stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

NW 40/77 MAART

P.J. VAN DER HOUWEN

RUNGE-KUTTA TYPE METHODS FOR THE INTEGRATION 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

Runge-Kutta type methods for the integration of hyperbolic differential equations *)

by

P.J. van der Houwen

ABSTRACT

First and second order Runge-Kutta formulas are presented for the integration of the large systems of second order differential equations arising from the semi-discretization of certain classes of hyperbolic differential equations. These formulas are characterized by their low storage requirements and their relatively large real stability interval. Numerical experiments are reported which show that the new formulas are superior to the stabilized Runge-Kutta formulas for first order equations both with respect to accuracy and to the computational effort involved.

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

---

*) This report will be submitted for publication elsewhere.

1. INTRODUCTION

Let

$$\frac{d^2\vec{y}}{dx^2} = \vec{f}(x,\vec{y}) \tag{1.1}$$

represent a set of differential equations of which the real-valued vector function $\vec{f}$ belongs to a class of sufficient differentiability. In order to solve the initial value problem for this special class of second order equations, we do not convert it into an initial value problem for a larger system of first order equations as is usually done in the case where the right hand side contains first derivatives, but we try to exploit the special form of the equation (cf. HENRICI [3,p.169]). We shall concentrate on the case where the Jacobian matrix of the right hand side has its eigenvalues in a long strip along the *negative* axis. Such equations arise when hyperbolic equations are discretized with respect to the space variables. In many hyperbolic initial value problems, it suffices to use a first or second order accurate time discretization. Therefore, we will confine our considerations to integration formulas of first and second order. In particular, formulas of Runge-Kutta type will be considered, i.e. formulas of the form

$$\begin{aligned}
&\vec{y}_{n+1}^{(0)} = \vec{y}_n, \\
&\vec{y}_{n+1}^{(j)} = \vec{y}_n + \mu_j h_n \vec{y}_n' + h_n^2 \sum_{\ell=0}^{j-1} \lambda_{j,\ell} \vec{f}(x_n+\mu_\ell h_n, \vec{y}_{n+1}^{(\ell)}), \qquad j = 1,2,\ldots m, \\
&\vec{y}_{n+1} = \vec{y}_{n+1}^{(m)}, \ \vec{y}_{n+1}' = \vec{y}_n' + h_n \sum_{\ell=0}^{m-1} \beta_\ell \vec{f}(x_n+\mu_\ell h_n, \vec{y}_{n+1}^{(\ell)}), \qquad \mu_0 = 0.
\end{aligned} \tag{1.2}$$

Here, h_n is the steplength $x_{n+1}-x_n$, and $\vec{y}_n'$, $\vec{y}_n'$, $n = 1,2,\ldots$ represent numerical approximations to $\vec{y}(x), d\vec{y}/dx$ at x_n.

The consistency conditions for scheme (1.2) are well known (see e.g. [3]); first order consistency is obtained for

$$\mu_m = \sum_{\ell=0}^{m-1} \beta_\ell = 1, \tag{1.3}$$

second order consistency when, in addition,

$$(1.4)\qquad \sum_{\ell=0}^{m-1} \lambda_{m,\ell} = \sum_{\ell=1}^{m-1} \beta_\ell \mu_\ell = \frac{1}{2}.$$

When scheme (1.2) is applied to the test equation

$$(1.5)\qquad \frac{d^2\vec{y}}{dx^2} = J\vec{y}$$

we obtain the relation (cf. ANSORGE and TORNIG [1])

$$(1.6)\qquad \begin{pmatrix} \vec{y}_{n+1} \\ h_n \vec{y}_{n+1} \end{pmatrix} = R(h_n^2 J) \begin{pmatrix} \vec{y}_n \\ h_n \vec{y}_n \end{pmatrix},$$

where R is a matrix-valued function of the argument $h_n^2 J$. This function is defined by the scheme

$$(1.7)\qquad \begin{aligned} R_0(z) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ R_j(z) &= \begin{pmatrix} 1 & \mu_j \\ 0 & 1 \end{pmatrix} + z \sum_{\ell=0}^{j-1} \lambda_{j,\ell} R_\ell(z), \qquad j = 1,2,\ldots,m-1, \\ R(z) = R_m(z) &= \begin{pmatrix} 1 & \mu_m \\ 0 & 1 \end{pmatrix} + z \sum_{\ell=0}^{m-1} \begin{pmatrix} \lambda_{m,\ell} & 0 \\ \beta_\ell & 0 \end{pmatrix} R_\ell(z). \end{aligned}$$

The elements of the matrix R(z) are polynomials of degree m in z. We shall call R(z) the *stability matrix* associated to scheme (1.2). Let z be a complex number and $\alpha_j(z)$ denote the eigenvalues of R(z), then we define the region

$$(1.8)\qquad \{z \mid |\alpha_j(z)| < 1,\ j = 1,2\}$$

as the *stability region* of scheme (1.2). When all points δh_n^2 with $\delta \in \Delta$, Δ being the set of eigenvalues of the Jacobian matrix $\partial\vec{f}/\partial\vec{y}$, are within the stability region, we call scheme (1.2) *strongly stable*. When the Jacobian has one or more eigenvalues δ such that the eigenvalues $\alpha_j(h_n^2\delta)$ of $R(h_n^2\delta)$ are on the unit circle we shall call (1.2) *weakly stable*.

