stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

NW 39/77 FEBRUARI

J.G. VERWER

A CLASS OF STABILIZED THREE-STEP RUNGE-KUTTA METHODS FOR THE NUMERICAL INTEGRATION OF PARABOLIC 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

A class of stabilized three-step Runge-Kutta methods for the numerical integration of parabolic equations *)

by

J.G. Verwer

ABSTRACT

A class of explicit three-step Runge-Kutta methods is discussed for the numerical solution of initial value problems for systems of ordinary differential equations. Attention is focussed on systems which originate from parabolic partial differential equations by applying the method of lines. New stabilized schemes of first and second order are presented. Some numerical examples are discussed.

KEYWORDS & PHRASES: Numerical analysis, Parabolic partial differential equations, Method of lines, Multistep Runge-Kutta methods, Stability.

---

*) This report will be submitted for publication elsewhere

1. INTRODUCTION

This paper deals with the construction of a class of stabilized explicit integration formulas for the numerical solution of systems of ordinary differential equations

$$y' = f(y), \tag{1.1}$$

Attention is focussed on systems which originate from semi-discretization of parabolic partial differential equations. Throughout this paper, it is assumed that the eigenvalues of the Jacobian matrix, say $J(y)$, of system (1.1) are spread out over a long narrow strip around the negative axis of the complex plane. For the majority of parabolic problems this assumption is satisfied (see Richtmeyer & Morton [4]).

The formulas we discuss are three-step formulas belonging to the wide class of multistep Runge-Kutta methods, a class of integration methods which was originally discussed by Gear [1]. In the terminology of Gear such a method is called a hybrid method. A thoroughly theoretical analysis of these methods has been given by Watt [10]. More recently, applications of multistep Runge-Kutta methods are discussed in Van der Houwen [7]. We shall also use the name multistep Runge-Kutta method of degree m, where m denotes the number of function evaluations. The discussion is confined to methods of order one and two. For a great deal of the problems of type (1.1), which occur in practice, methods of low order are valuable. The degree of the formulas constructed varies between two en twelve.

Stabilized one-step Runge-Kutta methods have already been discussed by Van der Houwen [6,7]. In the second reference he also pays attention to a special class of stabilized two-step methods. The change-over from

one-step to two- or three-step schemes is of course meant to enlarge the real boundary of absolute stability, say β. To state the most important result of our investigations, we present new schemes for which holds:

$$\beta(m) \simeq 5.15\ m^2,\quad 2 \le m \le 12,\ \text{order} = 1,$$
$$\beta(m) \simeq 2.29\ m^2,\quad 2 \le m \le 12,\ \text{order} = 2.$$

For comparison, these boundaries are nearly three times larger than corresponding boundaries for stabilized one-step methods, provided the same damping properties are required. This of importance because of the fact that the steplength for an explicit integration method for parabolic systems of type (1.1) is usually limited by stability requirements.

2. ANALYSIS OF THE INVESTIGATED CLASS OF METHODS

The class of methods we consider may be represented by the following formula:

$$y_{n+1}^{(0)} = y_n,$$
$$y_{n+1}^{(1)} = (1-b_1)\ y_n + b_1 y_{n-1} + c_1 hf(y_{n-1}) + \lambda_{1,0}\ hf(y_n),$$
$$y_{n+1}^{(j)} = (1-b_j) y_n + b_j y_{n-1} + c_j hf(y_{n-1}) + \lambda_{j,0}\ hf(y_n) + \tag{2.1}$$
$$+ \lambda_{j,j-1}\ hf(y_{n+1}^{(j-1)}),\quad j = 2,\ldots,m;\ m \ge 2,$$
$$y_{n+1} = d y_{n+1}^{(m)} + (1-d) y_{n-2},\quad n \ge 2.$$

The vector y_n always represents a numerical approximation to the analytical solution $y(x)$ at $x = x_n$. The points $x_j, j = n + 1, \ldots, n - 2$, denote the reference points of the three-step formula and h denotes the steplength, i.e. $h = x_{n+1} - x_n$, $n = 0,1,\ldots$. The steplength h is supposed to be constant. If $d = 1$ and $b_j = c_j = 0$, $j = 1,\ldots,m$, we have a one-step method which is discussed in Van der Houwen [6,7]. This one-step method may be used to provide the additional starting vectors y_1 and y_2. If $d = 1$, we have a two-step method which is discussed in Verwer [8].

There are two main reasons why we consider three-step formulas of the special class (2.1). Firstly, in order to reduce the storage requirements we only admit the derivatives $f(y_{n-1})$, $f(y_n)$ and $f(y_{n+1}^{(j-1)})$. By this choice our formulas need six arrays of storage. Secondly, in order to be able to apply the construction discussed in section 3.1, the vector y_{n-2} is not allowed to occur in the expressions for $y_{n+1}^{(j)}$ (cf. Verwer [9]).

As already noted in the introduction, this paper discusses the construction of stabilized formulas. As a consequence, we pay no attention to purely theoretical aspects. For a theoretical discussion of multistep Runge-Kutta methods the interested reader is referred to Watt [10], where also a convergence proof is given. For the usual definitions about convergence, consistency and stability we refer to Lambert [3].

2.1. CONVERGENCE AND CONSISTENCY CONDITIONS

The method is developed for the integration of partial differential equations. In a lot of applications of partial equations low order methods can be used succesfully. Therefore, we confine ourselves to methods of order $p = 1$ and $p = 2$. However, consistency conditions will be derived for $p \leq 3$. The third order conditions can then be used for a local error control in case of a second order method.

It is convenient to associate three-step method (2.1) with a non-linear difference operator

$$(2.2) \qquad y_{n+1} - E[y_n, y_{n-1}, y_{n-2}].$$

Now consider the initial value problem

$$(2.3) \qquad y' = f(y),\ y(x_0) = y_0,\ x \geq x_0,$$

where f is a vector function of sufficient differentiability. Let y(x) be the solution of (2.3). Then the higher derivatives of y(x) can be expressed in terms of the function f and its derivatives. By using the tensor notation in Taylor's theorem for functions of several variables (see Henrici [2], p. 118), we obtain

$$(2.4) \qquad y(x_{n+1}) - E[y(x_n), y(x_{n-1}), y(x_{n-2})] =$$

$$C_1 hf + C_2 h^2 f_j f^j + C_{31} h^3 f_j f^j_k f^k + C_{32} h^3 f_{jk} f^j f^k + O(h^4),$$

where

$$C_1 = 1 - \{d(-b_m+c_m+\lambda_{m,0}+\lambda_{m,m-1}) - 2(1-d)\},$$

$$C_2 = \frac{1}{2} - \{d(\tfrac{1}{2}b_m-c_m+\lambda_{m,m-1}(-b_{m-1}+c_{m-1}+\lambda_{m-1,0}+\lambda_{m-1,m-2})) + 2(1-d)\},$$

