MODELS OF QUASI–LINEAR PDAEs WITH CONVECTION

W. LUCHT, K. DEBRABANT

Martin-Luther-Universität Halle–Wittenberg
Fachbereich Mathematik und Informatik
Institut für Numerische Mathematik
Postfach, D–06099 Halle, Germany
email: lucht@mathematik.uni-halle.de

Abstract. For a class of partial differential algebraic equations (PDAEs) of quasi–linear type which include nonlinear terms of convection type a possibility to determine a time and spatial index is considered. Then we discuss the numerical solution of initial boundary value problems by means of a corresponding finite difference splitting procedure which is a modification of a well known fractional step method coupled with a matrix factorization. The convergence of the numerical solution towards the exact solution of the corresponding initial boundary value problem is investigated. Some results of a numerical solution of a nonlinear PDAE with convection from plasma physics are given.

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

Keywords: Partial differential algebraic equations, indexes for mixed nonlinear systems, numerical solution of PDAEs

1. Introduction

In this paper quasi–linear PDAEs for $u = u(t,x),\ x = (x_1, ..., x_d)$, of the form

$$Au_t + B\Delta u + \sum_{i=1}^{d} C_i[u]u_{x_i} + Du = f(t,x), \quad t \in (0,t_e),\ x \in \Omega \subset \mathbb{R}^d, \tag{1.1}$$

for some $t_e > 0,\ d = 1,2$ or 3, are considered. Ω is assumed to be an open bounded set whose boundary is sufficiently smooth for $d \geq 2$. If $d = 1$, we choose $\Omega = (0,1)$. u and f are mappings $u, f:\ [0,t_e] \times \bar{\Omega} \to \mathbb{R}^n,\ n \geq 1$, where f (supposed to be sufficiently smooth) is given. $A,\ B,\ C_i[u]$ and D are real $(n,n)-$matrices where $A,\ B$ and D are assumed to be constant. All matrices may be singular, but $A,\ B \neq 0$. $C_i[u]$ may depend on u. We suppose that it is linear in u, and that there is $z \in \mathbb{R}^n$ such that $C_i[z] \neq 0$ for at least one $i \in \{1, ..., d\}$ (i.e. there is at least one spatial derivative of first order of some component of u in (1.1)). Typically, when $C_i[u]$ is linear in u, vector $C_i[u]u_{x_i}$ describes physical convection. Terms of this form appear in many mathematical models, some applications are given in section 2.

For system (1.1) we study classical initial boundary value problems (IBVPs). Initial values (IVs) can be written in the form

$$u(0,x) = \Phi_a(x) + \Phi_c(x), \quad x \in \bar{\Omega} \tag{1.2}$$

with $(k = 1, ..., n)$

$$\Phi_{a,k}(x) = \begin{cases} u_k(0,x) & u_k(0,x) \quad \text{can be prescribed arbitrarily} \\ 0 & \text{otherwise,} \end{cases}$$

$$\Phi_{c,k}(x) = \begin{cases} u_k(0,x) & u_k(0,x) \quad \text{cannot be prescribed arbitrarily} \\ 0 & \text{otherwise.} \end{cases}$$

The boundary values (BVs) are of similar form,

$$u(t,x) = \Psi_a(t,x) + \Psi_c(t,x), \quad t \in [0,t_e], \quad x \in \partial\Omega. \tag{1.3}$$

This means the data which can be prescribed arbitrarily are in Φ_a, Ψ_a. The consistent data (see e.g. [4] or [10]) are collected in Φ_c, Ψ_c. Furthermore, we require that the compatibility relations

$$\Phi_a(x) + \Phi_c(x) = \Psi_a(0,x) + \Psi_c(0,x), \quad x \in \partial\Omega, \tag{1.4}$$

are satisfied. Because almost every PDAE has its own IV and BV distribution we avoid here to give a more detailed general description of these data. In section 5 we solve this problem for the plasma PDAE considered in the next section.

This paper is organized as follows. In section 2 nonlinear PDAEs of some applications from physics and medicine are presented. In section 3.1 known general index concepts applicable also to nonlinear PDAEs are considered. The numerical solution of IBVPs by a finite difference splitting method is studied in section 4. Especially, convergence results are given. A numerical example from plasma physics is presented in section 5.

2. Applications

In this section we cite three mathematical models (from physics and medicine) in which the underlying system of equations is a PDAE of type (1.1).

2.1. **A model from plasma physics.** We cite a simple model from plasma physics, see [7, Chapter 5]. It describes the space–time–movement of a system consisting of ions (with mass m_i, density n_i and positive electrical charge q) and electrons (with density n_e and electrical charge $-q$) in a bounded space domain Ω which is chosen for simplicity as one–dimensional, $d = 1$. The charge $q(n_e - n_i)$ produces an electrical potential ϕ such that the ions move under the corresponding electrical force $-q\phi_x$. The system of electrons is considered to be a gas with (constant) temperature T_e and pressure $p = k_B T_e n_e$ where k_B is Boltzmann's constant. The relation between n_e and ϕ is given by the equilibrium of electrical and mechanical forces,

$$q n_e \phi_x - k_B T_e n_{e,x} = 0,$$

and the equations of conservation of mass and momentum for the ions are

$$n_{i,t} + (n_i v_i)_x - D_i n_{i,xx} = 0,$$
$$m_i \left(\frac{\partial}{\partial t} + v_i \frac{\partial}{\partial x} \right) v_i + q\phi_x = 0$$

where D_i is a (constant) diffusion coefficient, and v_i is the velocity of the ions. The equation for ϕ is Poisson's equation

$$\phi_{xx} - 4\pi(n_e - n_i) = 0.$$

It is convenient to transform this system of equations by a linear transformation into a new system of the form (1.1) with $n = 4,\ f = 0$ and new dependend variables u_i and with matrices

$$A = \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{pmatrix}, \quad B = \begin{pmatrix} -b_0&0&0&0\\ 0&0&0&0\\ 0&0&0&0\\ 0&0&0&1 \end{pmatrix}, \quad C[u] = \begin{pmatrix} u_2&u_1&0&0\\ 0&u_2&0&d_1\\ 0&0&-1&u_3\\ 0&0&0&0 \end{pmatrix}, \quad D = \begin{pmatrix} 0&0&0&0\\ 0&0&0&0\\ 0&0&0&0\\ 1&0&-1&0 \end{pmatrix}. \tag{2.1}$$

Here $C[u] = C_1[u]$ (because $d = 1$). The new variables u_i are (up to a constant) given by $u_1 \sim n_i,\ u_2 \sim v_i,\ u_3 \sim n_e,\ u_4 \sim \phi$. $b_0 \geq 0$ and $d_1 > 0$ are constants.
Note that the matrices $A,\ B$ are diagonal and singular, and $C[u]$ is linear in the components of u (for short we say that $C[u]$ is linear in u).

2.2. **A poroelastic model of a living bone.** On the basis of poroelasticity theory, this model describes the interaction of mechanical bone deformation and bone fluid flow in a fluid–saturated porous medium. A survey is given by S. C. Cowin [5], see also [6]. A deformation of a bone influences the flow of the fluid and vice versa. Bone fluid has many functions. For example, it transports nutrients to, and carries waste from, the bone cells buried in the bony medium.

The poroelastic field variables are the stress σ_{ij} and the strain $\epsilon_{ij} = \frac{1}{2}(u_{i,j}+u_{j,i}),\ i,j,= 1,...,d$, where $u = (u_1,...,u_d)^T$ is the displacement vector of the solid phase, and p is the pore pressure which gives by Darcy's law the fluid mass flow rate $Q = -\kappa\nabla p$. (The superscript T means transposition.) As usual in the theory of porous media, κ is the coefficient of permeability. Using the representation given in [5] the set of equations of the model in two space dimensions is ($d = 2$, and we write $x = x_1,\ y = x_2$)

$$2Gu_{1,x} - \sigma_{11}(1 - \xi_2) + \xi_2\sigma_{22} + \xi_1 T = 0 \tag{2.2}$$

$$2Gu_{2,y} - \sigma_{22}(1 - \xi_2) + \xi_2\sigma_{11} + \xi_1 T = 0 \tag{2.3}$$

