ELSEVIER

Applied Numerical Mathematics 42 (2002) 297–314

APPLIED NUMERICAL MATHEMATICS

www.elsevier.com/locate/apnum

On quasi-linear PDAEs with convection: Applications, indices, numerical solution

W. Lucht *, K. Debrabant

Martin-Luther-Universität Halle-Wittenberg, Fachbereich Mathematik und Informatik, Institut für Numerische Mathematik, Postfach, D-06099 Halle, Germany

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. As a typical example we investigate an application from plasma physics. Especially 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 the plasma PDAE are given. © 2002 IMACS. Published by Elsevier Science B.V. All rights reserved.

Keywords: Partial differential algebraic equations; Indices for mixed nonlinear systems; Numerical solution of PDAEs

1. Introduction

In this paper quasi-linear PDAEs for $u = u(t, x), x \in \Omega := (0, 1)$, of the form

$$Au_t + Bu_{xx} + C[u]u_x + Du = f(t, x), \quad t \in (0, t_e),\ x \in \Omega, \tag{1}$$

for some $t_e > 0$ are considered. u and f are mappings $u,\ f : [0, t_e] \times \overline{\Omega} \to \mathbb{R}^n, n \geqslant 1$, where f (supposed to be sufficiently smooth) is given. $A, B, C[u]$ and D are real (n, n)-matrices where A, B and D are assumed to be constant. All matrices may be singular, but $A, B \neq 0$. $C[u]$ may depend on u. We suppose that it is linear in u. Typically, when $C[u]$ is linear in u, vector $C[u]u_x$ describes physical convection.

* Corresponding author.
E-mail address: lucht@mathematik.uni-halle.de (W. Lucht).

PII: S0168-9274(01)00157-X

For system (1) we study classical initial boundary value problems (IBVPs). Initial values (IVs) may be decomposed

$$u(0,x) = \Phi_a(x) + \Phi_c(x), \quad x \in \overline{\Omega}, \tag{2}$$

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

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

with $(k = 1, \ldots, n)$. The boundary values (BVs) are of similar form,

$$u(t,x) = \Psi_a(t,x) + \Psi_c(t,x), \quad t \in [0,t_e],\ x \in \partial\Omega = \{0,1\}. \tag{3}$$

This means the data which can be prescribed arbitrarily are in Φ_a, Ψ_a. The consistent data (see, e.g., [2] or [12]) are collected in Φ_c, Ψ_c. Furthermore, we assume that the compatibility relations

$$\Phi_a(x) + \Phi_c(x) = \Psi_a(0,x) + \Psi_c(0,x), \quad x \in \partial\Omega, \tag{4}$$

are satisfied. Because almost every PDAE has its own IV and BV distribution here we avoid giving 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 a nonlinear PDAE from physics is presented. In Section 3 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. An application

In this section we cite a mathematical model from plasma physics [7, Chapter 5] whose underlying system of equations is of type (1). 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 the space domain Ω. 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

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

respectively. 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) with $n = 4$, $f = 0$ and new dependent variables u_i and with matrices

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} -b_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$
$$C[u] = \begin{pmatrix} u_2 & u_1 & 0 & 0 \\ 0 & u_2 & 0 & d_1 \\ 0 & 0 & -1 & u_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{pmatrix}. \tag{5}$$

The new variables u_i are (up to a constant) given by $u_1 \sim n_i$, $u_2 \sim v_i$, $u_3 \sim n_e$, $u_4 \sim \phi$. $b_0 \geqslant 0$ 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).

Further examples of a system of type (1) are a poroelastic model of a living bone [5,6,10] and the incompressible Navier–Stokes equations if Eq. (1) is generalized in obvious manner to two and three space dimensions.

3. Indices

First we introduce two definitions of classical linear spaces. Let l, m be nonnegative integers, and let $C_{sc}^{l,m}$ 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 lth order and space derivatives up to the mth order) are continuous. With $C_{sc}^{l,m}$ the set

$$C_n^{l,m} := \left\{ w = (w_1, \ldots, w_n)^{\mathrm{T}};\ w_i \in C_{sc}^{l,m},\ i = 1, \ldots, n \right\}$$

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

$$C_{n,0}^{l,m} := \left\{ w \in C_n^{l,m};\ w(t,0) = w(t,1) = 0 \right\}.$$

In the analytical and numerical treatment of differential algebraic equations (DAEs), the so-called index plays an important role [1]. For example, a linear DAE of index 3 in general cannot be solved by the implicit Euler method. Therefore, we expect such a dependence also for PDAEs.

While the index of linear PDAEs has been considered by several authors, e.g., [2–4,8,12,13], the index for nonlinear systems is little investigated, see however [11], where a discretization based index definition has been given, and [14]. We mention that we do not transform system (1) to a first order system because in numerical calculations we prefer 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 indices. The basic notion of a time index ν_t and a spatial index ν_x can be found, e.g., in [3,12,14].

Throughout this paper the time index ν_t of (1) is of special interest. We assume that a solution u of the IBVP (1)–(4) exists and that u is sufficiently differentiable. Furthermore, we suppose that $u(t,x) = 0$ for $x \in \partial\Omega$. When Ψ_a and Ψ_c (see Eq. (3)) are known, zero BVs can be obtained by a suitable transformation of u.

3.1. Time index

Definition 1. If the matrix A is regular, the time index ν_t of the PDAE (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 indices for PDAEs of type (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 2. *Let 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* $C^{(1)}_{ik}[u_x] := C_{ij,k}u_{j,x}$,
(II) $\partial_t C^{(1)}[u_x]u_t = C^{(1)}[u_{tx}]u_t + C^{(1)}[u_x]u_{tt}$.

