Completion to Involution and Semi-Discretisations

Werner M. Seiler [1]

Lehrstuhl für Mathematik I, Universität Mannheim
68131 Mannheim, Germany [2]

Abstract

We discuss the relation between the completion to involution of linear over-determined systems of partial differential equations with constant coefficients and the properties of differential algebraic equations obtained by their semi-discretisation.

For a certain class of "weakly over-determined" systems, we show that the differential algebraic equations do not contain hidden constraints, if and only if the original partial differential system is involutive. We also demonstrate how the formal theory can be used to obtain an existence and uniqueness theorem for smooth solutions of strongly hyperbolic systems and to estimate the drift off the constraints, if an underlying equation is numerically solved.

Finally, we show for general linear systems how the index of differential algebraic equations obtained by semi-discretisations can be predicted from the result of a completion of the partial differential system.

Key words: partial differential equation, differential algebraic equation, index, involution, completion, semi-discretisation, existence and uniqueness of solutions
1991 MSC: 34A09, 35L45, 35N05, 65M20

1 Introduction

Differential algebraic equations have been studied intensively for more than twenty years. Only comparatively recently, attempts have been made to extend the theory to partial differential equations where new phenomena appear. One is that the interesting case is less that the given system contains equations of

[1] Work supported by Deutsche Forschungsgemeinschaft.

[2] E-mail: `werner.seiler@math.uni-mannheim.de`
WWW: `http://www.math.uni-mannheim.de/~wms`

different order but that it cannot be written in *Cauchy-Kovalevskaya form*, i. e. there does not exist a distinguished variable t such that each equation can be solved for a t-derivative.

Such systems appear in many places; classical examples are Maxwell's equations or the incompressible Navier-Stokes equations. For these special cases an extensive theory exists. It is hardly known that almost a century ago already several fairly general theories have been developed that can handle arbitrary systems of partial differential equations. Of course, such general theories never yield, say, existence and uniqueness results at the same level of a specialised theory covering only, for example, the Navier-Stokes equations. Nevertheless, these general theories allow us to derive some interesting results, as we hope to show in this article.

We will use the so-called *formal theory of differential equations* with its basic notion of an *involutive* system. For lack of space we cannot give a detailed introduction into this concept but must refer to the literature [1,12,15]. In an earlier article [16] we have shown how the formal theory can be used to define differentiation indices in a manner that is applicable to both partial and ordinary differential equations. Now we will show how the there defined *involution index* relates to the indices [3] of differential algebraic equations obtained by semi-discretisation of the original partial differential system.

Involution has two ingredients. The first one is simple. An involutive system is consistent; no integrability conditions exist. Neither by differentiating lower order equations nor by cross-differentiating equations of the same order it is possible to generate new equations either of the order of the system or of lower order. In the context of differential algebraic equations this means that there are no hidden constraints. This implies that the system is *formally integrable*: it possesses at least formal power series solutions.

For ordinary differential equations formal integrability is equivalent to involution and it is (at least for linear systems) no problem to check whether or not a given system is formally integrable. For partial differential equations the situation is more involved and we need a second ingredient: involution of the symbol. This is a combinatorial concept which we cannot fully explain here (for a simple introduction see [1]).

The idea is roughly the following one. Given a linear first order system for some functions $\vec{u}(x_1, \ldots, x_n)$, we solve as many equations as possible for an x_n-derivative. Of the remaining ones we solve as many as possible for an x_{n-1}-derivative etc. This leads to a *triangular form* of the system. For equations

[3] For a differential algebraic equation the involution index is simply the number of differentiations needed until no more hidden constraints appear. This is equivalent to the strangeness index [8] or the geometric index definitions in [13,14].

solved for an x_i-derivative, we call the variables $x_1, \ldots, x_i$ *multiplicative*, the remaining ones *non-multiplicative*. For an involutive system we require that differentiating an equation with respect to one of its non-multiplicative variables must lead to an equation which can be written as a linear combination of the original equations or of equations obtained by differentiation with respect to a multiplicative variable. If this holds, we say that the system is in *Cartan normal form*.

The article splits into two parts. In the first part – comprising the Sects. 2 to 4 – we study a class of linear systems which one might call "weakly over-determined", as their triangular form contains only equations solved for either x_n- or x_{n-1}-derivatives. The second part consists only of Sect. 5 and treats arbitrary linear systems.

In the first part we analyse explicitly the conditions under which such a "weakly over-determined" system is involutive. Then we use the results to prove – under some additional assumptions on the structure of the system – an existence and uniqueness theorem for smooth solutions. The proof consists essentially of lifting a classical result for hyperbolic systems in Cauchy-Kovalevskaya form to over-determined systems. For strongly hyperbolic systems we can furthermore estimate the drift off the constraints, if we simply solve an underlying equation by some standard techniques. Finally, we consider semi-discretisations both by finite difference and by spectral methods and show that they yield formally integrable differential algebraic equations, if and only if the original system was involutive.

In the second part we remove the restriction to "weakly over-determined" systems and study the differential algebraic equations one obtains by semi-discretising arbitrary linear systems of partial differential equations. It is proved that completing the partial differential system to involution allows us to predict the indices of these differential algebraic equations.

In the first part all calculations are performed in a very explicit manner and hardly any knowledge of the formal theory is required. In the second part, the proof of the main result can probably be fully understood only with some familiarity with the formal theory.

