stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE NW 58/78 AUGUSTUS
(DEPARTMENT OF NUMERICAL MATHEMATICS)

J.G. VERWER

# THE APPLICATION OF ITERATED DEFECT CORRECTION TO THE LOD METHOD FOR 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, 65M05, 65M20.

# The application of iterated defect correction to the LOD method for parabolic equations[*)]

by

J.G. Verwer

## ABSTRACT

The paper is concerned with the numerical solution of the initial boundary value problem for a class of multi-dimensional parabolic partial differential equations. In particular the time-integration of semi-discrete equations is investigated. An attempt is made to develop integration formulas being computationally attractive and of high accuracy, while possessing unconditional stability properties. To that end iterated defect correction is applied to the LOD method. The convergence properties of this process are investigated. Numerical experiments are reported.

KEY WORDS & PHRASES: *Numerical analysis, Parabolic partial differential equations, Locally one-dimensional method, Iterated defect correction.*

[*)] This report will be submitted for publication elsewhere.

## 1. INTRODUCTION

Let $\Omega$ denote a bounded and path-connected region in the k-dimensional $(x_1,\dots,x_k)$-space with sides parallel to the coordinate axes. Let $\delta\Omega$ be the boundary, and consider the parabolic partial differential equation of the *non-linear* type

$$(1.1) \qquad u_t = \sum_{i=1}^{k} F_i(t,x_1,\dots,x_k,u,u_{x_i},u_{x_ix_i}),$$

defined in the product set $(0,T] \times \Omega$. Let a boundary condition be given in the form

$$(1.2) \qquad a(t,x_1,\dots,x_k)u + b(t,x_1,\dots,x_k)u_n = c(t,x_1,\dots,x_k),$$

$(t,x_1,\dots,x_k) \in (0,T] \times \delta\Omega$, $u_n$ normal derivative, and assume an initial function is given at $t = 0$. In this paper we are concerned with the numerical solution of this initial boundary value problem when brought in an explicit, *semi-discretized* form, i.e., we primarily discuss the numerical integration of the system of *ordinary differential equations*

$$(1.3) \qquad y' = f(t,y), \quad t \in (0,T], \quad y(0) = y_0,$$

being obtained from discretizing the space variables in (1.1) - (1.2). In particular it is assumed that f satisfies the *linear splitting* relation [10]

$$(1.4) \qquad f(t,y) = \sum_{i=1}^{k} f_i(t,y),$$

where each *splitting function* $f_i$ approximates the operator $F_i$, which contains only space derivatives with respect to the variable $x_i$. This assumption can always be satisfied by semi-discretizing on a rectilinear grid with grid lines parallel to $\delta\Omega$. It is further assumed that in each direction we have a 3-*point coupling* at internal grid points and a 2-point coupling at points nearest to the boundary. This can be achieved by using 3-point symmetrical finite differences at internal grid points and 2-point or 3-point non-symmetrical finite differences at the other points [7]. In many cases it is also possible to satisfy our assumption with a finite element semi-descritization [2].

The paper has been written in order to investigate the application of iterated defect correction [3, 9] to the locally one-dimensional splitting formula [10, 14]

$$
\begin{aligned}
&y_{(0)} = y_\nu, \\
(1.5) \qquad &y_{(i)} = y_{(i-1)} + \tau f_i(t_{\nu+1}, y_{(i)}), \qquad i = 1(1)k, \\
&y_{\nu+1} = y_{(k)}.
\end{aligned}
$$

In this one-step integration formula $\tau = t_{\nu+1} - t_\nu$ denotes the steplength and $y_\nu$ denotes an approximation to the exact solution $y(t)$ of (1.3) at $t = t_\nu$. It is easy to see that the order of consistency of (1.5) is equal to 1 for every splitting (1.4). Observe that if $k = 1$, (1.5) reduces to the implicit Euler formula. For clarity, throughout the paper parenthesized subindices refer to intermediate results and not to approximations at step points $t_\nu$.

The purpose of the investigation is to find integration formulas for systems (1.3) - (1.4), which are more *accurate* than the LOD formula (1.5) and which possess its attractive *unconditional stability* property [10, 14], as well as its advantage of being *computationally attractive* (per integration step). The idea of iterated defect correction, when applied to (1.5), may lead to such integration formulas.

In our investigation we adopt the approach followed by Frank & Ueberhuber [3]. They investigated iterated defect correction for the efficient solution of stiff systems. Their basic formula is implicit Euler. Because our splitting formula is closely related to implicit Euler, many of their results carry over.

## 2. SOME PRELIMINARIES CONCERNING THE SPLITTING FORMULA

For future reference we discuss some properties of the splitting formula (1.5). When applied to linear systems (1.3), i.e.

$$(2.1)\qquad y' = Jy, \quad J = \sum_{i=1}^{k} J_i,$$

$J$ and $J_i$ constant matrices, (1.5) reads

$$(2.2)\qquad y_{\nu+1} = Ry_\nu,$$

where the *amplification matrix* R is given by the formal relation

$$(2.3)\qquad R = \prod_{i=1}^{k} (I - \tau J_{k-i+1})^{-1}.$$

We shall use the notation R for the amplification matrix (2.3), but also for the function $R\colon \mathbb{C}^k \to \mathbb{C}$,

$$R(z_1,\ldots,z_k) = \prod_{i=1}^{k} (1 - z_{k-i+1})^{-1},$$

which is called the *stability function* of the LOD formula [10]. It will be clear from the context, whether the amplification matrix, or the stability function is meant.

In the discussion of IDEC (iterated defect correction), we consider systems of the special type

$$(2.4)\qquad y' = g(t,y) = f(t,y) + d(t),$$

f satisfying (1.4) and d(t) being the *defect function*. For these systems we define the splitting functions $g_i$ by

$$(2.5)\qquad \begin{aligned} g_1(t,y) &= f_1(t,y) + d(t), \\ g_i(t,y) &= f_i(t,y), \qquad i = 2(1)k. \end{aligned}$$

For systems (2.4) - (2.5) formula (1.5) then yields

$$(2.6)\qquad \begin{aligned} y_{(1)} &= y_\nu + \tau[f_1(t_{\nu+1},y_{(1)}) + d(t_{\nu+1})], \\ y_{(i)} &= y_{(i-1)} + \tau f_i(t_{\nu+1},y_{(i)}), \qquad i = 2(1)k, \\ y_{\nu+1} &= y_{(k)}. \end{aligned}$$

For the linear system

$$(2.7)\qquad y' = Jy + d(t), \qquad J = \sum_{i=1}^{k} J_i,$$

(2.6) reduces to

$$(2.8)\qquad y_{\nu+1} = R[y_\nu + \tau d(t_{\nu+1})].$$

Observe that if the implicit Euler method is applied to (2.7), we have

$$(2.9)\qquad y_{\nu+1} = (I - \tau J)^{-1}[y_\nu + \tau d(t_{\nu+1})].$$

Because of the fact that d(t) is added to the first splitting function $f_1$, (2.8) and (2.9) only differ in the amplification matrix.