$$G(u_{1,y} + u_{2,x}) - \sigma_{12} = 0 \tag{2.4}$$

$$\sigma_{11,x} + \sigma_{12,y} + b_1 = 0 \tag{2.5}$$

$$\sigma_{12,x} + \sigma_{22,y} + b_2 = 0 \tag{2.6}$$

$$T_t - c\Delta T - q\gamma = 0 \tag{2.7}$$

where $T := Tr\{\sigma\} + \frac{3}{b_0}p$ and $\xi_1 := -\frac{\alpha b_0}{3}\frac{1-2\nu}{1+\nu}$, $\xi_2 := \frac{\nu}{1+\nu} - \xi_1$. γ is a source density of the bone fluid. $G,\ \nu$ and α are the shear modulus, Poisson's ratio of the solid material and a parameter, respectively. b_0 (with $0 < b_0 \leq 1$) is the compressibility coefficient which depends on the pore volume fraction (the volume of pores per unit total volume). q and c ($c > 0$) are physical parameters depending on $G,\ b_0$ and ν. (For details see [5].) $b_i,\ i = 1,2$, represents the i^{th} component of the body force.
Defining $U := (T, u_1, u_2, \sigma_{11}, \sigma_{22}, \sigma_{12})^T$ and $F := (q\gamma,\ 0,\ 0,\ 0,\ -b_1,\ -b_2)^T$, this system of

equations is a PDAE of the form (1.1), i.e. $AU_t + B\Delta U + C_1 U_x + C_2 U_y + DU = F$, with

$$A = \begin{pmatrix} 1 & \\ & O \end{pmatrix}, \qquad B = \begin{pmatrix} -c & \\ & O \end{pmatrix}, \tag{2.8}$$

$$C_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2G & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & G & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2G & 0 & 0 & 0 \\ 0 & G & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \xi_1 & 0 & 0 & (\xi_2 - 1) & \xi_2 & 0 \\ \xi_1 & 0 & 0 & \xi_2 & (\xi_2 - 1) & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where O is a $(5,5)-$zero matrix. Again note that A and B are diagonal and singular, and that the convection matrices C_i are constant in the approximation considered here.

2.3. **Incompressible Navier–Stokes equations.** As a third example we mention the system of Navier–Stokes equations of an incompressible medium for $d = 3$, i.e.

$$U_t + (U \cdot \nabla)U + \nabla p - \nu \Delta U = F, \qquad \nabla \cdot U = 0.$$

$U = (U_1, U_2, U_3)^T$ and p are the velocity vector and the pressure of the fluid, respectively. ν is the kinematic viscosity, and F represents the body force. If we introduce new vectors $f = (f_1, f_2, f_3, f_4)^T$ and $u = (u_1, u_2, u_3, u_4)^T$ with $f_i = F_i,\ u_i = U_i,\ i = 1,2,3,$ and $f_4 = 0,\ u_4 = p$, the system can be written as PDAE of type (1.1) with matrices

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -\nu & 0 & 0 & 0 \\ 0 & -\nu & 0 & 0 \\ 0 & 0 & -\nu & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad D = O, \tag{2.9}$$

$$C_1[u] = \begin{pmatrix} u_1 & 0 & 0 & 1 \\ 0 & u_1 & 0 & 0 \\ 0 & 0 & u_1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad C_2[u] = \begin{pmatrix} u_2 & 0 & 0 & 0 \\ 0 & u_2 & 0 & 1 \\ 0 & 0 & u_2 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad C_3[u] = \begin{pmatrix} u_3 & 0 & 0 & 0 \\ 0 & u_3 & 0 & 0 \\ 0 & 0 & u_3 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Again A and B are singular and diagonal matrices, and $C_i[u]$ is linear in u.

3. Indexes

In order to explain the method for defining indexes of the PDAE we introduce first two definitions of classical linear spaces. For simplicity, here we choose $d = 1,\ \Omega = (0,1)$. The extension to higher spatial dimensions is straightforward. Let $l,\ m,\ p,\ q$ be nonnegative integers, and let $C^{l,m}_{sc}$ be the space of scalar real functions $w(t,x),\ t \in [0, t_e],\ x \in \overline{\Omega}$, whose derivatives up to the $(l+m)^{th}$ order (time derivatives up to the l^{th} order and space derivatives up to the m^{th} order) are continuous. With $C^{l,m}_{sc}$ the set

$$C^{l,m}_n := \left\{ w = (w_1, ..., w_n)^T;\ w_i \in C^{l,m}_{sc},\ i = 1, ..., n \right\} \tag{3.1}$$

is defined. By $C^{l,m}_{n,0}$ we denote a set of vector valued functions with vanishing BVs,

$$C^{l,m}_{n,0} := \Big\{ w \in C^{l,m}_n; \quad w(t,0) = w(t,1) = 0 \Big\}. \tag{3.2}$$

The determination of indexes for nonlinear PDAEs is more cumbersome than for linear ones. While the index of linear PDAEs has been considered by several authors, e.g. [2], [3], [4], [8], [10], [12], the index for nonlinear systems is investigated little, see however [11] where a discretization based index definition has been given, and [13] where the index problem of a first order PDAE is reduced to the investigation of initial value problems. For these, indexes can be defined. We mention that we do not transform system (1.1) to a first order system because we prefer in numerial calculations the original second order form.

In this paper, a well known index concept for PDAEs is used to construct certain operators whose invertibility (if so) yields indexes. The basic notion of a time index ν_t and a spatial index ν_x can be found, e.g., in [2], [10], [13].

In this paper the time index ν_t of (1.1) is of special interest. Here, we define ν_t for one space dimension, i.e. $d = 1$. The definition in higher dimensions is analogous. We assume that a solution u of the IBVP (1.1) – (1.4) exists and that u is sufficiently differentiable. Furthermore, for $\Omega = (0, 1)$ we suppose that $u(t, x) = 0$ for $x \in \partial\Omega$. When Ψ_a and Ψ_c (see eq. (1.3)) are known, zero BVs can be obtained by a suitable transformation of u.

3.1. **Time index.**

Definition 3.1. *If the matrix A is regular, the time index ν_t of the PDAE (1.1) is defined to be zero. If A is singular, then ν_t is the smallest number of times the PDAE must be differentiated with respect to t in order to determine u_t as a continuous function of t, x, u and certain space derivatives of $u-$components.* □

Since in most practical applications of PDAEs the time index is $\nu_t = 1$ or $\nu_t = 2$, we give here the formalism for these two indexes for PDAEs of type (1.1) only. First, an auxiliary result is stated. To simplify the notation we use in the following the summation convention (summation about twofold indices from 1 to n).

Lemma 3.1. *Let $d = 1$ and u be sufficiently smooth, and suppose that $C[u]$ is linear in u. Then*

(I) $\partial_t C[u]u_x = C^{(1)}[u_x]u_t + C[u]u_{tx}$ *where $C^{(1)}_{ik}[u_x]$ is defined in (3.3),*
(II) $\partial_t C^{(1)}[u_x]u_t = C^{(1)}[u_{tx}]u_t + C^{(1)}[u_x]u_{tt}$. □

Proof. Let $C_{ij,k} := \frac{\partial C_{ij}}{\partial u_k}$, $i, j, k \in J_n$. Componentwise differentiation with respect to time yields

$$\partial_t C_{ij} u_{j,x} = C_{ij,k} u_{k,t} u_{j,x} + C_{ij} u_{j,tx} = (C_{ij,k} u_{j,x}) u_{k,t} + C_{ij} u_{j,tx}.$$

Defining

$$C^{(1)}_{ik}[u_x] := C_{ij,k} u_{j,x} \tag{3.3}$$

we get immediately statement (I). Eq. (3.3) and the linearity of C in u imply

$$\begin{aligned}\partial_t C^{(1)}_{ik}[u_x]u_{k,t} =& \partial_t C_{ij,k} u_{j,x} u_{k,t} = C_{ij,k} u_{j,tx} u_{k,t} + C_{ij,k} u_{j,x} u_{k,tt} \\ =& C^{(1)}_{ik}[u_{tx}]u_{k,t} + C^{(1)}_{ik}[u_x]u_{k,tt}\end{aligned}$$

which is (II) in component form. □

For example, $C[u]$ given in eq. (2.1) produces $C^{(1)}[u_x] = \begin{pmatrix} u_{2,x} & u_{1,x} & 0 & 0 \\ 0 & u_{2,x} & 0 & 0 \\ 0 & 0 & u_{4,x} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$.