2 A Class of Linear Over-determined Systems

In the first part we will study the following class of linear over-determined systems of partial differential equations with constant coefficients:

$$
\begin{aligned}
E\vec{u}_t &= \sum_{i=1}^{n} A_i \vec{u}_{x_i} + B\vec{u}\,, && (1)\\
0 &= \sum_{i=1}^{n} C_i \vec{u}_{x_i} + D\vec{u}\,. && (2)
\end{aligned}
$$

Here $\vec{u}$ is an m-dimensional vector of dependent variables and we have $n+1$ independent variables $\vec{x} = (x_1, \dots, x_n)$, t. For the formal analysis, we count t as x_{n+1}. The matrices E, A_i and B have q rows and m columns; the matrices C_i and D have p rows. In the sequel we will make three assumptions that are crucial for the analysis:

(a) The hypersurface $t = 0$ is non-characteristic.
(b) rank $E = q$.
(c) For at least one i we have rank $C_i = p$. Without loss of generality we assume that this is the case for C_n.

(a) explicitly excludes parabolic systems. We do this in order to facilitate the formal analysis. There is, however, no principal difficulty in analysing such systems. The other two assumptions are made in order to avoid distinguishing degenerate cases. (b) implies $q \leq m$; if $q < m$, we are dealing with an under-determined system. (c) restricts to "weakly over-determined" systems.

We impose on the system (1, 2) some initial conditions

$$\vec{u}(\vec{x}, 0) = \vec{f}(\vec{x}) \tag{3}$$

and either no or periodic boundary conditions. Due to the over-determinacy of (1, 2), we cannot prescribe arbitrary initial data $\vec{f}$ but they must satisfy some consistency conditions. For the purposes of this article, it is not necessary to specify these explicitly, but it suffices to say that they are straightforwardly determined during the formal analysis of the system. As we impose initial conditions, we implicitly assume that the subsystem (1) is in fact hyperbolic in t-direction, although this is of no relevance for the formal analysis.

Example 1 *A classical system within our class are* Maxwell's equations

$$
\begin{aligned}
\vec{E}_t &= \nabla \times \vec{B}\,, \quad \vec{B}_t = -\nabla \times \vec{E}\,, && (4)\\
\nabla \cdot \vec{E} &= 0\,, \quad \nabla \cdot \vec{B} = 0\,. && (5)
\end{aligned}
$$

Here $q = m = 6$, $p = 2$ and we have set all constants to one. It is well-known that (4) is indeed a hyperbolic system whereas the constraint equations (5) form an elliptic system.

Now we determine the conditions on the coefficient matrices of (1, 2) for the system being involutive. Because of Assumption (c), it suffices to analyse what happens, if (2) is differentiated with respect to its non-multiplicative variable t, as for (1) all variables are multiplicative whereas for (2) only the spatial variables are multiplicative. For an involutive symbol we must require that this does not yield any new second order equations; for a formally integrable system it may not lead to any new equations at all.

An obviously necessary condition for an involutive symbol is the existence of $p \times q$ matrices M_i such that

$$M_i E = C_i \,, \tag{6}$$

as otherwise we cannot eliminate all derivatives $\vec{u}_{x_i t}$. But this elimination uses equations obtained by differentiating (1) with respect to the spatial variables and thus introduces second order derivatives of the form $\vec{u}_{x_i x_j}$. They must be expressible as linear combinations of equations obtained by differentiating (2) with respect to the spatial variables. This yields as further necessary condition the existence of $p \times p$ matrices H_i such that

$$H_i C_j + H_j C_i = M_i A_j + M_j A_i \,. \tag{7}$$

It is easy to see that (6) and (7) are also sufficient for involution of the symbol.

For an involutive system, it must in addition be possible to eliminate all the first order derivatives of $\vec{u}$ introduced by the above operations using (1, 2). In a straightforward computation one obtains the following two necessary and sufficient conditions. There must exist a $p \times q$ matrix N such that

$$NE = D \tag{8}$$

and there must exist a $p \times p$ matrix K satisfying

$$H_i D + K C_i = M_i B + N A_i \,, \tag{9}$$

$$KD = NB \,. \tag{10}$$

Because of Assumption (b), rank $E = q$, and (c), rank $C_n = p$, it is not difficult to see that if such matrices M_i, H_i, N and K exist, then they are uniquely determined by the equations (6–10).

Finally, we derive for later use the so-called *compatibility conditions* of (1, 2). For this we assume that the system is involutive and add some perturbations $\vec{\delta}(\vec{x}, t)$, $\vec{\epsilon}(\vec{x}, t)$ of appropriate dimensions on the right hand side:

$$E\vec{u}_t = \sum_{i=1}^{n} A_i \vec{u}_{x_i} + B\vec{u} + \vec{\delta}\,, \tag{11}$$

$$0 = \sum_{i=1}^{n} C_i \vec{u}_{x_i} + D\vec{u} + \vec{\epsilon}. \tag{12}$$

Redoing the involution analysis with this perturbed system, we see that a necessary condition for the existence of solutions is that

$$\vec{\epsilon}_t - \sum_{i=1}^{n} H_i \vec{\epsilon}_{x_i} - K\vec{\epsilon} = \sum_{i=1}^{n} M_i \vec{\delta}_{x_i} + N\vec{\delta}\,. \tag{13}$$

Because of the assumed involution of the system, this is – at the level of *formal* solutions – also a sufficient condition.

Example 2 *It is easy to check that Maxwell's equations (4, 5) are involutive. For the compatibility conditions we consider the inhomogeneous equations*

$$\vec{E}_t - \nabla \times \vec{B} = \vec{J}_E\,, \quad \vec{B}_t + \nabla \times \vec{E} = \vec{J}_B\,, \tag{14}$$

$$\nabla \cdot \vec{E} = \rho_E\,, \quad \nabla \cdot \vec{B} = \rho_B\,. \tag{15}$$

Here $\vec{J}_E$, $\vec{J}_B$ may be interpreted as external currents and ρ_E, ρ_B as charge densities. The compatibility conditions (13) take now the form

$$\partial_t \rho_E - \nabla \cdot \vec{J}_E = 0\,, \quad \partial_t \rho_B - \nabla \cdot \vec{J}_B = 0\,. \tag{16}$$

Of course, these are nothing else but the classical continuity equations expressing charge conservation.