In case of non-linear problems the calculation of $y_{(i)}$, $i = 1(1)k$, in formula (1.5), or (2.6), involves the solution of a system of non-linear equations. In actual applications it is of no use to solve these systems very accurately, as the LOD formula is only of *first order*. At each stage i we perform 1 *Newton-type iteration* with predictor $y_{(i-1)}$. Formula (1.5) is then replaced by

$$\begin{aligned} & y_{(0)} = y_\nu, \\ (2.10)\qquad & y_{(i)} = y_{(i-1)} + \tau(I - \tau\bar{J}_i)^{-1} f_i(t_{\nu+1}, y_{(i-1)}), \qquad i = 1(1)k, \\ & y_{\nu+1} = y_{(k)}, \end{aligned}$$

where $\bar{J}_i$ are *tridiagonal* matrices approximating the partial derivatives $\partial f_i/\partial y$ at the point $(t_\nu, y_\nu)$. A similar formula then replaces (2.6). If the problem is linear, i.e. $\partial f_i/\partial y$ constant, $\bar{J}_i$ is always assumed to be equal to this constant derivative. We now proceed with formula (2.10), and the similar formula replacing (2.6). Formula (2.10) is also *first order* consistent, and is identical to (1.5) for linear equations (2.1).

## 3. THE IDEC-PROCESS FOR THE SPLITTING FORMULA

In this section we shortly describe the IDEC for the splitting formula (2.10). Details are omitted, as these are clearly discussed in [3]. For convenience we employ the same notation.

Let the solution of (1.3) be required on the interval $[0,T]$. Introduce the sequence of subintervals $[H_i,H_{i+1}]$, not necessarily equidistant, where $H_0 = 0$ and $H_{i_{max}} = T$ for a suitable integer $i_{max}$. We now restrict the discussion to the first subinterval $[0,H_1]$, on which we define the equidistant step points

$$(3.1) \qquad t_0 = 0, \quad t_\nu = \frac{\nu H_1}{m}, \quad \nu = 1(1)m,$$

where $1 \leq m \leq 4$. For practical reasons we do not consider values of $m > 4$.

Let j denote the iteration index of the IDEC. The process then consists of the following steps:

$1^o$. Set $j = 0$. Apply the splitting formula (2.10) to (1.3) on the grid (3.1) to obtain the row $\eta^0 = [\eta_0^0,\ldots,\eta_m^0]$ of approximation vectors $\eta_\nu^0$. The vectors $\eta_{\nu+1}^0$, $\nu = 0(1)m-1$, are thus defined by the scheme

$$\psi_{(0)} = \eta_\nu^0,$$

$$(3.2) \qquad \psi_{(i)} = \psi_{(i-1)} + \tau(I - \tau\bar{J}_i)^{-1} f_i(t_{\nu+1},\psi_{(i-1)}), \qquad i = 1(1)k,$$

$$\eta_{\nu+1}^0 = \psi_{(k)}.$$

where $\eta_0^0 = y_0$ and $\tau = H_1/m$.

$2^o$. Define the defect function

$$(3.3) \qquad d^j(t) = (P^j)'(t) - f(t,P^j(t)),$$

where $P^j(t)$ is the vector polynomial of degree $\leq m$ interpolating $\eta^j$, i.e.

$$(3.4) \qquad P^j(t_\nu) = \eta_\nu^j, \quad \nu = 0(1)m,$$

and compute the defects

$$d^j(t_\nu) = (P^j)'(t_\nu) - f(t_\nu,\eta^j_\nu), \qquad \nu = 1(1)m. \tag{3.5}$$

3$^o$. Apply the splitting formula (2.10) on the grid (3.1) to the initial value problem

$$y' = f(t,y) + d^j(t),\ t > 0,\ y(0) = y_0, \tag{3.6}$$

to obtain the row $\pi^j = [\pi^j_0,\dots,\pi^j_m]$. Thus the vectors $\pi^j_\nu$ are defined by

$$\begin{aligned} \bar\psi_{(1)} &= \pi^j_\nu + \tau(I-\tau\bar J_1)^{-1}[f_1(t_{\tau+1},\pi^j_\nu) + d^j(t_{\nu+1})], \\ \bar\psi_{(i)} &= \bar\psi_{(i-1)} + \tau(I-\tau\bar J_i)^{-1}f_i(t_{\nu+1},\bar\psi_{(i-1)}), \qquad i = 2(1)k, \\ \pi^j_{\nu+1} &= \bar\psi_{(k)}, \qquad \nu = 0(1)m-1, \end{aligned} \tag{3.7}$$

where $\pi^j_o = y_0$.

4$^o$. Improve, i.e. compute the (j+1)-th approximation row $\eta^{j+1}$ by

$$\eta^{j+1} = \eta^0 + \eta^j - \pi^j. \tag{3.8}$$

5$^o$. Increase j and proceed with 2$^e$.

We apply the *local connection* strategy [3], i.e., after the last iteration step on $[0,H_1]$ we simply repeat the whole process on $[H_1,H_2]$, and so on.

The polynomial $P^j(t)$ need not to be calculated explicitely. Its value and derivative values are only required at step points $t_\nu$, where $P^j(t_\nu) = \eta^j_\nu$. The values $(P^j)'(t_\nu)$ are easily determined from differentiation of Lagrange's formula [1, p. 878]. In our case we obtain weighted sums of the type

$$(P^j)'(t_\nu) = \tau^{-1}\sum_{\kappa=0}^{m} w_{\nu\kappa}P^j(t_\kappa), \qquad \nu = 1(1)m, \tag{3.9}$$

$w_{\nu\kappa}$ constant. For $m \leq 4$ these weights are given in [1, p. 914].

## 4. THE FIXED POINT OF THE IDEC

Again we consider the IDEC on the first subinterval $[0,H_1]$. Its fixed point is characterized by the following theorem which is the analogy of theorem 4.1 in [3]:

THEOREM 4.1. *Let* $\eta_0^* = y_0$. *The row* $\eta^* = [\eta_0,\ldots,\eta_m]$ *is a fixed point of the* IDEC *based on the splitting formula* (2.10), *iff*

$$d^*(t_\nu) = 0, \qquad \nu = 1(1)m,$$

*where*

$$d^*(t) = (P^*)'(t) - f(t,P^*(t)),$$

*and* $P^*(t)$ *interpolates* $\eta^*$.

The proof of this theorem is analogous to the proof of the corresponding theorem in [3]. The fact that our IDEC is based on formula (2.10), and not on implicit Euler, is of no importance for the proof.

From this theorem it follows that if the IDEC is iterated until convergence, we in fact apply the *polynomial collocation method* corresponding to the step points (3.1), which, in turn, may be interpreted as the *fully implicit Runge-Kutta method* (see e.g. [5, 13] and appendix 1 of [12])

$$(4.1) \qquad \eta_\nu^* = y_0 + \tau \sum_{\kappa=1}^{m} \bar{w}_{\nu\kappa} f(t_\kappa,\eta_\kappa^*), \qquad \nu = 1(1)m.$$

The coefficient matrix $\bar{W} = (\bar{w}_{\nu\kappa})_{\nu,\kappa=1(1)m}$ is the inverse of the weight matrix $W = (w_{\nu\kappa})_{\nu,\kappa=1(1)m}$. If $m > 1$, methods of this type are also called block methods. Each result $\eta_\nu^*$ is m-*th order consistent*, i.e. the local truncation errors are of order m+1 in $\tau$. If $m = 1$, (4.1) represents the implicit Euler method. An important feature of these methods is that they possess attractive stability properties for the integration of semi-discrete parabolic equations. When applied to the stability test-model

(4.2) $$s' = \lambda s, \quad t > 0, \quad s(0) = s_0, \quad \lambda \in \mathbb{C},$$

each scalar $\eta_\nu^*$ can be expressed as

(4.3) $$\eta_\nu^* = \phi_\nu(z)s_0, \quad z = m\tau\lambda,$$

$\phi_\nu$ being a rational function satisfying

(4.4) $$\phi_\nu(z) \sim 1/z, \ \mathrm{Re}(z) \to -\infty.$$

The stability function of the method is $\phi_m$. Stability regions $\{z \in C \,|\, |\phi_m(z)| < < 1\}$ for $m \leq 10$ are given in fig. 6 of [3]. All regions contain the whole negative axis and the greater part of the negative half-plane. In case of parabolic equations the $\lambda$-values are usually situated in a long narrow strip around the negative axis. As a consequence, formulas (4.1) possess unconditional stability properties for semi-discrete parabolic equations. The fact that $\phi_\nu(z) \sim 1/z$, $\mathrm{Re}(z) \to -\infty$, is of importance for the convergence of the IDEC. We return to these points in section 6.

