Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Factorization in block-triangularly implicit methods for shallow water applications

P.J. van der Houwen, B.P. Sommeijer

Modelling, Analysis and Simulation (MAS)

MAS-R9906 March 1999

Report MAS-R9906
ISSN 1386-3703

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

Factorization in Block-Triangularly Implicit Methods for Shallow Water Applications

P.J. van der Houwen & B.P. Sommeijer
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT
The systems of first-order ordinary differential equations obtained by spatial discretization of the initial-boundary value problems modelling phenomena in shallow water in 3 spatial dimensions have righthand sides of the form $\mathbf{f}(t,\mathbf{y}) := \mathbf{f}_1(t,\mathbf{y}) + \mathbf{f}_2(t,\mathbf{y}) + \mathbf{f}_3(t,\mathbf{y}) + \mathbf{f}_4(t,\mathbf{y})$, where $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$ contain the spatial derivative terms with respect to the x, y and z directions, respectively, and $\mathbf{f}_4$ represents the forcing terms and/or reaction terms. The number N of components of $\mathbf{f}$ is usually extremely large. It is typical for shallow water applications that the function $\mathbf{f}_4$ is nonstiff and that the function $\mathbf{f}_3$ corresponding with the vertical spatial direction is much more stiff than the functions $\mathbf{f}_1$ and $\mathbf{f}_2$ corresponding with the horizontal spatial directions. The reason is that in shallow seas the gridsize in the vertical direction is several orders of magnitude smaller than in the horizontal directions. In order to solve the initial value problem for the system of ordinary differential equations numerically, we need a stiff solver. Stiff IVP solvers are necessarily implicit, requiring the solution of large systems of implicit relations. In a few earlier papers, we considered implicit Runge-Kutta methods leading to *fully coupled*, implicit systems whose dimension is a multiple of N, and *block-diagonally* implicit methods in which the implicit relations can be decoupled into subsystems of dimension N. In the present paper, we analyse Rosenbrock type methods and the related DIRK methods (diagonally implicit Runge-Kutta methods) leading to *block-triangularly* implicit relations. In particular, we shall present a convergence analysis of various iterative methods based on approximate factorization for solving the triangularly implicit relations.

1991 Mathematics Subject Classification: 65L06
Keywords and Phrases: numerical analysis, shallow water applications, iteration methods, approximate factorization, parallelism.
Note. The investigations reported in this paper were partly supported by the Dutch HPCN Program.

1. Introduction

We consider initial-boundary value problems modelling phenomena in shallow water in 3 spatial dimensions. The systems of ordinary differential equations (ODEs) obtained by spatial discretization (method of lines) of the governing partial differential equations can be written in the form

$$\frac{d\mathbf{y}(t)}{dt} = \mathbf{f}(t,\mathbf{y}(t)), \quad \mathbf{f}(t,\mathbf{y}) := \mathbf{f}_1(t,\mathbf{y}) + \mathbf{f}_2(t,\mathbf{y}) + \mathbf{f}_3(t,\mathbf{y}) + \mathbf{f}_4(t,\mathbf{y}), \qquad \mathbf{y}, \mathbf{f}_k \in \mathbb{R}^N, \tag{1.1}$$

where $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$ contain the spatial derivative terms with respect to the x, y and z directions, respectively, $\mathbf{f}_4$ represents the forcing terms and/or reaction terms, and N is a large integer proportional to the number of spatial grid points used for the spatial discretization. It is typical for shallow water applications that the function $\mathbf{f}_4$ is nonstiff and that the function $\mathbf{f}_3$ corresponding with the vertical spatial direction is much more stiff than the functions $\mathbf{f}_1$ and $\mathbf{f}_2$ corresponding with the horizontal spatial directions. As a consequence, the spectral radius of the Jacobian matrix $\partial\mathbf{f}_3/\partial\mathbf{y}$ is much larger than the spectral radius of $\partial\mathbf{f}_1/\partial\mathbf{y}$ and $\partial\mathbf{f}_2/\partial\mathbf{y}$. The reason is that in shallow seas the

gridsize in the vertical direction is several orders of magnitude smaller than in the horizontal directions.

In order to solve the initial value problem (IVP) for the system (1.1) numerically, we need a stiff IVP solver, because the Lipschitz constants with respect to **y** associated with the functions $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$ become increasingly large as the spatial resolution is refined. Stiff IVP solvers are necessarily implicit, requiring the solution of large systems of implicit relations. In a few earlier papers, we considered implicit Runge-Kutta methods leading to *fully coupled*, implicit systems whose dimension is a multiple of N (cf. [3], [5] and [10]), and *block-diagonally* implicit methods in which the implicit relations can be decoupled into subsystems of dimension N (cf. [6]). In the present paper, we analyse Rosenbrock type methods and the related DIRK methods (diagonally implicit Runge-Kutta methods) leading to *block-triangularly* implicit relations (this is also the case for the DIRK methods, in spite of the terminology 'diagonally implicit'). Rosenbrock type methods, and in particular factorized versions of these methods, are quite popular in air pollution simulations (see e.g. [9], [12], and [13]). This motivated us to look whether Rosenbrock and the related DIRK methods can also be useful in shallow water modelling. First we show that in shallow water applications, factorized Rosenbrock methods are less suitable. However, iteration of Rosenbrock and DIRK methods using approximate factorization looks quite promising. This paper will focus on the convergence analysis of approximate factorization iteration of the triangularly implicit Rosenbrock and DIRK relations.

2. Rosenbrock methods and their factorization

We start with an example of a family of two-stage Rosenbrock methods:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + b\mathbf{k}_1 + (1\text{-}b)\mathbf{k}_2,$$

$$(I - \kappa_1 \Delta t J)\mathbf{k}_1 = \Delta t\, \mathbf{f}(\mathbf{y}_n), \tag{2.1}$$

$$(I - \kappa_2 \Delta t J)\mathbf{k}_2 = \Delta t\, \mathbf{f}(\mathbf{y}_n + \mu\mathbf{k}_1) + \nu \Delta t J \mathbf{k}_1, \quad \kappa_i > 0, \quad \mu := \frac{\frac{1}{2} - b\kappa_1 + (b-1)\kappa_2}{1 - b} - \nu.$$

Here, b, κ_1, κ_2 and ν are free parameters and J is an approximation to the Jacobian matrix $\partial\mathbf{f}/\partial\mathbf{y}$ at t_n. For simplicity of notation, we assumed the ODE of autonomous form. The nonautonomous version can be obtained by applying (2.1) to the augmented system $\{\mathbf{y}' = \mathbf{f}(y_0, \mathbf{y}),\ y_0' = 1\}$. The method (2.1) is triangularly implicit, that is, $\mathbf{k}_1$ and $\mathbf{k}_2$ can be computed by successively solving 2 linear systems of dimension N.

If $J = \partial\mathbf{f}/\partial\mathbf{y}(t_n) + O(\Delta t)$, then the formulas (2.1) are all second-order accurate Rosenbrock methods. The stability function for (2.1) is given by

$$R(z) = \frac{1 + (1 - \kappa_1 - \kappa_2)z + \frac{1}{2}(1 - 2\kappa_1 - 2\kappa_2 + 2\kappa_1\kappa_2)z^2}{(1 - \kappa_1 z)(1 - \kappa_2 z)}. \tag{2.2}$$

From this expression it follows that the methods (2.1) are A-stable if $\frac{1}{2} \le \kappa_1 + \kappa_2 \le 2\kappa_1\kappa_2 + \frac{1}{2}$ and L-stable if $\kappa_1 + \kappa_2 = \kappa_1\kappa_2 + \frac{1}{2}$.
The first examples of Rosenbrock methods were given by Rosenbrock [8] in 1962 and are obtained by choosing in (2.1)

$$(2.3) \qquad b = 0, \quad \kappa_1 = \kappa_2 = \kappa := 1 \pm \tfrac{1}{2}\sqrt{2}, \quad \nu = 0.$$

Of particular interest are the methods which remain second-order accurate if we choose an arbitrary matrix for J. Such methods are called Rosenbrock-W methods and were proposed by Steihaug and Wolfbrandt [11]. If we choose in (2.1) $\kappa_1 = \kappa_2 = \kappa$ and $\nu = -\kappa(1-b)^{-1}$, then (2.1) becomes a W-method (see Dekker and Verwer [2, p. 233]). The special case

$$(2.4) \qquad b = \tfrac{1}{2}, \quad \kappa_1 = \kappa_2 = \kappa := 1 \pm \tfrac{1}{2}\sqrt{2}, \quad \nu = -2\kappa$$

was used by Verwer et al. [12] for solving atmospheric transport problems. Note, however, that for stability reasons, J should be a reasonably close approximation to the true Jacobian $\partial\mathbf{f}/\partial\mathbf{y}$ at t_n.