3 The Underlying Equation

Let the system (1, 2) be involutive. If we exclude under-determined systems, we may assume without loss of generality that $E = \mathbb{I}_m$, the $m \times m$ identity matrix, what we will do for this section. In analogy to differential algebraic equations we may consider the subsystem (1) as an *underlying* equation. Using this point of view we will derive first an existence and uniqueness theorem for smooth solutions of the initial value problem for the case that (1) is hyperbolic and then study a simple approach for the numerical solution of (1, 2).

For *analytic* solutions the situation is very simple. If we assume that the initial data $\vec{f}$ in (3) is analytic and consistent, the existence and uniqueness of analytic solutions follows from the Cartan-Kähler Theorem (a generalisation of the well-known Cauchy-Kovalevskaya Theorem to arbitrary involutive systems), as the hypersurface $t = 0$ is non-characteristic by Assumption (a). However, for most applications analyticity is a too strong condition. We show now how formal techniques allow us to "lift" a classical result on smooth solutions of strongly hyperbolic systems to our class of over-determined systems.

By definition, the subsystem (1) is called *hyperbolic* in t-direction, if for any tuple $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ the matrix $A_\xi = \sum_{i=1}^{n} \xi_i A_i$ has only real eigenvalues and an eigenbasis. It is *strongly hyperbolic*, if there exists for any A_ξ a symmetric, positive definite matrix P_ξ, a *symmetriser*, such that $P_\xi A_\xi + A_\xi^t P_\xi = 0$.

Theorem 3 *Let (1) be strongly hyperbolic. Let the initial data in (3) be smooth and such that it satisfies the equations (2) for $t = 0$. Then the initial value problem (1, 2, 3) with periodic boundary conditions possesses a unique smooth solution.*

PROOF. The proof consists of two steps: first we show that the subsystem (1) has a unique smooth solution for the given initial data, then that this solution also satisfies (2). The first part is a classical result in the theory of strongly hyperbolic systems (cf. [7, Theorem 6.2.2]), so that there is nothing to prove.

For the second part, we take this solution of (1) and enter it into the subsystem (2) obtaining some residual $\vec{\epsilon}(\vec{x}, t)$. This residual must satisfy (13) with $\vec{\delta}(\vec{x}, t) \equiv 0$. Furthermore, we have the initial condition $\vec{\epsilon}(0, \vec{x}) = 0$, as by assumption our solution satisfies (2) for $t = 0$. It is obvious that under these conditions $\vec{\epsilon}(\vec{x}, t) \equiv 0$ is a solution of (13) and it follows from Holmgren's Theorem that no other continuously differentiable solution exists. Thus we have indeed a unique smooth solution of the full system (1, 2). $\square$

For Maxwell's equations our proof reduces to a well-known property. If the evolution equations (4) are solved for initial data satisfying (5), the solution fields $\vec{E}(\vec{x}, t)$ and $\vec{B}(\vec{x}, t)$ are divergence free for all times t. As we have seen, this is not a special property of Maxwell's equations, but a general fact of involutive systems which is the core of the proof of the Cartan-Kähler Theorem. The key is that only involution permits us to derive the necessary and sufficient conditions (13).

The idea used above also translates naturally into a simple approach for solving numerically the initial value problem (1, 2, 3). We require that the initial data $\vec{f}$ satisfy the constraint equations (2) and then we "forget" about the latter ones. We integrate numerically (1) which is a standard problem in numerical analysis. In the context of differential algebraic equations this is sometimes called the *method of the underlying equation.*

It is well-known that we must expect a "drift" of the numerical solution, i.e. the numerical solution will start to violate the constraint equations (2). But the above cited existence and uniqueness theorem for strongly hyperbolic systems also provides estimates for the solution of the perturbed system (cf. [7, Lemma 3.3.1]), namely that for $p = 0, 1, 2, \ldots$ and for $t \in [0, T]$

$$||\vec{u}(\cdot, t)||_{H^p} \leq K_p \left[||\vec{f}||_{H^p} + \int_0^t ||\vec{\delta}(\cdot, \tau)||_{H^p} \, d\tau \right] . \tag{17}$$

Here $|| \cdot ||_{H^p}$ is for a symmetric hyperbolic system the usual Sobolev norm of order p and the constant K_p depends only on the coefficients of the system and T. For a strongly hyperbolic system, the norm depends additionally on the symmetriser P_ξ (see [7, p. 187] for details).

The inequality (17) can be exploited to estimate the drift (if we assume that it is smooth). As a preparation, we must show how certain properties of the underlying equation (1) are inherited by the compatibility conditions (13). Recall that (2) is called an *elliptic* system, if the matrix $C_\xi = \sum_{i=1}^n \xi_i C_i$ defines for all tuples $0 \neq \xi \in \mathbb{R}^n$ a surjective mapping. This condition implies that *all* matrices C_i must possess maximal rank: $\operatorname{rank} C_i = p$.

Proposition 4 *If (1) is hyperbolic in t-direction and (2) is elliptic, then the compatibility conditions (13) – considered as a system for $\vec{c}$ only – are also hyperbolic in t-direction.*

PROOF. As we assume that the combined system (1, 2) is involutive, the matrices A_i, C_i, H_i satisfy the relations (7). This implies (with obvious notation) that $H_\xi C_\xi = C_\xi A_\xi$ for any tuple $\xi \in \mathbb{R}^n$. Let $\vec{v}$ be an eigenvector of A_ξ for the (real) eigenvalue λ. As $H_\xi C_\xi \vec{v} = C_\xi A_\xi \vec{v} = \lambda C_\xi \vec{v}$, we see that $C_\xi \vec{v}$ is an eigenvector of H_ξ for the eigenvalue λ.