(2.5)
$$C_{31} = \frac{1}{6} - \{d(-\tfrac{1}{6}b_m+\tfrac{1}{2}c_m+\tfrac{1}{2}\lambda_{m,m-1}(b_{m-1}-2c_{m-1}+2\lambda_{m-1,m-2}(-b_{m-2}+$$

$$+c_{m-2}+\lambda_{m-2,0}+\lambda_{m-2,m-3}))) - \frac{8}{6}(1-d)\},$$

$$C_{32} = \frac{1}{6} - \{d(-\tfrac{1}{6}b_m+\tfrac{1}{2}c_m+\tfrac{1}{2}\lambda_{m,m-1}(-b_{m-1}+c_{m-1}+\lambda_{m-1,0}+\lambda_{m-1,m-2})^2) +$$

$$- \frac{8}{6}(1-d)\}.$$

Thus method (2.1) is consistent of order $p = 1$ if $C_1 = 0$, and consistent of order $p = 2$ if, in addition, $C_2 = 0$.

As is the case for linear multistep methods, a necessary condition for multistep Runge-Kutta methods to be convergent is the condition of zero-stability. In fact, the well-known convergence theorem for linear multistep methods, which states that the method is convergent if and only if it is zero-stable and consistent, applies for multistep Runge-Kutta methods. The condition, necessary for zero-stability of method (2.1) is (see Verwer [8]).

(2.6) $\quad d(c_m+\lambda_{m,0}+\lambda_{m,m-1}) \neq 0.$

As is the case for linear multistep methods, it is recommended to use the left-hand side of (2.6) to normalize the error constants of the truncation error. We have chosen

(2.7) $\quad d(c_m+\lambda_{m,0}+\lambda_{m,m-1}) = 1,$

which is assumed throughout this paper. If the left-hand side of (2.6) is chosen smaller than one, we have in fact a method of the Du Fort-Frankel type (see Richtmyer & Morton [4]).

2.2. ABSOLUTE STABILITY PROPERTIES

In order to investigate the absolute stability properties of method (2.1) it is applied to the linear test-model

(2.8) $\quad y' = \delta y, \ \delta \in C.$

This yields the recurrence relation

(2.9) $\quad y_{n+1} = d\ S(z)\ y_n + d\ P(z)\ y_{n-1} + (1-d)\ y_{n-2},$

where $S(z)$ and $P(z)$ are polynomials of degree m in $z = h\delta$. For future reference, we shall call S and P the stability polynomials for method (2.1). Let us denote

where

$$C_1 = 1 - \{d(-b_m+c_m+\lambda_{m,0}+\lambda_{m,m-1}) - 2(1-d)\},$$

$$C_2 = \frac{1}{2} - \{d(\tfrac{1}{2}b_m-c_m+\lambda_{m,m-1}(-b_{m-1}+c_{m-1}+\lambda_{m-1,0}+\lambda_{m-1,m-2})) + 2(1-d)\},$$

(2.5)
$$C_{31} = \frac{1}{6} - \{d(-\tfrac{1}{6}b_m+\tfrac{1}{2}c_m+\tfrac{1}{2}\lambda_{m,m-1}(b_{m-1}-2c_{m-1}+2\lambda_{m-1,m-2}(-b_{m-2}+$$

$$+c_{m-2}+\lambda_{m-2,0}+\lambda_{m-2,m-3}))) - \frac{8}{6}(1-d)\},$$

$$C_{32} = \frac{1}{6} - \{d(-\tfrac{1}{6}b_m+\tfrac{1}{2}c_m+\tfrac{1}{2}\lambda_{m,m-1}(-b_{m-1}+c_{m-1}+\lambda_{m-1,0}+\lambda_{m-1,m-2})^2) +$$

$$- \frac{8}{6}(1-d)\}.$$

Thus method (2.1) is consistent of order $p = 1$ if $C_1 = 0$, and consistent of order $p = 2$ if, in addition, $C_2 = 0$.

As is the case for linear multistep methods, a necessary condition for multistep Runge-Kutta methods to be convergent is the condition of zero-stability. In fact, the well-known convergence theorem for linear multistep methods, which states that the method is convergent if and only if it is zero-stable and consistent, applies for multistep Runge-Kutta methods. The condition, necessary for zero-stability of method (2.1) is (see Verwer [8]).

(2.6) $\qquad d(c_m+\lambda_{m,0}+\lambda_{m,m-1}) \neq 0.$

As is the case for linear multistep methods, it is recommended to use the left-hand side of (2.6) to normalize the error constants of the truncation error. We have chosen

(2.7) $\qquad d(c_m+\lambda_{m,0}+\lambda_{m,m-1}) = 1,$

which is assumed throughout this paper. If the left-hand side of (2.6) is chosen smaller than one, we have in fact a method of the Du Fort-Frankel type (see Richtmyer & Morton [4]).

2.2. ABSOLUTE STABILITY PROPERTIES

In order to investigate the absolute stability properties of method (2.1) it is applied to the linear test-model

(2.8) $\qquad y' = \delta y, \ \delta \in C.$

This yields the recurrence relation

(2.9) $\qquad y_{n+1} = d\ S(z)\ y_n + d\ P(z)\ y_{n-1} + (1-d)\ y_{n-2},$

where $S(z)$ and $P(z)$ are polynomials of degree m in $z = h\delta$. For future reference, we shall call S and P the stability polynomials for method (2.1). Let us denote

$$(2.10)\qquad S(z) = \sum_{i=0}^{m} s_i z^i \quad \text{and} \quad P(z) = \sum_{i=0}^{m} p_i z^i.$$

Then we have

$$(2.11)\qquad \begin{aligned} s_0 &= 1 - b_m, \\ s_1 &= \lambda_{m,0} + \lambda_{m,m-1}(1-b_{m-1}), \\ s_i &= \prod_{j=m-i+2}^{m} \lambda_{j,j-1}(\lambda_{m-i+1,0}+\lambda_{m-i+1,m-i}(1-b_{m-i})), \quad i = 2,\ldots,m-1, \\ s_m &= \prod_{j=1}^{m} \lambda_{j,j-1}, \end{aligned}$$

and

$$(2.12)\qquad \begin{aligned} p_0 &= b_m, \\ p_1 &= \lambda_{m,m-1} b_{m-1} + c_m, \\ p_i &= \Big(\prod_{j=m-i+1}^{m} \lambda_{j,j-1}\Big) b_{m-i} + \Big(\prod_{j=m-i+2}^{m} \lambda_{j,j-1}\Big) c_{m-i+1}, \quad i = 2,\ldots,m-1, \\ p_m &= \Big(\prod_{j=2}^{m} \lambda_{j,j-1}\Big) c_1. \end{aligned}$$