2.1. General Rosenbrock methods

More generally, we consider Rosenbrock methods of the form (cf. [4, p. 111])

$$(2.5) \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + (\mathbf{b}^T \otimes I)\mathbf{K}, \quad (I - T \otimes \Delta t J)\mathbf{K} = \Delta t\, \mathbf{F}(\mathbf{e} \otimes \mathbf{y}_n + (L \otimes I)\mathbf{K}),$$

where $\mathbf{b}$ is an s-dimensional vector, $\mathbf{K} := (\mathbf{k}_1^T, \ldots, \mathbf{k}_s^T)^T$, and T and L are lower and strictly lower triangular s-by-s matrices, respectively. This property of T and L implies that (2.5) is triangularly implicit, so that the components $\mathbf{k}_i$ of $\mathbf{K}$ can be computed by successively solving s linear systems of dimension N with system matrices $I - \kappa_i \Delta t J$, where the κ_i denote the diagonal entries of T. If the order of the method (2.5) is independent of the choice of the Jacobian approximation J, then (2.5) is called a Rosenbrock-W method.
If T is not diagonal (as in (2.4)), then for an actual implementation one often transforms the linear system for $\mathbf{K}$ by a Butcher similarity transformation $\mathbf{U} = (T \otimes I)\mathbf{K}$, where T is assumed invertible (cf. [4, p. 120]). Writing $T^{-1} = S + D^{-1}$ with S strictly lower triangular and $D = \mathrm{diag}(T)$, (2.5) becomes

$$(2.6) \qquad \begin{aligned} &\mathbf{y}_{n+1} = \mathbf{y}_n + (\mathbf{b}^T T^{-1} \otimes I)\mathbf{U}, \\ &(I - D \otimes \Delta t J)\mathbf{U} = \Delta t\, (D \otimes I)\mathbf{F}(\mathbf{e} \otimes \mathbf{y}_n + (LT^{-1} \otimes I)\mathbf{U}) - (DS \otimes I)\mathbf{U}. \end{aligned}$$

As in (2.5) the components $\mathbf{u}_i$ of $\mathbf{U}$ can be computed by again successively solving s linear systems of dimension N. As an example of a transformed Rosenbrock method, we give the transformation of the method (2.4):

$$(2.4') \qquad \begin{aligned} &\mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{2}\kappa^{-1}(3\mathbf{u}_1 + \mathbf{u}_2), \\ &(I - \kappa\Delta t J)\mathbf{u}_1 = \kappa\Delta t\, \mathbf{f}(\mathbf{y}_n), \quad \kappa = 1 \pm \tfrac{1}{2}\sqrt{2}, \\ &(I - \kappa\Delta t J)\mathbf{u}_2 = \kappa\Delta t\, \mathbf{f}(\mathbf{y}_n + \kappa^{-1}\mathbf{u_1}) - 2\mathbf{u}_1. \end{aligned}$$

Note that unlike (2.5), no Jacobian multiplications are involved in transformed Rosenbrock methods. In general, this is considered as an advantage because such Jacobian multiplications can be quite expensive. However, it should be remarked that in the case of shallow water applications the matrix J is extremely sparse, so that Jacobian multiplications are not so costly.

2.2. Factorized Rosenbrock methods

In order to further reduce the linear algebra costs in the method (2.4), Sandu [9] and Verwer et al. [13] applied to the system matrix $I - \kappa\Delta tJ$ the technique of *approximate factorization* based on some splitting $\sum J_k$ of the Jacobian J. This leads to the *factorized Rosenbrock method*.

This technique goes back to Peaceman and Rachford [7] who used it for approximately solving the linear systems originating from a finite difference discretization of two-dimensional parabolic problems. In such problems, the system matrix is of the form $I - \frac{1}{2}\Delta tJ$, where J is the discretization of the Laplace operator $\partial^2/\partial x^2 + \partial^2/\partial y^2$. By writing $J = J_1 + J_2$, where J_1 and J_2 correspond with $\partial^2/\partial x^2$ and $\partial^2/\partial y^2$, respectively, Peaceman and Rachford replaced $I - \frac{1}{2}\Delta tJ$ by the approximate factorization $(I - \frac{1}{2}\Delta tJ_1)\,(I - \frac{1}{2}\Delta tJ_2)$.

The same approximate factorization technique can be applied to the matrix $I - T\otimes\Delta tJ$ in (2.5) or to the matrix $I - D\otimes\Delta tJ$ in (2.6). We shall illustrate this for the case (2.6). Since we are concerned with shallow water applications, we use the splitting $J = J_1 + J_2 + J_3$, where the matrices J_k denote the Jacobian matrices of the terms $\mathbf{f}_k$ at t_n occurring in the righthand side function **f** in (1.1) and where the nonstiff interaction terms are ignored. This leads to the factorized method

$$
\begin{aligned}
&\mathbf{y}_{n+1} = \mathbf{y}_n + (\mathbf{b}^T T^{-1}\otimes I)\mathbf{V},\\
&\Pi\mathbf{V} = \Delta t\,(D\otimes I)\mathbf{F}\big(\mathbf{e}\otimes\mathbf{y}_n + (LT^{-1}\otimes I)\mathbf{V}\big) - (DS\otimes I)\mathbf{V},
\end{aligned}
\tag{2.7}
$$

where Π is defined by

$$
\Pi := \big(I - D\otimes\Delta tJ_1\big)\big(I - D\otimes\Delta tJ_2\big)\big(I - D\otimes\Delta tJ_3\big),\quad D = \mathrm{diag}(T).
\tag{2.8}
$$

Each step of the factorized Rosenbrock method (2.7) requires the solution of 3s one-dimensional, linear systems. All LU-decompositions can be computed in parallel, but the 3s forward-backward substitutions have to be done sequentially.

Since $\Pi = I - D\otimes\Delta tJ + O((\Delta t)^2)$, we can interpret the factorized method as the original Rosenbrock method with an $O(\Delta t)$-perturbed matrix J. Hence, factorization will not affect the order of Rosenbrock-W methods. Furthermore, any factorized Rosenbrock method has at least order two if the original Rosenbrock method has at least order two.

2.3. Stability

Next, we define the stability region $\mathbb{S}$ for the factorized versions of the methods (2.5) and (2.6). We first define the stability function by applying them to the test equation $\mathbf{y}' = (J_1 + J_2 + J_3)\mathbf{y}$. Assuming that the matrices J_k commute and ignoring the interaction terms in **F**, the factorized versions of the methods (2.5) and (2.6) will reduce to recursions of the form

$$\mathbf{y}_{n+1} = R(\Delta tJ_1, \Delta tJ_2, \Delta tJ_3)\mathbf{y}_n,$$

where $R(z_1,z_2,z_3)$ is a rational function of its arguments. Using the identity

$$1 + \mathbf{p}^T M^{-1}\mathbf{q} = \frac{\det(M + \mathbf{q}\mathbf{p}^T)}{\det(M)},$$

which holds for any m-by-m matrix M and any two m-dimensional vectors **p** and **q** (cf. [1, p. 475]), we find that the stability functions corresponding to the factorized versions of (2.5) and (2.6) can be respectively expressed as

$$(2.9) \qquad R(z_1,z_2,z_3) = \frac{\det\left(P + z(\mathbf{e}\mathbf{b}^T - L)\right)}{\det(P)}, \qquad P := (I - z_1T)(I - z_2T)(I - z_3T),$$

$$(2.10) \qquad R(z_1,z_2,z_3) = \frac{\det\left(P + DS + zD(\mathbf{e}\mathbf{b}^T - L)T^{-1}\right)}{\det(P)}, \qquad P := (I - z_1D)(I - z_2D)(I - z_3D).$$