With this Lemma and with definitions

$$L := B\partial_x^2 + D, \qquad M[u, u_x] := L + C^{(1)}[u_x] + C[u]\partial_x \tag{3.4}$$

we find formally from eq. (1.1) under the assumptions of Lemma 3.1 after one time differentiation (this case is relevant for $\nu_t = 1$) the derivative array

$$\begin{pmatrix} A & 0 \\ M[u, u_x] & A \end{pmatrix} \begin{pmatrix} u_t \\ u_{tt} \end{pmatrix} + \begin{pmatrix} Lu + C[u]u_x \\ 0 \end{pmatrix} = \begin{pmatrix} f \\ f_t \end{pmatrix}. \tag{3.5}$$

Two time differentiations of (1.1) produce the derivative array (relevant for $\nu_t = 2$)

$$\begin{pmatrix} A & 0 & 0 \\ M[u, u_x] & A & 0 \\ 2C^{(1)}[u_{tx}] & M[u, u_x] & A \end{pmatrix} \begin{pmatrix} u_t \\ u_{tt} \\ u_{ttt} \end{pmatrix} + \begin{pmatrix} Lu + C[u]u_x \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} f \\ f_t \\ f_{tt} \end{pmatrix}. \tag{3.6}$$

By means of the PDAE, the mixed derivative u_{tx} in the system matrix of this equation should be expressed in terms of u and $x-$derivatives of u. For example, assume that A (supposed to be singular) can be written in diagonal form, $A = \left(\begin{smallmatrix} I_k & 0 \\ 0 & 0 \end{smallmatrix}\right)$ where I_k is a unit matrix of order $k < n$. Then the first k mixed derivatives $u_{j,tx}$, $j = 1, ..., k$, can be expressed immediately by means of the equation $Au_{tx} + Lu_x + \partial_x C[u]u_x = f_x$ which results from differentiation of the PDAE with respect to x.

When A is singular, the coefficient matrices (denoted by $\mathcal{A}$) of the systems (3.5), (3.6) are singular in the sense that $\mathcal{A}v = 0$ where $v \neq 0$ is an appropriate vector function (e.g. in the case of eq. (3.5), $v \in C^{0,2}_{2n}$). However, the coefficient matrices might uniquely determine u_t. The analogous problem for linear time varying DAEs (differential algebraic equations) is discussed in [1, p. 29], and for linear PDAEs of first order it is investigated in [13].
For the nonlinear PDAE (1.1), the derivative array (3.5) or (3.6) is used to write

$$Pu_t = F \tag{3.7}$$

where P is according to eq. (3.5) or (3.6) an operator valued matrix, and the vector F does not depend on u_t. First, we must find P and an appropriate domain of definition $D(P)$. Second, the invertibility of P must be studied. This means, the smallest number of time differentiations of the PDAE is needed in order to get P^{-1}.

3.2. **Spatial index.** If B is singular, we suppose that the PDAE is written in quasilinear form (with respect to second space derivatives), i.e. we transform B according to

$$\overline{B} = S_0 B S_1^{-1} = \left(\begin{smallmatrix} I_{n_1} & 0 \\ 0 & 0 \end{smallmatrix}\right) \tag{3.8}$$

where S_0, S_1 are constant regular $(n, n)-$matrices.

Definition 3.2. *The spatial index ν_x of a system with a regular matrix B is defined to be zero. When B is singular, ν_x is the smallest number of times the PDAE*

$$\overline{A}\overline{u}_t + \overline{B}\overline{u}_{xx} + \overline{C}(S_1^{-1}\overline{u})\overline{u}_x + \overline{D}\overline{u} = \overline{f}(t, x) \tag{3.9}$$

($\overline{u} = S_1 u$, $\overline{A} = S_0 A S_1^{-1}$, and so on) must be differentiated with respect to x in order to obtain

$$\overline{U} := (\underbrace{\overline{u}_{1,xx}, \overline{u}_{2,xx}, ..., \overline{u}_{n_1,xx}}_{n_1 \ elements}, \underbrace{\overline{u}_{n_1+1,x}, ..., \overline{u}_{n,x}}_{n-n_1 \ elements})^T \tag{3.10}$$

as a continuous function of t, x, $\overline{u}$, $\overline{u}_t$ *and certain (lower)* $x-$*derivatives of* $\overline{u}-$*components not present in* $\overline{U}$. $\square$

It is straightforward to derive for $\overline{U}$ a derivative array (by differentiations of the PDAE with respect to x) analogous to eq. (3.5) or (3.6). As a result, after a minimal number of $x-$differentiations one can find a representation of $\overline{U}$ of the form

$$\overline{Q}\,\overline{U} = \overline{G} \tag{3.11}$$

where $\overline{Q}$ is an operator valued matrix. The vector $\overline{G}$ is independent of the components of $\overline{U}$. This definition of the spatial index is according to the one given in [10]. It does not transform the PDAE (1.1) to a system of first order (as is often done).

3.3. **Time index of the plasma PDAE.** In general, one should study the time index ν_t of the PDAE (1.1) for $d = 1$ with zero BVs, $u(t,0) = u(t,1) = 0$, and $f \in C_4^{p,q}$ with suitable $p,\ q$. We assume that this IBVP has a solution $u \in C_{4,0}^{1,2}$ which possibly exists only local in time. The aim is to determine under this assumption the time index of system (1.1) with $C[u] \neq 0$. The idea is to generate for ν_t – when possible – system (3.7) with the condition that P defined on $D(P)$ is invertible. Here, we do not consider a general PDAE, but we develop the method by means of the plasma PDAE (see section 2), i.e. $n = 4$, $f = 0$ and eq. (1.1) with matrices (2.1).

In order to determine ν_t we differentiate the third and fourth equation of the PDAE with respect to t with the result $-u_{3,tx} + u_{3,t}u_{4,x} + u_3 u_{4,tx} = 0$ and $u_{4,txx} + u_{1,t} - u_{3,t} = 0$. A new system can be formed by means of these two equations and the first two equations of the original system. This is a closed system of four equations for the components of u_t in terms of u and derivatives which are not time derivatives of $u-$components. Obviously, the new system can be written

$$Pu_t = F(u) := \begin{pmatrix} -u_2u_{1,x} - u_1u_{2,x} + b_0u_{1,xx} \\ -u_2u_{2,x} - d_1u_{4,x} \\ 0 \\ 0 \end{pmatrix}, \tag{3.12}$$

$$Pu_t = P_u u_t \quad \text{where} \quad P_u := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & [-\partial_x + u_{4,x}] & u_3\partial_x \\ 1 & 0 & -1 & \partial_x^2 \end{pmatrix}. \tag{3.13}$$

Eq. (3.12) is of type (3.7). Now what is essential for the determination of the index of the PDAE is that u in P_u can be seen as a fixed element with the result that $P_u:\ D(P_u) \to R(P_u)$ is a linear operator ($R(P_u)$ denotes the range of P_u). Incorporating the time derivative into a new linear operator, say $\tilde{P}_u := P_u\partial_t$, the equation can be written also

$$\tilde{P}_u u = F(u), \qquad D(\tilde{P}_u) \subset C_{4,0}^{1,3} \subset C_{4,0}^{1,2}. \tag{3.14}$$

We can try to solve eq. (3.12) for u_t by standard linear theory. In particular, if P_u^{-1} exists we get ν_t as the least number of time differentiations which were necessary to derive equation (3.12) (in the example considered here, one differentiation with respect to t was needed).

Remark 3.1. Note that the right hand side of equation (3.12) does not play any role when ν_t is to be determined only. This means that ν_t is independent of the function f in eq. (1.1). This is a reasonable result because it is well known from linear PDAE theory that the right hand side function does not influence the indexes of a PDAE.

We ask whether the linear operator P_u with $D(P_u) \subset C^{0,3}_{4,0}$ has for fixed $u = u(t,x)$ an inverse. The answer comes from kernel $N(P_u)$ whose elements z fulfill

$$P_u z = \begin{pmatrix} z_1 \\ z_2 \\ -z_{3,x}+u_{4,x}z_3+u_3z_{4,x} \\ z_1-z_3+z_{4,xx} \end{pmatrix} = 0. \tag{3.15}$$

This reduces to $z_1 = z_2 = 0$ and a linear homogeneous coupled system of two ordinary differential equations for z_3 and z_4, $-z_{3,x} + u_{4,x}z_3 + u_3z_{4,x} = 0$ and $-z_3 + z_{4,xx} = 0$, with homogeneous BVs, $z_3(0) = 0$ and $z_4(0) = z_4(1) = 0$. z_4 is also a solution of the equation $z_{4,xxx} + u_{4,x}z_{4,xx} + u_3z_{4,x} = 0$ which can be reduced to $y_{xx} + u_{4,x}y_x + u_3y = 0$ where $y = y(t,x) := z_{4,x}(t,x)$. By φ_1 and φ_2 we denote two fundamental solutions of the equation for y, i.e. $y = K_1\varphi_1 + K_2\varphi_2$ is the general solution ($K_i,\ i = 1,2$, are independent of x). Then the following Lemma holds:

Lemma 3.2. *Let u be such that*

1. $\varphi_{2,x}(0) \neq 0,$
2. $\begin{vmatrix} \varphi_{1,x}(0) & \varphi_{2,x}(0) \\ \int_0^1 \varphi_1(\xi)\,d\xi & \int_0^1 \varphi_2(\xi)\,d\xi \end{vmatrix} \neq 0.$

