stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

NW 74/79 NOVEMBER

B.P. SOMMEIJER & P.J. VAN DER HOUWEN

ON THE ECONOMIZATION OF STABILIZED RUNGE-KUTTA METHODS WITH APPLICATIONS TO PARABOLIC INITIAL VALUE PROBLEMS

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).

1980 Mathematics subject classification: 65L05, 65L07, 65Mxx

On the economization of stabilized Runge-Kutta methods with applications to parabolic initial value problems *)

by

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

ABSTRACT

A modification of Runge-Kutta methods is analysed which leads for important classes of parabolic differential equations to a considerable reduction of the computational effort. The main characteristic of the modified methods is the replacement of the right hand side function of the differential equation by a "cheaper" function with roughly the same Jacobian matrix. Numerical experiments are reported and the results are compared with the results obtained by the unmodified Runge-Kutta method and by an ADI method.

KEY WORDS & PHRASES: *Numerical analysis, Runge-Kutta methods, internal stability.*

*) This report will be submitted for publication elsewhere.

1. INTRODUCTION

Suppose we are given a time-dependent partial differential equation defined on a domain Ω with boundary $\partial\Omega$ in the space of the space variable $\vec{x}$. By applying the method of lines, that is by replacing $\Omega \cup \partial\Omega$ by a set of grid points $\Gamma_h \cup \partial\Gamma_h$, h referring to the coarseness of the grid, such a partial differential equation (p.d.e.) is converted into a system of ordinary differential equations which frequently is of the form

$$\begin{aligned} \frac{d\vec{y}}{dt} &= \vec{f}(t,\vec{y}+\vec{b}) \\ \vec{b} &= \vec{g}(t,\vec{y}) \end{aligned} \qquad (1.1)$$

Here, $\vec{y}$ denotes an approximation to the solution of the p.d.e. at the internal grid points Γ_h, being zero at the boundary points $\partial\Gamma_h$, and $\vec{b}$ denotes an approximation to the solution at the boundary points $\partial\Gamma_h$ being zero at Γ_h. The functions $\vec{f}$ and $\vec{g}$ are given having zero-components at $\partial\Gamma_h$ and Γ_h, respectively. The function $\vec{g}$ originates from the boundary conditions and expresses the boundary values in terms of the solution at the internal grid points Γ_h.

In this paper we study a modification of *stabilized Runge-Kutta methods* (RK methods) for the solution of the initial value problem for equation (1.1). We shall call an RK method stabilized when extra function evaluations are added in order to increase the stability boundary. Large stability boundaries are desirable when the spectral radius of the Jacobian of $\vec{f}$ with respect to $\vec{y}$ is large which is just the situation when (1.1) originates from a p.d.e.. In the case of parabolic p.d.e.'s, stabilized RK methods have been constructed [3] with a (real) stability boundary $\beta \cong c_p m^2$ where m is the number of function evaluations per step and c_p is a constant depending on the order p of the RK method ($c_1 \cong 1.93$, $c_2 \cong .80$ [3]). The maximal stable integration step of these methods is given by

$$\tau_{max} = \frac{\beta}{\sigma(J_n)} \cong \frac{c_p m^2}{\sigma(J_n)} \quad , \qquad (1.2)$$

where $\sigma(J_n)$ denotes the spectral radius of the Jacobian matrix

$$(1.3) \qquad J_n = \left.\frac{\partial \vec{f}(t_n, \vec{y}+\vec{g}(t_n,\vec{y}))}{\partial \vec{y}}\right|_{\vec{y}=\vec{y}_n},$$

$\vec{y}_n$ denoting the numerical approximation to $\vec{y}$ at $t = t_n$ (note that in (1.3) $\vec{f}$ is only differentiated with respect to those components of $\vec{y}$ which correspond to the gridpoints of Γ_h). Although stabilized RK methods are available (for semi-discrete parabolic equations) in which the number of stages can be chosen arbitrarily large without danger for internal instabilities [3], these methods generally are more expensive in terms of right hand side evaluations than implicit or partial implicit integration methods. It is often possible, however, to reduce the work per integration step by exploiting the fact that a stabilized RK method usually contains a *small* number of $\vec{f}$-evaluations which ensure the order of accuracy and a *large* number of $\vec{f}$-evaluations which take care of the stability of the scheme. This means that these last $\vec{f}$-evaluations may be replaced by a function $\vec{f}^*(t,\vec{y})$ without affecting the *order* of accuracy. In order to obtain the same stability regions one should choose $\vec{f}^*$ such that its Jacobian matrix is a first order approximation in τ to the original Jacobian matrix. When the effort involved to compute the set of $\vec{f}^*$-vectors within an integration step is much less than the computational effort involved to evaluate the replaced $\vec{f}$-vectors, then computing time can be saved by this simple modification of the RK method. We call such methods *modified Runge-Kutta methods*. As an example, consider the function

$$(1.4) \qquad \vec{f}^*(t,\vec{y}) = J_n\vec{y}, \qquad t_n \le t \le t_{n+1},$$

which is a plausible choice in cases of linear problems where J_n is directly available. In [2] a few experiments were performed with this modification. It turned out that stability is preserved but the accuracy is considerably reduced. Here, we consider more general, and at the same time more accurate, $\vec{f}^*$-functions and investigate the effect on the accuracy of the numerical solution. Numerical experiments are reported and the results are compared with those obtained by the ADI method of Peaceman-Rachford. The modified Runge-Kutta method seems to be particularly advantageous for strongly non-

linear problems and if the costs per integration step are about 25% (according to our experiments) of the costs of the classical Runge-Kutta step.

