TECHNISCHE UNIVERSITÄT BERLIN

Index preserving polynomial representation of nonlinear differential-algebraic systems

Volker Mehrmann and Benjamin Unger

Preprint 2015/02

Preprint-Reihe des Instituts für Mathematik
Technische Universität Berlin
http://www.math.tu-berlin.de/preprints

Preprint 2015/02 **February 2015**

Index preserving polynomial representation of nonlinear differential-algebraic systems*

Volker Mehrmann† Benjamin Unger†

February 17, 2015

Abstract

Recently in [9] a procedure was presented that allows to reformulate nonlinear ordinary differential equations in a way that all the nonlinearities become polynomial on the cost of increasing the dimension of the system. We generalize this procedure (called 'polynomialization') to systems of differential-algebraic equations (DAEs). In particular, we show that if the original nonlinear DAE is regular and strangeness-free (i. e., it has differentiation index one) then this property is preserved by the polynomial representation. For systems which are not strangeness-free, i. e., where the solution depends on derivatives of the coefficients and inhomogeneities, we also show that the index is preserved for arbitrary strangeness index. However, to avoid ill-conditioning in the representation one should first perform an index reduction on the nonlinear system and then construct the polynomial representations. Although the analytical properties of the polynomial reformulation are very appealing, care has to be given to the numerical integration of the reformulated system due to additional errors. We illustrate our findings with several examples.

Keywords: differential-algebraic equation, strangeness index, differentiation index, polynomial representation of nonlinear differential-algebraic system, polynomialization, index preservation

AMS(MOS) subject classification: 34A09, 65L80

1 Introduction

In this paper we study nonlinear differential-algebraic equations (DAEs) of the form

$$0 = F(t, x, \dot{x}, g(t, x)), \tag{1.1}$$

where $F : \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \times \mathbb{G} \to \mathbb{R}^n$ is polynomial in its arguments and $g : \mathbb{I} \times \mathbb{D}_x \to \mathbb{G} \subset \mathbb{R}^m$ is a vector valued nonlinear function, for which each entry can be written as a simple combination of elementary nonlinear functions such as trigonometric functions, exponentials or logarithms, etc. with explicit derivatives available. The interval $\mathbb{I} \subset \mathbb{R}$ is closed and $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n, \mathbb{G}$ are open sets.

*Research supported by the Collaborative Research Center 910 *Control of self-organizing nonlinear systems: Theoretical methods and concepts of application.*

†Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, Fed. Rep. Germany, {mehrmann,unger}@math.tu-berlin.de.

Recently, in [9] a so called 'polynomialization' procedure was introduced for implicit nonlinear differential equations (and also control systems which we do not consider here) of the form

$$\frac{d}{dt}\left(q(x)\right) = f(x), \tag{1.2}$$

in the state vector $x : \mathbb{I} \to \mathbb{D}_x \subseteq \mathbb{R}^n$, where $q, f : \mathbb{D}_x \to \mathbb{R}^n$, and each entry of q, f is a function $q_i, f_i : \mathbb{D}_x \to \mathbb{R}$ for $i = 1, \ldots, n$, respectively. It is assumed that q_i and f_i are sums of the form,

$$\begin{aligned} q_i(x) &= q_{i,1}(x) + \ldots + q_{i,\ell_i}(x), \\ f_i(x) &= f_{i,1}(x) + \ldots + f_{i,k_i}(x), \quad i = 1, \ldots, n, \end{aligned} \tag{1.3}$$

where the functions $q_{i,j}, f_{i,j}$ are elementary nonlinear functions. For $i = 1, \ldots, n$, one introduces new variables

$$\begin{aligned} \phi_{i,1} &= q_{i,1}(x), \ldots, \phi_{i,\ell_i} = q_{i,\ell_i}(x), \\ \zeta_{i,1} &= f_{i,1}(x), \ldots, \zeta_{i,k_i} = f_{i,k_i}(x), \end{aligned} \tag{1.4}$$

and obtains a reformulated system (in the variables $\phi_{i,j}, \zeta_{i,r}$) given by

$$\frac{d}{dt}\left(\phi_{i,1}(t) + \ldots + \phi_{i,\ell_i}(t)\right) = \zeta_{i,1}(t) + \ldots + \zeta_{i,k_i}(t), \ i = 1, \ldots, n.$$

To have the same number of equations and unknowns, and to enforce the algebraic constraints (1.4) one adds the equations

$$\begin{aligned} \dot{\phi}_{i,1} &= \left(q_{i,1}\right)_x \dot{x}, \ldots, \dot{\phi}_{i,\ell_i} = \left(q_{i,\ell_i}\right)_x \dot{x}, \\ \dot{\zeta}_{i,1} &= \left(f_{i,1}\right)_x \dot{x}, \ldots, \dot{\zeta}_{i,k_i} = \left(f_{i,k_i}\right)_x \dot{x}, \quad i = 1, \ldots, n. \end{aligned}$$

Collecting all the variables $x_i,\ i = 1, \ldots, n,\ \phi_{i,j},\ j = 1, \ldots, \ell_i$, and $\zeta_{i,j},\ j = 1, \ldots, k_i$ in a vector y, one obtains an implicit system of the form

$$L(y)\dot{y} = R(y),$$

where R is a linear function in y. If in addition the Jacobians $\left(\phi_{i,j}\right)_x, \left(\zeta_{i,k}\right)_x$ have a polynomial representation, then the implicit nonlinear system has been turned into a *quasi-linear* DAE for y with a linear right-hand side and polynomial leading matrix L.

In this paper we extend this 'polynomialization' approach to more general DAEs of the form (1.1), but we refrain here from using this terminology because it may lead to confusion with similar concepts in other areas of mathematics. We rather speak of *polynomial representation of the DAE*.

In (1.1) we substitute the non-polynomial part g by $z(t) = g(t, x)$ and add the differential equation $\dot{z} = g_x\dot{x} + g_t$ to obtain an extended DAE system in the vector function $y = \begin{bmatrix} x^T & z^T \end{bmatrix}^T$ given by

$$0 = \tilde{F}(t, y, \dot{y}) = \begin{bmatrix} F(t, x, \dot{x}, z) \\ \dot{z} - g_x\dot{x} - g_t \end{bmatrix}, \tag{1.5}$$

where g_x, g_t denote the partial derivatives of g with respect to x, t, respectively.

