Helsinki University of Technology Institute of Mathematics Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 2001 A438

THE COMPLETE FORM OF A DIFFERENTIAL ALGEBRAIC EQUATION

Teijo Arponen

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics

Teijo Arponen: *The complete form of a differential algebraic equation*; Helsinki University of Technology Institute of Mathematics Research Reports A438 (2001).

Abstract: *We introduce a method to examine the structure of polynomial higher index differential algebraic equations (DAE), resulting in a so called complete form of the DAE. We argue that this approach reveals the structure of a DAE and is therefore an essential tool for handling higher index DAEs. Relations to other methods presented in literature are discussed.*
This work is continuation to earlier work by Tuomela and the present author [TA00], which is a geometrical approach based on the ideas of formal theory of partial differential equations. In this paper we construct an algebraic approach to DAEs, compatible with our geometrical one.

AMS subject classifications: 34A09,58A20

Keywords: higher index, overdetermined differential equations,index reduction, ideal decomposition, jet space

Teijo.Arponen@hut.fi

ISBN 951-22-5542-1
ISSN 0784-3143
Espoo, 2001

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics
P.O. Box 1100, 02015 HUT, Finland
email:math@hut.fi http://www.math.hut.fi/

1 Introduction

The differential algebraic equations (DAEs) are well known to be important in many contexts in engineering sciences, for example multibody dynamics, robotics and electric circuits. Their numerical solution has been extensively studied, beginning from [Gea71]. Some surveys are [BCP89, HW91, Mär92, AP98].

Several concepts of *index* have been developed to describe the structure of a DAE. The philosophy behind indeces is: "the higher the index, the more difficult to solve numerically". Usually the words 'higher index' refer to the probably most popular concept of an index: the differential index. Recent surveys of different indeces are [CG95a, Sei99].

On the other hand, the concept of *an involutive form (or involutivity)* of a system has several definitions which are more or less equivalent. Some relations between different definitions of involutiveness are studied in [Man96, Sei99]. In the formal theory of PDEs the involutivity of a system is a key concept. The philosophy behind involutiveness is: "all relevant information is explicitly visible". Now involutivity is defined in a geometrical way and a natural question arises: is there an algebraic, equivalent concept?

The purpose of this paper is to investigate more closely the (algebraic) structure of the DAE in the case where equations are multivariate polynomials. The structure is revealed by an algorithm whose output, called a complete form of the DAE, defines an algebraic counterpart for involutivity.

In case of a polynomial system it is known [Pom83, prop. 4.34] that involutivity implies the system to be a prime differential ideal (see remark 3.3). However, it is not clear if the converse holds, that is, is a prime differential system also involutive in the geometrical sense?

Therefore we cannot directly give an algebraic definition of involutivity of a system. Instead, we will construct an algebraic counterpart of "an involutive form" (and call it "complete form") compatible with the geometric definition presented in [TA00]. It turns out that in this polynomial case we can loosen our restrictions on f and give more detailed information on the structure of f.

For example, the phenomenon "index depends on the solution" (see e.g. [AP98]) is explained by ideal decomposition and becomes "index depends on the prime ideal the solution is in" (see example 3.6). We will also look at relations to other approaches in literature. In [TA00] relations to the *formal theory of partial differential equations* were established.

The paper is organized as follows: in section 2 we review the approach of [TA00] and specialize to the case where f is a set of polynomials, and see where the usual way of finding "an involutive form" of the system under consideration needs revision because of our definition of solution. In section 3 we recall some necessary algebraic preliminaries and give the new definition with an algorithm to compute it. In section 4 we briefly look at relations to other approaches. Finally in section 5 are some conclusions and comments.

2 Background

2.1 Review of geometric approach to DAEs

We will briefly review (about the first half of) the article [TA00], to which we refer for details and rigorous definitions. In that paper we considered ordinary differential equations of the form

$$\begin{cases} f^1(t, y, y_1, y_2, \dots, y_q) & = 0 \\ f^2(t, y, y_1, y_2, \dots, y_q) & = 0 \\ & \vdots \\ f^k(t, y, y_1, y_2, \dots, y_q) & = 0 \end{cases} \tag{1}$$

where $k \geq n$, each f^i is a smooth function, y is the n-vector of variables, subscripts denote derivatives. Especially, note that we allow $k > n$ which is sometimes called an overdetermined equation. Also, geometrically there is no distinction between ordinary differential equation and differential-algebraic equation, for reasons explained in [TA00, remark 3.6].

The locus of (1) is interpreted as a subset of $J_q(\mathbb{R} \times \mathbb{R}^n)$, a qth order jet space over $\mathbb{R} \times \mathbb{R}^n$. Then, the relevant equations deduced from (1) by differentiation and/or elimination, are those which define the locus as small as possible. Now the system (1) is defined to be involutive (or an involutive form) if it is a complete set of relevant equations. As a trivial example, consider

$$\begin{cases} y^1_1 - 1 & = 0 \\ y^2 - 7 & = 0 \end{cases} \tag{2}$$

whose locus is $\{(t, y^1, 7, 1, y^2_1) \in J_2(\mathbb{R} \times \mathbb{R}^2) \,|\, t, y^1, y^2_1 \in \mathbb{R}\}$. But, from $y^2 - 7 = 0$ follows $y^2_1 = 0$ which appended to (2) gives the locus $\{(t, y^1, 7, 1, 0) \in J_2(\mathbb{R} \times \mathbb{R}^2) \,|\, t, y^1 \in \mathbb{R}\}$ which is clearly smaller, hence $y^2_1 = 0$ is a relevant equation. Also, there are no other relevant equations. Hence an involutive form of (2) is

$$\begin{cases} y^1_1 - 1 & = 0 \\ y^2 - 7 & = 0 \\ y^2_1 & = 0. \end{cases} \tag{3}$$

In conventional DAE analysis (see e.g. [BCP89, HLR89]), it is customary to consider only first order equations

$$f(t, y, y') = 0.$$

This is because by introducing more variables one can transform a higher order equation to a first order one. However, we find it more convenient to consider equations in the form of (1), mainly for the following two reasons:

1. We want to keep n, the number of y-variables, as small as possible. In our article [TA00] it is shown that this reduces the cost of computation.

For, if we transformed (1) to a first order equation the number of y-variables would be increased from n to nq, which increases the cost of computation.

2. It is 'common folklore' that the highest derivatives decide the behaviour of the system, hence it is unillustrative to 'lose' those highest derivatives by lowering the order.

2.2 Solution of an (involutive) differential equation and its numerical computation

This section is very brief since, in this paper, we are not focusing on the numerical solution. We suppose that our equation $f = 0$ is involutive in the geometrical sense of [TA00]. Conventionally, solution is defined as a function $\phi : I \to \mathbb{R}^n$ s.t. $f(t, \phi(t), \phi'(t), \ldots, \phi^{(q)}(t)) = 0 \quad \forall t \in I$, an open subset of $\mathbb{R}$. However, we use a geometrical definition:

Denote $M := f^{-1}(0) \subset J_q$. On M we define a distribution

$$D_p := TM_p \cap \mathcal{C}_p$$

where $\mathcal{C}_p$ is the Cartan distribution at $p \in M$ and TM_p is the tangent plane at p. It is a well known fact from differential geometry that a one-dimensional distribution has an integral manifold, which then is a smooth curve.

Definition 2.1. if D is one-dimensional, the integral manifold of D through $p \in M$ is the solution of $f = 0$ at p.

If a solution function ϕ exists, the curve $(t, \phi(t), \phi'(t), \ldots, \phi^{(q)}(t))$ (which is also known as the lift of ϕ to J_q) is a geometrical solution. The converse does not hold in general, as the simple examples in [TA00, §2] show. Hence this geometrical solution is more general than conventional one.

In [TA00] an algorithm is described to solve $f = 0$ numerically. The algorithm is a nonlinear (low order) Runge-Kutta method: traditional Runge-Kutta equipped with certain orthogonal projections in the jet space to the locus $f = 0$. In [TA01] the theory of this method is extended to fourth order.

Remark 2.1. This Runge-Kutta with projections is *not* the same as the "projected Runge-Kutta" mentioned in [AP98] and [HW91]. Also, the concept of "solution manifold" is different: in [TA00] it refers to a subset of the jet space $J_q(\mathcal{E})$ where $\mathcal{E}$ is the $(t, y)-$space, while in most DAE literature it refers to a subset of $\mathcal{E}$.

2.3 The polynomial case

The DAE we are considering in this paper is as in (1) but now each f^i is a multivariate polynomial. We will continue to use a shorthand notation $f = 0$ for (1). It is well known that in this case the system is interpreted as a differential ideal (see remark 3.3) generated by $f^1, \ldots, f^s$.

