Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Stability of Approximate Factorization with theta-Methods

W.H. Hundsdorfer

Modelling, Analysis and Simulation (MAS)

MAS-R9721 September 30, 1997

Report MAS-R9721
ISSN 1386-3703

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

Stability of Approximate Factorization with θ-Methods

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

ABSTRACT
Approximate factorization seems for certain problems a viable alternative to time splitting. Since a splitting error is avoided, accuracy will in general be favourable compared to time splitting methods. However, it is not clear to what extent stability is affected by factorization. Therefore we study here the effects of factorization on a simple, low order method, namely the θ-method. For this simple method it is possible to obtain rather precise results, showing limitations of the approximate factorization approach.

1991 Mathematics Subject Classification: 65L20, 65M12, 65M20
Keywords and Phrases: Numerical analysis, splitting methods, approximate factorizations.
Note: Background research for the project LOTOS in the TASC Project HPCN for Environmental Applications. Work carried out under project MAS 1.4 "Exploratory research: Discretization of Initial Value Problems".

1. Introduction

Space discretization of multi-dimensional advection-diffusion-reaction equations leads to very large ODE systems

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

where F contains reaction terms and discretized spatial operators in the various directions. With standard implicit methods one has to solve at each time step a nonlinear system involving the whole function F. This may be troublesome with respect to computing time and memory.

Often this function F can be decomposed into simpler components,

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

For example, the individual F_j may contain discretized spatial derivatives in one direction, or a specific operation, such as chemistry. Fractional step (time splitting) and approximate factorization methods employ this decomposition by solving subsequently subproblems that involve only one of the components F_j in an implicit manner. In this paper it will be assumed that the term F_0 is nonstiff, or mildly stiff, so that this term can be treated explicitly. The other terms will be treated implicitly, in an approximate factorized fashion. Here, the effect of such an approximate factorization procedure on stability will be discussed.

As starting point we consider the so-called θ-method

$$u_{n+1} = u_n + (1-\theta)\tau F(u_n) + \theta\tau F(u_{n+1}), \tag{1.3}$$

where $\theta \geq \frac{1}{2}$ is a parameter. Here $u_n \approx u(t_n)$ and $\tau = t_{n+1} - t_n > 0$. The method is A-stable for any $\theta \geq \frac{1}{2}$. It has order 2 if $\theta = \frac{1}{2}$, and 1 otherwise. We will mainly look at the well known cases $\theta = \frac{1}{2}$, the trapezoidal rule, and $\theta = 1$, the implicit Euler method.

Linearization of (1.3) leads to the θ-Rosenbrock method

$$u_{n+1} = u_n + (I - \theta\tau A)^{-1}\tau F(u_n) \tag{1.4}$$

with $A = A(u_n) \approx F'(u_n)$. This method has order 2 if $\theta = \frac{1}{2}$ and $(A(u) - F'(u))F(u) = \mathcal{O}(\tau)$. Otherwise the order is 1. The linear stability properties of this Rosenbrock method are the same as those of the original θ-method; if $F(u) = Au$ then (1.3) and (1.4) are identical.

We consider the form where in the Jacobian approximation the nonstiff term is omitted and the rest is factorized in approximate fashion, that is

$$u_{n+1} = u_n + \Big(\prod_{j=1}^{s}(I - \theta\tau A_j)\Big)^{-1}\tau F(u_n) \tag{1.5}$$

with $A_j \approx F_j'(w_n)$. The order of this approximate factorization method is 1 in general. For second order we need $\theta = \frac{1}{2}$ and $F_0 = 0$. (For second order methods of this type with $F_0 \neq 0$ see [8].) Implementation of (1.5) only requires the solution of linear systems involving matrices $I - \theta\tau A_j$. Since, in general, there will be much decoupling, this makes such schemes attractive candidates for parallel computations.

In the following stability of this method will be analyzed for the scalar test equation where

$$F_j(w) = \lambda_j w, \qquad A_j = \lambda_j. \tag{1.6}$$

In applications for PDEs these λ_j will represent eigenvalues for the various components, found by inserting Fourier modes. Let $z_j = \tau\lambda_j$. Applied to this test equation method (1.5) yields $u_{n+1} = R\,u_n$ with amplification factor R given by

$$R = 1 + \Big(\prod_{j=1}^{s}(1 - \theta z_j)\Big)^{-1}(z_0 + z) \qquad \text{with} \qquad z = \sum_{j=1}^{s} z_j. \tag{1.7}$$

In this paper conditions on the z_j will be given to ensure that $|R| \leq 1$. Note that due to A-stability of method (1.4), it is sufficient for that method to have all z_j in the left half complex plane. As we shall see, for the factorized method (1.5) additional constraints in the z_j have to be imposed. We shall consider constraints of the type $|\arg(-z_j)| \leq \alpha$ with angle $\alpha \leq \frac{1}{2}\pi$. In the purely parabolic case it is sufficient to consider $\alpha = 0$, but with advection-diffusion problems we need $\alpha > 0$. The larger the angle α, the more advection is allowed to dominate.

Remark. Stabilty results for the scalar test equation can be easily generalized to linear systems if the matrices A_j commute. For linear PDE problems with constant coefficients, if the A_j contain discretized spatial derivatives in different directions, these matrices A_j may

indeed be assumed to commute. Some results for noncommuting negative definite matrices were given in [2], but under very restrictive step size conditions.

Approximate factorization methods were introduced by Beam and Warming [1]. The computational effort in such methods is comparable to time splitting methods, or fractional step methods, where subproblems $v'(t) = F_j(v(t))$ are solved sequentially on each time interval $[t_n, t_{n+1}]$. Such a fractional step approach introduces an additional error, the so called splitting error. This error is already present for stationary problems and may give rise, for example, to unphysical steady state solutions. This is avoided in the approximate factorization approach, but we will see that there stability may be affected.

Related methods can also be derived in the ADI approach of Douglas and Gunn [2]. For linear problems with $F_0 = 0$, method (1.5) with $\theta = \frac{1}{2}$ reduces to the ADI method of Brian and Douglas, whereas for $\theta = 1$ we reobtain the ADI Douglas-Rachford method, see [2, 5, 6]. In fact, the results presented here are extensions of results in [4] for the Douglas scheme.

