Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Accuracy and Stability of Splitting with Stabilizing Corrections

W. Hundsdorfer

Modelling, Analysis and Simulation (MAS)

MAS-R9935 December 31, 1999

Report MAS-R9935
ISSN 1386-3703

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Accuracy and Stability of Splitting with Stabilizing Corrections

Willem Hundsdorfer
CWI
P.O.Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

This paper contains a convergence analysis for the method of Stabilizing Corrections, which is an internally consistent splitting scheme for initial-boundary value problems. To obtain more accuracy and a better treatment of explicit terms several extensions are regarded and analyzed. The relevance of the theoretical results is tested for convection-diffusion-reaction equations.

1991 Mathematics Subject Classification: 65M06, 65M12, 65M20
Keywords and Phrases: Numerical analysis, initial-boundary value problems, splitting methods.
Note: Work carried out under the project MAS 1.4 – Discretization of Initial Value Problems.

1. Introduction

In this paper we consider several splitting schemes based on stabilizing corrections for initial-boundary value problems for partial differential equations (PDEs). It is assumed that the spatial derivatives are discretized with finite differences or finite volumes, resulting in very large systems of ordinary differential equations (ODEs), to which the splittings will be applied. In these stabilizing corrections splitting schemes all internal vectors are consistent with the underlying equations which greatly facilitates the implementation of correct boundary conditions.

Consider the initial value problem for a system of ODEs, arising as a semi-discrete system from an initial-boundary value problem for a multi-dimensional PDE problem,

$$u'(t) = F(t, u(t)) \tag{1.1}$$

with $0 \le t \le T$ and given initial value $u(0)$. We shall consider numerical time stepping schemes with step size τ yielding approximations u_n to the exact solution $u(t_n)$ at time levels $t_n = n\tau$ for $n = 0, 1, 2, \cdots$, starting with $u_0 = u(0)$.

It is often possible to decompose the function F into a number of simpler component functions,

$$F(t, w) = F_0(t, w) + F_1(t, w) + \cdots + F_s(t, w). \tag{1.2}$$

Splitting methods use this decomposition by treating in each stage of the calculations at most one of the components implicitly. We shall assume that the F_0 component can be treated explicitly, whereas the other F_j terms are stiff and need an implicit treatment.

We first consider the following scheme, with parameter $\theta \in [\frac{1}{2}, 1]$,

$$\left.\begin{aligned} &v_0 = u_n + \tau F(t_n, u_n), \\ &v_j = v_{j-1} + \theta\tau\Big(F_j(t_{n+1}, v_j) - F_j(t_n, u_n)\Big) \qquad (j = 1, 2, \cdots, s), \\ &u_{n+1} = v_s, \end{aligned}\right\} \tag{1.3}$$

where the v_j are internal vectors for the step from t_n to t_{n+1}. This scheme has been called the method of Stabilizing Corrections in [12, 18], since the calculation of the vectors v_j, $j = 1, 2, ..., s$, primarily serves to stabilize the forward Euler method that is used to calculate v_0. When applied with dimensional splitting, the ADI methods of Douglas and Brian with $\theta = \frac{1}{2}$, and Douglas-Rachford with $\theta = 1$, see [7, 20], are special cases of interest. For linear autonomous problems the scheme (1.3) fits in the general ADI framework of Douglas & Gunn [8]. The formulation for nonlinear problems is due to van der Houwen & Verwer [12].

Originally these methods were introduced for purely parabolic problems, $F_j \approx \partial^2/\partial x_j^2$, without an explicit term. We shall consider more general convection-diffusion-reaction equations, describing for instance transport with chemical reactions in porous media or in the atmosphere, where the term F_0 may contain discretized convection operators. The scheme (1.3) itself is not very suited for such applications, mainly because the explicit term is integrated with the forward Euler method, but (1.3) is a good starting point to derive more suitable schemes.

A big advantage of (1.3) over many other splitting methods [18, 19] is that all internal vectors v_i are consistent approximations to the exact solution, namely at time t_{n+1}. This implies that if we are in a steady state $F(u) = 0$, with F independent of t, then this steady state is also a stationary point of the scheme (1.3). Moreover, with time dependent boundary conditions (included in the time dependent part of the F_j) the formula retains its consistency.

Instead of solving the nonlinear algebraic equations in (1.3), by a Newton iteration, we can also linearize the equations (corresponding to 1 Newton iteration) to obtain the following scheme, expressed here in terms of $\delta_j = v_j - u_n$,

$$\delta_0 = \tau F(t_n, u_n), \quad \delta_j = \delta_{j-1} + \theta\tau\Big(A_j\delta_j + \tau b_j\Big), \quad u_{n+1} = u_n + \delta_s, \tag{1.4}$$

where $A_j \approx \partial F_j(t_{n+1/2}, u_n)/\partial u$ and $b_j \approx \partial F_j(t_{n+1/2}, u_n)/\partial t$. For autonomous equations this is equivalent to the Rosenbrock type scheme with factorized Jacobian matrix

$$u_{n+1} = u_n + (I - \theta\tau A_s)^{-1} \cdots (I - \theta\tau A_2)^{-1}(I - \theta\tau A_1)^{-1}\tau F(u_n).$$

For nonautonomous equations, arising for instance from a PDE problem with time dependent boundary conditions, (1.4) gives a proper way to deal with the time-dependent terms.

The contents of this paper is as follows. In Section 2 an error analysis for the stabilizing correction scheme (1.3) will be presented. The analysis will be restricted to linear problems and thus it also applies to (1.4). In order to improve accuracy and treatment of the explicit term we shall consider several extensions of (1.3) using the stabilizing corrections framework. In Section 3 a multistep extension based on the BDF2 method is considered, but stability restrictions make this extension only suited for purely parabolic equations if $s \geq 2$. In Section 4 the stabilizing correction concept will be incorporated in a two-stage Runge-Kutta

method. Along with a discussion on stability we shall consider in detail the local and global discretization errors of this new scheme. The final Section 5 contains numerical illustrations and concluding remarks.

2. Error analysis for the stabilizing correction scheme (1.3)

2.1. Preliminaries

The methods in this paper are all formulated for arbitrary nonlinear problems (1.1),(1.2). The analysis, however, will be mainly restricted to linear problems. In the following we present an error analysis of the scheme (1.3) for

$$F_j(t,w) = A_j w + g_j(t). \tag{2.1}$$

with $m \times m$ matrices A_j and $g_j(t) \in \mathbb{R}^m$. It is assumed that the problem represents a semi-discrete PDE, so the dimension depends on the mesh width in space h and some of the matrices A_j will contain negative powers of h. For nonhomogeneous boundary conditions, the terms g_j will contain the boundary values relevant to A_j, which will also lead to negative powers of h, see for example [13, 17] for a more detailed description.

In the following we consider on $\mathbb{R}^m$ the discrete L_2-norm $\|w\| = (w,w)^{1/2}$ with corresponding inner product $(v,w) = \frac{1}{m} v^T w$. The induced matrix norm is also denoted by $\|\cdot\|$. It will be assumed that we have for $0 \le j \le m$

$$(v, A_j v) \le 0 \quad \text{for all } v \in \mathbb{R}^m, \tag{2.2}$$

so that the ODE system (1.1) is stable. This assumption also implies that

$$\|(I - \theta\tau A_j)^{-1}\| \le 1 \qquad \text{for arbitrary } \theta, \tau > 0. \tag{2.3}$$

Frequently we shall need additional assumptions on the matrices A_j to ensure stability of the numerical schemes; those additional assumptions will be specified when needed. Some results could be extended to nonlinear systems as will be indicated in the text.

The error bounds derived in the next sections will not be adversely affected by the mesh width h in the spatial discretization. In particular, we shal write $\mathcal{O}(\tau^p)$ to denote vectors or matrices whose norm can be bounded by $C\tau^p$ with $C > 0$ independent of h. Further, we shall use throughout the paper the notation $Z_j = \tau A_j$, $Z = Z_0 + Z_1 + \cdots + Z_s$ and $Q_j = I - \theta Z_j$.

2.2. Error recursions

We consider along with (1.3) the perturbed scheme

$$\left.\begin{aligned} &\tilde v_0 = \tilde u_n + \tau F(t_n, \tilde u_n) + \rho_0, \\ &\tilde v_j = \tilde v_{j-1} + \theta\tau\Big(F_j(t_{n+1}, \tilde v_j) - F_j(t_n, \tilde u_n)\Big) + \rho_j \qquad (j = 1, 2, \cdots, s), \\ &\tilde u_{n+1} = \tilde v_s. \end{aligned}\right\} \tag{2.4}$$

The perturbations ρ_j may stand for round-off or errors introduced in the solution of the implicit systems, for instance. Further on we shall use them to derive an expression for the local discretization errors.

Let $e_n = \tilde{u}_n - u_n$, $\varepsilon_j = \tilde{v}_j - v_j$. Subtraction of (1.3) from (2.4) gives for the linear systems (2.1) the relations

$$\begin{aligned}
&\varepsilon_0 = e_n + Ze_n + \rho_0, \\
&\varepsilon_j = \varepsilon_{j-1} + \theta Z_j(\varepsilon_j - e_n) + \rho_j \qquad (j = 1, 2, \cdots, s), \\
&e_{n+1} = \varepsilon_s.
\end{aligned}$$

We can eliminate the internal quantities ε_j by using $\varepsilon_j - e_n = Q_j^{-1}(\varepsilon_{j-1} - e_n + \rho_j)$, leading to

$$e_{n+1} = Re_n + d_n \tag{2.5}$$

with stability matrix

$$R = I + Q_s^{-1} \cdots Q_2^{-1} Q_1^{-1} Z \tag{2.6}$$

and with d_n containing the internal perturbations,

$$d_n = Q_s^{-1} \cdots Q_1^{-1}(\rho_0 + \rho_1) + Q_s^{-1} \cdots Q_2^{-1}\rho_2 + \cdots Q_s^{-1}\rho_s. \tag{2.7}$$

So, the matrix R determines how an error already present at time t_n will be propagated to t_{n+1}, whereas d_n stands for the local error introduced during the step.