2. MODIFIED RUNGE-KUTTA METHODS

Consider the m-stage Runge-Kutta formula

$$
\begin{aligned}
&\vec{y}_{n+1}^{(0)} = \vec{y}_n, \\
&\vec{y}_{n+1}^{(j)} = \vec{y}_n + \tau \sum_{\ell=0}^{j-1} \lambda_{j\ell} \vec{f}^{**}(t_n+\theta_\ell\tau, \vec{y}_{n+1}^{(\ell)}), \qquad j = 1,2,\ldots,m, \\
&\vec{y}_{n+1} = \vec{y}_{n+1}^{(m)};\ \theta_0 = 0,\ \theta_j = \sum_{\ell=0}^{j-1} \lambda_{j\ell}, \qquad j = 1,2,\ldots,m-1.
\end{aligned} \tag{2.1}
$$

If the function $\vec{f}^{**}(t,\vec{y})$ is identical to the right hand side function $\vec{f}(t,\vec{y}+\vec{g}(t,\vec{y}))$, this scheme represents a classical Runge-Kutta method for equation (1.1). For the parameters $\lambda_{j\ell}$ we choose the values which define an appropriate classical stabilized RK method (cf. [3]). A few important schemes are specified in section 3.2. These are characterized by their limited storage requirements, an important feature in view of our purpose to apply them to the large systems arising from the space-discretization of partial differential equations (see also [4]). The formula (2.1) together with a prescribed function $\vec{f}^{**}$ define a *modified Runge-Kutta method (MRK method)*.

In this section the conditions are derived under which the modified formula has the same stability region and order of accuracy as the generating classical RK formula.

2.1. Stability

The stability region of the MRK method (2.1) is identical to that of the generating RK method because the stability polynomials are identical (the stability polynomial $R_m(z)$ is defined by the relation $y_{n+1} = R_m(\tau\lambda)y_n$ obtained by applying the method to the equation $y' = \lambda y$). In our case this polynomial is generated by the recurrence relation

$$R_0(z) = 1,$$

$$R_j(z) = 1 + \sum_{\ell=0}^{j-1} \lambda_{j\ell} z R_\ell(z), \qquad j = 1,2,\ldots,m. \tag{2.2}$$

When $J_n^* = \partial \vec{f}^{**}/\partial \vec{y}$ has negative eigenvalues λ then the modified scheme is called stable if (cf. (1.2))

$$\tau \le \frac{\beta}{\sigma(J_n^*)},$$

where β is the real stability boundary of $R_m(z)$. A p-th order RK method has a stability polynomial of the form

$$R_m(z) = 1 + z + \frac{1}{2}z^2 + \ldots + \frac{1}{p!}z^p + \beta_{p+1}z^{p+1} + \ldots + \beta_m z^m, \tag{2.3}$$

where the coefficients $\beta_{p+1},\ldots,\beta_m$ are expressions in the RK parameters. For instance,

$$\beta_1 = \sum_{\ell=0}^{m-1} \lambda_{m\ell}, \quad \beta_2 = \sum_{j=1}^{m-1} \lambda_{mj} \sum_{\ell=0}^{j-1} \lambda_{j\ell} = \sum_{j=1}^{m-1} \lambda_{mj}\theta_j. \tag{2.4}$$

Polynomials of the form (2.3) will be called p-*th order consistent*.

In this paper, we will use the first order consistent polynomial [3]

$$\tilde{R}_m^{(1)}(z) = \frac{T_m\left(w_0 + \frac{w_0+1}{\beta} z\right)}{T_m(w_0)}, \quad \beta = \frac{(w_0+1)T_m'(w_0)}{T_m(w_0)}, \qquad w_0 > 1 \tag{2.5}$$

and the second order consistent polynomial

$$\tilde{A}_m^{(2)}(z) = a + bT_m\left(w_0 + \frac{w_0+1}{\beta} z\right), \quad \beta = \frac{(w_0+1)T_m''(w_0)}{T_m'(w_0)}, \qquad w_0 > 1, \tag{2.6}$$

where

$$a = 1 - bT_m(w_0), \quad b = \frac{\varepsilon}{T_m(w_0)-1}, \quad \varepsilon = \frac{T_m''(w_0)[T_m(w_0)-1]}{[T_m'(w_0)]^2}.$$

In (2.5) and (2.6) T_m denotes the Chebyshev polynomial of degree m.

2.2. The order of accuracy of modified Runge-Kutta methods

Obviously, the function $\vec{f}^{*}$ in (2.1) should have some relation to the right hand side function $\vec{f}$ in order to present a consistent approximation to the equation (1.1). It will be assumed that in the interval $t_n \le t \le t_n + \tau$ this relation is of the form

$$\vec{f}^{*}(t_n + \theta_\ell \tau, \vec{y}_{n+1}^{(\ell)}) - \vec{f}(t_n + \theta_\ell \tau, \vec{y}_{n+1}^{(\ell)} + \vec{g}(t_n + \theta_\ell \tau, \vec{y}_{n+1}^{(\ell)})) = \delta_\ell \tau^q \vec{\phi}_n + \eta_\ell \tau^s \vec{\psi}_n + O(\tau^{q+1} + \tau^{s+1}), \tag{2.7}$$

where δ_ℓ, η_ℓ are scalars, q, s integers ≥ 1 and $\vec{\phi}_n$, $\vec{\psi}_n$ vectors only depending on $(t_n, \vec{y}_n)$. Evidently, $\vec{\phi}_n = \vec{\psi}_n = \vec{0}$ if $\vec{f}^{*} = \vec{f}$. In Section 2.3 examples are given of functions $\vec{f}^{*}$ satisfying (2.7). Here, we first derive the order equations.

