stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

NW 41/77 APRIL

K. DEKKER

# GENERALIZED RUNGE-KUTTA METHODS FOR COUPLED SYSTEMS OF HYPERBOLIC DIFFERENTIAL EQUATIONS

Preprint

2e boerhaavestraat 49 amsterdam

*Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.*

*The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).*

AMS(MOS) subject classification scheme (1970): 65L05, M10, M20

CONTENTS


1. Introduction. 1

2. The generalized Runge-Kutta method. 2

3. Conditions for the parameter matrices. 5
3.1. Consistency conditions. 5
3.2. Storage requirements. 7
3.3. Stability requirements. 9

4. Generalized second order formulas for a restricted class of equations. 16
4.1. Stabilized second order formulas. 17
4.2. Almost second order formulas using two arrays of storage. 19
4.3. Second order formulas using three arrays of storage. 22
4.4. Strongly stable formulas. 25

REFERENCES 28

# Generalized Runge-Kutta methods for coupled systems of hyperbolic differential equations. *)

by

K. Dekker

ABSTRACT

Runge-Kutta formulas are discussed for the integration of systems of differential equations. The parameters of these formulas are square matrices with component-dependent values. The systems considered are supposed to originate from hyperbolic partial differential equations, which are coupled in a special way. In this paper the discussion is concentrated on methods for a class of two coupled systems. For these systems first and second order formulas are presented, whose parameters are diagonal matrices. These formulas are further characterized by their low storage requirements, by a reduction of the computational effort per timestep, and by their relatively large stability interval along the imaginary axis. The new methods are compared with stabilized Runge-Kutta methods having scalar-valued parameters. It turns out that a gain factor of 2 can be obtained.

KEY WORDS & PHRASES: *Runge-Kutta formulas, ordinary differential equations, hyperbolic partial differential equations, extended stability region.*

*) This report will be submitted for publication elsewhere

# 1. INTRODUCTION

Runge-Kutta methods for second order differential equations with prescribed initial values are well known in literature (e.g. ZONNEVELD [8], FEHLBERG [1]). When the first derivative does not occur in the second equation, these special methods are more efficient than comparable methods for first order equations; for example, they may attain a higher *order of accuracy* with the same amount of derivative evaluations, or may possess a larger *stability region* (VAN DER HOUWEN [5]). When the second order equation is transformed into a system of two first order equations, these special methods may be considered as *generalized Runge-Kutta methods* whose parameters are square matrices. Evidently, these generalized methods derive their usefulness from taking into account the special structure of the Jacobian matrix of the resulting equations.

In this paper we will start to investigate generalized Runge-Kutta methods, which are not restricted to systems resulting from second order equations, but which apply to systems of the type

$$\frac{d\vec{y}_i}{dx} = \vec{f}_i(\vec{y}_1,\ldots,\vec{y}_k), \qquad i = 1,\ldots,k, \tag{1.1}$$

$\vec{y}_i$ , $i = 1,\ldots,k$, being prescribed at $x = x_0$. We observe that each component of this system in itself is a vector of a certain length, which is not necessarily the same for each component. Systems of this type may arise by applying the method of lines to a coupled system of hyperbolic or parabolic partial differential equations. When the Jacobian matrix of (1.1), given by

$$J_{ij} = \frac{\partial \vec{f}_i}{\partial \vec{y}_j}, \qquad i = 1,\ldots,k, \qquad j = 1,\ldots,k, \tag{1.2}$$

is sparse, it is likely that generalized Runge-Kutta methods are more efficient than ordinary Runge-Kutta methods.

In the next sections we will describe the generalized Runge-Kutta method, and derive conditions for consistency (up to order 2) and for low storage requirements. The stability analysis is performed by imposing con-

ditions on the Jacobian matrix, which are fulfilled for a wide class of hyperbolic systems. This particular choice is motivated by the fact that we want to investigate generalized Runge-Kutta methods for the *two-dimensional shallow water equation* (KREIS [6]) in a forthcoming paper.

In section 4 we restrict ourselves to problems consisting of two coupled systems (k=2 in 1.1), of which $\vec{f}_2$ does not depend on $\vec{y}_2$. Second order, m-point formulas using two or three arrays of storage are constructed. In the latter case the resulting stabibity condition reads

$$(1.3) \qquad h_n \leq \frac{m-1}{\sigma(J)}, \qquad m \text{ odd}.$$

Here $\sigma(J)$ denotes the spectral radius of the Jacobian matrix J. The number of derivative evaluations per time step for these formulas, however, is less than m, viz. $\frac{m+2}{2}$, so that we effectively have a stability limit of $2(m-1)/(m+2)$. Thus, asymptotically a factor 2 is gained over ordinary stabilized second order Runge-Kutta methods, which have an effective stability limit of $(m-1)/m$ (VAN DER HOUWEN [3]).

In the near future numerical results will be reported obtained by the new formulas, applied to the wave equation and the equation of the flow in a narrow canal. Also, we intend to construct generalized methods for problems consisting of three coupled systems.

## 2. THE GENERALIZED RUNGE-KUTTA METHOD

Consider the system of differential equations (1.1). In order to simplify the notation we introduce the variables

$$(2.1) \qquad y = (\vec{y}_1,\ldots,\vec{y}_k)^T \text{ and } f(y) = (\vec{f}_1(y),\ldots,\vec{f}_k(y))^T.$$

The generalized m-point Runge-Kutta method is defined by

$$
\begin{aligned}
& y_{n+1}^{(0)} = y_n, \\
(2.2) \qquad & y_{n+1}^{(j)} = M_j y_n + h_n \sum_{l=0}^{j-1} N_{jl}\, f(y_{n+1}^{(l)}), \qquad j = 1,\ldots,m, \\
& y_{n+1} = y_{n+1}^{(m)}.
\end{aligned}
$$

Here, $y_{n+1}$ denotes the numerical approximation to the solution y at the point $x = x_n + h_n$; The quantities $M_j$ and $N_{jl}$ stand for k x k matrices, whose entries are matrices, too, the size depending on the dimensions of $\vec{y}_1,\ldots,\vec{y}_k$.

EXAMPLE 2.1. Consider the method for second order differential equations

$$
(2.3) \qquad \frac{d^2\vec{w}}{dx^2} = \vec{g}(\vec{w}, \frac{d\vec{w}}{dx}),
$$

described by ZONNEVELD [1964]:

$$
\begin{aligned}
& \vec{z}_n = \vec{w}_n', \\
& \vec{w}_{n+1}^{(1)} = \vec{w}_n + h_n \vec{z}_n, \qquad \vec{z}_{n+1}^{(1)} = \vec{z}_n + h_n \vec{g}(\vec{w}_n, \vec{z}_n), \\
& \vec{w}_{n+1} = \vec{w}_n + h_n \vec{z}_n + \tfrac{1}{2} h_n^2 \vec{g}(\vec{w}_n, \vec{z}_n), \\
& \vec{w}_{n+1}' = \vec{z}_n + \tfrac{1}{2} h_n \{\vec{g}(\vec{w}_n, \vec{z}_n) + \vec{g}(\vec{w}_{n+1}^{(1)}, \vec{z}_{n+1}^{(1)})\}.
\end{aligned}
$$

When we define $y = (\vec{w},\vec{z})^T$ and $f = (\vec{z},\vec{g})^T$, this method can be represented by scheme (2.2) where m=2 and

$$
\begin{aligned}
& M_1 = M_2 = \begin{pmatrix} I & h_n I \\ 0 & I \end{pmatrix}, \qquad N_{10} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \\
(2.4a) \qquad & N_{20} = \begin{pmatrix} 0 & \tfrac{1}{2} h_n I \\ 0 & \tfrac{1}{2} I \end{pmatrix}, \qquad N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & \tfrac{1}{2} I \end{pmatrix}.
\end{aligned}
$$

In these expressions I denotes the unity matrix of appropriate order. The occurrence of this matrix in an off-diagonal position is allowed, because

the vectors $\vec{w}$ and $\vec{z}$ $(=\vec{w}')$ have the same number of components.

