Martin–Luther–Universität
Halle–Wittenberg
Fachbereich Mathematik und Informatik

LINEAR PARTIAL DIFFERENTIAL ALGEBRAIC EQUATIONS PART II: NUMERICAL SOLUTION

W. Lucht, K. Strehmel and C. Eichler-Liebenow

Report No. 18 (1997)

Reports of the Institute of
Numerical Mathematics

Editors:
Professors and Private Docents of the Department of Mathematics and Computer Science, Martin-Luther-University Halle-Wittenberg.

LINEAR PARTIAL DIFFERENTIAL ALGEBRAIC EQUATIONS PART II: NUMERICAL SOLUTION

W. Lucht, K. Strehmel and C. Eichler-Liebenow
Report No. 18 (1997)

Prof. Dr. Wenfried Lucht
Prof. Dr. Karl Strehmel
Dipl.-Math. Claudia Eichler-Liebenow

Fachbereich Mathematik und Informatik
Institut für Numerische Mathematik
Universität Halle
Theodor-Lieser-Str. 5
06120 Halle, Germany
Tel.:(0345) 55 24649, 55 24650, 55 24667

email: lucht@mail.mathematik.uni-halle.de
strehmel@mathematik.uni-halle.de
liebenow@mathematik.uni-halle.de

This paper has been submitted for publication.

LINEAR PARTIAL DIFFERENTIAL ALGEBRAIC EQUATIONS PART II: NUMERICAL SOLUTION

W. LUCHT, K. STREHMEL, C. EICHLER-LIEBENOW
Martin-Luther-Universität Halle–Wittenberg
Fachbereich Mathematik und Informatik
Institut für Numerische Mathematik
Postfach, D–06099 Halle, Germany
email: lucht@mail.mathematik.uni-halle.de

ABSTRACT. We consider the numerical solution of linear partial differential algebraic equations (PDAEs) of the general type $Au_t(t,x)+Bu_{xx}(t,x)+Cu(t,x)=f(t,x)$ (where A, B, C are constant $(n\times n)$–matrices) by means of the method of lines (MOL), the backward in time, centered in space (BTCS) and the Crank-Nicolson scheme. Using an algebraic quality (the proper quality) to characterize certain matrix pencils resulting from the PDAE, the convergence in norm of the numerical solutions (determined by the difference schemes) towards the exact solution is described in terms of two indexes of the PDAE. In two Theorems it is shown that there is a strong dependence of the order of convergence on these indexes. We present examples for the calculation of the order of convergence and give results of numerical calculations for several aspects encountered in the numerical solution of PDAEs.

AMS subject classification: 65M06, 65M15, 65M20

Keywords: Partial differential algebraic equations, coupled systems, indexes, convergence of difference schemes, method of lines

1. INTRODUCTION

Along the ideas of the well-known method of lines (MOL) approach we consider two discretization methods for the numerical solution of linear partial differential algebraic equations with constant coefficients of the form

$$A\,u_t(t,x)+B\,u_{xx}(t,x)+C\,u(t,x)=f(t,x) \tag{1.1}$$

with $t\in(0,t_e)$ and $x\in(-l,l)$, $A,B,C\in\mathbb{R}^{n\times n}$, $u,f:[0,t_e]\times[-l,l]\to\mathbb{R}^n$. The interest is in cases where at least one of the matrices A and B is singular. The two special cases $A=0$ or $B=0$ lead to ordinary differential equations (ODEs) or differential algebraic equations (DAEs) which are not considered here. Therefore, in this paper we assume that none of the matrices A and B is the zero matrix.

Eq. (1.1) is to be supplemented for all $t\in[0,t_e]$, $t_e>0$, by boundary conditions (BCs) for the components u_j of u for all $j\in\mathfrak{M}_{BC}$ (this set is defined in the following section), and for simplicity we assume Dirichlet BCs

$$\mathrm{R_B}\;u_j(t,x):=u_j(t,\pm l)\;=\;0,\quad j\in\mathfrak{M}_{BC}. \tag{1.2}$$

Furthermore, initial conditions (ICs)

$$u_i(0,x) = g_i(x) \quad \text{for} \quad x \in [-l, l], \quad i \in \mathfrak{M}_{IC}, \tag{1.3}$$

are assumed to be given. Any components g_i of g for all $i \in \mathfrak{M}_{IC} \subseteq \{1, 2, .., n\}$ can be chosen. For the components u_k, $k \notin \mathfrak{M}_{BC}$, BCs of only Dirichlet type are in order.

We also require the compatibility conditions

$$\mathrm{R_B}\ g(x) = \mathrm{R_B}\ u(0,x) \tag{1.4}$$

between the ICs and the BCs. In contrast to problems with regular matrices A and B (e.g. parabolic problems) the IC (1.3) and/or the BC (1.2) for problems with singular matrices A and/or B have to fulfill certain supplementary conditions (so-called consistency conditions).

Equations of the type (1.1) with BCs and ICs as given are investigated extensively in [7] which is the basis of this paper. Some facts necessary for a numerical solution of PDAEs are recapitulated in section 2.

In the third section a semidiscretization of the PDAE (1.1) with respect to the space coordinate (with space step size h), i.e. the method of lines (MOL) is considered. If this space discretization is of second order, then (under some assumptions) the order of convergence for $h \to 0$ is shown to be also of second order.

The full discretization of the PDAE is considered in section 4. In subsections 4.1 and 4.2 the BTCS scheme is investigated in detail. Of special interest is the influence of the differential spatial and the uniform differential time index on the convergence of the numerical solution to the exact solution. We derive order relations for the convergence by means of an algebraic characterization of the indexes. The Crank–Nicolson scheme is the subject of subsection 4.3. This investigation is shortened because the ideas are the same as for the BTCS scheme.

Numerical examples using the BTCS scheme are presented in subsection 4.4. First we determine numerically order relations for convergence. Hereby the theoretical results of subsections 4.1 and 4.2 are confirmed. In a second example we study numerically the influence of inconsistent BCs. A third example deals with the problem of an index jump which is considered also in [7]. Although the PDAE does not have an index jump the MOL-DAE may have such a jump. This may be the origin of an instability (as discussed in section 4.4) and of large errors in the numerical solution.

2. Preliminaries

In this section we collect for convenience from [7] some assumptions, definitions and facts necessary for the numerical solution of the initial boundary value problem (IBVP) (1.1)–(1.4). The general assumptions are

(I) : The IBVP (1.1)–(1.4) has one and only one solution.

(II) : Each component of the solution vector u satisfies a growth condition of the form

$$|u_i(t,x)| \le M e^{\alpha t}, \quad i = 1(1)n,$$

(M and α are independent of x).

(III) : The matrix pencil $(B, \xi A + C,\ \mathrm{Re}(\xi) > \alpha,$ is regular.

(IV) : The matrix pencil $(A, \mu_k B + C)$ is regular for all k. μ_k is an eigenvalue of the operator $\frac{\partial^2}{\partial x^2}$ with prescribed BCs (1.2).

(V) : The right hand side vector $f(t,x)$ and the initial vector $g(x)$ are sufficiently smooth.

(VI) : There is an uniform differential time index $\nu_{d,t}$ of the PDAE (1.1).
(VII) : The index sets $\mathfrak{M}_{BC}$ and $\mathfrak{M}_{IC}$ are known.

As in the case of DAEs which can be characterized conveniently by indexes [5],[10] it is useful to introduce indexes also for PDAEs. The basis are the first four assumptions which can be applied to define two indexes of the PDAE (1.1), namely the differential spatial space index $\nu_{d,x}$ and the uniform differential time index $\nu_{d,t}$. Furthermore, the two index sets entering assumption (VII) can be characterized by means of the IBVP considered.

The first index follows from a Laplace transformation, assumption (III) and a Kronecker transformation [4] of the matrices B and $\xi A + C$. This transformation reads with square regular transformation matrices $P_{L,\xi}$ and $Q_{L,\xi}$

$$P_{L,\xi} B Q_{L,\xi} = \begin{pmatrix} I_{m_1} & 0 \\ 0 & N_{L,\xi} \end{pmatrix}, \qquad P_{L,\xi}(\xi A + C) Q_{L,\xi} = \begin{pmatrix} R_{L,\xi} & 0 \\ 0 & I_{m_2} \end{pmatrix}.$$

$R_{L,\xi} \in \mathbb{R}^{m_1 \times m_1}$ is a square matrix and $N_{L,\xi} \in \mathbb{R}^{m_2 \times m_2}$ is a nilpotent Jordan chain matrix where $m_1 + m_2 = n$. $I_k \in \mathbb{R}^{k \times k}$ with $k \in \{m_1, m_2\}$ is an unit matrix. The Riesz index (or nilpotency) of $N_{L,\xi}$ is denoted by $\nu_{L,\xi}$ (i.e. $N_{L,\xi}^{\nu_{L,\xi}} = 0$, $N_{L,\xi}^{\nu_{L,\xi}-1} \neq 0$). Then the differential spatial index is defined by

Definition 2.1. Let $\alpha^* \in R^+$ be a number such that

(1) $\alpha^* > \alpha$ and
(2) the matrix $N_{L,\xi}$ with nilpotency $\nu_{L,\xi} \geq 1$ does not change the structure for all $\xi \in \mathcal{C}$ with $\mathrm{Re}(\xi) > \alpha^*$.

Then

$$\nu_{d,x} := 2\nu_L - 1 \qquad \text{with} \qquad \nu_L := \nu_{L,\xi}, \quad \mathrm{Re}(\xi) > \alpha^*,$$

is called the differential spatial index of the PDAE (1.1). If $\nu_L = 0$, then the differential spatial index of the PDAE is defined to be zero.