where $z := z_1 + z_2 + z_3$.
The *stability region* is defined by the region $\mathbb{S}$ in the (z_1, z_2, z_3)-space where $|R(z_1,z_2,z_3)| \le 1$. The method (2.7) is called *stable* if all eigenvalue triples $(\Delta t\lambda(J_1), \Delta t\lambda(J_2), \Delta t\lambda(J_3))$ are in $\mathbb{S}$. Since in shallow water applications, many of the eigenvalues of J_k, k = 1, 2, 3, are close to the imaginary axis, we are particularly interested in the most critical case where the eigenvalues of J_k are purely imaginary, i.e. $z_k = iy_k$ with y_k real-valued. Let us introduce for a given value of y_3 the stability boundary $\beta(y_3)$ which is such that the method is stable in a region of the form

$$(2.11) \qquad \mathbb{S}(y_3) := \{(y_1,y_2): |y_k| \le \beta(y_3), \ k = 1, 2\},$$

where the *stability boundary* $\beta(y_3)$ is not too small. Since the spectral radius of ΔtJ_1 and ΔtJ_2 is much smaller than that of ΔtJ_3, we would like stability in all regions $\mathbb{S}(y_3)$, $|y_3| \le \infty$. The corresponding timestep condition is given by

$$(2.12) \qquad \Delta t \le \frac{\beta}{\max\{\rho(J_1),\rho(J_2)\}}, \quad \beta := \min_{y_3} \beta(y_3).$$

Let us consider the stability of the factorized versions of (2.3) and (2.4'). It is easily verified that their stability functions respectively take the form

$$(2.13) \qquad R_1(z_1,z_2,z_3) := 1 + \frac{z}{(1 - \kappa z_1)(1 - \kappa z_2)(1 - \kappa z_3)} + \frac{\frac{1}{2}(1 - 2\kappa)z^2}{(1 - \kappa z_1)^2(1 - \kappa z_2)^2(1 - \kappa z_3)^2},$$

$$(2.14) \qquad R_2(z_1,z_2,z_3) := 1 + \frac{2z}{(1 - \kappa z_1)(1 - \kappa z_2)(1 - \kappa z_3)} + \frac{\frac{1}{2}z^2 - z}{(1 - \kappa z_1)^2(1 - \kappa z_2)^2(1 - \kappa z_3)^2}$$

and that $|R_1(0,0,iy_3)| < 1$, $|R_2(0,0,iy_3)| < 1$ for $y_3 \neq 0$. Hence, we have a nonzero stability boundary β. However, a numerical calculation reveals that β is quite small (less than 1/10). Hence, the factorized versions of (2.3), (2.4) and (2.4') are of no use in shallow water applications.

3. Approximate factorization iteration

The quite poor stability properties of the factorized Rosenbrock methods can be explained by observing that the vector $(T^{-1}\otimes I)\mathbf{V}$ defined by the factorized-Rosenbrock method (2.7) is too far away from the vector $\mathbf{K} = (T^{-1}\otimes I)\mathbf{U}$ defined by the Rosenbrock method (2.5). In this section, we improve the stability by really solving the implicit relations in the underlying Rosenbrock method by an iteration process. We shall also study the iterative solution of the related implicit methods

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t(\mathbf{b}^T\otimes I)\mathbf{F}(\mathbf{X}), \quad \mathbf{X} - \Delta t(A\otimes I)\mathbf{F}(\mathbf{X}) = \mathbf{e}\otimes\mathbf{y}_n, \tag{3.1}$$

where A is a lower triangular matrix. In [4, p.97] these methods are called DIRK methods (diagonally implicit Runge-Kutta methods). Like Rosenbrock methods, DIRK methods are triangularly implicit (in spite of the terminology 'diagonally implicit' now commonly accepted in the literature).
An advantage of the iterative approach is that we can rely on the stability of the underlying integration method. Thus, by choosing an A-stable integration method, we only have to deal with the region of *convergence* of the iteration method. The iteration processes considered below are based on the approximate factorization technique used in the preceding section and lead to acceptably large convergence regions.

3.1. Iterative solution of the Rosenbrock equations

We consider two iterative approximate factorization approaches for actually solving the implicit Rosenbrock relations. The first approach solves the components $\mathbf{u}_i$ from (2.6) one by one by *repeated* application of a linear system solver, the second approach solves all components $\mathbf{k}_i$ from (2.5) *simultaneously* by a nonlinear system solver. We shall refer to these iteration methods as repeated and simultaneous approximate factorization iteration of the Rosenbrock method, briefly, the RAF-Rosenbrock and SAF-Rosenbrock processes, respectively.

3.1.1. The RAF-Rosenbrock process. The s linear systems in (2.6) have the form

$$\begin{aligned} &(I - \kappa_i\Delta t J)\mathbf{u}_i = \mathbf{g}_i, \quad i = 1, \dots, s, \\ &\mathbf{g}_i := (\mathbf{e}_i^T\otimes I)\big(\Delta t\,(D\otimes I)\mathbf{F}\big(\mathbf{e}\otimes\mathbf{y}_n + (LT^{-1}\otimes I)\mathbf{U}\big) - (DS\otimes I)\mathbf{U}\big), \end{aligned} \tag{3.2}$$

where κ_i is the ith diagonal entry of T. Since L and S are strictly lower triangular, these s systems can be solved successively. We solve the ith linear system by the linear solver

$$(\mathbf{e}_i^T\otimes I)\Pi\left(\mathbf{u}_i^{(j)} - \mathbf{u}_i^{(j-1)}\right) = \mathbf{g}_i - (I - \kappa_i\Delta t J)\mathbf{u}_i^{(j-1)}, \quad j = 1, 2, \dots, m, \quad i = 1, \dots, s, \tag{3.3}$$

where Π is defined in (2.8). In this RAF-Rosenbrock process the initial iterate $\mathbf{u}_i^{(0)}$ should be provided by some predictor formula and the number of iterations m is assumed to be determined by some iteration strategy such that $\mathbf{u}_i^{(m)}$ may be considered as the solution $\mathbf{u}_i$ of (3.2).
If the iterates $\mathbf{u}_i^{(j)}$ converge, then they can only converge to this solution $\mathbf{u}_i$. Each iteration in (3.3) requires the solution of 3 linear systems with system matrices $I - \kappa_i \Delta t J_k$, k = 1, 2, 3, each of order N. Note that the three LU-decompositions of these system matrices can be done in parallel. These LU-decompositions and the corresponding forward-backward substitutions are relatively cheap, because the matrices J_k each correspond with a one-dimensional differential operator.
The convergence is determined by the error recursion satisfied by the iteration error $\varepsilon^{(j)}$:

$$\varepsilon^{(j)} := \mathbf{u}_i^{(j)} - \mathbf{u}_i \tag{3.4}$$

$$\varepsilon^{(j)} = Z_1 \varepsilon^{(j-1)}, \quad Z_1 := I - \Pi^{-1}(I - D\otimes\Delta t J), \quad j = 1, 2, \ldots, m. \tag{3.5}$$

Before analysing the matrix Z_1, we first derive the error recursion for the other iterative approaches.