## 5. THE ORDER OF CONSISTENCY OF THE ITERATES

General results concerning the order of consistency of IDEC iterates for one-step Runge-Kutta methods are given in Frank & Ueberhuber [4]. They show that if the order of consistency of the basic method is equal to p, the order of each IDEC iterate $\eta_\nu^j$ equals $\min(p(j+1),m)$. Hence, if $p = 1$, the order of consistency equals $\min(j+1,m)$. The results of Frank and Ueberhuber are proved using the theory of asymptotic expansions. If m is not too large, say $m \leq 4$, it is feasible to obtain this result in a more direct way, viz. by using elementary Taylor expansions. In this way it can be shown that if the basic formula is (2.10), the order of consistency of $\eta_\nu^j$ is also equal to $\min(j+1,m)$, $\nu = 1(1)m$. We did not investigate whether the theory of Frank and Ueberhuber can be used in our situation.

## 6. CONVERGENCE OF THE IDEC

We investigate the convergence for linear equations of the type (2.1), i.e.

$$(6.1) \qquad y' = Jy, \quad J = \sum_{i=1}^{k} J_i.$$

Again we consider the IDEC on the first subinterval $[0,H_1]$. When applied to (6.1), the IDEC may then be interpreted as the recurrence relation

$$(6.2) \qquad \tilde{\eta}^{j+1} = S\tilde{\eta}^{j} + V$$

for the vector of approximation vectors $\tilde{\eta}^j = [\eta_1^j,\ldots,\eta_m^j]^T$, where V is a constant m-block vector of length mn and S a constant m×m-block matrix of order mn, if n is the order of J.

LEMMA 6.1. *Let* R *be given by* (2.3). *Let* $S_{\nu\kappa}$, $1 \le \nu,\kappa \le m$ *denote the* $(\nu,\kappa)$*-th block of* S *and* $V_\nu$, $\nu = 1(1)m$, *the* $\nu$*-th block of* V. *Then*

$$S_{\nu\kappa} = -\sum_{\mu=1}^{\nu} w_{\mu\kappa} R^{\nu+1-\mu}, \qquad 1 \le \nu < \kappa \le m,$$

$$(6.3) \qquad S_{\nu\kappa} = \delta_{\nu\kappa} I - \sum_{\mu=1}^{\nu} w_{\mu\kappa} R^{\nu+1-\mu} + R^{\nu+1-\kappa}\tau J, \qquad 1 \le \kappa \le \nu \le m,$$

$$V_\nu = -\sum_{\mu=1}^{\nu} w_{\mu 0} R^{\nu+1-\mu} y_0,$$

*where* $\delta_{\nu\kappa}$ *is the Kronecker-symbol and the weights* $w_{\mu\kappa}$ *are given in* (3.9).

The proof of this lemma is completely analogous to the proof given in [3, appendix 1]. Obviously, we have convergence if the *spectral radius*

$$(6.4) \qquad \sigma(S) < 1.$$

To be able to investigate this spectral radius in a systematic way, we now impose the additional restrictions:

(6.5) $1^o$. *The matrices* $J_i$ *share the same eigensystem.*

$2^o$. *The matrices* $J_i$ *are symmetric and negative definite.*

In fact, we now consider the *test-model* being usually investigated in the stability analysis of splitting methods [10].

Let X denote the eigensystem of J, and let $\Lambda$ and $\Lambda_i$ be the diagonal matrices of eigenvalues of J and $J_i$, respectively, i.e.

$$\Lambda = \sum_{i=1}^{k} \Lambda_i, \quad J = X\Lambda X^{-1}, \quad J_i = X\Lambda_i X^{-1}. \tag{6.6}$$

From (6.6) it follows that

$$S = \tilde{X}\tilde{S}\tilde{X}^{-1}, \tag{6.7}$$

$\tilde{X}$ being the m×m-block diagonal matrix consisting of the blocks X, and $\tilde{S}$ the m×m-block matrix consisting of the $m^2$ diagonal blocks $\tilde{S}_{\nu\kappa} = X^{-1}S_{\nu\kappa}X$, $\nu,\kappa = 1(1)m$. The diagonal blocks $\tilde{S}_{\nu\kappa}$ are in fact defined by (6.3), if in the expressions J is replaced by $\Lambda$. Hence, if $\lambda$ denotes an eigenvalue of J, then

$$\sigma(S) = \sigma(\tilde{S}) = \max_{\lambda} \sigma(E_\lambda), \tag{6.8}$$

$E_\lambda$ being the m×m-matrix defined by expressions (6.3), if J is replaced by $\lambda$ and the matrix (2.3) by the scalar expression

$$R(\tau\lambda_1,\dots,\tau\lambda_k) = \prod_{i=1}^{k} (1-\tau\lambda_i)^{-1}, \tag{6.9}$$

where $\lambda_i$ is the corresponding eigenvalue of $J_i$, i.e. $\lambda = \lambda_1 + \dots + \lambda_k$. For a given $\lambda$, the eigenvalues $\lambda_1,\dots,\lambda_k$ are defined by the splitting of J. In the sequel we denote $z = \tau\lambda$ and $z_i = \tau\lambda_i$.

We now proceed with the investigation of $\sigma(E_\lambda)$ for $z < 0$, $z_i < 0$, $i = 1(1)k$, and arbitrary splittings $z = z_1 + \dots + z_k$. Let us begin with the simple case m = 1. We then have (in (6.3) $w_{11} = 1$ for m = 1)

$$(6.10) \qquad E_\lambda = 1 - (1-z)R(z_1,\ldots,z_k) = 1 - (1 - \sum_{i=1}^{k} z_i)/\prod_{i=1}^{k} (1-z_i).$$

It is easily seen that $0 < \sigma(E_\lambda) < 1$ for $z_i < 0$, provided $k \geq 2$, More precisely, if all $z_i \to 0$, then $\sigma(E_\lambda) \to 0$. On the other hand, if all $z_i \to -\infty$, then $\sigma(E_\lambda) \to 1$.

If $m > 1$, explicit expressions for $\sigma(E_\lambda)$ are not available. For the two limit cases we have the following theorem:

THEOREM 6.1. a) *For all* $m \geq 1$ *and* $k \geq 2$ *there holds:* $\lim \sigma(E_\lambda) = 0$ *for* $z_i \to 0$, $i = 1(1)k$.
b) *For all* $m \geq 1$ *and* $k \geq 2$ *there holds:* $\lim \sigma(E_\lambda) = 1$ *for* $z_i \to -\infty$, $i = 1(1)k$.

PROOF. a) The elements of $E_\lambda$ depend continuously on $z_i$. If $z_i = 0$, $i = 1(1)k$, is substituted, we obtain the matrix $S_m(0)$ investigated in theorem 5.1 of [3]. That theorem states that $\sigma(S_m(0)) = 0$.
b) If all $z_i \to -\infty$, then $R(z_1,\ldots,z_k)$ and $zR(z_1,\ldots,z_k)$ both tend to zero. Hence, if all $z_i \to -\infty$, the matrix $E_\lambda$ tends to the m-th order unit matrix. This property holds for all $k \geq 2$. □

REMARK. In theory, iterated defect correction can also be applied to other types of splitting methods, e.g., alternating direction methods. The stability functions $R(z_1,\ldots,z_k)$ of such methods, however, do not vanish at infinity (see [10]). Consequently, for such methods no IDEC convergence shall occur.