Remark 1.1 If the entries of the Jacobians g_x and g_t can be written as a polynomial in y and F is quasi-linear, i. e., it has the form

$$F(t, x, \dot{x}, g(t, x)) = E(t, x, g(t, x))\dot{x} - K(t, x, g(t, x)),$$

with entrywise polynomial functions $E : \mathbb{I} \times \mathbb{D}_x \times \mathbb{G} \to \mathbb{R}^{n,n}$ and $K : \mathbb{I} \times \mathbb{D}_x \times \mathbb{G} \to \mathbb{R}^n$, then the polynomial representation yields a quasi-linear DAE of the form $\tilde{E}(t,y)\dot{y} = \tilde{K}(t,y)$, where $\tilde{E}$ and $\tilde{K}$ are polynomial in y.

Example 1.2 The polynomial reformulation of the differential-algebraic equation

$$0 = F(t,x,\dot{x},g(t,x)) := \begin{bmatrix} \dot{x}_1 + \dot{x}_2 \\ x_1 - e^{x_1}x_2 \end{bmatrix}$$

with nonlinear term $g(t,x) = e^{x_1}$ is given by

$$0 = \tilde{F}(t,y,\dot{y}) = \begin{bmatrix} \dot{x}_1 + \dot{x}_2 \\ x_1 - zx_2 \\ \dot{z} - z\dot{x}_1 \end{bmatrix}, \tag{1.6}$$

where we substituted $z = g(t,x) = \exp(x_1)$ and set $y = \begin{bmatrix} x_1 & x_2 & z \end{bmatrix}^T$. The polynomial representation (1.6) is a quasi-linear DAE of the form $E(y)\dot{y} = K(y)$, where E is linear and K bilinear in y.

It is obvious that (1.2) is a special case of (1.1). The reformulation (1.5) generalizes the aforementioned 'polynomialization' procedure to general DAEs and allows for a time-dependency of the nonlinear term. In particular, the results obtained in Section 3 are also valid for the original method by Gu [9]. A further generalization to nonlinear terms of the form $g(t,x,\dot{x})$ seems possible but is beyond the scope of this paper. Note that in this situation the resulting system is of higher order, which requires additional treatment [18].

The Carleman bilinearization [7, 22] is another simplification procedure, which approximates general nonlinear ordinary differential equations (ODEs) by systems that are quadratic and even bilinear in the state variable x. In contrast, the discussed polynomial representation yields an exact representation on the cost of using a higher state space dimension. The motivation for such a reformulation procedure is the generation of a more accessible structure on the right-hand side of ODEs or quasi-linear DAEs that can be exploited in applications such as model order reduction [2, 4] or the regularization of delay differential-algebraic equations (DDAEs) [11]. Moreover, the polynomial structure can be used within computer algebra systems, for example to compute a smooth singular value decomposition [21].

Example 1.3 An application of the polynomial reformulation procedure is the regularization of quasi-linear DAEs with linear delay term of the form

$$E(t,x)\dot{x} = K(t,x) + B(t)\Delta_\tau x + f(t), \tag{1.7}$$

where $E : \mathbb{I} \times \mathbb{D}_x \to \mathbb{R}^{n,n}$ is a pointwise singular matrix function with constant rank r on some solution space $\mathbb{L} \subseteq \mathbb{I} \times \mathbb{D}_x$, $K : \mathbb{I} \times \mathbb{D}_x \to \mathbb{R}^n$ a nonlinear function, $B : \mathbb{I} \to \mathbb{R}^{n,n}$ a matrix function, $f : \mathbb{I} \to \mathbb{R}^n$ the inhomogeneity and for given $\tau > 0$, the operator Δ_τ is the (backward) shift operator, i. e., $(\Delta_\tau x)(t) = x(t-\tau)$. The regularization procedure for DDAEs requires a series of algebraic manipulations, differentiations, and applications of shift operators [11]. The extension of the regularization procedure of [11] to systems of the form (1.7) requires (loosely speaking) the following steps. Remove redundancies in E, K and B to

obtain the system

$$\begin{bmatrix} E_1(t,x) \\ 0 \\ 0 \\ 0 \end{bmatrix} \dot{x} = \begin{bmatrix} K_1(t,x) \\ K_2(t,x) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} B_1(t) \\ B_2(t) \\ B_3(t) \\ 0 \end{bmatrix} \Delta_\tau x + \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix},$$

where the last row gives a consistency condition for the inhomogeneity. Then iteratively remove the redundancies in

$$\begin{bmatrix} K_2(t,x)^T & (B_3(t+\tau)x)^T \end{bmatrix}^T \qquad \text{and} \qquad \begin{bmatrix} E_1(t,x)^T & K_{2,x}(t,x)^T & B_3(t+\tau)^T \end{bmatrix}^T$$

to obtain a *shift-* and *strangeness-free* system [11]. Numerically removing redundancies is performed via rank revealing decompositions, which however may be computationally infeasible, in particular for large-scale vector-valued nonlinear functions, which require a good approximations of the solution trajectory for the linearization and the computation of the Jacobians. Hence the numerical implementation of the regularization procedure is a difficult task. However, if we denote by $\bigotimes_{j=1}^{i} y$ the i times Kronecker product of y, then the polynomial representation of (1.7) can be written as

$$\tilde{E}(t,y)\dot{y} = \sum_{i=1}^{N_k} \tilde{K}_i(t) \bigotimes_{j=1}^{i} y + \tilde{B}(t)\Delta_\tau y + \tilde{f}(t) \qquad \text{with } \tilde{K}_i : \mathbb{I} \to \mathbb{R}^{n,n^i},$$

and the required row compressions can be accomplished in a much easier way, because the Jacobians are easily available and it is much easier to work pointwise.

If $\frac{\partial}{\partial \dot{x}} F$ is nonsingular, then the DAE (1.1) is (locally) equivalent to an ODE and so is the reformulated system (1.5). However, in the case of DAEs with singular $\frac{\partial}{\partial \dot{x}} F$, the solution x usually depends (implicitly) on derivatives of F and the required degree of differentiability is classified by one of many different index concepts [15, 17]. Furthermore, the set of possible initial conditions is restricted [15]. If more than one derivative is implicitly required, then this restricts the class of numerical integration methods for (1.1) [13, 15] and for such higher index problems one typically observes order reduction, ill-conditioning of the nonlinear systems or even divergence of the numerical method. For this reason an index reduction [15] is performed, which reformulates the system as a new DAE with the same solution set but that is better suited for numerical integration and control, and from which consistent initial conditions can be deduced.