In the following sections we try to maximize the length of the negative stability interval

$$(1.9) \qquad (-\beta,0) = \{z \mid z < 0, \quad |\alpha_j(z)| < 1, \ j = 1,2\}.$$

The corresponding stability condition becomes

$$(1.10) \qquad h_n < \sqrt{\frac{\beta}{|\delta|_{max}}},$$

provided that $\partial\vec{f}/\partial\vec{y}$ has a negative eigenvalue spectrum. It turns out that for all formulas of type (1.2) with optimal interval of stability, the maximal "step-length per function evaluation", to be denoted by h_{eff}, is close to the value $2/\sqrt{|\delta|_{max}}$, irrespective the number of stages involved. The formulas only differ by their order of accuracy and by the way of damping of the higher harmonics in the numerical solution $\vec{y}_n$. This result implies that within the class of stabilized formulas of type (1.2) we may choose a formula using only a few stages without lack of efficiency. When we compare the efficiency of the formulas proposed in this paper with that of the stabilized Runge-Kutta for first order equations, we may conclude that we have gained a factor greater than 2. For when equation (1.1) is converted into a first order system, application of an m-point, second order Runge-Kutta method with maximal *imaginary* stability boundary (note that the Jacobian matrix of the first order system has imaginary eigenvalues when $\partial\vec{f}/\partial\vec{y}$ has negative ones) results in the stability condition (cf.[5]).

$$(1.11) \qquad h_n < \frac{m-1}{\sqrt{|\delta|_{max}}}, \qquad m = 3,5,7, \quad \ldots \ .$$

Hence, $h_{eff} = (m-1) \ / \ (m\sqrt{|\delta|_{max}})$, whereas the new formulas yield $h_{eff} \cong 2/\sqrt{|\delta|_{max}}$.

In section 2.3 a modification of scheme (1.2) is discussed which is characterized by the fact that all function evaluations $\vec{f}(x_n+\mu_\ell h_n, \vec{y}^{(\ell)}_{n+1})$, preceded by parameters $\lambda_{j,\ell}$ which are not involved in the consistency conditions, are replaced by the vectors $J^* \vec{y}^{(\ell)}_{n+1}$ where J^* is some approximation to the Jacobian matrix $\partial\vec{f}/\partial\vec{y}$ at the point $(x_n,\vec{y}_n)$. For sufficiently close approximations the stability theory for these modified formulas is identical

to that for scheme (1.2). We shall maximize the interval of stability for a class of first and second order formulas requiring one $\vec{f}$ and (m-2) $J^* \vec{y}$ evaluations. These formulas contain a control function by which the damping of the higher harmonics can be monitored. In case of mild damping both classes have a stability boundary close to $4(m-1)^2$. Thus, when the evaluation of the vectors $J^* \vec{y}^{(\ell)}_{n+1}$ requires less computational effort than the evaluation of the vectors $\vec{f}(\vec{y}^{(\ell)}_{n+1})$, the modified formulas have a larger effective steplength h_{eff} than the original ones. In this connection we observe that other classes of stabilized Runge-Kutta formulas such as those given in [5], can also be economized in the way described above.

Finally, in section 3, some numerical experiments are reported. More extensive tests will be published in [2].

2. FORMULAS WITH EXTENDED REAL STABILITY INTERVAL

First of all we remark that the number of free parameters, relative to the number of right hand side evaluations, can be increased by one when we choose

$$\lambda_{j,0} = \beta_0 = 0, \qquad j = 1,2,\ldots,m, \qquad m \geq 2. \tag{2.1}$$

For by this choice we obtain an (m-1)-stage formula containing m free parameters μ_j, m(m-1)/2 free parameters $\lambda_{j,\ell}$ and m-1 free parameters β_ℓ. Thus, together $(m^2+3m-2)/2$ parameters at the price of m-1 right hand side evaluations, whereas scheme (1.2) contains $(m^2+5m)/2$ free parameters at the price of m right hand side evaluations, i.e. $(m^2+3m-4)/2$ parameters for (m-1) $\vec{f}$ evaluations. In the following we assume that (2.1) is satisfied.

Furthermore, since one of the most important classes of equations to which stabilized Runge-Kutta methods are applied, is the class of the very large systems originating from the semi-discretization of partial differential equations, we are interested in schemes with reduced storage requirements. Therefore, we tried to construct formulas with

$$\lambda_{j,\ell} = \beta_\ell = 0, \qquad \ell < j-1, \quad j = 1,\ldots,m. \tag{2.2}$$

It turns out that this choice does not restrict the interval of stability.