Proof. Let $C_{ij,k} := \frac{\partial C_{ij}}{\partial u_k}$, $i, j, k \in \{1, \ldots, n\}$. Componentwise differentiation with respect to time yields

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

Statement (I) follows from the definition of $C^{(1)}_{ik}$. Using this definition again and the linearity of C in u we get

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

which is (II) in component form. □

For example, $C[u]$ given in (5) 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

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

we find formally from Eq. (1) under the assumptions of Lemma 2 after one time differentiation (this case is relevant for $\nu_t = 1$) the derivative array

$$\begin{pmatrix} A & 0 \\ M[u, u_x] & A \end{pmatrix} \begin{pmatrix} u_t \\ u_{tt} \end{pmatrix} + \begin{pmatrix} Lu + C[u]u_x \\ 0 \end{pmatrix} = \begin{pmatrix} f \\ f_t \end{pmatrix}. \tag{6}$$

Two time differentiations of (1) produce the derivative array (for $\nu_t = 2$)

$$\begin{pmatrix} A & 0 & 0 \\ M[u, u_x] & A & 0 \\ 2C^{(1)}[u_{tx}] & M[u, u_x] & A \end{pmatrix} \begin{pmatrix} u_t \\ u_{tt} \\ u_{ttt} \end{pmatrix} + \begin{pmatrix} Lu + C[u]u_x \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} f \\ f_t \\ f_{tt} \end{pmatrix}. \tag{7}$$

When A is singular, the coefficient matrices (denoted by $\mathcal{A}$) of the systems (6), (7) 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. (6),

$v = (u_t^{\mathrm{T}}, u_{tt}^{\mathrm{T}})^{\mathrm{T}}, u \in C_n^{2,2}$). 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 [14].

For the nonlinear PDAE (1), the derivative array (6) or (7) is used to write

$$Pu_t = F, \tag{8}$$

where P is according to Eq. (6) or (7) 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.

We mention, that the matrix P in Eq. (8) is the analogue to the nonsingular diagonal matrix $D(x_j)$ defined in [14], Definition 3.5. However, here we need not require a diagonal form of P.

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 $\overline{B} = S_0 B S_1^{-1} = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}$ where S_0, S_1 are constant regular (n, n)-matrices.

Definition 3. 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

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

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

$$\overline{U} := \big(\underbrace{\overline{u}_{1,xx}, \overline{u}_{2,xx}, \ldots, \overline{u}_{m,xx}}_{m \text{ elements}}, \underbrace{\overline{u}_{n_1+1,x}, \ldots, \overline{u}_{n,x}}_{n-m \text{ elements}}\big)^{\mathrm{T}}$$

as a continuous function of $t, x, \overline{u}, \overline{u}_t$ and $\bar{u}_{1,x}, \ldots, \bar{u}_{m,x}$.

It is straightforward to derive for $\overline{U}$ a derivative array (by differentiations of the PDAE with respect to x) analogous to Eq. (6) or (7). As a result, after a minimal number of x-differentiations one can find a representation of $\overline{U}$ of the form $\overline{Q}\overline{U} = \overline{G}$ 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 [12]. It does not transform the PDAE (1) to a system of first order as is often done, e.g., in [14]. Such a transformation, in general, changes the indices. For example, one can show that the time index here is 0 if and only if the corresponding index in [14] is 1, a time index 1 or 2 here implies an index 2 there.

3.3. Time index of the plasma PDAE

Here, we study the time index ν_t of the PDAE (1) under the assumption of zero BVs, $u(t, 0) = u(t, 1) = 0$, and $f \in C_4^{p,q}$ with suitable p, q. To be specific we further assume that this IBVP has a solution $u \in C_{4,0}^{1,2}$ which possibly exists only local in time. To determine the time index of system (1) we generate for ν_t—if possible—system (8) with the condition that P defined on $D(P)$ is invertible. In

order to determine ν_t we differentiate the third and fourth equation of the PDAE with respect to t with the result

$$-u_{3,tx} + u_{3,t}u_{4,x} + u_3 u_{4,tx} = 0 \quad \text{and} \quad u_{4,txx} + u_{1,t} - u_{3,t} = 0.$$

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

$$Pu_t = F(u) := \begin{pmatrix} -u_2u_{1,x} - u_1u_{2,x} + b_0u_{1,xx} \\ -u_2u_{2,x} - d_1u_{4,x} \\ 0 \\ 0 \end{pmatrix}, \quad P := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & [-\partial_x + u_{4,x}] & u_3\partial_x \\ 1 & 0 & -1 & \partial_x^2 \end{pmatrix}. \tag{9}$$

Now what is essential for the determination of the index of the PDAE is that u in P can be seen as a fixed element with the result that $P : D(P) \to R(P)$, $D(P) \subset C^{1,2}_{4,0}$, is a linear operator ($R(P)$ denotes the range of P). We can try to solve the first equation in (9) for u_t by standard linear theory. In particular, if P^{-1} exists we get ν_t as the least number of time differentiations which are necessary to get the equation for u_t (in the example considered here, one differentiation with respect to t is needed).

Remark 4. The right-hand side of the first equation in (9) does not play any role when ν_t is to be determined. This implies that ν_t is independent of f in Eq. (1). This is a reasonable result because the right-hand side function should not influence the indices of a PDAE.

We ask whether the linear operator P with $D(P) \subset C^{1,3}_{4,0}$ has for fixed $u = u(t,x)$ an inverse. The answer comes from kernel $N(P)$ whose elements z fulfill

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

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 5. *Let u be such that*

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

Then $z_3 = z_4 = 0$.

Since the proof is simple, it is omitted here (for details see [10]).

Under the assumptions of this lemma we see that $N(P) = \{0\}$. Therefore, system (9) can be solved for u_t, and the time index is $\nu_t = 1$.

