r_i = u_h(t_n + c_i\tau) - u_h(t_n) - \tau \sum_{j=1}^{s} a_{ij} \dot{u}_h(t_n + c_j\tau) - \tau \sum_{j=1}^{s} a_{ij} \alpha(t_n + c_j\tau), \quad i = 1(1)s.$$

By straightforward Taylor expansion of u_h it follows that integers $p_i \geqslant 1$ (recall the convention made for (3.1)) and positive reals d_i, $i = 0(1)s$, exist such that uniformly in n

$$\|r_0\| \leqslant d_0 \tau^{p_0+1} + \tau \sum_{i=1}^{s} |b_i| \, \|\alpha(t_n + c_i\tau)\|, \tag{4.4}$$

$$\|r_i\| \leqslant d_i \tau^{p_i+1} + \tau \sum_{j=1}^{s} |a_{ij}| \, \|\alpha(t_n + c_j\tau)\|, \quad i = 1(1)s.$$

We note that all d_i are determined exclusively by bounds for one or more of the derivatives $\ddot{u}_h$, $\dddot{u}_h$,.... In the work of Frank, Schneid & Ueberhuber [9,10], the minimum of p_i, $\bar{p}$ say, is called the *stage order.*

Let us return to formulas (4.2) and subtract (3.1). If we define the intermediate errors $\epsilon_i = u_h(t_n + c_i\tau) - Y_i$, we then get the error scheme

$$\epsilon^{n+1} = \epsilon^n + \tau \sum_{i=1}^{s} b_i A_i \epsilon_i + r_0, \tag{4.5a}$$

$$\epsilon_i = \epsilon^n + \tau \sum_{j=1}^{s} a_{ij} A_j \epsilon_j + r_i, \quad i = 1(1)s, \tag{4.5b}$$

where, according to the mean value theorem for vector functions,

$$A_i = \int_0^1 F'(t_n + c_i\tau, \theta u_h(t_n + c_i\tau) + (1-\theta) Y_i) d\theta, \quad i = 1(1)s. \tag{4.6}$$

For convenience we suppress the dependence of A_i on n, like we did for Y_i, r_i and ϵ_i. Supposing that (4.5b) can be solved for $\epsilon_1, \ldots, \epsilon_s$ we thus arrive at the full error recurrence which is of the familiar form

$$\epsilon^{n+1} = R^{(n)}\epsilon^n + \beta^{n+1}, \tag{4.7}$$

with $R^{(n)}$ as the *amplification matrix* and β^{n+1} as the *full local error.*

The solution of the algebraic system (4.5b) is rather complicated for the general method (3.1) (see [7], Ch.5 for an extensive discussion), but fairly simple for DIRK schemes since then $a_{ij} = 0$, $j > i$, $i = 1(1)s$. We only need to assess the invertability of the matrices $I - \tau a_{ii} A_i$.

Lemma 4.1. Suppose (2.3) for a given norm $\|\cdot\|$. Then $I - \gamma\tau A_i$ is invertible for all $\tau > 0$ satisfying $\gamma\tau\mu_{\max} < 1$ while

$$\|(I-\gamma\tau A_i)^{-1}\| \leq \frac{1}{1-\gamma\tau\mu_{\max}}, \quad \|\gamma\tau A_i(I-\gamma\tau A_i)^{-1}\| \leq 1 + \frac{1}{1-\gamma\tau\mu_{\max}}. \tag{4.8}$$

Proof. The proof follows from known properties of the logarithmic norm (Dahlquist [6]). Also given in [7], Lemma 1.5.4 and Theorem 2.4.1. □

It follows that for the integration schemes and problem class under consideration the recurrence (4.7) is well defined. We may also conclude that for the DIRK schemes (3.2)-(3.5) the C-stability inequality

$$\|\epsilon^{n+1}\| \leq (1+C_0\tau)\|\epsilon^n\| + \|\beta^{n+1}\|, \quad \forall\tau \in (0,\tau_0], \tag{4.9}$$

holds due to Th.4.1 (provided the correct norm is chosen). This statement can be understood from the observation that if we subtract (3.1) from the perturbed RK step $\tilde{U}^n \to \tilde{U}^{n+1}$, where we only consider equal perturbations like in Def.4.1, that then $\tilde{U}^{n+1} - U^{n+1} = R^{(n)}(\tilde{U}^n - U^n)$ provided the definition of A is appropriately changed. Consequently, as C_0 is independent of τ and h, for finding error bounds of type (3.8) it suffices to prove that for the local error β^{n+1} a similar bound exist with the right hand side *multiplied* by τ.

By using (4.4) and (4.8) such local error bounds can be obtained in a straightforward manner for any DIRK scheme from the explicitly available expression for β^{n+1}. Rather than considering the general DIRK scheme, we shall carry out the computation for each of the four schemes (3.2)-(3.5). This enables us to discuss in greater detail the emerging order reduction phenomena. Finally we want to emphasize that the ideas behind the presented error analysis are borrowed from the B-convergence theory for stiff ODEs due to Frank, Schneid & Ueberhuber [9,10]. However, the derivation presented here is a bit shorter than in [9,10] and, in our opinion, also slightly more transparant. More details concerning this point can be found in a forthcoming paper with K. Burrage and W. Hundsdorfer [3].