We shall describe the Cartan-Kuranishi algorithm. First, a notation:

$$y_{\leq q} := (t, y, y_1, \ldots, y_q)$$

Step 1, prolongation. (differentiation)

Since $y = y(t)$ and $y_j = y^{(j)}(t)$ for all j, we have $\frac{df}{dt} = 0$. On the other hand,

$$\frac{df}{dt} = B(y_{\leq q})y_{q+1} + \tilde{f}(y_{\leq q}) \tag{4}$$

where

$$B = \begin{pmatrix} \frac{\partial}{\partial y_q^1} f^1 & \frac{\partial}{\partial y_q^2} f^1 & \ldots & \frac{\partial}{\partial y_q^n} f^1 \\ \vdots & & & \\ \frac{\partial}{\partial y_q^1} f^k & \frac{\partial}{\partial y_q^2} f^k & \ldots & \frac{\partial}{\partial y_q^n} f^k \end{pmatrix}, \tag{5}$$

$$\tilde{f} = \begin{pmatrix} \frac{\partial}{\partial t} f^1 & \frac{\partial}{\partial y^1} f^1 & \frac{\partial}{\partial y^2} f^1 & \ldots & \frac{\partial}{\partial y_{q-1}^n} f^1 \\ \vdots & & & & \\ \frac{\partial}{\partial t} f^k & \frac{\partial}{\partial y^1} f^k & \frac{\partial}{\partial y^2} f^k & \ldots & \frac{\partial}{\partial y_{q-1}^n} f^k \end{pmatrix} \cdot \begin{pmatrix} 1 \\ y_1 \\ \vdots \\ y_q \end{pmatrix} \tag{6}$$

Step 2, projection. (elimination)

Supposing $\ker(B^T)$ is constant, find a basis for it, denote it by $\{v^1, \ldots, v^\nu\}$. That is, each v^j is a map

$$v^j : (t, y, y_1, \ldots, y_q) \mapsto \mathbb{R}^k \tag{7}$$

Step 3, test surjectivity.

For $j \in \{1, \ldots, \nu\}$ multiply $\frac{df}{dt}$ from left by v^j:

$$\begin{aligned} 0 &= v^j \frac{df}{dt} \\ &= v^j B\, y_{q+1} + v^j \tilde{f} \\ &= v^j \tilde{f} =: u^j \end{aligned} \tag{8}$$

and check which ones, if any, of these ν equations $u^j = 0$ are algebraically independent of the k equations $f = 0$.

Step 4.

If there were no new equations to step 3, we are done. Otherwise, append the new equations after f and repeat from step 1 with this new f. (end of CK algorithm)

In terminology of the geometric theory of PDEs, step 1 is 'prolongation' (from J_q to J_{q+1}), step 2 is 'projection' (from J_{q+1} to J_q). In steps 3 and 4, the surjectivity of the projection $J_{q+1} \to J_q$ is checked: surjectivity is equivalent with no new equations. In the words of differential algebra, the differential ideal generated by f is the same as the one generated by f, $u^1, \ldots, u^\nu$. The set of equations achieved as an output of this algorithm is called *an involutive form of* f. However, later we shall reconsider this.

Steps 1 to 4 is known as the *Cartan-Kuranishi* algorithm (CK for short), or actually a special case of it: the original algorithm is more complicated and designed for *partial* differential equations. See for example [RLW01, Pom94, Man96] for more information about CK or other equivalent versions called e.g. Ritt-Kolchin or Janet-Spencer.

Example 2.1. This is example 2.3.1 in [TA00, §2.3]. Now $n = q = k = 1$ and $f = \frac{1}{2}(t^2 + y^2 + (y_1)^2 - 1)$, so $\mathcal{V}(f)$ is the unit ball in $J_1 = \mathbb{R}^3$. Now in CK algorithm we have

$$B = y_1, \quad \tilde{f} = t + y\, y_1 \tag{9}$$

and when $y_1 \neq 0$, $\ker(B^T)$ is trivial, hence f is in involutive form. If we continue the algorithm in the case $y_1 = 0$ we get an extra equation $\tilde{f} = 0$ which then becomes $t = 0$. But this leads to a conflict:

$$\begin{cases} f & = 0 \\ y_1 & = 0 \\ \tilde{f} & = 0 \end{cases} \Rightarrow \begin{cases} \frac{1}{2}(y^2 - 1) & = 0 \\ y_1 & = 0 \\ t & = 0 \end{cases} \overset{\delta}{\Rightarrow} \begin{cases} y\, y_1 & = 0 \\ y_2 & = 0 \\ 1 & = 0 \end{cases} \tag{10}$$

hence the "equator" $y_1 = 0$ is forbidden as far as CK algorithm is considered. However, we know from [TA00, §2.3] that also the "equator" $y_1 = 0$ is suitable for our definition of solution. So, we have to reconsider the concept of an involutive form.

Remark 2.2. In other words, in the example above, y_1 is not in the differential ideal generated by f, because $f = 0$ and $y_1 = 0$ led to a conflict. An immediate conclusion to be drawn from this is that, when constructing an algebraic counterpart for our geometrical approach from [TA00], *we cannot use differential ideals*! This is the motivation of the present paper.

3 The new definition

We recall the necessary definitions and results from commutative algebra. Proofs and further information can be found in any textbook on abstract algebra, we recommend [CLO92] and [Eis96]. Let $\mathcal{F}$ be a field and $\mathcal{R}$ a (nontrivial, that is, $0 \neq 1$) polynomial ring in m variables over $\mathcal{F}$:

$$\mathcal{R} = \mathcal{F}[y_1, y_2, \ldots, y_m]$$

3.1 Algebraic preliminaries: rings, ideals, varieties

An *ideal* of $\mathcal{R}$ is a subset $I \subset \mathcal{R}$ satisfying (i) $0 \in I$, (ii) if $f, g \in I$, then $f + g \in I$ (iii) if $f \in I$ and $h \in \mathcal{R}$, then $hf \in I$. Note that $I = \mathcal{R}$ if and only if $1 \in I$. An ideal *generated by* $f_1, \ldots, f_s \in \mathcal{R}$ is the smallest ideal containing $f_1, \ldots, f_s$. It is denoted by

$$\langle f_1, \ldots, f_s \rangle$$

and every element $x \in \langle f_1, \ldots, f_s \rangle$ can be represented by

$$x = \sum_{i=1}^{s} h_i \, f_i \quad \text{with } h_i \in \mathcal{R} \, \forall i.$$

Note that the h_i are non-unique. Every ring has at least two ideals: $\langle 0 \rangle$ and $\langle 1 \rangle$. We shall call both of these *trivial*. An ideal I is

- *maximal* if there is no non-trivial ideal containing it, i.e. if J is an ideal such that $I \subset J \subset \mathcal{R}$ and $I \neq J$, then $J = \mathcal{R}$.
- *prime* if whenever f, $g \in \mathcal{R}$ and $fg \in I$, then either $f \in I$ or $g \in I$.
- *radical* if $f \in \mathcal{R}$ and $f^m \in I$ for any integer $m \geq 1$ implies that $f \in I$.

These properties fulfill:

$$\text{maximal} \;\Rightarrow\; \text{prime} \;\Rightarrow\; \text{radical} \tag{11}$$

Theorem 3.1. For any ideal I, we can define in a natural way the quotient ring $\mathcal{R}/I$ which inherits its ring structure from $\mathcal{R}$. Properties: $\mathcal{R}/I$

- is a field if and only if I is maximal
- has no zero divisors if and only if I is prime
- has no nilpotent elements if and only if I is radical.

A convenient rule of thumb is "the bigger the ideal is, the simpler it is" (with the exception of the trivial 0 ideal). For every ideal I there is a corresponding unique *radical ideal of* I, denoted by $\sqrt{I}$, which is defined by

$$\sqrt{I} := \{f \in \mathcal{R} : f^m \in I \text{ for some integer } m \geq 1\}.$$

A *variety corresponding to* I is "the set of common zeros of elements of I", that is, a subset of $\mathcal{F}^m$:

$$\mathcal{V}(I) := \{(y_1, \ldots, y_m) \in \mathcal{F}^m \,|\, f(y_1, \ldots, y_m) = 0 \quad \forall f \in I\}.$$

If $I = \langle f_1, \ldots, f_s \rangle$ we will also use notation $\mathcal{V}(I) = \mathcal{V}(f_1, \ldots, f_s)$.

Now $\sqrt{I} \supset I$ so according to our rule of thumb above, operating with "$\sqrt{\ }$" means "make the ideal simpler such that its locus is unchanged". For any variety V, there is the *radical ideal corresponding to* V:

$$\mathcal{I}(V) := \{f \in \mathcal{R} \,|\, f(a_1, \ldots, a_n) = 0 \quad \forall (a_1, \ldots, a_n) \in V\}.$$

Definition 3.1. A set A is *quasialgebraic* (q.a. for short) if there exist varieties V and W such that $A = V - W$ where minus denotes the set-theoretic exclusion.