2.3. Stability

Stability in (2.5) requires that

$$\|R^n\| \le K \qquad \text{for all } n \ge 1 \tag{2.8}$$

with stability constant K independent of the mesh width h.

General results for matrices A_j satisfying (2.2) are lacking. However, if we assume that the matrices A_j are normal and commuting it suffices to consider the scalar test equation

$$u'(t) = (\lambda_0 + \lambda_1 + \cdots + \lambda_s)u(t). \tag{2.9}$$

Although normal commuting matrices hardly ever occur with practical problems, it is common in the stability analysis of numerical schemes for nonlinear PDEs to linearize, neglect boundary conditions and constant terms, freeze the coefficients and then apply a von Neumann analysis (Fourier transformation). The λ_j in the test equation then correspond with the eigenvalues of the linearized operators $A_j = \partial F_j(t,u)/\partial u$. Due to the freezing of coefficients and neglection of boundary conditions these A_j are assumed to be normal and commuting in the von Neumann analysis.

For the scalar test equation, the stability matrix (2.6) reduces to the factor $r = r(z_0, z_1, ..., z_s)$ given by

$$r = 1 + \frac{z}{p}, \tag{2.10}$$

where $z_j = \tau\lambda_j$, $z = \sum_{j=0}^{s} z_j$ and $p = \prod_{j=1}^{s}(1 - \theta z_j)$. In the following we consider closed wedges in the left half plane,

$$\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \le \alpha\}.$$

Ideally, we would like to have stability for arbitrary $z_j \in \mathcal{W}_{\pi/2}$, but if $s > 2$ this is not attainable. It was proved in [14, 16] that if $z_0 = 0$ then

$$|r| \le 1 \quad \text{for all } z_j \in \mathcal{W}_\alpha \qquad \Longleftrightarrow \qquad \alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,. \tag{2.11}$$

Moreover, it was shown in [16] that with $\theta = 1$ the above remains valid for $|1 + z_0| \le 1$ if $s \le 3$ (probably for all s but proof of that is lacking), but with $\theta = \frac{1}{2}$ and $|1 + z_0| \le 1$ we need $\alpha = 0$, that is real $z_j \le 0$, even if $s = 1$.

So it seems that the stronger stability of the underlying method with $\theta = 1$ is essential to obtain results with $z_0 \neq 0$. It should be noted, however, that considering arbitrary z_0 with $|1 + z_0| \le 1$ is a bit strict; better results are obtained with $\theta = \frac{1}{2}$ if we require $|\rho + z_0| \le \rho < 1$, for example. Moreover, in the above there is no relation assumed between z_0 and the other z_j. Often we will have that $\mathrm{Re}(z_1) << -1$ if z_0 is bounded away from 0, for instance if λ_0, λ_1 correspond with eigenvalues of convection and diffusion operators in the same spatial direction, and then the above can be much too pessimistic. If different directions are involved, for example if F_0 stands for horizontal convection and F_1 for vertical diffusion or reactions, then the corresponding eigenvalues should be considered as independent, and then the above results are relevant.

Results of a different type, valid for $s = 1$ but with noncommuting operators were obtained by Crouzeix [6], also for multi-step methods. Again, in these results it is required that $\theta > \frac{1}{2}$, so that the underlying implicit method is strongly A-stable.

In the following we shall simply assume that the general stability condition (2.8) is valid for the class of problems under consideration.

REMARK. Since the explicit term is integrated in (1.3) in a forward Euler fashion, we need in general $|1 + z_0| \le 1$ for stability. If F_0 represents a convection term then this eigenvalue condition can only be satisfied with first-order upwind spatial discretization, which is inaccurate and very diffusive. The extensions of method (1.3) that will be regarded in the following sections do allow higher order upwind convection discretizations, also with flux-limiters, see [22, 23]. ◇

2.4. Local discretization errors

We shall use the above error recursion (2.5)-(2.7) to obtain bounds on the discretization errors. The error bounds will be based on derivatives of the exact solution $u(t)$ and $\varphi_j(t) = F_j(t, u(t))$. If the PDE solution is smooth, we may assume that these derivatives are bounded uniformly in the mesh width h.

Let $\tilde{u}_n = u(t_n)$ so that $e_n = u(t_n) - u_n$ is the global discretization error. To derive an expression for the local discretization error d_n we are free to chose the $\tilde{v}_j$; it is only the global relation (2.5) that matters. Simple expressions for the residuals ρ_j are obtained by taking $\tilde{v}_j = u(t_{n+1})$ for $j = 0, 1, ..., s$. Then

$$\rho_0 = \tfrac{1}{2}\tau^2 u''(t_n) + \tfrac{1}{6}\tau^3 u'''(t_n) + \cdots,$$

$$\rho_j = -\theta\tau\Big(\varphi_j(t_{n+1}) - \varphi_j(t_n)\Big) = -\theta\tau^2\varphi_j'(t_n) - \tfrac{1}{2}\theta\tau^3\varphi_j''(t_n) - \cdots, \quad j = 1, ..., s.$$

Inserting these residuals into (2.7) yields the local discretization error

$$d_n = Q_s^{-1}\cdots Q_1^{-1}\cdot\tfrac{1}{2}\tau^2 u''(t_n) - \sum_{j=1}^{s} Q_s^{-1}\cdots Q_j^{-1}\cdot\theta\tau^2\varphi_j'(t_n) + \mathcal{O}(\tau^3). \tag{2.12}$$

Note that boundedness of the Q_j^{-1} factors, see (2.3), implies that $d_n = \mathcal{O}(\tau^2)$ uniformly in the mesh width h, and thus we always obtain at least first-order convergence of the global errors e_n independent of h.

This estimate can often be improved, but then we need to take a closer look on the error propagation. We shall elaborate this for $s \le 3$, where we have

$$\begin{aligned} d_n = Q_3^{-1}Q_2^{-1}Q_1^{-1}\Big(&\tfrac{1}{2}\tau^2\varphi_0'(t_n) + (\tfrac{1}{2}-\theta)\tau^2\varphi_1'(t_n) + [(\tfrac{1}{2}-\theta)I + \theta^2 Z_1]\tau^2\varphi_2'(t_n) + \\ &+ [(\tfrac{1}{2}-\theta)I + \theta^2(Z_1+Z_2) - \theta^3 Z_1 Z_2]\tau^2\varphi_3'(t_n)\Big) + \mathcal{O}(\tau^3). \end{aligned} \tag{2.13}$$

REMARK. The expressions for the residuals ρ_j leading to (2.12) are also valid for nonlinear systems. It can be shown that the $\mathcal{O}(\tau^2)$ bound for the local discretization errors d_n remains valid for more general equations than (2.1). However, the manipulations needed to show higher order convergence become very complicated for more general problems, even if stability is assumed a priori. Moreover, stability considerations based on the linear test equation (2.9) may not be very relevant anymore. These difficulties might be mainly theoretical; in nonlinear tests our numerical experiments did not show a decrease in accuracy or stability. ◇

2.5. Global discretization errors

Throughout this subsection it will be assumed that (2.2) is valid, so that the matrices Q_j^{-1} are bounded, and that the stability condition (2.8) is satisfied. As said before, this implies convergence with order at least one, which is not the case with many other splitting methods, see [13, 15] for instance. In this subsection some second-order convergence results will be derived. It is obvious that second order can only be achieved if $\theta = \frac{1}{2}$ and $F_0 = 0$, so this will be assumed in this subsection. Then, for $s \le 3$, expression (2.13) simplifies to

$$d_n = Q_3^{-1}Q_2^{-1}Q_1^{-1}\Big(\tfrac{1}{4}\tau^2 Z_1\varphi_2'(t_n) + \tfrac{1}{4}\tau^2(Z_1+Z_2-\tfrac{1}{2}Z_1Z_2)\varphi_3'(t_n)\Big) + \mathcal{O}(\tau^3).$$

In case $s = 2$ we can use this formula with $Z_3 = 0$, $\varphi_3 = 0$.

Consider the global error recursion (2.5) with a fixed time interval $[0, T]$. According to the general condition formulated in [13, Sect.5.1] we have second-order convergence if the local error can be decomposed as

$$d_n = (I - R)\xi_n + \eta_n \quad \text{with } \xi_n = \mathcal{O}(\tau^2),\ \eta_n = \mathcal{O}(\tau^3) \text{ and } \xi_{n+1} - \xi_n = \mathcal{O}(\tau^3). \tag{2.14}$$

We note that convergence directly follows from $e_n = d_{n-1} + Rd_{n-2} + \cdots + R^{n-1}d_0$ and that the above criterion is also necessary, see [13]. Using this framework, convergence results are now easily obtained.

Theorem 2.1. Let $\theta = \frac{1}{2}$, $F_0 = 0$. Consider scheme (1.3) with $s = 2$ and assume that $A^{-1}A_1\varphi_2^{(k)}(t) = \mathcal{O}(1)$ for $k = 1, 2$ and $t \in [0, T]$. Then $e_n = \mathcal{O}(\tau^2)$ for $t_n \in [0, T]$.