3.1.2. The SAF-Rosenbrock process. Instead of solving the linear systems in the Rosenbrock method (2.6) successively for the components $\mathbf{u}_i$ of **U**, we may iterate them simultaneously. Since in such an approach it is more convenient to go back to the untransformed method (2.5), we shall solve the components $\mathbf{k}_i$ of **K** simultaneously from (2.5). Consider the SAF-Rosenbrock process

$$\Pi\left(\mathbf{K}^{(j)} - \mathbf{K}^{(j-1)}\right) = -\left((I - T\otimes\Delta t J)\mathbf{K}^{(j-1)} - \Delta t\, \mathbf{F}\left(\mathbf{e}\otimes\mathbf{y}_n + (L\otimes I)\mathbf{K}^{(j-1)}\right)\right), \; j = 1, 2, \ldots, m. \tag{3.6}$$

Note that this method is a *nonlinear* system solver.
Evidently, if the iterates $\mathbf{K}^{(j)}$ converge and if (2.5) has a unique solution **K**, then they can only converge to this solution **K**. Each SAF-Rosenbrock iteration requires the solution of 3 linear systems with system matrices $I - D\otimes\Delta t J_k$, k = 1, 2, 3, each of order sN. The 3s LU-decompositions and the s forward-backward substitutions corresponding with each matrix $I - D\otimes\Delta t J_k$ can be done in parallel. Again, the LU-decompositions and the forward-backward substitutions are relatively cheap, because J_k corresponds with a one-dimensional differential operator. A drawback is the matrix-vector multiplication in the righthand side of (3.6). Note that applying the SAF-Rosenbrock iteration process to (2.6) instead of (2.5) does not avoid such a matrix-vector multiplication.
Let us consider the iteration error $\varepsilon^{(j)} := \mathbf{K}^{(j)} - \mathbf{K}$. From (2.5) and (3.6) it follows that

$$\begin{aligned} &\varepsilon^{(j)} = Z_2\varepsilon^{(j-1)} + \Delta t\, \Pi^{-1}\mathbf{G}(\varepsilon^{(j-1)}), \quad Z_2 := I - \Pi^{-1}(I - (T+L)\otimes\Delta t J), \quad j = 1, 2, \ldots, m, \\ &\mathbf{G}(\varepsilon) := \mathbf{F}\left(\mathbf{e}\otimes\mathbf{y}_n + (L\otimes I)(\mathbf{K} + \varepsilon)\right) - \mathbf{F}\left(\mathbf{e}\otimes\mathbf{y}_n + (L\otimes I)\mathbf{K}\right) - (L\otimes J)\varepsilon. \end{aligned} \tag{3.7}$$

Since $\mathbf{G}(\varepsilon)$ has a small Lipschitz constant in the neighbourhood of the origin, the error recursion (3.7) essentially behaves as the linearized recursion

$$\varepsilon^{(j)} \approx Z_2\, \varepsilon^{(j-1)}, \quad Z_2 := I - \Pi^{-1}(I - (T+L)\otimes\Delta t J), \quad j = 1, 2, \ldots, m. \tag{3.8}$$

3.2. Iterative solution of DIRK equations

As for Rosenbrock methods, we may consider repeated and simultaneous approximate factorization iteration of the DIRK method (3.1). These processes are respectively given by

$$(\mathbf{e}_i^T\otimes I)\Pi\left(\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(j-1)}\right) = \mathbf{g}_i, \quad j = 1, 2, \dots, m, \quad i = 1, \dots, s,$$

(3.9)

$$\mathbf{g}_i := (\mathbf{e}_i^T\otimes I)\left((\mathbf{e}\otimes I)\mathbf{y}_n - \mathbf{X}^{(j-1)} + \Delta t(A\otimes I)\mathbf{F}\left(\mathbf{X}^{(j-1)}\right)\right),$$

and

$$\Pi\left(\mathbf{X}^{(j)} - \mathbf{X}^{(j-1)}\right) = -\left(\mathbf{X}^{(j-1)} - \Delta t(A\otimes I)\mathbf{F}\left(\mathbf{X}^{(j-1)}\right) - (\mathbf{e}\otimes I)\mathbf{y}_n\right), \quad j = 1, 2, \dots, m, \tag{3.10}$$

where Π is again defined by (2.8) with $D := \mathrm{diag}(A)$. They will be referred to as the RAF-DIRK and SAF-DIRK processes. A comparison with (3.3) and (3.6) shows that we have the same iteration costs except for the Jacobian multiplication.

Defining the iteration error $\varepsilon^{(j)} := \mathbf{X}^{(j)} - \mathbf{X}$, we can write down the linearized error recursions. We find that the linearized error recursions associated with the RAF-DIRK method (3.9) and the SAF-DIRK method (3.10) are respectively given by

$$\varepsilon^{(j)} = Z_3\varepsilon^{(j-1)}, \quad Z_3 := Z_1, \quad j = 1, 2, \dots, m, \tag{3.11}$$

$$\varepsilon^{(j)} \approx Z_4\,\varepsilon^{(j-1)}, \quad Z_4 := I - \Pi^{-1}(I - A\otimes\Delta tJ), \quad j = 1, 2, \dots, m. \tag{3.12}$$

3.3. Convergence

The convergence of iterated Rosenbrock methods (3.2) and (3.6), and of the iterated DIRK methods (3.9) and (3.10) is determined by the amplification matrices

$$Z_1 := I - \Pi^{-1}(I - D\otimes\Delta tJ), \quad Z_2 := I - \Pi^{-1}(I - (T+L)\otimes\Delta tJ),$$
$$Z_3 = Z_1, \qquad\qquad Z_4 := I - \Pi^{-1}(I - A\otimes\Delta tJ),$$

occurring in the error recursions (3.5), (3.8), (3.11) and (3.12), respectively. They only differ by the matrix in front of ΔtJ (we recall that $D = \mathrm{diag}(T) = \mathrm{diag}(A)$). In the following subsections we respectively discuss the region of convergence where $\rho(Z_r) < 1$, the rate of convergence of the nonstiff iteration error components, and the stability of the iterated methods.

3.3.1. The region of convergence. The matrices Z_r are lower triangular block matrices with the same diagonal blocks

$$I - (I - \kappa_j\Delta tJ_1)^{-1}(I - \kappa_j\Delta tJ_2)^{-1}(I - \kappa_j\Delta tJ_3)^{-1}(I - \kappa_j\Delta tJ), \quad j = 1, \dots, s,$$

for all r. Here, the κ_i denote the diagonal entries of D. Hence, the eigenvalues of the matrices Z_r are identical. They act as amplification factors for the eigenvalue components of the iteration error and are given by

(3.13) $$\alpha_j = C(\kappa_j z_1, \kappa_j z_2, \kappa_j z_3), \quad C(x_1, x_2, x_3) := 1 - \frac{1 - x_1 - x_2 - x_3}{(1 - x_1)(1 - x_2)(1 - x_3)},$$

where $j = 1, \ldots, s$ and where z_k runs through the eigenvalues of $\Delta t J_k$. Evidently, we have convergence if $|\alpha_j| < 1$, $j = 1, \ldots, s$. We consider the most critical case where the eigenvalues of J_k are purely imaginary, that is we consider the values of $|\alpha_j| = |C(i\kappa_j y_1, i\kappa_j y_2, i\kappa_j y_3)|$. Recalling that the spectral radius of $\Delta t J_1$ and $\Delta t J_2$ is much smaller than that of $\Delta t J_3$, we are interested in convergence regions of the form (cf. (2.11))

(3.14) $$\mathbb{C}(y_3) := \{(y_1, y_2): |y_k| \le \frac{\gamma(y_3)}{\rho(D)}, \ k = 1, 2\}, \ |y_3| \le \infty.$$

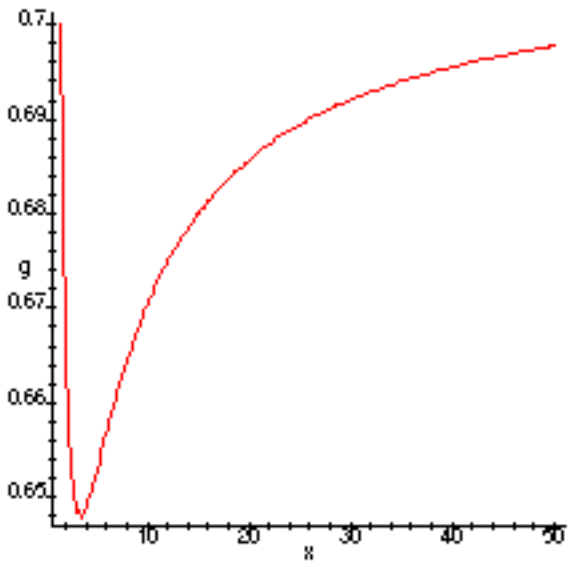

Figure 3.1. The function g(x) defined by (3.15).

Theorem 3.1. Let the function g(x) be defined by the relation

(3.15) $$4xg^3 + 2(x^2 - 1)g^2 - x^2 - 1 = 0.$$

Then, the convergence boundary $\gamma(y_3)$ in (3.14) is given by

(3.16) $$\gamma(y_3) = \rho(D) \min_j \frac{g(\kappa_j |y_3|)}{\kappa_j}$$

and the minimal value of $\gamma(y_3)$ is given by the positive root of the equation $4\gamma^4(\gamma^2 + 1) = 1$.