Any variety V is quasialgebraic: $V = V - \mathcal{V}(1)$, since $\mathcal{V}(1) = \varnothing$.

Theorem 3.2. Let I, J be ideals and V, W varieties. Some properties of $\sqrt{\ }$, $\mathcal{V}$ and $\mathcal{I}$:

- $\sqrt{I \cap J} = \sqrt{I} \cap \sqrt{J}$
- $\mathcal{V}(I \cap J) = \mathcal{V}(I) \cup \mathcal{V}(J)$
- $\mathcal{I}(V \cup W) = \mathcal{I}(V) \cap \mathcal{I}(W)$
- if $\mathcal{V}(I)$ is nonempty, then $\mathcal{I}(\mathcal{V}(I)) = \sqrt{I}$

An important tool for us is

Theorem 3.3. Suppose I is a radical ideal in $\mathcal{R}$. Then I can be written as a finite intersection of *prime* ideals:

$$I = I_1 \cap \cdots \cap I_m$$

Moreover, this decomposition is unique (up to the arrangement of I_i's, of course).

For any ideal I, the prime components of $\sqrt{I}$ are called *the associated primes of* I. Note that this decomposition depends also on $\mathcal{F}$, which can be seen for example in that the polynomial $(x^2 - 2)(x^2 + 1)$ factorizes over $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ to 2, 3 or 4 factors, respectively.

Remark 3.1. This decomposition can be done algorithmically, but is computationally quite costly and will be the dominating part of our algorithm for finding (and defining) the complete form.

Theorem 3.4. Every ideal I of $\mathcal{R}$ is finitely generated, that is, there exists a finite collection of elements $f_1, \ldots, f_s \in I$ such that $I = \langle f_1, \ldots, f_s \rangle$ (the s depends on I).

Examples:

1. take $\mathcal{R} = \mathbb{C}[y]$ and $I = \langle f \rangle$ where $f(y) = (y-a_1)^{e_1}(y-a_2)^{e_2} \cdots (y-a_r)^{e_r}$ with e_i positive integers and all $a_i \in \mathbb{C}$ distinct. Now $\sqrt{I} = \langle f_{red} \rangle$ where $f_{red}(y) = (y - a_1) \cdots (y - a_r)$. Prime decomposition:

$$\sqrt{I} = \langle y - a_1 \rangle \cap \langle y - a_2 \rangle \cap \cdots \cap \langle y - a_r \rangle$$

2. take $\mathcal{R} = \mathbb{R}[x, y, z]$ and $I = \langle xz, yz \rangle$. Now $\mathcal{V}(I) = \{\text{the plane } z = 0\} \cup \{\text{the line } x = y = 0\}$ in $\mathbb{R}^3$. Prime decomposition:

$$\sqrt{I} = I = \langle z \rangle \cap \langle x, y \rangle$$

which has a clear geometrical interpretation: $\mathcal{V}(z) = \{\text{the plane } z = 0\}$ and $\mathcal{V}(x, y) = \{\text{the line } x = y = 0\}$.

Suppose A is an $n \times k$ matrix, with $n \leq k$, over $\mathcal{R}$. Its *Fitting ideals* I_j are

$$\begin{aligned} I_{-1}(A) &:= 0 && (12) \\ I_j(A) &:= \sqrt{\langle (n-j) \text{ -sized minors of } A \rangle}, \quad \forall\, j = 0, \ldots, n-1 && (13) \\ I_n(A) &:= \mathcal{R} && (14) \end{aligned}$$

Clearly $I_j \subset I_{j+1}$ and

$$\mathcal{F}^m = \mathcal{V}(I_{-1}) \supset \cdots \supset \mathcal{V}(I_j) \supset \mathcal{V}(I_{j+1}) \supset \cdots \supset \mathcal{V}(I_n) = \varnothing \tag{15}$$

Also, $\mathcal{V}(I_j) = \{$ points where rank of $A < n - j\}$. Especially, $\mathcal{V}(I_0) = \varnothing \Leftrightarrow A(z)$ is of full rank $\forall\, z \in \mathcal{F}^m$.

Let $I = \langle f_1, \ldots, f_s \rangle$ and $f \in \mathcal{R}$. The *membership problem* is to decide whether $f \in I$. Now there is a natural generalization of the elementary euclidean algorithm, called "the division algorithm" in [CLO92, ch. 2], which computes for the ordered set $\{f_1, \ldots, f_s\}$ (unique) elements $r \in \mathcal{R}$ and $h_i \in \mathcal{R}$ such that $f = h_1 f_1 + \cdots + h_s f_s + r$. Clearly if $r = 0$ then $f \in I$. Unfortunately, the converse is not true in general. The r above is the *remainder of f with respect to the ordered set* $\{f_1, \ldots, f_s\}$. In general, r (and h_i) depends on the order in which the f_i are given as input to the division algorithm. That is, the remainder w.r.t. $\{f_1, f_2, f_3\}$ might be different than the remainder w.r.t. $\{f_2, f_1, f_3\}$. The tool to overcome these difficulties is a gröbner basis, which will be introduced next.

Suppose given an ordering (see e.g. [CLO92, ch. 2]) for $\mathcal{R}$. If $f \in \mathcal{R}$, the *leading term of f* is the term of f which is highest with respect to the ordering. Let I be a nonzero ideal. The leading terms of I, denoted $\mathrm{LT}(I)$, is the collection of leading terms of elements of I. A *gröbner basis* for I is a generating set $\{f_1, \ldots, f_s\}$ such that

$$\begin{aligned} I &= \langle f_1, \ldots, f_s \rangle \\ \langle \mathrm{LT}(I) \rangle &= \langle \mathrm{LT}(f_1),\ \ldots,\ \mathrm{LT}(f_s) \rangle. \end{aligned}$$

Remark 3.2. This is a bit abuse of language, since in usual mathematical terminology a "basis" means an *independent* generating set. However, it is common with ideals to call any generating set a basis. See also remark 3.4.

The reason we need gröbner bases is that they solve the membership problem:

Theorem 3.5. With notations as above, the remainder of f w.r.t. a gröbner basis of I is zero if and only if $f \in I$. Moreover, the remainder does not depend on the order the f_j's are presented.

Remark 3.3. On differential algebra: a *differential ring* is a ring with a distinquished linear mapping δ with property

$$\delta(ab) = (\delta a)b + a\delta b.$$

Also, differential ideal, prime differential ideal, radical differential ideal are defined as usual, with the additional requirement that they are closed w.r.t. δ, that is $\delta I \subset I$.

Theorem 3.3 has its counterpart in differential algebra: a radical differential ideal has a unique decomposition by prime differential ideals. However, in this paper we will not use differential algebra although we mention it in few occasions. See [Pom83, Kol73] for an introduction to differential algebra, or [Rit50, Kap57] if you want a more readable introduction to our case.

3.2 Algebraic preliminaries: modules

Recall that a module is defined like a vector space except that the set of scalars is only a ring, not necessarily a field. Note that an ideal of $\mathcal{R}$ is an example of an $\mathcal{R}-$module. Hence the concept of a module is a generalization of both vector spaces and ideals.

Remark 3.4. For modules, a *basis* is defined like for vector spaces: a generating, independent (over $\mathcal{R}$) set. Unfortunately, this does not coincide with the definition of a basis of an ideal, see remark 3.2.

Note that while all vector spaces have a basis (by the axiom of choice), a module usually has no basis at all. For example, a non-principal ideal has never a basis (as a module!): if $I = \langle a, b, c, \ldots, d \rangle$, then the set $\{a, b, c, \ldots, d\}$ is dependent over $\mathcal{R}$: namely, $b \cdot a + (-a) \cdot b + 0 \cdot c + \cdots + 0 \cdot d = 0$.

When a module has a basis, it is called *free*. All free modules are isomorphic to $\mathcal{R}^s$ (the direct sum) with some s. When s is a finite integer, a module isomorphic to $\mathcal{R}^s$ is called a *finite free module*. If $\varphi : A \to B$ is a homomorphism of modules, then $\ker(\varphi)$, $\mathrm{im}(\varphi)$, $\mathrm{coker}(\varphi)$ define $\mathcal{R}-$modules in a natural way. Also, $\ker(\varphi)$ is a submodule of A and $\mathrm{im}(\varphi)$ is a submodule of B.

As an example of a module without a basis, consider the matrix $A : \mathcal{R}^3 \to \mathcal{R}$ with $A = [a, b, c]$ (not all zero) and set $M := \ker(A)$. It can be shown that M is generated by the vectors $u := [-b, a, 0]^T$, $v := [c, 0, -a]^T$ and $w := [0, -c, b]^T$, and any two of these are not enough to span M. However, these are linearly dependent: $cu + bv + aw = 0$. Note that, if $\mathcal{R}$ was a field and $a \neq 0$, then a would be invertible and $w \in \mathrm{span}\{u, v\}$. But in the ring case, a nonzero element is not invertible in general.