Note, however, that the representation (2.4a) is not unique. Another choice for the parameter matrices reads

$$(2.4b)\qquad M_1 = M_2 = N_{10} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \qquad N_{20} = N_{21} = \begin{pmatrix} \frac{1}{2}I & 0 \\ 0 & \frac{1}{2}I \end{pmatrix},$$

so that the method reduces to an ordinary Runge-Kutta method for a system of equations.

A less trivial example is given in VAN DER HOUWEN [5]:

EXAMPLE 2.2.: Consider the method for second order differential equations without first derivatives, described by

$$(2.5)\qquad \begin{aligned} \vec{w}_{n+1} &= \vec{w}_n + h_n\vec{w}'_n + \tfrac{1}{2}h_n^2\vec{g}(\vec{w}_n + \tfrac{1}{2}h_n\vec{w}'_n + \tfrac{1}{16}h_n^2\vec{g}(\vec{w}_n + \tfrac{1}{2}h_n\vec{w}'_n)), \\ \vec{w}'_{n+1} &= \vec{w}'_n + h_n\vec{g}(\vec{w}_n + \tfrac{1}{2}h_n\vec{w}'_n + \tfrac{1}{16}h_n^2\vec{g}(\vec{w}_n + \tfrac{1}{2}h_n\vec{w}'_n)). \end{aligned}$$

Using the same conventions as in the previous example, we can represent this method by m=3 and

$$(2.5a)\qquad \begin{aligned} &M_1 = M_2 = \begin{pmatrix} I & \frac{1}{2}h_n I \\ 0 & I \end{pmatrix}, \qquad M_3 = \begin{pmatrix} I & h_n I \\ 0 & I \end{pmatrix}, \\ &N_{21} = \begin{pmatrix} 0 & \frac{1}{16}h_n I \\ 0 & 0 \end{pmatrix}, \qquad N_{32} = \begin{pmatrix} 0 & \frac{1}{2}h_n I \\ 0 & I \end{pmatrix}, \\ &N_{10} = N_{20} = N_{30} = N_{31} = 0, \end{aligned}$$

or, alternatively by m=5 and

$$(2.5b)\qquad \begin{aligned} &M_1 = M_2 = M_3 = M_4 = M_5 = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \qquad N_{10} = N_{32} = \begin{pmatrix} \frac{1}{2}I & 0 \\ 0 & 0 \end{pmatrix}, \\ &N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{8}I \end{pmatrix}, \qquad N_{43} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2}I \end{pmatrix}, \qquad N_{53} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \\ &N_{54} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad N_{20} = N_{30} = N_{31} = N_{40} = N_{41} = N_{42} = N_{50} = N_{51} = N_{52} = 0. \end{aligned}$$

We remark that, although (2.5b) defines a five-point formula, it evidently requires only two evaluations of the second derivative g, namely those with $y_{n+1}^{(1)}$ and $y_{n+1}^{(3)}$ as arguments.

## 3. CONDITIONS FOR THE PARAMETER MATRICES

In this section we will derive the conditions which should be imposed on the matrices $M_j$ and $N_{j1}$ in order to ensure second order consistency, minimal storage requirements and stability. Whereas we intend to derive schemes which are applicable to a restricted class of equations of type (1.1), we will formulate the conditions in terms of the variables $y_n$, $f_n(=f(y_n))$ and $J_n$, the Jacobian matrix at $(x_n,y_n)$. In general, the conditions are very complicated; therefore we will simplify them by imposing the following restrictions on the parameter matrices:

(3.1) The matrices $M_j$ and $N_{j1}$, $j = 1,\ldots,m$, $1 = 0,\ldots,j-1$, do not depend on the Jacobian $J_n$.

(3.2) The matrices $M_j$ satisfy the relation $M_j = I + 0(h_n)$, and the matrices $N_{j1}$ satisfy $h_n \times N_{j1} = 0(h_n)$, $j = 1,\ldots,m$, $1 = 0,\ldots,j-1$.

The examples of section 2 show that (3.2) need not be a severe restriction, whereas generalized RK-schemes whose parameter matrices depend on the Jacobian have already been analysed by several authors.

### 3.1. Consistency conditions.

The order equations for scheme (2.2) can be derived in the usual way (see e.g. ZONNEVELD [8]) by expanding $y_{n+1}$ and the analytic solution of (1.1) through the point $y(x_n) = y_n$ in a Taylor-series in $h_n$, and equating the corresponding terms.

The conditions for orders p up to 2 are listed in table 3.1.
It shouls be remarked that table 3.1 presents for p=2 "additional" conditions, i.e. the conditions for second order consistency are the conditions listed for both p=1 and p=2.

Table 3.1. Consistency conditions for scheme (2.2), applied to equation (1.1)

| | |
|---|---|
| p=1 | $M_m^{(1)} y_n + \sum_{l=0}^{m-1} N_{ml}^{(0)} f_n = f_n$ |
| p=2 | $\frac{1}{2} M_m^{(2)} y_n + \sum_{l=0}^{m-1} N_{ml}^{(1)} f_n + \sum_{l=0}^{m-1} N_{ml}^{(0)} J_n M_l^{(1)} y_n + \sum_{l=0}^{m-1} N_{ml}^{(0)} J_n \sum_{r=0}^{l-1} N_{lr}^{(0)} f_n = \frac{1}{2} J_n f_n,$ <br> where $M_l^{(i)} = \{\frac{d^i}{dh^i} M_l(h_n)\}_{h_n = 0}, \quad i = 0,1,2, \quad l = 1,\ldots,m,$ <br> and $N_{ml}^{(i)} = \{\frac{d^i}{dh^i} N_{ml}(h_n)\}_{h_n = 0}, \quad i = 0,1, \quad l = 0,\ldots,m-1.$ |

The conditions given in table 3.1 determine the consistency of scheme (2.2) for a particular differential equation at a specific point. Requiring that (2.2) is consistent for all problems of type (1.1) yields the conditions listed in table 3.2; these conditions can easily be derived from table 3.1 by suitable subsitutions for $y_n$, $f_n$ and $J_n$.

Table 3.2. Consistency conditions for scheme (2.2)

| | |
|---|---|
| p=1 | $M_m^{(1)} = 0,$ <br> $\sum_{l=0}^{m-1} N_{ml}^{(0)} = I.$ |
| p=2 | $M_m^{(2)} = 0,$ <br> $\sum_{l=0}^{m-1} N_{ml}^{(1)} = 0,$ <br> $\sum_{l=0}^{m-1} N_{ml}^{(0)}(p,q)\, M_l^{(1)}(r,s) = 0, \quad p, q, r, s \in [1,\ldots,k]$ <br> $\sum_{l=0}^{m-1} N_{ml}^{(0)}(p,q) \sum_{n=0}^{l-1} N_{ln}^{(0)}(r,s) = \frac{1}{2}\delta_{pq}\delta_{rs}, \quad p, q, r, s \in [1,\ldots,k]$ <br> Here, $N_{ml(p,q)}$ denotes the element in row p and column q of the matrix $N_{ml}$, whereas $\delta$ stand for the Kronecker function. |

EXAMPLE 3.1. The scheme determined by (2.5a) does not satisfy the conditions for p=1 given in table 3.2. As a consequence, scheme (2.5a) is generally not consistent of order one, when it is applied to an arbitrary system. However, scheme (2.5a) satisfies the conditions given in table 3.1, if the following equalities hold:

$$\begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix} y_n = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} f_n \qquad \text{and}$$

$$\begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix} f_n + \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} J_n \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} f_n = J_n f_n .$$

These equations are evidently satisfied when we substitute $y_n = (\vec{w}_n, \vec{w}'_n)$, $f_n = (\vec{w}'_n, \vec{g}(\vec{w}_n))$ and $J_n = \begin{pmatrix} 0 & I \\ \frac{dg}{dwn} & 0 \end{pmatrix}$, so (2.5a) is consistent of order 2 for second-order differential equations without first derivatives.