Starting with a p-th order classical RK formula our analysis can be confined to the derivation of an expression for the difference $\vec{y}_{n+1} - \tilde{\vec{y}}_{n+1}$, where $\tilde{\vec{y}}_{n+1}$ denotes the result which would be obtained if the classical RK method is applied at $t = t_n$. Let $\tilde{\vec{y}}_{n+1}^{(j)}$ denote the intermediate vectors obtained with the classical RK method. From (2.1) and (2.7) it follows that the intermediate deviations

$$\Delta\vec{y}_{n+1}^{(j)} = \vec{y}_{n+1}^{(j)} - \tilde{\vec{y}}_{n+1}^{(j)}$$

are approximately determined by the scheme

$$\Delta\vec{y}_{n+1}^{(0)} = \vec{0},$$

$$\Delta\vec{y}_{n+1}^{(j)} = \tau \sum_{\ell=0}^{j-1} \lambda_{j\ell}[J_n \Delta\vec{y}_{n+1}^{(\ell)} + \tau^q \delta_\ell \vec{\phi}_n + \tau^s \eta_\ell \vec{\psi}_n + O(\tau^{q+1} + \tau^{s+1})], \qquad j = 1,2,\ldots,m,$$

$$\vec{y}_{n+1} - \tilde{\vec{y}}_{n+1}^{(j)} = \Delta\vec{y}_{n+1}^{(m)}.$$

From these relations it follows that the *local deviation error* is

$$\vec{y}_{n+1} - \tilde{\vec{y}}_{n+1} = \tau^{q+1} Q_m(\tau J_n)[\vec{\phi}_n + \mathcal{O}(\tau)] + \tau^{s+1} S_m(\tau J_n)[\vec{\psi}_n + \mathcal{O}(\tau)], \tag{2.8}$$

where Q_m, S_m are polynomials in τJ_n of degree m-1 defined by the recurrence relations

$$\begin{aligned} Q_0(z) &= 0, \quad Q_j(z) = \sum_{\ell=0}^{j-1} \lambda_{j\ell}[zQ_\ell(z) + \delta_\ell] \\ &\qquad\qquad\qquad\qquad\qquad\qquad j = 1,2,\ldots,m. \\ S_0(z) &= 0, \quad S_j(z) = \sum_{\ell=0}^{j-1} \lambda_{j\ell}[zS_\ell(z) + \eta_\ell] \end{aligned} \tag{2.9}$$

Hence,

$$Q_m(z) = \sum_{\ell=0}^{m-1} \lambda_{m\ell}\delta_\ell + \sum_{j=1}^{m-1} \sum_{\ell=0}^{j-1} \lambda_{mj}\lambda_{j\ell}\delta_\ell z + \ldots \tag{2.10}$$

and a similar expression for $S_m(z)$.

From (2.8) it is immediate that a p-th order RK method generates an MRK method of order p if either

$$\min(q,s) = p \tag{2.11}$$

or

$$\min(q,s) < p,$$

$$\begin{aligned} &\frac{d^k}{dz^k} Q_m(z)\Big|_{z=0} \cdot \vec{\phi}_n = \vec{0}, \qquad k = 0,\ldots,p-1-q, \\ &\frac{d^\ell}{dz^\ell} S_m(z)\Big|_{z=0} \cdot \vec{\psi}_n = \vec{0}, \qquad \ell = 0,\ldots,p-1-s. \end{aligned} \tag{2.11'}$$

2.3. Modified right hand side functions

In this section we consider a few possibilities for choosing the modified right hand side function $\vec{f}^*(t,\vec{y})$. The order of the modification will be defined by min(q,s) (cf. (2.8)).

2.3.1. First order modifications

The first class of modified right hand side functions is of the form

$$\vec{f}^{**}(t,\vec{y}) = F(\vec{t}_n + \theta\tau, t, \vec{y}_n, \vec{y}), \qquad t_n \le t \le t_{n+1}, \tag{2.12}$$

where θ is a parameter in the interval $[0,1]$ and $\vec{F}$ a function satisfying the condition

$$\vec{F}(t,t,\vec{y},\vec{y}) = \vec{f}(t,\vec{y}+\vec{g}(t,\vec{y})). \tag{2.13}$$

EXAMPLES

(i) Simple examples of modifications satisfying (2.12) and (2.13) are defined by

$$\vec{F}(t^*,t,\vec{y}^{**},\vec{y}) = \vec{f}(t^*,\vec{y}+\vec{g}(t^*,\vec{y})), \tag{2.12a}$$

$$\vec{F}(t^*,t,\vec{y}^{**},\vec{y}) = \vec{f}(t^*,\vec{y}+\vec{g}(t,\vec{y})). \tag{2.12b}$$

These functions are of use if the time-dependency in $\vec{f}$ forms the main part of the computational effort in the evaluation of $\vec{f}$. For instance, in the case of a separable right hand side function $\vec{f}$, i.e.

$$\vec{f}(t,\vec{y}+\vec{g}(t,\vec{y})) = \sum_{i=1}^{r} T_i(t)\vec{Y}_i(\vec{y}+\vec{g}(t,\vec{y})), \tag{2.14}$$

T_i being matrices only depending on t and $\vec{Y}_i$ being vector functions depending on $\vec{y}$ and t, the functions (2.12a) and (2.12b) are suitable economizations if the matrices T_i require the greater part of the computational effort to evaluate $\vec{f}$. It should be remarked however, that generally (2.12b) yields considerably more accurate results than (2.12a). This is due to the equal levels of consistency of the g and y fields at all places where they simultaneously appear ([1], see also the experiments in Section 3.3).

(ii) When in (2.14) the matrices T_i also depend on $\vec{y}$, i.e.

$$\vec{f}(t,\vec{y}+\vec{g}(t,\vec{y})) = \sum_{i=1}^{r} T_i(t,\vec{y})\vec{Y}_i(\vec{y}+\vec{g}(t,\vec{y})), \tag{2.14'}$$

we may define

$$\vec{F}(t^*,t,\vec{y}^{**},\vec{y}) = \sum_{i=1}^{r} T_i(t^*,\vec{y}^{**})\vec{Y}_i(\vec{y}+\vec{g}(t,\vec{y})). \tag{2.12c}$$