The definition of the uniform differential time index $\nu_{d,t}$ is based on a Fourier transformation of eq. (1.1) and on assumption (IV). This implies that a Kronecker transformation of the matrices A and $\mu_k B + C$ with regular matrices $P_{F,k}$, $Q_{F,k}$ can be performed:

$$P_{F,k} A Q_{F,k} = \begin{pmatrix} I_{n_1} & 0 \\ 0 & N_{F,k} \end{pmatrix}, \qquad P_{F,k}(\mu_k B + C) Q_{F,k} = \begin{pmatrix} R_{F,k} & 0 \\ 0 & I_{n_2} \end{pmatrix}$$

where $R_{F,k} \in \mathbb{R}^{n_1 \times n_1}$. $N_{F,k} \in \mathbb{R}^{n_2 \times n_2}$ is again a nilpotent Jordan chain matrix with Riesz index $\nu_{F,k}$. Of course, $n_1 + n_2 = n$. The uniform differential time index $\nu_{d,t}$ is then defined by

Definition 2.2. The PDAE (1.1) is said to have the uniform differential time index $\nu_{d,t} = \nu_F$ if for all k the pencil $(A, \mu_k B + C)$ is regular and has Riesz index $\nu_{F,k} = \nu_F$ (independent of k) and if the Jordan chain matrix $N_{F,k}$ has the same structure for all k.

Remark 2.1. If the matrix $N_{F,k}$ does not have the same structure for all k, then it may be that the initial conditions which are needed for $\hat{u}_k(t)$ (the $k-$th Fourier coefficient of $u(t,x)$) are of different type for different k [7] (e.g. the $j-$th component $\hat{u}_{k,j}(0)$ may be prescribed, but

$\hat{u}_{\bar{k},j}(0)$ can not, $k, \bar{k} \in \mathbb{N}^+$, $k \neq \bar{k}$, and $j \in \{1, ..., n\}$, i.e. $\hat{u}_{\bar{k},j}(0)$ must be determined as a consistent IC). □

Furthermore, the two index sets $\mathfrak{M}_{BC}$ and $\mathfrak{M}_{IC}$ entering assumption (VII) are needed. The introduction of these sets rests on the fact that for the initial boundary value problem (IBVP) (1.1)–(1.4), in general, any BCs and ICs cannot be prescribed for all components of the solution vector u [7]. It can be shown that BCs and ICs can be assigned only for certain components whose indices are collected in the sets

$$\mathfrak{M}_{BC} \subseteq \{1, ..., n\} \qquad \text{and} \qquad \mathfrak{M}_{IC} \subseteq \{1, ..., n\}.$$

In general, the BCs for u_k, $k \notin \mathfrak{M}_{BC}$, and the ICs for u_i, $i \notin \mathfrak{M}_{IC}$, must be determined with the help of the PDAE (1.1), the BCs for u_j, $j \in \mathfrak{M}_{BC}$, and the ICs for u_m, $m \in \mathfrak{M}_{IC}$ (of course, there may be cases which need not all these informations). For the calculation of these ICs and BCs a coupled system of equations can be generated. This construction suggests the following

Definition 2.3. If the BCs for u_k, $k \notin \mathfrak{M}_{BC}$, and the ICs for u_i, $i \notin \mathfrak{M}_{IC}$, have been determined in this way, then these BCs and ICs are called consistent.

Note that if consistent BCs and ICs must be determined, then the two index sets $\mathfrak{M}_{BC}$ and $\mathfrak{M}_{IC}$ of the PDAE must be known.

3. Space discretization and convergence

In this section we consider the space discretization of the IBVP (1.1) – (1.4) by means of finite differences, and we investigate the convergence behavior in space.

Let us space discretize the PDAE (1.1) on an equidistant grid

$$\Omega_h = \left\{ x_k : x_k = -l + kh,\ k = 1, \ldots, N\ , h = \frac{2l}{N+1} \right\}$$

for a given integer $N \in \mathbb{N}^+$. Replacing $u_{xx}(t, x_k)$ by

$$u_{xx}(t, x_k) \approx \frac{1}{h^2} \left(u_{k+1}(t) - 2u_k(t) + u_{k-1}(t) \right), \quad k = 1, \ldots, N$$

we obtain the semi-discretized equation $(u_k(t) \approx u(t, x_k))$

$$A\, u_k'(t) + \frac{1}{h^2} B\ \left(u_{k+1}(t) - 2u_k(t) + u_{k-1}(t) \right) + C\, u_k(t) = f_k(t) \tag{3.1}$$

or in matrix representation (using the Kronecker product)

$$\left(I_N \otimes A \right) U'(t) + \left(\frac{1}{h^2} P \otimes B + I_N \otimes C \right) U(t) = F(t) - r(t). \tag{3.2}$$

where the component $u_k(t)$ of the vector $U(t) \in \mathbb{R}^{nN}$ is the continuous time approximation to $u(t, x_k)$ and the nN dimensional vectors $r(t)$, $F(t)$ are given by

$$r(t) = \left(\frac{1}{h^2} I_N \otimes B \right) \left(u^\top(t, -l), 0 \ldots, 0, u^\top(t, l) \right)^\top,$$
$$F(t) = \left(f_1^\top(t), \ldots, f_N^\top(t) \right)^\top, \quad f_k(t) = f(t, x_k).$$

I_N is the $(N \times N)$-unit matrix, and the matrix P is defined by

$$P = \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & & \ddots & \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{N \times N}. \tag{3.3}$$

Eq. (3.2) is (according to eq. (1.3)) completed by the consistent initial vector

$$U(0) = \left(\tilde{g}^\top(x_1), ..., \tilde{g}^\top(x_N)\right)^\top \in \mathbb{R}^{nN} \tag{3.4}$$

where the difference $\tilde{g} - g$ goes (component wise) to zero for $h \to 0$. If A is singular, then eq. (3.2) is for fixed h a DAE which may be handled by standard methods ([6]). Its differential index is identical (at least for small h) with the uniform differential time index $\nu_{d,t}$. This is easily proved by the expansion

$$U(t) = \sum_{k=1}^{N} \Phi_k \otimes w_k(t), \quad \Phi_k \in \mathbb{R}^N, \ w_k \in \mathbb{R}^n, \tag{3.5}$$

where Φ_k is an eigenvector of the matrix $\frac{1}{h^2}P$ ([11]), i.e.

$$\frac{1}{h^2} P \, \Phi_k = \lambda_k \Phi_k, \qquad k = 1, .., N. \tag{3.6}$$

The eigenvalues of $\frac{1}{h^2}P$ are given by $\lambda_k = -\frac{4}{h^2} \sin^2\left(\frac{k\pi}{2(N+1)}\right)$ with $h = \frac{2l}{N+1}$. These eigenvalues converge for $k \in \{1, ..., N\}$ in the limit $h \to 0$ towards the eigenvalues $\mu_k = -\left(\frac{k\pi}{2l}\right)^2$ with corresponding eigenfunctions $\cos\left(\frac{k\pi}{2l}x\right)$ of the operator $\frac{\partial^2}{\partial x^2}$ with homogeneous Dirichlet BCs on the interval $(-l, l)$. Furthermore the set $\{\Phi_k, \ k = 1, ..., N\}$ is orthogonal.

Let $U_h(t)$ be the restriction of $u(t, x)$ to the spatial grid Ω_h. The vector

$$\eta_h(t) = U_h(t) - U(t)$$

is referred to as the global space error in the semidiscretization.

Definition 3.1. If on an interval $I^* := [0\,,\, t^*),\ t^* > 0\,,$

$$\max_{t \in I^*} \|\eta_h(t)\| = o(1) \quad \text{as } h \to 0,$$

then the problem (3.2), (3.4) is called convergent on I^*. Hereby $\|\cdot\|$ is a suitable norm.

Convergence of order p^* is defined in an obvious way, i.e replacing $o(1)$ by $\mathcal{O}(h^{p^*})$. Of further interest is the space truncation error of (3.2) defined by

$$\alpha_h(t) = (I_N \otimes A)U_h'(t) + \left(\frac{1}{h^2} P \otimes B + I_N \otimes C\right) U_h(t) - F(t) + r(t). \tag{3.7}$$

Remark 3.1. $\alpha_h(t)$ represents the defect of the grid function $U_h(t)$ with respect to the DAE (3.2). Loosely speaking, it shows how good $U_h(t)$ satisfies the DAE (3.2). □

Definition 3.2. The semidiscretization is said to be consistent if

$$\max_{t\in I^*}\|\alpha_h(t)\| = o(1) \quad \text{as} \quad h\to 0,$$

where consistency of order p^* is defined again in an obvious way.

With a simple Taylor expansion (recall that $U_h(t)$ is smooth by assumption) we obtain

$$\left(\frac{1}{h^2}P\otimes B\right)U_h(t) - r(t) = (I_N\otimes B)U_{xx,h}(t) + h^2(I_N\otimes B)\gamma_h(t),$$

$$\gamma_h(t) = \frac{1}{12}\frac{\partial^4}{\partial x^4}(u^\top(t,\zeta_1),\dots,u^\top(t,\zeta_N))^\top\ , \quad \zeta_i\in(x_{i-1},x_i), i=1,..,N,$$

(ζ_i may be different for the each component of $u(t,x)$). Therefore we have

$$\alpha_h(t) = h^2(I_N\otimes B)\gamma_h(t). \tag{3.8}$$

Under the assumption that $\max_{-l\le x\le l}\left\|\frac{\partial^4}{\partial x^4}u(t,x)\right\|$ is bounded for $t\in I^*$ we get the estimate

$$\|\alpha_h(t)\| = \mathcal{O}(h^2)\ , \quad t\in I^*. \tag{3.9}$$

That means the discretization in space is consistent of second order. The space truncation error will enter the analysis in the following section.

From (3.2) and (3.7) it follows that the space discretization error $\eta_h(t)$ satisfies the DAE

$$(I_N\otimes A)\eta_h'(t) + \left(\frac{1}{h^2}P\otimes B + I_N\otimes C\right)\eta_h(t) = \alpha_h(t) \tag{3.10}$$

$$\eta_h(0) = U_h(0) - U(0). \tag{3.11}$$

This equation shows that $\eta_h(t)$ depends essentially on the truncation error $\alpha_h(t)$.

Now we derive for $t\in I^*$ a bound for $\eta_h(t)$ in terms of h from which we obtain convergence of $U(t)$ towards $U_h(t)$ provided the discretization in space is consistent and the truncation error $\alpha_h(t)$ is sufficiently smooth with respect to h and τ.