2.1 First oder formulas.

The order equations for first order accuracy are (cf.(1.3))

$$\mu_m = \beta_{m-1} = 1. \tag{2.3}$$

The corresponding stability matrix R(z) is given by

$$R(z) = \begin{pmatrix} 1 + \lambda_{m,m-1} R_{21}(z) & 1 - \lambda_{m,m-1} + \lambda_{m,m-1} R_{22}(z) \\ R_{21}(z) & R_{22}(z) \end{pmatrix}, \tag{2.4}$$

where ($m > 2$)

$$R_{21}(z) = z(1+\lambda_{m-1,m-2}z(1+\lambda_{m-2,m-3}z(1+\ldots(1+\lambda_{2,1}z)\ldots),$$

$$R_{22}(z) = 1 + z(\mu_{m-1}+\lambda_{m-1,m-2}z(\mu_{m-2}+\lambda_{m-2,m-3}z(\ldots$$

$$\ldots(\mu_2+\lambda_{2,1}\mu_1 z)\ldots);$$

for $m = 2$ we have $R_{21}(z) = z$ and $R_{22}(z) = \mu_1 z + 1$ (note that by our choice (2.1), the degree of the polynomials $R_{21}(z)$ and $R_{22}(z)$ is reduced to m-1). The eigenvalues of R(z) satisfy the equation

$$\alpha^2 - S(z)\alpha + P(z) = 0, \tag{2.5}$$

where

$$\begin{aligned} S(z) &= 2 + \sigma_1 z + \sigma_2 z^2 + \ldots + \sigma_{m-1} z^{m-1}, \\ P(z) &= 1 + (\sigma_1 - 1)z + \pi_2 z^2 + \ldots + \pi_{m-1} z^{m-1} \end{aligned} \tag{2.6}$$

and where the coefficients σ_j and μ_j are given by ($m \geq 2$)

$$\sigma_1 = \lambda_{m,m-1} + \mu_{m-1}, \quad \sigma_j = \prod_{i=m-j+1}^{m-1} \lambda_{i,i-1}(\lambda_{m,m-1} + \mu_{m-j})$$

$$\pi_1 = \sigma_1^{-1}, \quad \pi_j = \prod_{i=m-j+1}^{m-1} \lambda_{i,i-1}(\lambda_{m,m-1} + \mu_{m-j}^{-1}), \quad j = 2,3,\ldots,m-1.$$

These expressions for σ_j and π_j are easily converted to obtain the Runge-Kutta parameters in terms of σ_j and π_j:

$$\lambda_{m,m-1} = \sigma_1 - \mu_{m-1}, \quad \lambda_{m-1,m-2} = \sigma_2 - \pi_2, \quad \lambda_{j,j-1} = \frac{\sigma_{m-j+1} - \pi_{m-j+1}}{\sigma_{m-j} - \pi_{m-j}},$$

$$\mu_j = \frac{\sigma_{m-j}}{\sigma_{m-j} - \pi_{m-j}} - \sigma_1 + \mu_{m-1}, \quad \sigma_m = \pi_m = 0, \qquad j = 1,2,\ldots,m-2,$$

where apart from the coefficients σ_j and π_j, the parameter μ_{m-1} also is a free parameter. We shall choose

$$\mu_{m-1} = \frac{1}{2}, \tag{2.8}$$

by which one of the second order terms in the truncation error vanishes. The coefficients σ_j and π_j are at our disposal for maximizing the stability interval.

THEOREM 2.1. *The length of the negative stability interval of scheme* (1.2) *satisfying* (2.1) *and* (2.2) *cannot exceed the value* $4(m-1)^2$.

PROOF. The stability interval on the z-axis is determined by the condition that the roots of (2.5) are within the unit circle, i.e. by the inequalities

$$|S| < P+1, \qquad P < 1, \qquad z < 0. \tag{2.9}$$

Hence, a necessary condition is $|S| \leq 2$. Thus, we are looking for a polynomial $S(z)$ of degree $m-1$ in z which remains as long as possible between -2 and $+2$. This type of minimax problem is well known and is solved by

$$S(z) = 2\, T_{m-1}\left(1 + \frac{\sigma_1 z}{2(m-1)^2}\right), \quad T_{m-1}(w) = \cos[(m-1)\arccos w],$$

where σ_1 is still a free parameter.

This polynomial remains between -2 and +2 in the interval $[-\frac{4(m-1)^2}{\sigma_1}, 0]$. From the definition of P(z) it follows that $\sigma_1 \geq 1$; hence $\sigma_1 = 1$ is the optimal value and $\beta = 4(m-1)^2$ is the maximal length of the interval of stability.

The proof of this theorem suggests to choose.

$$S(z) = 2\, T_{m-1}(1 + \frac{z}{2(m-1)^2}), \qquad \rho(z) = 1.$$

However, this choice results in a weakly stable method since

$$|\alpha(z)| = |T_{m-1} \pm \sqrt{T_{m-1}^2 - 1}| = 1.$$

Therefore, we introduce a function $\rho = \rho(z)$ assuming positive values less than 1 in the stability interval $(-\beta,0)$, and we replace (2.9) by

$$(2.10) \qquad |S| \leq \frac{P}{\sqrt{\rho}} + \sqrt{\rho}, \qquad P \leq \rho, \qquad -\beta < z < 0.$$