(iii) As a last example we consider a modification based on the Jacobian matrix of $\vec{f}$:

$$\vec{F}(t^*,t,\vec{y}^{**},\vec{y}) = \vec{f}(t^*,\vec{y}^{**}+\vec{g}(t_n,\vec{y}^{**})) + K_\theta[\vec{y}+\vec{g}(t,\vec{y})-\vec{y}^{**}-\vec{g}(t_n,\vec{y}^{**})],$$

$$0 \le \theta \le 1,\ t_n \le t \le t_{n+1}. \tag{2.12d}$$

$$K_\theta = \left.\frac{\partial\vec{f}(t^*,\vec{v})}{\partial\vec{v}}\right|_{\vec{v}=\vec{y}^{**}+\vec{g}(t_n,\vec{y}^{**})}$$

This function is sort of linearization of $\vec{f}$ in the interval $[t_n,t_{n+1}]$ and is of practical value if the Jacobian K_θ is easily obtained. Several other linearizations are possible but (2.12d) was chosen because of its property to reduce to (2.12b) for right hand side functions of the form $A(t)(\vec{y}+\vec{g}(t,\vec{y}))$, $A(t)$ being a matrix operator. It should be remarked that similarly to (2.12b) the t-argument of the boundary function $\vec{g}$ occurring in (2.12d) is chosen according to the time-level of the corresponding $\vec{y}$-field. As already observed above the accuracy is improved by this choice. □

The function (2.12) satisfies condition (2.7) with

$$q = 1,\ \delta_\ell = \theta - \theta_\ell,\ \vec{\phi}_n = \left.\frac{\partial\vec{F}(t,t_n,\vec{y}_n,\vec{y}_n)}{\partial t}\right|_{t=t_n}$$

$$s = 1,\ \eta_\ell = -\theta_\ell,\ \vec{\psi}_n = \left.\frac{\partial\vec{F}(t_n,t_n,\vec{y},\vec{y}_n)}{\partial\vec{y}}\right|_{\vec{y}=\vec{y}_n} \vec{f}^{**}(t_n,\vec{y}_n) \tag{2.15}$$

Thus, (2.12) and in particular (2.12a,...,d), represents a class of first order modifications. Let p be the order of the generating RK formula, then the following theorem is immediate from (2.11), (2.15) and (2.4):

THEOREM 2.1.

(a) *The method* {(2.1); (2.12)} *is of first order for all* θ *if* $p = 1$.

(b) *The methods* {(2.1); (2.12a), (2.12b), (2.12d)} *are of second order if* $p = 2$ *and*

$$\theta = \frac{\sum_{\ell=1}^{m-1} \lambda_{m\ell}\theta_\ell}{\sum_{\ell=0}^{m-1} \lambda_{m\ell}} = \frac{\beta_2}{\beta_1} = \frac{1}{2}. \qquad \square \tag{2.16}$$

2.3.2. Second order modifications

Consider the modification

$$\vec{f}^*(t,\vec{y}) = \alpha(t)\vec{F}(t_n+\bar{\theta}\tau,t,\vec{y}_n,\vec{y}) + (1-\alpha(t))\vec{F}(t_n+\bar{\bar{\theta}}\tau,t,\vec{y}_n,\vec{y}),$$

$$\alpha(t) = \frac{t-t_n-\bar{\bar{\theta}}\tau}{(\bar{\theta}-\bar{\bar{\theta}})\tau}, \qquad 0 \le \bar{\theta} < \bar{\bar{\theta}} < 1, \tag{2.17}$$

where $\vec{F}$ satisfies (2.13). For $\vec{F}$ one may choose e.g. the function defined by (2.12a),...,(2.12d). These modifications will be indicated by (2.17a),..., (2.17d), respectively. A straightforward calculation yields that (2.7) is satisfied with

$$q = 2,\ \delta_\ell = \tfrac{1}{2}(\bar{\theta}-\theta_\ell)(\theta_\ell-\bar{\bar{\theta}}),\ \vec{\phi}_n = \left.\frac{\partial^2 F(t,t_n,\vec{y}_n,\vec{y}_n)}{\partial t^2}\right|_{t=t_n}$$

$$s = 1,\ \eta_\ell = -\theta_\ell,\ \vec{\psi}_n = \left.\frac{\partial\vec{F}(t_n,t_n,\vec{y},\vec{y}_n)}{\partial\vec{y}}\right|_{y=y_n} \vec{f}^*(t_n,\vec{y}_n) \tag{2.18}$$

A second order modification is obtained if the Jacobian matrix $L_n = \partial\vec{F}(t_n,t_n,\vec{y},\vec{y}_n)/\partial\vec{y}$ vanishes. We then have the following theorem:

THEOREM 2.2.

(a) *The method* {(2.1); (2.17)} *is of second order for all* $\bar{\theta}$ *and* $\bar{\bar{\theta}}$ *if* $p = 2$ *and if the Jacobian matrix* L_n *vanishes.*

(b) *The methods* {(2.1); (2.17)} *is of third order if* $p = 3$, *the matrix* L_n *vanishes and if* $\bar{\theta}$ *and* $\bar{\bar{\theta}}$ *satisfy the equation*

$$\bar{\theta}\bar{\bar{\theta}} - \frac{1}{2}(\bar{\theta}+\bar{\bar{\theta}}) + \frac{1}{3} = 0. \qquad \square \tag{2.19}$$

PROOF. Part (a) is immediate from (2.11) and (2.18). Part (b) is proved by applying (2.11'), i.e. $\bar{\theta}$ and $\bar{\bar{\theta}}$ have to satisfy the equation

$$\sum_{\ell=0}^{m-1} \lambda_{m\ell}(\bar{\theta} - \theta_\ell)(\bar{\bar{\theta}} - \theta_\ell) = 0.$$

Since every third order RK method satisfies the conditions

$$\sum_{\ell=0}^{m-1} \lambda_{m\ell} = 1, \quad \sum_{\ell=1}^{m-1} \lambda_{m\ell}\theta_\ell = \frac{1}{2}, \quad \sum_{\ell=1}^{m-1} \lambda_{m\ell}\theta_\ell^2 = \frac{1}{3}$$

we arrive at equation (2.19). □

REMARKS

(i) The application of the method {(2.1); (2.17)} with *non-vanishing* Jacobian matrix L_n is certainly possible and will be more accurate as $\vec{F}(t_n, t_n, \vec{y}, \vec{y}_n)$ is smoother with respect to $\vec{y}$.