Proof. If $s = 2$ we have

$$d_n = Q_2^{-1}Q_1^{-1}\tfrac{1}{4}\tau^2 Z_1\varphi_2'(t_n) + \mathcal{O}(\tau^3) = (R-I)\tfrac{1}{4}\tau^2 Z^{-1}Z_1\varphi_2'(t_n) + \mathcal{O}(\tau^3).$$

Thus we can apply criterion (2.14) with $\xi_n = Z^{-1}Z_1\varphi_2'(t_n) = A^{-1}A_1\varphi_2'(t_n)$ and η_n containing the remaining $\mathcal{O}(\tau^3)$ terms. □

For many splittings with standard convection-diffusion problems we will have $\|A^{-1}A_1\| \leq 1$, and hence the assumption $A^{-1}A_1\psi_2 = \mathcal{O}(1)$, $\psi_2 = \varphi_2', \varphi_2''$, in this theorem is natural. Furthermore we note that if A is singular, the above can be easily generalized: what we need to prove second-order convergence is the existence of a vector $v = \mathcal{O}(1)$ such that $Av = A_1\psi_2$. In all of the following such generalizations can be made.

Theorem 2.2. Let $\theta = \frac{1}{2}$, $F_0 = 0$. Consider scheme (1.3) with $s = 3$ and assume that $A^{-1}A_1\varphi_j^{(k)}(t) = \mathcal{O}(1)$ (j=2, 3), $A^{-1}A_2\varphi_3^{(k)}(t) = \mathcal{O}(1)$ and $A^{-1}A_1A_2\varphi_3^{(k)}(t) = \mathcal{O}(\tau^{-1})$ for $k = 1, 2$ and $t \in [0, T]$. Then $e_n = \mathcal{O}(\tau^2)$ for $t_n \in [0, T]$.

Proof. Since $R = I + Q_3^{-1}Q_2^{-1}Q_1^{-1}Z$, the local discretization error can be written as

$$d_n = (R-I)Z^{-1}\Big(\tfrac{1}{4}\tau^2 Z_1\varphi_2'(t_n) + \tfrac{1}{4}\tau^2(Z_1 + Z_2 - \tfrac{1}{2}Z_1Z_2)\varphi_3'(t_n)\Big) + \mathcal{O}(\tau^3).$$

Note that $A^{-1}A_1A_2\varphi_3^{(k)} = \mathcal{O}(\tau^{-1})$ implies $Z^{-1}Z_1Z_2\varphi_3^{(k)} = \mathcal{O}(1)$. Thus we can apply (2.14) in the same way as before with ξ_n containing the $\mathcal{O}(\tau^2)$ terms. □

Compared to the situation for $s = 2$, Theorem 2.1, the essential new condition here with $s = 3$ is $A^{-1}A_1A_2\psi_3 = \mathcal{O}(\tau^{-1})$, $\psi_3 = \varphi_3', \varphi_3''$, that is,

$$Z^{-1}Z_1Z_2\psi_3 = \mathcal{O}(1).$$

This may hold also if $Z^{-1}Z_1Z_2 \neq \mathcal{O}(1)$. As an example, consider A_j to be the standard second-order difference operator for $\partial^2/\partial {x_j}^2$, $j = 1, 2, 3$ with nonhomogeneous Dirichlet conditions at the boundaries. Then the matrices A_j commute and $\|A^{-1}A_j\| \leq 1$ for all h. Further it holds that $A_1^\gamma A_2^\gamma\psi_3 = \mathcal{O}(1)$ for any $\gamma < \frac{1}{4}$ (with $\gamma = \frac{1}{4}$ we have $A_1^\gamma A_2^\gamma\psi_3 = \mathcal{O}(\log(h))$, see [17, Appendix]. So we can write

$$Z^{-1}Z_1Z_2\psi_3 = \tau^{2\gamma}Z^{-1}Z_1^{1-\gamma}Z_2^{1-\gamma}[A_1^\gamma A_2^\gamma\psi_3].$$

Taking $\tau \sim h^{1+\epsilon}$ with $\epsilon = 1 - 4\gamma > 0$, it follows that

$$\|Z^{-1}Z_1Z_2\psi_3\| \sim \tau^{2\gamma}\Big(\frac{\tau}{h^2}\Big)^{1-2\gamma} = \mathcal{O}(1).$$

Thus the conditions in Theorem 2.2 are fulfilled under a step size restriction $\tau \sim h^{1+\epsilon}$ with $\epsilon > 0$ arbitrarily small. In a similar way it can also be shown that if $\tau \sim h$ then the global errors e_n can be bounded by $\tau^2\log(\tau)$, convergence with order practically equal to 2. Further we note that if φ_3', φ_3'' satisfy homogeneous boundary conditions on the boundaries relevant to A_1 and A_2 then no condition on the ratio τ/h is necessary, since then $A_1A_2\psi_3 = \mathcal{O}(1)$.

In conclusion, Theorem 2.2 indicates that also with $s = 3$ we will often have second-order convergence, although a mild restriction on the step size might be necessary in this case.

For larger values of s a similar analysis could be performed, but verification of the accuracy conditions becomes increasingly technical. For example, if $s = 4$ we get, in addition to conditions as in Theorem 2.2, the requirement $Z^{-1}Z_1Z_2Z_3\psi_4(t) = \mathcal{O}(1)$, that is, $A^{-1}A_1A_2A_3\psi_4(t) = \mathcal{O}(\tau^{-2})$. Although this may be fulfilled in many special cases, in general an order of convergence between 1 and 2 must now be expected.

2.6. Comparison with related results

The above order 1 and order 2 convergence results compare favourably with splitting methods that lack the internal consistency property for the internal vectors v_j, see [5, 13, 15] and also Section 5. The results for scheme (1.3) with $s = 2$, $\theta = \frac{1}{2}$ are similar to those given in [17] for the Peaceman-Rachford ADI method, but that ADI scheme cannot be generalized (in a consistent fashion) to $s > 2$.

On the other hand, if $s > 2$ stability restricts the class of problems to which the stabilizing correction scheme (1.3) can be applied: the eigenvalues λ_j associated with the implicit terms should not be too close to the imaginary axis. Convection dominated parts of the PDE should therefore be included in the explicit term F_0, with the corresponding CFL condition for stability. In its present form the stabilizing correction scheme is not very suited for this, since the explicit part F_0 is just treated as with the explicit Euler scheme. In the following sections generalizations will be considered where the explicit part is also integrated in a second-order manner.

3. A 2-step stabilizing correction scheme

To improve treatment of the explicit term, we consider in this section a 2-step extension of (1.3). Although this attempt to improve on (1.3) will turn out to have limited applicability, it is instructive nevertheless. We consider the following BDF2-type method:

$$\left.\begin{aligned} v_0 &= \tfrac{4}{3}u_n - \tfrac{1}{3}u_{n-1} + \tfrac{2}{3}\tau\Big(2F(t_n,u_n) - F(t_{n-1},u_{n-1})\Big),\\ v_j &= v_{j-1} + \tfrac{2}{3}\tau\Big(F_j(t_{n+1},v_j) - 2F_j(t_n,u_n) + F_j(t_{n-1},u_{n-1})\Big) \quad (j = 1,2,\cdots,s),\\ u_{n+1} &= v_s, \end{aligned}\right\} \tag{3.1}$$

In case $s = 1$ this leads to a familiar implicit-explicit BDF2 scheme,

$$u_{n+1} = \tfrac{4}{3}u_n - \tfrac{1}{3}u_{n-1} + \tfrac{2}{3}\tau\Big(2F_0(u_n) - F_0(u_{n-1})\Big) + \tfrac{2}{3}\tau F_1(u_{n+1})$$

in autonomous form. Stability and accuracy results for this scheme can be found in [1, 6, 10]. Stability of the explicit method now allows for example higher order upwind convection discretizations to be included in F_0.

Scheme (3.1) has a very nice structure of the local errors. Considering a perturbed scheme with $\tilde{v}_j = u(t_{n+1})$, $j = 0, 1, ..., s$ and $\tilde{u}_k = u(t_k)$, gives in (3.1) residual errors

$$\rho_0 = \tfrac{4}{9}\tau^3 u'''(t_n) + \mathcal{O}(\tau^4), \qquad \rho_j = -\tfrac{2}{3}\tau^3\varphi_j''(t_n) + \mathcal{O}(\tau^5) \quad (j = 1, ..., s).$$

Consequently, the local discretization error will contain only genuine $\mathcal{O}(\tau^3)$ terms and no order reduction will take place.

Unfortunately, if $s \geq 2$ the scheme is stable only for very restricted classes of problems. Consider the scalar test equation (2.9) with $s = 2$. Then the recursion (3.1) reduces to

$$(1 - \tfrac{2}{3}z_1)(1 - \tfrac{2}{3}z_2)u_{n+1} = (\tfrac{4}{3} + \tfrac{4}{3}z_0 + \tfrac{8}{9}z_1z_2)u_n - (\tfrac{1}{3} + \tfrac{2}{3}z_0 + \tfrac{4}{9}z_1z_2)u_{n-1}.$$

Stability of this recursion was found to be satisfactory in [10] for $s = 1$, that is $z_2 = 0$. However, if we let $z_2 \to -\infty$ we are left with

$$(1 - \tfrac{2}{3}z_1)u_{n+1} = -\tfrac{4}{3}z_1u_n + \tfrac{2}{3}z_1u_{n-1}.$$

According to the Schur criterion this recursion is only stable if

$$(\mathrm{Im}(\tfrac{2}{3}z_1))^2 \leq \tfrac{1}{4} - \mathrm{Re}(\tfrac{2}{3}z_1),$$

which is a region around the negative axis enclosed by a parabola. Therefore, already for the special situation $z_2 \to -\infty$ there is no wedge $\mathcal{W}_\alpha$ with positive angle α such that we have stability for all $z_1 \in \mathcal{W}_\alpha$. We note that MATLAB experiments suggest that if z_0 lies in the stability region of the explicit method that underlies (2.13) and the other z_j are real and negative then the scheme will be stable (easy to prove if $z_0 = 0$).