It is easily verified that these inequalities guarantee that $|\alpha(z)| \leq \sqrt{\rho(z)}$. Since it is sufficient for a stable behaviour that ρ is close to unity provided it is always less than unity, we simplify (2.10) by replacing it with the conditions

$$(2.10') \qquad |S| \leq 2\rho, \qquad P = \rho, \qquad -\beta < z < 0,$$

which is only slightly more restrictive than (2.10) as $\rho \to 1$. In (2.10') the function ρ may be freely chosen provided that ρ is a polynomial of degree m-1 in z with

$$\rho(0) = 1, \qquad \rho'(0) = \sigma_1 - 1.$$

The function ρ will be called the *damping function* of the method and is assumed to be an increasing function in the stability interval $(-\beta,0)$; the maximal deviation from unity will be denoted by ε, i.e. $\varepsilon = 1-\rho(-\beta)$.

For m=2 and m=3 the maximization of β in (2.10') is easily established. Omitting the details we obtain for m=2 (cf.[4]).

$$\sigma_1 = \frac{4-2\varepsilon}{4-3\varepsilon}, \quad \pi_1 = \frac{\varepsilon}{4-3\varepsilon}, \quad \beta = 4-3\varepsilon.$$

The corresponding integration formula reads

$$\begin{aligned} \vec{y}_{n+1} &= \vec{y}_n + h_n \vec{y}'_n + \frac{1}{2}\frac{4-\varepsilon}{4-3\varepsilon} h_n^2 \vec{f}(x_n + \tfrac{1}{2}h_n, \vec{y}_n + \tfrac{1}{2}h_n \vec{y}'_n), \\ \vec{y}'_{n+1} &= \vec{y}'_n + h_n \vec{f}(x_n + \tfrac{1}{2}h_n, \vec{y}_n + \tfrac{1}{2}h_n \vec{y}'_n). \end{aligned} \tag{2.11}$$

The stability condition reads

$$h_n \leq \sqrt{\frac{4-3\varepsilon}{|\delta|_{max}}} \simeq \frac{2-\frac{3}{4}\varepsilon}{\sqrt{|\delta|_{max}}} \text{ as } \varepsilon \to 0. \tag{2.12}$$

The damping function is given by (h_n maximal)

$$\rho = 1 + \frac{\varepsilon}{4-3\varepsilon} z = 1 + \varepsilon \frac{\delta}{|\delta|_{max}}. \tag{2.13}$$

For m=3 we obtain

$$\begin{aligned} \sigma_1 &= 1 + \pi_1, \quad \sigma_2 = \frac{\beta(1+\pi_1) - 2\varepsilon}{\beta^2}, \\ \pi_2 &= \frac{\pi_1\beta - \varepsilon}{\beta^2}, \quad \beta = 8\frac{1 + \sqrt{1-\varepsilon}}{1 + 3\pi_1}, \end{aligned} \tag{2.14}$$

where, apart from ε, the parameter π_1 also is a free parameter. When β is kept fixed we see that ε is maximal when we choose $\pi_1 = 0$. Therefore, it is expected that $\pi_1 = 0$ yields the strongest damping. It happens that this value of π_1 makes the integration formula *second order* accurate as may be seen from (2.7), (2.8) and (1.4). In the following section we will discuss formula (2.14) with $\pi_1 = 0$. Higher point formulas of first order will be discussed in section 2.3.

2.2. Second order formulas

To the conditions (2.2) and (2.8) we add condition (1.4), i.e.

$$\lambda_{m,m-1} = \frac{1}{2}, \tag{2.15}$$

to obtain a second order process. By substituting this value into (2.7), the corresponding Runge-Kutta parameters are expressed into the coefficients σ_j and π_j; note that in the present case where $\sigma_1 = 1$ and $\pi_1 = 0$, the other coefficients are still free.

For m=3 the coefficients σ_2 and π_2 directly follow from (2.14) by putting $\pi_1 = 0$ and $\sigma_1 = 1$:

$$\sigma_2 = \frac{\beta-2\varepsilon}{\beta^2}, \quad \pi_2 = -\frac{\varepsilon}{\beta^2}, \quad \beta = 8(1 + \sqrt{1 - \varepsilon}).$$

The integration formula is generated by

$$(2.16) \qquad (\lambda_{j,l}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_2 - \pi_2 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}, \quad (\beta_j) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad (\mu_j) = \begin{pmatrix} \frac{1}{2}\frac{\sigma_2 + \pi_2}{\sigma_2 - \pi_2} \\ \frac{1}{2} \\ 1 \end{pmatrix}$$

with the stability condition

$$(2.27) \qquad h_n \leq \sqrt{\frac{8(1+\sqrt{1-\varepsilon})}{|\delta|_{max}}} = \frac{4 - \frac{1}{2}\varepsilon}{\sqrt{|\delta|_{max}}} \text{ as } \varepsilon \to 0,$$

and damping function ($h_n = \sqrt{\beta/|\delta|_{max}}$)

$$(2.28) \qquad \rho = 1 - \frac{\varepsilon}{\beta^2} z^2 = 1 - \varepsilon\frac{\delta^2}{|\delta|^2_{max}}.$$