Let us now temporarily assume that indeed $\sigma(E_\lambda) < 1$ for $z_i < 0$, $i = 2(1)k$ and $m \geq 1$. It then follows from the preceding theorem that for small negative $z_i$-values a rapid IDEC convergence will occur, whereas for the larger ones the convergence is expected to be slow. We shall consider this point in more detail. The initial approximation $\tilde{\eta}^0$ in (6.2) reads

$$(6.11) \qquad \tilde{\eta}^0 = [Ry_0,\ldots,R^m y_0]^T,$$

where R is defined by (2.3). The fixed point, say $\tilde{\eta}^*$, is defined by (6.2), i.e. $\tilde{\eta}^* = (I-S)^{-1}V$. In section 4 it was pointed out that, in case of equation (6.1), the ν-th fixed point vector $\tilde{\eta}^*_\nu$ can be expressed as

(6.12) $$\tilde{\eta}^{*}_{\nu} = \phi_{\nu}(\tau mJ)y_{0}, \qquad \nu = 1(1)m,$$

$\phi_{\nu}$ being the rational function of expression (4.3). Hence, the $\nu$-th initial iteration error, say $\tilde{\varepsilon}^{0}_{\nu} = \tilde{\eta}^{*}_{\nu} - \tilde{\eta}^{0}_{\nu}$, reads

(6.13) $$\tilde{\varepsilon}^{0}_{\nu} = [\phi_{\nu}(\tau mJ) - R^{\nu}]y_{0}, \qquad \nu = 1(1)m,$$

or, equivalently,

(6.14) $$X^{-1}\tilde{\varepsilon}^{0}_{\nu} = [\phi_{\nu}(\tau m\Lambda) - \prod_{i=1}^{k} (I - \tau\Lambda_{i})^{-\nu}]X^{-1}y_{0}.$$

As $\phi_{\nu}(z) \sim 1/z$, $z \to -\infty$, components of $X^{-1}y_{0}$ belonging to large negative $\lambda$-values are *damped*. Often, these components (approximations to Fourier coefficients belonging to *higher harmonics*, see the example) are already small in the initial vector $y_{0}$. Hence, at least in the initial phase of the iteration, it is expected that the decrease of the iteration error $\tilde{\varepsilon}^{j} = [\tilde{\varepsilon}^{j}_{1},\ldots,\tilde{\varepsilon}^{j}_{m}]$, satisfying

$$\tilde{\varepsilon}^{j} = S^{j}\tilde{\varepsilon}^{0},$$

is governed by the small $z_{i}$-values. Unfortunately, the speed of convergence *decreases* if j increases. This is due to the fact that during the iteration $\tilde{\varepsilon}^{j}$ tends to lie in subspaces spanned by dominant eigenvectors of S. For these eigenvectors the convergence is slow. Consequently, it is of no use to perform a large number of iterations. We illustrate this phenomenon in the example at the end of this section.

For $k = 2$ and $m = 1(1)4$, we computed $\sigma(E_{\lambda})$ numerically at the set of points $(z_{1},z_{2})$, $z_{i} = -\ell/4$, $\ell = 0(1)80$. Observe that $E_{\lambda}(z_{1},z_{2}) = E_{\lambda}(z_{2},z_{1})$. All computed $\sigma$ are smaller than one. For almost all fixed values of $z_{1}$, the maximal $\sigma$ is found for $z_{1} = z_{2}$. These maximum values are given in fig. 6.1 which shows that for small z-values the speed of convergence decreases with increasing m. For large negative z-values the speed does not change with m.

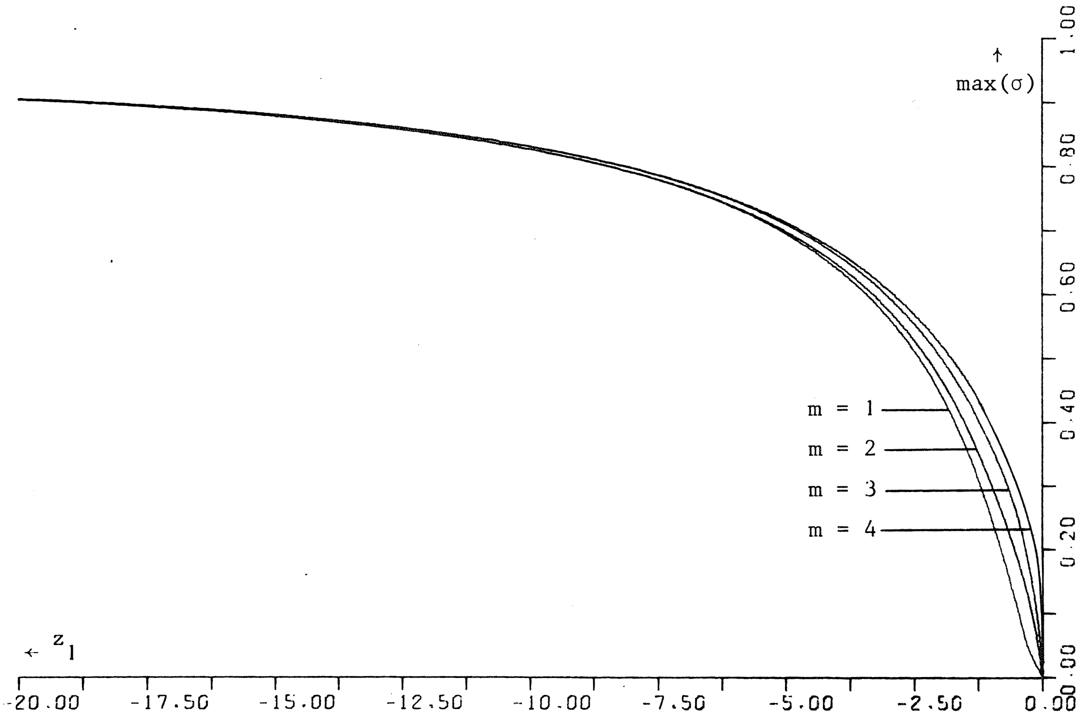

fig. 6.1. maximal σ-curves

An illustrative example

Let the integer $N \geq 1$, and denote $h = 1/(N+1)$. Let equation (6.1), with $k = 2$, originate from semi-discretization of $u_t = u_{x_1x_1} + u_{x_2x_2}$, defined on $(0,T] \times \{(x_1,x_2) \mid 0 < x_1,x_2 < 1\}$ with $u = 0$ on the boundary and initial function $s(x_1,x_2)$. Assume that the semi-discretization has been performed on a uniform grid of size h with second order symmetrical finite differences. The LOD matrices $J_1$ and $J_2$ are then given by $J_1 = I \otimes A$, $J_2 = A \otimes I$, where I denotes the unit matrix of order N, A the matrix of order N being defined by

$$(6.15) \qquad A = h^{-2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix},$$

and the symbol ⊗ denoting direct product as defined in [6, p. 216] (a discussion of the test-model can be found in appendix 2 of [12]). For the present matrices $J_i$ restrictions (6.5) are easily checked. The eigenvalues of both $J_1$ and $J_2$ equal

$$(6.16)\qquad -4h^{-2}\sin^2\frac{j\pi h}{2},\qquad j = 1(1)N.$$

Let $y^{[\ell]}(t)$ denote the $\ell$-th component of the solution vector $y(t)$. Then we have

$$(6.17)\qquad y^{[c+(r-1)N]}(t) = \sum_{i,j=1}^{N} a_{ij}\exp\{-4h^{-2}[\sin^2\frac{i\pi h}{2}+\sin^2\frac{j\pi h}{2}]t\} * \sin(i\pi ch)\sin(j\pi rh),\qquad c,r = 1(1)N,$$

$$a_{ij} = 4h^2\sum_{c,r=1}^{N} y_0^{[c+(r-1)N]}\sin(i\pi ch)\sin(j\pi rh),$$

where $y_0^{[c+(r-1)N]} = s(ch,rh)$. The coefficients $a_{ij}$ are in fact the components of the vector $X^{-1}y_0$ of expression (6.14), and approximate the exact Fourier coefficients

$$\bar{a}_{ij} = 4\int_0^1\int_0^1 s(x_1,x_2)\sin(i\pi x_1)\sin(j\pi x_2)dx_1dx_2.$$