In this paper we investigate whether the extension of the polynomial representation procedure of [9] to DAEs increases the (strangeness) index of the DAE and how it performs when combined with index reduction. The main contribution of the paper is given in section 3, where we show that under some further differentiability assumption on the nonlinear function g and a restriction of the solution manifold, the strangeness index is preserved during the reformulation. Section 4 is dedicated to some numerical observations and remarks. We illustrate the findings with examples.

2 Notation and Index Concept

In the following we denote by $x^{(j)}$ the j-th (time) derivative of a function $x : \mathbb{I} \to \mathbb{R}^n$, where $\mathbb{I} \subseteq \mathbb{R}$ is a closed interval, but use the short versions $\dot{x} = x^{(1)}$ and $\ddot{x} = x^{(2)}$. For partial

derivatives we use the short notation $\frac{\partial}{\partial x} f = f_x$. If $g : \mathbb{I} \times \mathbb{D}_x \to \mathbb{R}^m$, then the Hessian g_{xx} is a tensor and $g_{xx}xx$ is short hand for $(g_{xx}x)x$.

Consider a DAE of the form

$$F(t, x, \dot{x}) = 0 \tag{2.1}$$

with $F : \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \to \mathbb{R}^n$ and $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open. We use the concept of classical solutions, i. e., a one time continuously differentiable function $x \in C^1(\mathbb{I}, \mathbb{R})$ is a solution of (2.1) if x satisfies (2.1) pointwise. An initial condition $x(t_0) = x_0 \in \mathbb{R}^n$ is called *consistent*, if the associated initial value problem

$$F(t, x, \dot{x}) = 0, \qquad x(t_0) = x_0 \tag{2.2}$$

has at least one solution. For this paper, we assume that (2.1) is *regular*, i. e., (2.2) has a unique solution for every consistent initial condition. Note that (1.1) is a special case of (2.1), such that the concepts introduced in this section are valid for (1.1). As discussed in the introduction, the solution of (2.2) often implicitly depends on derivatives of F, and the degree of differentiability is classified by one of a variety of *index concepts* [17]. Here, we employ the *strangeness index* concept [15], which is based on *derivative arrays* [5]. Consider the derivative array of level ℓ defined by

$$F_\ell\left(t, x, \dot{x}, \ldots, x^{(\ell+1)}\right) = \begin{bmatrix} F(t, x, \dot{x}) \\ \frac{d}{dt} F(t, x, \dot{x}) \\ \vdots \\ \left(\frac{d}{dt}\right)^\ell F(t, x, \dot{x}) \end{bmatrix},$$

and the Jacobians

$$\begin{aligned} M_\ell\left(t, x, \dot{x}, \ldots, x^{(\ell+1)}\right) &= F_{\ell;\dot{x},\ldots,x^{(\ell+1)}}\left(t, x, \dot{x}, \ldots, x^{(\ell+1)}\right), \\ N_\ell\left(t, x, \dot{x}, \ldots, x^{(\ell+1)}\right) &= -\begin{bmatrix} F_{\ell;x}\left(t, x, \dot{x}, \ldots, x^{(\ell+1)}\right) & 0 & \cdots & 0 \end{bmatrix} \end{aligned}$$

both of dimension $(\ell+1)n \times (\ell+1)n$, where $F_{\ell;\dot{x},\ldots,x^{(\ell+1)}}$ is short hand for

$$F_{\ell;\dot{x},\ldots,x^{(\ell+1)}} = \begin{bmatrix} F_{\ell;\dot{x}} & F_{\ell;\ddot{x}} & \cdots & F_{\ell;x^{(\ell+1)}} \end{bmatrix}$$

and not to be confused with the tensor notation above. We introduce the following hypothesis, see [15, Hypothesis 4.2], written in slightly different form.

Hypothesis 2.1 *There exist integers μ, a_μ and d_μ such that the set*

$$\mathbb{L}_\mu = \left\{ \left(t, x, \dot{x}, \ldots, x^{(\mu+1)}\right) \in \mathbb{R}^{(\mu+2)n+1} \middle| F_\mu\left(t, x, \dot{x}, \ldots, x^{(\mu+1)}\right) = 0 \right\}$$

associated with F is nonempty and such that for every $(t_0, x_0, \dot{x_0}, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$, there exists a (sufficiently small) neighborhood in which the following properties hold:

1. *We have* $\operatorname{rank}(M_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)})) = (\mu+1)n - a_\mu$ *on $\mathbb{L}_\mu$ such that there exists a smooth matrix function $Z_{2,\mu}$ of size $(\mu+1)n \times a_\mu$ and pointwise maximal rank, satisfying $Z_{2,\mu}^T M_\mu = 0$ on $\mathbb{L}_\mu$.*

2. *We have* $\operatorname{rank}(A_{2,\mu}(t,x,\dot{x},\ldots,x^{(\mu+1)})) = a_\mu$, *where* $A_{2,\mu} = Z_{2,\mu}^T N_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T$ *such that there exists a smooth matrix function* $T_{2,\mu}$ *of size* $n \times d_\mu, d_\mu = n - a_\mu$ *and pointwise maximal rank, satisfying* $A_{2,\mu}T_{2,\mu} = 0$.

3. *We have* $\operatorname{rank}(F_{\dot{x}}(t,x,\dot{x})T_{2,\mu}(t,x,\dot{x},\ldots,x^{(\mu+1)})) = d_\mu$ *such that there exists a smooth matrix function* $Z_{1,\mu}$ *of size* $n \times d_\mu$ *and pointwise maximal rank, satisfying* $\operatorname{rank}(E_{1,\mu}T_{2,\mu}) = d_\mu$, *where* $E_{1,\mu} = Z_{1,\mu}^T F_{\dot{x}}$.

The strangeness index is defined by using Hypothesis 2.1, see Definition 4.4 in [15].

Definition 2.2 (Strangeness index) *Consider a DAE of the form* (2.1). *The smallest value of* μ *such that* F *satisfies Hypothesis 2.1 is called the* strangeness index *of* (2.1). *If* $\mu = 0$, *then the differential-algebraic equation is called* strangeness-free.

In the following, we suppress the subscript μ for better readability, whenever it is obvious from the context.

In section 3 we consider a special case of the hypothesis which assumes that each matrix M_k for $k = 0,\ldots,\mu$ has constant rank. This additional assumption was included in the original definition of the strangeness index [14] and is weaker than the general hypothesis where the rank only has to be constant for M_μ. The following Theorem 2.3 suggests a piecewise consideration of the problem if the constant rank assumption is violated. It is taken from [15] and is a Corollary of Theorem 10.5.2 from [6].