It may be interesting to compare the damping effect of the one-point formula (2.11) with that of formula (2.16). Let us denote the damping functions (2.13) and (2.18) by ρ_1 and ρ_2, respectively, and choose the value of ε in (2.18) equal to $3\varepsilon_{1,1}$ where $\varepsilon_{1,1}$ is the value of ε chosen in (2.13). From (2.12) and (2.17) it then follows that two maximal stable steps of the one-point formula covers the same integration interval as one maximal step of the two-point formula. The damping of the two formulas over this interval is given by

$$\rho_1^2 = (1 + \varepsilon_1 \frac{\delta}{|\delta|_{max}})^2 \text{ and } \rho_2 = 1 - 3\varepsilon_1 \frac{\delta^2}{|\delta|^2_{max}}.$$

It is easily seen that the two-point formula has a slightly stronger damping effect on the higher harmonics than the one-point formula.

For m=4 we only succeeded to solve the minimax problem for small values of ε(cf.[4]); we found

$$\sigma_2 = -\frac{2}{\gamma^2}(6 - \gamma - 3\varepsilon\frac{\gamma^2}{\beta^2}), \quad \sigma_3 = -\frac{1}{\gamma^3}(8 - \gamma - 4\varepsilon\frac{\gamma^3}{\beta^3}),$$

$$\pi_2 = -\frac{3\varepsilon}{\beta^2}, \quad \pi_3 = -2\frac{\varepsilon}{\beta^3}, \quad \beta = 36 - 9\varepsilon, \quad \gamma = 9 + \frac{9}{32}\varepsilon.$$

The corresponding integration formula is defined by the parameter matrices

$$(2.19) \qquad (\lambda_j,\ell) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_3-\pi_3}{\sigma_2-\pi_2} & 0 & 0 \\ 0 & 0 & \sigma_2-\pi_2 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}, (\beta_j) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, (\mu_j) = \begin{pmatrix} \frac{\sigma_3+\pi_3}{2(\sigma_3-\pi_3)} \\ \frac{\sigma_2+\pi_2}{2(\sigma_2-\pi_2)} \\ \frac{1}{2} \\ 1 \end{pmatrix}$$

with the stability condition

$$(2.20) \qquad h_n \leq \frac{6-\frac{3}{4}\varepsilon}{\sqrt{|\delta|_{max}}} \qquad \text{as } \varepsilon \to 0$$

and damping function

$$(2.21) \qquad \rho = 1 - 3\varepsilon \frac{\delta^2}{|\delta|^2_{max}} - 2\varepsilon \frac{\delta^3}{|\delta|^3_{max}}.$$

By comparing (2.17) and (2.20) it is seen that after three maximal steps with the two-point formula and two maximal steps with the three-point formula the same integration interval is covered. A comparison of the damping functions ρ_2^3 and ρ_3^2 of the two-point and three-point formula, respectively, reveals that the damping effect is not improved.

We conclude this section with the obsevation that for small values of ε all formulas derived so far approximately have the same maximal effective step $h_{eff} \doteq 2/\sqrt{|\delta|_{max}}$.

2.3 Modified Rung-Kutta formulas of first and second order

Let J^* be an approximation to the Jacobian matrix of the function $\vec{f}$ at the point $(x_n, \vec{y}_n)$. Instead of method (1.2) we now consider the modified formula

$$
\begin{aligned}
&\vec{y}_{n+1}^{(0)} = \vec{y}_n,\quad \vec{y}_{n+1}^{(j)} = \vec{y}_n + \mu_j h_n \vec{y}_n' + \sum_{\ell=0}^{j-1} \lambda_{j\ell} h^2 J^* \vec{y}_{n+1}^{(\ell)}, \qquad j = 1,2,\ldots m-1,\\
&\vec{y}_{n+1} = \vec{y}_n + \mu_m h_n \vec{y}_n' + \sum_{\ell=0}^{m-1} \lambda_{m\ell} h_n^2 \vec{f}(x_n + \mu_\ell h_n,\ \vec{y}_{n+1}^{(\ell)}), \qquad (2.22)\\
&\vec{y}_{n+1}' = \vec{y}_n' + \sum_{\ell=0}^{m-1} \beta_\ell h_n \vec{f}(x_n + \mu_\ell h_n,\ \vec{y}_{n+1}^{(\ell)}).
\end{aligned}
$$

It is easily verified that the consistency conditions (1.3) and (1.4) for first and second order also apply to this modified scheme. Furthermore, when J^* equals the Jacobian matrix $\partial\vec{f}/\partial\vec{y}$ of $\vec{f}$ at $(x_n, \vec{y}_n)$, method (2.22) has a stability matrix R which is identical to that defined by (1.7). Hence, the modified formula has a similar stability behaviour as the original formula. When J^* differs from $\partial\vec{f}/\partial\vec{y}$ the stability conditions should be carefully applied.