4.3. The first order implicit Euler scheme (3.2)

From (4.5) we immediately can write down the error recurrence (4.7), i.e.

$$\epsilon^{n+1} = (I-\tau A_1)^{-1}\epsilon^n + \beta^{n+1}, \tag{4.10}$$

$$\beta^{n+1} = (I-\tau A_1)^{-1} r_0, \quad r_0 = u_h(t_{n+1})-u_h(t_n)-\tau\dot{u}_h(t_{n+1})-\tau\alpha(t_{n+1}).$$

Hence, according to (4.8), for u_h in C^2,

$$\|\beta^{n+1}\| \leq \frac{1}{1-\tau\mu_{\max}}(\tfrac{1}{2}M_2\tau^2+\tau\|\alpha(t_{n+1})\|), \quad \tau\mu_{\max} < 1, \tag{4.11}$$

where M_2 is an upper bound for $\|\ddot{u}_h(t)\|$. In view of the C-stability of implicit Euler, the full error bound (3.8) exists with $q = p = 1$ (no order reduction). It shows convergence of order one in time as $\tau,h \to 0$ in any way and for any norm for which (2.3) holds. An interesting feature is that only $\ddot{u}_h$ enters into the bound. We emphasize that this convergence result for implicit Euler is well known in the PDE and stiff ODE literature.

4.4. The second order implicit midpoint scheme (3.3)

The error scheme (4.5) now reads

$$\epsilon^{n+1} = \epsilon^n + \tau A_1\epsilon_1 + r_0, \ \epsilon_1 = \epsilon^n + \tfrac{1}{2}\tau A_1\epsilon_1 + r_1, \tag{4.12}$$

and the local error β^{n+1} is given by

$$\beta^{n+1} = (I-\tfrac{1}{2}\tau A_1)^{-1}\tau A_1 r_1 + r_0, \tag{4.13}$$

$$r_0 = u_h(t_n+\tau)-u_h(t_n)-\tau\dot{u}_h(t_n+\tfrac{1}{2}\tau)-\tau\alpha(t_n+\tfrac{1}{2}\tau),$$

$$r_1 = u_h(t_n+\tfrac{1}{2}\tau)-u_h(t_n)-\tfrac{1}{2}\tau\dot{u}_h(t_n+\tfrac{1}{2}\tau)-\tfrac{1}{2}\tau\alpha(t_n+\tfrac{1}{2}\tau).$$

Using the second of inequalities (4.8), and (4.4), we find for u_h in C^3,

$$\|\beta^{n+1}\| \leq cd_1\tau^2 + d_0\tau^3 + \tau(1+\tfrac{1}{2}c)\|\alpha(t_n+\tfrac{1}{2}\tau)\|, \quad \tfrac{1}{2}\tau\mu_{\max} < 1,$$

where $c = 2(1+1/(1-\frac{1}{2}\tau\mu_{\max}))$. Consequently, for inner product norms the existence of a full error bound (3.8) has been shown, but only with q equal to the stage order $\tilde{p} = 1$. This result suggests that implicit midpoint may suffer from order reduction, unless the differential equation meets an additional requirement (cf. class D2 in [10]; in our setting this condition reads

$$\|F(t,u_h+\delta\tau^2\ddot{u}_h)-F(t,u_h)\| = O(\tau^2), \quad \delta = -\tfrac{1}{8}, \tag{4.14}$$

uniformly in h). Fortunately, this suggestion is false. The situation is that the local error β^{n+1} indeed may suffer from a reduction ([9,10],[7],Ch.7), but, quite unexpectedly, the global error ϵ^{n+1} does not. This last

point has been proved by Stetter [16] and just recently by Kraaijevanger [11] (see also Axelsson [1]). Kraaijevanger's proof fits best in our setting. His idea is to treat an appropriately perturbed error scheme where the defect of the intermediate stage has been removed.

Write r_0^n, r_1^n for r_0, r_1. Let $\tilde{\epsilon}^n = \epsilon^n + r_1^n$. Then $\tilde{\epsilon}^n$ satisfies

$$\tilde{\epsilon}^{n+1} = \tilde{\epsilon}^n + \tau A_1 \epsilon_1 + \tilde{\beta}^{n+1}, \quad \epsilon_1 = \tilde{\epsilon}^n + \tfrac{1}{2}\tau A_1 \epsilon_1 \tag{4.15}$$

and $\tilde{\beta}^{n+1} = r_1^{n+1} - r_1^n + r_0^n$ can be interpreted as a (perturbed) truncation error. Because $\tilde{\epsilon}^{n+1} = R^{(n)}\tilde{\epsilon}^n + \tilde{\beta}^{n+1}$, these new local errors, say for $n = 0(1)N-1$, can be transferred to $\tilde{\epsilon}^N$ in the standard way using the C-stability property. Herewith r_1^N should be defined as the zero vector so that $\epsilon^N = \tilde{\epsilon}^N$ and $\tilde{\beta}^N = r_0^{N-1} - r_1^{N-1}$. Note that $\tilde{\epsilon}^0 = r_1^0$ if $\epsilon^0 = \underline{0}$. Neglecting α in r_0^n, r_1^n it is easily verified that $\|\tilde{\epsilon}^0\|, \|\tilde{\beta}^N\| \leq M_2\tau^2/8$ and, for $n = 0(1)N-2$, $\|\tilde{\beta}^{n+1}\| \leq M_3\tau^3/12$. Here M_2, M_3 represent bounds for $\|\ddot{u}_h\|$ and $\|\dddot{u}_h\|$, respectively. In this way the *global* error bound (3.8) is proved with $q = p = 2$ (no reduction). Noteworthy is that C_1 is determined exclusively by M_2 and M_3. For more details, a.o. concerning variable stepsizes τ and the trapezoidal rule, we refer to [11].