In this paper we shall restrict ourselves to the one-stage, one-step θ-method as underlying scheme. Approximate factorization methods based on multi-step schemes were derived by Warming and Beam [9]. In that paper it was shown that unconditional stability may be lost with $s \geq 3$ if all eigenvalues are on the imaginary axis. In Verwer et al. [8] a similar approach was tested with a two-stage, second order Rosenbrock method for atmospheric transport-chemistry. A precise stability analysis for such, more sophisticated methods is difficult. The results in this paper may serve as a guideline for such methods, in the sense that we will show limitations of the factorized approximation approach that are already present for the simple θ-method with $\theta = 1$ or $\frac{1}{2}$.

Method (1.5) results if one Newton step with approximate factorization is applied to the θ-method. Another possibility is to solve the implicit relation in the θ-method (1.3) iteratively with such a modified Newton process (where the arising Newton Jacobian is factorized in approximate fashion). When applied to the test equation this iteration process has a convergence factor, see [3],

$$S = 1 - (\prod_{j=1}^{s}(1 - \theta z_j))^{-1}(1 - \theta(z_0 + z)) \qquad \text{with} \qquad z = \sum_{j=1}^{s} z_j, \tag{1.8}$$

and for the iteration to converge we need $|S| < 1$. As we shall see, for the linear test problem the stability results for the factorized Rosenbrock methods have close counterparts for the convergence of this iteration process. An analysis of such factorized iterations for more general methods, of multi-step or Runge-Kutta type, has been given by Eichler-Liebenow, van der Houwen and Sommeijer [3].

2. Stability of the factorized Rosenbrock method

For notation, put $w = \prod_{j=1}^{s}(1 - \theta z_j)$. Then the amplification factor (1.7) of the factorized Rosenbrock method can be written as

$$R = 1 + w^{-1}(z_0 + z). \tag{2.1}$$

Here z_0 corresponds to a term that is treated explicitly. For this term we consider the choices $z_0 = 0$ and $|1 + z_0| \leq 1$. The other z_j will be assumed to belong to the wedge

$W_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \le \alpha\}$ in the left half-plane. In the following it will always be tacitly assumed that $\alpha \le \frac{1}{2}\pi$.

If there is no explicit term, that is $z_0 = 0$, then the statement $|R| \le 1$ is equivalent with $|1 + w^{-1}z| \le 1$, or

$$|z + w| \le |w|. \tag{2.2}$$

Theorem 2.1. Let $z_0 = 0$ and $\theta \ge \frac{1}{2}$. We have

$$|R| \le 1 \text{ for all } z_j \in W_\alpha \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,.$$

Proof. For $\theta = \frac{1}{2}$ the result was proven in [4]. Let $\zeta_j = 2\theta z_j$, $\zeta = 2\theta z$. Since

$$1 + \frac{z}{\prod_j (1 - \theta z_j)} = (1 - \frac{1}{2\theta}) + \frac{1}{2\theta}\Big(1 + \frac{\zeta}{\prod_j (1 - \frac{1}{2}\zeta_j)}\Big),$$

it follows that $z_j \in W_\alpha$ is also sufficient for having $|R| \le 1$ with $\theta > \frac{1}{2}$.

Necessity can be shown as in [3, 4]. Here we give a slightly simpler proof. Note that

$$\text{Re } R > 1 \quad \Longleftrightarrow \quad \text{Re}\,\Big(\bar{z}\prod_j (1 - \theta z_j)\Big) > 0.$$

Now take $z_j = -t\,e^{i\alpha}$ with $t > 0$ large. Then $\bar{z} = -s\,t\,e^{-i\alpha}$ and

$$\text{Re}\,\Big(\bar{z}\prod_j (1 - \theta z_j)\Big) = -\text{Re}\,\Big(s\,t\,e^{-i\alpha}(\theta^s t^s e^{si\alpha} + \mathcal{O}(t^{s-1})\Big) = -s\theta^s t^{s+1}\cos((s-1)\alpha) + \mathcal{O}(t^s).$$

Thus we see that the real part of R can be larger than 1, and consequently also $|R| > 1$, if $(s-1)\alpha > \frac{1}{2}\pi$. □

In this theorem we get the same result for $\theta = \frac{1}{2}$ or 1. This is somewhat surprising since the underlying θ-method is merely A-stable for $\theta = \frac{1}{2}$, whereas it is L-stable for $\theta = 1$. The fact that we get the same angles is caused by large values of the z_j. Near the origin the case $\theta = 1$ allows room for improvement. This will be shown in the following by considering $z_0 \ne 0$, $|1 + z_0| \le 1$. This condition on z_0 corresponds to the stability requirement in case all other z_j are zero.

By observing that R can be written as $R = w^{-1}((1 + z_0) + (w + z - 1))$ it is easily seen that $|R| \le 1$ for all $|1 + z_0| \le 1$ iff it holds that

$$1 + |w + z - 1| \le |w|. \tag{2.3}$$

Theorem 2.2. Suppose $\theta = \frac{1}{2}$. Then $|R| \le 1$ for all $z_j \in W_\alpha$ and $|1 + z_0| \le 1$ iff

$$\alpha = 0.$$

Proof. If $s = 1$, then condition (2.3) reads

$$1 + \frac{1}{2}|z| \le |1 - \frac{1}{2}z|,$$

which can only hold if z is real and negative. So, already for $s = 1$ we get $\alpha = 0$ as necessary condition.

On the other hand, for arbitrary s, if all z_j are real and negative, then $z \le 0$, $w \ge 1 - \frac{1}{2}z$, and from this it easily follows that (2.3) is fulfilled. □

Theorem 2.3. Suppose $\theta = 1$ and $s \le 3$. Then $|R| \le 1$ for all $z_j \in W_\alpha$ and $|1 + z_0| \le 1$ iff

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

Proof. Necessity of the bound on α follows from Theorem 2.1. As for sufficiency, we have to show that inequality (2.3) holds for all $z_j \in W_\alpha$ with $(s-1)\alpha \le \frac{1}{2}\pi$. Since we are looking for the maximum of $|R|$, the maximum modulus theorem may be employed, so that we only have to verify the inequality for $z_j = -t_j e^{\pm i\alpha}$ with $t_j \ge 0$.