We introduce the following notation. Let

$$\tilde A = I_N\otimes A, \qquad \tilde B = \frac{1}{h^2}P\otimes B + I_N\otimes C,$$
$$\hat A = (c\tilde A+\tilde B)^{-1}\tilde A, \quad \hat B = (c\tilde A+\tilde B)^{-1}\tilde B, \qquad \hat\alpha_h(t) = (c\tilde A+\tilde B)^{-1}\alpha_h(t),$$

and c is such that $(c\tilde A+\tilde B)$ is invertible. Furthermore let $\hat A^D$ and $\hat B^D$ be the Drazin inverse of $\hat A$ and $\hat B$, respectively. Then the unique solution of (3.10) is given by

$$\begin{aligned}\eta_h(t) &= e^{-\hat A^D\hat B\,t}\hat A\hat A^D(U_h(0)-U(0)) + \hat A^D\int_0^t e^{-\hat A^D\hat B(t-s)}\,\hat\alpha_h(s)\,ds\\ &\quad + (I-\hat A\hat A^D)\sum_{l=0}^{\nu_{d,t}-1}(-1)^l(\hat A\hat B^D)^l\hat B^D\hat\alpha_h^{(l)}(t)\end{aligned} \tag{3.12}$$

where $U_h(0)-U(0)$ must satisfy

$$U_h(0)-U(0) = \hat A\hat A^D(U_h(0)-U(0)) + (I-\hat A\hat A^D)\sum_{l=0}^{\nu_{d,t}-1}(-1)^l(\hat A\hat B)^l\hat B^D\hat\alpha_h^l(0),$$

i.e. the initial values are consistent (see [3]). If A is regular, then it is $\hat{A}^D = \hat{A}^{-1}$ and the foregoing equation reduces to an identity.

Now let $\mu_h = \mu[-\hat{A}^D\hat{B}]$ be the logarithmic matrix norm of the matrix $(-\hat{A}^D\hat{B})$ corresponding to the chosen vector norm $\|\cdot\|$. Then we get from (3.12) the estimate

$$\begin{aligned}\|\eta_h(t)\| &\leq e^{\mu_h t}\|\hat{A}\hat{A}^D\|\|U_h(0)-U(0)\| + \|\hat{A}^D\|\int_0^t e^{\mu_h(t-s)}\|\hat{\alpha}_h(s)\|\,ds \\ &\quad + \|I-\hat{A}\hat{A}^D\|\|\hat{B}^D\|\sum_{l=0}^{\nu_{d,t}-1}\|(\hat{A}\hat{B}^D)\|^l\|\hat{\alpha}_h^{(l)}(t)\| \\ &\leq e^{\mu_h t}\|\hat{A}\hat{A}^D\|\|U_h(0)-U(0)\| + \frac{e^{\mu_h t}-1}{\mu_h}\|\hat{A}^D\|\max_{0\leq\tau\leq t}\|\hat{\alpha}_h(\tau)\| \\ &\quad + \|I-\hat{A}\hat{A}^D\|\|\hat{B}^D\|\max_{l=0}^{\nu_{d,t}-1}\|\hat{\alpha}_h^{(l)}(t)\|\sum_{l=0}^{\nu_{d,t}-1}\|(\hat{A}\hat{B}^D)\|^l, \quad 0\leq t\leq t_e\end{aligned}$$

Thus we have the following convergence result.

Theorem 3.1. Suppose that the discretization in space is consistent of second order and that the truncation error $\alpha_h(t)$ satisfies (3.9). Suppose also that $\|U_h(0)-U(0)\| = \mathcal{O}(h^2)$ for $h\to 0$ and that $|\mu_h|$ is bounded for $h\to 0$. Then the discretization in space is convergent of second order, i.e.

$$\|\eta_h(t)\| = \mathcal{O}(h^2) \;\text{ for }\; h\to 0, \quad t\in I^*.$$

Remark 3.2. If $(c\tilde{A}+\tilde{B})^{-1}$ exists for some c, then the solution (3.12) is independent of c (see [3]). □

Remark 3.3. If A and B are regular, then we have $\hat{A}^D = \hat{A}^{-1}$, $\hat{B}^D = \hat{B}^{-1}$ and with $U_h(0) - U(0) = 0$ we obtain the estimate

$$\|\eta_h(t)\| \leq \frac{e^{\mu_h t}-1}{\mu_h}\max_{0\leq\tau\leq t}\|\alpha_h(\tau)\|$$

(see [12]). □

4. Time discretization and convergence of the full discretization

For the time integration of the DAE (3.2) we use the implicit Euler and the trapezoidal rule. The result is the BTCS and the Crank-Nicolson scheme for the PDAE, respectively. For these two discretization methods we examine the behavior of the full local and global error associated with the exact solution of the IBVP (1.1) –(1.4). For easy of representation we restrict ourselves to a constant stepsize τ.

4.1. The BTCS scheme. This scheme for the PDAE (1.1) is given by

$$G(\tau,h^2)\,U_{m+1} = \left(\frac{1}{\tau}I_N\otimes A\right)U_m + F(t_{m+1}) - r(t_{m+1}) \tag{4.1}$$

with $U_0 = U(0)$ and

$$G(\tau,h^2) = \frac{1}{\tau}I_N\otimes A + \frac{1}{h^2}P\otimes B + I_N\otimes C. \tag{4.2}$$

Of course, if eq. (4.1) is to be used for the calculation of U_{m+1} then it should have an unique solution (G regular). The following lemma shows that the unique solution of eq. (4.1) follows essentially from the regularity of the $(n \times n)$–matrices

$$G_k(\tau, h^2) = \frac{1}{\tau} A + \lambda_k\, B + C, \qquad k = 1, .., N. \tag{4.3}$$

Lemma 4.1. Suppose

(1) the matrices $G_k(\tau, h^2)$ $(k = 1, .., N)$ are regular for $0 < \tau \leq \tau_0,\ 0 < h \leq h_0$ (τ, h fixed),

(2) the inhomogeneous term $\tilde{F}(t_{m+1}) := F(t_{m+1}) - r(t_{m+1})$ can be represented for each h as a series

$$\tilde{F}(t_{m+1}) = \sum_{k=1}^{N} \Phi_k \otimes \tilde{f}_{k,m+1} \ ,$$

with $\tilde{f}_{k,m+1} \in \mathbb{R}^n$.

Then there is an unique solution U_{m+1} of eq. (4.1).

Proof. The essential idea is to represent $U_{m+1} \in \mathbb{R}^{nN}$ also in terms of the N linear independent eigenvectors Φ_k,

$$U_{m+1} = \sum_{k=1}^{N} \Phi_k \otimes U_{k,m+1}.$$

The N vectors $U_{k,m+1} \in \mathbb{R}^n$ are to be determined. The last expression and the expansion for $\tilde{F}(t_{m+1})$ given in assumption (2) are inserted into eq. (4.1). Using then the orthogonality of the set of eigenvectors $\{\Phi_k\}_1^N$ it is easy to see that the coefficient vector $U_{k,m+1}$ satisfies the equation

$$G_k(\tau, h^2)\, U_{k,m+1} = \frac{1}{\tau} A\, U_{k,m} + \tilde{f}_{k,m+1}, \quad k = 1, .., N.$$

Assumption (1) implies that $U_{k,m+1}$ is unique. ⊠

Thus if the low order matrices $G_k(\tau, h^2),\ k = 1, ..., N,$ are regular, then $G(\tau, h^2)$ is regular. This property of $G(\tau, h^2)$ is important for the introduction of the proper quality (defined below) of the matrix $G^{-1}(\tau, h^2)$. This is seen from the following

Lemma 4.2. Let $G(\tau, h^2)$ be regular and $\nu_{d,t}$ and $\nu_{d,x}$ be the uniform differential time index and the differential space index of the PDAE (1.1), respectively. Then the following relations for the regular matrix $G^{-1}(\tau, h^2)$ hold

(i) for $\nu_{d,t} \geq 0$ and fixed $h > 0$

$$\lim_{\tau \to 0} \tau^{\nu_{d,t}}\, G^{-1}_{i,j}(\tau, h^2) = 0 \ , \qquad i, j = 1, \ldots, nN, \tag{4.4}$$

(ii) for $\nu_{d,t} \geq 1$ and fixed $h > 0$

$$\lim_{\tau \to 0} \tau^{\nu_{d,t}} \left(G^{-1}(\tau, h^2) \frac{1}{\tau} (I_N \otimes A) \right)_{i,j} = 0 \ , \qquad i, j = 1, \ldots, nN, \tag{4.5}$$

(iii) for $\nu_{d,x} \geq 0$ and fixed $\tau > 0$

$$\lim_{h\to 0} h^{\nu_{d,x}} \left(G^{-1}_{i,j}(\tau, h^2)\right)_{i,j} = 0 , \qquad i,j = 1,\dots,nN, \tag{4.6}$$

(iv) for $\nu_{d,x} \geq 1$ and fixed $\tau > 0$

$$\lim_{h\to 0} h^{\nu_{d,x}} \left(G^{-1}(\tau, h^2)\frac{1}{h^2}(I_N \otimes B)\right)_{i,j} = 0 , \qquad i,j = 1,\dots,nN. \tag{4.7}$$

Proof. First we consider the matrix pencil $(I_N \otimes A, \frac{1}{h^2}P \otimes B + I_N \otimes C)$ which can assumed to be regular. Thus we may perform a Kronecker transformation [4] of this pencil with regular matrices P_h, $Q_h \in \mathbb{R}^{nN\times nN}$, i.e.

$$I_N \otimes A = P_h^{-1}\begin{pmatrix} I_1 & 0 \\ 0 & \bar{N}\end{pmatrix} Q_h^{-1}, \quad \frac{1}{h^2}P\otimes B + I_N \otimes C = P_h^{-1}\begin{pmatrix} R_h & 0 \\ 0 & I_2\end{pmatrix} Q_h^{-1},$$

where $\bar{N}$ is a nilpotent matrix provided the matrix A is singular, and $I_1, R_h \in \mathbb{R}^{N_1\times N_1}$, I_2, $\bar{N} \in \mathbb{R}^{N_2\times N_2}$, $N_1 + N_2 = nN$ (I_k is an unit matrix). Thus G^{-1} can be written

$$G^{-1}(\tau, h^2) = Q_h \begin{pmatrix} \tau(I+\tau R_h)^{-1} & 0 \\ 0 & (I+\frac{1}{\tau}\bar{N})^{-1}\end{pmatrix} P_h.$$

The statements (i) and (ii) now follow easily from the transformation of $I_N \otimes A$ and from the identity $(I+\frac{1}{\tau}\bar{N})^{-1} = \sum_{l=0}^{\nu_{d,t}-1}(-\frac{1}{\tau}\bar{N})^l$. In the same manner we show with a Kronecker transformation of the matrix pencil $\left(P\otimes B, I_N \otimes \left(\frac{1}{\tau}A + C\right)\right)$ the statements (iii) and (iv). ⊠

Therefore, alternatively (cf. definitions 2.1 and 2.2) we can characterize for the BTCS scheme the differential space index $\nu_{d,x}$ and the uniform differential time index $\nu_{d,t}$ of the PDAE (1.1) by the following

Corollary 4.1. The differential space index $\nu_{d,x}$ and the uniform differential time index $\nu_{d,t}$ are for the BTCS scheme the smallest integers such that (4.6) and (4.4) hold, respectively.