Theorem 2.3 *Let* $\mathbb{I} \subseteq \mathbb{R}$ *be a closed interval and* $M \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$. *Then there exist open intervals* $\mathbb{I}_j \subseteq \mathbb{I}, j \in \mathbb{N}$, *with*

$$\overline{\bigcup_{j\in\mathbb{N}} \mathbb{I}_j} = \mathbb{I}, \qquad \mathbb{I}_i \cap \mathbb{I}_j = \emptyset \quad \text{for } i \neq j, \tag{2.3}$$

and integers $r_j \in \mathbb{N}_0$, $j \in \mathbb{N}$, *such that*

$$\operatorname{rank} M(t) = r_j \quad \text{for all } t \in \mathbb{I}_j.$$

In the case of DDAEs, see for example [1] and the references within, so called *breaking points* [10] enforce a piecewise smooth solution concept. For constant delay $\tau > 0$ the breaking points are given as the integer multiples of τ and the strangeness index (if defined) may differ on the intervals $[(\ell-1)\tau, \ell\tau]$ for $\ell \in \mathbb{N}$. Hence, a restriction to subintervals $\mathbb{I}_j \subseteq \mathbb{I}$ might be necessary and in general, the constant rank assumptions (even for Hypothesis 2.1) only hold on subintervals. A further reason for the restriction to subintervals is imparted in Example 2.4.

Example 2.4 Consider the DAE

$$0 = F(t,x,\dot{x},g(t,x)) = \begin{bmatrix} \dot{x}_1 - |x_1| - \exp(x_2) \\ \sin(x_2) \end{bmatrix} \qquad \text{with} \qquad g(t,x) = \begin{bmatrix} |x_1| \\ \exp(x_2) \\ \sin(x_2) \end{bmatrix} \tag{2.4}$$

for $t \in \mathbb{I} := [0,\infty)$. Obviously, g has no explicit derivative due to the absolute value in the first entry. If the initial condition $x(t_0) = x_0 = \begin{bmatrix} x_{0,1} & x_{0,2} \end{bmatrix}^T$ satisfies $x_{0,1} \geq 0$, then it is easy to see that $x_1(t) \geq 0$ for all t and hence $|x_1| = x_1$ is already in polynomial form. Otherwise,

the sign of x_1 changes at $t = \log(\exp(x_{0,2}) - x_{0,1}) - x_{0,2}$ from -1 to 1 and we cannot expect a global polynomial representation. Instead we handle (2.4) piecewise, such that g reduces to $g(t,x) = \begin{bmatrix}\exp(x_2) & \sin(x_2)\end{bmatrix}$, and a polynomial representation is given by

$$0 = \tilde{F}_1(t,y,\dot{y}) = \begin{bmatrix} \dot{x}_1 + x_1 - z_1 \\ z_2 \\ \dot{z}_1 - z_1\dot{x}_2 \\ \dot{z}_2 - z_3\dot{x}_2 \\ \dot{z}_3 + z_2\dot{x}_2 \end{bmatrix}$$

if $x_{0,1} < 0$ and $t < \log(\exp(x_{0,2}) - x_{0,1}) - x_{0,2}$, and

$$0 = \tilde{F}_2(t,y,\dot{y}) = \begin{bmatrix} \dot{x}_1 - x_1 - z_1 \\ z_2 \\ \dot{z}_1 - z_1\dot{x}_2 \\ \dot{z}_2 - z_3\dot{x}_2 \\ \dot{z}_3 + z_2\dot{x}_2 \end{bmatrix}$$

otherwise. Note that a second reformulation step with $z_3 = \cos(x_2)$ was applied to obtain the above quasi-linear systems.

The occurrence of an absolute value or a signum function is typical for *hybrid* systems, i. e., systems that switch between multiple models. The index, the number of algebraic constraints and free variables may change in such systems and discontinuities or jumps can occur at the switching points [19].

Lemma 2.5 *Suppose that* (2.1) *is regular and has strangeness index* μ. *If for* $k = 0, 1, \ldots, \mu$ *exist integers* a_k *such that rank*$(M_k) = (k+1) - a_k$ *on* $\mathbb{L}_k$, *then*

$$\operatorname{rank}(A_{2,k}) = a_k \text{ for all } \ k = 0, 1, \ldots, \mu.$$

Proof. To proceed with a proof by contradiction, assume that there exists $k_0 < \mu$ such that $\operatorname{rank}(A_{2,k_0}) < a_{k_0}$. Then the matrix M_{k_0+1} is of the form

$$M_{k_0+1} = \begin{bmatrix} M_{k_0} & 0 \\ * & M_0 \end{bmatrix},$$

and hence Z_{2,k_0+1} is given by

$$Z_{2,k_0+1} = \begin{bmatrix} Z_{2,k_0} & Y_{k_0+1} \\ 0 & Z_{s_{k_0+1}} \end{bmatrix} \in \mathbb{R}^{(k_0+2)n, a_{k_0+1}} \quad \text{with } \operatorname{rank}\left(Z_{s_{k_0+1}}\right) = a_{k_0+1} - a_{k_0} \geq 0,$$

where the columns of $Z_{s_{k_0+1}}$ span a subspace of $\operatorname{range}(Z_{2,0})$. We have

$$\begin{aligned} A_{2,k_0+1} &= Z_{2,k_0+1}^T N_{k_0+1} \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T = -Z_{2,k_0+1}^T F_{k_0+1;x} \\ &= -\begin{bmatrix} Z_{2,k_0}^T & 0 \\ Y_{k_0+1}^T & Z_{s_{k_0+1}}^T \end{bmatrix} \begin{bmatrix} F_{k_0;x} \\ \frac{\partial}{\partial x}\left(\frac{d}{dt}\right)^{k_0+1} F(t,x,\dot{x}) \end{bmatrix} \\ &= -\begin{bmatrix} Z_{2,k_0}^T F_{k_0;x} \\ Y_{k_0+1}^T F_{k_0;x} + Z_{s_{k_0+1}}^T \frac{\partial}{\partial x}\left(\frac{d}{dt}\right)^{k_0+1} F(t,x,\dot{x}) \end{bmatrix} \in \mathbb{R}^{a_{k_0+1}, n}. \end{aligned}$$

Hence, $\operatorname{rank}(A_{2,k_0+1}) \leq \operatorname{rank}(Z_{2,k_0}^T F_{k_0;x}) + a_{k_0+1} - a_{k_0} < a_{k_0} + a_{k_0+1} - a_{k_0} = a_{k_0+1}$, which is a contradiction to the existence of the strangeness index. □

As a direct consequence, we get the following result.