In a similar way one easily verifies that scheme (2.5b) is only consistent of order two is the equality

$$\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} J_n \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} J_n \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} = J_n$$

holds, which again is the case when second order equations without first derivatives are considered.

## 3.2. Storage Requirements

As we intend to apply the schemes to large systems originating from partial differential equations, attention should be paid to the storage requirements. We will derive here the conditions for schemes requiring two and three arrays of storage (confer VAN DER HOUWEN [2]).

*Schemes requiring two arrays of storage*

The flow of computation in schemes which require only two arrays of storage might be represented by the flow chart

$$\begin{array}{ccc} y_{n+1}^{(j-1)} & \longrightarrow & y_{n+1}^{(j)} \\ \downarrow & \nearrow & \\ f(y_{n+1}^{(j-1)}) & & \end{array}$$

or by the formula

$$(3.3) \qquad y_{n+1}^{(j)} = D_j\{y_{n+1}^{(j-1)} + h_n E_j f(y_{n+1}^{(j-1)})\}, \qquad j = 1,\ldots,m,$$

where $D_j$ denotes a general matrix and $E_j$ a sufficiently simple matrix in order to compute the product without auxilary storage.

Comparing (3.3) with scheme (2.2), we obtain the following relations for $M_j$ and $N_{j1}$:

$$(3.4) \qquad \begin{aligned} M_j &= \prod_{l=1}^{j} D_l, \qquad j = 1,\ldots,m, \\ N_{j1} &= \prod_{r=l+1}^{j} D_r E_{l+1}, \qquad j = 1,\ldots,m, \qquad l = 0,\ldots,j-1, \end{aligned}$$

Elimination of $D_j$ and $E_j$ yields

$$(3.5) \qquad N_{j1} = M_j M_{l+1}^{-1} N_{l+1,1}, \qquad j = 1,\ldots,m, \qquad l = 0,\ldots,j-1.$$

We note that $M_{l+1}^{-1}$ exists for sufficiently small $h_n$, in view of relation (3.2).

*Schemes requiring three arrays of storage*

Introducing an additional set of vectors $z_{n+1}$, we can construct schemes of type (2.2) by means of recurrence relations of the type

$$(3.6) \qquad \begin{aligned} y_{n+1}^{(0)} &= y_n, \qquad z_{n+1}^{(0)} = 0, \\ y_{n+1}^{(j)} &= A_j y_{n+1}^{(j-1)} + h_n E_j f(y_{n+1}^{(j-1)}) + B_j z_{n+1}^{(j-1)}, \\ z_{n+1}^{(j)} &= C_j y_{n+1}^{(j-1)} + h_n F_j f(y_{n+1}^{(j-1)}) + D_j z_{n+1}^{(j-1)}, \qquad j = 1,\ldots,m, \\ y_{n+1} &= y_{n+1}^{(m)} \end{aligned}$$

Assuming $A_j$ non-singular (this is implied by condition (3.2)), we may set $C_j = 0$ without loss of generality. In fact, given a recurrence relation (3.6), it is easy to contruct an equivalent relation (yielding the same $y_{n+1}$) with $C_j = 0$. Comparing (3.6) with scheme (2.2), we obtain the following relations for $M_j$ and $N_{jl}$:

$$
\begin{aligned}
&M_j = \prod_{l=1}^{j} A_l, \\
&N_{jl} = M_j M_{l+1}^{-1} N_{l+1,1} + \sum_{r=l+2}^{j} M_j M_r^{-1} B_r \prod_{s=l+2}^{r-1} D_s F_{l+1}, \qquad l = 0,\ldots,j-2 \\
&N_{jj-1} = E_j, \qquad j = 1,\ldots,m.
\end{aligned} \tag{3.7}
$$

We remark that, in general, the matrices $D_j$ cannot be eliminated from this formula, as they might be singular. However, it is easily verified that for a suitable transformation of $z_{n+1}^{(j)}$ the matrices $D_j$ will be of the form

$$
\begin{pmatrix} I & X \\ 0 & 0 \end{pmatrix}
$$

### 3.3 Stability requirements

To analyse the stability of scheme (2.2), we study the effect of a perturbation $\Delta y_n$ of $y_n$ on the resulting vector $y_{n+1}$. Let $J_n$ denote the Jacobian matrix of the right hand side $f(y_n)$; then this perturbation is approximately given by

$$
\Delta y_{n+1} = M_j \Delta y_n + \sum_{l=0}^{j-1} h_n N_{jl} J_n \Delta y_n^{(l)}, \qquad j = 1,\ldots,m, \tag{3.8}
$$

or alternatively by

$$
\begin{aligned}
&\Delta y_{n+1}^{(j)} = R_j(h_n J_n)\Delta y_n, \qquad j = 1,\ldots,m, \\
&R_j(h_n J_n) = M_j + \sum_{l=0}^{j-1} h_n N_{jl} J_n R_l(h_n J_n).
\end{aligned} \tag{3.8a}
$$

We will call method (2.2) stable if all the eigenvalues of $R_m(h_nJ_n)$ are within the unit circle; when one or more eigenvalues are on the unit circle, the method will be called weakly stable. Integrating problems with a constant Jacobian with a stable method, the effect of a perturbation $\Delta y_n$ will ultimately be damped out, as the k-th power of the amplification matrix $R_m(h_nJ)$ will tend to the zero matrix as k tends to infinity. Using a weakly stable method, the effect of $\Delta y_n$ will grow less than exponentially, the rate of growth depending on the number of coinciding eigenvalues on the unit circle.

Next, we consider a finite interval of integration and let $h_n$ tend to zero. Then, the eigenvalues of the matrix $R_m(h_nJ)$ with multiplicity greater than one may tend to one, as is illustrated in the following example.

EXAMPLE 3.2. Consider the second order method generated by (m=2):

$$(3.9)\qquad M_1 = M_2 = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}, \quad N_{10} = N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad N_{20} = \begin{pmatrix} 0 & \frac{1}{2}h \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Application of this method to the differential equations

$$(3.10)\qquad \frac{dy}{dx} = z, \ \frac{dz}{dx} = -4y - 4z, \ y(0) = y_0, \ z(0) = z_0,$$

yields the recurrence relation

$$(3.11)\qquad \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1-2h+h^2 & 5h-10h^2 \\ 0 & 1-2h+h^2 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} y_n \\ z_n \end{pmatrix} = \tilde{S}(h) \begin{pmatrix} y_n \\ z_n \end{pmatrix}.$$

Although $\tilde{S}(h)$ has a multiple eigenvalue, $\|\tilde{S}(h)^n\|$ is bounded by the constant $1 + 5hn$, as the off-diagonal elements of its Jordan-normal form are of order h. This suggests that we should consider amplification matrices, whose Jordan form have off-diagonal elements of order h. In the following lemma, we will show that scheme (2.2) has this property, provided that (3.1) and (3.2) are satisfied.