We mention that based on Definition 3 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)–(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, see, e.g., [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) as $C[u, \partial_x]u$ which is more convenient sometimes. $C[u, \partial_x]$ is an operator valued matrix. For example, with $C[u]$ given in (5) it is

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

We suppose that A is singular and is given as $A = \begin{pmatrix} I_{n_1} & 0 \\ 0 & 0 \end{pmatrix}$ where I_{n_1} is the unit matrix of order $n_1 < n$ ($n_1 \geqslant 1$). Let $n_2 = n - n_1$. Corresponding to this partition of A we introduce the notation $u = (u_1^{\mathrm{T}}, u_2^{\mathrm{T}})^{\mathrm{T}}$ 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 $u_k = (u_{k1}^{\mathrm{T}}, \ldots, u_{kn_k}^{\mathrm{T}})^{\mathrm{T}},\ k = 1, 2$. According to this partition, PDAE (1) is written

$$\begin{pmatrix} I_{n_1} & 0 \\ 0 & 0 \end{pmatrix} u_t + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} u_{xx} + \begin{pmatrix} C_{11}[u, \partial_x] & C_{12}[u, \partial_x] \\ C_{21}[u, \partial_x] & C_{22}[u, \partial_x] \end{pmatrix} u + \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} u = f. \tag{10}$$

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

$$L_1[u] := \begin{pmatrix} 0 & 0 \\ (B_{21}\partial_x^2 + C_{21} + D_{21}) & (B_{22}\partial_x^2 + C_{22} + D_{22}) \end{pmatrix}, \tag{11}$$

$$L_2[u] := \begin{pmatrix} (B_{11}\partial_x^2 + C_{11} + D_{11}) & (B_{21}\partial_x^2 + C_{21} + D_{21}) \\ 0 & 0 \end{pmatrix} \tag{12}$$

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

Remark 6. 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. (16).

In order to obtain an approximate numerical solution of IBVP (1)–(4) we consider it with zero BVs (3) by means of a difference method. The first time derivative is approximated with an equidistant time step size τ by

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

The corresponding IBVP is space discretized on an equidistant grid

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

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

$$u_{xx}(t, x_k) \approx \frac{\delta^2}{h^2} u_k(t) := \frac{1}{h^2}\big(u_{k-1}(t) - 2u_k(t) + u_{k+1}(t)\big),$$
$$u_x(t, x_k) \approx \frac{\delta_0}{2h} u_k(t) \quad \text{or} \quad \frac{\delta_+}{h} u_k(t) \quad \text{or} \quad \frac{\delta_-}{h} u_k(t), \tag{13}$$

where δ_0, δ_+ and δ_- are the usual central, forward and backward difference operators, respectively. Now suppose $t \in (0, t_e)$, $t_e > 0$. We approximate Eq. (1) by the difference equation ($m = 0, 1, \ldots$, $k = 1, \ldots, M-1$)

$$A\frac{u_k^{m+1} - u_k^m}{\tau} + L_h\big[u_k^m\big]u_k^{m+1} = f_k^{m+1}. \tag{14}$$

$L_h[u_k^m]$ denotes a discretization of $L[u(t_m, x_k)]$. We rewrite this equation as

$$\big(A + \tau L_h\big[u_k^m\big]\big)\frac{u_k^{m+1} - u_k^m}{\tau} = f_k^{m+1} - L_h\big[u_k^m\big]u_k^m \tag{15}$$

and factorize approximately for $\tau \to 0$

$$\begin{aligned} A + \tau L_h\big[u_k^m\big] &\approx \big(A + \tau L_{h1}\big[u_k^m\big]\big)\big(I + \tau L_{h2}\big[u_k^m\big]\big) \\ &= A + \tau\big(L_{h1}\big[u_k^m\big] + AL_{h2}\big[u_k^m\big]\big) + \mathrm{O}\big(\tau^2\big) = A + \tau L_h\big[u_k^m\big] + \mathrm{O}\big(\tau^2\big), \end{aligned} \tag{16}$$

where L_{h1} and L_{h2} are corresponding discretizations of (11) and (12), respectively (we used $AL_{h2} = L_{h2}$ and $L_{h1} + L_{h2} = L_h$). u_k^m is considered to be an approximation for $u(t_m, x_k)$. Therefore, we study for $m = 0, 1, \ldots$, $k = 1, \ldots, M-1$ fractional splitting

$$\big(A + \tau L_{h1}\big[u_k^m\big]\big)u_k^{m+1/2} = f_k^{m+1} - L_h\big[u_k^m\big]u_k^m, \tag{17}$$

$$\big(I + \tau L_{h2}\big[u_k^m\big]\big)\frac{u_k^{m+1} - u_k^m}{\tau} = u_k^{m+1/2}. \tag{18}$$

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

Remark 7. 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 (16) implies the following lemma.

Lemma 8. *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* (10), (2)–(4),
(2) *the system of algebraic equations* (17), (18) *has a unique solution.*

Then the method (17), (18) *and the system* (15) *are equivalent for* $\tau \to 0$, *i.e.,* (17), (18) *approximate the original system to the same* τ*-order as* (15).