4.5. The third and fourth order DIRK schemes (3.4),(3.5)

In view of the experiences in the field of stiff ODEs, see e.g. [7], §7.5 for numerical experiments with (3.5), we must reckon with eventual order reduction when a DIRK scheme of higher order is used for the time integration of a PDE. We shall discuss this now for the 3-rd and 4-th order schemes (3.4) and (3.5). In our analysis we hereby concentrate on (3.4) and remark that the 4-th order scheme can be dealt with in the same manner. For (3.4) the error scheme (4.5) reads

$$\epsilon^{n+1} = \epsilon^n + \tfrac{1}{2}\tau A_1 \epsilon_1 + \tfrac{1}{2}\tau A_2 \epsilon_2 + r_0, \tag{4.16}$$

$$\epsilon_1 = \epsilon^n + \gamma\tau A_1 \epsilon_1 + r_1,$$

$$\epsilon_2 = \epsilon^n + (1-2\gamma)\tau A_1 \epsilon_1 + \gamma\tau A_2 \epsilon_2 + r_2,$$

and the local error β^{n+1} is given by

$$\beta^{n+1} = r_0 + (\tfrac{1}{2}B_1 + \tfrac{1}{2}(1-2\gamma)B_2 B_1) r_1 + \tfrac{1}{2}B_2 r_2. \tag{4.17}$$

For convenience of notation we introduced the abbreviation $B_i = (I - \gamma\tau A_i)^{-1}\tau A_i$. The residuals r_i (cf.(4.3)) satisfy, for u_h in C^4 and for any γ,

$$r_0 = (-\frac{1}{12} + \frac{1}{2}\gamma - \frac{1}{2}\gamma^2)\tau^3 \dddot{u}_h(t_n) + O(\tau^4) - \frac{1}{2}\tau\alpha(t_n + \gamma\tau) - \frac{1}{2}\tau\alpha(t_n + (1-\gamma)\tau), \tag{4.18}$$

$$r_1 = \frac{1}{2}\gamma^2\tau^2 \ddot{u}_h(t_n) + \frac{1}{3}\gamma^3\tau^3 \dddot{u}_h(t_n) + O(\tau^4) - \gamma\tau\alpha(t_n + \gamma\tau),$$

$$r_2 = (-\frac{1}{2} + 3\gamma - \frac{7}{2}\gamma^2)\tau^2 \ddot{u}_h(t_n) - (\frac{1}{6} - \gamma + \gamma^2 + \frac{1}{3}\gamma^3)\tau^3 \dddot{u}_h(t_n) +$$

$$O(\tau^4)-(1-2\gamma)\tau\alpha(t_n+\gamma\tau)-\gamma\tau\alpha(t_n+(1-\gamma)\tau).$$

If $\gamma = \frac{1}{2}\pm\frac{1}{6}\sqrt{3}$, the τ^3-term of r_0 vanishes. This value of γ corresponds to the order $p=3$. Using the stability argument (the scheme is C-stable for $\gamma \geqslant 1/4$) and the boundedness of B_i it thus follows that for the DIRK scheme (3.4) a global error bound (3.8) exists with $q \geqslant 1$, i.e., $q \geqslant \tilde{p}$, the stage order. This result is disappointing as $p=2$ for any γ and $p=3$ for $\gamma = \frac{1}{2} \pm \frac{1}{6}\sqrt{3}$. For problems satisfying the condition (4.14), the order $q=2$ is obtained. However, the constant C_1 in (3.8) then will depend also on the size of $F(t,u_h+\delta\ddot{u}_h)-F(t,u_h)$ and no longer exclusively on the smoothness of u_h.

Extensive numerical experiments has led us to the *conjecture* that ϵ *always* satisfies a bound (3.8) with $q \geqslant 2$, rather than $q \geqslant 1$, although β may show a reduction which is more in line with $q \geqslant 1$. This means that we are in a similar situation as with the implicit midpoint rule. Note, however, that in the present case ϵ does suffer from a reduction. In fact, experiments showing a virtual 2-nd order in time for the global error are easily conducted.

When attempting to prove the *conjecture* the first approach which comes to mind is that of analysing an appropriate perturbation of (4.16), like Kraaijevanger did for the midpoint rule. A little reflection shows that this is easily done if the leading terms of r_1 and r_2 are equal, which is the case only if $\gamma = 1/2, 1/4$. For other values of γ the perturbation approach seems to lead to a rather complicated analysis, but is feasible for problems of the semi-linear type $\dot{U}=AU+G(t,U)$ [3].

A case study. We shall now outline an alternative method of proof for our conjecture for the constant coefficient problem

$$\dot{U} = AU + G(t). \tag{4.19}$$

The method of proof can be extended to problems of the above semi-linear type $\dot{U} = AU + G(t,U)$ where $\|G'(t,\zeta)\| < \infty$ uniformly in h.