LEMMA 3.1. *The amplification matrix belonging to a generalized Runge-Kutta method which satisfies the conditions* (3.1) *and* (3.2), *has a Jordan normal form with elements of order h in off-diagonal position.*

Proof. From the definition of the amplification matrix $R_m(h_n J_n)$ in (3.8a) and the conditions (3.1) and (3.2) it is obvious that there exists a matrix A, such that

$$R_m(h_n J_n) = I + A \text{ and } \|A\|_2 = O(h_n).$$

Now, let B be the Jordan normal form of A, given by the unitary transformation $B=T\,A\,T^{-1}$. From $|B_{ij}| \le \|Be_j\|_2 \le \|B\|_2 = \|A\|_2 = O(h_n)$ follows that all elements of B are of order h. As the Jordan normal form of $R_m(h_n J_n)$ is given by I + B, it is clear that all the off-diagonal elements of this Jordan form are of order h. □

COROLLARY. *The global discretization error of a generalized Runge-Kutta method for* $h \to 0$ *increases at most linearly with the number of steps, if the conditions* (3.1) *and* (3.2) *are satisfied, and the amplification matrix has eigenvalues on or within the unit circle.*

The above corollary suggests us to verify the stability of a generalized Runge-Kutta method for a given problem by proving that the eigenvalues of the amplification matrix are in modulus less or equal to one. However, this task is not as simple as in the case of ordinary Runge-Kutta methods, as was already observed by VAN DER HOUWEN [5]. The reason for this is that we cannot reduce a system of equations to a set of single equations, which are more easily analysed, because the eigenvectors of the Jacobian matrix differ in general from the eigenvectors of the amplification matrix. This behaviour may be illustrated in the following example:

EXAMPLE 3.3. Consider the generalized Runge-Kutta method defined by m=2 and

$$(3.12) \qquad M_1 = M_2 = I,\ N_{10} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},\quad N_{20} = 0,\quad N_{21} = I.$$

Application of this method to the *model problem for hyperbolic equations*

$$\frac{dy}{dx} = - cz, \qquad \frac{dz}{dx} = cy, \qquad y(0) = y_0, \qquad z(0) = z_0,$$

yields the recurrence relation

$$(3.14) \qquad \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} 1-\frac{1}{2}h^2c^2 & -hc \\ hc & 1-h^2c^2 \end{pmatrix} \begin{pmatrix} y_n \\ z_n \end{pmatrix}, \qquad n = 0,\ldots .$$

The eigenvalues of the amplification matrix are

$$\lambda_{1,2} = 1 - \frac{3}{4}h^2c^2 \pm i\, hc\sqrt{1-\frac{1}{4}h^2c^2};$$

these eigenvalues are in modulus less than or equal to 1 if $0 \leq hc \leq 1$.

When we try to uncouple system (3.13) by introducing the eigenvectors of the Jacobian matrix, $u = (1-i,1-i)^T$, $v = (1+i,1-i)^T$, we can rewrite (3.13) and (3.14) as

$$(3.13a) \qquad \frac{du}{dx} = - icu, \quad \frac{dv}{dx} = + icv, \quad u(0) = u_0 = (1+i)y_0 + (1-i)z_0,$$
$$v(0) = v_0 = (1-i)y_0 + (1+i)z_0,$$

and

$$(3.14a) \qquad \begin{pmatrix} u_{n+1} \\ v_{n+1} \end{pmatrix} = \begin{pmatrix} 1-\frac{3}{4}h^2c^2+i\,hc & -\frac{1}{4}i\,h^2c^2 \\ \frac{1}{4}i\,h^2c^2 & 1-h^2c^2-i\,hc \end{pmatrix} \begin{pmatrix} u_n \\ v_n \end{pmatrix}, \qquad n = 0,\ldots .$$

Evidently, the amplification matrix of (3.14a) is not a simple diagonal matrix, with as elements polynomials in ihc, as one would obtain in the case of ordinary Runge-Kutta methods.

From this example we may conclude that the stability analysis of a generalized scheme in terms of the eigenvalues of the Jacobian matrix is in general impossible. In order to derive a priori stability properties of a generalized scheme, we will restrict ourselves to a class of differential equations which is characterized by the fact that the Jacobian matrix has pairs of purely imaginary eigenvalues. This particular choice is motivated

by the observation that this situation frequently occurs after discretization of partial differential equations of hyperbolic type. For equations of this type the following lemma may be applied.

Lemma 3.2. *Let* $R(hJ_n)$ *be the amplification matrix of a generalized scheme* (2.2) *applied to a system of equations with Jacobian matrix* $J_n$ *of order* $2N$. *Assume that the eigenvectors of* $J_n$ *can be split into pairs* $(u_k,v_k)$, *having eigenvalues* $\lambda_k$ *and* $\overline{\lambda}_k$, *such that*

$$(3.15)\qquad R(hJ_n)\,(u_k,v_k) = (u_k,v_k)\begin{pmatrix} P(h\lambda_k) & \overline{Q(h\lambda_k)} \\ Q(h\lambda_k) & \overline{P(h\lambda_k)} \end{pmatrix} = (u_k,v_k)\,A(h\lambda_k)\,,$$
$$k = 1,\ldots,N,$$

*where* $(u_k,v_k)$ *denotes the matrix consisting of the columns of* $u_k$ *and* $v_k$ *and* $P$ *and* $Q$ *are polynomials. Then all the eigenvalues of* $R(hJ_n)$ *are in modulus less than or equal to* 1, *if both eigenvalues of all matrices* $A_k$ *are in modulus less than or equal to* 1.

Proof. According to the assumption there exists a similarity transformation

$$TJ_nT^{-1} = \begin{pmatrix} \lambda_1 & & & & 0 \\ & \overline{\lambda}_1 & & & \\ & & \ddots & & \\ & & & \lambda_N & \\ 0 & & & & \overline{\lambda}_N \end{pmatrix},$$

where the columns of $T$ are formed by the eigenvectors $u_k$ and $v_k$, $k = 1,\ldots,N$. Using (3.15) we find

$$R(hJ_n)\,T = T\begin{pmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_N \end{pmatrix},$$

so that $T^{-1}R(hJ_n)\,T$ transforms $R(hJ_n)$ into a $(2\times2)$ block-diagonal matrix. Thus each eigenvalue of $R(hJ_n)$ corresponds to an eigenvalue of $A_k$, for some index $k$, which implies the assertion of the lemma. □

Now we can define the *stability region* S of a scheme for which (3.15) holds as the region in the complex plane of z values, for which the eigenvalues of A(z) are within the unit circle. In particular we will be interested in the *imaginary stability boundary* $\beta_{im}$ ; that is the largest positive number such that $0 \le z \le \beta_{im}$ implies $iz \in S$. The most simple way to find S and $\beta_{im}$ is the application of the Hurwitz criterion: *the roots of the equation*

$$\alpha^2 - S\alpha + P = 0$$

*lie within or on the unit circle when the coefficients S and P are real and satisfy the inequalities*

$$|S| \le P + 1, \qquad P \le 1.$$

Application to (3.15) yields the *stability conditions*

$$|P(z)|^2 - |Q(z)|^2 \le 1 \tag{3.16}$$

and

$$2 \operatorname{Re} P(z) \le |P(z)|^2 - |Q(z)|^2 + 1.$$

EXAMPLE 3.4. When we consider the method described in example 3.3, we find $P(z) = 1+\frac{3}{4}z^2+z$ and $Q(z) = -\frac{1}{4}iz^2$. Substitution into the inequalities (3.16) yield the conditions

$$\tfrac{1}{2}z^2 + \tfrac{1}{2}z \le 0$$

and

$$z^2 \le \tfrac{1}{2}z^4,$$

These conditions are satisfied for $z = ia, \quad a \le 1$, so we find $\beta_{im} = 1$.

In lemma 3.2 we did choose a special form for the matrices A in order to obtain as characteristic equation a polynomial with real coefficients, on which the Hurwitz criterion was applicable. We might have chosen for the elements of A four different polynomials in $h\lambda_k$, and then we might have applied the Schur criterion (see e.g. LAMBERT [7]) on the complex characteristic polynomial of $A(h\lambda_k)$, thus relaxing the conditions of the lemma. However, the formulation chosen is more simple, and seems to leave enough freedom in the choice of the parameter matrices.

We now consider the question under what conditions the amplification matrix $R(hJ_n)$ can be written in the form (3.15). Obviously, a sufficient condition is, that all matrices $M_j$ and $N_{j1}$, $j = 1,\ldots,m$, $1 = 0,\ldots,j-1$, are of the form

$$(3.17) \qquad T \begin{pmatrix} aI & \overline{bI} \\ bI & \overline{aI} \end{pmatrix} T^{-1},$$

(T the matrix of eigenvectors of $J_n$, as in lemma 3.2). Substitution of these matrices in (3.8a) will yield (3.15). Thus the stability conditions (3.16) can be applied to generalized schemes of which the matrices are generated by (3.17), and the stability problem is reduced to the problem of finding suitable polynomials P(z) and Q(z).