From the above observations it may be concluded that the first and second order formulas derived in the preceding sections, are still legitimate integration formulas when modified in the sense of (2.22). These modified formulas require one evaluation of the right hand side (by virtue of (2.1) and (2.2)) and m-2 evaluations of $J^* \vec{y}_{n+1}^{(\ell)}$. Therefore, it is efficient to use the modified forms of the Runge-Kutta formulas when the evaluation of the vectors $J^* \vec{y}_{n+1}^{(\ell)}$ is cheaper than the evaluation of the right hand side; in particular, for large values of m the gain factor may be very large. This justifies to consider higher point formulas of first and second order. To that end, we have to optimize the polynomial S under the constraints (2.10) or the easier constraints (2.10'). For $m \geq 4$ this problem becomes increasingly more difficult and therefore we replace the constraints (2.10') by still more easy constraints; we look for polynomials S(z) and P(z) which satisfy the conditions

$$
\begin{aligned}
&|S| \le 2\rho, \qquad -\theta \le z < 0,\\
(2.23)\qquad &|S| \le 2(1-\varepsilon), \qquad -\beta \le z \le -\theta,\\
&p = \rho, \qquad -\beta \le z \le 0,
\end{aligned}
$$

for β as large as possible and ρ being the damping function as defined in section 2.1. For small values of ε these constraints are only slightly more restrictive than those of (2.10'). It can be shown that the optimal $S(z)$ satisfying (2.23) is given by (cf.[5, p.90])

$$
(2.24\qquad S(z) = 2\,\frac{T_{m-1}\left(w_0 + \frac{w_0+1}{\beta}\,z\right)}{T_{m-1}(w_0)},
$$

where

$$
\begin{aligned}
&(1-\varepsilon)\ T_{m-1}(w_0) = 1,\\
(2.25)\qquad &\theta = \frac{w_0-1}{w_0+1}\,\beta, \qquad \beta = \frac{2}{\sigma_1}(m-1)\sqrt{\frac{w_0+1}{w_0-1}}\,\tanh\left[(m-1)\ln\left(w_0+\sqrt{w_0^2-1}\right)\right].
\end{aligned}
$$

By choosing for ρ that polynomial which remains as long as possible close to the value $1-\varepsilon$ (optimal damping of the higher frequencies for given ε) we obtain for $P(z)$ the polynomial

$$
(2.26)\qquad P(z) = \rho(z) = 1 - \varepsilon + \left[\varepsilon + \left(\sigma_1 - 1 - \frac{(m-2)\varepsilon}{\beta}\right)z\right]\left[\frac{z+\beta}{\beta}\right]^{m-2}.
$$

In case of first order formulas, both parameters ε and σ_1 are free to select a suitable damping function ρ. By deriving the remaining coefficients σ_2, $\sigma_2,\sigma_3,\ldots,\sigma_{m-1}$ and $\pi_2,\pi_3,\ldots,\pi_{m-1}$ from (2.24),(2.25) and (2.26), the parameter matrices $(\lambda_{j,\ell}),(\beta_j)$ and (μ_j) directly follows from (2.7). In the second order case we have $\sigma_1 = 1$ by virtue of the consistency conditions (2.8) and (2.15). Again after deriving the remaining coefficients, the Runge-Kutta parameters are defined by (2.7).

We conclude this section with the derivation of a second order (m=5)-method of type (2.22) generated by (2.24),(2.25) and (2.26). An elementary calculation yields ($\sigma_1=1,\pi_1=0$)

$$\sigma_2 = 16(6w_0^2-1)(w_0+1)^2(1-\varepsilon)\beta^{-2}, \sigma_3 = 64w_0^2(w_0+1)^3(1-\varepsilon)\beta^{-3},$$

(2.27) $$\sigma_4 = 16(w_0+1)^4(1-\varepsilon)\beta^{-4},$$

$$\pi_2 = -6\varepsilon\beta^{-2}, \ \pi_3 = -8\varepsilon\beta^{-3}, \ \pi_4 = -3\varepsilon\beta^{-4},$$

where

$$\beta = 32w_0(w_0+1)(2w_0^2-1)(1-\varepsilon) \cong 64 - 20\varepsilon$$

and

$$w_0 = \tfrac{1}{2}\left(1+\sqrt{1+\frac{\varepsilon}{2(1-\varepsilon)}}\right) \cong 1 + \frac{1}{8}\varepsilon \ .$$

In terms of the coefficients σ_j and π_j, the integration formula is defined by

(2.28) $$(\lambda_{j\ell}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_4-\pi_4}{\sigma_3-\pi_3} & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma_3-\pi_3}{\sigma_2-\pi_2} & 0 & 0 \\ 0 & 0 & 0 & \sigma_2-\pi_2 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}, \ (\beta_j) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \ (\mu_j) = \tfrac{1}{2}\begin{pmatrix} \frac{\sigma_4+\pi_4}{\sigma_4-\pi_4} \\ \frac{\sigma_3+\pi_3}{\sigma_3-\pi_3} \\ \frac{\sigma_2+\pi_2}{\sigma_2-\pi_2} \\ 1 \\ 2 \end{pmatrix}$$

with the stability condition

(2.29) $$h_n \leq \frac{8-\frac{5}{4}\varepsilon}{\sqrt{|\delta|_{\max}}} \quad \text{as } \varepsilon \to 0$$

and damping function

$$\rho = 1 - \varepsilon + \varepsilon\left(1-3\frac{\delta}{|\delta|_{\max}}\right)\left(1+\frac{\delta}{|\delta|_{\max}}\right)^3.$$