Lemma 2.6 *Let F be regular with $F_{\dot{x}}$ having constant rank on $\mathbb{L}_0$. A necessary condition for F to have a well-defined strangeness index is that* $\operatorname{rank}(\begin{bmatrix} F_x & F_{\dot{x}} \end{bmatrix}) = n$ *on $\mathbb{L}_0$.*

Proof. Let Z_2 be as in Hypothesis 2.1 and complete Z_2 to a nonsingular matrix $Z = \begin{bmatrix} Z_2' & Z_2 \end{bmatrix}$. The assertion then follows directly from

$$\begin{bmatrix} F_x & F_{\dot{x}} \end{bmatrix} = \begin{bmatrix} -N_0 & M_0 \end{bmatrix} = Z^{-T} \begin{bmatrix} -(Z_2')^T N_0 & (Z_2')^T M_0 \\ -A_2 & 0 \end{bmatrix}. \qquad \square$$

The results of Lemmas 2.5 and 2.6 are illustrated in the following example.

Example 2.7 The algebraic equation

$$0 = F(t, x, \dot{x}) = x^2$$

is a special case of a DAE. We have $M_0 = 0$, $Z_2 = 1$ and $A_2 = Z_2^T N_0 = -2x$ has constant rank zero on $\mathbb{L}_0$. According to Lemma 2.5, F has no strangeness index, and we have indeed

$$\operatorname{rank}(\begin{bmatrix} F_x & F_{\dot{x}} \end{bmatrix}) = \operatorname{rank} \begin{bmatrix} 2x & 0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} 0 & 0 \end{bmatrix} = 0 < 1 \text{ on } \mathbb{L}_0$$

in agreement with Lemma 2.6.

3 Index Preservation in Polynomial Representations

In this section we discuss the effect of the polynomial representation on the strangeness index of the DAE. Let us first illustrate the difficulties with an example.

Example 3.1 Verifying Hypothesis 2.1 for Example 1.2 yields the characteristic values $\mu = 0$, $a_0 = 1$, and $d_0 = 1$. The matrices are given by

$$Z_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad T_2 = \begin{bmatrix} 1 \\ \frac{1-e^{x_1}x_2}{e^{x_1}} \end{bmatrix}, \qquad Z_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and the DAE is already in strangeness-free form. Considering the polynomial representation $\tilde{F}$, we check Hypothesis 2.1 for the extended system and mark the corresponding characteristic values and matrices with '$\sim$'. The Jacobians are given by

$$\tilde{M}_0 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ -y_3 & 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \tilde{N}_0 = \begin{bmatrix} 0 & 0 & 0 \\ -1 & y_3 & y_2 \\ 0 & 0 & \dot{y}_1 \end{bmatrix}.$$

Since $\mathbb{L}_0$ is nonempty, so is $\tilde{\mathbb{L}}_0$ by simply setting $y_3 = g(t, x), \dot{y_3} = g_t(t, x) + g_x(t, x)\dot{x}$ for $(t, x, \dot{x}) \in \mathbb{L}_0$. Moreover, $\operatorname{rank}(\tilde{M}_0) = 2$, which implies $\tilde{a}_0 = 1$ and accordingly, we have

$\operatorname{rank}(\tilde{A}_{2,0}) = 2 = \tilde{a}_0$. The relevant matrices in Hypothesis 2.1 for the reformulated system (1.5) are given by

$$\tilde{Z}_{2,0} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \tilde{T}_{2,0} = \begin{bmatrix} y_3 & y_2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \tilde{F}_{0,\dot{x}}\tilde{T}_{2,0} = \begin{bmatrix} y_3 + 1 & y_2 \\ 0 & 0 \\ -y_3^2 & -y_3 y_2 + 1 \end{bmatrix}.$$

The product $\tilde{F}_{0,\dot{x}}\tilde{T}_{2,0}$ has rank 2 for $y_1 = 0, y_2 = 0, y_3 = \exp(y_1)$ and rank 1 for $y_1 = 0$, $y_2 = 0$, $y_3 = -1$ and is hence not constant on $\tilde{\mathbb{L}}_0$. Thus, Hypothesis 2.1 does not hold for the extended system and $\tilde{\mu} = 0$. This means that in general we cannot expect that the polynomial representation is compatible with the standard Hypothesis and therefore with the standard index reduction procedure.

As Example 3.1 suggests, in order to compare the standard and the reformulated system we may have to modify the Hypothesis for the extended system. To do this and to analyze the behavior of the reformulated DAE (1.5), we consider the restricted solution manifold

$$\tilde{\mathbb{L}}_0^r = \left\{ (t, x, z, \dot{x}, \dot{z}) \in \mathbb{R}^{2(n+m)+1} \middle| (t, x, \dot{x}) \in \mathbb{L}_0,\ z = g(t, x) \right\},$$

and we have the following lemma.

Lemma 3.2 *Let* (1.1) *be regular and strangeness-free. Then the reformulated DAE* (1.5) *satisfies* $\operatorname{rank}(\begin{bmatrix} \tilde{F}_y & \tilde{F}_{\dot{y}} \end{bmatrix}) = n + m$ *on* $\tilde{\mathbb{L}}_0^r$.

Proof. We have

$$\operatorname{rank}(\begin{bmatrix} \tilde{F}_y & \tilde{F}_{\dot{y}} \end{bmatrix}) = \operatorname{rank}\left(\begin{bmatrix} F_x & F_z & F_{\dot{x}} & 0 \\ -g_{xx}\dot{x} - g_{tx} & 0 & -g_x & I_m \end{bmatrix} \right) = \operatorname{rank}\left(\begin{bmatrix} F_x & F_z & F_{\dot{x}} \end{bmatrix} \right) + m,$$

where I_m is the $m \times m$ identity matrix. Let Z_2 span the left nullspace of $F_{\dot{x}}$ (as in Hypothesis 2.1) and let Z_2' complete Z_2 to a nonsingular matrix. Then,

$$\begin{bmatrix} (Z_2')^T & Z_2^T \end{bmatrix} \begin{bmatrix} F_x & F_z & F_{\dot{x}} \end{bmatrix} = \begin{bmatrix} (Z_2')^T F_x & (Z_2')^T F_z & (Z_2')^T F_{\dot{x}} \\ Z_2^T F_x & Z_2^T F_z & 0 \end{bmatrix}$$

on $\tilde{\mathbb{L}}_0^r$, where $(Z_2')^T F_{\dot{x}}$ has full row rank by Lemma 2.6. Suppose that the assertion is false, i. e., $\begin{bmatrix} Z_2^T F_x & Z_2^T F_z \end{bmatrix}$ is rank deficient. This implies that also