Proof. We verified that for given values of y_3, $|C(i\kappa_j y_1, i\kappa_j y_2, i\kappa_j y_3)|$ increases most rapidly along the line $y_1 = y_2$, so that we may restrict our considerations to the values of

$$|\alpha_j| = |C(i\kappa_j y_1, i\kappa_j y_1, i\kappa_j y_3)| = \kappa_j^2 |y_1| \left(\frac{y_1^2(\kappa_j^2 y_3^2 + 1) + 4y_1 y_3 + 4y_3^2}{(1 + \kappa_j^2 y_1^2)^2 (1 + \kappa_j^2 y_3^2)} \right)^{1/2}.$$

If we set $|\alpha_j| = 1$, $x = \kappa_j y_3$ and $y = \kappa_j y_1$, then we find the relation

$$4xy^3 + 2(x^2 - 1)y^2 - x^2 - 1 = 0.$$

This relation determines a real-valued function $y = g(x)$. Hence, for given values of κ_j and y_3, i.e. of x, we have $|\alpha_j| \le 1$ provided that both $\kappa_j |y_1|$ and $\kappa_j |y_2|$ are bounded by $g(\kappa_j |y_3|)$. This proves (3.16).
In order to find the minimal value of $\gamma(y_3)$, we look at the plot of the function g(x) (see Figure 3.1). Let $x_1(y)$ and $x_2(y)$ denote the two solutions of the equation $4xy^3 + 2(x^2 - 1)y^2 - x^2 - 1 = 0$. Then, the minimal value of g(x) is determined by the relation $x_1(y) = x_2(y)$. This leads to the equation $4y^4(y^2 + 1) = 1$ whose only positive root determines the minimal value of $\gamma(y_3)$. ♦

Since the positive root of the equation $4\gamma^4(\gamma^2 + 1) = 1$ is given by $\gamma = 0.647...$ we derive from this theorem the following convergence condition:

Theorem 3.2. Let $\lambda(J_k)$, k = 1, 2, 3, be purely imaginary. Then, a sufficient condition for convergence of the iterated Rosenbrock methods (3.3) and (3.6), and of the iterated DIRK methods (3.9) and (3.10) is given by

$$\Delta t \le \frac{\gamma}{\rho(D)\max\{\rho(J_1),\rho(J_2)\}}, \quad \gamma = 0.647... \ . \ \blacklozenge$$

3.3.2. The rate of convergence of the nonstiff error components. The rate of convergence of the *nonstiff* error components can be studied by the behaviour of the nonstiff amplification factors, that is, the eigenvalues of Z_r corresponding with small values of $\Delta t\, \lambda(J_k)$. From (3.13) it can be deduced that

$$\alpha_j = \kappa_j^2(z_1z_2 + z_1z_3 + z_2z_3) + O((\Delta t)^3), \quad z_k = \Delta t\, \lambda(J_k), \quad j = 1, \dots, s.$$

Hence, after m iterations the amplification factors behave as $O((\Delta t)^{2m})$ for all m and irrespective the value of r. However, this is not true for the amplification matrices Z_r^m.

Theorem 3.3. The amplification matrices Z_r satisfy the relations

$$r = 1, 3: \qquad Z_r^m = O((\Delta t)^{2m}) \qquad \text{for all m,}$$

$$r = 2, 4: \qquad \begin{cases} Z_r^m = O((\Delta t)^m) & \text{for } m \le s-1 \\ Z_r^m = O((\Delta t)^{2m+1-s}) & \text{for } m \ge s \end{cases}.$$

Proof. The relation $Z_1 = Z_3 = O((\Delta t)^2)$ immediately follows from the definition of Z_1 and Z_3 in (3.5) and (3.11). For Z_2 and Z_4 it follows from (3.8) and (3.12) that

$$Z_2 = I - (I + D\otimes\Delta tJ)(I - (T+L)\otimes\Delta tJ) + O((\Delta t)^2) = (T + L - D)\otimes\Delta tJ + O((\Delta t)^2).$$
$$Z_4 = I - (I + D\otimes\Delta tJ)(I - A\otimes\Delta tJ) + O((\Delta t)^2) = (A - D)\otimes\Delta tJ + O((\Delta t)^2).$$

Hence, we certainly have $Z_r^m = O((\Delta t)^m)$ for r = 2, 4. However, writing $Z_r = A_r + B_r$ with $A_2 := (T + L - D)\otimes\Delta tJ$ and $A_4 := (A - D)\otimes\Delta tJ$, and observing that A_2 and A_4 are strictly lower block triangular, so that A_2^j and A_4^j vanish for $j \geq s$, we obtain for $m \geq s$

$$Z_r^m = \binom{m}{m-s+1} A_r^{s-1} B_r^{m-s+1} + \ldots + \binom{m}{m} B_r^m, \quad r = 2, 4.$$

Since $A_r = O(\Delta t)$ and $B_r = O((\Delta t)^2)$, we find that

$$Z_r^m = O((\Delta t)^{2m-s+1}), \quad r = 2, 4. \ \blacklozenge$$

From this theorem it follows that in all four approaches the nonstiff error components are rapidly removed from the iteration error. However, we may expect that the RAF processes (3.5) and (3.9) damp these nonstiff components stronger than the SAF processes (3.8) and (3.12).

3.3.3. The region of stability. Evidently, if the iteration process converges, then the stability of the iterated method is determined by the stability of the underlying integration method. Hence, with respect to the stability test equation, the stability region of the iterated method converges to the intersection of the convergence region and the stability region of the integration method, that is, to

$$\mathbb{S} := \mathbb{S}_0 \cap \mathbb{C}, \quad \mathbb{C} := \bigcap_{y_3} \mathbb{C}(y_3), \ |y_3| \leq \infty,$$

where $\mathbb{S}_0$ is the stability region of the integration method and $\mathbb{C}(y_3)$ is defined by (3.14). For A-stable integration methods, the stability region $\mathbb{S}$ equals the convergence region $\mathbb{C}$, so that the stability condition is given by the stepsize condition in Theorem 3.2. Thus, for iterated, A-stable integration methods we may define the stability boundary $\beta := \gamma\rho^{-1}(D)$.
For example, if the Rosenbrock methods (2.3) and (2.4) are iterated using the iteration matrix Π, then we find in both cases the stability boundary $\beta \approx 2.20$. If we choose $\kappa_1 = \kappa_2 = \frac{1}{4}$ in (2.1), then (2.1) is still A-stable with a slightly greater stability boundary $\beta \approx 2.59$.

4. Explicit treatment of the horizontal terms

The modest values of the stability boundary β raises the question whether it is necessary to treat the horizontal terms implicitly. Afterall, when applying the standard, explicit, fourth-order Runge-Kutta method, we have an imaginary stability boundary of comparable size, viz. $\beta = 2\sqrt{2}$.