A consequence of the nonexistence of a basis of a module is that we cannot define dimension of a module as in the vector space case. To define the dimension, we need to recall the following concept: a sequence of modules M_i and homomorphisms $\phi_i : M_i \to M_{i-1}$

$$\ldots \xrightarrow{\phi_{i+1}} M_i \xrightarrow{\phi_i} M_{i-1} \xrightarrow{\phi_{i-1}} \ldots \tag{16}$$

such that $\mathrm{im}\ \phi_i = \ker\ \phi_{i-1} \quad \forall i$, is called *exact*.

All modules we will consider are either submodules of $\mathcal{R}^s$, $s \in \mathbb{N}$ or of the form $\mathcal{R}^s/M$ where M is submodule of $\mathcal{R}^s$. Let M be a module. Then a *presentation of* M is a matrix A over $\mathcal{R}$ such that the sequence

$$\mathcal{R}^n \xrightarrow{A} \mathcal{R}^k \longrightarrow M \longrightarrow 0 \tag{17}$$

is exact, i.e. coker$A \simeq M$. On the other hand, given a $k \times n-$matrix A, it defines a module by the sequence (17). The presentation can be extended to an exact sequence of finite free modules:

$$0 \longrightarrow F_n \xrightarrow{\phi_n} F_{n-1} \xrightarrow{\phi_{n-1}} \dots \xrightarrow{\phi_1} F_0 \longrightarrow M \longrightarrow 0 \tag{18}$$

which is called a (finite free) resolution of length n. Now we can define the dimension: $\dim(M) = \min\{n \in \mathbb{N} |$ there exists a resolution of length $n\}$.

For example, if $M := \mathcal{R}/I$ where I is an ideal $I = \langle f^1, \dots, f^r \rangle$, it is presented by the column vector $[f^1, \dots, f^r]$ and $\dim(M) = 1$. Any free module is of dimension zero, since

$$0 \longrightarrow F_0 := \mathcal{R}^s \xrightarrow{id} \mathcal{R}^s \longrightarrow 0$$

is exact. In particular, any vector space, considered as a module, is of dimension zero.

3.3 An algorithm

Now $\mathcal{F} := \mathbb{Q}$ and $\mathcal{R} := \mathcal{F}[t, y, y_1, \dots, y_q]$ with $y = (y^1, \dots, y^n)$. That is, if I is an ideal then $\mathcal{V}(I) \subset \mathcal{F}^{nq+n+1}$. The *formal derivative* is the unique linear mapping $\delta : \mathcal{R} \to \mathcal{R}[y_{q+1}]$ such that

$$\begin{aligned} \delta(ab) &= (\delta a)b + a\delta b \\ \delta(y^i_j) &= y^i_{j+1} \quad \forall\, i \in \{1, \dots, n\} \quad \forall\, j \in \{0, \dots, q\} \\ \delta(r) &= r' \quad \text{(the usual derivative w.r.t. } t\text{), if } r \text{ independent of } y^i_j\text{'s} \end{aligned}$$

This coincides with the usual derivative in $\mathcal{R}$ when jet coordinates are interpreted as derivatives: $y^i_j \leftrightarrow (y^i)^{(j)}$ and, of course, y is a smooth enough function of t. The word "formal" refers to the fact that we are not concerned whether or not y is a (smooth) function of t.

The algorithm is presented in tables 1 and 2. Reasoning and comments for the algorithm *PRIMESYS* are presented in the remarks of this section. We have used *Singular* [GPS01] (we used version 1.3.8, actually) in our test runs.

Note that, by theorem 3.4, each Σ_i is presented by a finite set of generators.

Assumption. We assume that, in step 2 of PRIMESYS, all of the V_j and A' are quasialgebraic. The set A need not be such. Note that if A is q.a., then V_j's are also.

Remark 3.5. In step 2, the sets V_j are, in principle, point sets chosen according to rank(B). Note that $V_j \cap V_i = \varnothing$ for $j \neq i$ and

$$A = \bigcup_{j=-1}^{n-1} V_j. \tag{27}$$

INPUT:	a polynomial differential equation, that is, a finite subset $f \subset \mathcal{R}$.
OUTPUT:	a finite collection of pairs of ideal bases (A_j, B_j) such that $\mathfrak{IF} = \{\mathcal{V}(A_1) - \mathcal{V}(B_1), \ldots, \mathcal{V}(A_N) - \mathcal{V}(B_N)\}$
Step 1:	Set $\mathfrak{IF} := \varnothing$.
Step 2:	make the prime decomposition for f: $f =: \Sigma_1 \cap \ldots \cap \Sigma_r$.
Step 3:	For $\mu = 1$ to r do $\mathfrak{IF} := \mathfrak{IF} \cup$ *PRIMESYS*(Σ_μ) od

Table 1: Algorithm ANYSYS for arbitrary polynomial system

Remark 3.6. In step 2: the construction of A (or A') is in general case not immediate. This might cause a problem considering arbitrary polynomial systems: for example, suppose that A is a variety: $A = \mathcal{V}(I)$ for some ideal I, whose generators are found by inspecting B and $\tilde{f}$. Now if we fail to find all of the generators of I, say that we generate an ideal $J \subsetneqq I$, then $\mathcal{V}(J) \supsetneqq \mathcal{V}(I)$ and using $\mathcal{V}(J)$ as A would make $\mathfrak{IF}$ too big.

We expect the techniques introduced in [Sit92] to be helpful here. However, this aspect is beyond the scope of this paper and will be postponed to future work. In the examples of this paper we have been able to construct A and A'.

Remark 3.7. step 3: A' is defined to get the "vertical tangents" within, see below example 3.1. Now $\tilde{f}(z) \notin \operatorname{im}(B(z))$ could be due to an inconsistent evaluation point z. But, if the set A' is small enough compared to $\mathcal{V}(f)$, it probably is due to "vertical tangents" and is worth accepting. Hence we choose as the criteria of accepting A' the dimension condition in step 3. It might be possible to consider also other choices for the criteria.

Remark 3.8. In steps 3 and 4, appending A' or V_{-1}, respectively, means appending a q.a. representation of it. Note also that V_{-1} consists of points where the rank of B^T is maximal. Now A' or V_{-1} might be empty (the latter is empty if $A \subset \mathcal{V}(I_0)$) which produces an empty component to $\mathfrak{IF}$. This is of course harmless but it would be nicer to avoid such irritating sets in advance by, for example, setting step 4: if $A \not\subset \mathcal{V}(I_0)$, then $\mathfrak{IF} := \mathfrak{IF} \cup V_{-1}$. Likewise one could add to step 3 the condition "if $A' \neq \varnothing$ and...".

Remark 3.9. step 5: the construction of A_j and B_j's is a side product of step 2. In step 8: it might be difficult to check the condition $V_{jmi} \neq \varnothing$.

Remark 3.10. In step 6, B and $\tilde{f}$ are redefined but actually just updated because generators of A_j include f, because $\langle f \rangle$ is prime and because of the definition of A_j. Also, there is no need to construct corresponding sets A, A' because we already, by the definition of A_j, are limited to case A. Also, if B for f was $k \times n$, then this new B (for A_j) is $\tilde{k} \times n$ with $\tilde{k} \geq k$.

Remark 3.11. step 6: A_{jm} and B_{jm} are a closer look to B, in the sense that $\mathcal{V}(A_{jm}) = \mathcal{V}(A_j) \cap \mathcal{V}(I_{m-1})$ and $\mathcal{V}(B_{jm}) = \mathcal{V}(B_j) \cup \mathcal{V}(I_m)$. Note that $A_{j0} = A_j$ and $B_{jn} = B_j$.

INPUT:	a finite subset $f \subset \mathcal{R}$ such that $\langle f \rangle$ is prime.
OUTPUT:	a finite collection of pairs of ideal bases (A_j, B_j) such that $\mathfrak{IF} = \{\mathcal{V}(A_1) - \mathcal{V}(B_1), \ldots, \mathcal{V}(A_N) - \mathcal{V}(B_N)\}$
Step 1:	$\mathfrak{IF} := \varnothing, \quad B\, y_{q+1} + \tilde{f} := \delta f$ and $I_j := I_j(B^T)$, the Fitting ideals of B^T.
Step 2:	

$$A' := \{z \in \mathcal{F}^{nq+n+1} \mid \tilde{f}(z) \notin \operatorname{im} B(z)\} \cap \mathcal{V}(f) \tag{19}$$
$$A := \{z \in \mathcal{F}^{nq+n+1} \mid \tilde{f}(z) \in \operatorname{im} B(z)\} \cap \mathcal{V}(f) \tag{20}$$
$$V_j := A \cap (\mathcal{V}(I_j) - \mathcal{V}(I_{j+1}))\,, \quad j = -1, \ldots, n-1 \tag{21}$$

Step 3: Case A': if $\dim(\mathcal{V}(I_0) \cap A') < \dim(\mathcal{V}(f))$ then $\mathfrak{IF} := \mathfrak{IF} \cup A'$.