The characteristic equation of the three-step recurrence (2.9) is

$$(2.13)\qquad \alpha^3 - d\,S(z)\,\alpha^2 - d\,P(z)\,\alpha - 1 + d = 0.$$

Before analyzing the stability of (2.9) by means of the characteristic roots of (2.13), it is convenient to express the consistency conditions for orders p = 1 and p = 2 into the first coefficients of S and P. This

can be done by using equations (2.5), (2.11) and (2.12), or alternatively, by substituting the second order Padé-approximation $1 + z + z^2/2$ to exp(z) into characteristic equation (2.13). The conditions are given below:

$$
\begin{array}{c|l}
 & s_0 + p_0 = 1, \\
p = 1 & s_1 - p_0 + p_1 = (3-2d)/d; \\
\hline
p = 2 & s_2 + \frac{1}{2} p_0 - p_1 + p_2 = (-3/2+2d)/d;
\end{array}
\tag{2.14}
$$

Thus these conditions are equivalent to the conditions $C_1 = 0$ for first order and $C_1 = C_2 = 0$ for second order, which are stated in the preceding section.

It is further convenient to express equation (2.7) in terms of d and p_0. By using the first of relations (2.5) and equality $p_0 = b_m$, we find

$$p_0 = \frac{2(d-1)}{d}. \tag{2.15}$$

It is our aim to develop stabilized formulas for parabolic equations. As the eigenvalues of the Jacobian matrix of such equations are generally real or almost real, it is of interest to develop formulas whose stability regions contain a considerable part of the negative axis. As a consequence, we state the following

STABILITY PROBLEM: Let z be negative and let $\alpha_i(z)$, $i = 1,2,3$, denote the roots of equation (2.13). Let ρ: $(-\infty,0) \to (0,1)$, $\rho(0) = 1$, be given and assume that relation (2.15) is satisfied. Then determine the coefficients s_i, p_i, $i = 0,\ldots,m$, in such a way that $\max_i \mid \alpha_i(z) \mid \leq \rho(z)$, $z \in [-\beta,0]$, β maximal, where it is assumed that $p = 1$ and $p = 2$, respectively.

The function ρ is introduced in order to obtain a strong damping for the higher harmonics. Moreover, ρ may be considered as an aid to construct a method of which the absolute stability region contains a narrow strip around the negative axis. If $\rho(z)$, $z < 0$, is not too close to 1, it is immediately clear that we can find such a region. The boundary β is the real boundary of absolute stability.

The construction of approximate solutions to the optimization problem is discussed in section 3.1. In section 3.1 we also give the results and show some absolute stability regions.

2.3. INTERNAL STABILITY PROPERTIES

An important concept for stabilized methods of the Runge-Kutta type is the concept of internal stability (see Van der Houwen [7], section 2.6.10). Internal stability deals with the propagation of round-off errors during a single integration step. For Runge-Kutta methods, possessing a large degree and a large stability boundary, this propagation may be

considerable and may easily influence the local accuracy. Van der Houwen analyzes the internal stability for the class of one-step methods which is contained in class (2.1). He defines a so-called internal stability function, i.e. a function which approximately controls the propagation of round-off errors during a single step. It turns out that method (2.1) possesses the same internal stability function.

Let $\rho_{n+1}^{(j)}$ denote the local error entering at stage j of the Runge-Kutta process. Let $\varepsilon_{n+1}^{(j)}$ denote the accumulated local error at stage j. Further, let $\bar{y}_{n+1}^{(j)}$ denote the perturbed $y_{n+1}^{(j)}$. Instead of (2.1) we then have the process

$$\bar{y}_{n+1}^{(0)} = y_n,$$

$$\bar{y}_{n+1}^{(1)} = (1-b_1)y_n + b_1 y_{n-1} + c_1 hf(y_{n-1}) + \lambda_{1,0} hf(y_n) + \rho_{n+1}^{(1)},$$

$$\bar{y}_{n+1}^{(j)} = (1-b_j)y_n + b_j y_{n-1} + c_j hf(y_{n-1}) + \lambda_{j,0} hf(y_n) +$$

$$+ \lambda_{j,j-1} hf(\bar{y}_{n+1}^{(j-1)}) + \rho_{n+1}^{(j)}, \quad j = 2,\ldots,m, \quad m \geq 2,$$

$$\bar{y}_{n+1} = d\, \bar{y}_{n+1}^{(m)} + (1-d)\, y_{n-2}, \quad n \geq 2.$$

The errors $\varepsilon_{n+1}^{(j)} = \bar{y}_{n+1}^{(j)} - y_{n+1}^{(j)}$ then satisfy:

$$\varepsilon_{n+1}^{(1)} = \rho_{n+1}^{(1)},$$

$$\varepsilon_{n+1}^{(j)} = \lambda_{j,j-1}\, h[f(y_{n+1}^{(j-1)} + \varepsilon_{n+1}^{(j-1)}) - f(y_{n+1}^{(j-1)})] + \rho_{n+1}^{(j)}, \quad j = 2,\ldots,m,$$

$$\varepsilon_{n+1} = d\, \varepsilon_{n+1}^{(m)},$$

where $\varepsilon_{n+1} = \bar{y}_{n+1} - y_{n+1}$. By assuming that the Jacobian J(y) is slowly varying during one step, there approximately holds

$$\varepsilon_{n+1}^{(j)} \simeq \lambda_{j,j-1}\, hJ(y_n)\, \varepsilon_{n+1}^{(j-1)} + \rho_{n+1}^{(j)},\quad j = 2,\ldots,m.$$

After some elementary calculations we then arrive at the estimate

$$\|\varepsilon_{n+1}\| \le \Big[d + \sum_{k=1}^{m-1} d \prod_{j=m+1-k}^{m} |\lambda_{j,j-1}|\, \|(hJ(y_n))^k\|\Big] \max_{1\le k\le m} \|\rho_{n+1}^{(k)}\|,$$

where $\|\cdot\|$ is the spectral norm.

Following Van der Houwen we now define the internal stability function

$$Q(z) = d + \sum_{k=1}^{m-1} d \prod_{j=m+1-k}^{m} |\lambda_{j,j-1}|\, |z|^k. \tag{2.16}$$

In case of a normal matrix $J(y_n)$ there holds

$$\|\varepsilon_{n+1}\| \le Q(h\sigma(J(y_n))) \max_{1\le k\le m} \|\rho_{n+1}^{(k)}\|,$$

when σ denotes the spectral radius. Consequently, in actual computation the steplength h should at least satisfy the internal stability condition

$$Q(h\sigma(J(y_n))) \le \frac{\text{tolerance}}{\text{arithmetic precision}}, \tag{2.17}$$

where tolerance stands for the maximal local truncation error allowed. However, when cancellation of digits appears, this condition may even be too optimistic.