Hence, if $|\bar{a}_{ij}|$ decreases slowly with i and j, the convergence is also expected to be slow, even in the initial phase.

To get some insight in the convergence behaviour of the IDEC, we did some experiments for the special initial function [14, p. 127]

$$(6.18)\qquad s(x_1,x_2) = \frac{\sin\pi x_1\ \sin\pi x_2}{(1-2\alpha\ \cos(\pi x_1)+\alpha^2)(1-2\alpha\ \cos(\pi x_2)+\alpha^2)},\qquad \alpha^2 < 1,$$

where $\bar{a}_{ij} = \alpha^{i+j-2}$. If $\alpha \to 1$, we expect that the convergence becomes slower. We applied (6.2) for $m = 1(1)4$, each time on 4 subintervals $[0,H_1] = [0,m\tau]$ for $\alpha = 0, 0.1$ and $0.5$. In table 6.1 we listed the number of iterations necessary to satisfy

$$(6.19) \qquad \|\tilde{\eta}^{j+1} - \tilde{\eta}^{j}\|_{\infty} \leq \theta,$$

$\theta$ being given in the table. In all experiments $N = 10$, i.e. $h = 1/11$. As a consequence, the values $z_1 = -4\tau h^{-2}\sin^2(j\pi h/2)$, $j = 1(1)N$ (see fig. 6.1) are approximately lying between $-\pi^2\tau$ and $-484\tau$. The number $-\pi^2$ approximates the smallest eigenvalue (6.16). If only the first harmonic is present in the initial function, i.e. $\alpha = 0$ in (6.18), the speed of convergence is completely determined by the product of $\tau$ and this smallest eigenvalue. It should be observed that in this experiment the initial error $\tilde{\varepsilon}_0$ varies with $\tau$ and m.

| | | $\tau = 1/10$ | | | $\tau = 1/20$ | | | $\tau = 1/40$ | | | $\tau = 1/80$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| m | $\alpha \backslash \theta$ | $10^{-2}$ | $10^{-4}$ | $10^{-6}$ | $10^{-2}$ | $10^{-4}$ | $10^{-6}$ | $10^{-2}$ | $10^{-4}$ | $10^{-6}$ | $10^{-2}$ | $10^{-4}$ | $10^{-6}$ |
| 1 | 0.0 | 3 | 6 | 9 | 2 | 4 | 6 | 2 | 3 | 5 | 1 | 3 | 4 |
| | 0.1 | 3 | 7 | 14 | 2 | 5 | 10 | 2 | 4 | 7 | 2 | 3 | 5 |
| | 0.5 | 4 | 16 | 49 | 3 | 12 | 32 | 3 | 9 | 21 | 2 | 6 | 13 |
| 2 | 0.0 | 2 | 7 | 11 | 2 | 6 | 8 | 2 | 4 | 6 | 2 | 2 | 5 |
| | 0.1 | 2 | 7 | 14 | 2 | 6 | 10 | 2 | 5 | 8 | 2 | 3 | 6 |
| | 0.5 | 3 | 13 | 44 | 3 | 10 | 29 | 3 | 8 | 20 | 2 | 6 | 14 |
| 3 | 0.0 | 3 | 7 | 12 | 3 | 6 | 10 | 2 | 5 | 7 | 2 | 4 | 6 |
| | 0.1 | 3 | 7 | 13 | 3 | 6 | 10 | 3 | 5 | 8 | 2 | 5 | 8 |
| | 0.5 | 3 | 11 | 40 | 3 | 9 | 27 | 3 | 8 | 19 | 3 | 7 | 13 |
| 4 | 0.0 | 3 | 8 | 13 | 3 | 7 | 11 | 2 | 6 | 9 | 2 | 5 | 7 |
| | 0.1 | 3 | 8 | 13 | 3 | 7 | 12 | 2 | 6 | 11 | 2 | 5 | 9 |
| | 0.5 | 3 | 10 | 37 | 4 | 9 | 25 | 4 | 9 | 17 | 3 | 8 | 13 |

Table 6.1. Results of convergence experiment.

The results of the convergence experiment show that, despite the damping as pointed out in (6.14), the IDEC is rather sensitive with respect to the higher harmonics. The experiment also shows that mostly the speed of convergence decreases with the number of iterations (provided $\alpha \neq 0$). Because of these 2 unwanted phenomena, it seems of less use to apply the IDEC while iterating until convergence. In the next section we therefore discuss some more experiments being performed with a *fixed* number of iterations, viz. m-1. In this approach we fully rely on the *order* of the formulas.

## 7. NUMERICAL EXPERIMENTS

### 7.1. The examples used

We shall report numerical results for 3 examples of initial boundary value problems for 2-dimensional equations of type (1.1), i.e.

$$u_t = F_1(t,x_1,x_2,u_1,u_{x_1},u_{x_1x_1}) + F_2(t,x_1,x_2,u,u_{x_2},u_{x_2x_2}). \tag{7.1}$$

The equations have been chosen from 2 test families of parabolic problems suggested in [11]. We first list the two families (in reduced form), and then the 3 examples. In all examples (7.1) is assumed to be defined on $(0,1] \times \{(x_1,x_2) \mid 0 < x_1,x_2 < 1\}$. For simplicity, we confined ourselves to Dirichlet boundary conditions.

First family

$$\begin{aligned} F_1(t,x_1,x_2,u,u_{x_1},u_{x_1x_1}) &= u^{2\nu}[u_{x_1x_1} + a(t,x_1,x_2)] + g(t,x_1,x_2), \\ F_2(t,x_1,x_2,u,u_{x_2},u_{x_2x_2}) &= u^{2\nu}u_{x_2x_2}, \end{aligned} \tag{7.2}$$

where

$$a(t,x_1,x_2) = -2t^2(x_1 + \sin(2\pi t)),$$

$$g(t,x_1,x_2) = t[(x_1^2 + x_2)(2\sin(2\pi t) + 2\pi t\cos(2\pi t)) + 2x_1x_2^2],$$

and solution

$$u(t,x_1,x_2) = 1 + t^2[(x_1^2 + x_2)\sin(2\pi t) + x_1x_2^2].$$

Second family

$$
(7.3) \qquad \begin{aligned}
F_1(t,x_1,x_2,u,u_{x_1},u_{x_1x_1}) &= \sqrt{u}\, u_{x_1x_1} - \frac{u}{2(1+t)} - 2u\sqrt{u}\,,\\
F_2(t,x_1,x_2,u,u_{x_2},u_{x_2x_2}) &= \sqrt{u}\, u_{x_2x_2}\,,
\end{aligned}
$$

with solution

$$u(t,x_1,x_2) = \frac{e^{-x_1-x_2}}{\sqrt{1+t}}\ .$$

Example 1. The first example, being linear, is defined by equations (7.2) with $\nu = 0$ (the initial and boundary conditions were obtained from the exact solution).