Step 4: $\mathfrak{IF} := \mathfrak{IF} \cup V_{-1}$

Step 5: Study A: set $\Lambda := \varnothing$.
For $j = 0$ to $n-1$ do
if $V_j \neq \varnothing$, choose bases A_j and B_j for the corresponding (radical) ideals of the nonempty V_j's, that is, $V_j = \mathcal{V}(A_j) - \mathcal{V}(B_j)$.

Step 6: Update B and $\tilde{f}$ such that $B\, y_{q+1} + \tilde{f} := \delta(A_j)$.
For $m = 0$ to n do

$$I_m := I_m(B^T) \text{ (update the Fittings)} \tag{22}$$
$$A_{jm} := A_j \cup \{\text{generators of } I_{m-1}\} \tag{23}$$
$$B_{jm} := B_j \cap I_m \tag{24}$$

Step 7: make the prime decomposition for A_{jm}:

$$A_{jm} =: A_{jm0} \cap A_{jm1} \cap \cdots \cap A_{j,m,n_{jm}} \tag{25}$$

and let

$$V_{jmi} := \mathcal{V}(A_{jmi}) - \mathcal{V}(B_{jm}) \quad i = 0, \ldots, n_{jm} \tag{26}$$

Step 8: for $i = 0$ to n_{jm} do, if $V_{jmi} \neq \varnothing$, (steps 9 to 11)

Step 9: Reduce each entry of B with respect to A_{jmi} and then compute generators $\{v^1, \ldots, v^\nu\}$ for the module $\ker(B^T)$.

Step 10: for $k = 1$ to ν do
let $u^k := \sum_i v^{k,i} \tilde{f}^i$ where $v^{k,i}$ is the i^{th} component of v^k. If $u^k \notin A_{jmi}$ then $A_{jmi} := A_{jmi} \cup u^k$.
od (end of k-loop)

Step 11: if in the previous step all $u^k \in A_{jmi}$, then append $(\mathcal{V}(A_{jmi}), \mathcal{V}(B_{jm}))$ to $\mathfrak{IF}$
else $\Lambda := \Lambda \cup \{(j, m, i)\}$
od od od (end of i-loop) (end of m-loop) (end of j-loop)

Step 12: while $\Lambda \neq \varnothing$ do
pick a $(j, m, i) \in \Lambda$
$\Lambda := \Lambda - \{(j, m, i)\}$
$\mathfrak{IF} := \mathfrak{IF} \cup \mathit{ANYSYS}(A_{jmi})$
od

Table 2: Algorithm PRIMESYS for prime polynomial system

Remark 3.12. In step 7, we need prime ideals for two things: 1) to avoid zero divisors in step 9, and 2) to make $\mathcal{V}(A_{jmi})$ irreducible.

Remark 3.13. In steps 2 and 9: the idea is to mimic the idea "study separately each set where rank of ker B^T is constant" but this is not exactly so, because B^T (and B) is a mapping of modules and one can not even define its rank as in the case of a mapping between vector spaces. Indeed as noted in section 3.2, one can not define a dimension for modules as for vector spaces.

We consider this as a suitable way to overcome the "constant rank" assumptions which are a severe limitation in most other DAE approaches, see section 4.1.

Remark 3.14. Step 10: construct the (possibly) new generators for ideals; i.e. if we have a new generator u^k, it means we found a hidden equation $u^k = 0$ for *this component*. In step 11, all $u^k \in A_{jmi}$ means no new generators.

Proof. **Proof of termination of** *PRIMESYS*: the only place that needs to be checked is the recursive loop in step 12. But there, if the algorithm is needed for A_{jmi}, it means that $A_{jmi} \supsetneqq A_{jm} \supset A_j \supset f$ where A_{jm} is from (23), A_j is from Step 5 and f from Step 1. Hence we have a strictly ascending chain of ideals. Now $\mathcal{R}$ is a nötherian ring and the process terminates in finitely many steps. ☐

Definition 3.2. Given equation $f = 0$ with f as in (1) such that each $f^i \in \mathcal{R}$, the set $\mathfrak{IF}$ constructed in the algorithm *ANYSYS* of table 1 is called the complete form of f.

As a point set, it is of the form

$$\mathfrak{IF} = \bigcup_{j=1}^{N} (V_j - W_j) \tag{28}$$

where each V_j and W_j are varieties, the latter possibly empty and the former irreducible. Each variety V is presented by a finite generating set of $\mathcal{I}(V)$.

Now, the numerical solution of (1) will be done to *each component separately*. Here one can use the methods described in [TA00, TA01]. Especially, an initial point is consistent if and only if it belongs to some $V_j - W_j$.

The theory of existence and uniqueness of solutions immediately reduces to the theory of [TA00].

3.4 Examples

Example 3.1. The sphere (example 2.1) revisited: in step 2

$$\begin{aligned}
A &= \{\, \{(t,y,y_1) \,|\, t^2+y^2+(y_1)^2-1=0, \quad y_1 \neq 0\},\ \{(0,\pm 1,0)\}\,\} && (29)\\
A' &= \{(t,y,0) \,|\, y_1 = 0, \quad t + y\,y_1 \neq 0, \quad t^2+y^2+(y_1)^2-1=0\} && (30)\\
I_0 &= \langle y_1 \rangle, \quad I_1 = \langle 1 \rangle \\
V_{-1} &= \mathcal{V}(t^2+y^2+(y_1)^2-1) - \mathcal{V}(y_1) \\
&= \{(t,y,y_1) \,|\, t^2+y^2+(y_1)^2-1=0\} - \{(t,y,y_1) \,|\, y_1 = 0\} && (31)\\
V_0 &= \{(0,\pm 1,0)\}. && (32)
\end{aligned}$$

Step 3: $\dim(A') = 1 < \dim(\text{sphere})$ hence $\mathfrak{I}\mathfrak{F} := \mathfrak{I}\mathfrak{F} \cup A'$. Here $A' = \mathcal{V}(y_1,\, t^2{+}y^2{-}1) - \mathcal{V}(y_1,\, t)$. Step 4: $\mathfrak{I}\mathfrak{F} := \mathfrak{I}\mathfrak{F} \cup V_{-1}$. At this point we note that (as a point set) $\mathfrak{I}\mathfrak{F} = \mathcal{V}(f) - \{(0, \pm 1, 0)\}$. Now one could immediately see that the remaining components (namely V_0) are zero dimensional and therefore cannot have a 1-dimensional distribution. Hence they could be discarded right now and we are done. But let us see anyway how the algorithm works in this simple example: Step 5: $A_0 = \{f,\, y^2 - 1\}$, $B_0 = \{1\}$, as can be seen from above. Updated B and $\tilde{f}$:

$$B = \begin{pmatrix} y_1 \\ 0 \end{pmatrix} \quad \tilde{f} = \begin{pmatrix} t + y\, y_1 \\ 2y\, y_1 \end{pmatrix} \tag{33}$$

step 6:

$$\begin{aligned} A_{00} &= A_0 = \{f, y^2 - 1\} \\ A_{01} &= A_0 \cup \{y_1\} \\ B_{00} &= \{y_1\} \\ B_{01} &= \{1\} \end{aligned}$$

decompositions give: $A_{00} = A_{000} \cap A_{001}$ and $A_{01} = A_{010} \cap A_{011}$ where

$$\begin{aligned} A_{000} &= \{y + 1,\, t^2 + (y_1)^2\} \\ A_{001} &= \{y - 1,\, t^2 + (y_1)^2\} \\ A_{010} &= \{y + 1,\, t,\, y_1\} \\ A_{011} &= \{y - 1,\, t,\, y_1\} \end{aligned}$$

step 8: now $V_{000} = V_{010} = \{(0, -1, 0)\}$ and $V_{001} = V_{011} = \{(0, 1, 0)\}$. Step 9 (for V_{010}): $\ker(B^T)$ is generated by $\left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$ and $u^1 := 2y\, y_1 \in A_{010}$ hence $\mathfrak{I}\mathfrak{F} := \mathfrak{I}\mathfrak{F} \cup V_{010}$. Step 9 for V_{011} works likewise and $\mathfrak{I}\mathfrak{F} := \mathfrak{I}\mathfrak{F} \cup V_{011}$. Hence we are done and the output is

$$\begin{aligned} \mathfrak{I}\mathfrak{F} = \big\{ &\mathcal{V}(t^2 + y^2 + (y_1)^2 - 1) - \mathcal{V}(y_1), \\ &\mathcal{V}(t^2 + y^2 - 1) - \mathcal{V}(t), \\ &\{(0, -1, 0)\} - \varnothing, \\ &\{(0, 1, 0)\} - \varnothing \big\} \end{aligned} \tag{34}$$

Here the set A' brings in the 'formerly forbidden' equator of the sphere.