(ii) Suitable choices of the parameters $\bar{\theta}$ and $\bar{\bar{\theta}}$ are

Case (a): $\bar{\theta} = 0; \ \bar{\bar{\theta}} = 1$

Case (b):
$$\begin{cases} \bar{\theta} = 0; \ \bar{\bar{\theta}} = \frac{2}{3} & \text{for n even} \\ \bar{\theta} = \frac{1}{3}; \ \bar{\bar{\theta}} = 1 & \text{for n odd} \end{cases}.$$

It is easily seen that these parameter values save computational effort when we are dealing with separable right hand side functions.

2.4. The difference between the classical and modified Runge-Kutta solution

In the preceding section we derived the conditions for p-th order accuracy of the modified RK formula. In actual computations, however, the MRK methods usually are less accurate than the generating RK methods (of the same order). The reason is that the error (2.8) is generally larger than the local trucation error of the generating formula although the orders in τ are equal. Therefore, it is of interest to see how the polynomials $Q_m(z)$ and $S_m(z)$ behave because these polynomials determine largely the magnitude of the deviation (2.8). Considering modifications which satisfy (2.15), we can express the polynomials $Q_m(z)$ and $S_m(z)$ in terms of the stability polynomial $R_m(z)$.

THEOREM 2.3. *If* (2.15) *is satisfied then*

$$Q_m(z) = \theta \frac{R_m(z)-1}{z} - \frac{R_m(z)-1-\theta_m z}{z^2},$$

(2.20)

$$S_m(z) = -\frac{R_m(z)-1-\theta_m z}{z^2}. \qquad \square$$

PROOF. Substitution of $\delta_\ell = \theta - \theta_\ell$ into (2.9) yields

$$Q_0(z) = 0, \quad Q_j(z) = \theta \sum_{\ell=0}^{j-1} \lambda_{j\ell} + \sum_{\ell=0}^{j-1} \lambda_{j\ell}[zQ_\ell(z) - \theta_\ell].$$

Let us define the polynomials $A_j(z)$ and $B_j(z)$ by writing

$$Q_j(z) = \theta A_j(z) - B_j(z).$$

These polynomials satisfy the recurrence relations

$$A_0(z) = 0, \quad A_1(z) = \lambda_{10}, \quad A_j(z) = \sum_{\ell=0}^{j-1} \lambda_{j\ell}(1 + zA_\ell(z)), \qquad j = 2,3,\ldots,m. \tag{2.21}$$

$$B_0(z) = B_1(z) = 0, \quad B_j(z) = \sum_{\ell=0}^{j-1} \lambda_{j\ell}(\theta_\ell + zB_\ell(z)), \tag{2.22}$$

It is easily verified that

$$A_j(z) = \frac{R_j(z)-1}{z} \quad \text{and} \quad B_j(z) = \frac{R_j(z)-1-\theta_j z}{z^2} \tag{2.23}$$

satisfy (2.21) and (2.22), respectively, provided that $R_j(z)$ is defined by (2.2). From (2.23) the expression (2.20) for $Q_m(z)$ follows. In a similar way we find the expression for $S_m(z)$ as given by (2.20). $\square$

From this theorem it follows that

$$\frac{1}{z} \le Q_m(z) \le \frac{2+(1-2\theta)z}{z^2}, \qquad \frac{1}{z} \le S_m(z) \le \frac{2+z}{z^2}, \qquad -\beta \le z \le 0, \tag{2.24}$$

where we have put $\theta_m = 1$, that is we assume the formula at least first order exact. These inequalities show that $|Q_m(z)|$ and $|S_m(z)|$ are small for large $|z|$-values so that we concentrate on their behaviour (relatively) close to the origin. Again by theorem 2.3 we have

$$(2.25)\qquad Q_m(z) \cong (\theta - \beta_2) + (\theta\beta_2 - \beta_3)z + (\theta\beta_3 - \beta_4)z^2 + (\theta\beta_4 - \beta_5)z^3$$

for sufficiently small $|z|$-values. The behaviour of $S_m(z)$ follows from this approximation by putting $\theta = 0$. In our case where we choose for $R_m(z)$ either $\tilde{R}_m^{(1)}(z)$ or $\tilde{A}_m^{(2)}(z)$ as defined by (2.5) and (2.6), respectively, the coefficients $\beta_2,\ldots,\beta_5$ can be shown to be almost independent of m and hence the polynomials $Q_m(z)$ and $S_m(z)$ have for all m a more or less identical behaviour near the origin. For $\theta = \beta_2$ the polynomial $|Q_m(z)|$ assumes its maximum value .07 at $z = -10$ and .20 at $z = -3.5$ in the respective cases $\tilde{R}_m^{(1)}(z)$ and $\tilde{A}_m^{(2)}(z)$. The polynomial $|S_m(z)|$ assumes its maximum value β_2 at the origin. Thus, it is expected that *first order methods will loose less accuracy when modified than second order methods*.

2.5. Boundary conditions

The last aspect of scheme (2.1) to be discussed in this paper concerns the boundary conditions $\vec{b} = \vec{g}(t,\vec{y})$. It is well-known that in algorithms for time-dependent p.d.e.'s with intermediate stages such as RK methods or splitting methods [1], the boundary values and the internal solution values should form a sufficiently smooth grid function on $\Gamma_h \cup \partial\Gamma_h$ as soon as they simultaneously appear as argument of the function $\vec{f}$. This requirement restricts the choice of the functions $\vec{f}$ and is the reason for choosing the t-arguments of $\vec{g}(t,\vec{y})$ as done in (2.12b,c,d). We will illustrate this by the following heuristic analysis of what happens when we choose (2.12a) and $\theta = 0$.