4.1. Convergence of the fractional step method

We consider the convergence of the numerical solution u_k^m calculated by the scheme (14) towards the exact solution $v(t_m, x_k)$ of the corresponding IBVP for $\tau \to 0$ and $h \to 0$ under the condition $(m+1)\tau = t$ (t fixed). The basic assumption is that $C[v] = C^0 + C^1[v]$ is linear in v where the matrix $C^0 = \text{const}$ is chosen in such a manner that the matrices $G_{0k},\ k = 1, \ldots, M-1$, defined below in (39) are regular. For example, a proper choice may be $C^0 = C[\bar{v}],\ C^1[v] = C[v] - C[\bar{v}]$ 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 $m = 0, 1, \ldots,\ k = 1, \ldots, M-1$ by

$$\alpha_k^{m+1} := A\frac{v_k^{m+1} - v_k^m}{\tau} + L_h\big[v_k^m\big]v_k^{m+1} - f_k^{m+1}, \tag{19}$$

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

$$\begin{aligned}\alpha_k^{m+1} \;=\; & A\left(v_{k,t}^{m+1} - \frac{1}{2}v_{k,tt}^{m+1}\tau + \mathrm{O}\big(\tau^2\big)\right) + B\left(v_{k,xx}^{m+1} + \frac{1}{12}v_{k,xxxx}^{m+1}h^2 + \mathrm{O}\big(h^4\big)\right)\\ & + C\big[v_k^{m+1} - v_{k,t}^{m+1}\tau + \mathrm{O}\big(\tau^2\big)\big]\big(v_{k,x}^{m+1} + \overline{\mathrm{O}}^{m+1}\big(h^q\big)\big) + Dv_k^{m+1} - f_k^{m+1}.\end{aligned}$$

Since

$$C\big[v_k^{m+1} - v_{k,t}^{m+1}\tau\big] = C^0 + C^1\big[v_k^{m+1}\big] - \tau C^1\big[v_{k,t}^{m+1}\big]$$

and

$$Av_{k,t}^{m+1} + Bv_{k,xx}^{m+1} + \big(C^0 + C^1\big[v_k^{m+1}\big]\big)v_{k,x}^{m+1} + Dv_k^{m+1} = f_k^{m+1},$$

$$\begin{aligned}\alpha_k^{m+1} \;=\; & A\overline{\mathrm{O}}^{m+1}(\tau) + B\overline{\mathrm{O}}^{m+1}\big(h^2\big) + C\big[v_k^{m+1}\big]\overline{\mathrm{O}}^{m+1}\big(h^q\big)\\ & + C^1\big[v_{k,t}^{m+1}\big]\overline{\mathrm{O}}^{m+1}\big(\tau h^q\big) + C^1\big[v_{k,t}^{m+1}\big]\overline{\mathrm{O}}^{m+1}(\tau).\end{aligned} \tag{20}$$

Here, e.g., $\overline{\mathrm{O}}^{m+1}(\tau) \to \tau w$ for $\tau \to 0$, $w \in \mathbb{R}^n$, w independent of τ. Second, scheme (14) is written by means of the Kronecker product in matrix form,

$$\begin{aligned}&\left(I_{M-1} \otimes \tfrac{A}{\tau} + Q_h\big[U^m\big]\right)U^{m+1} = \left(I_{M-1} \otimes \tfrac{A}{\tau}\right)U^m + F^{m+1},\\ &Q_h\big[U^m\big] := \tfrac{1}{h^2}P \otimes B + \tfrac{1}{qh}\widetilde{P}_q \otimes C\big[U^m\big] + I_{M-1} \otimes D,\end{aligned} \tag{21}$$

where $U^m := (u_1^{m\mathrm{T}}, \ldots, u_{M-1}^{m\mathrm{T}})^{\mathrm{T}}$, $F^m := (f_1^{m\mathrm{T}}, \ldots, f_{M-1}^{m\mathrm{T}})^{\mathrm{T}}$, and

$$P := \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)}. \tag{22}$$

q and $\widetilde{P}_q$ depend on the discretization of the first space derivative. For example, using central difference formula in (13), it is $q = 2$ and

$$\widetilde{P} := \begin{pmatrix} 0 & 1 & & \\ -1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 0 \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)}$$

(another differencing is chosen in (35)). The $(n(M-1), n(M-1))$-matrix $\widetilde{P}_q \otimes C[U^m]$ is a block matrix whose block at position (j,k) is $\widetilde{P}_{q,jk} C[u_j^m]$, $u_j^m \in \mathbb{R}^n$. Eqs. (19) and (20) imply that V^{m+1} satisfies equation

$$\left(I_{M-1} \otimes \frac{A}{\tau} + Q_h\big[V^m\big]\right) V^{m+1} = \left(I_{M-1} \otimes \frac{A}{\tau}\right) V^m + F^{m+1} + \mathcal{A}^{m+1}, \tag{23}$$

$$\begin{aligned} \mathcal{A}^{m+1} :=\ & E_1 \mathrm{O}^{m+1}(\tau) + E_2 \mathrm{O}^{m+1}\big(h^2\big) + E_3^{m+1} \mathrm{O}^{m+1}\big(h^q\big) \\ & + E_4^{m+1}\big[\mathrm{O}^{m+1}\big(\tau h^q\big) + \mathrm{O}^{m+1}(\tau)\big], \end{aligned} \tag{24}$$

where

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

and, e.g., $\mathrm{O}^{m+1}(\tau)$ means $\mathrm{O}^{m+1}(\tau) \in \mathbb{R}^{n(M-1)}$. 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 $v = (v_1, \ldots, v_{n(M-1)})^{\mathrm{T}}$ by $||v|| := [h \sum_{k=1}^{n(M-1)} v_k^2]^{1/2}$.