Remark 3.15. A surprising side effect is that we found the two singularity points $\{(0, \pm 1, 0)\}$, cf. [TA00].

Example 3.2. This is example 4 in the help file "overview of rifsimp package" of *Maple*. Here $n = 1$, $k = q = 3$, and we denote y instead of y^1.

$$y^3 + y + 1 = 0 \tag{35}$$

$$y_2 - 3y_1 = 0 \tag{36}$$

$$y_3 - 2y_1 = 0 \tag{37}$$

The system is prime.

$$B = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad \tilde{f} = \begin{pmatrix} 3y^2y_1 + y_1 \\ -3y_2 + y_3 \\ -2y_2 \end{pmatrix} \tag{38}$$

$A' = \varnothing$, $V_{-1} = A = \mathcal{V}(f,\, 3y^2y_1 + y_1,\, -3y_2 + y_3)$, $V_0 = \varnothing$ and a gröbner basis of $\mathcal{I}(A)$ is $\{y_3, y_2, y_1, y^3 + y + 1\}$ in agreement of rifsimp (more precisely, in our notation: $\mathfrak{I}\mathfrak{F} = \{\mathcal{V}(y_3, y_2, y_1, y^3 + y + 1),\, \varnothing)\}$. Especially, one can see that the solution y is constant.

Remark 3.16. In the previous example, we can also see that the problem can be projected from J_3 to J_0, since y_3, y_2, y_1 are clearly consequences from the generator $y^3 + y + 1$ which says that y is constant. However, we do not consider this aspect in this paper.

Example 3.3. This is example 1 in the rifsimp package of *Maple*. Here $n = 1$, $k = 2$, $q = 3$.

$$\begin{aligned} t\,(y_1)^2\,(y_2)^2 - 2t\,y\,y_1\,y_2\,y_3 + t\,y^2\,(y_3)^2 - y\,y_2 + (y_1)^2 &= 0 \qquad (39) \\ -y_1\,y_2 + y\,y_3 + 2\,y^2\,(y_2)^2 - 4y\,y_2\,(y_1)^2 + 2(y_1)^4 &= 0 \qquad (40) \end{aligned}$$

decomposition gives 3 components:

$$\begin{aligned} \Sigma_1 &= \langle 5 \text{ generators} \rangle \qquad (41) \\ \Sigma_2 &= \begin{cases} (y_2)^2 - y_1\,y_3 \\ y_1\,y_2 - y\,y_3 \\ (y_1)^2 - y\,y_2 \end{cases} \qquad (42) \\ \Sigma_3 &= \langle y, y_1 \rangle \qquad (43) \end{aligned}$$

Now Σ_2 is already complete, and Σ_3 gives: $\langle y, y_1\rangle \Rightarrow \langle y, y_1, y_2\rangle \Rightarrow \langle y, y_1, y_2, y_3\rangle$ and the last form is complete. We note that as in remark 3.16, the last form is clearly equivalent with $y \equiv 0$ but we do not consider (methods to find) such reductions in this paper.

With $f := \Sigma_1$ we get:

$$B = \begin{pmatrix} y \\ -2tyy_1y_2 + 2ty^2y_3 \\ -2ty(y_1)^2 + 2ty^2y_2 \\ -2ty(y_1)^2y_2 - 2ty^2(y_2)^2 + 4ty^2y_1y_3 \\ 2ty(y_1)^3 - 6ty^2y_1y_2 + 4ty^3y_3 + tyy_2 \end{pmatrix} \tag{44}$$

A: now if $y = 0$ then $B = 0$, hence for A is needed also $\tilde{f} = 0$ and $A \cap \mathcal{V}(y) = \mathcal{V}(y, f, \tilde{f}) = \mathcal{V}(1) = \varnothing$ so we can suppose $y \neq 0$. One can also show that the other elements of B do not vanish in $\mathcal{V}(f)$, hence

$$\begin{aligned} A &= \mathcal{V}(f) \qquad (45) \\ A' &= \mathcal{V}(f, y) - \mathcal{V}(\tilde{f}) \qquad (46) \\ V_{-1} &= A - V_0 \qquad (47) \\ V_0 &= A \cap \langle y \rangle = \mathcal{V}(f, y) \qquad (48) \end{aligned}$$

so we can choose $A_0 = \{f, y\}$, $B_0 = \{1\}$. Now $\langle f, y\rangle \supsetneqq \langle f\rangle$ hence $\dim(A') < \dim\mathcal{V}(f)$ and A' is accepted. For A_0 we get

$$B = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad \tilde{f} = \begin{pmatrix} y_1 \\ 2ty_2y_3 + (y_2)^2 \\ 6(y_1)^2y_2 - y_3 \end{pmatrix} \tag{49}$$

hence $A_{00} = A_0 = A_{01}$ and are primes, $B_{00} = 0$, $B_{01} = \{1\}$. $V_{000} = \varnothing$, $V_{010} \neq \varnothing$, step 9: new generators are $\tilde{f}$ from (49), and now it turns out that $\mathcal{V}(A_{000}, \tilde{f}) = \varnothing$. Hence the algorithm (for Σ_1) stops and we get that $\mathfrak{IF}$ for Σ_1 is $A' \cup V_{-1}$ which can be shown to simplify to $\mathcal{V}(f)$. Hence $\mathcal{V}(\Sigma_1)$ contains 'vertical tangents' but it is accepted as a whole.

Example 3.4. *An ODE.* Let's look at the situation

$$f(t, y, y_1) := y_1 - g(t, y) = 0 \quad \text{with } n = k \tag{50}$$

which is what most people mean by "a (non-constrained) ordinary differential equation". Step 2 gives B = identity matrix, $\tilde{f} = -\frac{\partial}{\partial t}g - (\frac{\partial}{\partial y}g)\, g$. Now $\ker(B^T)$ is trivial, hence step 9 gives only the zero vector. In step 10 we have $u^1 := 0$ which certainly belongs to $\langle f\rangle$, hence we are done and (50) already is a complete form.

Example 3.5. In [BCP89, p. 34] is described a semiexplicit DAE:

$$\begin{cases} x_1' - F_1(x_1, x_2, t) & = 0 \\ F_2(x_1, x_2, t) & = 0 \end{cases} \tag{51}$$

and it has been said that this is index one if and only if $\frac{\partial}{\partial x_2}F_2$ is nonsingular. Let us see how this looks like in our algorithm: first, we suppose that F_1, F_2 are polynomials and f is prime. Then,

$$B = \begin{pmatrix} \frac{\partial}{\partial x_1'}F_1 & \frac{\partial}{\partial x_2'}F_1 \\ \frac{\partial}{\partial x_1'}F_2 & \frac{\partial}{\partial x_2'}F_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \tag{52}$$

$$\tilde{f} = \begin{pmatrix} -\delta F_1 \\ \delta F_2 \end{pmatrix} = \begin{pmatrix} -\delta F_1 \\ (\frac{\partial}{\partial x_1}F_2)x_1' + (\frac{\partial}{\partial x_2}F_2)x_2' + \frac{\partial}{\partial t}F_2 \end{pmatrix} \tag{53}$$

B is constant, hence the fittings are trivial and do not affect, and step 10 gives the new generators δF_2, hence the new f is

$$\begin{cases} f \\ \delta F_2 \end{cases} \tag{54}$$

Now again, to proceed as in [BCP89], we have to suppose that the ideal $\langle f, \delta F_2\rangle$ is prime. Then, B is

$$\begin{pmatrix} I & 0 \\ 0 & 0 \\ \frac{\partial}{\partial x_1}F_2 & \frac{\partial}{\partial x_2}F_2 \end{pmatrix} \tag{55}$$

now $\ker(B^T)$ is trivial if and only if $\frac{\partial}{\partial x_2}F_2$ is nonsingular.

We conclude that the definition of the index in [BCP89] does not take into account the prime structure. The following example shows a side effect of this.