First, consider $s = 2$. Then $w = 1 - z + z_1 z_2$, and thus (2.3) reads

$$1 + |z_1 z_2| \le |1 - z_1||1 - z_2|.$$

It is easily verified that this holds for arbitrary $z_j = \pm i t_j$ on the imaginary axis.

Now let $s = 3$, and denote $\gamma = \cos\alpha$. It has to be shown that (2.3) holds with $\alpha = \frac{1}{4}\pi$, that is $\gamma = \frac{1}{2}\sqrt{2}$. We have

$$|w| = \prod_{j=1}^{3} |1 - z_j| = \Big(\prod_{j=1}^{3}(1 + 2\gamma t_j + t_j^2)\Big)^{1/2}.$$

By a straightforward calculation we get

$$|w| = \sqrt{1 + \eta + \xi} \tag{2.4}$$

where

$$\eta = 2\gamma \sum_j t_j + \sum_j t_j^2 + 4\gamma^2 \sum_{j<k} t_j t_k + 2\gamma\Big(t_1(t_2^2 + t_3^2) + t_2(t_1^2 + t_3^2) + t_3(t_1^2 + t_2^2)\Big) + 8\gamma^3 t_1 t_2 t_3,$$

$$\xi = \sum_{j<k} t_j^2 t_k^2 + 4\gamma^2 (\sum_j t_j) t_1 t_2 t_3 + 2\gamma (\sum_{j<k} t_j t_k) t_1 t_2 t_3 + (t_1 t_2 t_3)^2,$$

with all summations from 1 to s. Since $\gamma^2 = \frac{1}{2}$ this last term can be written as

$$\xi = (\sum_{j<k} t_j t_k)^2 + 2\gamma (\sum_{j<k} t_j t_k) t_1 t_2 t_3 + (t_1 t_2 t_3)^2. \tag{2.5}$$

Further we have

$$\eta \ge 2\Big(\sum_{j<k} t_j t_k + t_1 t_2 t_3\Big). \tag{2.6}$$

To estimate $|w + z - 1|$ we consider two cases separately: case (I) where $z_j = -t_j e^{i\alpha}$ $(j = 1, 2, 3)$, and case (II) where $z_j = -t_j e^{i\alpha}$ $(j = 1, 2)$, $z_3 = -t_3 e^{-i\alpha}$. These two cases cover in essence all possibilities $z_j = -t_j e^{\pm i\alpha}$.

Case (I): We have

$$w+z-1 = e^{2i\alpha}\Big(\sum_{j<k} t_j t_k + e^{i\alpha} t_1 t_2 t_3\Big), \qquad |w+z-1| = \sqrt{\xi}.$$

From (2.5),(2.6) it easily follows that $4\xi \le \eta^2$. Hence

$$1+\sqrt{\xi} \le \sqrt{1+\eta+\xi},$$

and the inequality (2.3) follows.

Case (II): Here we have

$$w+z-1 = t_1 t_3 + t_2 t_3 + e^{i\alpha} t_1 t_2 t_3 + e^{2i\alpha} t_1 t_2.$$

For arbitrary real $p, q, r > 0$ it holds that

$$|p + e^{i\alpha} q + e^{2i\alpha} r| \le |p + r + e^{i\alpha} q|,$$

since $\arg(p+e^{i\alpha}q)$ will be closer to $\arg(r)$ than to $\arg(e^{2i\alpha}r)$. Thus it is seen that $|w+z-1| \le \sqrt{\xi}$. Hence (2.3) follows in the same way as in the previous case, which concludes the proof for $s = 3$. □

Note. Some numerical calculations suggest that the result of the above theorem is also valid for $s > 3$, but a proof of this is lacking.

Similar as in [4], we can also consider the case where we assume a priori that several z_j are real, negative. Then one may hope that for the other, complex z_j a wider angle α will be allowed.

Theorem 2.4. Let $1 \le r < s$. Assume either $\{\theta \ge \frac{1}{2}, z_0 = 0\}$ or $\{\theta = 1, |1+z_0| \le 1, s \le 3\}$. Further assume $z_1, \cdots, z_{s-r} \in W_\alpha$ and $z_{s-r+1}, \cdots, z_s \le 0$. We have $|R| \le 1$ for all such z_j iff

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

Proof. To begin with, suppose that $z_s \le 0$. If we consider fixed $z_0, z_1, \cdots, z_{s-1}$ then R is fractional linear in z_s with real denominator,

$$R = (1-\theta z_s)^{-1}(\xi - \eta\theta z_s),$$

where ξ, η correspond to the values of R for $z_s = 0, \infty$, respectively. If $z_s \le 0$ this is a convex combination of ξ and η, and thus $|R| \le 1$ for all $z_s \le 0$ iff this holds for $z_s = 0$ and $z_s = -\infty$. In case $z_s = 0$ we get a same inequality as before, only with s replaced by $s-1$. For $z_s = -\infty$ we have to verify whether

$$|1 - \frac{1}{\theta}\prod_{j=1}^{s-1}(1-\theta z_j)^{-1}| \le 1.$$

Continuing in this fashion, assuming that r of the terms z_j are real and negative, it is seen that $|R| \le 1$ for all $z_s, z_{s-1}, \cdots, z_{s-r+1} \le 0$ iff

$$|1 + \Big(\prod_{j=1}^{s-r}(1-\theta z_j)\Big)^{-1}(z_0 + \sum_{j=1}^{s-r} z_j)| \le 1 \quad \text{and} \quad |1 - \frac{1}{\theta}\prod_{j=1}^{s-r}(1-\theta z_j)^{-1}| \le 1.$$

The first inequality is of the same form as considered before, only with s replaced by $s-r$. As in [4, Sect. 2.2] the latter inequality is easily shown to hold for all $z_1, \cdots, z_{s-r} \in W_\alpha$ iff $\alpha \le \pi/(2(s-r))$. So, combining this with the results of the Theorems 2.1, 2.3, the proof follows. □

Note that for $r = 1$ we have the same result as for $r = 0$. To get a wider angle for the complex eigenvalues we need at least two negative, real z_j. Another consequence is the following: if $s \ge 3$, $z_1, ..., z_{s-1} \in W_\alpha$ with $\alpha \le \pi/(2(s-2))$ then we have stability in case $z_s = 0$, but letting $z_s < 0$ requires the tighter bound $\alpha \le \pi/(2(s-1))$. In other words, adding a purely diffusive term may destroy stability.