In the following section we will construct some pairs of polynomials, which are optimal in the sence that $\beta_{im}$ is maximized. Here, we remark that the resulting scheme may be of little value if the matrices generated by (3.17) are not very sparse. However, for a model problem the matrices (3.17) may turn out to be sparse, and we may hope that the thus constructed scheme has good stability properties for less trivial problems, too.

EXAMPLE 3.5. Assume that the Jacobian matrix $J_n$ of a problem has a matrix of normalized eigenvectors T, which consists of 4 blocks,

$$(3.18) \qquad T = \begin{pmatrix} U & U \\ V & -V \end{pmatrix} .$$

Then it is easily verified that the matrices $T \begin{pmatrix} aI & bI \\ bI & aI \end{pmatrix} T^{-1}$ are sparse. In fact, we obtain

$$\begin{pmatrix} U & U \\ V & -V \end{pmatrix} \begin{pmatrix} aI & bI \\ bI & aI \end{pmatrix} \begin{pmatrix} U^H & V^H \\ U^H & -V^H \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(a+b)I & 0 \\ 0 & \frac{1}{2}(a-b)I \end{pmatrix},$$

being a diagonal matrix; the number of non-zero elements may be further reduced by the choices a=b and a=-b.

Matrices of the form $\begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}$, with purely imaginary eigenvalues, have property (3.18), for we may choose U to be the eigensystem of the matrix BC, and V as $\Lambda^{-1}CU$, where $\Lambda$ is the matrix of eigenvalues of $\begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}$. The stability analysis of problems of this type may thus be performed by analyzing the stability of the model problem (3.13) with approriate values of c (equal to the modulus of the eigenvalues of the Jacobian matrix).

## 4. GENERALIZED SECOND ORDER FORMULAS FOR A RESTRICTED CLASS OF EQUATIONS

In this section we will construct m-point formulas, which are of second order for problems with a Jacobian matrix of the form $J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & 0 \end{pmatrix}$. We only considered formulas using at most two or three arrays of storage; amongst those we tried to optimize the *effective imaginary stability boundary*, the quotient of the imaginary stability boundary and the number of derivative evaluations (which need not be equal to m, as was shown in example 2.2). The optimization was done by choosing optimal pairs of polynomials P(z) and Q(z). Schemes corresponding to these polynomials are found by (3.17), setting

$$T = \begin{pmatrix} (1+i)I & (1-i)I \\ (1-i)I & (1+i)I \end{pmatrix},$$

the matrix of eigenvectors of problem (3.13).

## 4.1. Stabilized second order formulas

We assume that the Jacobian matrix J may be written as a 2 by 2 matrix with matrices as entries (possibly originating from a system of two partial differential equations) and that $J_{22}$ is the zero matrix. Considering generalized schemes (2.2) with k=2, we derive from table 3.2 (and partially 3.1) the relations for second order consistency:

$$
\begin{aligned}
& M_m^{(1)} = 0, \qquad M_m^{(2)} = 0 \\
& \sum_{l=0}^{m-1} N_{ml}^{(0)} = I, \qquad \sum_{l=0}^{m-1} N_{ml}^{(1)} = 0 \\
& \sum_{l=0}^{m-1} N_{ml}^{(0)}(p,q)\, M_l^{(1)}(r,s) = 0 , \qquad p,q,r,s \in \{1,2\} \qquad q + r \neq 4, \\
& \sum_{l=0}^{m-1} N_{ml}^{(0)}(p,q) \sum_{k=0}^{l-1} N_{lk}^{(0)}(r,s) = \tfrac{1}{2}\delta_{pq}\delta_{rs}, \quad p,q,r,s \in \{1,2\} \quad q + r \neq 4.
\end{aligned}
\tag{4.1}
$$

From definition (3.8a) it follows that the amplification matrix is given by

$$
\begin{aligned}
R_m(hJ_n) = M_m + h \sum_{l=0}^{m-1} N_{ml} J_n R_l(hJ_n) &= M_m + h \sum_{l=0}^{m-1} N_{ml} J_n M_l \\
&+ h^2 \sum_{l=0}^{m-1} N_{ml} J_n \sum_{k=0}^{l-1} N_{lk} J_n M_k + h^3 \ldots,
\end{aligned}
$$

so that, using (3.1), (3.2) and (4.1) we find

$$
R_m(hJ_n) = I + hJ_n + \tfrac{1}{2}h^2 J_n^2 + O(h^3). \tag{4.2}
$$

Assuming that notation (3.15) is applicable, we see that the polynomials P and Q can be written as

$$
\begin{aligned}
& P(z) = 1 + z + \tfrac{1}{2}z^2 + p_3 z^3 + \ldots + p_m z^m, \\
& Q(z) \equiv q_3 z^3 + \ldots + q_m z^m.
\end{aligned}
\tag{4.3}
$$

Now, we try to choose the parameters $p_3,\ldots,p_m$ and $q_3,\ldots,q_m$ in such a way, that the conditions (3.16) are fulfilled for $0 \le -iz \le \beta_{im}$, for a value of $\beta_{im}$ as large as possible.

*2-point formulas:*

For a 2-point formula we have $P_2(z) = 1+z+\frac{1}{2}z^2$ and $Q_2(z) = 0$, and substitution in (3.16) shows that the Hurwitz conditions are violated for small imaginary values of z.

*3-point formulas*

Substitution of $P_3(z)$ and $Q_3(z)$ in (3.16) yields the conditions

$$1 + z^4\,(\tfrac{1}{4}-2p_3) + z^6\,(q_3^2-p_3^2) \le 1$$

and

$$|2 + z^2| \le 2 + z^4\,(\tfrac{1}{4}-2p_3) + z^6\,(q_3^2-p_3^2).$$

It is easily verified that the choice $p_3 = q_3 = \frac{1}{8}$ results in an optimal stability boundary $\beta_{im} = 2$.

*Multi-point formulas*

Let us define the polynomials

$$V_m(z^2) = |P_m(z)|^2 - |Q_m(z)|^2 = 1 + v_4 z^4 + \ldots + v_{2m} z^{2m}$$

(4.4) and

$$W_m(z^2) = 2 \operatorname{Re} P_m(z) = 2 + z^2 + w_4 z^4 + \ldots + w_{2k} z^{2k}, \qquad 2k \le m.$$

We now can express the Hurwitz-conditions (3.16) in terms of $V_m$ and $W_m$ as follows:

$$V_m(s) \le 1 \qquad (s = z^2 \le 0)$$

(4.5) $$W_m(s) \le V_m(s) + 1,$$

$$-W_m(s) \le V_m(s) + 1.$$

An optimum is achieved for

(4.6) $V_m(s) = 1$ and $W_m(s) = 2\,T_k(\frac{2s^2}{\beta} + 1)$, where $\beta = 4k^2$, and $T_k(x)$ is the Chebyshev polynomial of degree k.

Unfortunately, we cannot find $P_m$ and $Q_m$ related according (4.4) to the polynomials $V_m$ and $W_m$ as given by (4.6) for all values of m. However, for odd values of m we did find polynomials satisfying (4.4), namely

(4.6a) and

$$P_m(z) = \frac{1}{z}\{1 + z + \tfrac{1}{2}z^2)\,T_k(\frac{z^2}{2k^2} + 1) - 1 - \frac{1}{8}z^4\}$$

$$Q_m(z) = P_m(z) - 1 - z - \tfrac{1}{2}z^2.$$

In table 4.1 we list the polynomials $P_m(z)$ for m=3,5 and 7.