3. NUMERICAL EXPERIMENTS

In this section some of the formulas derived in the proceeding sections are applied to a simple linear, hyperbolic system of the form

$$\frac{d^2\vec{y}}{dt^2} = J\vec{y} + \vec{v}(t), \tag{3.1}$$

where J is a matrix with constant coefficients. This type of equation was chosen in order to illustrate the advantages of the modified formulas when compared with the original ones; the reduction of the computational labour is greater as the evaluation of the vector $J\vec{y} + \vec{v}(t)$ is more expensive than the matrix-vector multiplication $J\vec{y}$.

In particular we have choosen the system

$$\begin{aligned}
\frac{d^2}{dt^2}\, y_0 &= 2g\, d_0 (\underline{\Delta}x)^{-2}(y_1 - y_0) + \frac{1}{4}\lambda^2 y_0 + e^{\frac{1}{2}\lambda t}\, w_0, \\
\frac{d^2}{dt^2}\, y_j &= g\, d_j (\underline{\Delta}x)^{-2}\left(2y_{j+1} - y_j + 2y_{j-1}\right) + \frac{1}{4}\lambda^2 y_j + e^{\frac{1}{2}\lambda t}\, w_j, \\
\frac{d^2}{dt^2}\, y_r &= 2g\, d_r (\underline{\Delta}x)^{-2}(y_{r-1} - y_r) + \frac{1}{4}\lambda^2 y_r + e^{\frac{1}{2}\lambda t}\, w_r,
\end{aligned} \tag{3.1}$$

where $j = 1,2,\ldots,r-1$. This system is derived from the partial differential equations describing the water elevation at the points $j \underline{\Delta} x$ in a river of length $r \underline{\Delta} x$; the depth and the wind field in these points are given by d_j and w_j, respectively; furthermore, λ and g denote the friction coefficient of the bottom and the accelaration of gravity. The following specifications were used

$$y_j(0) = \frac{dy_j}{dt}(0) = 0, \quad j = 0,\ldots,r,$$

$\underline{\Delta}x = 10\ 000$ and $\underline{\Delta}x = 1000$, respectively

$r = 100\ 000/\Delta x,$

$$
\begin{aligned}
&d_j = 10(2 + \cos(2\pi j\Delta x 10^{-5}),\\
&w_j = 10^{-3}\sin(\pi j\Delta x 10^{-5}),\\
&\lambda = .000025,\\
&g = 9.81.
\end{aligned}
\tag{3.2}
$$

Although in this test problem, the computational effort to compute the vector $\vec{v}(t) = \exp(\frac{1}{2}\lambda t)\,\vec{w}$ is relatively small, it serves its purpose to compare the modified and unmodified formulas.

In order to illustrate the increased efficiency of the new formulas with respect to Runge-Kutta formulas for first order equations, we also integrated the first order form of system (3.1') by a few stabilized Runge-Kutta formulas for first order systems of hyperbolic type. Writing the first order equations in the general form

$$\frac{d\vec{y}}{dx} = \vec{F}(x,\vec{y}), \tag{3.3}$$

the Runge-Kutta formulas we used, are given by (cf.[5])

$$\vec{y}_{n+1} = \vec{y}_n + h_n\vec{F}(x_n + h_n,\ \vec{y}_n + h_n\vec{F}(x_n,\ \vec{y}_n)) \tag{3.4}$$

with the stability condition

$$h_n \le \frac{1}{|\lambda|_{max}}, \tag{3.5}$$

$$\vec{y}_{n+1} = \vec{y}_n + h_n\vec{F}(x_n + \tfrac{1}{2}h_n,\ \vec{y}_n + \tfrac{1}{2}h_n\vec{F}(x_n + \tfrac{1}{2}h_n,\ \vec{y}_n + \tfrac{1}{2}h_n\vec{F}(x_n,\vec{y}_n))), \tag{3.6}$$

with the stability condition

$$h_n \le \frac{2}{|\lambda|_{max}}, \tag{3.7}$$

and finally, the Runge-Kutta formula

$$
\begin{aligned}
\vec{y}^{(1)}_{n+1} &= \vec{y}_n + \tfrac{1}{6}h_n\vec{F}(x_n, \vec{y}_n),\\
\vec{y}^{(2)}_{n+1} &= \vec{y}_n + \tfrac{1}{12}h_n\vec{F}(x_n + \tfrac{1}{6}h_n, \vec{y}^{(1)}_{n+1}),\\
\vec{y}^{(3)}_{n+1} &= \vec{y}_n + \tfrac{2}{9}h_n\vec{F}(x_n + \tfrac{1}{12}h_n, \vec{y}^{(2)}_{n+1}),\\
\vec{y}^{(4)}_{n+1} &= \vec{y}_n + \tfrac{4}{19}h_n\vec{F}(x_n + \tfrac{2}{9}h_n, \vec{y}^{(3)}_{n+1}),\\
\vec{y}^{(5)}_{n+1} &= \vec{y}_n + \tfrac{19}{54}h_n\vec{F}(x_n + \tfrac{4}{19}h_n, \vec{y}^{(4)}_{n+1}),\\
\vec{y}^{(6)}_{n+1} &= \vec{y}_n + \tfrac{1}{2}h_n\vec{F}(x_n + \tfrac{19}{54}h_n, \vec{y}^{(5)}_{n+1}),\\
\vec{y}_{n+1} &= \vec{y}_n + h_n\vec{F}(x_n + \tfrac{1}{2}h_n, \vec{y}^{(6)}_{n+1}),
\end{aligned}
\tag{3.8}
$$