Then $z_3 = z_4 = 0.$ □

Proof. The general solution y implies

$$z_4(x) = K_1 \int_0^x \varphi_1(\xi)\,d\xi - K_2 \int_x^1 \varphi_2(\xi)\,d\xi + K_3 \tag{3.16}$$

(K_3 independent of x). The first assumption and $z_3(0) = z_{4,xx}(0) = 0$ give $K_2 = -K_1\varphi_{1,x}(0)/\varphi_{2,x}(0)$. This combined with (3.16) yields for the first boundary condition of z_4

$$0 = z_4(0) = K_1 \frac{\varphi_{1,x}(0)}{\varphi_{2,x}(0)} \int_0^1 \varphi_2(\xi)\,d\xi + K_3.$$

The second boundary condition of z_4 gives by means of (3.16)

$$0 = z_4(1) = K_1 \int_0^1 \varphi_1(\xi)\,d\xi + K_3.$$

These are two linear homogeneous equations for $K_1,\ K_3$. By the second assumption, the coefficient determinant is different from zero. Therefore, $K_1 = K_3 = 0$. Hence, also $K_2 = 0$, and $z_4 = 0$. Because $-z_3 + z_{4,xx} = 0$ it follows $z_3 = 0$.
Note that (with a little change of the proof) the first assumption can be replaced by $\varphi_{1,x}(0) \neq 0$ as well. □
From this Lemma we see that $N(P_u) = \{0\}$. Therefore, system (3.12) can be solved for u_t, and the time index is $\nu_t = 1$.
Remark 3.2. We mention that based on Definition 3.2 the spatial index of the plasma PDAE can be determined similarly. The result is $\nu_x = 0$. □

4. Numerical solution by a finite difference method

In this section we consider the numerical solution of IBVPs (1.1) – (1.4) by means of a fractional step difference method which is combined with a matrix factorization. The fractional step method and numerous variants of it are well known [9]. By this method, the order of the system of equations which must be solved can be reduced considerably. The effort can be reduced further by a proper partition of the original system matrix (denoted by L below) into two splitting matrices ($L = L_1 + L_2$).

To describe the general procedure, we first rewrite the nonlinear part $C[u]u_x$ of the PDAE (1.1) as $C[u, \partial_x]u$ which is more convenient sometimes. $C[u, \partial_x]$ is an operator valued matrix. For example, with the matrix $C[u]$ given in (2.1) it is

$$C[u]u_x = \begin{pmatrix} u_2 & u_1 & 0 & 0 \\ 0 & u_2 & 0 & d_1 \\ 0 & 0 & -1 & u_3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_{1,x} \\ u_{2,x} \\ u_{3,x} \\ u_{4,x} \end{pmatrix} = \begin{pmatrix} \partial_x u_2 & 0 & 0 & 0 \\ 0 & u_2\partial_x & 0 & d_1\partial_x \\ 0 & 0 & -\partial_x & u_3\partial_x \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} =: C[u, \partial_x]u. \tag{4.1}$$

Note that there may be different representations of $C[u, \partial_x]$.

We suppose that the matrix A is singular and is given in the form

$$A = \begin{pmatrix} I_{n_1} & 0 \\ 0 & 0 \end{pmatrix} \tag{4.2}$$

where I_{n_1} is the unit matrix of order $n_1 < n$, $(n_1 \geq 1)$. Let $n_2 = n - n_1$. Corresponding to this partition of A we introduce the notation $u = (u_1,\ u_2)^T$ and $M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$ where the $(n, n)-$matrix M may be one of the matrices $B,\ C,\ D$. The component representation of u_k is $u_k = (u_{k1}^T, ..., u_{kn_k}^T)^T$, $k = 1, 2$. According to this partition, PDAE (1.1) is written

$$\begin{pmatrix} I_1 & 0 \\ 0 & 0 \end{pmatrix} u_t + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} u_{xx} + \begin{pmatrix} C_{11}[u, \partial_x] & C_{12}[u, \partial_x] \\ C_{21}[u, \partial_x] & C_{22}[u, \partial_x] \end{pmatrix} u + \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} u = f. \tag{4.3}$$

Furthermore, let the operator valued matrices $L_k[u]$, $k = 1, 2$, be defined by ($C_{kj} = C_{kj}[u, \partial_x]$)

$$L_1[u] := \begin{pmatrix} 0 & 0 \\ (B_{21}\partial_x^2 + C_{21} + D_{21}) & (B_{22}\partial_x^2 + C_{22} + D_{22}) \end{pmatrix}, \tag{4.4}$$

$$L_2[u] := \begin{pmatrix} (B_{11}\partial_x^2 + C_{11} + D_{11}) & (B_{21}\partial_x^2 + C_{21} + D_{21}) \\ 0 & 0 \end{pmatrix} \tag{4.5}$$

such that $L[u] \ := \ B\partial_x^2 + C[u, \partial_x] + D \ = \ L_1[u] + L_2[u]$.

Remark 4.1. Sometimes, other operators L_1, L_2 with $L = L_1 + L_2$ may be more convenient because the solution of the equations may be easier. In every case, the partition should be such that the identity $AL_2 = L_2$ does hold. This is needed in a factorization as explained below in eq. (4.11). $\square$

In order to obtain an approximate numerical solution of IBVP (1.1) – (1.4) we consider it with zero BVs (1.3) (these BVs can be achieved by a suitable transformation) by means of a difference method. The first time derivative is approximated with an equidistant time step size τ by

$$u_t(t_{m+1}, x_k) \approx \frac{1}{\tau}(u^{m+1}(x_k) - u^m(x_k)), \quad t_{m+1} = (m+1)\tau, \quad m = 0, 1, ...$$

The corresponding IBVP is space discretized on an equidistant grid

$$\Omega_h := \{x_k = kh,\ k = 0, 1, ..., M;\ h = 1/M,\ M > 1\} \tag{4.6}$$

by using difference formulas ($k = 1, ..., M-1$)

$$u_{xx}(t, x_k) \approx \frac{\delta^2}{h^2} u_k(t) := \frac{1}{h^2}\big(u_{k-1}(t) - 2u_k(t) + u_{k+1}(t)\big), \tag{4.7}$$

$$u_x(t, x_k) \approx \frac{\delta_0}{2h} u_k(t) \quad \text{or} \quad \frac{\delta_+}{h} u_k(t) \quad \text{or} \quad \frac{\delta_-}{h} u_k(t) \tag{4.8}$$

where δ_0, δ_+ and δ_- are the usual central, forward and backward difference operators, respectively. Now suppose $t \in (0, t_e)$ and $d = 1$. We approximate eq. (4.3) by the difference equation

$$A\frac{u_k^{m+1} - u_k^m}{\tau} + L_h[u_k^m]\left(\gamma u_k^{m+1} + (1-\gamma)u_k^m\right) = f_k^{m+1}, \quad m = 0, 1, ...,\ k = 1, ..., M-1. \tag{4.9}$$

γ is a real parameter, e.g. $\gamma \in (0, 1]$, and $L_h[u_k^m]$ denotes a discretization of $L[u(t_m, x_k)]$. We rewrite this equation as

$$(A + \tau\gamma L_h[u_k^m])\,\frac{u_k^{m+1} - u_k^m}{\tau} = f_k^{m+1} - L_h[u_k^m]u_k^m \tag{4.10}$$

and factorize approximately for $\tau \to 0$

$$\begin{aligned} A + \tau\gamma L_h[u_k^m] &\approx (A + \tau\gamma L_{h1}[u_k^m])(I + \tau\gamma L_{h2}[u_k^m]) \qquad (4.11)\\ &= A + \tau\gamma(L_{h1}[u_k^m] + AL_{h2}[u_k^m]) + O(\tau^2) = A + \tau\gamma L_h[u_k^m] + O(\tau^2), \end{aligned}$$

where L_{h1} and L_{h2} are corresponding discretizations of (4.4) and (4.5) respectively (we used $AL_{h2} = L_{h2}$ and $L_{h1} + L_{h2} = L_h$). Therefore, we consider for $m = 0, 1, ...,\quad k = 1, ..., M-1$ fractional splitting

$$(A + \tau\gamma L_{h1}[u_k^m])\, u_k^{m+1/2} = f_k^{m+1} - L_h[u_k^m]u_k^m, \tag{4.12}$$

$$(I + \tau\gamma L_{h2}[u_k^m])\,\frac{u_k^{m+1} - u_k^m}{\tau} = u_k^{m+1/2}. \tag{4.13}$$

u_k^0 is (for every k) given as IV.

Remark 4.2. This discussion shows that the fractional step method requires the solution of two linear systems of coupled equations, but each of them is, in numerous applications, of a considerably reduced order. □

The approximate factorization (4.11) implies the following Lemma.