Example 3.6. This is example 9.2 in [AP98]:

$$y_1^1 - y^3 = 0 \tag{56}$$

$$y^2(1 - y^2) = 0 \tag{57}$$

$$y^1y^2 + y^3(1 - y^2) - t = 0 \tag{58}$$

First, decomposition gives:

$$\sqrt{\langle f \rangle} = \langle y^2,\, y^3 - y_1^1,\, f^3\rangle \cap \langle y^2 - 1,\, y^3 - y_1^1,\, f^3\rangle \tag{59}$$

after running the algorithm, we have

$$\begin{aligned}\mathfrak{IF} = {} & (\mathcal{V}(y_1^2,\, y_1^3 - 1,\, y^2,\, y^3 - y_1^1,\, t - y_1^1), \varnothing) \\ & \cup (\mathcal{V}(y_1^2,\, -y_1^1 + 1,\, y^3 - y_1^1,\, y^2 - 1,\, t - y^1,\, y_1^3), \varnothing).\end{aligned} \tag{60}$$

Here one can see an explanation for the effect of the initial value (of a solution) on the index as noted in [AP98]: any consistent initial point must belong to one (and only one, in this case) of the varieties $\mathcal{V}(y_1^2,\, y_1^3 - 1,\, y^2,\, y^3 - y_1^1,\, t - y_1^1)$ or $\mathcal{V}(y_1^2,\, -y_1^1 + 1,\, y^3 - y_1^1,\, y^2 - 1,\, t - y^1,\, y_1^3)$. In the former case, $y^2 = 0$ and we are solving the system $y_1^2 = 0, y_1^3 = 1, y^2 = 0, y^3 = y_1^1, t = y_1^1$ which originally came from the component $\langle y^2,\, y^3 - y_1^1,\, y^1y^2 + y^3(1 - y^2) - t\rangle$ in (59). This component is in notation of [AP98] $y_2 = 0, y_3 = y_1', y_1y_2 + y_3(1 - y_2) - t = 0$. In the latter case, $y^2 = 1$ and we are solving the system $y_1^2 = 0, y_1^1 = 1, y^3 = y_1^1, y^2 = 1, t = y^1, y_1^3 = 0$ which originally came from the component $\langle y^2 - 1,\, y^3 - y_1^1,\, y^1y^2 + y^3(1 - y^2) - t\rangle$ in (59). This component is in notation of [AP98] $y_2 = 1, y_3 = y_1', y_1y_2 + y_3(1 - y_2) - t = 0$. So the explanation for the dependence of the index on the initial value is that the index depends on the (prime) component! Any consistent initial value belongs to a variety of some prime component.

Example 3.7. *A "triangular" example.* An equation of the form $Ag = 0$ where A is $m \times m$ upper triangular matrix with nonzero diagonal and g is an $m-$vector, both A and g with elements from $\mathcal{R}$ and the element A_{mm} a nonzero constant, is equivalent with $g = 0$. Although this seems trivial, this kind of equation is considered in examples in [KM98] and also in [CG95b] as examples of the difficulties with defining indeces. It also shows that in (most of) the conventional definitions of indeces, the index depends on the chosen representation of the equations. Hence it is not intrinsic.

4 Relations to other approaches

In this section we compare our method to others in literature. For that, we assume that $q = 1$ in (1). Due to vast amount of articles on DAEs it is clear that we cannot do an exhaustive survey but we have chosen few papers that, in our opinion, represent quite well the conventional approaches.

Let us first note a recent paper [PS] which includes few pages of comparison in the same spirit as ours, although they have not considered the paper [KM98] which contains a generalization to nonlinear case.

4.1 Relations considering numerical solving

In this section we consider the point of view of numerics. In [KM96] the strangeness index is defined for certain linear DAEs and the definition is generalized to nonlinear case in [KM98]. In these articles the system is not assumed to be a polynomial, and they present an algorithm for transforming the system into a so called strangeness free from, but we claim that their approach is of limited applicability. The strangeness index is not always defined: indeed there are strong requirements (hypothesis 3.2 in [KM98]) for the rank of B, where B refers to our notation in *PRIMESYS* step 1, to be constant which means that (20) reduces to $A = V_j$ for some j. Also, A' is not considered at all.

The algorithm for converting into strangeness free form requires finding suitable coordinates in intermediate steps of the algorithm. Although in many systems in practical applications this can be done “by inspection”, this is generally not constructive. Also checking the rank conditions is a non-trivial problem for which they do not present a constructive solution.

In the approach of Campbell et al., e.g. [CHYZ98, CG95b, BCP89], the derivative array is formed. This is essentially prolongation without projection (compare to CK algorithm in section 2.3). Prolongation is continued μ times, where μ is defined to be the global index of the system, until y' is uniquely determined by (t, y). This definition is also extended to a local version: index of the system along a solution, see [AP98, p. 236]. Here one could also interprete the projection step of the CK algorithm as a procedure which automatically picks up the relevant equations from a derivative array.

Campbell et al use the following assumptions (here G is the derivative array):

(A1) sufficient smoothness of G

(A2) consistency of $G = 0$ as an algebraic equation

(A3) $\overline{J} := [\frac{\partial}{\partial y_1}G \ldots \frac{\partial}{\partial y_q}G]$ is 1-full and has constant rank

(A4) $J := [\frac{\partial}{\partial y}G \frac{\partial}{\partial y_1}G \ldots \frac{\partial}{\partial y_q}G]$ has full rank everywhere

In our case, assumption A1 becomes trivial and A2 is implicitly assumed in “if $V_j \neq \varnothing$, then...”. But A3 and A4 are quite different compared to ours. Indeed we have no assumptions for constant rank or full rank, on the contrary we decompose the system by Fitting ideals, see remark 3.13.

As noted in [CHYZ98, p. 78], checking the 1-fullness in a neighborhood is generally not constructive. Although, as noted there, one can compute the "symbolic rank" by computer algebra programs and then compare to it the "numerical rank" at a point, the problem remains: how can one determine the symbolic rank to be constant? Therefore their approach has the same problem as that in [KM98].

The approach of Rabier and Rheinboldt [RR91, RR94] is geometrical and closely related to that of Reich's [Rei91]. The main ideas are similar as in the papers we have considered in this section. Their definitions are more intrinsic, due to their geometrical nature without referring to equations. However, two main problems remain: first, to actually handle the system, even if the definitions are geometrical, one needs to handle *equations* after all. Second, they are forced to use similar constant rank conditions as in the papers considered earlier in this section. More precisely, they assume that the "interstage" manifolds in their definition are of constant dimension.

We also note that in [RLW01] is proven that the approach of Rabier and Rheinboldt is equivalent to a version of the geometrical theory of PDEs.

4.2 Relations to computer algebra approaches

There has been developed in the last decade several computer algebra approaches, that is methods based on symbolic manipulation of the equations, to DAEs. In this section we consider relations of those to our method. Although the algorithm in this paper is also a computer algebra approach, our aim is to get a form which is suitable for numerical integration. We also remind the reader that [TA00] is a lengthy exposition of what "suitable" in this case means.

Now almost all of these symbolic approaches consider the case of *partial* differential equations and are viewing DAEs as only a special case. Like Kolchin puts it [Kol73, p. xiii]:

> ...there is no special distinction made between ordinary and partial differential equations. The governing philosophy is that 1 is merely a special case of m, a case neither requiring nor greatly benefitting from special treatment.

However, we do feel that the ordinary (DAE) case does deserve a special attention. We also like to recall that we do not make difference between "DAE"s and "ODE"s, cf. [TA00, remark 3.6].

The symbolic approaches, see e.g. [Hub97, RLW01] and references therein, are mostly based on differential algebra (see remark 3.3): the system defines a differential ideal. However, we saw in a very simple example (remark 2.2) that we cannot base our method on differential ideals.

Another property of (the implementations of) these approaches is that they assume each equation to be solvable for its highest derivative term. This causes some 'pivoting' problems, that is, if there is an equation whose highest derivative term is multiplied by a nonconstant term g, then the system splits to two cases: whether $g = 0$ or $g \neq 0$. This has some resemblance to our approach but it is not the same.

Also, changing the ranking might lead to a different splitting of cases. That is, their case splitting depends on the chosen ranking. It is not clear what is the geometrical interpretation of different case splittings.

Note that some of these approaches are implemented in Maple, for example packages rifsimp, diffgrob2 or diffalg.

We note that there seems to be a desire to have algorithms which avoid prime decomposition, due to its computational cost. See for example [Hub00] and references therein. We admit that it is an advantage to avoid the prime decomposition(s) but here is the same problem as in the splitting mentioned above: the choice of ranking decides what the separants and initials are, and it is not clear what choice, if any, is (geometrically) a "right one". However, it is an interesting question to pose also to our method: with what could the prime decompositions be replaced to reduce the computational cost?

Finally we mention the concept of an *algebraic index* defined in Pritchard and Sit [PS]. They have done a nice survey on DAE approaches but it is not clear how the algebraic index is related to others. On the other hand, they concentrate on quasilinear first order systems ($q = 1$ and linear with respect to y_1). They demonstrate how a system can be converted to a quasilinear one by adding more variables. It is not clear how such transformation would affect in case of approaches considered in the previous section. Moreover, as discussed in section 2.1 we like to avoid such transformations.