So, unconditional stability for the implicit z_j with the scalar test equation will essentially only hold if these z_j, $j = 1, ..., s$ are real and negative. Although this might be sufficient for certain specific purely parabolic applications, in general it is too tight. For example, if F_j represents a stiff chemistry term then the largest eigenvalues, corresponding to radical elements, are often negative indeed, but the eigenvalues somewhat closer to the origin are complex in general.

As a further word of warning, experiments in [14] with scheme (1.3) have shown that instabilities with stabilizing corrections can occur after a long time period, due to the fact that the growth factors can be very close to 1. Therefore these instabilities might be difficult to detect. For this reason it is advisable to use such schemes only for problem classes where one can be very sure about the behaviour of the eigenvalues of the Jacobian matrices F_j' over the integration interval. Thus it seems that the range of problems to which the 2-step scheme (2.13) can be applied if $s \geq 2$ is essentially limited to multi-dimensional heat equations (with dimension splitting), a very restricted class of problems.

The above stability analysis was repeated for a class of 2-step Adams methods considered in [10], but it was found that this class does not yield methods with better stability properties than the above BDF2 type scheme if $s > 1$. Further attempts to find more stable 2-step extensions with stabilizing corrections have not been tried. Since (3.1) is based on the implicit BDF2 scheme, which itself is in a certain sense the most stable 2-step scheme possible, it seems unlikely that such extensions exist.

REMARK. The above scheme (3.1) can be viewed to fit in the framework of Douglas & Gunn [8], although in that paper only linear equations without explicit terms were considered. Different splittings with multi-step schemes applied to linear equations can be found in Warming & Beam [24]. These splittings appear to be more stable, see van der Houwen & Sommeijer [11], but they contain internal vectors v_j that are not consistent with the full problem and for which special boundary conditions have to be constructed. $\diamond$

4. A 2-stage stabilizing correction scheme

In this section we shall consider the following extension of (1.3) with two stabilizing correction stages,

$$\left.\begin{aligned}
&v_0^* = u_n + \tau F(t_n, u_n),\\
&v_j^* = v_{j-1}^* + \theta\tau\Big(F_j(t_{n+1}, v_j^*) - F_j(t_n, u_n)\Big) \qquad (j = 1, 2, \cdots, s),\\
&u_{n+1}^* = v_s^*,\\
&v_0 = u_n + \tfrac{1}{2}\tau\Big(F(t_n, u_n) + F(t_{n+1}, u_{n+1}^*)\Big),\\
&v_j = v_{j-1} + \theta\tau\Big(F_j(t_{n+1}, v_j) - F_j(t_{n+1}, u_{n+1}^*)\Big) \qquad (j = 1, 2, \cdots, s),\\
&u_{n+1} = v_s.
\end{aligned}\right\} \tag{4.1}$$

Linearization of this scheme similar as in (1.4), with increments $\delta_j^* = v_j^* - u_n$ and $\delta_j = v_j - u_{n+1}^*$, leads to a Rosenbrock type method that was successfully used in [23]. The above scheme has been constructed with this Rosenbrock method as reference. As with (1.4), linearization of scheme (4.1) gives a proper way to include time dependent boundary conditions.

In case $s = 1$ scheme (4.1) reduces to the following implicit-explicit Runge-Kutta method, given here in autonomous form,

$$\begin{aligned}
u_{n+1}^* &= u_n + \tau F_0(u_n) + \tau\Big((1-\theta)F_1(u_n) + \theta F_1(u_{n+1}^*)\Big),\\
u_{n+1} &= u_n + \tfrac{1}{2}\tau\Big(F_0(u_n) + F_0(u_{n+1}^*)\Big) + \tau\Big(\tfrac{1}{2}F_1(u_n) + (\tfrac{1}{2}-\theta)F_1(u_{n+1}^*) + \theta F_1(u_{n+1})\Big).
\end{aligned}$$

Several other methods of this type were studied in [2], but the accuracy analysis in that paper was confined to classical ODE convergence, without taking order reduction effects into account that can already occur with standard Runge-Kutta methods of either implicit or explicit type, see [3, 21].

In the following we shall derive accuracy results for (4.1) by considering systematically the errors introduced in the various stages of the scheme. It will be assumed in this section that $\theta \geq \frac{1}{4}$, so that the underlying implicit method (with $s = 1$, $F_1 = F$) is A-stable.

4.1. Error recursions

As before, we consider a perturbed version of (4.1) with perturbed internal vectors $\tilde{v}_j^* = v_j^* + \varepsilon_j^*$, $\tilde{v}_j = v_j + \varepsilon_j$ and perturbations ρ_j^*, ρ_j on the corresponding stages. For the linear systems (2.1) this leads to a recursion for the errors $e_n = \tilde{u}_n - u_n$,

$$\begin{aligned}
\varepsilon_0^* = (I+Z)e_n + \rho_0^*, \quad \varepsilon_j^* = e_n + Q_j^{-1}(\varepsilon_{j-1}^* - e_n + \rho_j^*), \quad e_{n+1}^* = \varepsilon_s^*,\\
\varepsilon_0 = e_n + \tfrac{1}{2}Z(e_n + e_{n+1}^*) + \rho_0, \quad \varepsilon_j = e_{n+1}^* + Q_j^{-1}(\varepsilon_{j-1} - e_{n+1}^* + \rho_j), \quad e_{n+1} = \varepsilon_s,
\end{aligned}$$

with $j = 1, 2, \cdots, s$. In the following we denote $P = Q_1 Q_2 ... Q_s$.

From the previous results in Section 2 we know that

$$e_{n+1}^* = (I + P^{-1}Z)e_n + d_n^*, \qquad d_n^* = P^{-1}\rho_0^* + \sum_{j=1}^{s} Q_s^{-1} ... Q_j^{-1}\rho_j^*. \tag{4.2}$$

Writing

$$\varepsilon_0 = (I+Z)e^*_{n+1} + \sigma_0, \qquad \sigma_0 = (I+\tfrac{1}{2}Z)(e_n - e^*_{n+1}) + \rho_0.$$

we get in a similar way

$$e_{n+1} = (I+P^{-1}Z)e^*_{n+1} + P^{-1}\sigma_0 + \sum_{j=1}^{s} Q_s^{-1}...Q_j^{-1}\rho_j.$$

Elimination of σ_0 and e^*_{n+1} now gives

$$e_{n+1} = Re_n + d_n \tag{4.3}$$

with stability matrix

$$R = I + 2P^{-1}Z - P^{-2}Z + \tfrac{1}{2}(P^{-1}Z)^2 \tag{4.4}$$

and local error

$$d_n = (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)d^*_n + P^{-1}\rho_0 + \sum_{j=1}^{s} Q_s^{-1}...Q_j^{-1}\rho_j. \tag{4.5}$$

4.2. Stability

Stability will be discussed here under the assumption that the matrices A_j are normal and commuting, so that it suffices to consider the scalar test equation (2.8). The stability factor $r = r(z_0, z_1, ..., z_s)$ with scheme (4.1) is given by

$$r = 1 + 2\frac{z}{p} - \frac{z}{p^2} + \tfrac{1}{2}\frac{z^2}{p^2} \tag{4.6}$$

where $z = \sum_{j=0}^{s} z_j$ and $p = \prod_{j=1}^{s}(1-\theta z_j)$. This is the scalar counterpart of (4.4).

If $s = 0$ we obtain $r = 1 + z_0 + \frac{1}{2}z_0^2$, which is a familiar stability factor with explicit two-stage Runge-Kutta methods. In contrast to the the forward Euler method, this now allows explicit treatment of convective terms discretized in space with second-order upwind or third order upwind-biased spatial discretizations, see [22]. In this subsection we discuss stability if implicit terms are included. As in Section 2, we assume that $z_1, z_2, ..., z_s$ are in a wedge $\mathcal{W}_\alpha$ with angle α.

General upper bound: To begin with, it will be shown that the angles given in (2.11) for scheme (1.3) cannot be improved with scheme (4.1) or related schemes. Consider $z_0 = 0$, $z = z_1 + z_2 + \cdots + z_s$ and $q = q(z_1, ..., z_s)$ given by

$$q = 1 + \frac{\phi(z_1, z_2, ..., z_s)}{\psi_1(z_1)\psi_2(z_2)\cdots\psi_s(z_s)}\, z,$$

where ϕ is a polynomial in each z_j and the ψ_j are all polynomials without zeros in the left half plane. This is the general form of a stability factor q of a method with a factorized denominator and with the property that $\{z = 0 \Rightarrow q = 1\}$, which is necessary to maintain stationary solutions for the trivial case of (2.8) with $\sum_{j=0}^{s} \lambda_j = 0$.

Consider equal $z_j = -t\,e^{i\beta}$, $(j = 1, ..., s)$, with $0 \le \beta \le \alpha$, and assume that

$$\frac{\phi(-t\,e^{i\beta}, ..., -t\,e^{i\beta})}{\psi_1(-t\,e^{i\beta}) \cdots \psi_s(-t\,e^{i\beta})} = C(t\,e^{i\beta})^{-k} + \mathcal{O}(t^{-k-1}), \quad t \to \infty$$

with integer k and constant $C \neq 0$. Then, for $t \to \infty$,

$$q = 1 - s\,C\,t^{1-k}\,e^{i(1-k)\beta} + \mathcal{O}(t^{-k}).$$

Hence stability for all $\beta \le \alpha$ requires that $C > 0$ and

$$|k-1|\,\alpha \le \tfrac{1}{2}\pi.$$

Now, stability for fixed $z_i < 0$ $(i \neq j)$ and $z_j \to -\infty$ implies that the degree of ϕ in z_j is less than the degree of ψ_j. Consequently $k \ge s$ and thus we get the condition

$$\alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,,$$

the same upper bound as in (2.11).