Example 2. The second example is defined by equations (7.2) with $\nu = 1$, hence it is non-linear.

Example 3. The third example, also being non-linear, is defined by equations (7.3).

The problems were semi-discretized on a uniform grid, using second order symmetrical finite differences, with grid size $h = 1/(N+1)$, $N = 19$. The boundary expressions, appearing in the ordinary differential equations for the internal grid points nearest to the boundary, were evaluated at $t = t_{\nu+1}$ (see formula (2.10)). Note that the space errors for examples 1 - 2 are equal to zero.

## 7.2. The algorithms used

The basic formula for our IDEC is (2.10) with $k = 2$. The tridiagonal matrices $\bar{J}_1$ and $\bar{J}_2$, approximating the partial derivatives $\partial f_1/\partial y$ and $\partial f_2/\partial y$ at the point $(t_\nu, y_\nu)$, were computed using first order finite differences. In case of non-linear problems, this computation was performed only at the beginning of each IDEC step, hence at the step points $t_{\ell m}$, $\ell = 0,1,\ldots$ .

As discussed in the previous section, it is of no use to perform a large number of IDEC iterations. Consequently, we applied the technique with a fixed number of m-1 iterations, so that the order of consistency of the resulting algorithm is equal to m. Observe that for m = 1 we thus applied the LOD formula (2.10) with k = 2.

To be able to compare the results of the various algorithms we need a measure, say $ce_m$, for the *computational effort per integration step* of length $\tau$. It is convenient to express $ce_m$ in the effort of the LOD method. Therefore, we set $ce_1 = 1$. We now assume that the effort of 1 IDEC iteration, using m points, is equal to $2 * m * ce_1 = 2m$. This is justified by the observation that the defect calculations require the evaluation of a derivative and a weighted sum (3.9). Consequently, we have $ce_m = 2m-1$. The computational labour involved in the calculation of the matrices $\bar{J}_1$ and $\bar{J}_2$ has been left out of consideration. For non-linear problems this favours the schemes where m is small, especially the LOD formula, because it integrates with matrices being updated every integration step. This will influence the accuracy and stability of the formula.

### 7.3. The results

The 3 examples were integrated with all algorithms, i.e. with m = 1(1)4, for $\tau$ = 1/12, 1/24, 1/48, 1/96. In the tables of results one finds, for t = 0.5 and t = 1,

$$ae = -{}^{10}\log(\text{maximum over all grid points of the absolute errors at the point } t),$$

$$sce_m = ce_m * t/\tau .$$

In the tables the symbol * means instability.

Table 7.1 (ae,$sce_m$) - values for example 1.

| | t = 0.5 | | | | t = 1 | | | |
|---|---|---|---|---|---|---|---|---|
| $\tau$ \ m | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1/12 | 1.73,6 | 2.13,18 | 2.43,30 | 2.73,42 | 0.96,12 | 1.36,36 | 1.81,60 | 2.07,84 |
| 1/24 | 1.94,12 | 2.51,36 | 2.89,60 | 3.12,84 | 1.16,24 | 1.76,72 | 2.23,120 | 2.46,168 |
| 1/48 | 2.18,24 | 2.87,72 | 3.27,120 | 3.49,168 | 1.42,48 | 2.15,144 | 2.61,240 | 2.84,336 |
| 1/96 | 2.46,48 | 3.21,144 | 3.67,240 | 3.92,336 | 1.69,96 | 2.51,288 | 3.02,480 | 3.28,672 |

Table 7.2 $(ae,sce_m)$ - values for example 2.

| | t = 0.5 | | | | t = 1 | | | |
|---|---|---|---|---|---|---|---|---|
| $\tau \setminus m$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1/12 | 1.67,6 | 0.83,18 | * ,30 | * ,42 | 0.36,12 | * ,36 | * ,60 | * ,84 |
| 1/24 | 1.83,12 | 1.97,36 | * ,60 | * ,84 | 0.99,24 | * ,72 | * ,120 | * ,168 |
| 1/48 | 2.06,24 | 2.57,72 | 2.53,120 | 1.60,168 | 1.25,48 | * ,144 | * ,240 | * ,336 |
| 1/96 | 2.34,48 | 2.95,144 | 3.37,240 | 3.69,336 | 1.49,96 | 1.72,288 | * ,480 | * ,672 |

Table 7.3 $(ae,sce_m)$ - values for example 3.

| | t = 0.5 | | | | t = 1 | | | |
|---|---|---|---|---|---|---|---|---|
| $\tau \setminus m$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1/12 | 1.67,6 | 2.07,18 | 2.39,30 | 1.50,42 | 1.76,12 | 2.17,36 | 2.50,60 | * ,84 |
| 1/24 | 1.89,12 | 2.39,36 | 2.74,60 | 2.97,84 | 1.98,24 | 2.48,72 | 2.84,120 | 3.06,168 |
| 1/48 | 2.13,24 | 2.71,72 | 3.10,120 | 3.34,168 | 2.22,48 | 2.81,144 | 3.21,240 | 3.45,336 |
| 1/96 | 2.39,48 | 3.05,144 | 3.52,240 | 3.78,336 | 2.48,96 | 3.15,288 | 3.63,480 | 3.89,672 |

The results of the computations, given in tables 7.1 - 7.3, clearly indicate that the following conclusions are justified:

$1^e$. For non-linear problems the IDEC formulas are less stable than the LOD formula. This conclusion is justified by the results for example 2, and the result for example 3 obtained for $\tau = 1/12$ and $m = 4$. We emphasize, however, that the LOD formula updates the Jacobian matrices every integration step. In practice this is very costly and will seldom be done. Nevertheless, if the updating is not performed every step, it is still expected that the LOD formula is more stable.

$2^e$. In case of stable computations the results become better with increasing m. This can be immediately verified by putting the $(ae,sce_m)$-values of examples 1, 3 in an accuracy-efficiency diagram. Between succeeding values of m the improvement is not large. If we compare the results obtained for $m = 4$ with the results obtained by the LOD formula, however, the improvement is significant. Let us, for example, consider the results given in table 7.1 for t = 1. Now, if we assume that further halving the stepsize in the LOD formula also halves the error - for $\tau$ small enough this is inevitable - we can write down the following $(ae,sce_1)$-values:

1.99,192
2.29,384
2.59,768
2.89,1536
3.19,3072

It is immediately seen that the corresponding m = 4 - results are significantly better.

$3^e$. The order of consistency of the IDEC formulas cannot be recovered from the results (note that in example 1 the space errors are equal to zero; further, from the experiments described below it can be seen that in the errors of table 7.3 the time integration errors clearly dominate). To indicate that this phenomenon is probably not due to the defect correction, but *inherent in the collocation schemes*, we performed two further experiments. We integrated examples 1, 3 with the m = 4-formula, using $\tau$ = 1/24, 1/48 and 1/96, but now performing 10 IDEC iterations in order to obtain a numerical approximation more closely to the collocation approximation. The ae-values obtained at t = 1, are given below:

| $\tau$ | 1/24 | 1/48 | 1/96 |
|---|---|---|---|
| example 1 | 3.18 | 3.67 | 4.33 |
| example 3 | 3.95 | 4.39 | 5.04 |

All errors are significantly smaller than the corresponding errors of the preceding experiments. Again, however, the order p = 4 can not be recovered (this will be the case after an unacceptable decrease of $\tau$). As in all computations the inequality $\|\eta_\nu^{10} - \eta_\nu^{9}\|_\infty < 10^{-ae}$, $\nu = 1(1)1/\tau$, was satisfied, we believe that the *effective order* of the collocation schemes itself - when applied to semi-discrete parabolic equations with realistic stepsizes - is significantly smaller than the *theoretical* *order*. The reader should observe that this conclusion is in disagreement with the results reported in [3, section 6].