Lemma 4.1. *Suppose*

1. $v = v(t, x) \in \mathbb{R}^n$, $t \in (0, t_e)$, *is for some* $t_e > 0$ *the exact solution (sufficiently smooth) of the IBVP (4.3), (1.2) – (1.4),*
2. *the system of algebraic equations (4.12), (4.13) has an unique solution.*

Then the method (4.12), (4.13) and the system (4.10) are equivalent for $\tau \to 0$, i.e. eqs. (4.12), (4.13) approximate the original system to the same $\tau-$order as eq. (4.10). $\square$

4.1. **Convergence of the fractional step method.** We consider the convergence of the numerical solution u_k^m calculated by the scheme (4.9) for $\gamma = 1$ towards the exact solution $v(t_m, x_k)$ of the corresponding IBVP for $\tau \to 0$ and $h \to 0$ under the condition $m\tau = t$. The basic assumption is that $C[v] = C^0 + C^1[v]$ is linear in v where the matrix C^0 is constant and is chosen in such a manner that the matrices $G_{0k},\ k = 1, ..., M-1$, defined below in eq. (4.39) are regular. For example, a proper choice may be

$$C^0 = C[\bar{v}], \qquad C^1[v] = C[v] - C[\bar{v}] \tag{4.14}$$

where vector $\bar{v}$ does not depend on t and x, e.g. $\bar{v}$ is a suitable mean value of v.

First, we need the full truncation error α_k^{m+1} defined for $m = 0, 1, ...,\ k = 1, ..., M-1$ by

$$\begin{aligned}\alpha_k^{m+1} :=& A\frac{v_k^{m+1} - v_k^m}{\tau} + L_h[v_k^m]v_k^{m+1} - f_k^{m+1}, \\ L[v_k^m] :=& B\frac{\delta^2}{h^2} + C[v_k^m]\frac{\delta}{qh} + D\end{aligned} \tag{4.15}$$

where $q = 1$ (for one–sided differences) or $q = 2$ (for central differences), and δ stands for δ_0 or δ_+ or δ_-. Under the assumption that v is sufficiently smooth we Taylor expand $v_{k\pm1}^m$ in t_{m+1} and x_k to get in lowest order with respect to τ and h

$$\begin{aligned}\alpha_k^{m+1} =& A(v_{k,t}^{m+1} - \frac{1}{2}v_{k,tt}^{m+1}\tau) + B(v_{k,xx}^{m+1} + \frac{1}{12}v_{k,xxxx}^{m+1}h^2) \\ &+ C[v_k^{m+1} - v_{k,t}^{m+1}\tau](v_{k,x}^{m+1} + \bar{O}^{m+1}(h^q)) + Dv_k^{m+1} - f_k^{m+1}.\end{aligned}$$

Since $C[v_k^{m+1} - v_{k,t}^{m+1}\tau] = C^0 + C^1[v_k^{m+1}] - \tau C^1[v_{k,t}^{m+1}]$ and $Av_{k,t}^{m+1} + Bv_{k,xx}^{m+1} + \big(C^0 + C^1[v_k^{m+1}]\big)v_{k,x}^{m+1} + Dv_k^{m+1} = f_k^{m+1}$, the truncation error can be written

$$\begin{aligned}\alpha_k^{m+1} =& A\bar{O}^{m+1}(\tau) + B\bar{O}^{m+1}(h^2) + C[v_k^{m+1}]\bar{O}^{m+1}(h^q) \\ &+ C^1[v_{k,t}^{m+1}]\bar{O}^{m+1}(\tau h^q) + C^1[v_{k,t}^{m+1}]\bar{O}^{m+1}(\tau).\end{aligned} \tag{4.16}$$

Here e.g., $\bar{O}^{m+1}(\tau)$ represents a vector in $\mathbb{R}^n$ (independent of τ) which has a factor τ.
Second, scheme (4.9) (with $\gamma = 1$) is written by means of the Kronecker product in matrix form,

$$\Big(I_{M-1} \otimes \frac{A}{\tau} + Q_h[U^m]\Big)U^{m+1} = \Big(I_{M-1} \otimes \frac{A}{\tau}\Big)U^m + F^{m+1}, \tag{4.17}$$

$$Q_h[U^m] := \frac{1}{h^2}P \otimes B + \frac{1}{qh}\tilde{P}_q \otimes C[U^m] + I_{M-1} \otimes D \tag{4.18}$$

where $U^m := (u_1^{mT}, ..., u_{M-1}^{mT})^T$. P is defined by

$$P := \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & & \ddots & \\ & & 1 & -2 \end{pmatrix}. \tag{4.19}$$

q and $\tilde{P}_q$ depend on the discretization of the first space derivative. For example, using central difference formula in (4.8), it is $q=2$ and $\tilde{P} := \begin{pmatrix} 0 & -1 & & \\ 1 & 0 & -1 & \\ & & \ddots & \\ & & 1 & 0 \end{pmatrix}$ (another differencing is chosen in (4.34)). The $(n(M-1), n(M-1))-$matrix $\tilde{P}_q \otimes C[U^m]$ is a block matrix whose block at position (j,k) is $\tilde{P}_{q,jk} C[u_j],\ u_j \in \mathbb{R}^n$.

Eqs. (4.15) and (4.16) imply that V^{m+1} satisfies equation

$$\Big(I_{M-1} \otimes \frac{A}{\tau} + Q_h[V^m]\Big) V^{m+1} = \Big(I_{M-1} \otimes \frac{A}{\tau}\Big) V^m + F^{m+1} + \mathcal{A}^{m+1}, \tag{4.20}$$

$$\mathcal{A}^{m+1} := E_1 O^{m+1}(\tau) + E_2 O^{m+1}(h^2) + E_3 O^{m+1}(h^q) + E_4 O^{m+1}(\tau h^q) + E_5 O^{m+1}(\tau) \tag{4.21}$$

where

$$\begin{aligned} E_1 :=& I_{M-1} \otimes A, \quad E_2 := I_{M-1} \otimes B, \quad E_3 := I_{M-1} \otimes C[V^{j+1}], \\ E_4 :=& I_{M-1} \otimes C^1[V_t^{j+1}], \quad E_5 := I_{M-1} \otimes C^1[V_t^{j+1}], \end{aligned} \tag{4.22}$$

and, e.g., $O^{m+1}(\tau)$ means $O^{m+1}(\tau) \in \mathbb{R}^{n(M-1)}$. The foregoing relations can be used to estimate a norm of the global error $\eta^m := V^m - U^m$. In the following we choose the discrete L_2-norm defined for a vector $v = (v_1, ..., v_{n(M-1)})^T$ by $||v|| := \left[h \sum_{k=1}^{n(M-1)} v_k^2\right]^{1/2}$.

Subtracting eq. (4.17) from eq. (4.20), we obtain

$$\left(I_{M-1} \otimes \frac{A}{\tau}\right) \eta^{m+1} + Q_h[V^m] V^{m+1} - Q_h[U^m] U^{m+1} = \left(I_{M-1} \otimes \frac{A}{\tau}\right) \eta^m + \mathcal{A}^{m+1}.$$

By the identity

$$Q_h[V^m] V^{m+1} - Q_h[U^m] U^{m+1} = \big(Q_h[V^m] - Q_h[U^m]\big) V^{m+1} + Q_h[U^m] \big(V^{m+1} - U^{m+1}\big),$$

the foregoing equation can be written

$$G^m \eta^{m+1} = \left(I_{M-1} \otimes \frac{A}{\tau}\right) \eta^m - \left(Q_h[V^m] - Q_h[U^m]\right) V^{m+1} + \mathcal{A}^{m+1}, \tag{4.23}$$

$$G^m := I_{M-1} \otimes \frac{A}{\tau} + \frac{1}{h^2} P \otimes B + \frac{1}{qh} \tilde{P}_q \otimes C[U^m] + I_{M-1} \otimes D. \tag{4.24}$$