According to our assumptions, all matrices A_ξ possess only real eigenvalues and an eigenbasis and all matrices C_ξ define surjective mappings. Thus it is possible to obtain an eigenbasis of H_ξ consisting of vectors of the form $C_\xi \vec{v}$ where $\vec{v}$ is an eigenvector of A_ξ. This implies furthermore that all eigenvalues of H_ξ are real, as they are also eigenvalues of A_ξ. □

Proposition 5 *Let the assumptions of Prop. 4 be satisfied and let furthermore (1) be* strongly *hyperbolic with symmetriser P_ξ. Then (13) is strongly hyperbolic, too. As symmetriser we may use the matrix Q_ξ defined by*

$$Q_\xi^{-1} = C_\xi P_\xi^{-1} C_\xi^t . \tag{18}$$

PROOF. We must show that Q_ξ satisfies the three properties of a symmetriser. Obviously, it inherits the symmetry of P_ξ. For the positive definiteness we note that $\vec{v}^t Q_\xi^{-1}\vec{v} = (C_\xi^t\vec{v})^t P_\xi^{-1}(C_\xi^t\vec{v})$. As C_ξ defines a surjective mapping, its transpose C_ξ^t defines an injective mapping. Thus the expression above vanishes only for $\vec{v}=0$. This also implies that Q_ξ^{-1} is indeed invertible. Finally, using again (7)

$$\begin{aligned} H_\xi Q_\xi^{-1} + Q_\xi^{-1}H_\xi^t &= H_\xi C_\xi P_\xi^{-1}C_\xi^t + C_\xi P_\xi^{-1}C_\xi^t H_\xi^t \\ &= C_\xi(A_\xi P_\xi^{-1} + P_\xi^{-1}A_\xi^t)C_\xi^t. \end{aligned} \tag{19}$$

But this last expression vanishes, as P_ξ is a symmetriser for A_ξ. □

Now we apply the estimate (17) to (13). We have seen above that the residual $\vec{\epsilon}$ in the constraint equations (2) satisfies the compatibility conditions (13) with vanishing initial conditions. Thus Prop. 5 yields the following result.

Theorem 6 *Let the subsystem (1) be strongly hyperbolic and (2) elliptic. Let $\vec{v}(\vec{x},t)$ be an approximate solution of (1) with residual $\vec{\delta}(\vec{x},t)$ satisfying the constraint equations (2) for $t=0$. Then the constraint residual $\vec{\epsilon}(\vec{x},t)$ obtained by entering $\vec{v}(\vec{x},t)$ in (2) satisfies*

$$||\vec{\epsilon}(\cdot,t)||_{H^p} \le K_p \int_0^t ||\sum_{i=1}^n M_i\vec{\delta}_{x_i}(\cdot,\tau) + N\vec{\delta}(\cdot,\tau)||_{H^p}\, d\tau\,. \tag{20}$$

We have a kind of H^{p+1} norm of $\vec{\delta}$ on the right hand side (because of the ellipticity none of the matrices M_i vanishes). In order to control the drift, it does not suffice to keep the residual $\vec{\delta}$ small; its spatial derivatives $\vec{\delta}_{x_i}$ enter the estimate as well. Even for an involutive system we are in a similar situation as for a differential algebraic equation with perturbation index 2 and not 1 as one may expect!

Example 7 *For Maxwell's equations the method of the underlying equation is a popular approach to their numerical solution. As the evolution part (4) forms a symmetric hyperbolic system, we can apply the theory developed above. The integrand on the right hand side of (20) is now simply $||\nabla\cdot\vec{\delta}(\cdot,\tau)||_{H^p}$. Thus for the preservation of the constraint equations (5), it is decisive to keep the* divergence *of the residuals as small as possible whereas the size of the residuals themselves is of no importance for the drift.*

Neglecting this effect can lead to numerical problems (see e. g. [6,11] and references therein). One has shown experimentally that in many methods a significant violation of the constraints occurs. Furthermore, it seems that ignoring *the constraint equations yields spurious modes in some computations. The "mimetic" discretisations introduced by Hyman and Shashkov [5] are examples of methods that take care that discrete analogues of the decisive relations are preserved and thus no drift appears.*

4 Semi-Discretisation

4.1 Finite Differences

We discretise the spatial derivatives on a grid where the points $\vec{x}_{\vec{k}}$ are labelled by integer vectors $\vec{k} = [k_1, \dots, k_n]$. $\vec{u}_{\vec{k}}(t)$ denotes the value of the function $\vec{u}$ at the point $\vec{x}_{\vec{k}}$ at time t. We approximate in the system (1, 2) the spatial derivative $\vec{u}_{x_i}(\vec{x}_{\vec{k}}, t)$ by the finite difference

$$\delta_i \vec{u}_{\vec{k}}(t) = \sum_{\ell_i=-a_i}^{b_i} \alpha^{(i)}_{\ell_i} \vec{u}_{[k_1,\dots,k_i+\ell_i,\dots,k_n]}(t) \tag{21}$$

with some real coefficients $\alpha^{(i)}_{\ell_i}$. Thus different discretisations are allowed for different values of i, but $\vec{u}_{x_i}$ is everywhere discretised in the same way.

Entering the approximation (21) into the system (1, 2) yields the following differential algebraic equation:

$$E\dot{\vec{u}}_{\vec{k}} = \sum_{i=1}^{n} A_i \delta_i \vec{u}_{\vec{k}} + B\vec{u}_{\vec{k}}\,, \tag{22}$$

$$0 = \sum_{i=1}^{n} C_i \delta_i \vec{u}_{\vec{k}} + D\vec{u}_{\vec{k}}\,. \tag{23}$$

An obvious question is whether this system is formally integrable, i.e. whether or not differentiating the constraint equations (23) leads to any new equations (differential or algebraic).