Subtracting Eq. (21) from Eq. (23), we obtain

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

By the identity

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

the foregoing equation can be written

$$G^m \eta^{m+1} = \left(I_{M-1} \otimes \frac{A}{\tau}\right) \eta^m - \big(Q_h\big[V^m\big] - Q_h\big[U^m\big]\big) V^{m+1} + \mathcal{A}^{m+1}, \tag{25}$$

$$G^m := I_{M-1} \otimes \frac{A}{\tau} + \frac{1}{h^2} P \otimes B + \frac{1}{qh} \widetilde{P}_q \otimes C\big[U^m\big] + I_{M-1} \otimes D. \tag{26}$$

Now we use the fact that

$$Q_h\big[V^m\big] - Q_h\big[U^m\big] = \frac{1}{qh} \widetilde{P}_q \otimes \big(C\big[V^m\big] - C\big[U^m\big]\big) = \frac{1}{qh} \widetilde{P}_q \otimes C^1\big[\eta^m\big]$$

(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 $\widetilde{C}[V^{m+1}]$ such that

$$\big(\widetilde{P}_q \otimes C^1\big[\eta^m\big]\big) V^{m+1} = \widetilde{C}\big[V^{m+1}\big] \eta^m.$$

Therefore, Eq. (25) takes the form ($G^{-m} = (G^m)^{-1}$)

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

or for short

$$\eta^{m+1} = H^m\eta^m + R^{m+1}, \quad H^m := G^{-m}\left(I_{M-1} \otimes \frac{A}{\tau} - \frac{1}{qh}\widetilde{C}\big[V^{m+1}\big]\right),$$

$R^{m+1} := G^{-m}\mathcal{A}^{m+1}$. It follows for $m = 0, 1, \ldots$

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

$$\begin{aligned}\big||\eta^{m+1}\big|| \leqslant{}& \big||H^m H^{m-1}\cdots H^0\big|\big||\eta^0\big|| + \big||H^m H^{m-1}\cdots H^1\big|\big||R^1\big|| \\ &+ \big||H^m H^{m-1}\cdots H^2\big|\big||R^2\big|| + \cdots + \big||H^m\big|\big||R^m\big|| + \big||R^{m+1}\big||.\end{aligned} \tag{27}$$

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

$$\sup_{m\in\mathbb{N}}\big\{\big||H^m H^{m-1}\cdots H^j\big||,\ j = 1, \ldots, m\big\} < \infty, \tag{28}$$

possibly under a restriction of one of the forms

$$\kappa_0 \leqslant \frac{\tau}{h} \leqslant \kappa_1, \qquad \kappa_2 \leqslant \frac{\tau}{h^2} \leqslant \kappa_3. \tag{29}$$

Assumption (28) is discussed shortly in Remark 13. With this condition, the right-hand side in (27) can be estimated further to give

$$\begin{aligned}\big||\eta^{m+1}\big|| &\leqslant K_0\big||\eta^0\big|| + \overline{K}_1\sum_{j=0}^{m}]\big||R^{j+1}\big|| \leqslant K_0\big||\eta^0\big|| + (m+1)\overline{K}_1\max_{j\in[0,m]}\big||R^{j+1}\big|| \\ &\leqslant K_0\big||\eta^0\big|| + K_1\frac{t}{\tau}\max_{j\in[0,m]}\big||R^{j+1}\big||.\end{aligned} \tag{30}$$

K_0, K_1, $\overline{K}_1$ are constants. Using $||R^{j+1}|| = ||G^{-j}\mathcal{A}^{j+1}||$, relation (24) yields

$$\begin{aligned}\big||R^{j+1}\big|| \leqslant{}& \big||G^{-j}E_1\big|\big|\mathrm{O}(\tau) + \big||G^{-j}E_2\big|\big|\mathrm{O}\big(h^2\big) + \big||G^{-j}E_3^{j+1}\big|\big|\mathrm{O}\big(h^q\big) \\ &+ \big||G^{-j}E_4^{j+1}\big|\big|\big[\mathrm{O}\big(\tau h^q\big) + \mathrm{O}(\tau)\big]\end{aligned} \tag{31}$$

or also

$$\begin{aligned}\big||R^{j+1}\big|| \leqslant \big||G^{-j}\big|\big(&||E_1||\mathrm{O}(\tau) + ||E_2||\mathrm{O}\big(h^2\big) + \big||E_3^{j+1}\big|\big|\mathrm{O}\big(h^q\big) \\ &+\big||E_4^{j+1}\big|\big|\big[\mathrm{O}\big(\tau h^q\big) + \mathrm{O}(\tau)\big]\big).\end{aligned} \tag{32}$$

Note that in Eq. (31) is sharper than in Eq. (32). Since the matrices E_l are block diagonal, their norms can be estimated easily. For example, E_3^{j+1} is of the form $E_3^{j+1} = \mathrm{diag}\{C[v_1^{j+1}], \ldots, C[v_{M-1}^{j+1}]\}$, and

$$\big||E_3^{j+1}\big|| = \Big[\max_k\big[\lambda_{\max}\big(C^{\mathrm{T}}\big[v_k^{j+1}\big]C\big[v_k^{j+1}\big]\big)\big]\Big]^{1/2},$$

where $\lambda_{\max}(E)$ is the largest eigenvalue of E. Especially, if v is sufficiently smooth for some time $t \leqslant t_e$, then all norms $||E_l||$ are bounded. Therefore, (32) can be also written

$$\big||R^{j+1}\big|| \leqslant K_3\big||G^{-j}\big|\big|\big(\mathrm{O}(\tau) + \mathrm{O}\big(h^2\big) + \mathrm{O}\big(h^q\big) + \mathrm{O}\big(\tau h^q\big)\big)$$

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, $\frac{1}{h^2}P\phi_k = \lambda_k\phi_k,\ k = 1,\dots,M-1$. The eigenvalues are given by $\lambda_k = -\frac{4}{h^2}\sin^2(\frac{k\pi}{2M})$. They fulfill for $h \to 0$ the asymptotic relation $\lambda_k = -(k\pi)^2 + \mathrm{O}(h^2)$ (because $h = 1/M$). The set $\{\phi_k,\ k = 1,\dots,M-1\}$ is assumed to be orthonormal. Let $\Phi := \sqrt{h}(\phi_1,\dots,\phi_{M-1})$ be the matrix of the eigenvectors and $\Lambda := \mathrm{diag}(\lambda_1,\dots,\lambda_{M-1})$. Since $\Phi^{-1} = \Phi^{\mathrm{T}} = \Phi$, it follows that (for short $\Psi_n := \Phi \otimes I_n$)

$$\frac{1}{h^2}P \otimes B = \Psi_n(\Lambda \otimes B)\Psi_n, \tag{33}$$

$$I_{M-1} \otimes \left(\frac{1}{\tau}A + D\right) = \Psi_n\left(I_{M-1} \otimes \left(\frac{1}{\tau}A + D\right)\right)\Psi_n. \tag{34}$$