$$Z_2^T \left(F_x + F_z g_x \right) = Z_2^T \begin{bmatrix} F_x & F_z \end{bmatrix} \begin{bmatrix} I_n \\ g_x \end{bmatrix}$$

is rank deficient, which is a contradiction to Lemma 2.6. Hence, the matrix function

$$\begin{bmatrix} Z_2^T F_x & Z_2^T F_z \end{bmatrix} = Z_2^T \begin{bmatrix} F_x & F_z \end{bmatrix} \tag{3.1}$$

has pointwise full row rank and the result follows. □

Theorem 3.3 *Suppose that the DAE* (1.1) *is regular and strangeness-free with characteristic values* a *and* d. *If* F *is continuously differentiable and* g *is twice continuously differentiable, i. e.,* $g \in C^2(\mathbb{I} \times \mathbb{D}_x, \mathbb{G})$, *then* (1.5) *is strangeness-free on* $\tilde{\mathbb{L}}_0^r$ *with characteristic values* $\tilde{a} = a$ *and* $\tilde{d} = d + m$.

Proof. First note that, if the initial condition for z is given by $z_0 = g(t_0, x_0)$, then the unique solution of $\dot{z}(t) = g_t(t,x) + g_x(t,x)\dot{x}(t)$ is given by $z = g(t,x)$. Hence, for such initial conditions the reformulated DAE (1.5) is regular and we can check Hypothesis 2.1. As before, all characteristic values and matrices corresponding to (1.5) are marked with '$\sim$'. The restricted solution manifold $\tilde{\mathbb{L}}_0^r$ is nonempty by definition and in the remainder of this proof, all investigations are performed on $\tilde{\mathbb{L}}_0^r$, i. e., we have $z = g(t,x)$ for all $(t,x,\dot{x}) \in \mathbb{L}_0$. We have

$$\operatorname{rank}(\tilde{M}_0) = \operatorname{rank}\begin{bmatrix} F_{\dot{x}} & 0 \\ -g_x & I_m \end{bmatrix} = \operatorname{rank}(M_0) + m = (n+m) - a.$$

Let Z_2 be as in Hypothesis 2.1 with pointwise maximal rank. Then $\tilde{Z}_2 = \begin{bmatrix} Z_2^T & 0 \end{bmatrix}^T \in \mathbb{R}^{n+m,a}$ has pointwise maximal rank and spans the left nullspace of $\tilde{M}_0$. The Jacobians with respect to x and y of the original and the reformulated system are given by

$$N_0 = -F_x - F_z g_x, \qquad \tilde{N}_0 = -\tilde{F}_y = -\begin{bmatrix} F_x & F_z \\ -g_{tx} - g_{xx}\dot{x} & 0 \end{bmatrix},$$

respectively. Hence, the matrix-function $\tilde{A}_2$ is given by

$$\tilde{A}_2 = \tilde{Z}_2^T \tilde{N}_0 = -Z_2^T \begin{bmatrix} F_x & F_z \end{bmatrix}$$

with $\operatorname{rank}(\tilde{A}_2) = a$ by Lemma 3.2 and (3.1), and thus $\tilde{a} = a$. It follows that the kernel of $\tilde{A}_2$ has dimension $n + m - \tilde{a} = d + m$, i. e., $\tilde{d} = d + m$. For $v \in \operatorname{kernel}(A_2)$ we have

$$\tilde{A}_2 \begin{bmatrix} v \\ g_x v \end{bmatrix} = -Z_2^T \left(F_x v + F_z g_x v\right) = -Z_2^T \left(F_x + F_z g_x\right) v = A_2 v = 0.$$

Since $\dim(\tilde{T}_2) = d + m$, there exists a matrix $V = \begin{bmatrix} V_1^T & V_2^T \end{bmatrix}^T$ such that its columns span a subspace of $\operatorname{kernel}(\tilde{A}_2)$, and at the same time

$$g_x(t,x)V_1 - V_2$$

is pointwise nonsingular. Setting $\begin{bmatrix} \tilde{V}_1 & \tilde{V}_2 \end{bmatrix} = \tilde{V} = V\left(-g_x(t,x)V_1 + V_2\right)^{-1}$ and

$$\tilde{T}_2 = \begin{bmatrix} T_2 & \tilde{V}_1 \\ g_x(t,x)T_2 & \tilde{V}_2 \end{bmatrix}, \qquad \tilde{Z}_1 = \begin{bmatrix} Z_1 & 0 \\ 0 & I_m \end{bmatrix},$$

we have

$$\tilde{F}_{0,\dot{y}}\tilde{T}_2 = \begin{bmatrix} F_{\dot{x}} & 0 \\ -g_x(t,x) & I_m \end{bmatrix} \begin{bmatrix} T_2 & \tilde{V}_1 \\ g_x T_2 & \tilde{V}_2 \end{bmatrix} = \begin{bmatrix} F_{\dot{x}} T_2 & F_{\dot{x}} \tilde{V}_1 \\ 0 & I_m \end{bmatrix},$$

which clearly has rank $d + m = \tilde{d}$ on $\tilde{\mathbb{L}}_0^r$. It is obvious that $\operatorname{rank}(\tilde{Z}_1^T \tilde{F}_{\dot{y}} \tilde{T}_2) = d + m$ and Hypothesis 2.1 is satisfied for $\tilde{\mu} = 0, \tilde{a} = a$ and $\tilde{d} = d + m$. In particular, $\tilde{F}$ is strangeness-free on $\tilde{\mathbb{L}}_0^r$. □

Remark 3.4 If g acts only on the differential equations, i. e., $Z_2^T F_z = 0$, then the kernel of $\tilde{A}_2$ can be chosen as

$$\tilde{T}_2 = \begin{bmatrix} T_2 & 0 \\ 0 & I_m \end{bmatrix} \in \mathbb{R}^{n+m,d+m}.$$

In particular, this is true in the case of implicit ordinary differential equations.

Remark 3.5 For the numerical integration of general DAEs, a strangeness-free form is desirable. If $\mu > 0$, then the strangeness-free form can be obtained (locally) via choosing pointwise orthonormal matrices Z_1 and Z_2. Since most numerical integration methods are invariant under multiplications with nonsingular matrices from the left, it is not even necessary to choose smooth projections Z_1, Z_2, it is enough that there exist such smooth transformations [8, 15]. These matrices are commonly choosen to be orthonormal to avoid ill-conditioning. Performing the polynomial representation prior to the strangeness-free reformulation in general then again leads to a non-polynomial system (see Example 3.6 below). Thus, it is advisable to first apply the index reduction procedure before computing the polynomial representation.