For the sake of simplicity $\vec{g}$ will be assumed only to depend on t (i.e. Dirichlet boundary conditions). The formula (2.1) will contain terms

$$(2.26)\qquad \vec{f}(t_n,\vec{y}_{n+1}^{(\ell)}+\vec{g}(t_n)) = \vec{f}(t_n,\vec{y}_n+\tau \sum_{i=0}^{\ell-1} \lambda_{\ell i}\vec{f}(t_n,\vec{y}_{n+1}^{(i)}+\vec{g}(t_n))+\vec{g}(t_n)).$$

Let us start the integration step at $t = t_n$ with a grid function $\vec{y}_n + \vec{g}(t_n)$ which converges to a smooth function as the grid is refined, i.e. as $h \to 0$, and τ is kept fixed. We expect that the expression (2.26) remains only bounded as $h \to 0$ if its argument $\vec{y}_{n+1}^{(\ell)} + \vec{g}(t_n)$ is a sufficiently smooth function on the grid $\Gamma_h \cup \partial\Gamma_h$, that is if the grid functions

$$\vec{f}(t_n, \vec{y}^{(i)}_{n+1} + \vec{g}(t_n))$$

converge to a smooth function on $\Omega \cup \partial\Omega$. However, since $\vec{f}$ has zero-components in all boundary points $\partial\Gamma_h$, this grid function converges to a discontinuous function, and therefore we cannot expect that (2.26) is bounded as $h \to 0$. The modification (2.12b) does not give this singular behaviour, because (2.26) now reads (for general boundary functions $\vec{g}(t,\vec{y})$)

$$\begin{aligned} &\vec{f}(t_n + \theta\tau, \vec{y}^{(\ell)}_{n+1} + \vec{g}(t_n + \theta_\ell \tau, \vec{y}^{(\ell)}_{n+1})) \\ &= \vec{f}(t_n + \theta\tau, \vec{y}_n + \vec{g}(t_n, \vec{y}_n) + \tau \sum_{i=0}^{\ell-1} \lambda_{\ell i} \vec{f}(t_n + \theta\tau, \vec{v}^{(i)}) \\ &+ \vec{g}(t_n + \theta_\ell \tau, \vec{y}_n + \tau \sum_{i=0}^{\ell-1} \lambda_{\ell i} \vec{f}(t_n + \theta\tau, \vec{v}^{(i)})) - \vec{g}(t_n, \vec{y}_n)) \end{aligned} \tag{2.26'}$$

where

$$\vec{v}^{(i)} = \vec{y}^{(i)}_{n+1} + \vec{g}(t_n + \theta_i \tau, \vec{y}^{(i)}_{n+1}).$$

By the same reasoning as above we should now require that the grid functions

$$\vec{f}(t_n + \theta\tau, \vec{v}^{(i)}) + \frac{\vec{g}(t_n + \theta_\ell \tau, \vec{y}_n + \tau \sum_{j=0}^{\ell-1} \lambda_{\ell j} \vec{f}(t_n + \theta\tau, \vec{v}^{(j)})) - \vec{g}(t_n, \vec{y}_n)}{\tau \sum_{i=0}^{\ell-1} \lambda_{\ell i}}$$

are sufficiently smooth for small h-values. If the functions $\vec{f}$ and $\vec{g}$ slowly vary with their arguments, this grid function approximates the grid function $d\vec{y}/dt + d\vec{b}/dt$ at $t = t_n$ and therefore may be assumed to be sufficiently smooth.

By a similar argument we are led to the function (2.12d) instead of the usual linearization of $\vec{f}(t,\vec{y})$.

3. NUMERICAL EXPERIMENTS

The aim of our numerical experiments was to get information to what extent the results of the MRK methods differ from the results of the generating classical methods. In order to demonstrate the relevance of the modified

schemes in practical problems we also present results obtained by the ADI method which is accepted in the literature as one of the more efficient integration techniques. To compare the efficiency of the MRK and ADI methods we define the *efficiency rate* of the MRK methods with respect to the ADI method: Let τ be the stepsize for which the m-stage MRK method produces an accuracy A and let N(A) be the total number of $\vec{f}$-evaluations which is needed by the ADI method to produce the same accuracy A. Then $\tau N(A)/m$ will be called the efficiency rate of the MRK method corresponding to the integration step τ. This value indicates the fraction to which the computational effort of the generating RK method should be reduced in order to make its modification as efficient as the ADI method. The rate of efficiency of an MRK method provides insight in its practical relevance. Of course, one should only apply the modification when a substantial reduction of computing time can be achieved. For the sake of illustration however this is not important and therefore we have chosen relatively simple test problems.

3.1. The initial-boundary value problems

The following two equations were chosen:

$$u_t = u_{x_1x_1} + u_{x_2x_2} - e^{-t}(x_1^2 + x_2^2 + 4), \qquad 0 \le t \le 1 \tag{3.1}$$

and

$$u_t = \frac{x_1+x_2}{2(1+t)}\left[(u^3)_{x_1x_1} + (u^3)_{x_2x_2}\right] + \pi(x_1+x_2)\cos(2\pi t) - \frac{3(x_1+x_2)^2}{4(1+t)}\sin^3(2\pi t), \quad 0 \le t \le 1. \tag{3.2}$$

The functions

$$u(t,x_1,x_2) = 1 + e^{-t}(x_1^2 + x_2^2) \tag{3.3}$$

and

(3.4) $\quad u(t,x_1,x_2) = \sin(2\pi t)(x_1 + x_2)/2$

satisfy (3.1) and (3.2), respectively. Both equations were considered on the unit square $0 \le x_1,x_2 \le 1$ with Dirichlet-boundary conditions along its boundary $\partial\Omega$. These boundary conditions and the initial condition at $t = 0$ are defined by the exact solutions (3.3) and (3.4).