We note that the same upper bound is obtained with $z_s = -\infty$ (similar proof as above). Hence, a priori knowledge of one z_j to be real and negative does not lead to a larger angle α; for that we need at least two $z_j \le 0$, see also [14, 16].

Experimental bounds: Analytical results for the stability factor (4.6) with lower bounds for the admissible angles α are very difficult to obtain. Therefore we have determined in an experimental way the maximal α for which stability holds.

This has been done by considering $z_j = -t_j e^{\pm i\alpha}$ with t_j assuming the values $-1 + \exp(20k^3/m^3)$ for $k = 0, 1, ..., m$. For $\frac{1}{4} \le \theta \le \frac{1}{2}$ some extra points $z_1 \approx (2\theta - 1)/(2\theta^2)$ were added. For z_0 we consider 2 cases: $|1 + z_0 + \frac{1}{2}z_0^2| \le 1$ and $z_0 = 0$. In the first case $2m$ points are taken on the corresponding boundary curve. The calculations were done with $m = 50$ and $m = 100$, and the estimated accuracy of the results is below 1%. The experimental results are found in the following Figures 4.1, 4.2 where the fraction a, corresponding to an angle $\alpha = a\,\pi/2$, is plotted as function of θ.

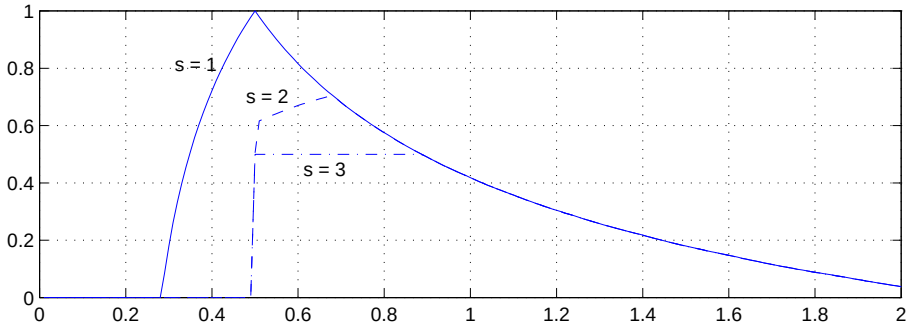

FIGURE 4.1. Fractions $a = 2\alpha/\pi$ versus θ for $s = 1, 2, 3$ and $|1 + z_0 + \frac{1}{2}z_0^2| \le 1$.

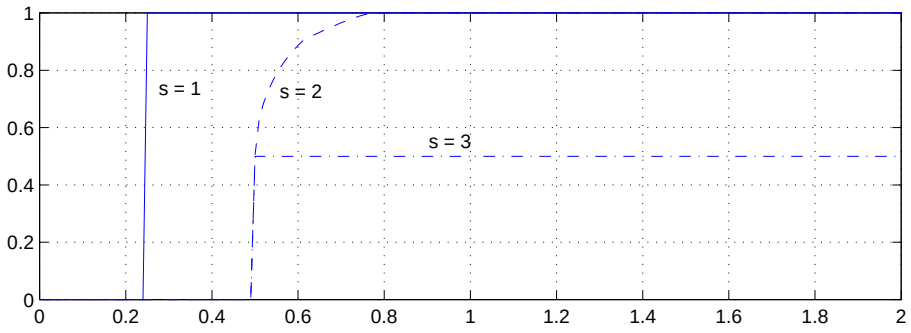

FIGURE 4.2. Fractions $a = 2\alpha/\pi$ versus θ for $s = 1, 2, 3$ and $z_0 = 0$.

Note that if $s = 1$ then A-stability of the implicit term is maintained in the presence of an explicit term with parameter value $\theta = \frac{1}{2}$. This is rather surprising since the underlying stability function of the implicit method with $\theta = \frac{1}{2}$ is "just" A-stable (no damping at ∞). We note that this fact can also be easily shown analytically by writing out the stability factor for $s = 1$,

$$r(z_0, z_1) = \frac{(1 + z_0 + \frac{1}{2}z_0^2) + (1 - 2\theta)(1 + z_0)z_1 + (\frac{1}{2} - 2\theta + \theta^2)z_1^2}{(1 - \theta z_1)^2}.$$

If $\theta = \frac{1}{2}$ and $|1 + z_0 + \frac{1}{2}z_0^2| \leq 1$, then $|r| \leq (1 + \frac{1}{4}|z_1|^2)/|1 - \frac{1}{2}z_1|^2 \leq 1$ whenever Re $z_1 \leq 0$.

Further we note that if $s \geq 2$ we get an angle $\alpha = 0$ for $\frac{1}{4} \leq \theta \leq \frac{1}{2}$, which is due to $z_1 \approx (2\theta - 1)/(2\theta^2)$ and $z_2 \to -\infty$. In Figure 4.2 we also see that if $s = 2$, $z_0 = 0$ then $\frac{1}{2}\pi$ whenever θ exceeds some threshold value; in the recent report [4] it was shown that this threshold value equals $\frac{1}{2} + \frac{1}{6}\sqrt{3}$.

From the Figures 4.1, 4.2 we conclude that if the implicit terms are splitted, $s = 2, 3$, values of θ between 0.6 and 0.8 seem optimal. If there is no explicit term and $s = 2$ it is advisable to take $\theta \geq \frac{1}{2} + \frac{1}{6}\sqrt{3}$, but with $s = 3$ all values $\theta \geq \frac{1}{2}$ give an angle $\alpha = \frac{1}{4}\pi$.

Additional tests were performed to study the admissible angles α if $z_1 \in \mathcal{W}_\alpha$ with real $z_j \leq 0$ for $j \geq 2$. There identical results were obtained with $s = 2$ and $s = 3$. In case $z_0 = 0$ the optimal angles correspond with the curve for $s = 2$ in Figure 4.2. In case $|1 + z_0 + \frac{1}{2}z_0^2| \leq 1$ we now get optimal angles corresponding to the minimum of the curves for $s = 1$ in Figure 4.1 and $s = 2$ in Figure 4.2, which is for $0.5 \leq \theta \leq 0.7$ somewhat larger than the $s = 2$ curve in Figure 4.1.

Comparison with the first-order method (1.3), $\theta = 1$, reveals that method (4.1) allows a larger stability region for z_0, but, as shown by Figure 4.1, we lose some stability in the implicit terms if $\frac{1}{2} \leq \theta \leq 0.8$ (with other values of θ we lose a lot).

In the remainder of this section it will be tacitly assumed that the general stability condition (2.8) is valid with stability matrix R given by (4.4).

4.3. Discretization errors

To obtain bounds on the discretization errors, we consider a perturbed version of (4.1) with $\tilde{v}_j^* = u(t_{n+1})$ and $\tilde{v}_j = u(t_{n+1})$ for $j = 0, 1, ..., s$. Then the corresponding perturbations in the various stages are

$$\rho_0^* = \tfrac{1}{2}\tau^2 u''(t_n) + \cdots, \qquad \rho_j^* = -\theta\tau^2\varphi_j'(t_n) + \cdots \quad (j = 1, ..., s),$$

$$\rho_0 = -\tfrac{1}{12}\tau^3 u'''(t_n) + \cdots, \qquad \rho_j = 0 \quad (j = 1, ..., s).$$

From (4.5) we thus obtain

$$d_n = (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)d_n^* + \mathcal{O}(\tau^3) \tag{4.7}$$

where d_n^* is the local discretization error of the 1-stage method (1.3), given by formula (2.12). As before, stability implies that the order of convergence is at least 1.

In the following we shall consider the condition for second-order convergence with $s \leq 3$. Denoting

$$a_n = \tfrac{1}{2}\varphi_0'(t_n) + (\tfrac{1}{2} - \theta)\Big(\varphi_1'(t_n) + \varphi_2'(t_n) + \varphi_3'(t_n)\Big),$$

$$b_n = \theta^2\Big(\varphi_2'(t_n) + \varphi_3'(t_n)\Big), \quad c_n = \theta^2\varphi_3'(t_n),$$

we have, in view of (2.13),

$$d_n^* = P^{-1}\Big(\tau^2 a_n + Z_1\tau^2 b_n + (I - \theta Z_1)Z_2\tau^2 c_n\Big) + \mathcal{O}(\tau^3),$$

Hence the local discretization error of scheme (4.1) is given by

$$d_n = P^{-1}(P - I + \tfrac{1}{2}Z)P^{-1}\Big(\tau^2 a_n + Z_1\tau^2 b_n + (I - \theta Z_1)Z_2\tau^2 c_n\Big) + \mathcal{O}(\tau^3). \tag{4.8}$$

Further

$$R - I = P^{-1}(2P - I + \tfrac{1}{2}Z)P^{-1}Z.$$

Therefore, see (2.14), the essential criterion for second-order convergence is

$$Z^{-1}P(2P - I + \tfrac{1}{2}Z)^{-1}(P - I + \tfrac{1}{2}Z)P^{-1}\Big(a_n + Z_1 b_n + (I - \theta Z_1)Z_2 c_n\Big) = \mathcal{O}(1). \tag{4.9}$$

To simplify the situation, we assume in the following that the matrices A_j commute. We consider the cases $s = 1$ and $s = 2, 3$ separately.