Now we use the fact that $Q_h[V^m] - Q_h[U^m] = \frac{1}{qh} \tilde{P}_q \otimes (C[V^m] - C[U^m]) = \frac{1}{qh} \tilde{P}_q \otimes C^1[\eta^m]$ (because the linearity of $C[u]$ implies $C[V^m] - C[U^m] = C^1[\eta^m]$). Furthermore, one can construct a $(n(M-1), n(M-1))-$matrix $\tilde{C}[V^{m+1}]$ such that $\left(\tilde{P}_q \otimes C^1[\eta^m]\right) V^{m+1} = \tilde{C}[V^{m+1}] \eta^m$. Therefore, eq. (4.23) takes the form ($G^{-m} = (G^m)^{-1}$)

$$\eta^{m+1} = G^{-m} \left(I_{M-1} \otimes \frac{A}{\tau} - \frac{1}{qh} \tilde{C}[V^{m+1}]\right) \eta^m + G^{-m} \mathcal{A}^{m+1} \tag{4.25}$$

or for short $\eta^{m+1} = H^m \eta^m + R^{m+1}, \quad H^m := G^{-m} \left(I_{M-1} \otimes \frac{A}{\tau} - \frac{1}{qh} \tilde{C}[V^{m+1}]\right)$,
$R^{m+1} := G^{-m} \mathcal{A}^{m+1}$. It follows for $m = 0, 1, ...$

$$\eta^{m+1} = H^m H^{m-1} ... H^0 \eta^0 + H^m H^{m-1} ... H^1 R^1 + H^m H^{m-1} ... H^2 R^2 + ... + H^m R^m + R^{m+1},$$

$$||\eta^{m+1}|| \leq ||H^m H^{m-1} ... H^0|| \; ||\eta^0|| + ||H^m H^{m-1} ... H^1|| \; ||R^1|| + ||H^m H^{m-1} ... H^2|| \; ||R^2|| + ... + ||H^m|| \; ||R^m|| + ||R^{m+1}||. \tag{4.26}$$

Now we require for $0 < m\tau = t \in (0, t_e]$ that

$$\sup_{m \in \mathbb{N}} \left\{ ||H^m H^{m-1} ... H^j||, \; j = 1, ..., m \right\} \; < \; \infty, \tag{4.27}$$

possibly under a restriction of one of the forms

$$\kappa_0 \leq \frac{\tau}{h} \leq \kappa_1, \qquad \kappa_2 \leq \frac{\tau}{h^2} \leq \kappa_3. \tag{4.28}$$

The basic assumption (4.27) is discussed shortly in the first item of Remark 4.3. With this condition, the right hand side in (4.26) can be estimated further to give

$$\begin{aligned} ||\eta^{m+1}|| &\leq K_0 ||\eta^0|| + \bar{K}_1 \sum_{j=0}^{m} ||R^{j+1}|| \leq K_0 ||\eta^0|| + (m+1) \bar{K}_1 \max_{j \in [0,m]} ||R^{j+1}|| \\ &\leq K_0 ||\eta^0|| + K_1 \frac{t}{\tau} \max_{j \in [0,m]} ||R^{j+1}|| \end{aligned} \tag{4.29}$$

where $K_0, \; K_1, \; \bar{K}_1$ are constants. Using $||R^{j+1}|| = ||G^{-j} \mathcal{A}^{j+1}||$, relation (4.21) yields

$$||R^{j+1}|| \leq ||G^{-j} E_1|| O(\tau) + ||G^{-j} E_2|| O(h^2) + ||G^{-j} E_3|| O(h^q) + ||G^{-j} E_4|| O(\tau h^q) + ||G^{-j} E_5|| O(\tau) \tag{4.30}$$

or also

$$||R^{j+1}|| \leq ||G^{-j}|| \Big(||E_1|| O(\tau) + ||E_2|| O(h^2) + ||E_3|| O(h^q) + ||E_4|| O(\tau h^q + ||E_5|| O(\tau)) \Big). \tag{4.31}$$

Note that ineq. (4.30) is sharper than ineq. (4.31). Since the matrices E_l are block diagonal, their norms can be estimated easily. For example, E_3 is of the form
$E_3 = diag\{C[v_1^{j+1}], ..., C[v_{M-1}^{j+1}]\}$, and

$$||E_3|| = \left[\max_k \left[\lambda_{max}(C^T[v_k^{j+1}] C[v_k^{j+1}]) \right] \right]^{1/2}$$

where $\lambda_{max}(E)$ is the largest eigenvalue of E. Especially, if v is sufficiently smooth for some time $t \leq t_e$, then all norms $||E_l||$ are bounded. Therefore, (4.31) can be written also

$$||R^{j+1}|| \leq K_3 ||G^{-j}|| \left(O(\tau) + O(h^2) + O(h^q) + O(\tau h^q) \right) \tag{4.32}$$

with some constant K_3.
An estimate of $||G^{-j}||$ can be based on the eigenvectors ϕ_k of the symmetric matrix $\frac{1}{h^2} P$. By definition, $\frac{1}{h^2} P \phi_k = \lambda_k \phi_k, \; k = 1, ..., M-1$. The eigenvalues are given by $\lambda_k = -\frac{4}{h^2} \sin^2 \left(\frac{k\pi}{2M} \right)$. They fulfill for $h \to 0$ the asymptotic relation $\lambda_k = -(k\pi)^2 + O(h^2)$ (because $h = 1/M$). The set $\{\phi_k, \; k = 1, ..., M-1\}$ is assumed to be orthonormal. Let $\Phi := \sqrt{h}(\phi_1, ..., \phi_{M-1})$ be the matrix of the eigenvectors and $\Lambda := diag(\lambda_1, ..., \lambda_{M-1})$. Since $\Phi^{-1} = \Phi^T = \Phi$, it is easy to

show that

$$
\begin{aligned}
\frac{1}{h^2} P \otimes B =& (\Phi \otimes I_n)(\Lambda \otimes B)(\Phi \otimes I_n), \qquad (4.33)\\
I_{M-1} \otimes \left(\frac{1}{\tau} A + D\right) =& (\Phi \otimes I_n) \left(I_{M-1} \otimes \left(\frac{1}{\tau} A + D\right)\right) (\Phi \otimes I_n).
\end{aligned}
$$

For simplicity, we assume that the first order derivative ∂_x is discretized by an one sided difference approximation ($q = 1$), say

$$
\frac{1}{h}\tilde{P}_q \equiv \frac{1}{h}\tilde{P} = \frac{1}{h} \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \\ & & & & -1 \end{pmatrix} \equiv \frac{1}{h}\left(-I_{M-1} + H_{M-1}\right). \qquad (4.34)
$$

Inserting this into the right hand side of eq. (4.24) and using eqs. (4.33), G^j can be written

$$
\begin{aligned}
G^j =& G_0 + G_1^j, \qquad (4.35)\\
G_0 =& I_{M-1} \otimes \left(\frac{1}{\tau} A - \frac{1}{h} C^0 + D\right) + \frac{1}{h^2} P \otimes B\\
=& (\Phi \otimes I_n) \left[I_{M-1} \otimes \left(\frac{1}{\tau} A - \frac{1}{h} C^0 + D\right) + \Lambda \otimes B\right] (\Phi \otimes I_n), \qquad (4.36)\\
G_1^j =& \frac{1}{h}\left[H_{M-1} \otimes C^0 + \tilde{P} \otimes C^1[U^j]\right]. \qquad (4.37)
\end{aligned}
$$

Eq. (4.36) implies that the $(n(M-1), n(M-1))-$matrix G_0 can be reduced to the set of $(n,n)-$matrices $G_{0k},\ k = 1, ..., M-1$,

$$
\begin{aligned}
G_0 =& (\Phi \otimes I_n) \begin{pmatrix} G_{01} & & \\ & \ddots & \\ & & G_{0M-1} \end{pmatrix} (\Phi \otimes I_n), \qquad (4.38)\\
G_{0k} :=& \frac{1}{\tau} A - \frac{1}{h} C^0 + D + \lambda_k B. \qquad (4.39)
\end{aligned}
$$

Therefore, if the low order matrices $G_{0k},\ k = 1, ..., M-1$, are regular (this is the case because of the choice of C^0), then G_0 is regular. If this is true, we can estimate $||G^{-j}||$ by writing $G^j = G_0\left(I_{n(M-1)} + G_0^{-1} G_1^j\right)$, and it follows

$$
||G^{-j}|| \ \leq || \left(I_{n(M-1)} + G_0^{-1} G_1^j\right)^{-1} || \ ||G_0^{-1}|| \qquad (4.40)
$$

provided $I_{n(M-1)} + G_0^{-1} G_1^j$ is also a regular matrix. The second factor on the right hand side can be written

$$
||G_0^{-1}|| = \max_k ||G_{0k}^{-1}||_n \ = \ \max_k \frac{1}{\left[\lambda_{min}\left(G_{0k}^T G_{0k}\right)\right]^{1/2}} \qquad (4.41)
$$

where $||Q||_n$ is the spectral norm of a real $(n,n)-$matrix Q, and $\lambda_{min}\left(G_{0k}^T G_{0k}\right)$ is the smallest eigenvalue of the matrix $G_{0k}^T G_{0k}$. The first factor on the right of ineq. (4.40) can be estimated by means of

$$
\left|\left|\left(I_{n(M-1)} + G_0^{-1} G_1^j\right)^{-1}\right|\right| \leq \frac{1}{1 - ||G_0^{-1} G_1^j||} \qquad (4.42)
$$

provided $||G_0^{-1}G_1^j|| < 1$. It may be appropriate to estimate further $||G_0^{-1}G_1^j|| \leq ||G_0^{-1}||\ ||G_1^j||$. $||G_0^{-1}||$ is given already in (4.41), and $||G_1^j||$ is

$$
\begin{aligned}
||G_1^j|| =& \frac{1}{h}\left|\left|H_{M-1}\otimes C^0 + \tilde{P}\otimes C^1[U^j]\right|\right| \qquad (4.43)\\
=& \frac{1}{h}\left|\left|-I_{M-1}\otimes C^1[U^j] + H_{M-1}\otimes C[U^j]\right|\right| \leq \frac{1}{h}\left(\left|\left|I_{M-1}\otimes C^1[U^j]\right|\right| + \left|\left|H_{M-1}\otimes C[U^j]\right|\right|\right)\\
=& \frac{1}{h}\left(\max_{k\in[1,M-1]}\left|\left|C^1[u_k^j]\right|\right|_n + \max_{k\in[1,M-2]}\left|\left|C[u_k^j]\right|\right|_n\right).
\end{aligned}
$$