Consider the error scheme (4.5). Let us write r_i^n for r_i. Put $\tilde{\epsilon}^n = \epsilon^n + r_1^n$. Then (note that $R^{(n)}, B_1, B_2$ are independent of n in this case)

$$\tilde{\epsilon}^N = R^N\tilde{\epsilon}^0 + \sum_{n=0}^{N-1} R^n\tilde{\beta}^{n+1}, \quad \tilde{\beta}^{n+1} = (r_0^n - r_1^n + r_1^{n+1}) + \tfrac{1}{2}B(r_2^n - r_1^n), \tag{4.20}$$

which we write as

$$\tilde{\epsilon}^N = R^N\tilde{\epsilon}^0 + \sum_{n=0}^{N-1} \tfrac{1}{2}R^nB\hat{r}^n + \sum_{n=0}^{N-1} R^n(\tilde{\beta}^{n+1} - \tfrac{1}{2}B\hat{r}^n), \tag{4.21}$$

where $\hat{r}^n$ is the difference of the leading terms of r_1 and r_2, i.e., $\hat{r}^n = (\frac{1}{2} - 3\gamma + 4\gamma^2)\tau^2\ddot{u}_h(t_n)$. Using the stability argument on R and the boundedness of B it thus can be seen that for proving (3.8) with $q=2$ it suffices to prove that the second term, say S, satisfies $\|S\| \leqslant C\tau^2$ for all $\tau(0,\bar{\tau}]$ uniformly in h.

In what follows we now consider the most simple case where $\ddot{u}_h$ is constant, i.e., $\hat{r}^n = \hat{r}^0$ for all n. Also suppose that $I - R$ is regular (both restrictions are not essential and can be removed). Then S can be brought in the form

$$S = (I - R^N)(I - R)^{-1}\frac{1}{2}B\hat{r}^0 \tag{4.22}$$

$$= \tfrac{1}{2}(\tfrac{1}{2} - 3\gamma + 4\gamma^2)(I - R^N)(I + (\tfrac{1}{2} - 2\gamma)\tau A)^{-1}(I - \tau A)\tau^2\ddot{u}_h(0),$$

where we used the expression $R = I + B + \frac{1}{2}(1-2\gamma)B^2$. Again using the stability argument to cope with R^N and inequalities (4.8) for the rational expression in τA, finally shows that S is of second order in τ, uniformly in h, for all $\gamma > 1/4$ (for $\gamma = 1/4, 1/2$ we have $S = 0$).

For clarity, the essence of the proof is to bound the whole series S rather than its individual terms $R^n B\hat{r}^n$. The philosophy here it to attack directly the global error rather than following the standard approach of the convergence analysis which consists of first bounding locally and then adding all bounds via the stability argument. We also emphasize that no additional condition, such as (4.14), has been made and that the constant C_1 in the resulting bound (3.8) for ϵ^N is determined exclusively by $\mu_{\max}$ and bounds for $d^2u_h/dt^2, d^3u_h/dt^3$ and d^4u_h/dt^4 (only if $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$). Note that for problem (4.19), (4.14) implies that A and u_h should satisfy $A\ddot{u}_h = 0(1)$ uniformly in h. In the example below we will show, that already for simple PDE problems, leading to (4.19), (4.14) is a too severe restriction. □

Example 4.1. The objective of this example is two-fold. We want to show, for a concrete but simple problem, that the local error β indeed may suffer from more reduction than the global error, thus motivating the global approach we followed in the case study. In the second place we want to illustrate in which cases order reduction is to be expected for the method (3.4) (and (3.5)).

Let the semi-discrete system be of type (4.19) and suppose that u_h is a quadratic polynomial (this restriction is not essential and can be removed). Let $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$, so we have $p = 3$ in (3.4). The local error $\hat{\beta}^{n+1}$ given by (4.17) then takes the form $\beta^{n+1} = \hat{\beta} +$ space error, where $\hat{\beta}$ is the time error

$$\hat{\beta} = \tfrac{1}{4}\gamma^2(1-2\gamma)B^2\tau^2\ddot{u}_h(0) = \tfrac{1}{4}\gamma^2(1-2\gamma)(I - \gamma\tau A)^{-2}\tau^4A^2\ddot{u}_h(0), \tag{4.23}$$

which is independent of n. We now confine our attention to $\hat{\beta}$. Clearly, in order that $\hat{\beta} = O(\tau^{p+1}) = O(\tau^4)$, uniformly in h, it is sufficient and necessary that $(I - \gamma\tau A)^{-2}A^2\ddot{u}_h(0) = 0(1)$, uniformly in τ and h. However, this boundedness condition is rather restrictive and essentially requires that $A^2\ddot{u}_h(0) = O(1)$, uniformly in h. As A contains negative powers of h, u_{tt} then should not only be smooth enough in x, but also satisfy the boundary conditions imposed by A^2. However these b.c. are not natural (see also [2],p.7). To show this we consider the simple parabolic equation

$$u_t = u_{xx} + g(x,t), \quad t > 0, \quad 0 \leq x \leq 1, \tag{4.24}$$

with the exact solutions (imposed by adapting $g(x,t)$)

$$u(x,t) = t^2 x(1-x) \text{ and (homogeneous) Dirichlet b.c.,} \tag{4.25a}$$

$$u(x,t) = t^2(x+\tfrac{1}{2})(\tfrac{3}{2}-x) \text{ and (inhomogeneous) Dirichlet b.c..} \tag{4.25b}$$

For the discretization in space we select 2-nd order finite differences on a uniform grid. Then (4.24) is converted into (4.19) where A is the finite difference matrix

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ \ddots & \ddots & \ddots & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & \end{bmatrix}_{m\times m}, \qquad h = \frac{1}{m+1}. \tag{4.26}$$

The definition of G in (4.19) shall be clear. Note that the discretization in space is exact since u is a quadratic polynomial in x, in both cases. Hence, $\beta^{n+1} = \hat{\beta}$, $n = 0,1,....$

Now let $\tau = h \to 0$. Then the following asymptotic behaviour is observed:

$$\|\hat{\beta}\|_2 \sim \begin{matrix} \tau^{3.25} & \text{for (4.25}a\text{),} \\ \tau^{2.25} & \text{for (4.25}b\text{),} \end{matrix} \tag{4.27}$$

where $\|\cdot\|_2 = (h<.,.>)^{1/2}$, the standard l^2 norm. In the homogeneous case we have a reduction in local order from 4 to 3.25, and in the inhomogeneous case even from 4 to 2.25.