3. Convergence of factorized iterations

By denoting again $w = \prod_{j=1}^{s}(1-\theta z_j)$, the convergence factor (1.8) of the Newton iteration with approximate factorization can be written as

$$S = 1 - w^{-1}(1 - \theta z_0 - \theta z). \tag{3.1}$$

First we consider the case where $z_0 = 0$, that is, the explicit term is absent.

Theorem 3.1. Let $z_0 = 0$ and $\theta \ge \frac{1}{2}$. We have

$$|S| \le 1 \text{ for all } z_j \in W_\alpha \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,.$$

Further, if $1 \le r < s$ we have

$$|S| \le 1 \text{ for all } z_1, \cdots, z_{s-r} \in W_\alpha,\ z_{s-r+1}, \cdots, z_s \le 0 \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-r}\,\frac{\pi}{2}\,.$$

Proof. Without loss of generality, we may take $\theta = 1$. The results for $r \ge 1$ follow from those for $r = 0$ as in Theorem 2.4, so we consider here only $r = 0$. In the following, all summations are from 1 to s, unless indicated otherwise.

We have $S = 1 + w^{-1}(z-1)$. Thus $|S| \le 1$ iff $|w + z - 1|^2 \le |w|^2$, that is

$$\text{Re}\, Q \ge 0 \qquad \text{with} \qquad Q = -(\bar{z} - 1)(2w + z - 1). \tag{3.2}$$

This last form will be used here to show sufficiency. Note that according to the maximum modulus theorem it is only necessary to consider $z_j = -t_j e^{\pm i\alpha}$ with $t_j \ge 0$. By some calculations we get

$$Q = (1-\bar{z})(1 - z + 2\sum_{j<k} z_j z_k - 2\sum_{j<k<l} z_j z_k z_l + \cdots + 2(-1)^s z_1 z_2 ... z_s).$$

First, assume that $z_j = -t_j e^{i\alpha}$ for all j. Let

$$q_1 = \sum_j t_j, \quad q_2 = \sum_{j<k} t_j t_k, \quad q_3 = \sum_{j<k<l} t_j t_k t_l, \quad \dots \quad , q_s = t_1 t_2 ... t_s .$$

Then

$$\text{Re } Q = \text{Re } (1 + e^{-i\alpha} q_1)(1 + e^{i\alpha} q_1 + 2e^{2i\alpha} q_2 + 2e^{3i\alpha} q_3 + \cdots + 2e^{si\alpha} q_s) =$$

$$= \cos\alpha \; q_1 + 1 + q_1^2 + \cos\alpha (q_1 + 2q_1 q_2) + 2 \sum_{j=2}^{s-1} \cos(j\alpha)(q_j + q_1 q_{j+1}) + 2\cos(s\alpha) q_s .$$

If $(s-1)\alpha \le \frac{1}{2}\pi$ then $\cos(j\alpha) \ge 0$ for $1 \le j \le s-1$ and $\cos((s-2)\alpha) \ge |\cos(s\alpha)|$. Further, $q_1 q_{s-1} \ge q_s$, and so it follows that Re $Q \ge 0$.

Next, consider the case where some z_j are $-t_j e^{i\alpha}$ and some are $-t_j e^{-i\alpha}$. Then it easily follows that Re $(2w + 2z - 1)$ will be a sum of $\cos(j\alpha)$ terms with positive coefficients and with $0 \le j \le s-1$. Hence Re $(2w + 2z - 1) \ge 0$. Further, since $|w + z| \le |w|$, see Theorem 2.1 and (2.2), we also know that Re $\bar{z}(2w + z) \le 0$. Therefore, by writing

$$\text{Re } Q = \text{Re } \Big((2w + z - 1) + \bar{z} - \bar{z}(2w + z) \Big) = \text{Re } \Big((2w + 2z - 1) - \bar{z}(2w + z) \Big),$$

it again follows that Re $Q \ge 0$.

Finally we note that necessity of the bound $(s-1)\alpha \le \frac{1}{2}\pi$ follows as in proof of Theorem 2.1 by considering the inequality Re $S > 1$. □

The result in the above theorem for the θ-method is a generalization of Lemma 2.1 and Theorem 2.6 in [3], where sufficiency of the bound on α was shown for $s = 3$, $r = 0$, and necessity for $s \ge 2$, $r = 0$. In [3] more general ODE methods were considered.

From the Figures 1, 2 it can be seen that the estimations with wedges in the theorems are in fact quite close. In these figures, with $\theta = 1$ and $\theta = \frac{1}{2}$, respectively, the boundaries of the stability region $|R| \le 1$ and convergence region $|S| \le 1$ are plotted for the special case with $s = 3$ and all z_j equal. Left and right pictures are on different scales. Also included in the plots, as dotted curved lines, are contour lines for $|S|$ at $0.1, 0.2, \cdots, 0.9$. From this it is seen that we will have fast convergence of the factorized iteration only relatively close to the origin. (To accelerate convergence some kind of smoothing seems necessary, see also Verwer [7].)

Similar pictures with $s > 3$ for the special case where all z_j are equal also show that away from the origin there is a tight fit with the wedges of the above theorem.

The fact that we get approximately the same region of convergence $|S| \le 1$ and stability $|R| \le 1$ is somewhat disappointing for the factorized iteration approach: roughly spoken, if the iteration converges then the cheaper method (1.5) is stable. On the other hand, if the iteration converges then it yields the A-stable method (1.3) and to verify whether an iteration converges is easier than detecting instability.

We now take $z_0 \ne 0$. As condition on z_0 we consider $|\theta z_0| \le 1$. Note that for such z_0 we just have $|S| \le 1$ if the other z_j are zero, so this is the case were the explicit term is taken as large as possible.