Here we used the identities

$$
\left|\left|H_{M-1}\otimes C[U^j]\right|\right| = \left[\lambda_{max}\left(\left(H_{M-1}\otimes C[U^j]\right)^T\left(H_{M-1}\otimes C[U^j]\right)\right)\right]^{1/2}
$$

$$
= \left[\ \lambda_{max}\begin{pmatrix} 0 & & & \\ & C_1^TC_1 & & \\ & & \ddots & \\ & & & C_{M-2}^TC_{M-2}\end{pmatrix}\ \right]^{1/2} = \max_k\left[\ \lambda_{max}\left(C_k^TC_k\right)\right]^{1/2} = \max_k ||C_k||_n
$$

(for short $C_k = C[u_k^j]$). Relations (4.41) and (4.43) show that the norms $||G_0^{-1}||$ and $||G_1^j||$ can be estimated by norms of low order matrices (of order n where n is usually small, e.g. $n = 4$ for the plasma PDAE given in section 2).

We summarize the foregoing result in the following Lemma.

Lemma 4.2. *Suppose that for $\tau \in (0,\tau_0]$ and $h \in (0,h_0]$ $(\tau_0,\ h_0 > 0)$, $m\tau = t \in (0,t_e]$,*

1. *the $(n,n)-$matrices $G_{0k},\ k = 1,...,M-1,\ j = 1,...,m,$ are regular,*
2. *$||u_k^j||_n \leq K_0,\ k = 1,...,M-1,\ j = 1,...,m,$ where K_0 is some constant independent of τ and h,*
3. *there is a positive constant $\delta_0 < 1$ such that the condition*

$$
\begin{aligned}
||G_0^{-1}||\ ||G_1^j|| =& \left(\max_{k\in[1,M-1]}\ ||G_{0k}^{-1}||_n\right)\ \frac{1}{h}\left(\max_{k\in[1,M-1]}\ ||C^1[u_k^j]||_n\right. \qquad (4.44)\\
&\left.+\max_{k\in[1,M-2]}\left|\left|C[u_k^j]\right|\right|_n\right)\ \leq \delta_0
\end{aligned}
$$

is satisfied.

Then

$$
\left|\left|G^{-j}\right|\right|\ \leq \frac{1}{1-\delta_0}\left|\left|G_0^{-1}\right|\right|. \qquad \square
$$

This condition plays a crucial role when proving the convergence of the difference scheme considered here. We illustrate this by the following simple example.

Example 4.1. Inequality (4.44) is for the case $n = 1,\ A = 1,\ B = -1,\ C = C^0 = const. < 0,\ D = 0$ and Dirichlet BVs of type of a CFL–condition known from the discretization theory of hyperbolic differential equations. In this case $G = \frac{1}{\tau}I_{M-1} - \frac{1}{h^2}P + \frac{C}{h}\tilde{P}$,

$$
G = G_0 + G_1, \qquad G_0 = \Phi\left[\left(\frac{1}{\tau}-\frac{C}{h}\right)I_{M-1} - \Lambda\right]\Phi, \quad G_1 = \frac{C}{h}H_{M-1}.
$$

Therefore, $G_{0k} = \frac{1}{\tau} - \frac{C}{h} - \lambda_k$ which implies

$$||G_0^{-1}|| = \frac{\tau}{\min_k \left(1 + |C|\frac{\tau}{h} + \tau|\lambda_k|\right)} \leq \tau.$$

Furthermore, $H_{M-1}^T H_{M-1}$ is a diagonal matrix which has eigenvalues 0 and 1, hence $||H_{M-1}|| = 1$, and

$$||G_0^{-1}||\ ||G_1|| \leq \tau \frac{|C|}{h}.$$

If this is required to be less than 1 (according to (4.44)), it resembles the well known CFL condition $|C|\tau/h < 1$. $\square$

In order to prove convergence, the estimate (4.30) should be applied in the error inequality (4.29). To estimate the norms $||G^{-j}E_l||$, $l = 1, ..., 5$, we use again the decomposition (4.35) under the assumption that G_0^{-1} exists. Then the relation $G^j = G_0\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)$ gives

$$\begin{aligned} ||G^{-j}E_l|| &= ||(G_0 + G_1^j)^{-1}E_l|| = \left|\left|\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)^{-1} G_0^{-1}E_l\right|\right| \\ &\leq \left|\left|\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)^{-1}\right|\right|\ ||G_0^{-1}E_l||\,. \end{aligned}$$

Since E_l is (for all l) a block diagonal matrix, the representation (4.36) implies that also the norm $||G^{-j}E_l||$ can be estimated by the calculation of norms of $(n, n)-$matrices. Because

$$||G_0^{-1}E_l|| = \left|\left|\left[I_{M-1} \otimes \left(\frac{1}{\tau}A - \frac{1}{h}C^0 + D\right) + \Lambda \otimes B\right]^{-1} E_l\right|\right|,$$

it is in obvious notation

$$\begin{aligned} ||G^{-j}E_l|| &\leq \left|\left|\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)^{-1}\right|\right|\ \left|\left|\begin{pmatrix} G_{01}^{-1} & & \\ & \ddots & \\ & & G_{0M-1}^{-1} \end{pmatrix}\begin{pmatrix} E_{l1} & & \\ & \ddots & \\ & & E_{lM-1} \end{pmatrix}\right|\right| \\ &= \left|\left|\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)^{-1}\right|\right|\ \max_k\ ||G_{0k}^{-1}E_{lk}||_n. \end{aligned} \tag{4.45}$$

where, for example, for $l = 1$, $E_{1k} = A + C^1[V_t(t_{j+1}, x_k)]$, see (4.22).
The result of the foregoing estimates is the following Lemma.

Lemma 4.3. *Suppose*

1. $C[u] = C^0 + C^1[u] \in \mathbb{R}^{n\times n}$ *is linear in u (see also eq. (4.14)),*
2. *the exact solution* $v = v(t, x)$ *of the IBVP is sufficiently smooth for* $t \in (0, t_e]$, $t_e > 0$, *especially* $||V||$, $||V_t|| < \infty$,
3. $m\tau = t \in (0, t_e]$, $m \in \mathbb{N}$, *t fixed,*
4. $||U^j|| < \infty$, $j = 1, ...m$,
5. $||\eta^0|| = 0$,
6. $\sup_{m\in\mathbb{N}} \{||H^m H^{m-1} ... H^j||,\ j = 1, ..., m\} < \infty$,
7. *inequality (4.44) with* $\delta_0 < 1$ *is valid.*

Then an upper bound of the global error (see ineq. (4.29)) can be expressed by norms of the low order matrices $G_{0k}^{-1}E_{lk},\ l = 1, ..., 5$:

$$||\eta^{m+1}|| \ \leq \frac{K_1}{1-\delta_0}\ \frac{t}{\tau} \max_{j\in[0,m]} \Big\{ \max_{k\in[1,M-1]} \Big[\Big(||G_{0k}^{-1}E_{1k}||_n + ||G_{0k}^{-1}E_{5k}||_n\Big)\ O(\tau) \tag{4.46}$$
$$+||G_{0k}^{-1}E_{2k}||_n\ O(h^2) + ||G_{0k}^{-1}E_{3k}||_n\ O(h) + ||G_{0k}^{-1}E_{4k}||_n\ O(\tau h)\Big]\Big\}. \qquad \square$$

Remark 4.3.

1. The sixth requirement is a stability condition. Conditions of this type are well known in the theory of difference methods of time dependent partial differential equations (see also [10] where the assumption is discussed for linear PDAEs without convection). Unfortunately, a general easy method to verify the assumption when a convection term is present is not available. It should be studied separately for each problem of type (1.1) with given matrices $A,\ B,\ C,\ D$.
2. In most cases, the seventh assumption can be satisfied only when the step sizes τ and h are restricted in a certain way and/or when the convection term is not dominant, see also example 4.1.