The initial-boundary value problems were semi-discretized on a uniform grid in the (x_1,x_2)-plane using the standard symmetric differences. As grid-size we chose $h = 1/20$ for equation (3.1) and $h = 1/20$, $h = 1/40$ for equation (3.2). This results in systems of respectively 361 and 1521 ordinary differential equations of the form (1.1). The solutions of these systems can de derived from $u(t,x_1,x_2)$ by restricting (x_1,x_2) to the grid points (this is a consequence of x_1 and x_2 occurring quadratically in (3.3) and linearly in (3.4)).

The spectral radius $\sigma(J)$ needed in the stability condition was approximated by respectively

$$\text{(3.5)} \qquad \sigma = \frac{8}{h^2}, \quad \sigma = \max_{\Gamma_h} \frac{24u^2}{h^2} \cong \frac{24}{h^2}.$$

3.2. Methods used

Two RK formulas combined with several modified right hand side functions $\vec{f}^{**}$ were applied to the problems (3.1) and (3.2). In addition, we used the standard ADI method of Peaceman and Rachford.

The RK formulas used were proposed in [3] and of the form

$$\text{(3.6)} \qquad \begin{aligned} &\vec{y}^{(0)}_{n+1} = \vec{y}_n, \quad \vec{y}^{(1)}_{n+1} = \vec{y}_n + \tilde{\mu}_1 \tau \vec{f}^{**}(t_n,\vec{y}_n), \\ &\vec{y}^{(j)}_{n+1} = \mu_j \vec{y}^{(j-1)}_{n+1} + (1-\mu_j)\vec{y}^{(j-2)}_{n+1} + \tilde{\gamma}_j \tau \vec{f}^{**}(t_n,\vec{y}_n) + \tilde{\mu}_j \tau \vec{f}^{**}(t_n+\theta_{j-1}\tau, \vec{y}^{(j-1)}_{n+1}), \\ &\theta_0 = 0, \quad \theta_1 = \tilde{\mu}_1, \quad \theta_j = \mu_j\theta_{j-1} + (1-\mu_j)\theta_{j-2} + \tilde{\gamma}_j + \tilde{\mu}_j, \quad j = 1,2,\ldots,m, \\ &\vec{y}_{n+1} = \vec{y}^{(m)}_{n+1}. \end{aligned}$$

The first formula is of first order, has $\tilde{R}_m^{(1)}(z)$ as its stability polynomial (cf. (2.5)) and is defined by

$$\tilde{\mu}_1 = \frac{w_0+1}{\beta w_0}, \quad \mu_j = 2w_0 \frac{T_{j-1}(w_0)}{T_j(w_0)}, \quad \tilde{\mu}_j = \frac{w_0+1}{\beta w_0}\mu_j, \quad \tilde{\gamma}_j = 0,$$

(3.7)

$$w_0 = 1 + \frac{1}{20m^2}, \quad \beta \cong 1.93m^2, \quad R_m(z) = \tilde{R}_m^{(1)}(z).$$

The second formula is second order accurate, has $\tilde{A}_m^{(2)}(z)$ as its stability polynomial (cf. (2.6)) and is defined by

$$\tilde{\mu}_1 = \frac{b(w_0+1)T_m(w_0)}{\beta w_0}, \quad \mu_j = 2w_0 \frac{T_{j-1}(w_0)}{T_j(w_0)}, \quad \tilde{\mu}_j = \frac{w_0+1}{\beta w_0}\mu_j,$$

(3.8)

$$\tilde{\gamma}_j = -a\tilde{\mu}_j,$$

$$w_0 = 1 + \frac{2}{13m^2}, \quad \beta \cong 0.65m^2, \quad R_m(z) = \tilde{A}_m^{(2)}(z).$$

Both formulas are internally stable for unlimited large m-values [3].

The modified right hand side functions $\vec{f}^{**}$ used in the experiments are taken from Section 2.3 and specified in the tables of results given in Section 3.3.

The number of stages in the RK methods was chosen according to the stability condition (1.2), i.e.

$$m = \left[\sqrt{\frac{\tau\sigma}{c_p}} + 1\right], \qquad c_1 \cong 1.93, \; c_2 \cong .65, \tag{3.9}$$

where [x] denotes the greatest integer less than or equal to x.

3.3. Numerical results

In the tables of results given below the values of the pair m\A are listed where A denotes the number of correct digits in the numerical solution, i.e.

(3.10) $$A = \min_{\Gamma_h} (-^{10}\log |\text{exact solution} - \text{numerical solution}|).$$

In the case of the ADI method m will denote the number of $\vec{f}$-evaluations per step needed in the Newton iteration processes. The various methods in the tables will be denoted by {integration formula; modification $\vec{f}^*$}. To compare the efficiency of the MRK and ADI methods we list for the most efficient MRK method the efficiency rate with respect to the ADI method. Actually the listed efficiency rates can be increased because in the ADI method we did not take into account the effort involved in evaluating the Jacobian matrix and the solution of the implicit equations.

In table 3.1 results are listed for the linear problem (3.1). It shows that the constant-time modification (2.12a) causes an unacceptably large drop in accuracy in spite of the fact that the generating formula (3.7) is of first order and the deviation from the classical RK formula of second order in τ (note that we choose $\theta = \beta_2$, cf. (2.16)). The modification (2.12b) in which the boundary field $\vec{g}(t,\vec{y})$ and the internal field $\vec{y}$ are tuned to the same time level, performs better but evidently the local deviation error (2.8) is still much larger than the local error of the RK formula. The second order version (2.17b) of (2.12b) turns out to produce a deviation error which is small with respect to the errors of the classical formulas (3.7) and (3.8) (we note that the A-values of the RK solution and the MRK solution differ by at most .30 if the global deviation error is roughly equal to the global error of the generating formulas). Since in this example the modification (2.17b) with $\bar{\theta} = 0$, $\bar{\bar{\theta}} = 1$ is hardly more expensive than (2.12b) we may conclude that the *second order modification (2.17b) combined with the second order RK formula is the most efficient MRK method*. The efficiency rate listed in table 3.1 shows that this MRK method is competative with the ADI method if in the low accuracy range the evaluation of $\vec{f}^*$ costs 20% of the evaluation of $\vec{f}$ and about 45% in the high accuracy range.