5 Conclusions

We have presented a method which continues our earlier work [TA00, TA01] and is between numerical and symbolic computations: we use symbolic computation to achieve a form, here called complete form, suitable to numerical computation. There are already methods aiming at same goal, but we demonstrate some problems they have.

As noted in [TA00], the conventional approaches to DAEs lack the fundamental property of *involutivity*, and this lack causes for example the well known problems of drift-off and finding consistent initial values for the system. One could think of the involutivity (or involutive form), as preconditioning the system: find all hidden equations. Now our complete form is aimed to be a kind of algebraic counterpart to involutivity, in the more general case where the system has components.

We assume that the system under consideration is a multivariate polynomial. This assumption is not very restrictive, since most applications in literature either are polynomials or can be converted to polynomials. On the other hand, this assumption makes it possible to define the complete form in such a way that we can, in particular, avoid the constant rank assumptions in conventional approaches. We claim that the constant rank assumptions are the main problem in those approaches.

Our tools come from commutative algebra and the computationally most costly operation is prime decomposition. We note that the decomposition depends on the chosen ground field, but we have restricted the ground field to be $\mathbb{Q}$.

Still comparing to literature, one could also think our method of "finding the complete form" as some kind of index reduction technique, but we take into account *all* equations instead of "choosing n eqns", what is done in index-reductions.

Finally, we have noted about constructivity: it seems to us that most 'algorithmic' approaches to numerics of DAEs include some steps which are, in general, nonconstructive. These are discussed in section 4.1. On the other hand, in section 4.2 we note that those working in symbolic algebra seem to have completely constructive algorithms but they are not concerned with numerical solution; i.e. what properties should the chosen form of the system have to be suitable to numerical computations? Also our algorithm has, at this level of implementation, a gap in constructivity, see remark 3.6. The next immediate task to do is to fill that gap with techniques mentioned in the remark.

References

[AP98] U. Ascher and L. Petzold. *Computer Methods for ODEs and DAEs*. SIAM, 1998.

[BCP89] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in DAEs*. North-Holland, 1989.

[CG95a] S.L. Campbell and C.W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

[CG95b] S.L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. *SIAM J. Sci. Comp*, 16(2):257–270, 1995.

[CHYZ98] S. Campbell, R. Hollenbeck, K. Yeomans, and Y. Zhong. Mixed symbolic-numerical computations with general DAEs I: system properties. *Numer. Algor.*, 19:73–83, 1998.

[CLO92] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms*. Springer, 1992.

[Eis96] D. Eisenbud. *Commutative Algebra*, volume 150 of *Graduate Texts in Mathematics*. Springer, 1996. corr. 2nd printing.

[Gea71] C. Gear. Simultaneous numerical solution of DAEs. *IEEE Trans. Circ. Th.*, 18(1):89–95, 1971.

[GPS01] G.-M. Greuel, G. Pfister, and H. Schönemann. SINGULAR 2.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern, 2001. `http://www.singular.uni-kl.de`.

[HLR89] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods.* Springer-Verlag, 1989.

[Hub97] E. Hubert. *Étude Algébrique et Algorithmique des Singularités des Équations Différentielles Implicites.* PhD thesis, l'Institut National Polytechnique Grenoble, 1997. in French and English.

[Hub00] E. Hubert. Factorization-free decomposition algorithms in differential algebra. *J. Symb. Comp.*, 29:641–662, 2000.

[HW91] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: stiff and differential-algebraic problems*, volume 14 of *Computational Mathematics.* Springer, 1991.

[Kap57] I. Kaplansky. *An introduction to differential algebra.* Hermann, Paris, 1957.

[KM96] P. Kunkel and V. Mehrmann. Local and global invariants of linear DAEs and their relation. *Electr. Trans. Num. Anal.*, 4:138–157, 1996.

[KM98] P. Kunkel and V. Mehrmann. Regular solutions of nonlinear DAEs and their numerical determination. *Numer. Math.*, 79:581–600, 1998.

[Kol73] E.R. Kolchin. *Differential Algebra and Algebraic Groups*, volume 54 of *Pure and Applied Mathematics.* Academic Press, 1973.

[Man96] E. Mansfield. A simple criterion for involutivity. *J. London Math. Soc.*, 2(54):323–345, 1996.

[Mär92] R. März. Numerical methods for differential-algebraic equations. *Acta numerica*, 1:141–198, 1992.

[Pom83] J. F. Pommaret. *Differential Galois Theory*, volume 15 of *Mathematics and Its Applications.* Gordon and Breach Science Publishers, 1983.

[Pom94] J. F. Pommaret. *Partial Differential Equations and Group Theory: New perspectives for applications*, volume 293 of *Mathematics and Its Applications.* Kluwer Academic Publishers, 1994.

[PS] F. Pritchard and W. Sit. On initial value problems for ordinary DAEs. *to appear in J. Symb. Comp.*

[Rei91] S. Reich. On an existence and uniqueness theory for nonlinear DAEs. *Circuits Systems Signal Process*, 10(3), 1991.

[Rit50] J. Ritt. *Differential Algebra.* Dover, 1950.

[RLW01] G. Reid, P. Lin, and A. Wittkopf. Differential elimination-completion algorithms for DAE and PDAE. *Stud. Appl. Math.*, 106:1–45, 2001.

[RR91] P.J. Rabier and W. Rheinboldt. A general existence and uniqueness theory for implicit DAEs. *Diff. Int. Eqns*, 4:563–582, 1991.

[RR94] P.J. Rabier and W. Rheinboldt. A geometric treatment of implicit DAEs. *J. Diff. Eqns*, 109:110–146, 1994.

[Sei99] W. Seiler. Indices and solvability for general systems of differential equations. In V. Ghanza, E. Mayr, and E. Vorzthsov, editors, *Computer Algebra in Scientific Computing – CASC 99*, pages 355–385. Springer, 1999.

[Sit92] W. Sit. An algorithm for solving parametric linear systems. *J. Symb. Comp.*, 13(4):353–394, 1992.

[TA00] J. Tuomela and T. Arponen. On the numerical solution of involutive ordinary differential systems. *IMA J. Numer. Anal.*, 20:561–599, 2000.

[TA01] J. Tuomela and T. Arponen. On the numerical solution of involutive ordinary differential systems 2. *BIT*, 41:599–628, 2001.

(continued from the back cover)

A428 Jarmo Malinen
Discrete Time H^{∞} Algebraic Riccati Equations, Mar 2000

A427 Marko Huhtanen
A Hermitian Lanczos method for normal matrices, Jan 2000

A426 Timo Eirola
A Refined Polar Decomposiotion: $A = UPD$, Jan 2000

A425 Hermann Brunner, Arvet Pedas and Gennadi Vainikko
Piecewise polynomial collocation methods for linear Volterra integro-differential equations with weakly singular kernels, Dec 1999

A424 Olavi Nevanlinna
Resolvent conditions and powers of operators, Dec 1999

A423 Jukka Liukkonen
Generalized Electromagnetic Scattering in a Complex Geometry, Nov 1999

A422 Marko Huhtanen
A matrix nearness problem related to iterative methods, Oct 1999

A421 Ph. Clément, G. Gripenberg, V. Högnäs, and S-O. Londen
Some remarks on the method of sums, Oct 1999

A420 Ville Havu and Juhani Pitkäranta
An Analysis of Finite Element Locking in a Parameter Dependent Model Problem, Aug 1999

A419 G. Gripenberg, Ph. Clément, and S-O. Londen
Smoothness in fractional evolution equations and conservation laws, Aug 1999

A418 Marko Huhtanen
Pole assignment problems for error bounds for GMRES, Aug 1999

A417 Kirsi Peltonen
Examples of uniformly quasiregular mappings, Jul 1999

A416 Saara Hyvönen
On the Iterative Solution of Nonnormal Problems, Jul 1999

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The list of reports is continued inside. Electronical versions of the reports are available at *http://www.math.hut.fi/reports/* .

A435 Timo Salin
On Quenching with Logarithmic Singularity, May 2001

A434 Ville Havu
An Analysis of Asymptotic Consistency Error in a Parameter Dependent Model Problem, Mar 2001

A433 Ville Havu and Juhani Pitkäranta
Analysis of a bilinear finite element for shallow shells I: Consistency error, Jan 2001

A432 Tuomas Hytönen
Optimal Wrap-Araoud Network Simulation, Dec 2000

A430 Ville Havu and Juhani Pitkäranta
Analysis of a bilinear finite element for shallow shells I: Approximation of Inextensional Deformations, May 2000

ISBN 951-22-5542-1
ISSN 0784-3143
Espoo, 2001