In the *homogeneous case* the reduction originates from the fact that u_{ttxx} is not zero on the boundary $x = 0,1$. To see this, u_{ttxx} is approximated by $A\ddot{u}_h$. Here, $A\ddot{u}_h(0) = 2[-2,...,-2]^T$. However, this implies that $A^2\ddot{u}_h(0) = 2[2h^{-2},0,...,0,2h^{-2}]^T$, i.e., the nearby boundary components of $A^2\ddot{u}_h$ are unbounded. Fortunately, these extremely large boundary errors are smeared out and diminished through the multiplication by $(I-\gamma\tau A)^{-2}$. In passing we note that $\hat{\beta} = O(\tau^3)$, uniformly in h, as $A\ddot{u}_h(0) = 0(1)$ (see (4.8)).

In the *inhomogeneous case* we have a similar situation, but here the reduction is larger because already u_{tt} does not vanish at $x = 0,1$ which implies that the nearby boundary components of $A\ddot{u}_h$, and $A^2\ddot{u}_h$, are unbounded in h. Notice that now condition (4.4) does not hold and that $\hat{\beta} = O(\tau^2)$, uniformly in h, as $\ddot{u}_h(0) = 0(1)$.

At first sight one might think now that we have to face a reduction in global order from , respectively, 3 to 2.25 and 3 to 1.25 as $\tau = h \to 0$. However, a direct consequence from our case study is that for both solutions (4.25) the global error is at least $O(\tau^2)$, uniformly in h. To illustrate this numerically we have integrated the problems (4.24)-(4.26) in time over the interval [0,1] using the 3-rd and 4-th order method (the latter was applied for the sake of comparison). Table 4.1 shows the quantity

$$p_2(N) = \log_2 \|\epsilon^N\|_2 / \|\epsilon^{2N}\|_2, \quad N\tau = 1, \tag{4.28}$$

i.e., the *order of accuracy* measured using $\tau = h = N^{-1}, (2N)^{-1}$. Recall that no space error is present. The floating point numbers are $\|\epsilon^{10}\|_2$.

Table 4.1. Order test for methods (3.4), (3.5) applied to problems (4.24)-(4.26). The left table corresponds to the homogeneous b.c., the right table to the inhomogeneous ones.

N	10	20	40	80	160
(3.4)	1.6_{10}^{-4}	2.56	2.72	2.83	2.90
(3.5)	8.7_{10}^{-4}	2.99	3.28	3.40	3.33

N	10	20	40	80	160
(3.4)	8.2_{10}^{-4}	2.34	2.34	2.29	2.26
(3.5)	5.0_{10}^{-4}	2.38	2.25	2.21	2.22

We see that in the case of the homogeneous b.c., p_2 tends to $p = 3$ for method (3.4) (no virtual reduction visible), whereas for method (3.5) the p_2-values indicate clearly that reduction occurs. In the case of the inhomogeneous b.c. both methods suffer from reduction. Noticeable is that it is larger for the 4-th order method (3.5) (from 4 to approximately 2.2). This experiment shows that even for simple parabolic problems with smooth solutions and inhomogeneous b.c. there may be no advantage at all in using high order in time. Finally, it is worthwhile to remark that the same results are found when we keep h fixed and consider a finite, realistic range of τ-values. □

Remark 4.1. Brenner, Crouzeix & Thomée [2] reported earlier on the phenomenon of order reduction for RK methods applied to PDEs. They restrict their analysis to constant coefficient linear problems (in Banach space) and examine only reduction of the local error. They also use problem (4.24)-(4.25) as an example. □

Remark 4.2. The case study and the example treated in this paragraph were meant to give insight into the local and global error behaviour of higher order DIRK schemes. It is noted that a proof of our conjecture that $q \geqslant 2$ in (3.8) has not yet been obtained for the general nonlinear problem (2.1)-(2.3). The method of proof followed in the case study can probably not extended to this general nonlinear problem (see also [3]). In the example we have shown the origin of the order reduction. We want to remark that the restriction to constant A is not essential. Also for A time dependent, thus covering the most general situation, an expression similar to (4.23) can be derived from (4.17). However, this expression is lengthy and complicated and renders no more insight. □

Example 4.2. The objective of this example is to call attention for another source of inaccuracy, viz., *non-smooth coefficients* in the PDE operator (non-smooth in the sense of having large gradients). It is best illustrated from a concrete, simple problem. Consider the parabolic equation

$$u_t = (d(x)u_x)_x + g(x,t), \quad t > 0, \quad 0 \leqslant x \leqslant 1, \tag{4.29}$$

with Dirichlet boundary conditions. Let u be a quadratic polynomial in t. Any (finite difference) semi-discrete approximation takes the form (4.19) and, like in Example 4.1, the time error part $\hat{\beta}$ of the local error (4.17) is given by (4.23).