Table 3.1. Results for equation (3.1) when discretized on a grid with h = 1/20.

method	$\tau = 1$	$\tau = \frac{1}{12}$	$\tau = \frac{1}{35}$	$\tau = \frac{1}{70}$
{(3.7); $\vec{f}^{\,*} = \vec{f}$}	41\1.39	12\2.74	7\3.52	
{(3.7); (2.12a), $\theta = 0$}	41\-.06	12\.63	7\1.21	
{(3.7); (2.12a), $\theta = \beta_2$}	41\.07	12\.63	7\1.22	
{(3.7); (2.12b), $\theta = \beta_2$}	41\.55	12\2.19	7\3.26	
{(3.7); (2.17b), $\bar{\theta} = 0$, $\bar{\bar{\theta}} = 1$}	41\1.15	12\2.71	7\3.51	
{(3.8); $\vec{f}^{\,**} = \vec{f}$}	71\2.02	21\3.70	12\4.49	9\5.08
{(3.8); (2.12b), $\theta = \frac{1}{2}$}	71\0.82	21\2.01	12\2.61	9\3.27
{(3.8); (2.17b), $\bar{\theta} = 0$, $\bar{\bar{\theta}} = 1$}	71\2.29	21\3.53	12\4.43	9\5.02
{ADI; $\vec{f}$}	2\.79	2\2.81	2\3.74	2\4.34
Efficiency rate	0.19	0.22	0.37	0.49

Table 3.2. Results for equation (3.2) with h = 1/20.

method	$\tau = 1$	$\tau = \frac{1}{10}$	$\tau = \frac{1}{20}$	$\tau = \frac{1}{40}$	$\tau = \frac{1}{80}$	$\tau = \frac{1}{160}$
{(3.7); $\vec{f}^{\,**} = \vec{f}$}	71\-.23	23\.87	16\1.25	12\1.56	8\1.86	
{(3.7); (2.12b), $\theta = \beta_2$}	71\-.16	23\.83	16\1.24	12\1.56	8\1.86	
{(3.8); $\vec{f}^{\,**} = \vec{f}$}	122\-.26	38\1.41	28\2.05	20\2.89	14\3.66	10\4.26
{(3.8); (2.12b), $\theta = \frac{1}{2}$}	122\-.13	38\.78	28\1.51	20\2.39	14\3.45	10\4.23
{ADI; $\vec{f}$}	10*	10*	10*	10*	2\2.11 4\3.18	2\2.73 4\4.25
Efficiency rate				0.25	0.38	0.38

Table 3.2 shows that the deviation error of the first order modifications is relatively small in comparison with the error of the generating formulas. It is also clearly seen that the second order RK formula (3.8) is considerably more sensitive to modification than the first order RK formula (3.7). The ADI method developed instabilities for $\tau \geq 1/40$ (indicated

by * in the tables of results). In the higher accuracy region the MRK method {(3.8); (2.12b)} is competitive with the ADI method if the computation time to evaluate $\vec{f}^*$ is reduced to 38% of that of $\vec{f}$. In the lower accuracy region (A $\in$ [1,3] say) the MRK method is at least competitive if this percentage is 25%.

Table 3.3. Results for equation (3.2) with h = 1/40.

method	$\tau = 1$	$\tau = \frac{1}{10}$	$\tau = \frac{1}{20}$	$\tau = \frac{1}{40}$	$\tau = \frac{1}{80}$	$\tau = \frac{1}{160}$
{(3.7); $\vec{f}^* = \vec{f}$}	142\-.21	45\.85	32\1.24	23\1.56	16\1.86	12\2.16
{(3.7); (2.12b), $\theta = \beta_2$}	142\-.15	45\.81	32\1.23	23\1.56	16\1.86	12\2.16
{(3.8); $\vec{f}^* = \vec{f}$}	244\-.31	77\1.36	55\2.00	39\2.83	28\3.67	20\4.28
{(3.8); (2.12b), $\theta = \frac{1}{2}$}	244\-.14	77\.71	55\1.50	39\2.36	28\3.27	20\4.24
{ADI; $\vec{f}$}	10*	10*	10*	10*	4\2.67	4\3.71
Efficiency rate					0.23	0.37

In table 3.3 the experiments listed in table 3.2 are repeated for h = 1/40. We see that for larger values of $\tau\sigma$ the accuracy of the RK and MRK methods is nearly preserved whereas the ADI method shows a substantial decrease in accuracy.

The final conclusion (based on our experiments) is that

(i) the second order MRK method is the most efficient one;

(ii) the MRK method is advantageous in problems where the modification results in a substantial reduction of computing time when compared with the generating RK method. We found reduction to 25% adequate in order to be competative with the ADI method;

(iii) the stability behaviour of the MRK method is insensitive to large $\tau\sigma$-values; hence low accuracy results can be obtained for strongly non-linear problems, while the ADI method behaves unstable in this situation.

REFERENCES

[1] FAIRWEATHER, G. & A.R. MITCHELL, *A new computational procedure for ADI-methods*, SIAM J. Num. Anal. 4, 1967, 163-170.

[2] HOUWEN, P.J. van der, *Stabilized Runge-Kutta methods for second order differential equations without first derivatives*, SIAM J. Num. Anal. 16, 1979, 523-537.

[3] HOUWEN, P.J. van der & B.P. SOMMEIJER, *On the internal stability of explicit, m-stage Runge-Kutta methods for large values of* m, Report NW 72/79, Mathematisch Centrum, Amsterdam (prepublication) 1979.

[4] SHAMPINE, L.F., *Storage reduction for Runge-Kutta codes*, ACM Trans. Math. Software 5, 245-250, 1979.