4.1. Fully explicit treatment of the horizontal terms

We once again consider the iteration methods (3.3), (3.6), (3.9) and (3.10), but we replace the iteration matrix Π by the matrix $\Pi_3 := I - D\otimes\Delta tJ_3$. As a consequence, the amplification factors are now given by (cf. (3.13))

(4.1) $$\alpha_j = C_3(\kappa_j z_1, \kappa_j z_2, \kappa_j z_3), \quad C_3(x_1, x_2, x_3) := 1 \; - \frac{1 - x_1 - x_2 - x_3}{1 - x_3} \; .$$

Hence, $\left| C_3(i\kappa_j y_1, i\kappa_j y_2, i\kappa_j y_3) \right|^2 = \kappa_j^2 (y_1 + y_2)^2 (1 + \kappa_j^2 y_3^2)^{-1}$, so that we have convergence in regions of the form (3.14) with

(4.2) $$\gamma(y_3) = \tfrac{1}{2}\sqrt{1 + y_3^2 \rho^2(D)} \; , \quad |y_3| \leq \infty.$$

Thus, the convergence boundary γ in the timestep condition in Theorem 3.2 changes from $\gamma \approx 0.647$ to $\gamma = \frac{1}{2}$. Since the iterations in (3.3), (3.6), (3.9) and (3.10) with Π replaced by Π_3 are cheaper, the modest reduction of the convergence boundary seems to be a small price. Moreover, the convergence boundary (4.2) quickly increases with $|y_3|$, whereas (3.16) approaches a constant value $\gamma(\infty) \approx 0.7$ (see Figure 3.1). Hence, the *stiff* error components will be more strongly damped when using the matrix Π_3. On the other hand, if we look at the behaviour of the amplification factor α_j as $\Delta t \to 0$, then we find $\alpha_j^m = O((\Delta t)^m)$, so that the *nonstiff* iteration error components are expected to require more iterations to be removed from the iteration error (see Theorem 3.3).
Thus, the iteration processes with Π and Π_3 both have advantages. This suggests a combination of the two iteration methods, for example, by iterating successively with iteration matrices $\Pi_3, \Pi, \Pi_3, \Pi, \ldots$ (the $\Pi_3\Pi$ process) or with $\Pi_3, \Pi, \Pi, \Pi_3, \Pi, \Pi, \ldots$ (the $\Pi_3\Pi^2$ process). Evidently, in such combined processes, the averaged iteration costs are still slightly higher than in the Π_3 process and slightly lower than in the Π process.

4.1.1. Convergence boundaries of the $\Pi_3\Pi$ and $\Pi_3\Pi^2$ processes. Let us first consider the $\Pi_3\Pi$ process more closely. After each two iterations, the corresponding amplification matrix Z_r in the linearized error recursion has diagonal blocks of the form

$$\left(I - ((\mathbf{e}_j^T \otimes I)\Pi_3)^{-1}(I - \kappa_j \Delta t J)\right)\left(I - ((\mathbf{e}_j^T \otimes I)\Pi)^{-1}(I - \kappa_j \Delta t J)\right), \quad j = 1, 2, \ldots, s,$$

irrespective the value of r. Hence, the amplification factors for two iterations are given by

(4.3) $$\alpha_j^2 = C_3(\kappa_j z_1, \kappa_j z_2, \kappa_j z_3)\, C(\kappa_j z_1, \kappa_j z_2, \kappa_j z_3),$$

where C and C_3 are defined in (3.13) and (4.1). Again, we restrict z_k to imaginary values iy_k and again it turns out that for given values of y_3 the amplification factor $|\alpha_j^2|$ increases most rapidly along the line $y_1 = y_2$. Along the lines of the proof of Theorem 3.1 it can be shown that the convergence boundary $\gamma(y_3)$ in (3.14) is given by (3.16) where the function g(x) is defined by

(4.4) $$4(x^2 + 1)g^6 + 16xg^5 - (x^4 - 14x^2 + 1)g^4 - 2(x^2 + 1)^2 g^2 - (x^2 + 1)^2 = 0.$$

Figure 4.1 presents a plot of g(x). It is easily seen that $g(x) \approx \frac{1}{2}x$ as $x \to \infty$, so that the combined iteration process has the property that the convergence boundary $\gamma(y_3)$ increases with y_3 to infinity (compare (4.2)). In order to find the minimal value γ of g(x), we write (4.4) as $F(g,x) = 0$. Then, γ is determined by (4.4) and by the equation

(4.5) $$\frac{\partial F(g,x)}{\partial x} = 4\big(2xg^6 + 4g^5 - x(x^2 - 7)g^4 - 2x(x^2 + 1)g^2 - x(x^2 + 1)\big) = 0.$$

Solving (4.4) and (4.5) yields the minimal value $\gamma \approx 0.72$. This is even slightly greater than the value $\gamma \approx 0.65$ given in Theorem 3.2 for the Π process.
Similarly, we find for the $\Pi_3\Pi^2$ process a still greater value $\gamma \approx 0.75$.

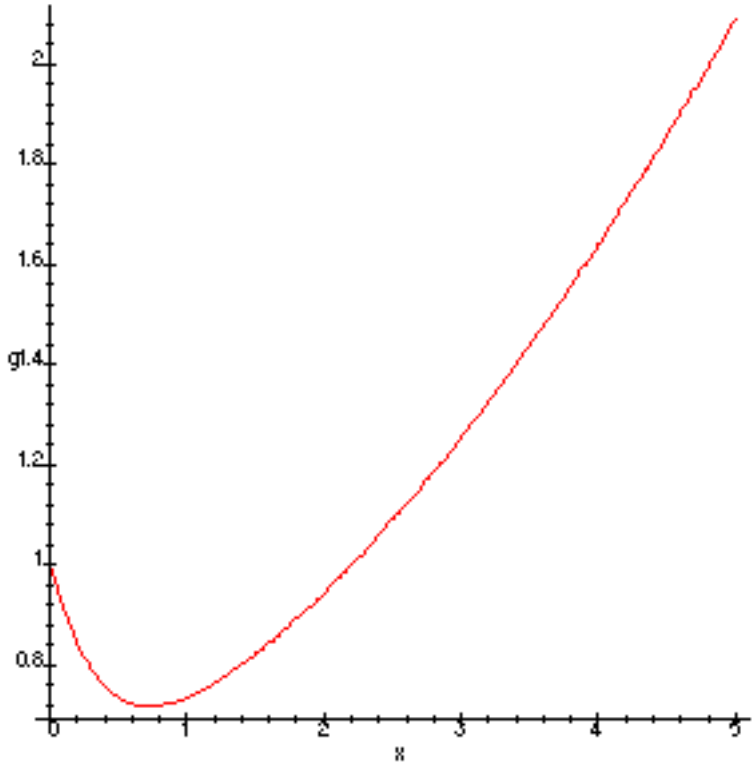

Figure 4.1. The function g(x) defined by (4.4).

4.1.2. The rate of convergence. The size of the convergence region is not the only issue to be considered. The *rate* of convergence, that is, the magnitude of the amplification factors, is equally important. We consider the behaviour of the amplification factors for small values of $\Delta t\, \lambda(J_k)$ and their overall behaviour. For the $\Pi_3\Pi$ process it can be deduced from (4.3) that for $\Delta t \to 0$

$$\alpha_j^2 = \kappa_j^3(z_1z_2 + z_1z_3 + z_2z_3)(z_1 + z_2) + O((\Delta t)^4), \quad z_k = \Delta t\, \lambda(J_k), \quad j = 1, \dots, s,$$

so that after an even number of m iterations the amplification factors behave as $O((\Delta t)^{3m/2})$, irrespective the value of r. Similarly, we have for the $\Pi_3\Pi^2$ process, $\alpha_j^m = O((\Delta t)^{5m/3})$, where m is assumed to be a multiple of 3. Recalling that the Π and Π_3 processes yield after m iterations amplification by $\alpha_j^m = O((\Delta t)^{2m})$ and $\alpha_j^m = O((\Delta t)^m)$, respectively, we see that with respect to damping of the nonstiff error components, the order of preference is Π, $\Pi_3\Pi^2$, $\Pi_3\Pi$, and Π_3.

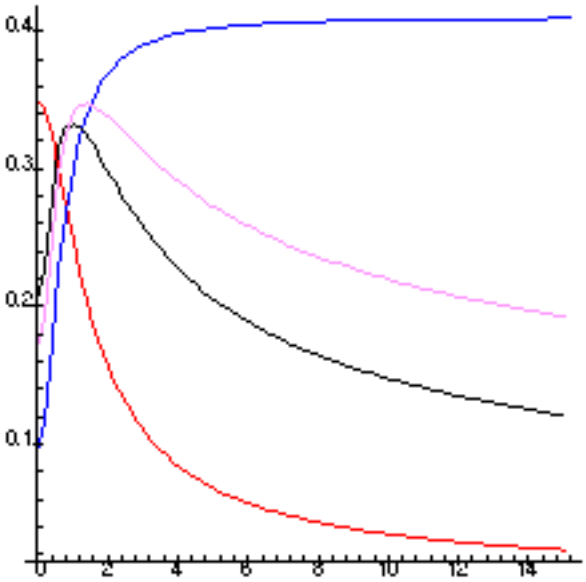

Figure 4.2a. The functions $\alpha_{aver}(x)$ for the Π_3, $\Pi_3\Pi$, $\Pi_3\Pi^2$ and Π processes.