Summarizing our conclusions: when compared with the basic LOD formula the IDEC formulas are more efficient, especially the higher order ones (provided they remain stable when integrating non-linear problems). A disappointment is that the effective order of the formulas is significantly smaller than the theoretical order when considering realistic stepsizes. Because this may imply that the additional computational effort, needed to obtain the higher theoretical orders, is better used when integrating - using relatively small stepsizes - with a simple *second* order splitting method, such as the method of alternating directions or the line hopscotch method [10]. These methods also possess unconditional stability properties and are computationally attractive (per integration step). Some results of the line hopscotch method, applied to examples 1, 3, are given in appendix 3 of [12]. It appears that for example 1 our m = 4-formula is slightly better, whereas for the non-linear example 3 the line hopscotch method is to be preferred. Unfortunately, if we apply defect correction to the line hopscotch method, or the ADI method, the resulting schemes do not possess unconditional stability properties (see the remark in section 6).

## ACKNOWLEDGEMENTS

The author wishes to thank Mrs. M. Louter-Nool for her programming assistence and Dr. C den Heyer for carefully reading the manuscript.

## REFERENCES

[1] ABRAMOWITZ, M. & I.A. STEGUN, *Handbook of mathematical functions*, National Bureau of Standards Applied Mathematics Series 55, U.S. Government Printing Office, Washington, 1964.

[2] DOUGLAS, J. & T. DUPONT, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal. 7, 575-626, 1970.

[3] FRANK, R. & C.W. UEBERHUBER, *Iterated defect correction for the efficient solution of stiff systems of ordinary differential equations*, Report No. 17/76, Institute for Numerical Analysis, Technical University, Vienna, 1976 (in a condensed form published in BIT 17, 146-159, 1977).

[4] FRANK, R. & C.W. UEBERHUBER, *Iterated defect correction for Runge-Kutta methods*, Report No. 14/75, Institute for Numerical Analysis, Technical University, Vienna, 1975.

[5] HULME, B.L., *Discrete and related one-step methods for ordinary differencial equations*, Math. Comp. 26, 881-891, 1972.

[6] LANCASTER, P., *Theory of matrices*, Academic Press, New York and London, 1969.

[7] RICHTMYER, R.D. & K.W. MORTON, *Difference methods for initial value problems*, Interscience, New York, 1967.

[8] ROTTMANN, K., *Mathematische Formelsammlung*, Bibliographisches Institut, Mannheim, 1960.

[9] STETTER, H.J., *The defect correction principle and discretization methods*, Num. Math. 29, 425-443, 1978.

[10] VAN DER HOUWEN, P.J. & J.G. VERWER, *One-step splitting methods formulated for semi-discrete parabolic equations*, Report NW 55/78, Mathematisch Centrum, Amsterdam, (prepublication) 1978.

[11] VAN DER HOUWEN, P.J., B. SOMMEIJER & J.G. VERWER, *Comparing time-integrators for parabolic equations in two space dimensions with a mixed derivative*, in preparation.

[12] VERWER, J.G., *The application of iterated defect correction to the LOD method for parabolic equations*, Report NW 58/78, Mathematisch Centrum, Amsterdam, (prepublication) 1978.

[13] WRIGHT, K., *Some relationships between implicit Runge-Kutta, Collocation and Lanczos* $\tau$*-methods and their stability properties*, BIT 10, 217-227, 1970.

[14] YANENKO, N.N., *The method of fractional steps*, Springer-Verlag, Berlin, 1971.

## appendix 1. SOME KNOWN RESULTS ON THE COLLOCATION METHODS

Polynomial collocation methods for initial value problems for systems of ordinary differential equations form a subclass of the wide class of implicit Runge-Kutta methods [5, 13]. In case of the grid of step points (3.1) this equivalence is not difficult to show. Let $P^*(t)$ denote the collocation polynomial of theorem 4.1. From (3.9) we have

$$(P^*)'(t_\nu) = \tau^{-1} \sum_{\kappa=0}^{m} w_{\nu\kappa} P^*(t_\kappa), \qquad \nu = 1(1)m.$$

Using $P^*(t_\kappa) = \eta_\kappa^*$ and $(P^*)'(t_\nu) = f(t_\nu, P^*(t_\nu))$, then yields

$$\sum_{\kappa=0}^{m} w_{\nu\kappa}\eta_\kappa^* = \tau f(t_\nu, \eta_\nu^*), \qquad \nu = 1(1)m,$$

or, equivalently,

$$\sum_{\kappa=1}^{m} w_{\nu\kappa}\eta_\kappa^* = -w_{\nu 0} y_0 + \tau f(t_\nu, \eta_\nu^*), \qquad \nu = 1(1)m.$$

From the invertability of the weight-matrix

$$W = (w_{\nu\kappa})_{\nu,\kappa=1(1)m},$$

and the identity

$$W^{-1}W_0 = -E, \quad W_0 = [w_{10}, \ldots, w_{m0}]^T, \quad E = [1, \ldots, 1]^T,$$

the Runge-Kutta formula (4.1) is obtained:

$$\eta_\nu^* = y_0 + \tau \sum_{\kappa=1}^{m} \bar{w}_{\nu\kappa} f(t_\kappa, \eta_\kappa^*), \qquad \nu = 1(1)m.$$

When applied to the scalar, stability test-equation

$$s' = \lambda s, \quad t > 0, \quad s(0) = s_0, \quad \lambda \in \mathbb{C},$$

this formula yields the expressions

$$\eta_\nu^* = \phi_\nu(z)s_0, \quad z = m\tau\lambda, \quad \nu = 1(1)m,$$

$\phi_\nu$ being a rational function satisfying $\phi_\nu(z) \sim 1/z$, $z \to -\infty$. Using the technique of [13], it can be shown that

$$\phi_\nu(z) = (\sum_{r=1}^{m} r!\, a_{r,\nu} z^{m-r})/(\sum_{r=0}^{m} r!\, b_r z^{m-r}),$$

of which the coefficients $a_{r,\nu}$ and $b_r$ are defined by

$$\prod_{r=1}^{m} (x - r/m) = \sum_{r=0}^{m} b_r x^r,$$

$$\prod_{r=1}^{m} (x + \nu/m - r/m) = \sum_{r=1}^{m} a_{r,\nu} x^r.$$

The stability function of the implicit Runge-Kutta method is the rational function $\phi_m$.