5. Numerical example

For an example we choose the plasma PDAE (with $n = 4,\ d = 1$) described in section 2.The matrices are defined in (2.1). Let the parameters in B and $C[u]$ be defined by $b_0 = 0.02$ and $d_1 = 1$. First, according to (1.2) and (1.3) IVs and (Dirichlet) BVs must be specified. Since $\nu_t = 1$ (see section 3.3), both terms $\Phi_a,\ \Phi_c$ are different from zero. One can prescribe arbitrary IVs for two components of $u = u(t,x)$ only, say u_2 and u_4 (these are the non zero components of Φ_a). On the other hand, the spatial index is zero (i.e. $\Psi_c = 0$), and therefore we can assign especially the BV $u_3(0,0)$ which is assumed to be non zero. Given this BV and

$$g_4(x) := u_4(0,x) = K_4\cos(2\pi x) \tag{5.1}$$

(where the constant K_4 is defined below), it follows from the third equation of the system (1.1), (2.1) that the consistent IV of u_3 is

$$g_3(x) := u_3(0,x) = u_3(0,0)e^{g_4(x)}/e^{g_4(0)}. \tag{5.2}$$

This can be inserted into the fourth equation of the system with the result that the consistent IV for u_1 is

$$g_1(x) := u_1(0,x) = g_3(x) - g_{4,xx}(x) = u_3(0,0)e^{g_4(x)}/e^{g_4(0)} + 4\pi^2 g_4(x). \tag{5.3}$$

The IV for u_2 can be given arbitrarily, here we choose

$$g_2(x) := u_2(0,x) = K_2x(x-0.5), \qquad K_2 = const., \tag{5.4}$$

and BVs are for $t \in \bar{I}_t$

$$u_1(t,0) = u_1(t,1) = 0, \quad u_2(t,0) = 0, \quad u_3(t,0) = u_3(0,0), \quad u_4(t,0) = u_4(t,1) = K_4. \tag{5.5}$$

According to the compatibility relation (1.4), the BVs for u_1 imply $K_4 = -u_3(0,0)/(4\pi^2)$. With these relations the IBVP (1.1) – (1.4) with matrices (2.1) is defined completely, and we may solve it by a finite difference method. The two parameters K_2 and $u_3(0,0)$ can be chosen arbitrarily, for an example let $u_3(0,0) = 0.2$ and $K_2 = 0.4$. Furthermore, let the two step sizes τ and h be related by $\tau = K_0 h$, (or less restrictive by $\tau \le K_0 h$) where K_0 is a positive constant ($K_0 = 0.5$ below). This restriction has its origin in the second equation which is of hyperbolic type (inhomogeneous Burgers equation). In the example, the CFL condition for the second equation is satisfied for the chosen data because

$$\frac{\tau}{h} \max_k \{|u_{2k}^m|_{m\tau=1}\}$$

is much smaller than 1. Some values of this expression are given in Table 5.1 in the line CFL_2 for different values of $N = 1/h$. Furthermore, in the table e_i is defined by $e_i := ||U_{i,h} - U_{i,h/2}||$, $i = 1, 2$, where $U_{i,h}$ and $U_{i,h/2}$ are the numerical approximations of u_i at time $t = 1$ and at the space grid points for the two different step sizes h and $h/2$. Note that u_1 and u_2 are the ion density and the ion velocity, respectively (see section 2), and therefore, u_1 must be nonnegative.

In Figure 5.1, a typical result of the finite difference solution of the IBVP with the given data at time $t = 1$ is shown. The components u_1 and u_2 which are the most interesting ones are presented.

FIGURE 5.1

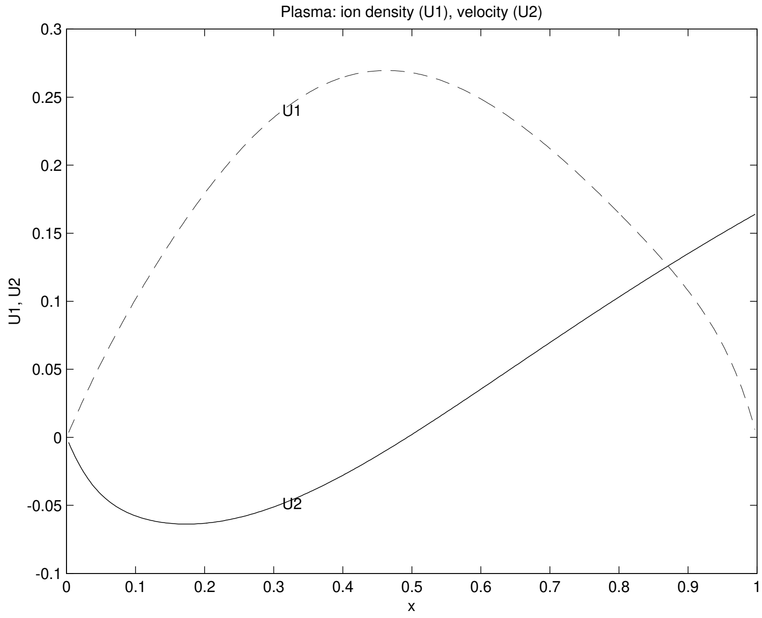

TABLE 5.1

N	20	40	80	160	320
CFL_2	0.0771	0.0799	0.0811	0.0817	0.0819
e_1	0.0075	0.0051	0.0027	0.0013	0.0006
e_2	0.0141	0.0113	0.0076	0.0049	0.0031

6. Conclusion

After some important applications of PDAEs with nonlinear convection terms we first considered the determination of the time and spatial index of such systems. These were based on definitions given earlier in the literature, e.g. by Martinson and Barton [13]. We reduced this problem to the question whether there is an inverse of some operator defined on a properly defined solution space. For the case of the plasma PDAE (a system with time index 1) this was shown in detail. In most practical applications, PDAEs have time index 1 or 2.

Then we studied the numerical solution of corresponding IBVPs by means of a finite difference splitting method familiar from the treatment of partial differential equations. Especially the convergence was considered for the case that both the time and space step sizes tend to zero. The main problem in these investigations is (in one space dimension) to handle the limit $h \to 0$ in relation to the time step size τ. For fixed h and $\tau \to 0$ this problem is easy. Some results concerning the problem when both h and τ go to zero were given.

Finally, some results of a numerical solution of an IBVP from plasma physics were presented. This example is interesting because the PDAE which is nonlinear is a mixture of partial differential equations of parabolic, elliptic and hyperbolic type.

References

[1] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. North-Holland Publ. Co., Amsterdam, 1989.

[2] S.L. Campbell and W. Marszalek. The index of an infinite dimensional implicit system. *Math. Mod. of Syst.*, 1(1):1–25, 1996.

[3] S.L. Campbell and W. Marszalek. DAEs arising from traveling wave solutions of PDEs. *J. Comput. Appl. Math.*, 82(1-2):41–58, 1997.

[4] S. L. Campbell and W. Marszalek. ODE/DAE integrators and MOL problems. ICIAM 95 Minisymposium on MOL, 1995.

[5] S. C. Cowin. Bone poroelasticity. *Journ. of Biomechan.*, 32:217–238, 1999.

[6] E. Detournay and H.-D. A. Cheng. *Fundamentals of poroelasticity. In: Hudson, J. A. (Ed.), Comprehensive rock engineering: principles, practice and projects*. Pergamon, Oxford, 1993.

[7] R.K. Dodd, J.C. Eilbeck, J.D. Gibbon, and H.C. Morris. *Solitons and nonlinear wave equations*. Academic Press, New York, 1982.

[8] M. Günther and Y. Wagner. Index concepts for linear mixed systems of differential-algebraic and hyperbolic-type equations. *SIAM Journ. on Sci. and Statist. Comp., in press*, 1999.

[9] N.N. Janenko. *Die Zwischenschrittmethode zur Lösung mehrdimensionaler Probleme der mathematischen Physik*. Springer-Verlag, Berlin, 1969.

[10] W. Lucht, K. Strehmel, and C. Eichler-Liebenow. Indexes and special discretization methods for linear partial differential algebraic equations. *BIT*, 39, No. 3:484–512, 1999.

[11] W. Lucht and K. Strehmel. Discretization based indices for semilinear partial differential algebraic equations. *Appl. Numer. Math.*, 28:371–386, 1998.

[12] W. Marszalek. *Analysis of partial differential algebraic equations*. PhD thesis, North Carolina State University, Raleigh, 1997.

[13] W.S. Martinson and P.I. Barton. A differentiation index for partial differential equations. *SIAM Journ. on Sci. Comp., in press*, 1998.