Theorem 3.2. The results of Theorem 3.1 remain valid if $|\theta z_0| \le 1$, $s \le 3$.

Proof. Again, without loss of generality, we may take $\theta = 1$. Let $\tilde{z}_0 = z_0 - 1$. Then $|1+\tilde{z}_0| \leq 1$ and $S = 1 + w^{-1}(\tilde{z}_0 + z)$, the same form as considered in the previous section. Therefore we may apply the Theorems 2.3 and 2.4. □

As with Theorem 2.3, this result will probably also be valid for $s > 3$, but a proof of this is lacking.

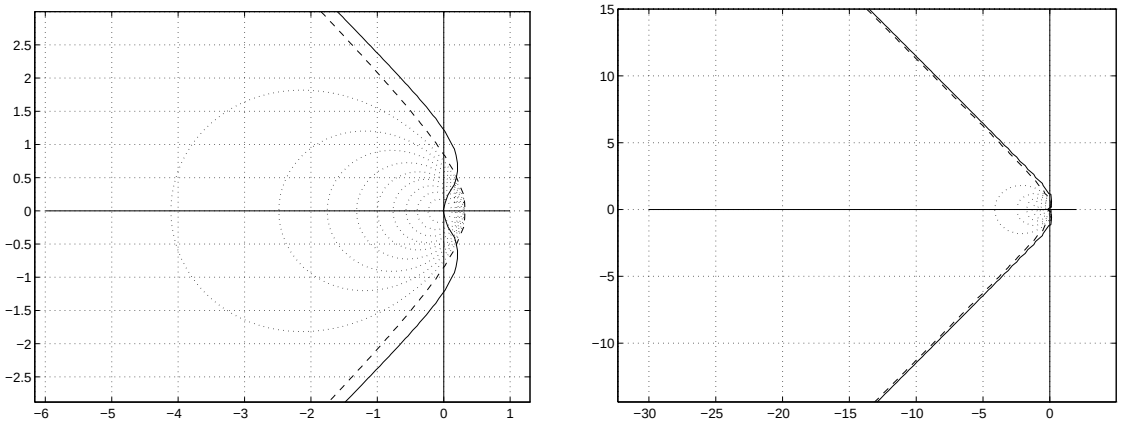

Figure 1. Regions of stability (solid) and convergence (dashed) for $s = 3$, $\theta = 1$ with special choice $z_1 = z_2 = z_3$.

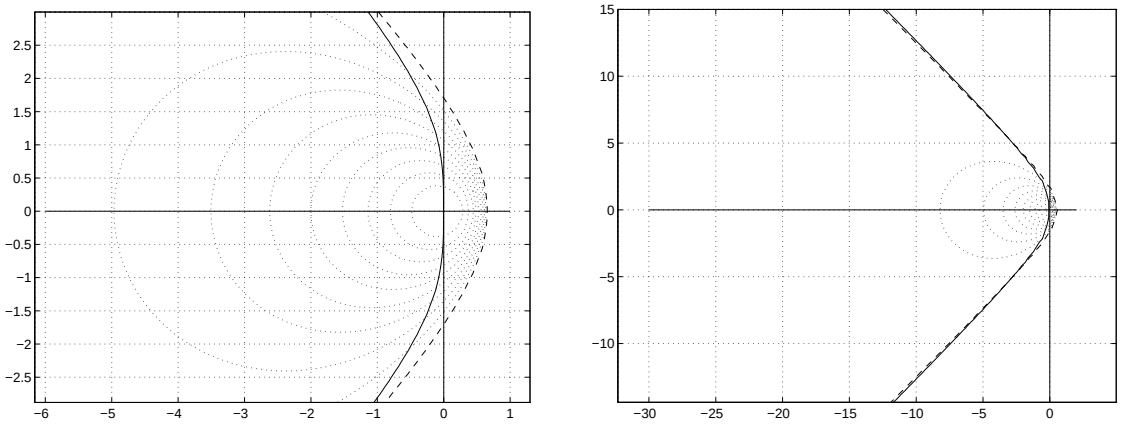

Figure 2. Regions of stability (solid) and convergence (dashed) for $s = 3$, $\theta = \frac{1}{2}$ with special choice $z_1 = z_2 = z_3$.

4. General remarks and conclusions

Similar as for $|S|$, see the contour lines in Figure 1 and 2, also $|R|$ will often assume values close to 1 for z_j in the stability region. If two of the z_j tend to $-\infty$ then $R \to 1$, so there will

be no damping. Such lack of damping might manifest itself as inaccuracy for solutions with discontinuities or steep wave fronts.

To see whether this leads to an additional step size restriction for nonsmooth problems, some numerical tests were performed on the parabolic equation

$$u_t = \frac{1}{\kappa}(u_{xx} + u_{yy}) + \kappa u^2(1-u), \quad 0 \le x, y \le 1,\ 0 \le t \le 1,$$

with solution $u(x,y,t) = (1 + \exp(\frac{1}{2}\kappa(x+y-t)))^{-1}$. This is a traveling wave that crosses the region diagonally. If κ becomes large the wave becomes steeper, and one might expect the lack of damping to become visible. However, in the experiments the schemes (1.5) did produce good results for those step sizes for which the ODE $u_t = \kappa u^2(1-u)$ could be solved with reasonable accuracy. For large κ, solution of the ODE part gives here a far more severe step size restriction than the lack of damping.

A related, but potentially more dangerous phenomenon is the fact that for many z_j outside the stability regions the value of $|R|$ will only be slightly larger than 1. Then the scheme is unstable but this instability will be difficult to detect. A numerical example is given in [4] for a 2D advection problem with a stiff chemistry term. The advection term gives z_1, z_2 close to the imaginary axis and the chemistry term gives rise to $z_3 << 0$. According to Theorem 2.4 we can expect instability. Experiments in [4] showed that these instabilities sometimes build up very slowly and may only become visible on long time intervals.

There are some practical conclusions that can be drawn from the results presented in this paper:

- For purely parabolic problems ($z_j \le 0$, $1 \le j \le s$), the schemes (1.5) are stable for $\theta \ge \frac{1}{2}$, also with an explicit term. (With $z_0 = 0$ this is a well known result, see [2, 5, 6], for instance.)