The significance of the internal stability condition becomes clear when one realizes that Q is a strongly increasing function in its argument, and that for stabilized methods large arguments occur. Values of $Q(\beta)$ will be given in section 3.2.

The significance of condition (2.17) is corroborated by practical experiments. It turns out that when (2.17) is satisfied the internal stability is generally under control. A numerical experiment illustrating the significance of the internal stability function is discussed in section 3.2.

3. CONSTRUCTION OF THE ALGORITHMS

In section 3.1 we discuss a heuristic solution technique which yields approximate solutions to the stability problem stated in section 2.2. Once the parameter d and the coefficients s_i and p_i are determined, it is easy to derive parameters for a three-step scheme. A class of schemes of first and second order is presented in section 3.2.

3.1. A SOLUTION TECHNIQUE FOR THE STABILITY PROBLEM

In order to save space we will confine ourselves to the main features of the technique. For details we refer to Verwer [9].

Suppose a damping function $\rho(z)$ is given. Then define $\alpha = \rho\xi$ and substitute into equation (2.13). This yields a cubic equation in ξ:

$$\rho^3\xi^3 - dS\rho^2\xi^2 - dP\rho\xi - (1-d) = 0. \tag{3.1}$$

Let

$$\xi = \frac{1 + \eta}{1 - \eta}, \tag{3.2}$$

which maps the interior of the unit circle $|\xi| = 1$ into the half-plane $\mathrm{Re}(\eta) < 0$. Substitution of (3.2) into (3.1) yields a cubic equation in η:

$$a_0\eta^3 + a_1\eta^2 + a_2\eta + a_3 = 0, \tag{3.3}$$

where

$$\begin{aligned} a_0 &= \rho^3 + dS\rho^2 - dP\rho + 1 - d, \\ a_1 &= 3\rho^3 + dS\rho^2 + dP\rho - 3(1-d), \\ a_2 &= 3\rho^3 - dS\rho^2 + dP\rho + 3(1-d), \\ a_3 &= \rho^3 - dS\rho^2 - dP\rho - (1-d). \end{aligned} \tag{3.4}$$

Sufficient conditions for the roots of (3.1) to lie inside or on the unit circle can be obtained by applying the Routh-Hurwitz criterion to (3.3) (see Lambert [3]). These conditions read:

$$a_i \geq 0, \quad i = 0,\ldots,3; \quad a_1a_2 - a_0a_3 \geq 0. \tag{3.5}$$

Observe that without the equality signs conditions (3.5) are necessary for the roots of (3.1) to lie inside the unit circle. In terms of S, P, d and ρ conditions (3.5) give:

$$\rho S - P \geq \frac{d - 1 - \rho^3}{d\rho},$$

$$\rho S + P \geq \frac{3(1-d) - 3\rho^3}{d\rho},$$

(3.7)
$$-\rho S + P \geq \frac{3(d-1) - 3\rho^3}{d\rho},$$

$$-\rho S - P \geq \frac{1 - d - \rho^3}{d\rho},$$

$$\frac{1 - d}{\rho^2} S + P \geq \frac{(1-d)^2 - \rho^6}{d\rho^4}.$$

The problem stated in section 2.2 thus reads: Let the function ρ be given and let condition (2.15) be satisfied. Then determine the coefficients s_i and p_i, compatible to an imposed order of accuracy, in such a way that (3.7) is satisfied for $z \in [-\beta,0]$, β maximal.

The technique is based on the following heuristic idea: Suppose the parameter d is fixed beforehand. Then discretize the variable z on an interval $[-\bar{\beta},0]$, i.e. define points $z_j = j\Delta z$, $\Delta z = \bar{\beta}/N$, $j = 1,\ldots,N$, where N is prescribed. Next replace the five non-linear inequalities by a system of linear inequalities by substituting $z = z_j$, $j = 1,\ldots,N$. After adding 4 + 2p inequalities associated with consistency conditions (2.14) and relation (2.15), we arrive at the system

(3.8)
$$AX \geq C,$$

A being a $(5N+2p+4) * (2m+2)$ matrix, X being a $2m + 2$-vector of unknowns

$p_i, s_i, i = 0,\ldots,m$, and C being the $5N + 2p + 4$-vector of right-hand sides. If $\overline{\beta} \leq \beta$, β being the optimal real stability boundary, a feasible solution to (3.8) must exist. On the other hand, if $\overline{\beta} > \beta$ and N large enough, a feasible solution cannot exist. Such a feasible solution is easy to determine by using a linear programming method. Summarizing, once the optimal d is known, an approximate and almost optimal solution is easy to determine by solving a sequence of linear programming problems, e.g. by performing bisection on $\overline{\beta}$.

In actual calculation it is recommended to expand the polynomials S and P in orthogonal polynomials in order to prevent numerical difficulties for higher values of m. The vector X of unknowns then consists of the coefficients of the polynomial expansions.

Another remark of practical interest is the following. In order to satisfy (3.7) for arguments z between points z_j, it is necessary to choose N rather large. As a consequence, the number of constraints of (3.8) is much larger than the number of variables. With regard to computational efficiency, it is then more effective to solve a linear programming problem belonging to the transposed of (3.8). In order to realize this, consider problem

$$\min B^T X, \quad B = [1,\ldots,1]^T, \tag{3.9}$$

subject to

$$AX \geq C, \quad -\infty \leq X_i \leq \infty, \tag{3.10}$$

and its dual problem

(3.11) $\max C^T Y,$

subject to

(3.12) $A^T Y = B,\ Y_i \geq 0.$

From the foregoing it is clear that we are primarily interested in the existence or non-existence of a feasible solution to (3.8). Well, according to the duality theorem, the dual solution to (3.11) - (3.12) is the primal solution to (3.9) - (3.10), which is a feasible solution to (3.8), provided system (3.8) has a feasible solution. Thus in actual calculation it is recommended to use (3.11) - (3.12) for the determination of the solution to (3.8) which is optimal with respect to β.

There remains to describe the determination of the parameter d. It is trivial that d is restricted to $0 < d < 2$, and the assumption is that the optimal d is independent of m. This assumption has been confirmed by practical experiments. The idea is then to determine an approximation to the optimal value of d by a numerical search technique for low values of m. Another assumption, also confirmed by practical experiments, is that $\beta(m) \simeq K\, m^2$, K constant. This means that the bisection process on $\overline{\beta}$ has to be performed only for some low values of m. For further details we refer to Verwer [9].