Case $s = 1$. Since $b_n = 0, c_n = 0$ if $s = 1$ we are left with the condition

$$(I + \tfrac{1}{2}Z_0 - (2\theta - \tfrac{1}{2})Z_1)^{-1}Z^{-1}(\tfrac{1}{2}Z_0 + (\tfrac{1}{2} - \theta)Z_1)a_n = \mathcal{O}(1).$$

Therefore second-order convergence is obtained in this case if we assume that $Z^{-1}Z_1 a_n = \mathcal{O}(1)$ and

$$(I + \tfrac{1}{2}Z_0 - (2\theta - \tfrac{1}{2})Z_1)^{-1} = \mathcal{O}(1). \tag{4.10}$$

Note that even if $Z_1 = 0$, in which case we are dealing with an explicit Runge-Kutta method, we have to impose a condition on Z_0 so that it is bounded away from $-2I$, see Sanz-Serna et al. [21]. We assume here that there is a $\gamma > 0$ such that

$$(Z_0 v, v) \geq 2(\gamma - 1)\|v\|^2 \quad \text{for all } v \in \mathbb{R}^m. \tag{4.11}$$

Then for any $\theta \geq \frac{1}{4}$ we have

$$((I + \tfrac{1}{2}Z_0 - (2\theta - \tfrac{1}{2})Z_1)v, v) \geq \gamma\|v\|^2 \quad \text{ or all } v \in \mathbb{R}^m,$$

which implies

$$\|(I + \tfrac{1}{2}Z_0 - (2\theta - \tfrac{1}{2})Z_1)^{-1}\| \leq \frac{1}{\gamma}.$$

Summarizing, we have shown that the condition (4.9) for second-order convergence with $s = 1$ is fulfilled if A_0 and A_1 commute, $A^{-1}A_1 a_n = \mathcal{O}(1)$ and (4.11) is satisfied.

Case $s = 2, 3$. For $s > 1$ no clear analytical results were found. Some computer experiments were performed for scalar problems, similar as with the stability question.

Let $z = \sum_{j=0}^{s} z_j$, $p = \prod_{j=1}^{s}(1 - \theta z_j)$, as before, and $v = p - 1 + \frac{1}{2}z$, $w = 2p - 1 + \frac{1}{2}z$. Further let

$$\psi_1 = \frac{v}{wz}, \qquad \psi_2 = \frac{vz_1}{wz}, \qquad \psi_3 = \frac{v(1 - \theta z_1)z_2}{wz}.$$

Then the condition (4.9) reads

$$\psi_1 a_n = \mathcal{O}(1), \quad \psi_2 b_n = \mathcal{O}(1), \quad \psi_3 c_n = \mathcal{O}(1).$$

Note that $c_n = 0$ if $s < 3$, so then the last condition vanishes.

Experiments with MATLAB for this scalar case did show that if $|1 + z_0 + \frac{1}{2}z_0^2| \leq 1$, $|z_0 + 2| > \delta > 0$ and the other z_j are in wedge $\mathcal{W}_\alpha$ with angle α such that we have stability according to Figure 4.1, then ψ_1 and ψ_2 are bounded. (The critical part here is to verify that $w \neq 0$ for such z_j). Hence the experiments indicate that the conditions for $s = 2$ will be satisfied.

For $s = 3$ we need an additional assumption of the type $z^{-1}z_1 z_2 c_n = \mathcal{O}(1)$, in order to deal with the situation where both $z_1, z_2 \to \infty$. This is completely similar to the situation as discussed after Theorem 2.2 for the 1-stage method with $\theta = \frac{1}{2}$ without an explicit term. Therefore, as concluded in Section 2.5, for many problems this will hold, but a mild step size restriction might be necessary.

In conclusion, although we were not able to prove second order convergence beyond the $s = 1$ case, the experimental results strongly suggest that in general scheme (4.1) will be second order accurate.

4.4. Remarks

The two-stage scheme (4.1) is more expensive per step than (1.3), whereas with $\theta = \frac{1}{2}$ the implicit terms are already integrated in a second-order fashion in (1.3). A cheaper way to improve the treatment of the explicit term would be to replace the first line in (1.3) by

$$v_0^* = u_n + \tau F(t_n, u_n), \qquad v_0 = v_0^* + \tfrac{1}{2}\tau\Big(F_0(t_{n+1}, v_0^*) - F_0(t_n, u_n)\Big), \tag{4.12}$$

so that now the explicit term is treated as with the second-order explicit Runge-Kutta method underlying (4.1). A similar effect is obtained by replacing the final line in (1.3) by

$$u_{n+1} = v_s + \tfrac{1}{2}\tau\Big(F_0(t_{n+1}, v_s) - F_0(t_n, u_n)\Big). \tag{4.13}$$

However, with respect to improving stability of the explicit term both these modifications fail. Considering the scalar test equation $u'(t) = (\lambda_0 + \lambda_1)u(t)$, with $s = 1$, the modifications yield $u_{n+1} = r(z_0, z_1)u_n$ with stability factor

$$r(z_0, z_1) = 1 + \frac{1 + \frac{1}{2}z_0}{1 - \frac{1}{2}z_1}(z_0 + z_1).$$

If $z_1 \to -\infty$ then $r(z_0, z_1) \to -(1 + z_0)$, and therefore we then still need the condition $|1 + z_0| \le 1$ on the explicit term, the same as with the forward Euler method.

The actual ratio of the CPU costs of (4.1) and (1.3) will depend on the implementation for specific problems. If the two stabilizing correction stages in (4.1) are treated as a separate step in (1.3) this ratio will be 2, but the linear algebra costs with (4.1) may be relatively lower since preconditioners or LU factorizations can be reused more often than with (1.3).

Scheme (4.1) was constructed to obtain a method with properties similar to the Rosenbrock type method that was successfully used in [23] for atmospheric transport-chemistry models. With (4.1) time-dependent boundary conditions are included in a natural way, so that second order convergence is obtained in general. It is not clear however whether (4.1) is an optimal scheme in terms of stability-accuracy properties: the stabilizing correction procedure can be applied to any diagonally implicit Runge-Kutta method and (4.1) could also be applied with different values of θ for the different functions F_j. The results in this section show that both stability and convergence should be very carefully examined.

5. Numerical tests

The second order convergence criterion for scheme (4.1) turned out to be difficult to analyze if $s \ge 2$. In this section some numerical illustrations are given for the stabilizing correction schemes applied to 2D convection-diffusion-reaction equations. We shall refer to the 1-stage scheme (1.3) as SC1 and to the 2-stage scheme (4.1) as SC2, and for both schemes parameter values $\theta = \frac{1}{2}$ and 1 are considered.

We consider here the following 2D equation, on spatial domain $\Omega = [0, 1]^2$ and $t \in [0, 1]$,

$$\mathbf{u}_t + \alpha\Big(\mathbf{u}_x + \mathbf{u}_y\Big) = \epsilon\Big(\mathbf{u}_{xx} + \mathbf{u}_{yy}\Big) + \gamma\mathbf{u}^2(1 - \mathbf{u}). \tag{5.1}$$

The solution is the traveling wave

$$\mathbf{u}(x, y, t) = \Big(1 + \exp(a(x + y - bt) + c)\Big)^{-1}, \tag{5.2}$$

with $a = \sqrt{\gamma/4\epsilon}$ determining the smoothness of the solution, $b = 2\alpha + \sqrt{\gamma\epsilon}$ the velocity of the wave and $c = a(b - 1)$ a shift parameter. Initial and Dirichlet boundary conditions are prescribed so as to correspond with this solution.

For this scalar test example splitting is not really necessary, but the structure of the equations is similar to many real-life problems where splitting cannot be avoided with present

day computer (memory) capacities. In [23] a linearized version of scheme (4.1) has been tested for a large scale application from atmospheric dispersion. In the present tests we found little difference between forms (1.3), (4.1), as used here, and the linearized (Rosenbrock) counterparts, except that instabilities are sometimes more pronounced with the latter schemes.

5.1. Reaction-diffusion test

We consider (5.1) with $\alpha = 0$. To give an illustration of the convergence behaviour of the various methods we take $\gamma = 1/\epsilon = 10$, which gives a relatively smooth solution and allows comparison with tests in [15].

The spatial derivatives are discretized with standard second order finite differences. Let $G^{(x)}(t,u) = A^{(x)}u + g^{(x)}(t)$ stand for the finite difference approximation of $\epsilon \mathbf{u}_{xx}$ with the associated time-dependent boundary conditions for $x = 0$ and $x = 1$. Likewise $G^{(y)}(t,u)$ approximates $\epsilon \mathbf{u}_{yy}$ with boundary conditions at $y = 0$, $y = 1$. Further, $H(t,u)$ represents the reaction term $\gamma \mathbf{u}^2(1-\mathbf{u})$ on the spatial grid.

We consider the following two splittings with $s = 3$ and $F_0 = 0$,

$$F_1 = G^{(x)}, \quad F_2 = G^{(y)}, \quad F_3 = H, \tag{5.3}$$

and

$$F_1 = H, \quad F_2 = G^{(x)}, \quad F_3 = G^{(y)}. \tag{5.4}$$

Since the reaction term in (5.1) with $\gamma = 10$ is not stiff, we also consider here the case where this term is taken explicitly,

$$F_0 = H, \quad F_1 = G^{(x)}, \quad F_2 = G^{(y)}. \tag{5.5}$$

The spatial grid is uniform with mesh width h in both directions. The errors in the L_2-norm are calculated at output time $T = 1$ with $\tau = h = 1/N$, $N = 10, 20, 40, 80$. In the Figure 5.1 these errors are plotted versus τ on a logarithmic scale. The results for the SC1 scheme (1.3) are indicated by dashed lines with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$. Likewise, the results for the SC2 scheme (4.1) are indicated by solid lines with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$.

For comparison, results of the well-known fractional step method of Yanenko [25] are included, indicated by dotted lines with stars. With this method fractional steps are taken with the implicit trapezoidal rule $v_j = v_{j-1} + \frac{1}{2}\tau F_j(t_n, v_{j-1}) + \frac{1}{2}\tau F_j(t_{n+1}, v_j)$, with $v_0 = u_n$. After each step the order of the F_j is interchanged to achieve symmetry and second order (in the classical ODE sense). If an explicit term F_0 is present, the implicit trapezoidal rule is replaced by its explicit counterpart (equal to method (4.1) with $s = 0$) for the fractional step with F_0.