Theorem 8 *The differential algebraic equation (22, 23) is formally integrable, if and only if the partial differential system (1, 2) is involutive.*

PROOF. The proof consists of some rather lengthy but straightforward calculations which we can only sketch for lack of space. It is easy to see that no new *differential* equations are generated, if and only if matrices M_i and N satisfying (6) and (8), respectively, exist.

If these conditions are fulfilled, we can eliminate all derivatives from the differentiated equations. Then we must try to express the arising *algebraic* equations

$$\sum_{i=1}^{n}\left[(M_iB+NA_i)\delta_i\vec{u}_{\vec{k}}+\sum_{j=1}^{n}M_iA_j\delta_i\delta_j\vec{u}_{\vec{k}}\right]+B\vec{u}_{\vec{k}}=0 \tag{24}$$

as linear combinations of the equations (23) (taken at different grid points). Thus we are looking for matrices $H_{\ell_i}^{(i)}$ such that the left hand side of (24) can be written in the form [4]

$$\sum_{i=1}^{n}\sum_{\ell_i=-a_i}^{b_i}H_{\ell_i}^{(i)}\left[\sum_{j=1}^{n}C_j\delta_j\vec{u}_{[k_1,\dots,k_i+\ell_i,\dots,k_n]}+D\vec{u}_{[k_1,\dots,k_i+\ell_i,\dots,k_n]}\right]. \tag{25}$$

Analysing point by point, one can show that this is possible, if and only if matrices H_i and K satisfying (7) and (9, 10), respectively, exist. One obtains that $H_\ell^{(i)}=\alpha_\ell^{(i)}H_i$ for $\ell\neq 0$ and $H_0^{(i)}=\alpha_0^{(i)}H_i+K_i$ with any matrices K_i such that $\sum_{i=1}^{n}K_i=K$. This arbitrariness reflects the redundancy in the ansatz (25). □

The Assumptions (b) and (c) on the ranks of the matrices E and C_n are important here. They guarantee the uniqueness of the matrices M_i, N, H_i, K both for the partial differential system and for the differential algebraic equation. Only this uniqueness yields the "if and only if" of the theorem.

Example 9 *The following linear system for two unknowns $v(x,t)$, $w(x,t)$*

$$\begin{pmatrix}v\\w\end{pmatrix}_t=\begin{pmatrix}0&1\\0&0\end{pmatrix}\begin{pmatrix}v\\w\end{pmatrix}_x,\qquad 0=\begin{pmatrix}1&0\end{pmatrix}\begin{pmatrix}v\\w\end{pmatrix}_x \tag{26}$$

is of the form (1, 2). It is easy to see that it is formally integrable, i. e. that no integrability conditions exist, as all terms are of the same order. However, (26) is not involutive, as differentiating the constraint equation with respect to t yields (after simplification) a new second order equation: $w_{xx}=0$.

This effects immediately the semi-discretisation. Using backward differences for the spatial derivatives, we obtain the differential algebraic equation

$$\Delta x\begin{pmatrix}\dot v_n\\\dot w_n\end{pmatrix}=\begin{pmatrix}0&1\\0&0\end{pmatrix}\begin{pmatrix}v_n-v_{n-1}\\w_n-w_{n-1}\end{pmatrix},\quad 0=\begin{pmatrix}1&0\end{pmatrix}\begin{pmatrix}v_n-v_{n-1}\\w_n-w_{n-1}\end{pmatrix}. \tag{27}$$

[4] There is a slight redundancy in this ansatz, as all matrices $H_0^{(i)}$ are multiplied by the same vector and thus only their sum is important.

As predicted by Theorem 8, this system is not formally integrable, as it contains the hidden constraint $w_n - 2w_{n-1} + w_{n-2} = 0$ which is obviously nothing but an approximation of the differential equation $w_{xx} = 0$.

4.2 Spectral Methods

Finite differences are of course not the only possibility for discretising differential equations. If we prescribe periodic boundary conditions, a spectral discretisation with Fourier series also appears very natural. Thus we use the following approximation:

$$\vec{u}(\vec{x},t) \approx \sum_{\vec{k}\in\mathcal{G}} \left[\vec{a}_{\vec{k}}(t) + \mathrm{i}\vec{b}_{\vec{k}}(t)\right] \mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{x}} \ . \tag{28}$$

Here $\mathcal{G}$ represents some finite grid of wave vectors and we have split the complex Fourier coefficients in their real and imaginary part.

The arising differential algebraic equation possesses the following block structure where the vector $\vec{A}$ consists of the matrices A_i

$$\begin{pmatrix} E & 0 \\ 0 & E \end{pmatrix} \begin{pmatrix} \dot{\vec{a}}_{\vec{k}} \\ \dot{\vec{b}}_{\vec{k}} \end{pmatrix} = \begin{pmatrix} B & -\vec{k}\cdot\vec{A} \\ \vec{k}\cdot\vec{A} & B \end{pmatrix} \begin{pmatrix} \vec{a}_{\vec{k}} \\ \vec{b}_{\vec{k}} \end{pmatrix} , \tag{29}$$

$$0 = \begin{pmatrix} D & -\vec{k}\cdot\vec{C} \\ \vec{k}\cdot\vec{C} & D \end{pmatrix} \begin{pmatrix} \vec{a}_{\vec{k}} \\ \vec{b}_{\vec{k}} \end{pmatrix} . \tag{30}$$

As in the last section, we want to know what happens when we differentiate the constraint equations (30). The answer is exactly the same.