- For more general problems (z_j in the left half plane), there will be a substantial loss of stability if $s \ge 3$, even without explicit term. If an explicit term is included, the loss of stability may occur with $\theta = \frac{1}{2}$ already for $s = 1$. The scheme with $\theta = 1$ seems rather insensitive to the inclusion of an explicit term.

- Instabilities may be slow and difficult to detect. In situations where these might occur, the factorized iteration approach will be more robust, but there slow convergence must be expected for solutions that are not very smooth.

Based on this, it seems that approximate factorizations are only suited for restricted classes of problems, where characteristics of the solution are more or less known in advance, so that stability can be well predicted. Within a general purpose environment, the factorized Rosenbrock schemes (1.5) are not sufficiently robust and the factorized iteration approach will often be too slow. On the other hand, for restricted classes of problems, the Rosenbrock schemes with approximate factorization will lead to codes that are easy to program, very efficient and potentially well suited for parallel computations.

Acknowledgement. The author thanks J.G. Blom and J.G. Verwer for helpful discussions on the relevance of the theoretical results.

References

[1] R.M. Beam, R.F. Warming, *An implicit finite-difference algorithm for hyperbolic systems in conservation-law form.* J. Comp. Phys. 22, pp. 87-110 (1976).

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

[3] C. Eichler-Liebenow, P.J. van der Houwen & B.P. Sommeijer *Analysis of approximate factorization in iteration methods.* CWI Report MAS-R97··, to appear.

[4] W. Hundsdorfer, *A note on stability of the Douglas splitting method.* CWI Report NM-R9606, to appear in Math. Comp. 1998.

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

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

[7] J.G. Verwer, *On iterated defect correction and the LOD-method for parabolic equations.* In: MC Syllabus 44, Colloquium Numerical Solution of Partial Differential Equations, J.G. Verwer (ed.), Mathematical Center, Amsterdam, 1980.

[8] J.G. Verwer, E. Spee, J.G. Blom & W. Hundsdorfer, *A second order Rosenbrock method applied to photochemical dispersion problems*, CWI Report MAS-R97··, to appear.

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

Appendix
(not for publication)

In this appendix some details and figures are given concerning the numerical results mentioned in Section 4 on the loss of damping and the slow onset of instabilities.

As for lack of damping, in the following figure the values of R are plotted for $z_j = -x$, all equal on the negative axis for $s = 1, 2, 3$. Near the origin, R approximates e^{-sx}, but for larger values of x we see that R tends to 1 if $s \geq 2$. Note that with $\theta = 1$ the damping is more quickly lost if $s > 1$ than with $\theta = \frac{1}{2}$.

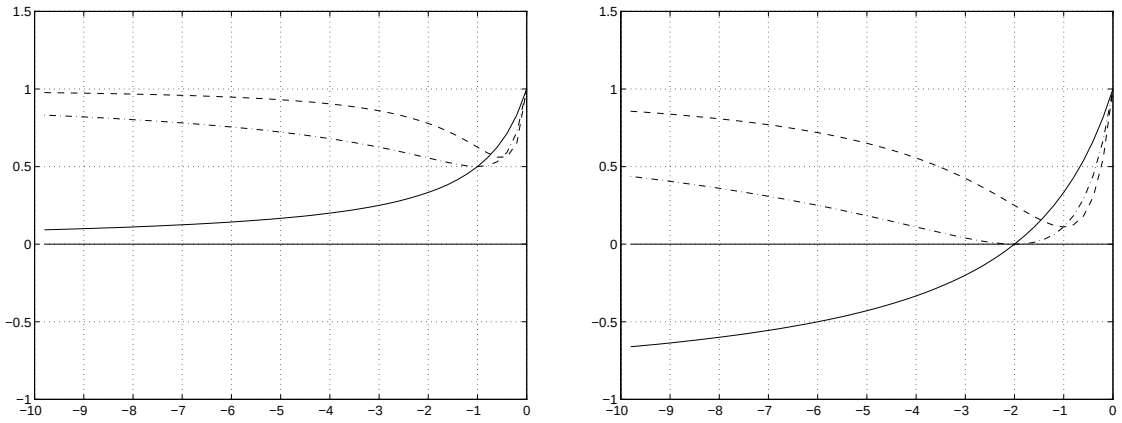

Figure A.1. Values of R for s=1(solid), s=2(dash-dot) and $s = 3$(dashed) with negative $z_j = -x$, $1 \leq j \leq s$. Left picture $\theta = 1$, right picture $\theta = \frac{1}{2}$.

The lack of damping might lead to inaccurate results for discontinuous solutions or steep wave fronts, since high frequency Fourier modes may be poorly treated. To verify the relevance of this, several numerical tests were performed for the parabolic equation

$$u_t = \frac{1}{\kappa}(u_{xx} + u_{yy}) + \kappa u^2(1-u), \quad (x,y) \in [0,1]^2,\ 0 \leq t \leq 1, \tag{A.1}$$

with solution $u(x,y,t) = (1 + \exp(\frac{1}{2}\kappa(x + y - t)))^{-1}$. If $\kappa > 0$ becomes large the traveling wave becomes steeper, and one might expect the lack of damping to become visible.

Numerical solutions at $t = 1$ with $\kappa = 100$ are given in Figure A.2 for $\theta = \frac{1}{2}$ and Figure A.3 for $\theta = 1$. The solutions are computed on a 40×40 grid with step size $\tau = 1/40$. Space discretization is done with second order central differences, and at the boundaries Dirichlet conditions are prescribed. We consider splitting with F_1, F_2 defined by the finite difference operators for diffusion in the x and y direction, respectively, and with F_3 defined by the nonlinear source term. The lay-out in these figures is as follows: left top picture exact solution, right top picture numerical solution, left bottom picture contour lines (dotted for exact solution) and bottom right picture the cross section for $x = y$.

The results for $\theta = \frac{1}{2}$ are good, but for $\theta = 1$ we have a wrong propagation speed of the wave. Note that the wave travels too fast, whereas lack of damping could be expected to give a wave speed that is too slow.