Definition 4.1. The matrices $\tau^{\nu_{d,t}}G^{-1}(\tau,h^2)$ satisfying (4.4) and $h^{\nu_{d,x}}G^{-1}(\tau,h^2)$ satisfying (4.6) are called *τ-proper* and *h-proper*, respectively.

From (4.5) and (4.7) we see that for $\nu_{d,t}, \nu_{d,x} \geq 1$ the matrices

$$\tau^{\nu_{d,t}}G^{-1}(\tau,h^2)\frac{1}{\tau}(I_N \otimes A) \quad \text{and} \quad h^{\nu_{d,x}}G^{-1}(\tau,h^2)\frac{1}{h^2}(I_N\otimes B)$$

have the same proper quality as the matrices $\tau^{\nu_{d,t}}G^{-1}(\tau,h^2)$ and $h^{\nu_{d,x}}G^{-1}(\tau,h^2)$, respectively.

Remark 4.1. The characterization of the indexes of the PDAE by means of the proper quality is of a purely algebraic nature. In contrast, its definitions given earlier in definitions 2.1 and 2.2 use the differential properties of the algebraic solution components obtained after a Kronecker transformation. □

Remark 4.2. The algebraic x-index for the regular matrix $R(s,z) = (sA + z^2B + C)^{-1}$ (s and z some parameters) defined in ([2]) is equal to the differential space index $\nu_{d,x}$. □

Remark 4.3. The proper quality of the matrix $R(s,z) = (sA + z^2B + C)^{-1}$ for $|s| \to \infty$ or $z^2 \to \infty$ was introduced by Campbell/Marszalek ([2]). □

Now let us introduce the *full truncation error*

$$le_{m+1} = U_h(t_{m+1}) - \hat{U}_{m+1} \tag{4.8}$$

with
$$\hat{U}_{m+1} = G^{-1}(\tau, h^2)\left(\left(\frac{1}{\tau}I_N \otimes A\right)U_h(t_m) + \tilde{F}(t_{m+1})\right).$$

We emphasize that this error is defined with respect to the true PDAE solution $U_h(t)$ and not for the true DAE solution U(t). We introduce the following

Definition 4.2. The BTCS scheme (4.1) is consistent with the PDAE (1.1) in a norm $\|\cdot\|$ if the full truncation error le_{m+1} satisfies

$$\frac{1}{\tau}\|le_{m+1}\| \to 0 \quad \text{for} \quad \tau,\ h \to 0, \quad m = 0, ..., N. \tag{4.9}$$

Definition 4.3. The BTCS scheme is said to be accurate of order (p,q), $p, q \geq 1$, if

$$\frac{1}{\tau}\|le_{m+1}\| = \mathcal{O}(h^p) + \mathcal{O}(\tau^q) \quad \text{for} \quad \tau,\ h \to 0. \tag{4.10}$$

Remark 4.4. It is easy to see that if the BTCS scheme is accurate of order (p,q), then it is consistent. If the relations (4.9) and (4.10) hold under a time space restriction of the form (4.14) we say that the BTCS scheme is conditionally consistent and conditionally accurate of order (p,q), respectively. □

With eq. (3.7) we get

$$\begin{aligned} le_{m+1} =& G^{-1}(\tau, h^2)\Big\{(\frac{1}{\tau}I_N \otimes A)\Big[U_h(t_{m+1}) - U_h(t_m)\Big] \\ & - (I_N \otimes A)U_h'(t_{m+1}) + \alpha_h(t_{m+1})\Big\} \end{aligned}$$

and from this follows with a Taylor expansion of $U_h(t_{m+1})$ and $U_h'(t_{m+1})$ in t_m

$$\begin{aligned} le_{m+1} =& \tau G^{-1}(\tau, h^2)(I_N \otimes A)\Big[\frac{1}{2}U_h''(t_m + \zeta\tau) - U_h''(t_m + \overline{\zeta}\tau)\Big] \\ & + h^2G^{-1}(\tau, h^2)(I_N \otimes B)\gamma_h(t_{m+1}) \end{aligned} \tag{4.11}$$

where $\zeta, \overline{\zeta} \in (0,1)$. Therefore, for the full truncation error we have for $t \in I^*$ (cf. the foregoing section) the estimate

$$\|le_{m+1}\| \leq C_0\left(\tau\|G^{-1}(\tau, h^2)(I_N \otimes A)\| \ + \ h^2\|G^{-1}(\tau, h^2)(I_N \otimes B)\|\right) \tag{4.12}$$

where C_0 is a positive constant independent of τ and h. The terms on the right hand side of this inequality can be estimated easily with Lemma 4.2 because it implies that the following relations hold for $\tau, h \to 0$

$$\begin{aligned}
\tau^{\nu_{d,t}-1}\|G^{-1}(\tau,h^2)\| &\le \kappa_1 \quad \text{for } \nu_{d,t} \ge 0, \\
h^{-2}\|G^{-1}(\tau,h^2)\| &\le \kappa_2 \quad \text{for } \nu_{d,x} = 0, \\
\tau^{\nu_{d,t}-2}\|G^{-1}(\tau,h^2)(I_N \otimes A)\| &\le \kappa_3 \quad \text{for } \nu_{d,t} \ge 1, \\
h^{\nu_{d,x}-3}\|G^{-1}(\tau,h^2)(I_N \otimes B)\| &\le \kappa_4 \quad \text{for } \nu_{d,x} \ge 1
\end{aligned} \tag{4.13}$$

whereby possibly τ and h cannot go independent towards zero. This means it may be that one (or none) of the following conditions must be satisfied

$$c_0 \le \frac{\tau}{h^2} \quad \text{or} \quad \frac{\tau}{h^2} \le c_1 \quad \text{or} \quad c_0 \le \frac{\tau}{h^2} \le c_1, \qquad c_0, c_1 \in \mathbb{R}^+. \tag{4.14}$$

(For illustration see examples 4.1 and 4.2 below.)

Remark 4.5. If $c_0 = 0$ and $c_1 = \infty$, then there is no restriction on τ and h. □

The estimate (4.12) and the asymptotic relations (4.13) can now be used to state (in the sense of Definition 4.3) an accuracy result:

Theorem 4.1. Let for a given PDAE (1.1) the asymptotic relations (4.13) be valid (eventually under one of the conditions (4.14)). Then the BTCS scheme is for $t \in I^*$ (conditionally) accurate of order (2,1) for

- an uniform time index $\nu_{d,t} = 0$ and a space index $\nu_{d,x} \ge 0$,
- an uniform time index $\nu_{d,t} \ge 0$ and a space index $\nu_{d,x} = 0$ under a condition $0 < \tilde{c}_0 \le \frac{\tau}{h^2}$,
- an uniform time index $\nu_{d,t} = 1$ and a space index $\nu_{d,x} = 1$ under a condition $0 < \tilde{c}_0 \le \frac{\tau}{h^2}$.

Remark 4.6. The condition $0 < \tilde{c}_0 \le \frac{\tau}{h^2}$ does not follow from the conditions (4.14) being eventually necessary for the relations (4.13) to be valid. □

Proof. In the following K is a constant which is independent of τ and h and possibly different in different expressions.

First we consider the case $\nu_{d,t} = 0$, $\nu_{d,x} \ge 0$. Using in (4.13) the first inequality we can estimate the right hand side of ineq. (4.12) further by (the arguments of G are suppressed)

$$K\left(\tau^2\|\tau^{-1}G^{-1}\| + \tau h^2\|\tau^{-1}G^{-1}\|\right) = \mathcal{O}(\tau^2) + \mathcal{O}(\tau h^2).$$

According to Definition 4.3 we see that the BTCS scheme is accurate of order (2,1) for these indexes.

In the same way we obtain in the case $\nu_{d,t} \ge 0$, $\nu_{d,x} = 0$ by means of the second inequality in (4.13) for the right hand side of (4.12) the upper bound

$$K\left(\tau h^2\|h^{-2}G^{-1}\| + h^4\|h^{-2}G^{-1}\|\right) = \mathcal{O}(\tau h^2) + \mathcal{O}(h^4).$$

If we now take into account the condition in the second case, then the right hand side reduces to $\mathcal{O}(\tau^2) + \mathcal{O}(\tau h^2)$. Using again Definition 4.3 we obtain the second statement of the theorem.