Theorem 10 *The differential algebraic equation (29, 30) is formally integrable, if and only if the partial differential system (1, 2) is involutive.*

PROOF. This result is proven in a similar manner as Theorem 8. The condition that the differentiation of (30) does not lead to any new *differential* equations implies that there must exist matrices Λ_i such that

$$\begin{pmatrix} D & -\vec{k}\cdot\vec{C} \\ \vec{k}\cdot\vec{C} & D \end{pmatrix} = \begin{pmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_3 & \Lambda_4 \end{pmatrix} \begin{pmatrix} E & 0 \\ 0 & E \end{pmatrix} . \tag{31}$$

Because of Assumption (b) this system has the unique solution $\Lambda_1 = \Lambda_4 = N$ and $\Lambda_2 = -\Lambda_3 = \vec{k} \cdot \vec{M}$, if there exist matrices M_i and N such that (6) and (8), respectively, are satisfied. Otherwise it has no solution at all.

Similarly, the condition that no new *algebraic* equations arise implies that there must exist matrices Σ_i such that

$$\begin{pmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_3 & \Lambda_4 \end{pmatrix} \begin{pmatrix} B & -\vec{k} \cdot \vec{A} \\ \vec{k} \cdot \vec{A} & B \end{pmatrix} = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_3 & \Sigma_4 \end{pmatrix} \begin{pmatrix} D & -\vec{k} \cdot \vec{C} \\ \vec{k} \cdot \vec{C} & D \end{pmatrix} . \tag{32}$$

Some straightforward calculations show that because of Assumption (c) this system possesses the unique solution $\Sigma_1 = \Sigma_4 = K$ and $\Sigma_2 = -\Sigma_3 = \vec{k} \cdot \vec{H}$ provided there exist matrices H_i and K satisfying (7) and (9, 10), respectively. Otherwise no solution at all exists. □

5 Semi-Discretisation of General Linear Systems

We consider now an arbitrary homogeneous linear first order system with constant coefficients for some unknowns $\vec{u}(\vec{x})$ where $\vec{x}$ is now $(n+1)$-dimensional

$$\sum_{i=1}^{n+1} A_i \vec{u}_{x_i} + B\vec{u} = 0 \, . \tag{33}$$

The matrices A_i and B are here completely arbitrary. Instead of (33) we complete to involution the *perturbed* system

$$\sum_{i=1}^{n+1} A_i \vec{u}_{x_i} + B\vec{u} = \vec{\epsilon}(\vec{x}) \, . \tag{34}$$

This yields in general a system of higher order, but this does not concern us here. However, we will assume that (33) is a solvable system, i.e. that no inconsistencies appear during the completion.

Definition 11 *The* involution index ν_ℓ in direction x_ℓ *of the partial differential system (33) is the maximal number of differentiations with respect to x_ℓ appearing in an $\vec{\epsilon}$-derivative contained in the involutive completion of the perturbed system (34). The* (total) involution index ν *of (33) is the highest order of an $\vec{\epsilon}$-derivative appearing in the completion.*

A more intrinsic definition of the (total) involution index that can also be applied to non-linear systems was given in [16]. If the system (33) is already involutive, all involution indices ν_i vanish. Thus the following result – the main result of this article – comprises Theorem 8 as a special case.

Theorem 12 *Let $x_\ell = 0$ be a non-characteristic surface. The semi-discretisation of the system (33) with respect to x_ℓ leads to a differential algebraic equation with (involution) index ν_ℓ.*

PROOF. In order to simplify the subsequent calculations, we rewrite (33). We denote x_ℓ by t and renumber the remaining independent variables as x_i with $1 \leq i \leq n$. Then we solve as much equations as possible for a t-derivative. This transforms (33) into the form (1, 2). While the Assumptions (a) and (b) still hold, we cannot make any statement about the ranks of the matrices C_i.

For the semi-discretisation we approximate the spatial derivatives $\vec{u}_{x_i}$ in (1, 2) by some finite differences $\delta_i \vec{u}_{\vec{k}}$ and obtain the differential algebraic equation

$$E\dot{\vec{u}}_{\vec{k}} = \sum_{i=1}^{n} A_i \delta_i \vec{u}_{\vec{k}} + B\vec{u}_{\vec{k}}\,, \tag{35}$$

$$0 = \sum_{i=1}^{n} C_i \delta_i \vec{u}_{\vec{k}} + D\vec{u}_{\vec{k}}\,. \tag{36}$$

We must now compare what happens during the completion of (11, 12) and (35, 36), respectively. Determining the involution index ν_ℓ requires keeping track of when the perturbations $\vec{\epsilon}$ are differentiated with respect to t. We analyse the completion process by looking at the so-called derivative arrays (or the prolonged systems in the language of formal theory) of the original system and of the differential algebraic equation.

For the partial differential system the differentiations with respect to t lead to the second order equations

$$E\vec{u}_{tt} = \sum_{i=1}^{n} A_i \vec{u}_{x_i t} + B\vec{u}_t + \vec{\delta}_t\,, \tag{37}$$

$$0 = \sum_{i=1}^{n} C_i \vec{u}_{x_i t} + D\vec{u}_t + \vec{\epsilon}_t\,. \tag{38}$$

As t is non-multiplicative for (2), the equations (38) are decisive for the involution analysis. The spatial differentiations yield for $1 \leq j \leq n$

$$E\vec{u}_{x_j t} = \sum_{i=1}^{n} A_i \vec{u}_{x_i x_j} + B\vec{u}_{x_j} + \vec{\delta}_{x_j}\,, \tag{39}$$

$$0 = \sum_{i=1}^{n} C_i \vec{u}_{x_i x_j} + D\vec{u}_{x_j} + \vec{\epsilon}_{x_j}\,. \tag{40}$$

All equations in (39) are the result of differentiations with respect to multiplicative variables. In (40) we cannot make a similar statement, as we have not transformed our original system into a full triangular form and thus do not know which variables are multiplicative for which equation. However, this information is not needed later.