with the stability condition

$$
h_n \le \frac{6}{|\lambda|_{\max}}. \tag{3.9}
$$

In the stability conditions (3.5), (3.7) and (3.9), $|\lambda|_{\max}$ denotes the spectral radius of the Jacobian matrix $\partial\vec{F}/\partial\vec{y}$ of the right hand side $\vec{F}$, where it is assumed that the eigenvalues of $\partial\vec{F}/\partial\vec{y}$ are imaginary (note that $|\lambda|_{\max} = \sqrt{|\delta|_{\max}}$ when $\vec{F}$ corresponds to the first order form of equation (1.1)).

Formula (3.4) is first order accurate, (3.6) and (3.8) are both second order accurate. This also holds for the modified forms of these formulas, that is when in (3.4) and (3.6) the vector $\vec{F}(x_n, \vec{y}_n)$ is replaced by $K\vec{y}_n$, K being some approximation to the Jacobian matrix $\partial\vec{F}/\partial\vec{y}$ at the point $(x_n, \vec{y}_n)$, and when in (3.8) the formulas for the vectors $\vec{y}^{(1)}_{n+1},\ldots,\vec{y}^{(5)}_{n+1}$ are modified in the same sense.

Table 3.1 Number of correct significant digits (*sd*) and number of right hand side evaluations (*fev*) for several stabilized Runge-Kutta formulas and their modified forms in the case $\Delta x = 10^4$

fev \ sd	0.7	0.8	1.3	1.8	1.9	2.1	2.2	2.3	2.4	3.1	3.2
3	(2.19*)										
4			(2.16*)								
6	(3.8*)			(2.11)							
7						(2.19*)	(2.16*)		(2.16)		
8											
9							(2.19)				
11											
13						(2.11)					
14										(2.16)	
21					(3.8)			(3.6)			(2.19)
26		(3.4)									

In order to compare the efficiency of the various formulas, we have arranged them in an (accuracy-computational effort) - diagram (see tables 3.1 and 3.2), that is the pairs (*sd*, *fev*), *sd* being the number of correct significant digits and *fev* the number of right hand side evaluations involved, are indicated in a diagram by the reference number of the corresponding formula. Tables 3.1 and 3.2 present the results for the respective cases $\Delta x = 10000$ and $\Delta x = 1000$ at $t = 3600$. The number of correct digits was determined by using the numerical values produced by a higher order Runge-Kutta method wiht extreme small step sizes ($\Delta t = 100$ for $\Delta x = 10000$ and $\Delta t = 10$ for $\Delta x = 1000$). The results of the modified formulas are indicated by adding an asterix to the reference number of the corresponding unmodified formula. All formulas were applied with the maximal stable integration step. Moreover, the new formulas (2.11), (2.16) and (2.19) are also applied with the integration step used by the formulas (3.4) and (3.6), respectively.

Table 3.2 Number of correct significant digits (*sd*) and number of right hand side evaluations (*fev*) for several stabilized Runge-Kutta formulas and their modified forms in the case $\Delta x = 10^3$

fev \ sd	1.8	2.8	3.1	3.6	4.1	4.2	4.3	4.4	5.0	5.3
21			(2.19*)							
32				(2.16*)						
42		(3.8*)								
65		(2.11)			(2.19*)	(2.16*)	(2.19)	(2.16)		
124			(2.11)						(2.16)	
148						(3.8)				
186						(3.6)				(2.19)
248	(3.4)									

From the tables 3.1 and 3.2 the superiority of the formulas (2.11), (2.16) and (2.19) is evident. Furthermore, we see that the modified formulas make it possible to find a less accurate solution for considerable less computational effort.

REFERENCES

[1] R. ANSORGE, R. and W. TORNIG, *Zür Stabilität des Nyström̈schen Verfahrens*, ZAMM 40(1960), pp. 368-370.

[2] W.J. GERRITSEN, *Experiments with stabilized Runge-Kutta methods for second order differetial equations without first derivatives* (to appear in the NW-series of the Mathematisch Centrum, Amsterdam).

[3] P. HENRICI. *Discrete variable methods in ordinary differential equations*. John Wiley, New York, 1962.

[4] P.J. VAN DER HOUWEN, *Stabilized Runge-Kutta methods for second order differential equations without first derivatives*, Report NW26 Mathematisch Centrum, Amsterdam, 1975.

[5] - . *Construction of integration formulas for initial value problems*. North-Holland Publishing Company, Amsterdam, 1977.