Table 4.1. Optimal polynomials $P_m(z)$ for m=3,5 and 7.

| | |
|---|---|
| m=3 | $P_3(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{8}z^3$ |
| m=5 | $P_5(z) = 1 + z + \frac{1}{2}z^2 + \frac{5}{32}z^3 + \frac{1}{32}z^4 + \frac{1}{64}z^5$ |
| m=7 | $P_7(z) = 1 + z + \frac{1}{2}z^2 + \frac{35}{216}z^3 + \frac{1}{27}z^4 + \frac{14}{729}z^5 + \frac{1}{1458}z^6 + \frac{1}{2916}z^7.$ |

4.2. Almost second order formulas using two arrays of storage.

In the derivation of the Runge-Kutta matrices $M_j$ and $N_{jl}$ we will only consider the case that these matrices do not depend on the stepsize h. From the condition (3.5) for schemes only using two arrays of storage, together with the conditions (3.1) and (3.2), we find that $R_m(hJ_n)$ is given by

(4.7) $$R_m(hJ_n) = \prod_{l=0}^{m-1} (I + hN_{l+1,l}J_n).$$

Writing $N_{l+1,l} = \begin{pmatrix} \mu_l I & 0 \\ 0 & \beta_l I \end{pmatrix}$, and substituting for T and $J_n$ in the relation

$$T\begin{pmatrix} P_m(h\Lambda) & Q_m(h\Lambda) \\ Q_m(h\Lambda) & P_m(h\Lambda) \end{pmatrix} T^{-1} = R_m(hJ_n)$$

the matrix of eigenvectores and the Jacobian of the model problem (3.13), we obtain the stability conditions

$$(4.8)\qquad \begin{aligned} p_k + q_k &= \sum_{j=1}^{m} \mu_j \sum_{r=1}^{j-1} \beta_r \sum_{s=1}^{r-1} \mu_s \cdots \sum_{t=1} . \quad (k \text{ sums, last term is } \beta \text{ for } k \text{ even, else } \mu) \\ p_k - q_k &= \sum_{j=1}^{m} \beta_j \sum_{r=1}^{j-1} \mu_r \sum_{s=1}^{r-1} \mu_s \cdots \sum_{t=1} . \quad (\text{last term, is } \mu \text{ for } k \text{ even, else } \lambda),\ k = 3,\ldots,m. \end{aligned}$$

Moreover, we derive from (4.1) the conditions for consistency of order two:

$$(4.9)\qquad \sum_{j=1}^{m} \mu_j = 1, \quad \sum_{j=1}^{m} \beta_j = 1, \quad \sum_{j=2}^{m} \beta_j \sum_{k=1}^{j-1} \mu_k = \tfrac{1}{2}, \quad \sum_{j=2}^{m} \mu_j \sum_{k=1}^{j-1} \beta_k = \tfrac{1}{2},$$
$$, \quad \sum_{j=2}^{m} \mu_j \sum_{k=1}^{j-1} \mu_k = \tfrac{1}{2}.$$

Obviously, (4.8) and (4.9) consist of (2m+1) equations in the (2m) unknowns $\beta_j$ and $\mu_j$, so that we may not expect to find a solution yielding a second order scheme with optimal polynomials $P_m(z)$ and $Q_m(z)$. Of course we could have found second order schemes by admitting less optimal $P_m$ and $Q_m$. However, we did remove the last consistency condition, so that the schemes constructed are only of second order if $J_{11} = 0$ (e.g. which is the case by second order equations without first derivatives, written as a system of first order equations).

Now, we can easily calculate the parameters $\beta_j$ and $\mu_j$ from (4.8) and (4.9) for polynomials $P_m$ and $Q_m$ given by (4.6a). However, a more efficient set of formulas is given by the relations

$$(4.10)\qquad \mu_1 = \mu_m = \frac{1}{2k},\ \mu_{even} = \beta_{odd} = 0,\ \mu_{odd} = \beta_{even} = \frac{1}{k}, \qquad m = 2k+1.$$

It is easily verified that the first four relations of (4.9) are satisfied, whereas substitution in (4.8) yields for m=3,5 and 7 the polynomials

$$
\begin{aligned}
&\tilde{P}_3(z) = 1 + z + \tfrac{1}{2}z^2 + \tfrac{1}{8}z^3, \qquad \tilde{Q}_3(z) = \tfrac{1}{8}z^3,\\
&\tilde{P}_5(z) = 1 + z + \tfrac{1}{2}z^2 + \tfrac{5}{32}z^3 + \tfrac{1}{32}z^4 + \tfrac{1}{256}z^5, \qquad \tilde{Q}_5(z) = \tfrac{1}{32}z^3 + \tfrac{1}{256}z^5,\\
&\tilde{P}_7(z) = 1 + z + \tfrac{1}{2}z^2 + \tfrac{35}{216}z^3 + \tfrac{1}{27}z^4 + \tfrac{1}{162}z^5 + \tfrac{1}{1458}z^6 + \tfrac{1}{17496}z^7,\\
&\tilde{Q}_7(z) = \tfrac{1}{72}z^3 + \tfrac{1}{486}z^5 + \tfrac{1}{17496}z^7.
\end{aligned}
\tag{4.11}
$$

These polynomials are different from those listed in table 4.1 for m=5 and m=7; however, they satisfy the relations (4.4) and (4.6), so that the imaginary stability boundary $\beta_{im}$ is again equal to 2 (m=3), 4 (m=5) and 6 (m=7). The advantage of the formulas given by (4.10) lies in the fact that per Runge-Kutta step only (k+1) evaluations of the first component of the right hand side of (1.1) are required, and k evaluations of the second component. Thus, the computational work is approximately m/2 right hand side evaluations. Defining *the effective stability boundary* $\beta_{eff,im}$ as the quotient of $\beta_{im}$ and the number of right hand side evaluations per Runge-Kutta step, we obtain values as listed in table 4.2.

Table 4.2. $\beta_{im}$, $\beta_{eff,im}$ and the number of right hand side evaluations

| m | Scheme | generated by | $\beta_{im}$ | number of r.h.s.eval. | $\beta_{eff,im}$ |
|---|---|---|---|---|---|
| 3 | $P_3$ and | $Q_3$ from table 4.1 | 2 | ≤3 | ≥.67 |
| 3 | $\tilde{P}_3$ and | $\tilde{Q}_3$ from 4.11 | 2 | 1.5 | 1.33 |
| 5 | $P_5$ and | $Q_5$ from table 4.1 | 4 | ≤5 | ≥.80 |
| 5 | $\tilde{P}_5$ and | $\tilde{Q}_5$ from 4.11 | 4 | 2.5 | 1.60 |
| 7 | $P_7$ and | $Q_7$ from table 4.1 | 6 | ≤7 | ≥.86 |
| 7 | $\tilde{P}_7$ and | $\tilde{Q}_7$ from 4.11 | 6 | 3.5 | 1.71 |

From these results we expect that all schemes generated by (4.10) have, for odd values of m, a $\beta_{im}$ equal to m-1. For large values of m we would then obtain a $\beta_{eff,im}$ which is approximately equal to 2. However, we did not succeed in proving this relation for all odd m. Finally, we remark that the schemes generated by (4.10) looks like the "symmetrized-scheme" proposed by VAN DER HOUWEN [4] for the integration of the shallow water equations. We intend to apply our schemes to these equations in the near future.

## 4.3. Second order formulas using three arrays of storage.

When we allow ourselves three arrays of storage, it turns out to be possible to satisfy the conditions for second order consistency (4.1) and for stability (4.8). Thus, we will try to reduce the number of right hand side evaluations per step. For that purpose we consider two subclasses of schemes defined by (3.7).