The third statement can be proved by using the last two inequalities of (4.13). We find for the right of (4.12) the upper bound

$$K\left(\tau^{3-\nu_{d,t}}\|\tau^{\nu_{d,t}-2}G^{-1}(I_N \otimes A)\| + h^{5-\nu_{d,x}}\|h^{\nu_{d,x}-3}G^{-1}(I_N \otimes B)\right).$$

Inserting $\nu_{d,t} = \nu_{d,x} = 1$ into this bound it reduces to an expression of the form $\mathcal{O}(\tau^2) + \mathcal{O}(h^4)$ which is equivalent to $\mathcal{O}(\tau^2) + \mathcal{O}(\tau h^2)$ because $0 < \tilde{c}_0 \le \frac{\tau}{h^2}$. This proves the third statement of the theorem. ⊠

Now we introduce the *full discretization error* at time level t_{m+1}

$$\varepsilon_h(t_{m+1}) = U_h(t_{m+1}) - U_{m+1}$$

and

Definition 4.4. If the order relation

$$\|\varepsilon_h(t_{m+1})\| = O(h^p) + \mathcal{O}(\tau^q) \quad \text{for} \quad \tau,\, h \to 0,\; m = 0, 1, ..., N, \tag{4.15}$$

is valid, then the BTCS scheme is called convergent of order (p, q). If this relation holds under a time space condition (4.14), then the BTCS scheme is said to be conditionally convergent of order (p, q).

With eq. (4.8) for the truncation error we obtain the recursion

$$\varepsilon_h(t_{m+1}) = G^{-1}(\tau, h^2)\left(\frac{1}{\tau} I_N \otimes A\right)\varepsilon_h(t_m) + le_{m+1}, \tag{4.16}$$

and by induction we get

$$\varepsilon_h(t_{m+1}) = \left(G^{-1}(\tau, h^2)\frac{1}{\tau} I_N \otimes A\right)^{m+1}\varepsilon_h(0) + \sum_{i=0}^{m}\left(G^{-1}(\tau, h^2)\frac{1}{\tau} I_N \otimes A\right)^{i} le_{m+1-i}. \tag{4.17}$$

It follows

$$\begin{aligned}\|\varepsilon_h(t_{m+1})\| \le{} & \left\|\left(G^{-1}(\tau, h^2)\frac{1}{\tau} I_N \otimes A\right)^{m+1}\right\| \|\varepsilon_h(0)\| \\ & + \max_{i=0}^{m} \|le_{i+1}\| \sum_{i=0}^{m}\left\|\left(G^{-1}(\tau, h^2)\frac{1}{\tau} I_N \otimes A\right)^{i}\right\|\end{aligned} \tag{4.18}$$

We require for $0 < m\,\tau \le t^*$

$$\sup\left\{\left\|\left(G^{-1}(\tau, h^2)\frac{1}{\tau} I_N \otimes A\right)^{i}\right\|,\ i \in \mathbb{N}, \quad \tau,\, h \text{ possibly restricted according to (4.14)}\right\} < \infty. \tag{4.19}$$

Remark 4.7. If the matrices A and B are regular, then this requirement is the Lax stability condition. □

Assuming $\|\varepsilon_h(0)\| = \mathcal{O}(h^2)$, we get for $t \in I^*$ the estimate

$$\|\varepsilon_h(t_{m+1})\| \le \overline{K}(m+1)\max_{i=0}^{m}\|le_{i+1}\| \;\le\; K \max_{i=0}^{m}\frac{\|le_{i+1}\|}{\tau}$$

where $\overline{K}$ and K are positive constants independent of τ and h for $\tau,\, h \;\to 0$ under the condition $\tau\, m = t \in I^*$ (t fixed) . With Theorem 4.1 we obtain immediately

Theorem 4.2. Suppose the assumptions of Theorem 4.1 and the estimate (4.19) are valid. Then the BTCS scheme is (conditionally) convergent with respect to $\|\cdot\|$ of order (2,1) for $t \in I^*$ and for

- $\nu_{d,t} = 0$ and $\nu_{d,x} \ge 0$,

- $\nu_{d,t} \geq 0$ and $\nu_{d,x} = 0$ under a condition $0 < \tilde{c}_0 \leq \frac{\tau}{h^2}$,
- $\nu_{d,t} = \nu_{d,x} = 1$ under a condition $0 < \tilde{c}_0 \leq \frac{\tau}{h^2}$.

4.2. Finite Fourier representation of the full discretization error. Beside the full discretization error $\varepsilon_h(t_{m+1})$ we define a generalized Fourier component of this error in terms of the low order $(n \times n)$–matrices $G_k^{-1}\frac{A}{\tau}$ and $G_k^{-1}B$. This Fourier error component is (as shown below) convenient for an investigation of necessary conditions for the convergence of the numerical scheme considered.

In the following we use in the vector space $\mathbb{R}^N$ the discrete L_2-norm corresponding to the inner product

$$< x, y >= hx^T y = \sum_{i=1}^{N} h\, x_i\, y_i\ , \quad x, y \in \mathbb{R}^N,\ h = \frac{2l}{N+1}.$$

The eigenvectors Φ_k (cf. eq. (3.6)) are supposed to be normed, $\|\Phi_k\| = \sqrt{< \Phi_k, \Phi_k >} = 1$.

Lemma 4.3. Let for fixed $\tau, h > 0$

(i) G_k^{-1} be regular for $k = 1, .., N$;
(ii) $u(t, x)$ (and hence $U_h(t)$ and $\gamma_h(t)$) be sufficiently smooth for $t \in I^*$, $x \in [-l, l]$;
(iii) $\varepsilon_h(t_j)$ be represented as a finite generalized Fourier series with respect to Φ_k (in analogy to eq. (3.5))

$$\varepsilon_h(t_j) = \sum_{k=1}^{N} \Phi_k \otimes e_{h,k}^j, \quad j = 0, 1, ..., \tag{4.20}$$

where $e_{h,k}^j \in \mathbb{R}^n$.

Then $e_{h,k}^{m+1}$ can be estimated for $t \in I^*$ by

$$\|e_{h,k}^{m+1}\| \leq \left\| \left(G_k^{-1}\frac{A}{\tau} \right)^m \right\| \|e_{h,k}^0\| + \omega \sum_{i=0}^{m} \left\| \left(G_k^{-1}\frac{A}{\tau} \right)^i \right\| \left[h^2 \|G_k^{-1}B\| + \tau^2 \left\| G_k^{-1}\frac{A}{\tau} \right\| \right] \tag{4.21}$$

where the positive coefficient ω is independent of τ and h.

Proof. We rewrite eq. (4.16) in the form

$$G(\tau, h^2)\, \varepsilon_h(t_{m+1}) = \frac{1}{\tau}(I_N \otimes A)\, \varepsilon_h(t_m) + G(\tau, h^2)\, le_{m+1}$$

and multiply it from the left with $h(\Phi_{k'}^T \otimes I_n)$, $k' \in \{1, ..., N\}$. Using now for $G\, le_{m+1}$ eq. (4.11), one gets with the Fourier representation indicated in assumption (iii)

$$\begin{aligned} G_k\, e_{h,k}^{m+1} =& \frac{A}{\tau}\, e_{h,k}^m + h^2(h\, \Phi_k^\top \otimes B)\gamma_h(t_{m+1}) \\ &+ \tau(h\, \Phi_k^\top \otimes A) \left[\frac{1}{2}U_h''(t_m + \zeta\tau) - U_h''(t_m + \overline{\zeta}\tau) \right]. \end{aligned}$$

Combining this formula with assumptions (i) and (ii) the assertion of the lemma follows by taking norm estimates. ⊠

Remark 4.8. The estimate (4.21) has formally the same τ- and h-dependence as the estimate (4.18), as it should be. □

The Fourier error components $e_{h,k}^{m+1}$ can be used to derive a necessary condition for the convergence of the numerical scheme considered.

Lemma 4.4. Suppose that the assumptions of Lemma 4.3 are valid for $0 < \tau \le \tau_0$, $0 < h \le h_0$ and that $\|\varepsilon_h(t_{m+1})\|$ goes to zero for $\tau,\, h \to 0$. Then the relation

$$\|e_{h,k}^{m+1}\| \to 0 \quad \text{for} \quad \tau,\, h \to 0$$

holds for $t \in I^*$. Furthermore the convergence order (with respect to τ and h) of $\|e_{h,k}^{m+1}\|$ is the same as the convergence order of $\|\varepsilon_h(t_{m+1})\|$.

Proof. We multiply eq. (4.20) from the left by $h(\Phi_{k'}^T \otimes I_n)$ and use the orthogonality of the eigenvectors Φ_k, $\Phi_{k'}$ to get

$$e_{h,k'}^{m+1} = h(\Phi_{k'}^T \otimes I_n)\varepsilon_h(t_{m+1}).$$

From this formula the statement of the lemma follows by taking norms. ⊠

In the following examples it is assumed that the initial error component $\|e_{h,k}^0\|$ in ineq. (4.21) is $\mathcal{O}(h^2)$.

Example 4.1. Let $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} b_1 & b_2 \\ 0 & 0 \end{pmatrix}$ with $b_1 < 0$, $C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The uniform differential time index and the differential spatial index are easily shown to be $\nu_{d,t} = 1$ and $\nu_{d,x} = 1$, respectively [7]. $\|G_k^{-1}\frac{A}{\tau}\|$ (we take spectral norms) is of order $\mathcal{O}(1)$ for $\tau, h \to 0$, while $\|G_k^{-1}B\|$ is of order $\mathcal{O}(\tau)$ as well for $\frac{\tau}{h^2} \ge c_0$ as for $\frac{\tau}{h^2} \le c_1$ (cf. the ineqs. in (4.14)). Therefore, no restrictions for $\frac{\tau}{h^2}$ are needed in this example. Of special interest is the norm

$$\left\|G_k^{-1}\frac{A}{\tau}\right\| = \frac{1}{|1+\tau(\lambda_k b_1 + 1)|}$$

which must be summed over in the estimate of the right hand side of ineq. (4.21). Because $\|G_k^{-1}\frac{A}{\tau}\| \le 1$ for $\tau,\, h \to 0$,

$$\sum_{i=0}^{m}\left\|\left(G_k^{-1}\frac{A}{\tau}\right)^i\right\| \le m+1$$

and $m\tau = t \in I^*$ (t fixed), we find the order relation of the sum

$$\sum_{i=0}^{m}\left\|\left(G_k^{-1}\frac{A}{\tau}\right)^i\right\| = \mathcal{O}\left(\frac{1}{\tau}\right)$$

for $\tau, h \to 0$. Inserting this into ineq. (4.21) and taking $\|e_{h,k}^0\| = \mathcal{O}(h^2)$ into account we get the result that the implicit Euler method for this example gives convergence of the error component $e_{h,k}^{m+1}$ of the following order

$$\|e_{h,k}^{m+1}\| = \mathcal{O}(\tau) + \mathcal{O}(h^2).$$

Example 4.2. Let the PDAE (1.1) be generated by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Its indexes are $\nu_{d,t} = 2, \nu_{d,x} = 3$. For convenience the basic matrices $G_k^{-1}\frac{A}{\tau}$ and $G_k^{-1}B$ are given,

$$G_k^{-1}\frac{A}{\tau} = \begin{pmatrix} (1+\tau)^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & (\tau - \tau\lambda_k)^{-1} & 0 \end{pmatrix}, \quad G_k^{-1}B = \begin{pmatrix} 0 & -\frac{\tau}{1+\tau} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (\lambda_k - 1)^{-1} \end{pmatrix}.$$