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

$$\frac{1}{h}\widetilde{P}_q \equiv \frac{1}{h}\widetilde{P} = \frac{1}{h}\begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \\ & & & & -1 \end{pmatrix} \equiv \frac{1}{h}(-I_{M-1} + H_{M-1}). \tag{35}$$

Inserting this into Eq. (26) and using Eqs. (33) and (34), G^j can be written

$$G^j = G_0 + G_1^j, \tag{36}$$

$$G_0 = \Psi_n\left[I_{M-1} \otimes \left(\frac{1}{\tau}A - \frac{1}{h}C^0 + D\right) + \Lambda \otimes B\right]\Psi_n, \tag{37}$$

$$G_1^j = \frac{1}{h}\left[H_{M-1} \otimes C^0 + \widetilde{P} \otimes C^1\left[U^j\right]\right]. \tag{38}$$

Eq. (37) implies that the $(n(M-1), n(M-1))$-matrix G_0 can be reduced to the set of (n,n)-matrices $G_{0k},\ k = 1,\dots,M-1$,

$$G_0 = \Psi_n\begin{pmatrix} G_{01} & & \\ & \ddots & \\ & & G_{0M-1} \end{pmatrix}\Psi_n, \qquad G_{0k} := \frac{1}{\tau}A - \frac{1}{h}C^0 + D + \lambda_k B. \tag{39}$$

If the low order matrices $G_{0k},\ k = 1,\dots,M-1$, are regular (this is due to the choice of C^0), then G_0 is regular. Therefore, $G^j = G_0(I_{n(M-1)} + G_0^{-1}G_1^j)$, and it follows

$$\left|\left|G^{-j}\right|\right| \leqslant \left|\left|\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)^{-1}\right|\right|\left|\left|G_0^{-1}\right|\right| \tag{40}$$

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

$$\left|\left|G_0^{-1}\right|\right| = \max_k\left|\left|G_{0k}^{-1}\right|\right|_n = \max_k \frac{1}{[\lambda_{\min}(G_{0k}^{\mathrm{T}}G_{0k})]^{1/2}}, \tag{41}$$

where $||Q||_n$ is the spectral norm of a real (n,n)-matrix Q, and $\lambda_{\min}(G_{0k}^{\mathrm{T}}G_{0k})$ is the smallest eigenvalue of the matrix $G_{0k}^{\mathrm{T}}G_{0k}$. The first factor on the right of in Eq. (40) can be estimated by means of

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

provided $||G_0^{-1}G_1^j|| < 1$. It may be appropriate to estimate further $||G_0^{-1}G_1^j|| \leqslant ||G_0^{-1}||||G_1^j||$. $||G_0^{-1}||$ is given already in (41), and $||G_1^j||$ is

$$\begin{aligned}\big||G_1^j|\big| &= \frac{1}{h}\big||-I_{M-1}\otimes C^1\big[U^j\big] + H_{M-1}\otimes C\big[U^j\big]\big|\big| \\ &\leqslant \frac{1}{h}\big(\big||I_{M-1}\otimes C^1\big[U^j\big]\big|\big| + \big||H_{M-1}\otimes C\big[U^j\big]\big|\big|\big) \\ &= \frac{1}{h}\Big(\max_{k\in[1,M-1]}\big||C^1\big[u_k^j\big]\big|\big|_n + \max_{k\in[1,M-2]}\big||C\big[u_k^j\big]\big|\big|_n\Big). \end{aligned} \tag{42}$$

Here we used the identities

$$\begin{aligned}\big||H_{M-1}\otimes C\big[U^j\big]\big|\big| &= \big[\lambda_{\max}\big(\big(H_{M-1}\otimes C\big[U^j\big]\big)^{\mathrm{T}}\big(H_{M-1}\otimes C\big[U^j\big]\big)\big)\big]^{1/2} \\ &= \left[\lambda_{\max}\begin{pmatrix} 0 & & & \\ & C_1^{\mathrm{T}}C_1 & & \\ & & \ddots & \\ & & & C_{M-2}^{\mathrm{T}}C_{M-2}\end{pmatrix}\right]^{1/2} \\ &= \max_k\big[\lambda_{\max}\big(C_k^{\mathrm{T}}C_k\big)\big]^{1/2} = \max_k||C_k||_n\end{aligned}$$

(for short $C_k = C[u_k^j]$). Relations (41) and (42) 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 9. *Suppose that for $\tau \in (0, \tau_0]$ and $h \in (0, h_0]$ ($\tau_0, h_0 > 0$), $(m+1)\tau = t \in (0, t_e]$:*

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

$$\big||G_0^{-1}|\big|\,\big||G_1^j|\big| \leqslant \Big(\max_{k\in[1,M-1]}\big||G_{0k}^{-1}|\big|_n\Big)\frac{1}{h}\Big(\max_{k\in[1,M-1]}\big||C^1\big[u_k^j\big]|\big|_n + \max_{k\in[1,M-2]}\big||C\big[u_k^j\big]|\big|_n\Big) \leqslant \delta_0 \tag{43}$$

is satisfied.

Then $||G^{-j}|| \leqslant \frac{1}{1-\delta_0}||G_0^{-1}||$.