First we choose $A_j=I$, $B_j=+I$, $D_j=0$, $E_j=-F_j$, $j=1,\ldots,m$, which yields for $N_{j1}$ and $M_j$ the relations

$$(4.12)\qquad M_j = I, \qquad N_{j1} = 0, \qquad j = 1,\ldots,m, \qquad 1 = 0,\ldots,j-2.$$

Writing $N_{j,j-1} = \begin{pmatrix} \mu_j I & 0 \\ 0 & \beta_j I \end{pmatrix}$, as in the previous section, we find the following relations for $\mu_j$ and $\beta_j$:

$$(4.13)\qquad \mu_m = \beta_m = 1, \qquad \mu_m\mu_{m-1} = \mu_m\beta_{m-1} = \beta_m\mu_{m-1} = \tfrac{1}{2},$$
$$p_k = (\mu_m\beta_{m-1}\mu_{m-2}\cdots\cdot_{m-k+1} + \beta_m\mu_{m-1}\beta_{m-2}\cdots\cdot_{m-k+1})/2$$
$$q_k = (\mu_m\beta_{m-1}\mu_{m-2}\cdots\cdot_{m-k+1} - \beta_m\mu_{m-1}\beta_{m-2}\cdots\cdot_{m-k+1})/2.$$

For given $p_k$ and $q_k$ the parameters $\mu_j$ and $\beta_j$ can be determined easily. The results, corresponding to the polynomials $\tilde{P}_m(z)$ and $P_m(z)$ as listed in formula (4.11) and table 4.1 are given below:

$$(4.14a)\qquad m = 3,\ N_{32} = I,\ N_{21} = \tfrac{1}{2}I,\ N_{10} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

$$(4.14b)\qquad m = 5,\ N_{54} = I,\ N_{43} = \tfrac{1}{2}I,\ N_{32} = \begin{pmatrix} \frac{3}{8} & 0 \\ 0 & \frac{1}{4} \end{pmatrix},\ N_{21} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{6} \end{pmatrix},$$
$$,\ N_{10} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 0 \end{pmatrix}.$$

$$(4.14c)\qquad m = 7,\ N_{76} = I,\ N_{65} = \tfrac{1}{2}I,\ N_{54} = \frac{1}{54}\begin{pmatrix} 19 & 0 \\ 0 & 16 \end{pmatrix},\ N_{43} = \frac{1}{76}\begin{pmatrix} 19 & 0 \\ 0 & 16 \end{pmatrix},$$
$$,\ N_{32} = \frac{1}{9}\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},\ N_{21} = \frac{1}{12}\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},\ N_{10} = \frac{1}{6}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

$$(4.14d)\qquad m = 5,\ N_{54} = I,\ N_{43} = \tfrac{1}{2}I,\ N_{32} = \begin{pmatrix} \frac{5}{8} & 0 \\ 0 & 0 \end{pmatrix},\ N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{5} \end{pmatrix},$$

$$, \; N_{10} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

(4.14e) $$m = 7, \; N_{76} = I, \; N_{65} = \tfrac{1}{2}I, \; N_{54} = \begin{pmatrix} \frac{35}{54} & 0 \\ 0 & 0 \end{pmatrix}, \; N_{43} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{8}{35} \end{pmatrix},$$

$$, \; N_{32} = \begin{pmatrix} \frac{14}{27} & 0 \\ 0 & 0 \end{pmatrix}, \; N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{28} \end{pmatrix}, \; N_{10} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

Obviously, the schemes (4.14d) and (4.14e) are more efficient than (4.14b) and (4.14d), as more zeros appear in the matrices. In table 4.3 we mention the effective stability boundary for these schemes.

When we try to maximize the number of zeros in the parameter matrices $N_{j1}$, (3.7) seems to impose a too severe condition. However, schemes requiring only (m+1)/2 right hand side evaluations can be constructed, when we consider the class of formulas given by

(4.15) $$N_{j1} = 0, \; j = 1,\ldots,m, \; 1 = 1,\ldots,j-2, \text{ and } N_{m0} = 0, \; M_j = I,$$

$$j = 1,\ldots,m, \; N_{j0} = N_{10}, \; j = 2,\ldots,m-1.$$

The consistency conditions now read

$$\mu_m = \beta_m = 1, \; \mu_m(\mu_1+\mu_{m-1}) = \mu_m(\beta_1+\beta_{m-1}) = \beta_m(\mu_1+\mu_{m-1}) = \tfrac{1}{2}$$

whereas the coeeficients of the stability functions are given by

$$p_k + q_k = \mu_m\beta_{m-1}\cdots(\cdot_1+\cdot_{m-k+1}); \quad (\cdot \text{ stand for } \mu, \text{ if } k \text{ is odd, else } \beta)$$

$$p_k - q_k = \beta_m\mu_{m-1}\cdots(\cdot_1+\cdot_{m-k+1}), \quad (\cdot \text{ stand for } \beta \text{ if } k \text{ is odd, else } \mu).$$

Choosing the coefficients $p_k$ and $q_k$ as given in formula (4.11), we obtain the following schemes:

(4.16a) $m = 3,\ N_{32} = I,\ N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},\ N_{20} = N_{10} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$

(4.16b) $m = 5,\ N_{54} = I,\ N_{43} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},\ N_{32} = \begin{pmatrix} \frac{1}{8} & 0 \\ 0 & 0 \end{pmatrix},\ N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$

$N_{40} = N_{30} = N_{20} = N_{10} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$

(4.16c) $m = 7,\ N_{76} = I,\ N_{65} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},\ N_{54} = \begin{pmatrix} \frac{4}{27} & 0 \\ 0 & 0 \end{pmatrix},\ N_{43} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$

$N_{32} = \begin{pmatrix} \frac{1}{54} & 0 \\ 0 & 0 \end{pmatrix},\ N_{21} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},\ N_{60}=N_{50}=N_{40}=N_{30}=N_{20}=N_{10} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}.$

The effective stability boundaries of these schemes are given in table 4.3.

As the matrices $N_{j1}$ are sparse, implementation of these schemes using not more than three arrays is possible, too.

Table 4.3. The effective stability boundary of the schemes (4.14) and (4.16)

| m | Scheme | $\beta_{im}$ | number of r.h.s.eval. | $\beta_{eff,im}$ |
|---|---|---|---|---|
| 3 | 4.14a | 2 | $2\frac{1}{2}$ | .80 |
| 3 | 4.16a | 2 | 2 | 1.00 |
| 5 | 4.14b | 4 | $4\frac{1}{2}$ | .89 |
| 5 | 4.14d | 4 | $3\frac{1}{2}$ | 1.14 |
| 5 | 4.16b | 4 | 3 | 1.33 |
| 7 | 4.14c | 6 | $6\frac{1}{2}$ | .92 |
| 7 | 4.14e | 6 | $4\frac{1}{2}$ | 1.33 |
| 7 | 4.16c | 6 | 4 | 1.67 |

REMARK. When we apply the schemes determined by (4.10) to the second order equation $\frac{d^2y}{dx^2} = g(y)$, we obtain the relations:

(a) $m = 3:\ y_{n+1} = y_n + hy'_n + \frac{1}{2}h^2 g(y_n + \frac{1}{2}hy'_n),$

$$y'_{n+1} = 2\,\frac{y_{n+1} - y_n}{h} - y'_n,$$

(b) $m = 5$:

$$y_{n+1}^{(1)} = y_n + \tfrac{1}{4}hy'_n,$$

$$y_{n+1}^{(2)} = y_{n+1}^{(1)} + \tfrac{1}{2}hy' + \tfrac{1}{4}h^2 g(y_{n+1}^{(1)})$$

$$y_{n+1} = y_n + hy'_n + \tfrac{3}{8}h^2 g(y_{n+1}^{(1)}) + \tfrac{1}{8}h^2 g(y_{n+1}^{(2)}),$$

$$y'_{n+1} = y'_n + \tfrac{1}{2}hg(y_{n+1}^{(1)}) + \tfrac{1}{2}hg(y_{n+1}^{(2)}).$$

As these formulas require only two arrays of storage, they are more economical than formula (2.5), which requires three arrays. We mention that this formula, which was devised for second order equations without first derivatives in VAN DER HOUWEN [5], can be constructed by the methods described in this report, too. In fact, let us consider almost second order formulas (the condition $\sum_j N_{ml}(p,q) \sum N_{lk}(r,s) = \tfrac{1}{2}\delta_{pq}\delta_{rs}$ need not be satisfied for r=s=p=q=1), which use three arrays of storage. Setting

$$N_{54} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \text{ and } N_{53} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \text{ and using } \tilde{P}_5(z) \text{ and } \tilde{Q}_5(z)$$

from (4.11), we obtain formula (2.5) by a relation similar to (4.13).