Similarly, we differentiate the differential algebraic equation (35, 36)

$$E\ddot{\vec{u}}_{\vec{k}} = \sum_{i=1}^{n} A_i \delta_i \dot{\vec{u}}_{\vec{k}} + B\dot{\vec{u}}_{\vec{k}}\,, \tag{41}$$

$$0 = \sum_{i=1}^{n} C_i \delta_i \dot{\vec{u}}_{\vec{k}} + D\dot{\vec{u}}_{\vec{k}}\,. \tag{42}$$

In addition, we write down – in formal analogy to the equations (39, 40) obtained by spatial differentiations – the following equations obtained by applying the difference operators δ_j for $1 \le j \le n$ to (35, 36)

$$E\delta_j \dot{\vec{u}}_{\vec{k}} = \sum_{i=1}^{n} A_i \delta_i \delta_j \vec{u}_{\vec{k}} + B\delta_j \vec{u}_{\vec{k}}\,, \tag{43}$$

$$0 = \sum_{i=1}^{n} C_i \delta_i \delta_j \vec{u}_{\vec{k}} + D\delta_j \vec{u}_{\vec{k}}\,. \tag{44}$$

They are obviously just linear combinations of the equations (35, 36); however, their addition simplifies the subsequent analysis.

Given all these equations, we can analyse where either integrability conditions or obstructions to involution might occur. We distinguish two cases. Case I analyses (38) and (42), respectively. New equations arise whenever it is not possible to express all equations in (38) (or (42), respectively) as linear combinations of the equations in (39) (or (43), respectively). Obviously, the number of new equations is the same in both systems in view of their identical structure and those for the differential algebraic equation will be semi-discretisations of those for the original partial differential system. For the partial differential system each of them will contain a term $\vec{\epsilon}_t$. In the case of the differential algebraic equation all new equations are integrability conditions; for the partial differential system some of them might represent obstructions to involution, namely those of second order.

Case II occurs only for the partial differential system: integrability conditions or obstructions to involution could arise *within* the subsystem (40). There is no simple criterion for this, as we have not transformed to a full triangular form. But this is not important for us, as all these new equations contain only spatial derivatives of the perturbations and thus do not affect our theorem. There are no corresponding equations for the differential algebraic equation, as (43, 44) are algebraically dependent of the original system.

If Case I does not lead to any new equations, no hidden constraints exist in the differential algebraic equation. Thus it is formally integrable and of index 0. The partial differential system is not necessarily involutive, as Case II might still lead to new equations. Our theorem would be false, if differentiating these with respect to the non-multiplicative variable t yielded again new equations. However, this cannot happen: by interchanging the order of differentiations one can see that any equation arising this way can also be generated by combining spatial derivatives of equations stemming from Case I. [5] This implies that no additional differentiations with respect to t of the perturbations are performed – no matter how many further iterations within Case II may be necessary. Hence the involution index ν_ℓ vanishes in accordance with our claim.

If Case I yields any new equations, both the index of the differential algebraic equation and the involution index ν_ℓ are at least 1. We may then continue in precisely the same manner. For the partial differential system we take all equations in (37–40) and first differentiate them with respect to t. Then we differentiate them with respect to the spatial variables x_k. Similarly, we take the equations in (41–44) and differentiate them with respect to t. Finally, we consider what we get by applying the difference operators δ_k to them.

The analysis of these third order systems proceeds as above. We distinguish the same two cases: Case I involves time derivatives and thus adds *one* differentiation with respect to t to the perturbations; Case II concerns only the partial differential system and does not lead to any t-differentiations. New equations arising in Case II do not lead at later stages of the completion process to any additional t-differentiations of the perturbations, as we may always interchange the order of differentiations and thus all the differentiations with respect to t have already been studied at some earlier stage in Case I.

In this manner we proceed to higher order systems until no new equations appear. It is a clear that at stage k new equations arise only out of the new equations of stage $k-1$, so that those of Case I in stage k contain derivatives of the perturbations with precisely k differentiations with respect to t. Thus we may conclude that the (involution) index of the differential algebraic equation (35, 36) is indeed equal to the maximal number of differentiations with respect to t applied to the perturbations $\vec{\epsilon}$. But this is just the definition of the involution index ν_ℓ in direction x_ℓ. $\square$

[5] This is exactly the same argument as in the proof of Theorem 2.14 in [15] where all calculations involved are performed in full detail.

Corollary 13 *Let ν be the involution index of the partial differential system (33) and assume that all coordinate hyperplanes $x_i = 0$ are non-characteristic. Then any differential algebraic equation arising by a semi-discretisation of (33) has an index less than or equal to ν.*

PROOF. We have obviously $\nu_i \leq \nu$ for all i. $\square$

Example 14 *We consider again the system of Example 9. We may write its perturbed form as follows:*

$$v_t - w_x = \gamma\,, \quad w_t = \delta\,, \quad v_x = \epsilon\,. \tag{45}$$

While this system is formally integrable, it becomes involutive only after adding the obstruction

$$w_{xx} = \epsilon_t - \gamma_x\,. \tag{46}$$

We see that in the final involutive system (45, 46) only first order derivatives of the perturbations γ, δ, ϵ appear. Thus the (total) involution index of (45) is one and as both a t- and an x-derivative of perturbations appear, we find that also $\nu_t = \nu_x = 1$. This agrees with the observation above that the semi-discretised system (27) has a hidden constraint.

Example 15 *As a further example we consider a system in three independent variables. Its perturbed form is*

$$v_t = \gamma\,, \quad w_t - v_y - w = \delta\,, \quad v_x - w = \epsilon\,. \tag{47}$$

The completion to involution reveals two integrability conditions but no obstructions to involution

$$v_y + w = \gamma_x - \epsilon_t - \delta\,, \quad w_x + w_y = \gamma_{xx} - \epsilon_{tx} - \delta_x - \epsilon_y\,. \tag{48}$$

This implies that the (total) involution index is two, whereas $\nu_t = \nu_y = 1$ and $\nu_x = 2$. And indeed, if we semi-discretise with respect to t using backward differences, we obtain the differential algebraic equation

$$\dot{v}_{i,j} = 0\,, \quad \dot{w}_{i,j} = v_{i,j} - v_{i,j-1} - w_{i,j}\,, \quad v_{i,j} - v_{i-1,j} - w_{i,j} = 0\,. \tag{49}$$

It has one hidden constraint $v_{i,j} - v_{i,j-1} - w_{i,j} = 0$ and thus its index is given by ν_t. This constraint is a discretisation of the first equation in (48); a discretisation of the second one can be constructed by linearly combining all the constraints (assuming periodic boundary conditions).