From this it is seen that the norm of $G_k^{-1}\frac{A}{\tau}$ is

$$\left\| G_k^{-1}\frac{A}{\tau} \right\| = \max\left\{ \frac{1}{1+\tau}, \frac{1}{\tau(1-\lambda_k)} \right\}$$

which goes to 1 for τ, $h \to 0$ (this is needed for convergence, cf. ineq. (4.21)) only under the condition $-\tau\lambda_k = \kappa_k \frac{\tau}{h^2} \geq 1$ where $\kappa_k = 4\sin^2\left(\frac{\pi k h}{4l}\right)$ (i.e. $c_0 = \kappa_k^{-1}$). But this implies

$$\|G_k^{-1}\, B\| = \max\left\{ \frac{\tau}{1+\tau}, \frac{1}{(1-\lambda_k)} \right\} \leq \max\left\{ \frac{\tau}{1+\tau}, \frac{\kappa_k \tau}{\kappa_k + h^2} \right\} = \mathcal{O}(\tau).$$

Since $m\,\tau = t$ is fixed for convergence and

$$\sum_{i=0}^{m} \left\| \left(G_k^{-1}\frac{A}{\tau} \right)^i \right\| = \mathcal{O}\left(\frac{1}{\tau}\right) \quad \text{for } \tau, h \to 0$$

we get from ineq. (4.21) the result that the error norm $\|e_{h,k}^{m+1}\|$ tends to zero according to $\|e_{h,k}^{m+1}\| = \mathcal{O}(\tau) + \mathcal{O}(h^2)$ for $\tau, h \to 0$ under the condition $0 < c_0 = \frac{1}{\kappa_k} \leq \frac{\tau}{h^2}$ (i.e. the first of the conditions in (4.14) must be satisfied).

By the way of this example we note that asymptotic estimates (or order relations) can also be gained from ineq. (4.21) by means of the relations (4.13) which can easily be written in terms of $\|G_k^{-1}\|$, $\|G_k^{-1}\frac{A}{\tau}\|$ and $\|G_k^{-1}B\|$. However, the asymptotic estimates based on these relations are in general only crude ones. In the example considered here we would obtain by this method

$$\|e_{h,k}^{m+1}\| = \mathcal{O}(h^2) + \mathcal{O}(h^{3-\nu_{d,x}}) + \mathcal{O}(\tau^{2-\nu_{d,t}}) \quad \text{for} \quad \tau,\, h \to 0$$

under the condition $c_0 \leq \frac{\tau}{h^2}$. The expression on the right hand side goes not to zero for τ, $h \to 0$ (because $\nu_{d,t} = 2$, $\nu_{d,x} = 3$). In contrast, the estimates given before are stronger. For these the proper quality (cf. definition 4.1) which is the basis of the relations (4.13) is not used.

4.3. The Crank–Nicolson scheme. This discretization scheme can be generated if the trapezoidal rule (in the following expressed by the lower index T) is applied to eq. (3.2). Using the grid Ω_h defined in section 3 the analogue to eq. (4.1) for the Crank–Nicolson scheme is

$$G_T(\tau,h^2)U_{T,m+1} = \left[\frac{1}{\tau}I_N \otimes A - H_h\right] U_{T,m} + \frac{1}{2}\left(\tilde{F}(t_{m+1} + \tilde{F}(t_m)\right) \tag{4.22}$$

with

$$G_T(\tau,h^2) = \frac{1}{\tau}I_N \otimes A + H_h, \qquad H_h = \frac{1}{2}\left(\frac{1}{h^2}P \otimes B + I_N \otimes C\right), \tag{4.23}$$

and $\tilde{F}(t) = F(t) - r(t)$. We consider again the full truncation error $le_{T,m+1} = U_h(t_{m+1}) - \hat{U}_{T,m+1}$ where now

$$\hat{U}_{T,m+1} = G_T^{-1}(\tau,h^2)\left[\frac{1}{\tau}I_N \otimes A - H_h\right] U_h(t_m) + \frac{1}{2}G_T^{-1}(\tau,h^2)\left[\tilde{F}(t_{m+1}) + \tilde{F}(t_m)\right]. \tag{4.24}$$

In the same manner as in subsection 4.1 we get for $le_{T,m+1}$ the representation

$$\begin{aligned} le_{T,m+1} =& G_T^{-1}(\tau,h^2)\Big\{\left(\frac{1}{\tau}I_N \otimes A\right)\left[U_h(t_{m+1}) - U_h(t_m) - \frac{\tau}{2}\left(U_h'(t_{m+1}) - U_h'(t_m)\right)\right] \\ &+ \frac{1}{2}\left(\alpha_h(t_{m+1}) + \alpha_h(t_m)\right)\Big\}. \end{aligned} \tag{4.25}$$

which leads to the estimate

$$\|le_{T,m+1}\| \leq C_1\left(\tau^3\|G_T^{-1}(\tau,h^2)\frac{1}{\tau}(I_N \otimes A)\| + h^2\|G_T^{-1}(\tau,h^2)(I_N \otimes B)\|\right) \tag{4.26}$$

where C_1 is positive and independent of τ and h. This is the analogue of ineq. (4.12) for the BTCS scheme.

Taking into account that the proper quality is valid also for the matrix G_T^{-1} (cf. Lemma 4.2) we obtain from the foregoing estimate the following accuracy result:

Theorem 4.3. Let for a given PDAE (1.1) the asymptotic relations (4.13) (where G must be replaced by G_T) be valid (possibly under one of the conditions (4.14)). Then the Crank–Nicolson scheme is for $t \in I^*$ (conditionally) accurate of order (2,2) for

- $\nu_{d,t} = 0$ and $\nu_{d,x} \geq 0$,
- $\nu_{d,t} \geq 0$ and $\nu_{d,x} = 0$ under a condition $0 < \hat{c}_0 \leq \frac{\tau}{h^2}$,
- $\nu_{d,t} = \nu_{d,x} = 1$ under a condition $0 < \hat{c}_0 \leq \frac{\tau}{h^2}$.

It is (conditionally) accurate of order (2,1) for

- $\nu_{d,t} = 2$ and $\nu_{d,x} = 1$ under a condition $0 < \hat{c}_0 \leq \frac{\tau}{h^2}$.

This Theorem is analogous to Theorem 4.1 for the BTCS scheme.

In order to obtain a convergence result also for the Crank–Nicolson scheme (like Theorem 4.2 for the BTCS scheme) we consider the full discretization error $\varepsilon_{T,h}(t_{m+1}) = U_h(t_{m+1}) - U_{T,m+1}$ under the assumption that the matrix $G_T(\tau, h^2)$ is regular. Because

$$U_{T,m+1} = G_T^{-1}\Big(\Big[\frac{1}{\tau} I_N \otimes A - H_h\Big] U_{T,m} + \frac{1}{2}\left(\tilde{F}(t_{m+1}) + \tilde{F}(t_m)\right)\Big)$$

(this follows from eq. (4.22)) and $U_h(t_{m+1}) = le_{T,m+1} + \hat{U}_{T,m+1}$ (see eq. (4.8)) we can write

$$\varepsilon_{T,h}(t_{m+1}) = G_T^{-1}\left(\frac{1}{\tau} I_N \otimes A - H_h\right)\varepsilon_{T,h}(t_m) + le_{T,m+1} \tag{4.27}$$

which is analogous to eq.(4.16). Thus we also get the analogue of ineq. (4.18) which reads for the Crank–Nicolson method

$$\|\varepsilon_{T,h}(t_{m+1})\| \le \left\| \left(G_T^{-1}(\tau, h^2)\left[\frac{1}{\tau} I_N \otimes A - H_h\right]\right)^{m+1} \right\| \cdot \|\varepsilon_{T,h}(0)\| \tag{4.28}$$

$$+ \max_{j=0}^{m} \|le_{j+1}\| \sum_{i=0}^{m} \left\| \left(G_T^{-1}(\tau, h^2)\left[\frac{1}{\tau} I_N \otimes A - H_h\right]\right)^{i} \right\| \tag{4.29}$$

Requiring that for $0 < m\tau \le t^*$

$$\sup\Big\{ \Big\| \left(G_T^{-1}(\tau, h^2)\left[\frac{1}{\tau} I_N \otimes A - H_h\right]\right)^{i} \Big\|,\ i \in \mathbb{N}^+, \tag{4.30}$$

$$\tau,\ h \quad \text{possibly restricted according to (4.14)}\Big\} < \infty$$

we obtain the following convergence theorem.

Theorem 4.4. Suppose the assumptions of Theorem 4.3 and the estimate (4.30) are valid. Then the Crank–Nicolson scheme is for $t \in I^*$ (conditionally) convergent of order (2,2) for

- $\nu_{d,t} = 0$ and $\nu_{d,x} \ge 0$,
- $\nu_{d,t} \ge 0$ and $\nu_{d,x} = 0$ under a condition $0 < \hat{c}_0 \le \frac{\tau}{h^2}$,
- $\nu_{d,t} = \nu_{d,x} = 1$ under a condition $0 < \hat{c}_0 \le \frac{\tau}{h^2}$.