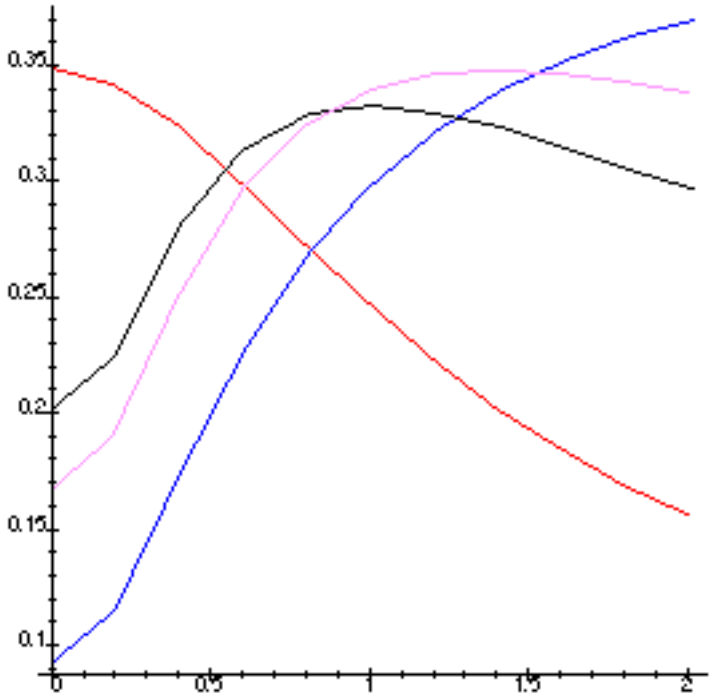

Figure 4.2b. Details near the origin of $\alpha_{aver}(x)$ for the Π_3, $\Pi_3\Pi$, $\Pi_3\Pi^2$ and Π processes.

As to the overall convergence, we compare the averaged amplification per iteration in the region $\mathbb{C}(y_3)$ defined in (3.14), that is, the averaged values of the functions $|C_3|$, $|C|$, $|C_3C|^{1/2}$ and $|C_3C^2|^{1/3}$ in $\mathbb{C}(y_3)$, with C and C_3 defined by (3.13) and (4.1). Writing $x := \kappa_j y_3$ and denoting these averaged values by $\alpha_{aver}(x)$, we compute $\alpha_{aver}(x)$ in the square $\{|\kappa_j y_1| \le \gamma, |\kappa_j y_2| \le \gamma\}$, where $\gamma \approx 0.50$, 0.65, 0.72, 0.75 in the Π_3, Π, $\Pi_3\Pi$ and $\Pi_3\Pi^2$ processes, respectively. In Figure 4.2a, the function

$\alpha_{aver}(x)$ is plotted for $0 \le x \le 15$. Details in the interval $0 \le x \le 2$ are given by Figure 4.2b. In these plots, the left and right end point values of the graphs respectively decrease and increase for the Π_3, $\Pi_3\Pi$, $\Pi_3\Pi^2$ and Π processes. Thus, the Π_3 and Π processes clearly are extreme cases, the Π_3 process being superior for larger values of x, the Π process for small values of x. The $\Pi_3\Pi$ and $\Pi_3\Pi^2$ processes are good compromises. Moreover, they possess substantially larger convergence boundaries respectively allowing about 45% and 10% larger stepsizes than the Π_3 and Π processes.

4.2. Partially explicit treatment of the horizontal terms

Instead of inserting iterations that are fully explicit with respect to the horizontal terms, one may consider the use of iterations with iteration matrices $\Pi_{13} := (I - D\otimes\Delta t J_1)(I - D\otimes\Delta t J_3)$ and $\Pi_{23} := (I - D\otimes\Delta t J_2)(I - D\otimes\Delta t J_3)$. Since in shallow water applications, the two horizontal directions introduce a comparable degree of stiffness, we shall alternate these two iteration matrices. It should be remarked that we already used the iteration matrices Π_{13} and Π_{23} in [6] in another type of iteration methods for use in shallow water computations.

4.2.1. Convergence boundary of the $\Pi_{13}\Pi_{23}$ process. After each two iterations, the corresponding amplification matrix Z_r in the linearized error recursion has diagonal blocks of the form

$$\big(I - ((\mathbf{e}_j^T\otimes I)\Pi_{13})^{-1}(I - \kappa_j\Delta t J)\big)\big(I - ((\mathbf{e}_j^T\otimes I)\Pi_{23})^{-1}(I - \kappa_j\Delta t J)\big), \quad j = 1, 2, \dots, s,$$

irrespective the value of r. Defining the functions

$$C_{k3}(x_1, x_2, x_3) := 1 - \frac{1 - x_1 - x_2 - x_3}{(1 - x_k)(1 - x_3)}, \quad k = 1, 2, \tag{4.6}$$

the amplification factors for two iterations are given by

$$\alpha_j^2 = C_{13}(\kappa_j z_1, \kappa_j z_2, \kappa_j z_3)\, C_{23}(\kappa_j z_1, \kappa_j z_2, \kappa_j z_3). \tag{4.7}$$

Restricting z_k to imaginary values iy_k, we obtain

$$|\alpha_j^2| = \left(\frac{(\kappa_j^2 y_1^2 + \kappa_j^4 y_3^2 y_2^2)(\kappa_j^2 y_2^2 + \kappa_j^4 y_3^2 y_1^2)}{(1 + \kappa_j^2 y_1^2)(1 + \kappa_j^2 y_2^2)(1 + \kappa_j^2 y_3^2)^2}\right)^{1/2}. \tag{4.8}$$

We verified that for given values of $\kappa_j^2 y_3^2 \ge 0.01$ the amplification factor $|\alpha_j^2|$ increases most rapidly along the lines $y_1 = 0$ and $y_2 = 0$. For $\kappa_j^2 y_3^2 < 0.01$, we found numerically that $\gamma(y_3) \ge 10$. Hence, proceeding as in the proof of Theorem 3.1, it can be shown that for $\kappa_j |y_3| \ge 0.1$ the convergence boundary $\gamma(y_3)$ in (3.14) is given by (3.16) where the function g(x) is defined by

$$x^2 g^4 - (x^2 + 1)^2 g^2 - (x^2 + 1)^2 = 0, \quad x \ge 0.1. \tag{4.9}$$

A comparison of the plot of this function in Figure 4.3 with Figure 4.1 shows that the convergence boundary $\gamma(y_3)$ is considerably larger than that of the $\Pi_3\Pi$ process. This also results in a much larger value for γ. To see this, we again write the implicit equation for g as $F(g,x) = 0$ and solve γ from the

equations $F(\gamma,x) = 0$ and $\partial F(\gamma,x)/\partial x = 0$ to obtain $\gamma = \sqrt{2 + 2\sqrt{2}} \approx 2.19$ which compares favourably with the value $\gamma \approx 0.65$ of Theorem 3.2.

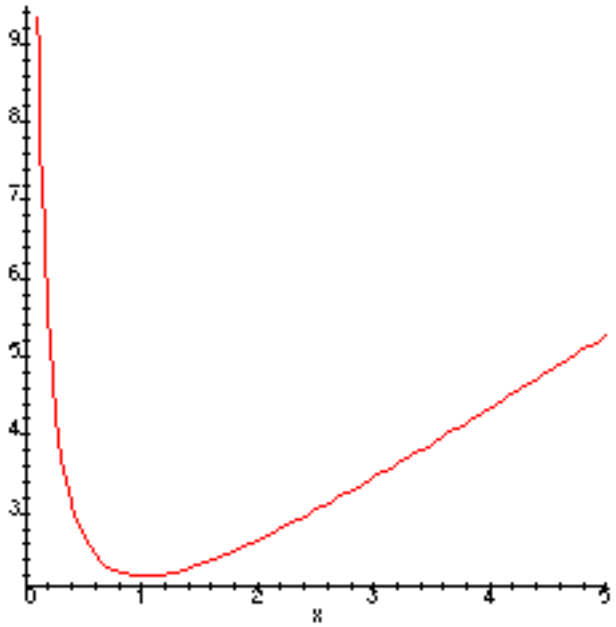

Figure 4.3. The function g(x) defined by (4.9).