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

Example 10. Inequality (43) is for the case $n = 1,\ A = 1,\ B = -1,\ C = C^0 = \text{const} < 0,\ D = 0$ and Dirichlet BVs of type of a CFL-condition known from the discretization theory of hyperbolic differential equations. In this case

$$G = \frac{1}{\tau}I_{M-1} - \frac{1}{h^2}P + \frac{C^0}{h}\widetilde{P} = G_0 + G_1,$$

$$G_0 = \Phi\left[\left(\frac{1}{\tau} - \frac{C^0}{h}\right) I_{M-1} - \Lambda\right]\Phi, \qquad G_1 = \frac{C^0}{h} H_{M-1}.$$

Therefore, $G_{0k} = \frac{1}{\tau} - \frac{C^0}{h} - \lambda_k$ and $||G_0^{-1}|| = \tau/\min_k(1 + |C^0|\frac{\tau}{h} + \tau|\lambda_k|) \leqslant \tau$. Furthermore, $H_{M-1}^{\mathrm{T}} H_{M-1}$ is a diagonal matrix which has eigenvalues 0 and 1, hence $||H_{M-1}|| = 1$, and

$$\left|\left|G_0^{-1}\right|\right| ||G_1|| \leqslant \tau \frac{|C^0|}{h}.$$

If this is required to be less than 1 (according to (43)), it resembles the well-known CFL condition $|C^0|\tau/h < 1$.

In order to prove convergence, the estimate (31) should be applied to the error inequality (30). To estimate the norms $||G^{-j}E_l||$, $l = 1, \ldots, 4$, we use again the decomposition (36) under the assumption that G_0^{-1} exists. Then the relation $G^j = G_0(I_{n(M-1)} + G_0^{-1}G_1^j)$ gives

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

Since E_l is (for all l) a block diagonal matrix, the representation (37) implies that also the norm $||G^{-j}E_l||$ can be estimated further by the calculation of norms of (n,n)-matrices,

$$\left|\left|G^{-j}E_l\right|\right| \leqslant \left|\left|\left(I_{n(M-1)} + G_0^{-1}G_1^j\right)^{-1}\right|\right| \max_k \left|\left|G_{0k}^{-1}\right|\right|_n \max_{k'} \left|\left|E_{lk'}\right|\right|_n.$$

Sometimes, this estimate may be useful.

The result of the foregoing estimates is the following Theorem.

Theorem 11. *Suppose*

(1) $C[u] = C^0 + C^1[u] \in \mathbb{R}^{n\times n}$ *is linear in* u,
(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) $(m+1)\tau = t \in (0, t_e]$, $m \in \mathbb{N}$, t *fixed*,
(4) $||U^j|| < \infty$, $j = 1, \ldots, m$,
(5) $||\eta^0|| = 0$,
(6) $\sup_{m\in\mathbb{N}}\{||H^m H^{m-1} \ldots H^j||,\ j = 1, \ldots, m\} < \infty$,
(7) $||G_0^{-1}G_1^j|| \leqslant \delta_0 < 1$ *for* $j = 1, \ldots, m$.

Then an upper bound of the global error (*see in Eq.* (30)) *can be expressed in terms of norms containing* G_0^{-1} *only*:

$$\begin{aligned}\left|\left|\eta^{m+1}\right|\right| \leqslant \frac{K_1}{1-\delta_0}\frac{t}{\tau}\Big\{&\max_{k\in[1,M-1]}\left[\left|\left|G_{0k}^{-1}A\right|\right|_n \mathrm{O}(\tau) + \left|\left|G_{0k}^{-1}B\right|\right|_n \mathrm{O}\left(h^2\right)\right] \\ &+ \max_{j\in[1,m+1]}\left[\left|\left|G_0^{-1}E_3^j\right|\right|\mathrm{O}(h) + \left|\left|G_0^{-1}E_4^j\right|\right|\left(\mathrm{O}(\tau) + \mathrm{O}(\tau h)\right)\right]\Big\}. \qquad (44)\end{aligned}$$

From this theorem, one may get convergence results possibly under a restriction of the type given in (29). $||G_{0k}A||_n$, $||G_{0k}^{-1}B||_n$ and $||G_0^{-1}E_l||$, $l = 3, 4$, depend on τ, h and on the indices of the PDAE (see also [12]).

Example 12. We illustrate this for the plasma system ($\nu_t = 1$) with the choice

$$C^0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The condition $\tau/h^2 \geqslant \kappa$ (where κ is independent of τ and h) implies $||G_{0k}^{-1}A||_4 = \mathrm{O}(\tau)$, $||G_{0k}^{-1}B||_4 = \mathrm{O}(\tau)$, but $||G_{0k}^{-1}||_4 = \mathrm{O}(\tau^{1/2})$ only. But with this result we are not able to show convergence for $\tau, h \to 0$.

The condition $\tau/h \geqslant \kappa$ is more satisfactorily. It implies $||G_{0k}^{-1}||_4 = \mathrm{O}(\tau)$, $||G_{0k}^{-1}A||_4 = \mathrm{O}(\tau)$, $||G_{0k}^{-1}B||_4 = \mathrm{O}(\tau^2)$ and also (by the second assumption of the Theorem) $||G_0^{-1}E_3^j|| = \mathrm{O}(\tau)$, $||G_0^{-1}E_4^j|| = \mathrm{O}(\tau)$. Therefore, in (44) the factor $\frac{t}{\tau}\{\ldots\}$ tends to zero for $\tau, h \to 0$.

Remark 13.

(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 [12] 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) with given matrices A, B, C, D.

(2) In most cases, the seventh assumption can be satisfied only if the step sizes τ and h are related by a restriction of the form as given in (29). Note that the seventh assumption is valid if inequality (43) with $\delta_0 < 1$ is satisfied.