It is known that Yanenko's method needs boundary corrections to obtain second-order convergence for initial-boundary value problems, otherwise order of convergence can be lower, see [5, 13]. In the present test we get convergence with order $\frac{1}{2}$ approximately. The same test with boundary corrections was performed in [15] for the splittings (5.3) and (5.4), but still the results have less accuracy than with the second-order stabilizing correction schemes. It should be noted that the errors for Yanenko's method given here are slightly different from those in [15, Table 6.1, 6.3], due to the fact that we used here only time levels t_n and t_{n+1} in the fractional steps.

Finally we note that boundary corrections were also attempted on the scheme (1.3), similar to formula (101) in Mitchell & Griffiths [19]. In the above test this did lead to smaller errors, reduction with a factor ranging between 1.2 and 2, but the convergence behaviour did not change fundamentally. Since boundary corrections have to be derived for each individual problem, it is a favourable property of the stabilizing correction schemes that such corrections are not necessary to get a genuine second-order behaviour.

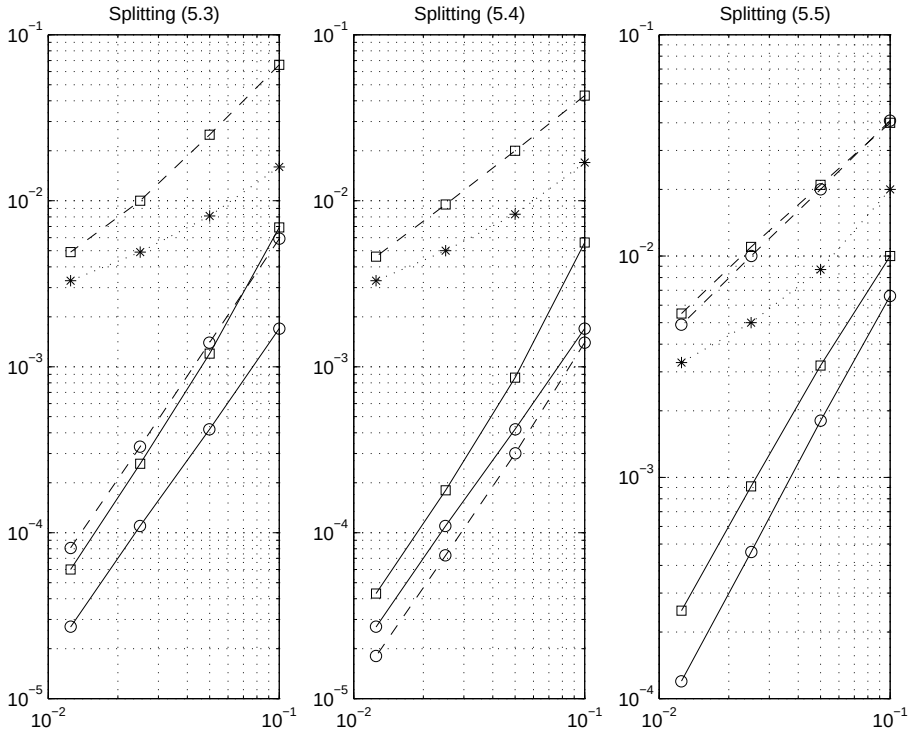

FIGURE 5.1. L_2-errors versus $\tau = h$ for the splittings (5.3), (5.4) and (5.5). Methods SC1 (dashed lines) and SC2 (solid lines) with $\theta = \frac{1}{2}$ (circles) and $\theta = 1$ (squares). Results for Yanenko's method are indicated with stars.

5.2. Convection-diffusion-reaction test

To illustrate the improved stability behaviour of the 2-stage scheme SC2 over SC1 if a substantial explicit term is present, we now consider equation (5.1) with a convection term with $\alpha = -1$ that will be taken explicitly. Further we choose $\gamma = 100$, as before, and $\epsilon = 0.01, 0.001$ which gives solutions that have a steep gradient, relative to the mesh widths used here.

The splitting is such that F_0 contains the convective terms, F_1, F_2 diffusion in x and y direction, respectively, and F_3 the nonlinear reaction term. The convective terms are discretized with third-order upwind-biased differences (4-point stencil). For the diffusion terms standard second-order differences are used as before.

The results with $\epsilon = 0.01$ are given in the Figures 5.2, 5.3. In the plots of Figure 5.2 the solutions $h = 1/40$ and $\tau = 1/80$ are found, represented as contour lines at the levels 0.1, 0.2,...,0.9, with solid lines for the numerical solution and dotted lines for the exact solution.

Quantitative results are given in Figure 5.3, where the L_2-errors are plotted as function of the time step for a 40×40 and 80×80 grid with $\tau = h, \frac{1}{2}h$ and so on. As in Figure 5.1 results for SC1 are indicated with dashed lines, for SC2 with solid lines, and with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$.

It is obvious that the 2-stage schemes SC2 give much better results than the corresponding 1-stage schemes SC1. To achieve a level of accuracy comparable to the SC2 schemes we need much smaller time steps with the SC1 schemes, see Figure 5.3. This is primarily due to the more stable treatment of the explicit convection term with the SC2 schemes. The explicit 2-stage Runge-Kutta method underlying (4.1) is stable for third-order convection discretization up to Courant number 0.87 (experimental bound). On the other hand, some of the eigenvalues associated with this discretization are always outside the stability region of the explicit Euler scheme. In this test it is the (implicit) diffusion part that provides a stabilization for the smaller step sizes. (In fact, for $\epsilon = 0.01$ similar results were obtained with second order central convection discretization, but not anymore with $\epsilon = 0.001$). Further we note that instabilities do not lead to overflow since the solutions are pushed back to the range [0,1] by the reaction term, but the resulting numerical solutions are qualitatively wrong.

Decreasing the value of the diffusion coefficient ϵ gives a clearer distinction between the methods. Results with $\epsilon = 0.001$ are given in the Figures 5.4 and 5.5. The grids chosen are 80×80 and 160×160, since the 40×40 grid gives quite large spatial errors with this small ϵ. The results are essentially the same as above: the 1-stage schemes SC1 need much smaller time steps than the SC2 schemes to obtain reasonable solutions.

For more realistic problems, with stiff reaction terms, nonlinear convection discretizations with flux limiters are recommended to avoid oscillations, and this fits easily into the present framework, see [23] for instance.

5.3. Concluding remarks

In this paper it has been shown that the stabilizing correction scheme (1.3) is in general for $s \leq 3$ convergent with order 2 if $\theta = \frac{1}{2}$ and $F_0 = 0$, and with order 1 otherwise. This the same as the classical ODE order which is only valid for nonstiff equations. The fact that the same orders are obtained for initial-boundary value problems for PDEs is due to the consistent treatment of all internal vectors v_j in (1.3).

Multi-step extensions of (1.3) seem not sufficiently stable in case $s \geq 2$. On the other hand for problems with $F = F_0 + F_1$, that is $s = 1$, such multi-step methods do perform well, see for instance [22] for tests with the BDF2 type scheme.

The 2-stage stabilizing correction scheme (4.1) is capable to deal with splittings of implicit terms, $s \geq 2$. Numerical tests on problems more complicated than considered here are found in [23] with the linearized (Rosenbrock) version. In that paper a parameter value $\theta = 1 + \frac{1}{2}\sqrt{2}$ was used, in order to obtain fast damping of the implicit terms (L-stability). In the present tests the exact solution was known and it was found that smaller values of θ give much better accuracy due to smaller error constants. Therefore, a recommended parameter range is $\frac{1}{2} \leq \theta \leq 1$, where $\theta = \frac{1}{2}$ seems most accurate but $\theta = 1$ gives more damping of the implicit terms.

The stability results in this paper show that the large eigenvalues of the implicit terms should have a dominant negative real part. For this reason the stabilizing correction schemes are not suited as black-box solvers. However, with many practical problems splittings of

implicit terms that meet the above requirement are possible (diffusion dominated terms or stiff reactions), and for such problems stabilizing correction schemes provide consistent splitting methods that do not need elaborate boundary corrections to achieve reasonable accuracies.

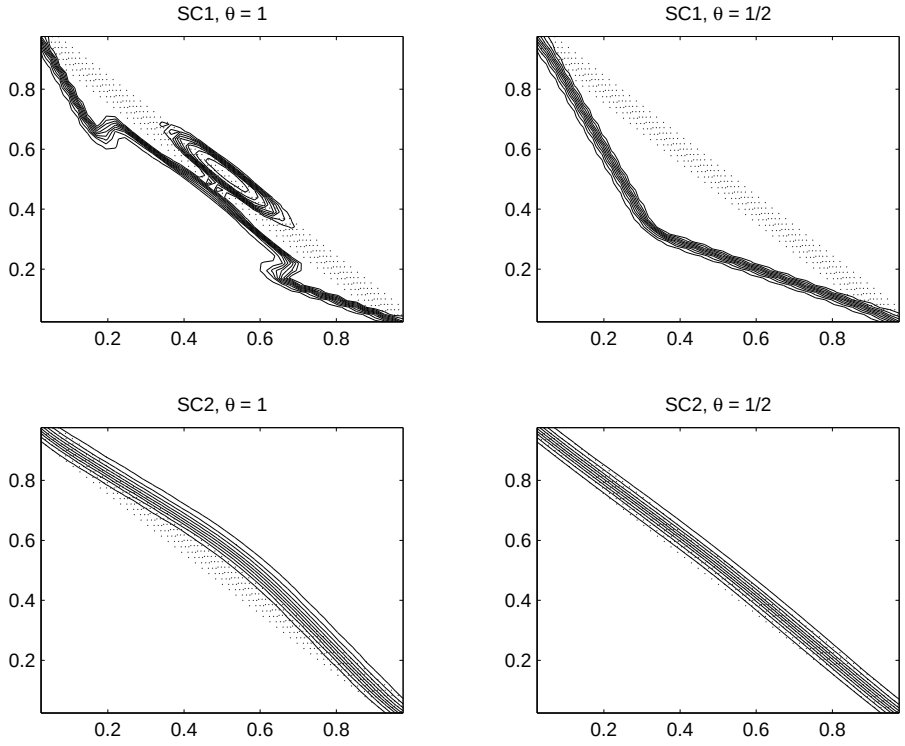

FIGURE 5.2 Contour lines for $\epsilon = 0.01$ with $h = 1/40$, $\tau = 1/80$.