4.2.2. The rate of convergence. As to the behaviour of the amplification factors for small values of $\Delta t\, \lambda(J_k)$, it is immediate from (4.6) and (4.7) that $\alpha_j^m = O((\Delta t)^m)$. Thus, the $\Pi_{13}\Pi_{23}$ process damps the nonstiff error components with the same order in Δt as Π_3, but less strongly than Π, $\Pi_3\Pi$ and $\Pi_3\Pi^2$. Next, we look for a given value of $x := \kappa_j y_3$ at the averaged value $\alpha_{aver}(x)$ of $|C_{13}C_{23}|^{1/2}$ in the square $\{|\kappa_j y_1| \le \gamma, |\kappa_j y_2| \le \gamma\}$ with $\gamma = 2.19$, and we compare this function with that obtained for the $\Pi_3\Pi$ with $\gamma = 0.72$. The lowest and highest graph in the Figures 4.4a and 4.4b presents $\Pi_3\Pi$ and $\Pi_{13}\Pi_{23}$, respectively. Hence, we may conclude that the three times larger convergence boundary of the process $\Pi_{13}\Pi_{23}$ is paid by a considerably worse convergence rate. Nevertheless, the question remains which iteration method is most efficient. To answer this question, we may either count three $\Pi_{13}\Pi_{23}$ iterations for one iteration, that is, we define $\alpha_{aver}(x)$ by means of $|C_{13}C_{23}|^{3/2}$, or we apply $\Pi_{13}\Pi_{23}$ with the same convergence boundary as used in the $\Pi_3\Pi$ process, i.e. $\gamma = 0.72$. The middle graphs in the Figures 4.4a and 4.4b present these cases, where the lowest middle one corresponds with the $\gamma = 0.72$ approach. Thus, our final conclusion is that on the basis of the $\alpha_{aver}(x)$ profiles and the damping of the nonstiff error components, the $\Pi_3\Pi$ is the most efficient iteration strategy.

5. Conclusions

In this paper we analysed the convergence of iterative solution methods for the linear systems arising in Rosenbrock integration methods and the nonlinear systems arising in DIRK integration methods. In

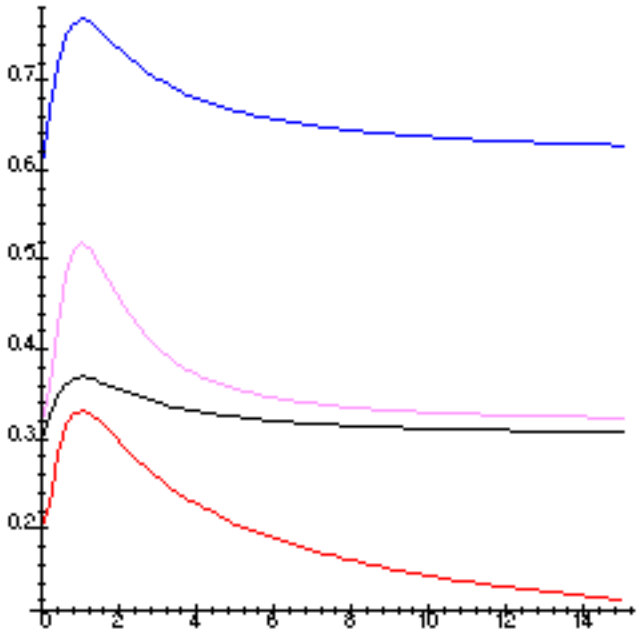

Figure 4.4a. The functions $\alpha_{aver}(x)$ for the $\Pi_3\Pi$ and $\Pi_{13}\Pi_{23}$ processes.

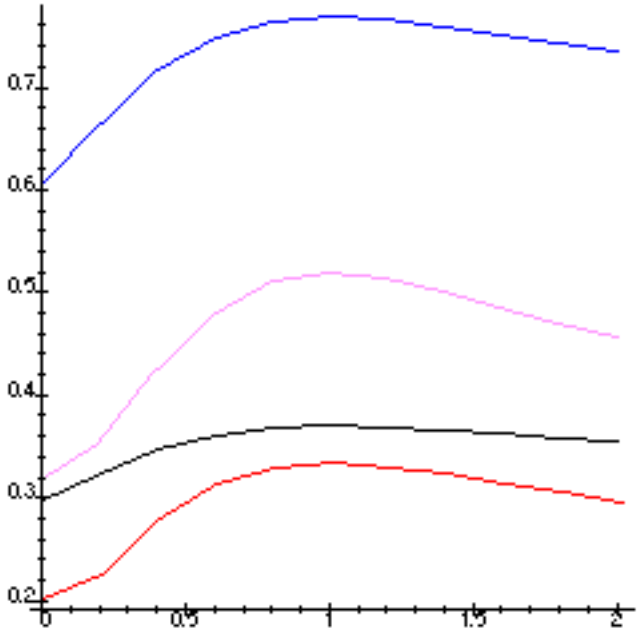

Figure 4.4b. Details near the origin of $\alpha_{aver}(x)$ for the $\Pi_3\Pi$ and $\Pi_{13}\Pi_{23}$ processes.

particular, we focused on the integration of ODEs originating from shallow water applications, where the ODE system contains a highly stiff part corresponding with the vertical derivative terms and a moderately stiff part corresponding with the horizontal derivative terms. We considered iteration methods based on the approximate factorizations of the system matrix associated with the application of modified Newton. The resulting convergence boundaries γ are given by

Process	Π	Π_3	$\Pi_3\Pi$	$\Pi_3\Pi^2$	$\Pi_{13}\Pi_{23}$
$\gamma \approx$	0.65	0.5	0.72	0.75	2.19

The best convergence characteristics were obtained for the $\Pi_3\Pi$ process where alternatingly only the vertical direction and all coordinate directions are treated implicitly.

References

[1] Brunner, H. & Houwen, P.J. van der [1986]: The numerical solution of Volterra equations, North-Holland.

[2] Dekker, K. & Verwer, J.G. [1984]: Stability of Runge-Kutta methods for stiff nonlinear differential equations, North-Holland.

[3] Eichler-Liebenow, C., Houwen, P.J. van der & Sommeijer, B.P. [1998]: Analysis of approximate factorization in iteration methods, Appl. Numer. Math. 28, 245-258.

[4] Hairer, E. & Wanner, G. [1991]: Solving ordinary differential equations II. Stiff and differential-algebraic problems, Springer.

[5] Houwen, P.J. van der, Sommeijer, B.P. & Kok, J. [1997]: The iterative solution of fully implicit discretizations of three-dimensional transport models, Appl. Numer. Math. 25, 243-256.

[6] Houwen, P.J. van der & Sommeijer, B.P. [1998]: Approximate factorization in shallow water applications, Report MAS R9835, CWI, Amsterdam, submitted for publication.

[7] Peaceman, D.W. & Rachford, H.H. Jnr. [1955]: The numerical solution of parabolic and elliptic differential equations, J. Soc. Indust. App. Math. 3, 28-41.

[8] Rosenbrock, H.H. [1962-1963]: Some general implicit processes for the numerical solution of differential equations, Computer J. 5, 329-330

[9] Sandu, A. [1997]: Numerical aspects of air quality modelling, PhD thesis, University of Iowa.

[10] Sommeijer, B.P. [1999]: The iterative solution of fully implicit discretizations of three-dimensional transport models, in: Parallel Computational Fluid Dynamics - Implementation and Results Using Parallel Technology (eds. C.A. Lin, A. Ecer, J. Periaux, N. Satofuka), Proceedings of the 10th Int. Conf. on Parallel CFD, May 1998, Hsinchu, Taiwan, Elsevier.

[11] Steihaug, T. & Wolfbrandt, A. [1979]: An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations, Math. Comp. 33, 521-534.

[12] Verwer, J.G., Spee, E.J., Blom, J.G. & Hundsdorfer, W. [1999]: A second-order Rosenbrock method applied to photochemical dispersion problems, to appear in SIAM J. Sci. Comput.

[13] Verwer, J.G., Hundsdorfer, W. & Blom, J.G. [1998]: Numerical time integration for air pollution models, Report MAS R9825, CWI, Amsterdam, submitted for publication.