Now examine the grid functions $A\ddot{u}_h$, $A^2\ddot{u}_h$. Clearly, $A\ddot{u}_h$ represents an approximation to $(du_{tx})_x$ and

$A^2\ddot{u}_h$ to $(d(du_{ttx})_{xx})_x$. Next suppose that d has much larger gradients than u so that

$$|(d(du_{ttx})_{xx})_x| >> |(du_{ttx})_x| >> |u_{tt}|,$$

which will imply that for all components $(\cdot)_j$, and for any realistic value of h,

$$|(A^2\ddot{u}_h)_j| >> |(A\ddot{u}_h)_j| >> |(\ddot{u}_h)_j|.$$

This observation suggests that the bound $\|\hat{\beta}\| \leq C\tau^2$, derived from the expression

$$\hat{\beta} = \tfrac{1}{4}\gamma^2(1-2\gamma)(\tau^2A^2(I-\gamma\tau A)^{-2})\tau^2\ddot{u}_h(0), \tag{4.23'}$$

and thus with C *independent of the non-smooth coefficient* d, will be in better accord with the true error behaviour than a higher order bound where the constant involved does depend on d.

To test this we have repeated the numerical experiment of Example 4.1 using the *non-smooth coefficient* $d(x) = (2+x)^8$ and the solutions (4.25). For the discretization in space we here used the standard 4-th order finite difference formula, except at the nearby boundary points where a 3-rd order approximation was applied. Table 4.2 shows the results in exactly the same way as Table 4.1. These results indeed reveal a distinct *2-nd order behaviour* for both solutions (4.25) (and both methods). □

Table 4.2. (same information as in table 4.1).

N	10	20	40	80	160	N	10	20	40	80	160
(3.4)	$2.1_{10}-4$	1.62	1.90	1.98	2.00	(3.4)	$1.3_{10}-3$	1.86	1.97	1.99	2.00
(3.5)	$1.7_{10}-4$	1.60	1.88	1.97	2.00	(3.5)	$1.1_{10}-3$	1.90	1.97	1.99	2.00

5. A NUMERICAL STUDY

Our DIRK schemes of order $p > 2$ do suffer from order reduction as the numerical experiments of §4 clearly illustrate. One then should question whether the extra computational work needed to reach this order p pays off. The answer to this question is not easy to give since in general there are many factors involved (type of problem, level of accuracy, stability, eventual stepsize control , iteration strategy). Despite this inherent uncertainty we have conducted numerical experiments on some more problems in an attempt to supplement our theoretical results with a conclusion which is of some value to the numerical practice. The present section is devoted to three of these problems (scalar parabolic PDEs from practice, but with smooth solutions). For the sake of comparison all four DIRK methods discussed in this paper were applied.

For the discretization in space we used a uniform grid and the standard 4-th order finite difference technique, except at the nearby boundary points where a 3-rd order formule was implemented. Further in all cases $h = \tau_B$, so h decreases with the stepsize τ. In the tables of result we have listed the full error

$\|\epsilon^N\|_2, N\tau = T$ and the quantity $p_2(N)$ given by (4.28). In each experiment we selected one basic stepsize τ_B and then used $\tau = \tau_B$ for the two 1-stage schemes and $\tau = 2\tau_B, 3\tau_B$ for the 2-stage and 3-stage scheme, respectively, thus accounting the extra work of the latter ones.

Noteworthy is that according to this way of presentation, the 2-stage (DIRK2) and 3-stage (DIRK3) method are considered to be 2 and 3 times as expensive as EULER and MIDPOINT, respectively. Thus we tacitly assume that the costs per stage are equal, for all four methods and all stepsizes. For nonlinear problems this may be somewhat in favour of the higher stage methods because these become attractive only when they are capable of yielding sufficient accuracy for relatively large stepsizes. In order to reach this accuracy it then may be necessary, in practice, to do some more Newton iterations per stage, which, to some extent, then annihilates the advantage of a greater stepsize.

Problem I. The Burger's equation

$$u_t = \nu u_{xx} - u u_x, \quad 0 < t \leq T = 1, \quad 0 < x < 1. \tag{5.1}$$

This equation has been studied by many authors (e.g. by Varah [17]). For $\nu << 1$, steep gradients may exist in the solution u. We used the "large" value $\nu = 0.1$ and defined initial and Dirichlet boundary values from the exact solution given by Whitham [19], Ch.4.

$$u(x,t) = 1 - 0.9 \frac{r_1}{r_1 + r_2 + r_3} - 0.5 \frac{r_2}{r_1 + r_2 + r_3}, \tag{5.2}$$

where $r_1 = e^{-\frac{x-.5}{20\nu} - \frac{99t}{400\nu}}$, $r_2 = e^{-\frac{x-.5}{4\nu} - \frac{3t}{16\nu}}$, $r_3 = e^{-\frac{x-3/8}{2\nu}}$.

Table 5.1. Results for Burger's equation (5.1)-(5.2).