Using the heuristic solution technique described in this section, approximate solutions to the stability problem were computed for $2 \leq m \leq 12$ for the function

$$\rho(z) = \begin{cases} 1, & -1.5 < z \leq 0, \\ 0.85, & z \leq -1.5. \end{cases} \tag{3.13}$$

In the neighbourhood of the origin no damping is prescribed, the consistency and zero-stability of the method will take care off. The restriction to $m \leq 12$ will be explained in the next section. As a result of our calculations we found

$$\begin{aligned} \beta(m) &\simeq 5.15\ m^2, \quad p = 1, \\ \beta(m) &\simeq 2.29\ m^2, \quad p = 2. \end{aligned} \tag{3.14}$$

These boundaries are nearly three times larger than corresponding boundaries of one-step methods with the same damping. As a consequence of the discretization, the continuous damping for $-\beta \leq z \leq -1.5$ is not exactly 0.85, but approximately 0.9. To save space we do not list the resulting values of the parameter d and the coefficients s_i and p_i; they are given in Verwer [9].

In order to illustrate that the absolute stability regions $\{z \mid z \in \mathbb{C}, |\alpha_i(z)| < 1, i = 1,2,3\}$, belonging to the constructed stability polynomials S and P, contain a long narrow strip around the negative axis, four of such regions are given in fig. 3.1.

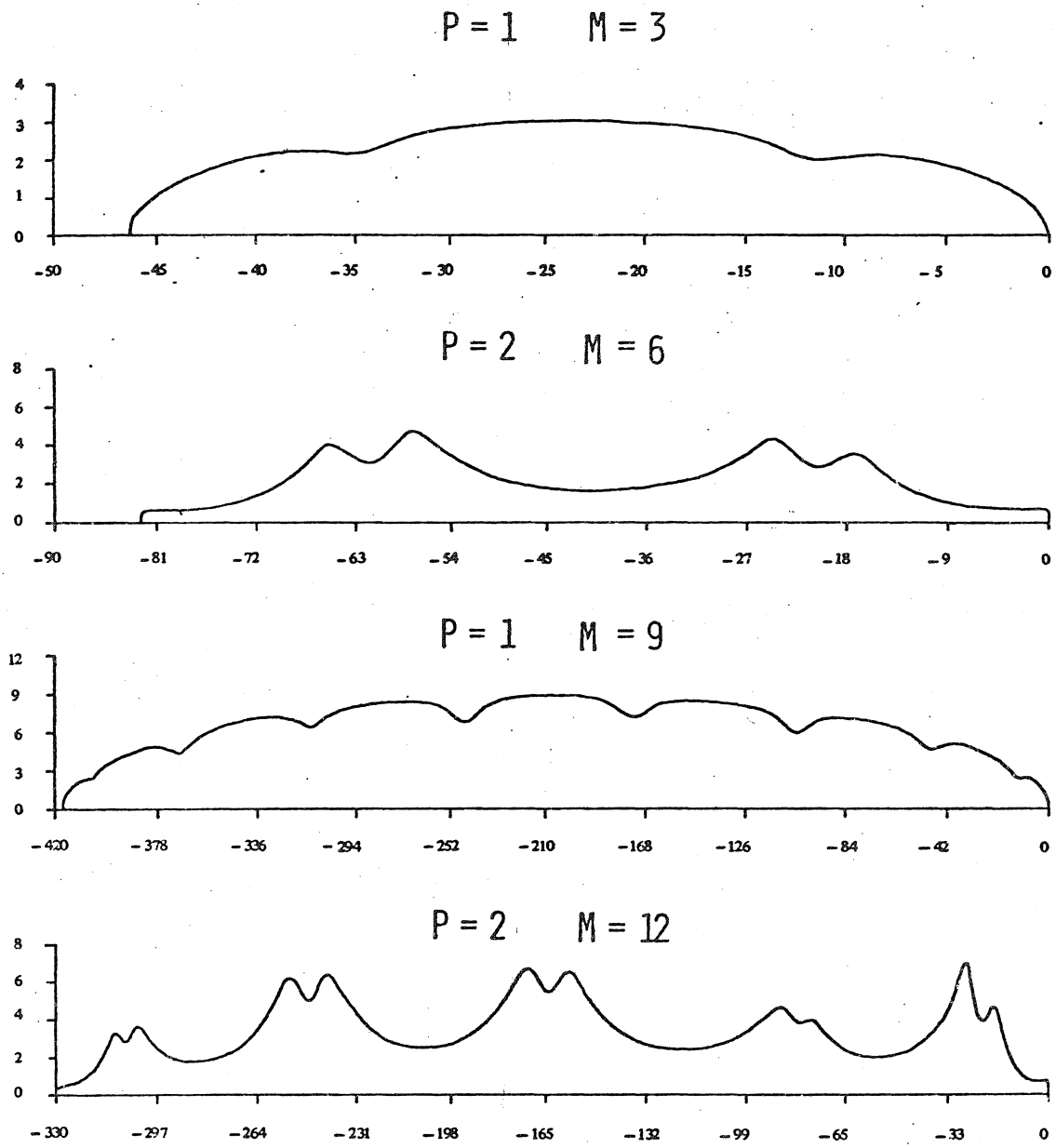

Fig. 3.1 Absolute stability regions

3.2. A CLASS OF FORMULAS OF FIRST AND SECOND ORDER

Once the coefficients s_i and p_i are given, integration parameters can be determined by using relations (2.11) - (2.12). From these relations it is easily seen that there exists more than one solution. We shall use this freedom by requiring that the local truncation error (2.4) satisfies a relation of the form

$$(3.15) \qquad y(x_{n+1}) - E[y(x_n), y(x_{n-1}), y(x_{n-2})] = C_{p+1}\, h^{p+1}\, y^{(p+1)}(x_n) + O(h^{p+2}),$$

where C_{p+1} is a constant. Representation (3.15) is convenient for a local error control based on interpolation with backward values. In the near future we intend to supply methods of type (2.1) with steplength and automatic error control based on such an interpolation.

If $p = 1$, relation (3.15) is always satisfied (see expansion (2.4)). In order to obtain such a relation for $p = 2$, we have to require

$$(3.16) \qquad C_{31} = C_{32}.$$

Because of the equality $y''' = f_j f^j_k f^k + f_{jk} f^j f^k$, relation (3.15) is then satisfied with $C_3 = C_{31}$. Observe that for linear equations the error constants C_2 and C_3 are always determined by the coefficients

s_i and p_i. As a consequence of (3.16), the same holds for a non-linear equation too. It turns out that the error constants are almost independent of m, and thus also the accuracy of the schemes. The constants are rather large and approximately satisfy $C_2 \simeq 1.27$, $C_3 \simeq 0.44$.