Example. For m = 2 (4.1) reads

$$\eta_1^* = y_0 + \frac{3}{2}\tau f(t_1,\eta_1^*) - \frac{1}{2}\tau f(t_2,\eta_2^*),$$

$$\eta_2^* = y_0 + 2\tau f(t_1,\eta_1^*).$$

The functions $\phi_\nu$ are given by

$$\phi_1(z) = \frac{1 - \frac{1}{4}z}{1 - \frac{3}{4}z + \frac{1}{4}z^2} ,$$

$$\phi_2(z) = \frac{1 + \frac{1}{4}z}{1 - \frac{3}{4}z + \frac{1}{4}z^2} .$$

## appendix 2. A DESCRIPTION OF A TEST MODEL FOR SEMI-DISCRETE TWO-DIMENSIONAL PARABOLIC EQUATIONS

Let $\Omega$ denote the unit square $\{(x_1,x_2) | \; 0 < x_1,x_2 < 1\}$ with boundary $\delta\Omega$. Consider the initial boundary value problem

$$u_t = u_{x_1x_1} + u_{x_2x_2}, \qquad (t,x_1,x_2) \in (0,T] \times \Omega.$$

(A2.1) $$u = 0 \quad \text{on } \delta\Omega.$$

$$u = s(x_1,x_2) \quad \text{at } t = 0.$$

Assume that the initial function s can be expanded in a 2-dimensional Fourier series. Thus we have

$$u(t,x_1,x_2) = \sum_{i,j=1}^{\infty} \bar{a}_{ij} e^{-\pi^2(i^2+j^2)t} \sin(i\pi x_1)\sin(j\pi x_2),$$

where

$$\bar{a}_{ij} = 4 \int_0^1 \int_0^1 s(x_1,x_2)\sin(i\pi x_1)\sin(j\pi x_2)dx_1dx_2.$$

Now, impose a uniform grid of size $h = 1/(N+1)$, on $\Omega \cup \delta\Omega$. Let $U_{cr}(t)$ denote the approximation for $u(t,x_1,x_2)$ at the grid point $(x_1,x_2) = (ch,rh)$, $c,r = 1(1)N$, which is obtained from replacing $u_{x_1x_1} + u_{x_2x_2}$ by standard symmetrical finite differences. Further, let y denote the $N^2$-dimensional vector function

(A2.2) $$[U_{11},U_{21},\ldots,U_{N1},\ldots,U_{1r},U_{2r},\ldots,U_{Nr},\ldots,U_{1N},U_{2N},\ldots,U_{NN}]^T,$$

and define $y_0 = y(0)$. The semi-discrete version of (A2.1) is then given by

(A2.3) $$y' = Jy, \quad t \in (0,T], \quad y(0) = y_0,$$

where J is the $N^2$-th order matrix

$$J = I \otimes A + A \otimes I,$$

with I being the unit matrix of order N, A the matrix of order N being given by

$$A = h^{-2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} ,$$

and the symbol ⊗ denoting direct product as defined in [6, p. 256].

Now, if we define

$$J_1 = I \otimes A, \quad J_2 = A \otimes I,$$

we have the LOD matrices of (A2.3). These matrices are easily shown to satisfy restrictions (6.5). Firstly, from direct product properties it follows that $J_1$ and $J_2$ commute. Because they are also simple, they have a set of $N^2$ linearly independent right eigenvectors in common [6, p. 265]. Secondly, the symmetry and negative definiteness of $J_1$ and $J_2$ follow immediately from their definition.

Let $X_A$ and $\Lambda_A$ denote the eigensystem and eigenvalue matrix of A, i.e. $A = X_A \Lambda_A X_A^{-1}$. From the relations [6, p. 258]

$$J_1 = (I \otimes X_A)(I \otimes \Lambda_A)(I \otimes X_A)^{-1},$$

$$J_2 = (X_A \otimes I)(\Lambda_A \otimes I)(X_A \otimes I)^{-1},$$

and the commutativity of $I \otimes X_A$ and $X_A \otimes I$, it follows that

$$(X_A \otimes I)^{-1}(I \otimes X_A)^{-1} J_1 (I \otimes X_A)(X_A \otimes I) = I \otimes \Lambda_A,$$

$$(X_A \otimes I)^{-1}(I \otimes X_A)^{-1} J_2 (I \otimes X_A)(X_A \otimes I) = \Lambda_A \otimes I.$$

Hence, the common eigensystem X of $J_i$ and J is

$$X = (I \otimes X_A)(X_A \otimes I).$$

The eigenvalue matrix $\Lambda$ of J is

$$\Lambda = I \otimes \Lambda_A + \Lambda_A \otimes I.$$

From the well-known expressions for the i-th eigenvalue of A,

$$-4h^{-2}\sin^2 \frac{i\pi h}{2} ,$$

and its corresponding eigenvector

$$[\sin(i\pi h),\sin(2i\pi h),\ldots,\sin(Ni\pi h)]^T,$$

it is not difficult to see that the eigenvalues of J are expressions of the type

$$\text{(A2.4)} \qquad -4h^{-2}[\sin^2 \frac{i\pi h}{2} + \sin^2 \frac{j\pi h}{2}],$$

while the components of the corresponding eigenvector are given by

$$\sin(i\pi ch)\sin(j\pi rh), \qquad c,r = 1(1)N.$$

These expressions are immediately recognized in the exact solution of (A2.3) (componentwise, as defined by (A2.2)):

$$\text{(A2.5)} \qquad \begin{aligned} U_{cr}(t) &= \sum_{i,j=1}^{N} a_{ij} \exp\{-4h^{-2}[\sin^2 \frac{i\pi h}{2} + \sin^2 \frac{j\pi h}{2}]t\} * \\ &\qquad \sin(i\pi ch)\sin(j\pi rh), \\ a_{ij} &= 4h^2 \sum_{c,r=1}^{N} s(kh,rh)\sin(i\pi ch)\sin(j\pi rh). \end{aligned}$$

The coefficients $a_{ij}$ are in fact the components of the vector $X^{-1}y_0$ and approximate the exact Fourier coefficients $\bar{a}_{ij}$. The eigenvalues (A2.4), occurring in (A2.5), are approximations to $-\pi^2(i^2 + j^2)$.

## appendix 3. SOME RESULTS OF THE LINE HOPSCOTCH FORMULA

For reasons of comparison we also performed some experiments with an implementation of Gourlay's line hopscotch formula, which is described in [11]. With respect to computational effort per integration step this implementation is comparable with our implementation of the LOD formula (2.10). Using the measure of section 7.2, we can set $ce_{LHS} = 1$ for linear problems and $ce_{LHS} = 2$ for non-linear problems. With this measure we only count the number of functions evaluations. We observe that for linear problems the number of tridiagonal inversions of the LHS implementation is half the number of inversions used by the LOD implementation. For non-linear problems they are equal. Hence, our measure is slightly in favour of the LOD implementation.

We integrated example 1 with the LHS implementation using the stepsize $\tau = 1/84, 1/168, 1/336, 1/672$, and example 3 with $\tau$ as twice as large. By this choice - using our measure - the total computational effort of the LHS implementation equals the total effort of the $m = 4$-formula, as used in section 7. Because we compare the LHS results with the results of the $m = 4$-formula, the tridiagonal Jacobian matrices, required by the LHS implementation, were computed every 4 integration steps. The $(ae, sce_{LHS})$-values at $t = 0.5$ and $t = 1$ are given below. For example 1 the IDEC results are slightly better, whereas for example 3 the line hopscotch method is to be preferred (see the $m = 4$-results in tables 7.1 and 7.3).

example 1

| $\tau$ \ $t$ | 0.5 | 1 |
|---|---|---|
| 1/84 | 1.99,42 | 1.55,84 |
| 1/168 | 2.60,84 | 2.16,168 |
| 1/336 | 3.20,168 | 2.77,336 |
| 1/672 | 3.81,336 | 3.37,672 |

example 3

| $\tau$ \ $t$ | 0.5 | 1 |
|---|---|---|
| 1/42 | 3.36,42 | 3.68,84 |
| 1/84 | 3.98,84 | 4.33,168 |
| 1/168 | 4.70,168 | 5.22,336 |
| 1/336 | 5.72,336 | 5.36,672 |