It is convergent of order (2,1) for

- $\nu_{d,t} = 2$ and $\nu_{d,x} = 1$ under a condition $0 < \hat{c}_0 \le \frac{\tau}{h^2}$.

This theorem is the analogue of Theorem 4.1 for the BTCS scheme.

Furthermore, as for the BTCS method a finite Fourier representation of the full discretization error $\varepsilon_{T,h}(t_{m+1})$ can be given also for the Crank–Nicolson scheme. It is based on the eigenvectors of the matrix $\frac{1}{2h^2}P$ (for P see eq. (3.3)),

$$\frac{1}{2h^2} P\Phi_{T,k} = \lambda_{T,k}\Phi_{T,k}, \qquad k = 1, ..., N.$$

For completeness we give the analogue of Lemma 4.3:

Lemma 4.5. Let for fixed τ, $h > 0$

(i) $G_{T,k}^{-1} = \left(\frac{1}{\tau}A + \lambda_{T,k}B + \frac{1}{2}C\right)^{-1}$ be regular for $k = 1, ..., N$;
(ii) $u(t,x)$ (and hence $U_h(t)$ and $\gamma_h(t)$) be sufficiently smooth for $t \in I^*$, $x \in [-l, l]$;
(iii) $\varepsilon_{T,h}(t_j)$ be represented as a finite Fourier series with respect to $\Phi_{T,k}$,

$$\varepsilon_{T,h}(t_j) = \sum_{k=1}^{N} \Phi_{T,k} \otimes e_{T,h,k}^{j}, \quad j = 0, 1, ..., \quad e_{T,h,k}^{j} \in \mathbb{R}^n.$$

Then $e_{T,h,k}^{m+1}$ can be estimated for $t \in I^*$ by

$$\begin{aligned} \|e_{T,h,k}^{m+1}\| \leq & \left\| \left(G_{T,k}^{-1}\left[\frac{1}{\tau}A - \lambda_{T,k}B - \frac{1}{2}C\right]\right)^m \right\| \cdot \|e_{T,h,k}^{0}\| \\ & + \omega_T \sum_{i=0}^{m} \left\| \left(G_{T,k}^{-1}\left[\frac{1}{\tau}A - \lambda_{T,k}B - \frac{1}{2}C\right]\right)^i \right\| \cdot \left[h^2\|G_{T,k}^{-1}B\| + \tau^3\|G_{T,k}^{-1}\frac{1}{\tau}A\|\right] \end{aligned}$$

where ω_T is a positive constant independent of τ and h.

Proof. We multiply eq. (4.27) from the left by G_T and apply eq. (4.25) to express $G_T(\tau, h^2)le_{T,m+1}$ (after a Taylor expansion, cf. section 4.2) in terms of the matrices A and B. The resulting equation is then multiplied from the left by $h\left(\Phi_{T,k'}^T \otimes I_n\right)$. Using the orthogonality of the system $\{\Phi_{T,k}\}_1^N$, the definition of the scalar product (cf. subsection 4.2) and assumption (i) we obtain the statement of the lemma by taking norms. ⊠

We note that the analogue of Lemma 4.4 holds also for the Crank–Nicolson scheme ($\varepsilon_h(t_{m+1})$ must be replaced by $\varepsilon_{T,h}(t_{m+1})$ and so on). Lemma 4.5 is useful especially because the norms of only low order $(n \times n)$−matrices must be analyzed (provided n is low as is often the case in applications).

4.4. Numerical examples. In the following some results of numerical test calculations with the BTCS scheme are presented. For specific examples we confirm the convergence results derived in subsections 4.1 and 4.2, and we use this scheme also to demonstrate that the consistency of the BCs is essential for an accurate solution of the PDAE (1.1). Furthermore the problem of an index jump discussed in [7] is considered. Using again the BTCS scheme it is shown that if an index-jump-free PDAE is semidiscretized not properly, then this semidiscretized version may have an index jump. This can cause large errors in the corresponding numerical solution.

In the examples considered below Dirichlet BCs are used throughout, and we take $l = 1$. Further the inhomogeneity $f(t,x)$ of the PDAE (1.1) is chosen such that an exact solution is known. BCs (1.2) and ICs (1.3) needed for numerical calculations are then taken from the exact solution. Therefore the BCs and ICs can (if necessary) assumed to be consistent.

Order relations. First we note that the full discretization error ε_h defined in section 4 can be considered to be a function of τ (see, e.g., eq. (4.17)). We therefore write here $\varepsilon_{h,\tau}$.

In order to show the convergence order clearly, we consider example 4.1 with $b_1 = b_2 = -1$. Let $f(t,x)$ be such that

$$u_I(t,x) = \left(x(x^2 - l^2)\, e^{-t},\ (x^2 - l^2)\, e^{10\,t}\right)^\top \in \mathbb{R}^2$$

is the exact solution. Because the components of $u_I(t,x)$ are polynomials in x of degree ≤ 3, there will be no spatial error by discretization of u_{xx} with standard second order finite differences, i.e. $\alpha_h(t) = 0$ (cf. (3.8)). Comparing $\varepsilon_{h,\tau}$ with $\varepsilon_{h,\frac{\tau}{2}}$ under, e.g., the condition $\tau\, m = t = 1$ we can determine the numerical order of convergence $p_{num,\tau} = \log_2 \|\varepsilon_{h,\tau}\| - \log_2 \|\varepsilon_{h,\frac{\tau}{2}}\|$ with respect to τ. Analogously if we choose $f(t,x)$ such that

$$u_{II}(t,x) = \Big(x^6(x^2-l^2)\,t,\ x^4(x^2-l^2)\,t\Big)^\top \in \mathbb{R}^2,$$

is the exact solution, then the second order derivatives in time of $U_h(t)$ and $F(t)$ in eq. (4.10) vanish identically, and the error is only an error in space. Its convergence order can be determined numerically by means of the expression $p_{num,h} = \log_2 \|\varepsilon_{h,\tau}\| - \log_2 \|\varepsilon_{\frac{h}{2},\tau}\|$ with respect to h. Table 4.1 shows (in agreement with the investigation in the foregoing subsections) that the τ-order is 1, and the h-order is 2.

$10^{-1}\,\tau^{-1}$	$p_{num,\tau}$ using $u_I(t,x)$: 2^2	2^3	2^4	2^5	2^6	2^7	$p_{num,h}$ using $u_{II}(t,x)$: 2^2	2^3	2^4	2^5	2^6	2^7
$10^{-1}\,h^{-1}$												
1	1.01	1.01	1.00	1.00	1.00	1.00	1.92	1.92	1.92	1.92	1.92	1.92
2^1	1.01	1.01	1.00	1.00	1.00	1.00	1.98	1.98	1.98	1.98	1.98	1.98
2^2	1.01	1.01	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00
2^3	1.01	1.01	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00

TABLE 4.1. Numerical order of convergence for example 4.1

Considering analogously example 4.2 with

$$u_I(t,x) = \Big(x(x^2-l^2)e^{-t},\ (x^2-l^2)e^{t},\ (x+l)(x^2-l^2)e^{10\,t}\Big)^\top \in \mathbb{R}^3$$
$$u_{II}(t,x) = \Big(x^6(x^2-l^2)\,t,\ x^4(x^2-l^2)\,t,\ (x^4-l^4)\,t\Big)^\top \in \mathbb{R}^3$$

we get numerical convergence orders for $m\tau = 1$ as given in table 4.2, i.e. the convergence is of order $\mathcal{O}(\tau) + \mathcal{O}(h^2)$. The same convergence order has been derived for the error norm $\|e^m_{h,k}\|$ (under the condition $m\tau = 1$ for $\tau \to 0$), cf. the foregoing subsection.

Consistency of BCs. The following example shows that it is important to know those components of u for which boundary conditions can be prescribed arbitrarily. For this we consider the example

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} u_t + \begin{pmatrix} -1 & -1 \\ 0 & 0 \end{pmatrix} u_{xx} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} u = f \tag{4.31}$$

whose right hand side function $f = (f_1, f_2)^\top$ is chosen such that the exact solution is given by

$$u(t,x) = \begin{pmatrix} v(t,x) \\ w(t,x) \end{pmatrix} := \begin{pmatrix} (x^2-l^2)(x^2+4)x\cos(\pi\,t) \\ (x^2-l^2)(x^5-2x^2+5)e^{-t} \end{pmatrix}.$$

	$p_{num,\tau}$ using $u_I(t,x)$						$p_{num,h}$ using $u_{II}(t,x)$					
$10^{-1}\,\tau^{-1}$	2^2	2^3	2^4	2^5	2^6	2^7	2^2	2^3	2^4	2^5	2^6	2^7
$10^{-1}\,h^{-1}$												
1	0.99	0.99	1.00	1.00	1.00	1.00	1.77	1.77	1.77	1.77	1.77	1.77
2^2	0.99	0.99	1.00	1.00	1.00	1.00	1.95	1.95	1.95	1.95	1.95	1.95
2^4	0.99	0.99	1.00	1.00	1.00	1.00	1.99	1.99	1.99	1.99	1.99	1.99
2^6	1.03	1.01	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00

TABLE 4.2. Numerical order of convergence for example 4.2

Since the matrix A is regular the corresponding MOL-system (3.2) is an ordinary differential equation which can be treated without problems. The set $\mathfrak{M}_{BC}$ can be chosen as $\mathfrak{M}_{BC} = \{1\}$ (another possibility could be $\mathfrak{M}_{BC} = \{2\}$).

In order to illustrate the difference between consistent and inconstant BCs for the second component of the solution vector, we compare the numerical solution $u_h(t,x)$ with another numerical solution $\tilde{u}_h(t,x)$ where $\tilde{v}_h(t,\pm l) = v(t,\pm l)$ and $\tilde{w}_h(t,\pm l) = w(t,\pm l) + e^{2t} - 1$, i.e. $\tilde{w}_h(t,\pm l)$ is an inconsistent BC because f vanishes at the boundary. The ICs $u_h(0,x)$ and $\tilde{u}_h(0,x)$ are the same (i.e. equal to $u(0,x)$) such that the compatibility condition (1.4) is guaranteed.