τ_B	$EULER(\tau_B)$		$MIDPOINT(\tau_B)$		$DIRK2(2\tau_B)$		$DIRK3(3\tau_B)$	
	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2
1/24	1.2_{10}^{-3}		4.6_{10}^{-5}		5.6_{10}^{-5}		8.8_{10}^{-5}	
1/48	6.0_{10}^{-4}	1.0	1.2_{10}^{-5}	2.0	9.8_{10}^{-6}	2.52	1.4_{10}^{-5}	2.63
1/96	3.0_{10}^{-4}	1.0	2.9_{10}^{-6}	2.0	1.8_{10}^{-6}	2.43	2.8_{10}^{-6}	2.35
1/192	1.5_{10}^{-4}	1.0	7.3_{10}^{-7}	2.0	3.6_{10}^{-7}	2.34	5.9_{10}^{-7}	2.25
1/384	7.6_{10}^{-4}	1.0	1.8_{10}^{-7}	2.0	7.4_{10}^{-8}	2.28	1.3_{10}^{-7}	2.23

The results, collected in Table 5.1, reveal a distinct order reduction of the 3-rd order DIRK2 and the 4-th order DIRK3. In contrast, the order one and two of EULER and MIDPOINT clearly shows up. An interesting observation is that the p_2-values of the 3-rd and 4-th order method again are nearly equal (compare with Table 4.1, right table, and Table 4.2) and approach 2. A consequence is that these two methods do not perform better than MIDPOINT.

Problem II. Again the Burger's equation

$$u_t = \pi^{-2}\nu u_{xx} - \pi^{-1}uu_x, \quad 0 < t \leqslant T = 1, \quad 0 \leqslant x \leqslant 1, \tag{5.3}$$

but now with homogeneous boundary conditions $u(0,t) = u(1,t) = 0$ and with the initial function $u(x,0) = u_0\sin(\pi x)$. The exact solution of this problem was obtained by Cole [4] and reads

$$u(x,t) = \frac{4\nu \sum_{n=1}^{\infty} e^{-\nu n^2 t} \mathfrak{I}_n(\frac{u_0}{2\nu})\sin(n\pi x)}{\mathfrak{I}_0(\frac{u_0}{2\nu}) + 2\sum_{n=1}^{\infty} e^{-\nu n^2 t} \mathfrak{I}_n(\frac{u_0}{2\nu})\cos(n\pi x)}, \tag{5.4}$$

where $\mathfrak{I}_n(y)$ is the modified Bessel function of the first kind. In this example we chose $\nu = \pi^2/10$ and $u_0 = 1$. Results are given in Table 5.2

Table 5.2. Results for Burger's equation (5.3)-(5.4).

τ_B	$EULER(\tau_B)$		$MIDPOINT(\tau_B)$		$DIRK2(2\tau_B)$		$DIRK3(3\tau_B)$	
	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2
1/24	8.1_{10}^{-3}		6.5_{10}^{-5}		3.0_{10}^{-5}		6.4_{10}^{-5}	
1/48	4.1_{10}^{-3}	.97	1.6_{10}^{-5}	2.06	4.8_{10}^{-6}	2.67	7.8_{10}^{-6}	3.04
1/96	2.1_{10}^{-3}	.99	3.9_{10}^{-6}	2.01	6.9_{10}^{-7}	2.79	7.3_{10}^{-7}	3.42
1/192	1.0_{10}^{-3}	.99	9.7_{10}^{-7}	2.00	9.2_{10}^{-8}	2.91	5.7_{10}^{-8}	3.68
1/384	5.2_{10}^{-4}	1.00	2.4_{10}^{-7}	2.00	1.2_{10}^{-8}	2.94	4.4_{10}^{-9}	3.69

It is striking that for the solution (5.4) the observed orders p_2 of DIRK2 and DIRK3 are in much better agreement with their orders p than for the solution (5.2). This indicates that for (5.4) the contamination of their local errors with large elementary differentials is much less than for (5.2) due to the zero boundary values. We again refer to Table 4.1 for comparison. Also note that for the larger τ_B-values DIRK2 and DIRK3 are hardly more efficient than MIDPOINT.

Problem III. The nonlinear problem

$$u_t = (u^5)_{xx}, \quad 0 < t \leqslant T = 1, \quad 0 \leqslant x \leqslant 1, \tag{5.5}$$

discussed by Richtmyer & Morton [15], §8.6. They consider the running wave solution implicitly defined by $\frac{5}{4}(u-u_0)^4 + \frac{20}{3}u_0(u-u_0)^3 + 15u_0^2(u-u_0)^2 + 20u_0^3(u-u_0) + 5u_0^4\ln(u-u_0) = v(vt-x)$, where v, u_0 are constants. This is a wave running to the right if $v > 0$. Following [15], initial and (Dirichlet) boundary values were taken from (5.6) by Newton-Raphson solution. Results are given in Table 5.3 for $v = 10, u_0 = 1$.

Table 5.3. Results for the nonlinear problem (5.5).

τ_B	$EULER(\tau_B)$		$MIDPOINT(\tau_B)$		$DIRK2(2\tau_B)$		$DIRK3(3\tau_B)$	
	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2	$\|\epsilon^N\|_2$	p_2
1/12	$1.1\bar{}_{10}{}^{5}$		$3.9\bar{}_{10}{}^{3}$		$3.1\bar{}_{10}{}^{3}$		$4.4\bar{}_{10}{}^{3}$	
1/24	$5.4\bar{}_{10}{}^{6}$	1.03	$5.5\bar{}_{10}{}^{4}$	2.82	$3.6\bar{}_{10}{}^{4}$	3.09	$7.7\bar{}_{10}{}^{4}$	2.50
1/48	$2.7\bar{}_{10}{}^{6}$	1.02	$9.7\bar{}_{10}{}^{5}$	2.50	$7.8\bar{}_{10}{}^{5}$	2.22	$1.5\bar{}_{10}{}^{4}$	2.40
1/96	$1.3\bar{}_{10}{}^{6}$	1.01	$1.8\bar{}_{10}{}^{5}$	2.44	$1.9\bar{}_{10}{}^{5}$	2.01	$3.6\bar{}_{10}{}^{5}$	2.02
1/192	$6.7\bar{}_{10}{}^{7}$	1.00	$3.4\bar{}_{10}{}^{6}$	2.39	$4.8\bar{}_{10}{}^{6}$	2.02	$8.9\bar{}_{10}{}^{6}$	2.03