6 Conclusions

While the concept of formal integrability is natural and familiar already from differential algebraic equations, involution appears at first sight as somewhat strange and artificial. Why should it be important how many multiplicative variables an equation has? We tried to give a partial answer to this question in this article. As the results in Sects. 4 and 5 demonstrate, a semi-discretisation "feels" whether the original system was involutive or merely formally integrable: obstructions to involution turn into integrability conditions upon semi-discretisation.

The idea to define "directional" indices for general systems of partial differential equations by considering differential algebraic equations that can be derived by various methods like semi-discretisations, integral transforms etc. is not new. In the contrary, most index definitions for partial differential systems that can be found in the literature (see e. g. [3,9,10]) are based on this simple and fairly obvious idea.

The main difference between these works and our results is that we compute our indices directly from the *partial* differential system and not from any derived *ordinary* differential system. This implies that the results depend only on the system and not in addition on a specific method of treating it. However, we have seen in Theorem 12 that for the linear constant coefficient systems considered in this article the two approaches yield the same results.

As we have shown in [16], the formal theory can also be applied to non-linear systems. Here it is not difficult to predict that different semi-discretisation methods will in general lead to different index definitions and thus do not provide intrinsic information about the structure of the partial differential system. This does not imply that such indices are useless. In the contrary, comparing them with intrinsic indices like the involution index should be important for analysing the appropriateness of a given numerical approach. If the two indices differ, the numerical approximation has probably destroyed some structural properties of the partial differential system.

We considered in this article only problems with no or periodic boundary conditions. This is of course not realistic for most applications. Formal theory cannot help us in analysing the effect of boundary conditions in a semi-discretisation, as this depends on the precise way the conditions are implemented in the discretised system. Thus it is not an intrinsic property of the partial differential system. But as above one can probably say that a change of the index of the differential algebraic equation due to the implementation of the boundary conditions is a bad sign, as it indicates a structural change (foremost in the number of degrees of freedom of the system).

Finally, we want to comment on the extension of our results to more general systems. The obvious next class are linear systems with variable coefficients. Here a new problem arises: differential and difference operators possess different commutation rules with functions. Whereas our proof of Theorem 12 implicitly also showed that the semi-discretisation of a solvable partial differential system leads to solvable differential algebraic equations, the situation is less clear for systems with variable coefficients. One can easily construct examples of semi-discretisations where the algebraic equations are inconsistent. However, it seems that the inconsistencies appear only at an order that is higher than the order of the discretisation scheme. So it is unclear whether these inconsistencies pose any real problems.

For linear systems it makes actually no difference whether one completes the system (33) or its perturbed form (34). The perturbations do not affect the completion but serve only as a device for book-keeping. For non-linear systems this is no longer true: the perturbations can have a profound effect on the completion procedure. As we will show elsewhere [4], this is the deeper reason why one needs such concepts as the uniform differentiation index, if one wants to estimate the perturbation index by a differentiation index [2].

References

[1] J. Calmet, M. Hausdorf, and W.M. Seiler. A constructive introduction to involution. In *Proc. Int. Symp. Applications of Computer Algebra – ISACA '2000*. World Scientific, Singapore, to appear.

[2] S.L. Campbell and C.W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

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

[4] M. Hausdorf and W.M. Seiler. Perturbation versus differentiation indices. Preprint Universität Mannheim, 2001.

[5] J.M. Hyman and M. Shashkov. Mimetic discretizations for Maxwell's equations. *J. Comp. Phys.*, 151:881–909, 1999.

[6] B.N. Jiang, J. Wu, and L.A. Povelli. The origin of spurious solutions in computational electrodynamics. *J. Comp. Phys.*, 125:104–123, 1996.

[7] H.-O. Kreiss and J. Lorenz. *Initial-Boundary Value Problems and the Navier-Stokes Equations.* Pure and Applied Mathematics 136. Academic Press, Boston, 1989.

[8] P. Kunkel and V. Mehrmann. A new look at pencils of matrix valued functions. *Lin. Alg. Appl.*, 212/213:215–248, 1994.

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

[10] W. Lucht, K. Strehmel, and C. Eichler-Liebenow. Indexes and special discretization methods for linear partial differential algebraic equations. *BIT*, 39:484–512, 1999.

[11] C.D. Munz, P. Omnes, R. Schneider, E. Sonnendrücker, and U. Voß. Divergence correction techniques for Maxwell solvers based on a hyperbolic model. *J. Comp. Phys.*, 161:484–511, 2000.

[12] J.F. Pommaret. *Systems of Partial Differential Equations and Lie Pseudogroups*. Gordon & Breach, London, 1978.

[13] P.J. Rabier and W.C. Rheinboldt. A geometric treatment of implicit differential algebraic equations. *J. Diff. Eq.*, 109:110–146, 1994.

[14] S. Reich. On an existence and uniqueness theory for nonlinear differential-algebraic equations. *Circ. Sys. Sig. Proc.*, 10:343–359, 1991.

[15] W.M. Seiler. *Analysis and Application of the Formal Theory of Partial Differential Equations*. PhD thesis, School of Physics and Materials, Lancaster University, 1994.

[16] W.M. Seiler. Indices and solvability for general systems of differential equations. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC '99*, pages 365–385. Springer, Berlin, 1999.