Next we shall express the parameters of the schemes into the coefficients s_i and p_i. As there exists more than one solution, it is recommended to look for a positive, or almost positive solution in order to avoid a possible cancellation of digits. For p = 1 an almost positive solution is easily found, because of the fact that all s_i and p_i are positive. On the other hand, for p = 2 such a solution does not exist because of the fact that all p_i are negative, and all s_i are positive. Therefore, for p = 2, we select a solution which reduces the computational effort. To that end we set

$$b_i = 0,\ i = 1,\ldots,m-2;\quad \lambda_{i,0} = 0,\ i = 2,\ldots,m.$$

By using the consistency relations for p = 2 and relations (2.11) - (2.12), an elementary calculation then yields that condition (3.16) is satisfied, if and only if

$$c_m = \frac{(1-\tfrac{1}{2}p_0)(p_1-2p_2+2p_3+2s_3)-(\tfrac{1}{2}+\tfrac{1}{4}p_0)^2}{2+p_1-2p_2+2p_3+2s_3}. \tag{3.17}$$

By performing some further elementary calculations, the remaining parameters can be solved from (2.11) - (2.12). Summarizing, we find the expressions:

$$b_i = 0, \quad i = 1,\ldots,m-2,$$

$$b_{m-1} = \frac{p_1 - c_m}{\lambda_{m,m-1}},$$

$$b_m = p_0;$$

$$c_i = \frac{p_{m+1-i}}{s_{m-i}}, \quad i = 1,\ldots,m-2,$$

$$c_{m-1} = \frac{p_2}{\lambda_{m,m-1}},$$

$$c_m = \frac{(1-\tfrac{1}{2}p_0)(p_1-2p_2+2p_3+2s_3)-(\tfrac{1}{2}+\tfrac{1}{4}p_0)^2}{2+p_1-2p_2+2p_3+2s_3}; \tag{3.18}$$

$$\lambda_{i,0} = 0, \quad i = 2,\ldots,m;$$

$$\lambda_{i,i-1} = \frac{s_{m+1-i}}{s_{m-i}}, \quad i = 1,\ldots,m-2,$$

$$\lambda_{m-1,m-2} = \frac{s_2}{\lambda_{m,m-1}},$$

$$\lambda_{m,m-1} = 1 - \tfrac{1}{2}p_0 - c_m.$$

In order to present the first and second order methods in a uniform way, we define the first order formulas also by expressions (3.18). The resulting parameter solution is almost positive.

Finally a remark about the range of the m-values. We restricted the m-values to $m \leq 12$, for we have to take into account the internal stability behaviour. To illustrate this, in table 3.1 we list the values $Q(5.15\ m^2)$, $m = 3,\ldots12$, for the first order formulas. The corresponding values $Q(2.29\ m^2)$ for the second order formulas are only slightly smaller. According to the in-

ternal stability condition (2.17), the smallest value of tolerance, allowed for a certain value of m, is then given by: tolerance = $Q(\beta)*$ arithmetic precision. Now suppose that the arithmetic precision is 14 digits, being a relevant machine precision nowadays. From table 3.1 it then follows that the local error may not be expected to be smaller than 10^{-5} if $m \geq 12$. Herewith it is assumed of course, that the maximally stable integration step is used. To our opinion a margin of 5 digits is acceptable, however it is recommended to choose the margin for the local accuracy not smaller.

As an illustration of the concept of internal stability we discuss an experiment for the simple linear system

$$
\begin{aligned}
y_1' &= (-2y_1+y_2+1)*10^4, \\
y_j' &= (y_{j-1}-2y_j+y_{j+1})*10^4, \quad j = 2,\ldots,99, \qquad (3.19) \\
y_{100}' &= (y_{99}-2y_{100}+1)*10^4.
\end{aligned}
$$

The Jacobian of (3.19) is well-known and a normal matrix. By prescribing the initial values $y_j(0) = 1$, $j = 1,\ldots,100$, we have the solution $y_j(x) \equiv 1$, $x \geq 0$. Thus in case of exact computations, i.e. no rounding errors occur, the integration schemes will yield the exact solutions, provided the parameters are exactly representable. In order to get an indication about the significance of the values $Q(\beta)$ we did perform one integration step with schemes of first and second order, for several values of m, using the steplength $h = \beta(m)/\sigma$. Here σ is equal to $4_{10}4$. The experiment has been carried out on a CDC 73/28 computer using an arithmetic precision of about 14 digits. Therefore the additional starting values are chosen as $1 + 10^{-14}*rn$, where rn denotes a random number between -1 and 1. In table 3.1 we have listed $10^{14} * \text{error}_p(m)$, where

$$\text{error}_p(m) = \max_j (y_j - 1.0), \quad p = 1,2.$$

The results of table 3.1 clearly indicate that the behaviour of the propagation of rounding errors with increasing m is in accordance with the error analysis of section 2.3. For the first order formulas, the internal stability condition for equation 3.19) is somewhat too pessimistic. On the other hand, for the second order formulas the internal stability condition is too optimistic. Here cancellation of digits appears, which is due to the fact that the second order formulas possess positive and negative parameters. This means that second order formulas of a high degree must be used with some caution.

m	$Q(\beta)$	$10^{14} * \text{error}_1(m)$	$10^{14} * \text{error}_2(m)$
3	$.1_{10}3$	$.9_{10}1$	$.9_{10}2$
4	$.7_{10}3$	$.6_{10}2$	$.4_{10}3$
5	$.4_{10}4$	$.4_{10}3$	$.1_{10}5$
6	$.3_{10}5$	$.1_{10}4$	$.8_{10}5$
7	$.2_{10}6$	$.7_{10}4$	$.6_{10}6$
8	$.9_{10}6$	$.3_{10}5$	$.1_{10}7$
9	$.5_{10}7$	$.2_{10}6$	$.9_{10}7$
10	$.3_{10}8$	$.2_{10}7$	$.2_{10}9$
11	$.2_{10}9$	$.3_{10}7$	$.4_{10}10$
12	$.7_{10}10$	$.4_{10}8$	$.4_{10}11$

Tabel 3.1

4. A NUMERICAL EXAMPLE

As already noted in section 3.2, in the near future we intend to supply the first and second order schemes defined by (3.18) with automatic error, steplength and order control. Intentionally we do not discuss these matters in this paper, because of the fact that constructing a formula and supplying an existing formula with control mechanisms leads to quite distinct problems. Therefore we discuss a numerical example where formulas are applied using a constant steplength.