Example 3.6 We perform the polynomial reformulation to the dynamical system given by

$$0 = F(t, x, \dot{x}, g(t, x)) = \begin{bmatrix} \dot{x}_1 - e^{x_2} \\ x_2\dot{x}_1 + x_2(1 - e^{x_2}) - 1 \end{bmatrix}$$

with $g(t, x) = \exp(x_2)$ and obtain the system

$$0 = \tilde{F}(t, y, \dot{y}) = \begin{bmatrix} \dot{x}_1 - z \\ x_2\dot{x}_1 + x_2(1 - z) - 1 \\ \dot{z} - z\dot{x}_2 \end{bmatrix}.$$

The computation of an orthonormal matrix function $\tilde{Z}_2$ as in Hypothesis 2.1 can be performed via (smooth) QR-decomposition of $\tilde{M}_0 = \tilde{F}_{\dot{x}}$ given by

$$\tilde{M}_0 = \begin{bmatrix} 1 & 0 & 0 \\ x_2 & 0 & 0 \\ 0 & -z & 1 \end{bmatrix} = \frac{1}{\sqrt{1 + x_2^2}} \begin{bmatrix} 1 & 0 & -x_2 \\ x_2 & 0 & 1 \\ 0 & \sqrt{1 + x_2^2} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{1 + x_2^2} & 0 & 0 \\ 0 & -z & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

which implies $\tilde{Z}_2^T = \left(1 + x_2^2\right)^{-\frac{1}{2}} \begin{bmatrix} -x_2 & 1 & 0 \end{bmatrix}$. The strangeness-free form is given as

$$0 = \begin{bmatrix} \tilde{Z}_1^T \tilde{F} \\ \tilde{Z}_2^T \tilde{F}_0 \end{bmatrix} = \begin{bmatrix} \dot{x}_1 - z \\ \dot{z} - z\dot{x}_2 \\ \frac{x_2 - 1}{\sqrt{1 + x_2^2}} \end{bmatrix} \qquad \text{with} \qquad \tilde{Z}_1^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which is not in polynomial form. Introducing a second variable $z_2 = (1 + x_2^2)^{-\frac{1}{2}}$ would then lead to a strangeness-free system in polynomial form of dimension four. On the other side, performing the strangeness-free reformulation analytically (i. e., we do not require $\tilde{Z}_2$ to be orthonormal), we obtain the three dimensional system

$$0 = \begin{bmatrix} \tilde{Z}_1^T \tilde{F} \\ \tilde{Z}_2^T \tilde{F}_0 \end{bmatrix} = \begin{bmatrix} \dot{x}_1 - z \\ \dot{z} - z\dot{x}_2 \\ x_2 - 1 \end{bmatrix}.$$

For a DAE (1.1) with strangeness-index $\mu > 0$, one can first perform a strangeness-free reformulation $\hat{F}$ and then determine the polynomial representation and hence Theorem 3.3 applies. On the other hand, most algorithms, see e. g. [16], perform the index reduction along the time integration. A polynomial reformulation is not possible due to the local character of

the time integration and the reformulation must be applied beforehand. We first investigate the case $\mu = 1$. Similarly as before, we introduce the set

$$\tilde{\mathbb{L}}_1^r = \left\{ (t, x, z, \dot{x}, \dot{z}, \ddot{x}, \ddot{z}) \in \mathbb{R}^{3(n+m)+1} \;\middle|\; (t, x, \dot{x}, \ddot{x}) \in \mathbb{L}_0, z = g(t, x), \dot{z} = g_x(t, x)\dot{x} + g_t(t, x) \right\}$$

and perform all computation on the restriction to $\tilde{\mathbb{L}}_1^r$. We have

$$M_0 = F_{\dot{x}}, \qquad M_1 = \begin{bmatrix} F_{\dot{x}} & 0 \\ F_{\dot{x}t} + F_{\dot{x}x}\dot{x} + F_x + F_{\dot{x}\dot{x}}\ddot{x} + F_{\dot{x}z}g_t + F_{\dot{x}z}g_x\dot{x} + F_z g_x & F_{\dot{x}} \end{bmatrix}$$

$$\tilde{M}_0 = \begin{bmatrix} F_{\dot{x}} & 0 \\ -g_x & 1 \end{bmatrix}, \qquad \tilde{M}_1 = \begin{bmatrix} F_{\dot{x}} & 0 & 0 & 0 \\ -g_x & I_m & 0 & 0 \\ F_{\dot{x}t} + F_{\dot{x}x}\dot{x} + F_x + F_{\dot{x}\dot{x}}\ddot{x} + F_{\dot{x}z}\dot{z} & F_z & F_{\dot{x}} & 0 \\ -2g_{xt} - 2g_{xx}\dot{x} & 0 & -g_x & I_m \end{bmatrix}$$

and as in the proof of Lemma 2.5 we choose $Z_{2,1}$ in the form

$$Z_{2,1} = \begin{bmatrix} Z_{2,0} & Y_1 \\ 0 & Z_{s_1} \end{bmatrix}$$

and Z_{s_1} spans a subspace of range$(Z_{2,0})$. Moreover, from the proofs of Lemma 2.5 and Theorem 3.3, we know that $\tilde{Z}_{2,1}$ takes the form

$$\tilde{Z}_{2,1}^T = \begin{bmatrix} Z_{2,0}^T & 0 & 0 & 0 \\ \tilde{Y}_{1,1}^T & \tilde{Y}_{1,2}^T & \tilde{Z}_{s_1}^T & 0 \end{bmatrix}.$$