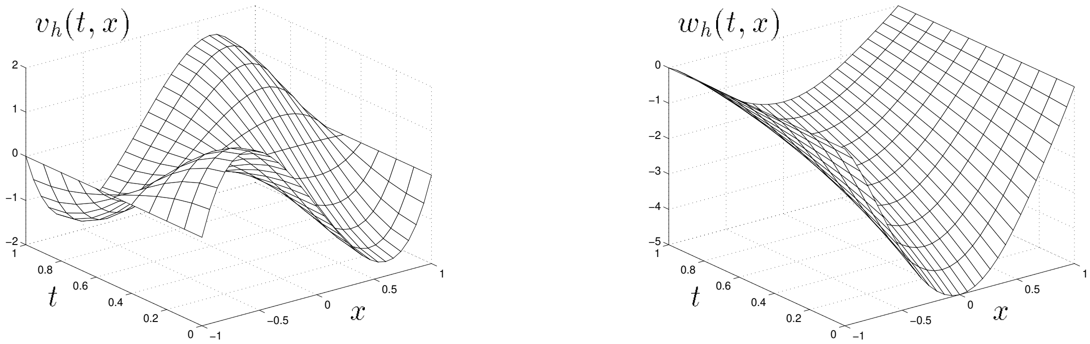

FIGURE 4.1. Example: PDAE 4.31, consistent BC, $h = \frac{1}{15}$, $\tau = 0.1$

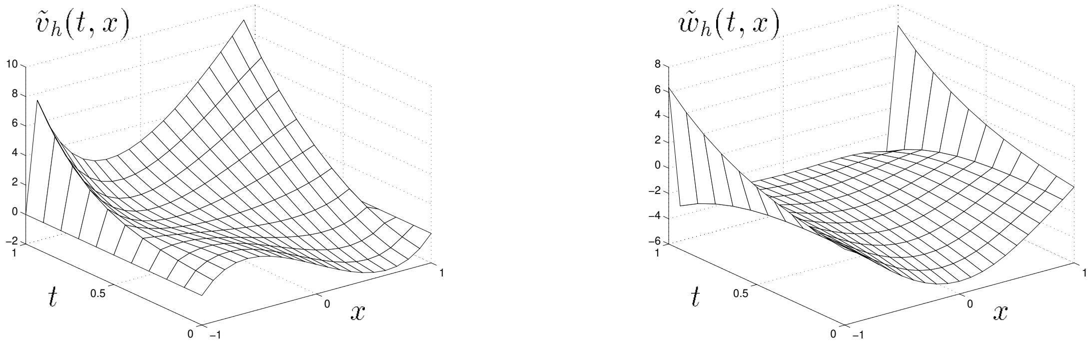

FIGURE 4.2. Example: PDAE (4.31), inconsistent BC, $h = \frac{1}{15}$, $\tau = 0.1$

Comparing figures 4.1 and 4.2 we see that the solutions for consistent and inconsistent BCs are very different. Of course, this is expected because two different problems were solved.

Especially the second component $\tilde{w}_h(t,x)$ of the numerical solution calculated with an inconsistent BC has near the boundary points $\pm l$ in part very large gradients.

The problem of an index jump in the MOL-DAE. Suppose the matrices A, B, C are defined by

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 \\ 1 & c \end{pmatrix}.$$

Let further $\mu_k := -\left(\frac{k\pi}{2l}\right)^2$, $k = 1, 2, \ldots$, be the k-th eigenvalue mentioned in assumption (IV) of section 2 (cf. also [7], subsection 2.2).

It is easy to see that the matrix pencil $(A, \mu_k B + C)$ has Riesz index 1 for $c \neq \mu_k$ and Riesz index 2 for $c = \mu_{\bar{k}}$ for some $\bar{k}$, i.e. the PDAE has uniform differential time index $\nu_{d,t} = 1$ if $c \neq \mu_k \ \forall k \in \mathbb{N}^+$ and an index jump if $c = \mu_{\bar{k}}$ for one $\bar{k} \in \mathbb{N}^+$. Of course it may be that the MOL-DAE (3.2) has an index jump for some h and special matrices A, B, C even if the PDAE has an uniform differential time index. In this example this would happen if $c = \lambda_{\bar{k}}$ ($c \neq \mu_k \forall\, k$) for a certain $h > 0$ (this can be shown by a linear transformation of the MOL-DAE using the Kronecker product).

Let $f(t,x)$ be chosen such that the exact solution of the PDAE is

$$u(t,x) = \begin{pmatrix} v(t,x) \\ w(t,x) \end{pmatrix} := \begin{pmatrix} x^5\,(x^2 - l^2)\cos(\pi t) \\ x^2\,(x^2 - l^2)\,e^{-t} \end{pmatrix}.$$

and define $c := \lambda_1(h) = -\frac{4}{h^2}\sin^2\left(\frac{\pi\, h}{4\, l}\right)$ for $h = h_0 = 0.05$ ($N = 39$). (The numerical value of $\lambda_1(h_0)$ is -2.466133.) Figure 4.3 shows the absolute error err_h^w of the w-component between the numerical solution and the exact solution in each space and time point.

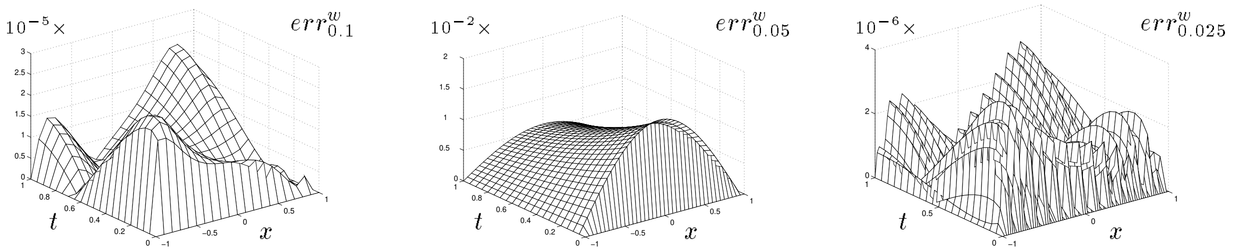

FIGURE 4.3. $|w_{num}(t,x) - w(t,x)|$, $h = 0.1, 0.05, 0.025$, $\tau = 0.05$

We see that the error for $h = h_0$ is greater than for $h \neq h_0$. Effects like these can also occur in other more complicated examples. We note that because $\lambda_k(h) = \mu_k + \mathcal{O}(h^2)$ for $h \to 0$ there is no index jump of the MOL-DAE for sufficiently small h. At the whole this example demonstrates that we must be careful in the numerical treatment of PDAEs.

We note that the error shown in Figure 4.3 can be explained also from another point of view. For this we insert the expansion (3.5) into the MOL-eq. (3.2) to get the DAE for the time dependent coefficient vector w_k

$$Aw_k'(t) + [\lambda_k(h)\, B + C]w_k(t) = \tilde{f}_k(t), \quad k = 1, \ldots, N,$$

where $\tilde{f}_k(t)$ is the generalized Fourier coefficient of $\tilde{F}(t) = F(t) - r(t)$ (analogous to eq. (3.5)). In the example considered here the DAE for $w_k = (w_{k1}, w_{k2})^\top$ can be written

$$\begin{aligned} w'_{k1} + [1 - \lambda_k(h)]w_{k1} + w_{k2} &= f_{k1} \\ w_{k1} + [c - \lambda_k(h)]w_{k2} &= f_{k2} \end{aligned}$$

which gives for $w_{k1}(t)$ the simple ODE

$$w'_{k1} + \sigma_k(h)w_{k1} = f_{k1} - \frac{f_{k2}}{c - \lambda_k(h)}$$

where for short $\sigma_k(h) = 1 - \lambda_k(h) + 1/[\lambda_k(h) - c]$. For our example we choose $k = 1$. The stability of w_1 and the numerical method considered here depend on the sign of $\sigma_1(h)$ which is $\sigma_1(h) > 0$ for $h > h_0$ and $\sigma_1(h) < 0$ for $\underline{h}_0 \leq h < h_0$ where $\underline{h}_0$ is some step size near h_0. Furthermore, it is

$$\lim_{h \to h_0} |\sigma_1(h)| = \infty.$$

From these relations we get the result that the numerical method is unstable for $\underline{h}_0 \leq h < h_0$. This is reflected in the last figure for the case $h = 0.025$. A similar phenomenon is considered also in [1], [9].

References

1. M. Arnold, *A note on the uniform perturbation index*, interner Bericht, Universität Rostock, 1995.
2. S.L. Campbell and W. Marszalek, *The index of an infinite dimensional implicit system*, Preprint, 1996.
3. S.L. Campbell, C.D. Meyer Jr., and N.J. Rose, *Applications of the drazin inverse to linear systems of differential equations with singular constant coefficients*, SIAM J. Appl. Math. **31** (1976), no. 3, 411–425.
4. F.R. Gantmacher, *Matrizentheorie*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1986.
5. E. Hairer, Ch. Lubich, and M. Roche, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, vol. 1409, Springer-Verlag, 1989, Lecture Notes in Mathematics.
6. E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II*, Springer-Verlag, 1996.
7. W. Lucht, K. Strehmel and C. Eichler-Liebenow, *Linear partial differential algebraic equations, Part I: Indexes, consistent boundary/initial conditions*, Report, Martin-Luther-Universität Halle, Fachbereich Mathematik und Informatik, 1997, submitted for publication.
8. R. D. Richtmyer and K. W. Morton, *Difference methods for initial–value problems*, Interscience Publ., New York, 1967.
9. G. Söderling, *Remarks on the stability of high–index DAEs with respect to parametric perturbations*, Computing,**49** (1992), 303–314.
10. K. Strehmel and R. Weiner, *Numerik gewöhnlicher Differentialgleichungen*, B.G. Teubner Verlagsgesellschaft, Stuttgart, 1995.
11. J.W. Thomas, *Numerical partial differential equations: Finite difference methods*, Springer, New York, 1995.
12. J.G. Verwer and J.M. Sanz-Serna, *Convergence of method of lines approximations to partial differential equations*, Computing **33** (1984), 297–313.