We consider the non-linear initial-boundary value problem (cf. Sincovec & Madsen [5]):

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u\frac{\partial u}{\partial x}\right) - u^2, \quad 0 \le x \le 1, \quad t \ge 0,$$

$$u(t,0) = 50, \quad u_x(t,1) = 1 - \sin(u), \quad t \ge 0, \tag{4.1}$$

$$u(0,x) = 50, \quad 0 \le x \le 1.$$

By using central differencing with respect to x the initial-boundary value problem (4.1) can be semi-discretized to the initial value problem

$$u_1' = [-(2+2(\Delta x)^2)\,u_1^2 + u_2^2 + 2500]/2(\Delta x)^2,$$

$$u_j' = [u_{j-1}^2 - (2+2(\Delta x)^2)u_j^2 + u_{j+1}^2]/2(\Delta x)^2, \quad j = 2,\ldots,N-1, \tag{4.2}$$

$$u_N' = [2u_{N-1}^2 - (2+2(\Delta x)^2)u_N^2 + 4\,\Delta x\, u_N(1-\sin(u_N))]/2(\Delta x)^2,$$

$$u_j(0) = 50, \quad j = 1,\ldots,N,$$

where $u_j(t)$ approximates $u(t,x_j)$, and where $x_j = j\Delta x$, $\Delta x = 1/N$, N prescribed.

When applied to semi-discretized problems such as (4.2), the step-length for a stabilized explicit formula is usually limited by stability. Thus to apply such a formula efficiently a rough overestimate of $\sigma(J(y))$ is necessary. In fact, a program embodying an explicit method for semi-discretized problems will have to calculate $\sigma(Jy))$ and possess a control mechanism for it.

For equation (4.2) an estimation of the spectral radius σ is easily found by using elementary matrix theory. If $u_j(t) > 0$ for $t \geq 0$, the elements of the lower and upper diagonal of the tridiagonal Jacobian are positive. Thus, if $u_j(t) > 0$, the Jacobian has real eigenvalues for $t \geq 0$. Further, by applying Gershgorin's circle theorem a small investigation of the Jacobian yields

$$\sigma \simeq 4(\Delta x)^{-2} \max_j u_j(t).$$

At $t = 0$ we then have

$$\sigma \simeq 200(\Delta x)^{-2}. \tag{4.3}$$

For the calculations we assume that $\max_j u_j(t) \simeq 50$, $t \geq 0$. With this assumption, and the assumption that $u_j(t) > 0$, approximation (4.3) can be used for all t, and for all t we have real eigenvalues. After the integration of (4.2) both assumptions are easily verified.

We have carried out two experiments with problem (4.2) for $\Delta x = 1/30$. For problem (4.2) no analytical solution is available. Therefore, a computed reference solution at some selected times and points is given in table 4.1. In both experiments we applied a second order three-step scheme and, for comparison, a second order one-step scheme. The one-step is defined by (cf. section 2)

$$y_{n+1}^{(0)} = y_n,$$

$$y_{n+1}^{(j)} = y_n + \lambda_{j,j-1}\, h\, f(y_{n+1}^{(j-1)}),\ j = 1,\ldots,m, \tag{4.4}$$

$$y_{n+1} = y_{n+1}^{(m)}.$$

The integration parameters $\lambda_{j,j-1}$ are given by

$$\lambda_{j,j-1} = \bar{s}_{m+1-j}/\bar{s}_{m-j},\ j = 1,\ldots,m-2;\ \lambda_{m-1,m-2} = \tfrac{1}{2},\ \lambda_{m,m-1} = 1,$$

where $\bar{s}_j, j = 3,\ldots,m$, are coefficients of the strongly stable stability polynomial

$$\tilde{R}_m^{(2)}(z) = 1 + z + \tfrac{1}{2} z^2 + \bar{s}_3 z^3 + \ldots + \bar{s}_m z^m$$

of (4.4), which are listed in Van der Houwen [7, table 2.6.7']. The extrema of $\tilde{R}_m^{(2)}$ are bounded by 0.95 in the real stability interval.

Experiment I: The integration has been performed with schemes of degree 12 using the maximally stable steplength. Thus in this experiment we ignore any accuracy condition. Let h_1 and h_3 denote the steplength, and let β_1 and β_3 denote the real stability boundary of the one-step and three-step scheme, respectively. Then (cf. Van der Houwen [7, table 2.6.7'])

$$\beta_1 = 0.8 * 144 = 115.20,\ h_1 = \frac{115.20}{200} * (\Delta x)^{-2} = 0.00064,$$

and

$$\beta_3 = 2.29 * 144 = 329.76, \quad h_3 = \frac{329.96}{200} * (\Delta x)^{-2} = 0.001832.$$

The integration is stopped as soon as $t \geq 0.1$. At about $t = 0.1$ the steady state solution, i.e. the solution of the related two-point boundary value problem, is obtained. The additional starting vector for the three-step scheme are obtained from the computed reference solution. The start of the three-step scheme is counted as two integration steps. Then the three-step scheme needs 55 steps, and the one-step scheme needs 157 steps to reach $t = 0.1$. In table 4.2 we give relative errors with respect to the computed reference solution at some selected times and points. In order to calculate relative errors at the selected times, quadratic interpolation has been used between the solutions computed by the schemes. An error equal to zero means that after rounding the interpolated value is equal to the computed reference solution within the specified number of digits.

t \ x	0.0	0.2	0.4	0.6	0.8	1.0
.010	50.000	45.091	41.471	39.041	37.708	37.429
.025	50.000	44.506	40.253	37.262	35.577	35.229
.050	50.000	44.403	40.024	36.891	35.058	34.576
.100	50.000	44.383	39.979	36.816	34.952	34.442

Table 4.1. Reference solution for problem (4.2), $\Delta x = 1/30$.

	one-step					three-step				
t \ x	0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
.010	$2_{10}{}^{-4}$	$1_{10}{}^{-4}$	$4_{10}{}^{-4}$	$2_{10}{}^{-4}$	$5_{10}{}^{-4}$	$4_{10}{}^{-3}$	$4_{10}{}^{-3}$	$6_{10}{}^{-4}$	$3_{10}{}^{-3}$	$9_{10}{}^{-3}$
.025	$2_{10}{}^{-5}$	$3_{10}{}^{-5}$	$8_{10}{}^{-5}$	$3_{10}{}^{-5}$	$9_{10}{}^{-5}$	$7_{10}{}^{-4}$	$1_{10}{}^{-3}$	$7_{10}{}^{-4}$	$2_{10}{}^{-3}$	$3_{10}{}^{-3}$
.050	0	0	0	$3_{10}{}^{-5}$	$6_{10}{}^{-5}$	$5_{10}{}^{-5}$	$2_{10}{}^{-4}$	$2_{10}{}^{-4}$	$4_{10}{}^{-4}$	$5_{10}{}^{-4}$
.100	0	0	0	0	0	0	0	0	$3_{10}{}^{-5}$	$3_{10}{}^{-5}$

Table 4.2. Results of experiment I.