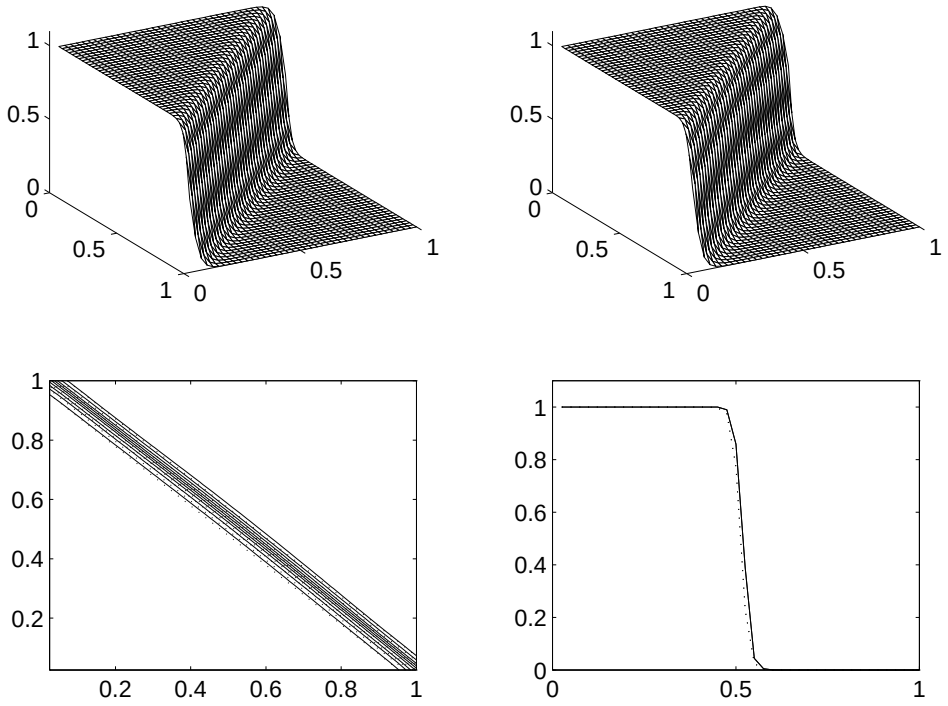

Figure A.2. Numerical solutions reaction-diffusion problem (A.1) for $\kappa = 100$ and $\theta = \frac{1}{2}$.

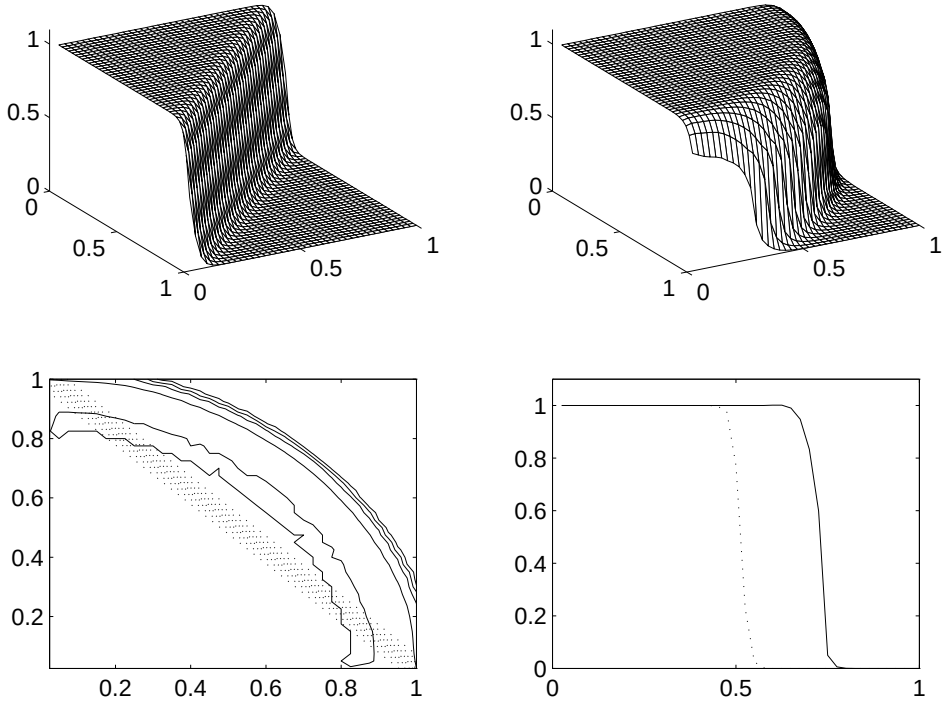

Figure A.3. Numerical solutions reaction-diffusion problem (A.1) for $\kappa = 100$ and $\theta = 1$.

Indeed the results in Figure A.3 are not caused by lack of damping, but by inaccuracy in the ODE part

$$u_t = \kappa u^2(1-u).$$

Numerical solutions with $\kappa = 100$ and $\tau = 1/40, 1/80$ with the Rosenbrock method (1.4) are presented in Figure A.4. We see that the numerical results for $\theta = 1$ tend much too fast to the stable steady state solution $u = 1$. For accurate solution of the ODE part we need smaller step sizes, and with smaller step sizes also the results for equation (A.1) become correct. The same was also observed with larger values of κ.

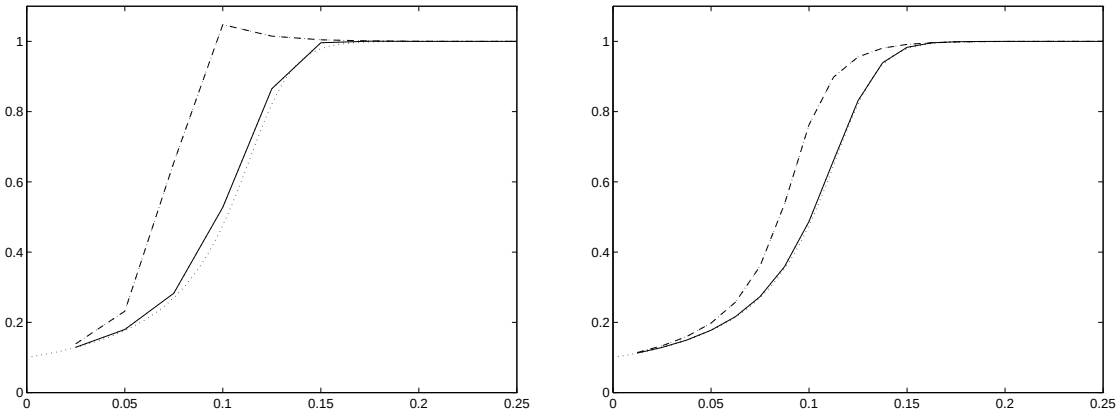

Figure A.4. Numerical solutions of $u_t = 100u^2(1-u), u(0) = 0.1$ with $\theta = \frac{1}{2}$ (solid), $\theta = 1$ (dashed) and exact solution (dotted). Left picture $\tau = 1/40$, right picture $\tau = 1/80$.