Also for this problem DIRK2 and DIRK3 both suffer from a distinct order reduction. However, EULER and MIDPOINT behave uncommon, too. The observed order of MIDPOINT is clearly higher than two, while, not withstanding its order one, EULER yields remarkably accurate results. The explanation lies in the fact that u is non-smooth in the sense that higher derivatives of u are much larger than the lower ones (differentiate, e.g., the solution for $u_0 = 0$). In such situations EULER may operate more accurately than higher order schemes because for EULER the error depends essentially on the size of u_{tt}. The peculiar behaviour of MIDPOINT must be due to some lucky cancellation. Finally, the appearance of $p_2 = 2.0$ for DIRK2 and DIRK3 indicates that the reduction is dominated by a phenomenon as discussed in Example 4.2.

Our numerical experiments lead us to the following conclusions: (i) The experiments support our conjecture of §4 which states that the order q of DIRK2 and DIRK3 in the error bound (3.8) is at least 2. We proved this for semi-linear problems of the type $\dot{U} = AU + G(t,U)$ [3]. (ii) For many problems order reduction will decrease seriously the performance of DIRK2 and DIRK3. In case of time dependent boundary conditions the quantity p_2 given in (4.28) will be nearly equal for these two methods and close to the conjectured lower bound 2. (iii) DIRK2 and DIRK3 shall in general not perform better than MIDPOINT, neither in the high accuracy region due to order reduction. Our experiments strongly indicate that mostly the three schemes will be competitive to each other.

Acknowledgements This paper is a sequel to [18] which was written jointly with Prof. Chus Sanz-Serna from the University of Valladolid. With great pleasure I acknowledge many stimulating discussions with him. I also wish to thank Dr. Kevin Burrage from the University of Auckland and Dr. Willem Hundsdorfer for many helpful discussions on the order reduction phenomenon. Margreet Louter-Nool has taken care of the numerical experiments. She is to be acknowledged for her patience in doing a lot of trial and error testing.

References

1 Axelsson, O., Error estimates over infinite intervals of some discretizations of evolution equations, BIT 24, (1984) 413-424.

2 Brenner, P., M. Crouzeix & V. Thomée, Single step methods for inhomogeneous linear differential

equations in Banach space, R.A.I.R.O. Analyse numérique 16, (1982) 5-26.

3 Burrage, K., W.H. Hundsdorfer & J.G. Verwer, A study of B-convergence of Runge-Kutta methods, in press.

4 Cole, J.D., On a quasilinear parabolic equation occurring in aerodynamics, Quart. Appl. Math. 3, (1951) 225-236.

5 Crouzeix, M., Sur l'approximation des équations différentielle opérationelles linéaires parr des méthodes de Runge-Kutta, These, Université Paris VI, 1975.

6 Dahlquist, G., Stability and error bounds in the numerical integration of ordinary differential equations, Trans. Royal Inst. of Technology, No 130, Stockholm, 1959.

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

8 Desoer, C. & H. Haneda, The measure of a matrix as a tool to analyze computer algorithms for circuit analyses, IEEE Trans. Circuit Theory 19, (1972) 480-486.

9 Frank, R., J. Schneid & C.W. Ueberhuber, The concept of B-convergence, SIAM J. Numer. Anal. 18, (1981) 753-780.

10 Frank, R., J. Schneid & C.W. Ueberhuber, Order results for implicit Runge-Kutta methods applied to stiff systems, Bericht Nr.53/82, Institut für Numerische Mathematik, TU Wien, 1982 (to appear in SIAM J. Numer. Anal.).

11 Kraaijevanger, H.F.B.M., B-convergence of the implicit midpoint rule and the trapezoidal rule, Report 01-1985, Inst. of Appl. Math. and Comp. Sc., University of Leiden, 1985.

12 Kreiss, H.O., Ueber die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren, BIT 2, (1962) 153-181.

13 Montijano, J.I., Estudio de los metodos SIRK para la resolucion numérica de ecuaciones differenciales de tipo stiff, Thesis, University of Zaragoza, 1983.

14 Norsett, S.P., Semi-explicit Runge-Kutta methods, Rep. Math. and Comp. No. 6/74, Dept. of Math., Univ. of Trondheim, 1974.

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

16 Stetter, H.J., Zur B-Konvergenz der impliziten Trapez-und Mittelpunktregel, Unpublished Note.

17 Varah, J.M., Stability restrictions on second order, three level finite difference schemes for parabolic equations, Technical Report 78-9, The University of British Columbia, Vancouver, 1978.

18 Verwer, J.G. & J.M. Sanz-Serna, Convergence of method of lines approximations to partial differential equations, Computing 33, (1984) 297-313.

19 Whitham, G.B., Linear and nonlinear waves, Wiley-Interscience, New York, 1974.

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