From table 4.2 we conclude that the schemes yield the steady state solution at $t = 0.1$ with almost the same accuracy. In the initial phase of the integration the one-step scheme yields more accurate results than the three-step scheme. This is what we expect, as $h_3 \simeq 3h_1$. In order to get some indication about the accuracy behaviour of our three-step methods, when compared with stabilized one-step methods, we did another integration with two schemes using the same steplength. Before discussing this integration in experiment II, we observe that we only consider results of the time integrations. For the reference solution, given in table 4.1, belongs to the system of ordinary differential equations (4.2).

Experiment II: Problem (4.2) has been integrated with a one-step scheme and a three-step scheme with steplength $h = 0.0005$ over the interval $[0,0.1]$. As observed in section 3.2, the accuracy of the three-step schemes is independent of the degree. The same holds for the one-step schemes. Therefore it is allowed to select the degree m of the schemes in such a way that

(4.5) $\qquad h\sigma \leq \beta(m), \quad m$ minimal.

Let m_1 and m_3 denote the degree of the one-step and three-step scheme that satisfy (4.5), respectively. According to Van der Houwen [7, table 2.6.7'], there holds $\beta_1(10) \simeq 79.70$, $\beta_1(11) \simeq 96.66$. With the specified h and σ we then find $m_1 = 11$ and $m_3 = 7$.

The additional starting vectors for the three-step scheme are obtained from the computed reference solution. The start of the three-step scheme is again counted as two integration steps. Then the three-step scheme needs 1400, and the one-step scheme 2200 function evaluations to reach t = 0.1. Relative errors are listed in table 4.3 and are computed in the same way as in experiment I.

	one-step					three-step				
t \ x	0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
.010	$2_{10}{}^{-5}$	$5_{10}{}^{-5}$	$3_{10}{}^{-5}$	$2_{10}{}^{-4}$	$5_{10}{}^{-5}$	$9_{10}{}^{-5}$	$2_{10}{}^{-4}$	$3_{10}{}^{-4}$	$5_{10}{}^{-4}$	$5_{10}{}^{-4}$
.025	0	0	0	$3_{10}{}^{-5}$	0	$2_{10}{}^{-5}$	$5_{10}{}^{-4}$	$1_{10}{}^{-4}$	$1_{10}{}^{-4}$	$2_{10}{}^{-4}$
.050	0	0	0	$3_{10}{}^{-5}$	$6_{10}{}^{-5}$	0	0	0	$3_{10}{}^{-5}$	$9_{10}{}^{-5}$
.100	0	0	0	0	0	0	0	0	0	0

Table 4.3. Results of experiment II.

From table 4.3 we conclude that the results of the one-step integration are more accurate than the results of the three-step integration. This is corroborated by other practical experiments. However, due to condition (4.5), in many situations the degree of the three-step scheme can be chosen smaller. With other words, the three-step scheme may integrate with a smaller step-length than the one-step scheme with the same computational effort. This

may lead to comparable, or even higher accuracy. As an illustration, problem (4.2) has been integrated another time with the three-step scheme of degree four using the steplength h = 1/5500 over the interval [0,.1]. This integration also costs 2200 function evalutions. The relative errors are listed in table 4.4.

t \ x	0.2	0.4	0.6	0.8	1.0
.010	0	$2_{10}{}^{-5}$	$5_{10}{}^{-5}$	$8_{10}{}^{-5}$	$8_{10}{}^{-5}$
.025	0	0	$3_{10}{}^{-5}$	0	$3_{10}{}^{-5}$
.050	0	0	0	0	$6_{10}{}^{-5}$
.100	0	0	0	0	0

Table 4.4. Results of the three-step scheme with m = 4.

Summarizing, when the steplength is completely determined by stability and not by accuracy conditions, the three-step schemes can integrate with stepsizes which are nearly three times larger than the stepsizes allowed for the one-step schemes. If the steplength is completely determined by accuracy, the one-step schemes will generally yield more accurate results than the three-step schemes when using the same steplength. With respect to computational efficiency however, we expect that in general the three-step schemes can compete with the one-step schemes when an accurate time integration is necessary.

Perhaps it is needless to say that the integration formulas discussed in this paper, can equally be applied to parabolic problems in two, or

possibly three dimensions. The only mathematical restriction is that the eigenvalues of the Jacobian of the system of ordinary differential equations, which exists by semi-discretization, are almost real. For a lot of problems, this means no restriction. A practical restriction on the use of stabilized explicit methods may arise when $\sigma(J(y))$ is extremely large. In spite of the relatively large stability boundaries, very small timesteps are then required to maintain stability. In particular, this disadvantage applies when the integration has to be executed over a relatively large interval. In such a situation it may be necessary to apply an unconditionally stable method. For one-dimensional problems this is easy to realize (see e.g. Sincovec & Madsen [5]). However, for multi-dimensional problems direct-grid methods, such as ADI, are in general difficult to implement, whereas explicit methods are very easy to implement.

ACKNOWLEDGEMENTS. The author wishes to thank Prof. P.J. Van der Houwen for his constructive comments during the preparation of this paper. The author is also grateful to Mr C. den Heyer for his careful reading of the manuscript.

REFERENCES

[1] C.W. Gear, Hybrid methods for initial value problems in ordinary differential equations, J. SIAM Numer., Anal., Ser. B, 2(1964), pp. 69-86.

[2] P. Henrici, Discrete variable methods in ordinary differential equations, John Wiley & Sons, New York, 1962.

[3] J.D. Lambert, Computational methods in ordinary differential equations, John Wiley & Sons, London, 1973.

[4] R.D. Richtmeyer & K.W. Morton, Difference methods for initial value problems, Interscience, New York, 1967.

[5] R.F. Sincovec N.K. Madsen, Software for nonlinear partial differential equations, ACM Transactions on mathematical software, 1(1975), pp. 232-260.

[6] P.J. van der Houwen, Explicit Runge-Kutta formulas with increased stability boundaries, Numer. Math., 20(1972), pp. 149-164.

[7] P.J. van der Houwen, Construction of integration formulas for initial value problems, North-Holland Publishing Company, Amsterdam, 1976.

[8] J.G. Verwer, Multipoint multistep Runge-Kutta methods I: On a class of two-step methods for parabolic equations, Report NW 30/76, Mathematical Centre, Amsterdam, 1976.

[9] J.G. Verwer, Multipoint multistep Runge-Kutta methods II: The construction of a class of stabilized three-step methods for parabolic equations, NW 31/76, Mathematical Centre, Amsterdam, 1976.

[10] J.M. Watt, The asymptotic discretization error of a class of methods for solving ordinary differential equations, Proc. Camb. Phil. Soc., 61(1967), pp. 461-472.