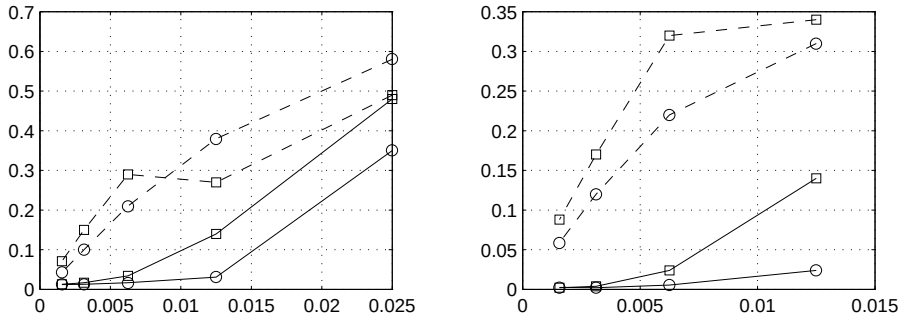

FIGURE 5.3. L_2-errors versus time step τ on 40×40 grid (left) and 80×80 grid (right) for $\epsilon = 0.01$. Various methods indicated as in Figure 5.1.

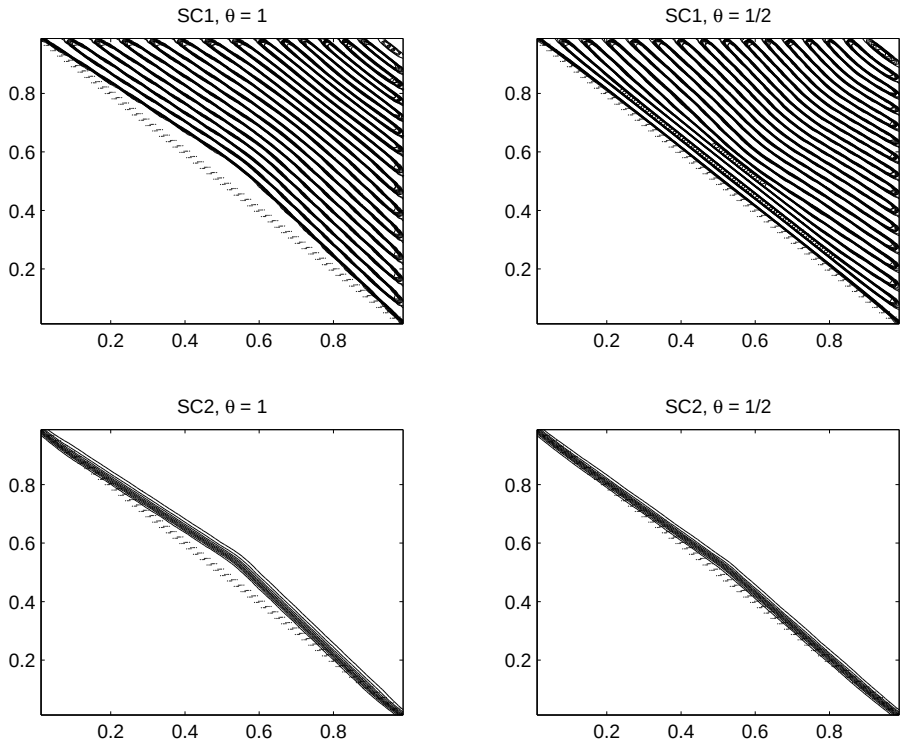

FIGURE 5.4 Contour lines for $\epsilon = 0.001$ with $h = 1/80$, $\tau = 1/160$.

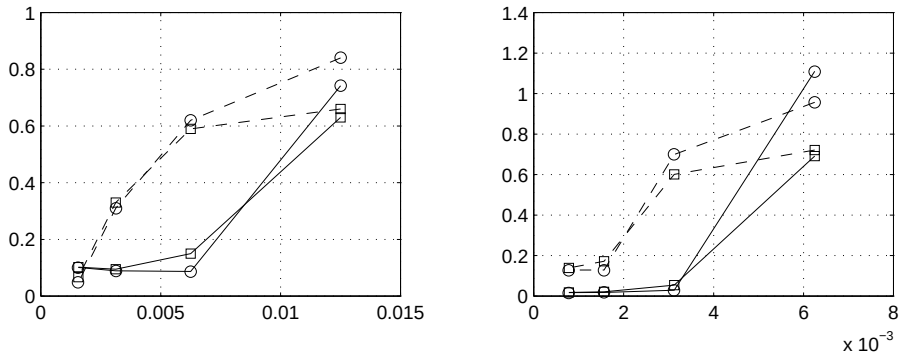

FIGURE 5.5. L_2-errors versus time step τ on 80×80 grid (left) and 160×160 grid (right) for $\epsilon = 0.001$. Various methods indicated as in Figure 5.1.

References

[1] U.M. Ascher, S.J. Ruuth and B. Wetton, *Implicit-explicit methods for time-dependent PDEs*, SIAM J. Numer. Anal. 32, pp. 797–823 (1995).

[2] U.M. Ascher, S.J. Ruuth and R.J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math. 25, pp. 151–167 (1997).

[3] P. Brenner, M. Crouzeix and V. Thomée, *Single step methods for inhomogeneous linear differential equations*. RAIRO Numer. Anal. 16, pp. 5-26 (1982).

[4] J.G. Blom, D. Lanser and J.G. Verwer, *Time integration of the shallow water equations in spherical geometry*. CWI Report MAS-00xx (in preparation).

[5] R. Čiegis, K. Kiškis, *On the stability of LOD difference methods with respect to boundary conditions*. Lithuanian Academy of Sciences, Informatica 5, pp. 297-323 (1994).

[6] M. Crouzeix , *Une méthode multipas implicite-explicite pour l' approximation des équations d' évolution paraboliques*. Numer. Math. 35 , pp. 257-276 (1980).

[7] J. Douglas, *Alternating direction method for three space variables*. Numer. Math. 4, pp. 41-63 (1962).

[8] J. Douglas, J.E. Gunn, *A general formulation of alternating direction methods*. Numer. Math. 6, pp. 428-453 (1964).

[9] C. Eichler-Liebenow, P.J. van der Houwen & B.P. Sommeijer, *Analysis of approximate factorization in iteration methods*. Appl. Num. Math. 28, pp. 245-258 (1998).

[10] J. Frank, W. Hundsdorfer and J.G. Verwer , *On the stability of implicit-explicit linear multistep methods*. Appl. Num. Math. 25 , pp. 193-205 (1997).

[11] P.J. van der Houwen, B.P. Sommeijer, *Approximate factorization for time-dependent partial differential equations*. CWI report MAS-R9915, 1999.

[12] P.J. van der Houwen, J.G. Verwer, *One-step splitting methods for semi-discrete parabolic equations*. Computing 22, pp. 291-309 (1979).

[13] W. Hundsdorfer, *Unconditional convergence of some Crank-Nicolson LOD methods for initial-boundary value problems*. Math. Comp. 58, pp. 35-53 (1992).

[14] W. Hundsdorfer, *A note on stability of the Douglas splitting method*. Math. Comp. 67, pp 183-190 (1998).

[15] W. Hundsdorfer, *Trapezoidal and midpoint splittings for initial-boundary value problems*. Math. Comp. 67, pp. 1047-1062 (1998).

[16] W. Hundsdorfer, *Stability of approximate factorization with θ-methods*. BIT 39, pp. 473-483 (1999).

[17] W. Hundsdorfer, J.G. Verwer, *Stability and convergence of the Peaceman-Rachford ADI method for initial-boundary value problems*. Math. Comp. 53, pp. 81-101 (1989).

[18] G.I. Marchuk, *Splitting and alternating direction methods*. Handbook of Numerical Analysis 1 (P.G. Ciarlet. J.L. Lions, eds.), North-Holland, Amsterdam, pp. 197-462, 1990.

[19] A.R. Mitchell, D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*. John Wiley & Sons, Chichester, 1980.

[20] D.W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*. Developments in Petroleum Science 6, Elsevier, Amsterdam, 1977.

[21] J.M. Sanz-Serna, J.G. Verwer, W. Hundsdorfer, *Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations*. Numer. Math. 50, pp. 405-418 (1987).

[22] J.G. Verwer, J.G. Blom and W. Hundsdorfer, *An implicit-explicit approach for atmospheric transport-chemistry problems*, Appl. Numer. Math. 20, pp. 191-209 (1996).

[23] J.G. Verwer, E.J. Spee, J.G. Blom and W. Hundsdorfer, *A second order Rosenbrock method applied to photochemical dispersion problems*, SIAM J. Sci. Comput. 20, pp. 1456-1480 (1999).

[24] R.F. Warming, R.M. Beam, *An extension of A-stability to alternating direction methods*. BIT 19, pp. 395-417 (1979).

[25] N.N. Yanenko, *The Method of Fractional Steps*. Springer Verlag, Berlin, 1971.