## 4.4. Strongly stable formulas.

The formulas generated in the preceding sections are only weakly stable, as their associated polynomials P(z) and Q(z) satisfy (3.16) with the equality sign. Indeed, it is easily verified that their amplification factors $\alpha$ are exactly in modulus 1. Strongly stable formulas, whose amplification factors are bounded by a damping function $\sqrt{\rho(z)}$, can be constructed as described by VAN DER HOUWEN [5].

Instead of the conditions (3.16) we now satisfy

$$|P(z)|^2 - |Q(z)|^2 \leq \rho(z)$$

(4.17) and

$$|\mathrm{Re}\ P(z)| \leq \rho(z).$$

Setting again (cf. (4.6)) $Q(z) = P(z) - 1 - z - \frac{1}{2}z^2$, the derivation of Re $P(z)$ is completely analogous to the derivation of $S(z)$ in [5]. Therefore, we suffice with giving the resulting polynomials, together with their $\beta_{im}$, for m=3,5 and 7.

$$(4.18) \qquad P_{3,\varepsilon}(z) = 1 + z + \tfrac{1}{2}z^2 + (\frac{1}{8} + \frac{\varepsilon}{2\beta^4})z^3, \qquad \beta_{im} = \beta = \sqrt{4-2\varepsilon},$$

The related damping function is $\rho(z) = 1 - \frac{\varepsilon z^4}{\beta^4}$ .

$$(4.19) \qquad P_{5,\varepsilon}(z) = 1 + z + \tfrac{1}{2}z^2 + (\frac{1}{8} + \frac{\varepsilon}{2\beta^4} + \frac{\beta^2-2\varepsilon}{2\beta^4})z^3 + \frac{\beta^2-2\varepsilon}{2\beta^4} + \frac{\beta^2-2\varepsilon}{4\beta^4}z^5,$$

$$\beta_{im} = \beta = \sqrt{8(1+\sqrt{1-\varepsilon})}.$$

Again, the related damping function is $\rho(z) = 1 - \frac{\varepsilon z^4}{\beta^4}$ .

$$P_{7,\varepsilon}(z) = 1 + z + \tfrac{1}{2}z^2 + (\frac{35}{216} + \frac{\varepsilon}{216})z^3 + (\frac{1}{27} + \frac{\varepsilon}{288})z^4 +$$

$$+ (\frac{14}{729} + \frac{5\varepsilon}{2592})z^5 + (\frac{4+\varepsilon}{5832})z^6 + \tfrac{1}{2}(\frac{4+\varepsilon}{5832})z^7, \quad \beta_{im} = \beta \approx 6 - \frac{3}{4}\varepsilon$$

(terms of order $\varepsilon^2$ are neglected).

The damping function related to $P_{7,\varepsilon}(z)$ is $\rho(z) = 1 - \frac{3\varepsilon}{\beta^4}z^4 - \frac{2\varepsilon z^6}{\beta^6}$.

Now, using the above values of the coefficients $p_k$ and $q_k$ of the polynomials $P(z)$ and $Q(z)$, we can compute the Runge-Kutta parameters $\mu_j$ and $\beta_j$ by means of (4.8) or (4.13). As the schemes computed by using (4.8) contains only few zeros compared to the schemes (4.10), which are exactly the same ones for $\varepsilon = 0$, the use of a damping substitute for (4.10) does not seem appropriate.

However, using (4.13) for the calculation of the $\mu_j$ and $\beta_j$, we find only slight modifications of ($4.14^a$), ($4.14^d$) and ($4.14^e$) . The resulting parameter values are listed in table 4.4.

Table 4.4 Runge-Kutta parameters for second order strongly stable schemes

| Associated polynomials | Damping function | $\beta = \beta_{im}$ | RK parameters | |
|---|---|---|---|---|
| $P_{3,\varepsilon}(z)$ | $1-\frac{\varepsilon z^4}{\beta^4}$ | $\sqrt{4-2\varepsilon}$ | $\mu_3 = 1$ | $\beta_3 = 1$ |
| | | | $\mu_2 = \frac{1}{2}$ | $\beta_2 = \frac{1}{2}$ |
| | | | $\mu_1 = \frac{1}{2} + \frac{2\varepsilon}{\beta^4}$ | $\beta_1 = 0$ |
| $P_{5,\varepsilon}(z)$ | $1-\frac{\varepsilon z^4}{\beta^4}$ | $\sqrt{8(1+\sqrt{1-\varepsilon})}$ | $\mu_5 = 1$ | $\beta_5 = 1$ |
| | | | $\mu_4 = \frac{1}{2}$ | $\beta_4 = \frac{1}{2}$ |
| | | | $\mu_3 = \frac{1}{2} + \frac{2\beta^2-2\varepsilon}{\beta^4}$ | $\beta_3 = 0$ |
| | | | $\mu_2 = 0$ | $\beta_2 = \frac{4\beta^2 - 8\varepsilon}{\beta^4 + 4\beta^2 - 4\varepsilon}$ |
| | | | $\mu_1 = \frac{1}{2}$ | $\beta_1 = 0$ |
| $P_{7,\varepsilon}(z)$ | $1-\frac{3\varepsilon z^4}{\beta^4}+\frac{2\varepsilon z^6}{\beta^6}$ | $\sqrt{36-9\varepsilon}$ | $\mu_7 = 1$ | $\beta_7 = 1$ |
| | | | $\mu_6 = \frac{1}{2}$ | $\beta_6 = \frac{1}{2}$ |
| | | | $\mu_5 = \frac{35+\varepsilon}{54}$ | $\beta_5 = 0$ |
| | | | $\mu_4 = 0$ | $\beta_4 = \frac{32+3\varepsilon}{140+4\varepsilon}$ |
| | | | $\mu_3 = \frac{1}{27}\frac{448+45\varepsilon}{32+3\varepsilon}$ | $\beta_3 = 0$ |
| | | | $\mu_2 = 0$ | $\beta_2 = \frac{16+4\varepsilon}{448+45\varepsilon}$ |
| | | | $\mu_1 = \frac{1}{2}$ | $\beta_1 = 0$ |

SUMMARY

The first order formulas defined by (4.10) are the most efficient ones as they yield optimal values of $\beta_{eff,im}$, especially for large values of m, and have minimal storage requirements. These formulas might be used when only low accuracy is requested, or when the Jacobian matrix of the system to be solved has a small component matrix $J_{11}$.

When one is interested in higher accuracies, the second order formulas given by (4.16) and in table (4.4) may come into consideration. The former formulas are weakly stable, just as the first order schemes the latter strongly stable at the cost of an additional $\frac{1}{2}$ function evaluation. We expect that the weakly stable formulas will be the most efficient if the range of integration is short, whereas the strongly stable schemes will be superior for long ranges. In a next paper these suggestions will be verified.

REFERENCES

1. FEHLBERG, E., *Classical eight- and lower-order Runge-Kutta-Nyström formulas with step-size control for special second-order differential equations*, Technical Report, NASA TR R-381, Marshall Space Flight Centre, Alabama, 1972.

2. VAN DER HOUWEN, P.J., *Stabilized Runge-Kutta methods with limited storage requirements*. Report TW 124/71, Mathematisch Centrum, Amsterdam, 1971.

3. VAN DER HOUWEN, P.J., *Explicit Runge-Kutta formulas with increased stability boundaries*. Numer Math. 20 (1972), pp.149-164.

4. VAN DER HOUWEN, P.J., *Two-level difference schemes with varying mesh sizes for the shallow water equations*. Report NW 22/75, Mathematisch Centrum, Amsterdam 1975.

5. VAN DER HOUWEN, P.J., *Stabilized Runge-Kutta methods for second-order differential equations without first derivatives*. Report NW 26/75, Mathematisch Centrum, Amsterdam, 1975.

6. KREIS, H., and J. OLIGER, *Methods for the approximate solution of time dependent problems*. GARP publication series no 10, Geneva, 1973.

7. LAMBERT, J.D., *Computational methods in ordinary differential equations*. John Wiley & Sons, London, 1973.

8. ZONNEVELD, J.A., *Automatic numerical integration*. MC Tract 8, Mathematisch Centrum, Amsterdam, 1964.