As an illustration of the slow onset of instability, we repeated an experiment of [4] on the following advection equation with a simple linear reaction term,

$$u_t = au_x + bu_y + G\,u, \quad (x,y) \in [0,1]^2,\ 0 \le t. \tag{A.2}$$

The velocities are given by $a(x,y,t) = 2\pi(y - \frac{1}{2})$, $b(x,y,t)) = 2\pi(\frac{1}{2} - x)$. Further,

$$u = u(x,y,t) = \begin{pmatrix} u_1(x,y,t) \\ u_2(x,y,t) \end{pmatrix}, \qquad G = \begin{pmatrix} -k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix}.$$

We take $k_1 = 1$. The second reaction constant k_2 can be used to vary the stiffness of the reaction term, and is taken here as 2000. Note that the matrix G has eigenvalues 0 and $-(k_1 + k_2)$, and we have a chemical equilibrium if $u_1/u_2 = k_2/k_1$.

The initial condition is chosen as

$$u_1(x,y,0) = c, \quad u_2(x,y,0) = (1-c) + 100\,k_2^{-1}\exp(-80((x-\tfrac{1}{2})^2 - 80(y-\tfrac{3}{4})^2),$$

with $c = k_2/(k_1 + k_2)$. After the short transient phase, where most of the Gaussian pulse is transfered from u_2 to u_1, this is purely an advection problem, and the velocity field gives a

rotation around the center of the domain. At $t = 1$ one rotation is completed. The exact solution is easily found by superimposing the solution of the reaction term onto the rotation caused by the advection terms, see [4].

Dirichlet conditions are prescribed at the inflow boundaries. At the outflow boundaries we use standard upwind discretization, in the interior second order central differences are used. We consider splitting with F_1, F_2 the finite difference operators for advection in the x and y direction, respectively, and with F_3 defined by the linear reaction term. The corresponding eigenvalues λ_1, λ_2 will be close to the imaginary axis whereas $\lambda_3 = 0$ or $-(k_1 + k_2)$. The test has been performed on a fixed 80×80 grid, and with $\tau = 1/160$.

The numerical solution of the first component u_1 for the scheme with $\theta = \frac{1}{2}$ is given in in Figure A.5 at time $t = 1$ (top left), $t = 2$ (top right), $t = 3$ (bottom left) and $t = 4$ (bottom right). There are some smooth oscillations in the wake of the Gaussian pulse, but these are caused by the spatial discretization with central differences. The instabilities occur near the corners where both advection speeds, in x and y direction, are large. The build up of the instabilities is very slow, and therefore it will be difficult to detect this with error estimators.

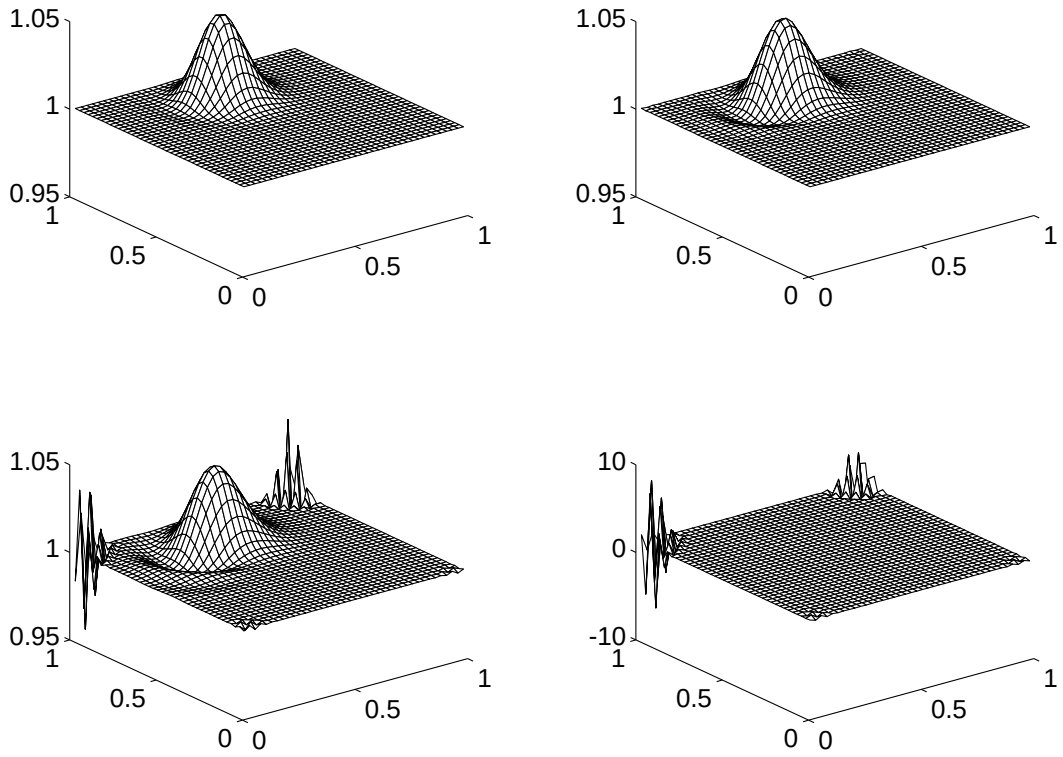

Figure A.5. Numerical solutions advection-reaction problem (A.2) at $t = 1, 2, 3, 4$.