5. Numerical example

For an example we choose the plasma PDAE again. Let the parameters in B and $C[u]$ be defined by $b_0 = 0.02$ and $d_1 = 1$ (see (5)). First, according to (2) and (3) IVs and (Dirichlet) BVs must be specified. Since $\nu_t = 1$, both terms Φ_a, Φ_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). Furthermore, it can be shown that $u_3(0, 0)$ can be chosen arbitrarily. Let $u_3(0, 0) \neq 0$. Given this BV and

$$g_4(x) := u_4(0, x) = K_4 \cos(2\pi x)$$

(where the constant K_4 is defined below), it follows from the third equation of the system (1) that the consistent IV of u_3 is

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

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

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

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

$$g_2(x) := u_2(0,x) = K_2 x(x-0.5), \quad K_2 = \text{const},$$

and BVs are for $t \in [0, t_e]$

$$u_1(t,0) = u_1(t,1) = u_2(t,0) = 0, \qquad u_3(t,0) = u_3(0,0), \qquad u_4(t,0) = u_4(t,1) = K_4.$$

According to relation (4), the BVs for u_1 imply $K_4 = -u_3(0,0)/(4\pi^2)$. With these relations the IBVP (1)–(4) with matrices (5) is defined completely, and we may solve it by a finite difference method. As mentioned, 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 $t_e = 1$ and the two step sizes τ and h be related by $\tau = \kappa h$ where κ is a positive constant ($\kappa = 0.5$ below). In the example, the CFL condition for the second equation is satisfied for the data chosen because

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

is much smaller than 1. Some values of this expression are given in Table 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

Table 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

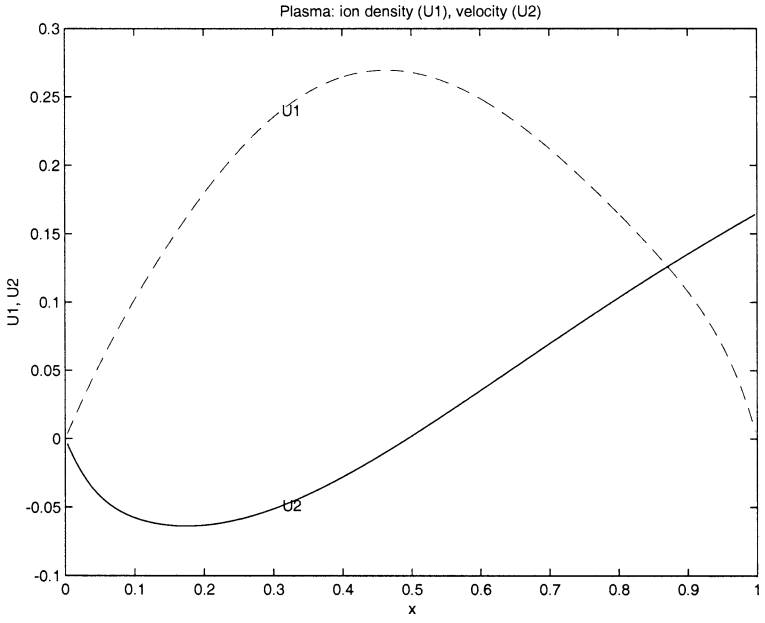

Fig. 1.

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

6. Conclusion

After an application of PDAEs with a nonlinear convection term 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., [14]. 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 a plasma PDAE (a system with time index 1) this was shown in detail.

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 of the space step size $h \to 0$ 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, L.R. Petzold, Numerical Solution of Initial-Value Problems in Differential–Algebraic Equations, North-Holland, Amsterdam, 1989.
[2] S.L. Campbell, W. Marszalek, ODE/DAE integrators and MOL problems, Z. Angew. Math. Mech. (1996) 251–254; also in: ICIAM '95 Minisymposium on MOL, 1995.
[3] S.L. Campbell, W. Marszalek, The index of an infinite dimensional implicit system, Math. Mod. Syst. 1(1) (1996) 1–25.
[4] S.L. Campbell, W. Marszalek, DAEs arising from traveling wave solutions of PDEs, J. Comput. Appl. Math. 82 (1–2) (1997) 41–58.
[5] S.C. Cowin, Bone poroelasticity, J. Biomech. 32 (1999) 217–238.
[6] E. Detournay, H.-D.A. Cheng, Fundamentals of poroelasticity, in: J.A. Hudson (Ed.), Comprehensive Rock Engineering: Principles, Practice and Projects, Pergamon, Oxford, 1993.
[7] R.K. Dodd, J.C. Eilbeck, J.D. Gibbon, H.C. Morris, Solitons and Nonlinear Wave Equations, Academic Press, New York, 1982.
[8] M. Günther, Y. Wagner, Index concepts for linear mixed systems of differential-algebraic and hyperbolic-type equations, SIAM J. Sci. Statist. Comput. 22 (2000) 1610–1629.
[9] N.N. Janenko, Die Zwischenschrittmethode zur Lösung mehrdimensionaler Probleme der mathematischen Physik, Springer, Berlin, 1969.
[10] W. Lucht, K. Debrabant, Models of quasi-linear PDAEs with convection, Technical Report, Martin-Luther-Universität Halle, Fachbereich Mathematik und Informatik, 2000.
[11] W. Lucht, K. Strehmel, Discretization based indices for semilinear partial differential algebraic equations, Appl. Numer. Math. 28 (1998) 371–386.
[12] W. Lucht, K. Strehmel, C. Eichler-Liebenow, Indexes and special discretization methods for linear partial differential algebraic equations, BIT 39(3) (1999) 484–512.

[13] W. Marszalek, Analysis of partial differential algebraic equations, Ph.D. Thesis, North Carolina State University, Raleigh, 1997.
[14] W.S. Martinson, P.I. Barton, A differentiation index for partial differential equations, SIAM J. Sci. Comput. 21 (6) (2000) 2295–2315.