Comparing the products $Z_{2,1}^T M_1$ and $\tilde{Z}_{2,1}^T \tilde{M}_1$ given by

$$Z_{2,1}^T M_1 = \begin{bmatrix} Z_{2,0}^T F_{\dot{x}} & 0 \\ Y_1^T F_{\dot{x}} + Z_{s_1}^T \left(F_{\dot{x}t} + F_{\dot{x}x}\dot{x} + F_x + F_{\dot{x}\dot{x}}\ddot{x} + F_{\dot{x}z}g_t + F_{\dot{x}z}g_x\dot{x} + F_z g_x \right) & Z_{s_1}^T F_{\dot{x}} \end{bmatrix},$$

$$\tilde{Z}_{2,1}^T \tilde{M}_1 = \begin{bmatrix} Z_{2,0}^T F_{\dot{x}} & 0 & 0 & 0 \\ \tilde{Y}_{1,1}^T F_{\dot{x}} - \tilde{Y}_{1,2}^T g_x + \tilde{Z}_{s_1}^T \left(F_{\dot{x}t} + F_{\dot{x}x}\dot{x} + F_x + F_{\dot{x}\dot{x}}\ddot{x} + F_{\dot{x}z}\dot{z} \right) & \tilde{Y}_{1,2}^T + \tilde{Z}_{s_1}^T F_z & \tilde{Z}_{s_1}^T F_{\dot{x}} & 0 \end{bmatrix},$$

yields $\tilde{Y}_{1,2}^T = -\tilde{Z}_{s_1}^T F_z$. Since $\dot{z}(t) = g_x(t, x)\dot{x} + g_t(t, x)$ on $\tilde{\mathbb{L}}_1^r$, we see that the choices $\tilde{Y}_{1,1} = Y_1$ and $\tilde{Z}_{s_1} = Z_{s_1}$ satisfy the condition $\tilde{Z}_{2,1}^T \tilde{M}_1 = 0$. In particular, the columns of $\tilde{Z}_{2,1}$ span the left nullspace of $\tilde{M}_1$, because the columns of $Z_{2,1}$ span the left nullspace of M_1. Therefore, $\tilde{Z}_{2,1}$ is given by

$$\tilde{Z}_{2,1}^T = \begin{bmatrix} Z_{2,0}^T & 0 & 0 & 0 \\ Y_1^T & -Z_{s_1}^T F_z & Z_{s_1}^T & 0 \end{bmatrix},$$

and we have $\operatorname{rank}(\tilde{Z}_{2,1}) = \operatorname{rank}(Z_{2,1}) = a_1$, since Z_{s_1} has full row rank. Consider the matrices $A_{2,1}$, $T_{2,1}$ and $\tilde{A}_{2,1}$ of Hypothesis 2.1. Then we have

$$\tilde{A}_{2,1} \begin{bmatrix} T_{2,1} \\ g_x T_{2,1} \end{bmatrix} = 0.$$

Assume that (1.5) satisfies $\operatorname{rank}(\tilde{F}_{1;y,\dot{y},\ddot{y}}) = 2(n+m)$ on $\tilde{\mathbb{L}}_1^r$. Complete $\tilde{Z}_{2,1}$ to a nonsingular matrix $\tilde{Z}_0 = \begin{bmatrix} \tilde{Z}'_{2,1} & \tilde{Z}_{2,1} \end{bmatrix}$ and scale $\tilde{F}_{1;y,\dot{y},\ddot{y}}$ from the left with $\tilde{Z}_0^T$ to obtain

$$\begin{aligned} 2(n+m) &= \operatorname{rank}(\tilde{Z}_0^T \tilde{F}_{1;y,\dot{y},\ddot{y}}) = \operatorname{rank} \begin{bmatrix} (\tilde{Z}'_{2,1})^T \tilde{F}_{1;y} & (\tilde{Z}'_{2,1})^T \tilde{F}_{1;\dot{y},\ddot{y}} \\ \tilde{Z}_{2,1}^T \tilde{F}_{1;y} & 0 \end{bmatrix} \\ &= \operatorname{rank}(\tilde{Z}'_{2,1})^T \tilde{F}_{1;\dot{y},\ddot{y}} + \operatorname{rank} \tilde{Z}_{2,1}^T \tilde{F}_{1;y}, \end{aligned}$$

which implies that $\operatorname{rank}(\tilde{A}_{2,1}) = a$. Hence, $\dim \operatorname{kernel}(\tilde{A}_{2,1}) = d + m$, and the construction of $\tilde{T}_{2,1}$ proceeds analogously as in the proof of Theorem 3.3. Using $\tilde{Z}_{1,1}$ as before, it is easy to verify part three of Hypothesis 2.1. Summarizing the previous discussion, we have proved the following theorem.

Theorem 3.7 *Suppose that* (1.1) *is regular and satisfies Hypothesis 2.1 with characteristic values* $\mu = 1$, *a and d, and that* (1.5) *satisfies* $\operatorname{rank}(\tilde{F}_{1;y,\dot{y},\ddot{y}}) = 2(n+m)$ *on* $\tilde{\mathbb{L}}_1^r$. *Then the strangeness index of* (1.5) *is* $\tilde{\mu} = \mu = 1$ *on* $\tilde{\mathbb{L}}_1^r$ *with characteristic values* $\tilde{a} = a$ *and* $\tilde{d} = d + m$.

Example 3.8 Consider a DAE (1.1) and its reformulation (1.5) given by

$$F(t,x,\dot{x},g(t,x)) = \begin{bmatrix} e^{x_1}x_2\dot{x}_1 + e^{x_1}\dot{x}_2 - x_1 + 1 \\ e^{x_1}x_2 \end{bmatrix}, \qquad \tilde{F}(t,y,\dot{y}) = \begin{bmatrix} zx_2\dot{x}_1 + z\dot{x}_2 - x_1 + 1 \\ zx_2 \\ \dot{z} - e^{x_1}\dot{x}_1 \end{bmatrix}$$

with $g(t,x) = \exp(x_1)$. The system is regular, since the second equation implies $x_2 \equiv 0$, and therefore $x_1 \equiv 1$ using the first equation. It is easy to verify Hypothesis 2.1 for $\mu = 1$, with characteristic values $a = 2$, and $d = 0$ and matrices

$$Z_{2,1}^T = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}, \quad A_{2,1} = -\begin{bmatrix} e^{x_1}x_2 & e^{x_1} \\ 1 & 0 \end{bmatrix}, \quad T_{2,1} = \begin{bmatrix}\,\end{bmatrix}, \quad Z_{1,1} = \begin{bmatrix}\,\end{bmatrix}.$$

The strangeness-free form is given by

$$0 = \begin{bmatrix} Z_{1,1}^T F \\ Z_{2,1}^T F_1 \end{bmatrix} = \begin{bmatrix} e^{x_1}x_2 \\ 1 - x_1 \end{bmatrix}.$$

For the reformulated system $\tilde{F}$ we proceed as suggested. Since $Z_{s_1}^T F_z = x_2$ we have

$$\tilde{Z}_{2,1}^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & -x_2 & 0 & 1 & 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{A}_{2,1} = -\begin{bmatrix} 0 & z & x_2 \\ 1 + e^{x_1}\dot{x}_1 x_2 & \dot{z} - z\dot{x}_1 & -x_2\dot{x}_1 \end{bmatrix}.$$

On $\tilde{\mathbb{L}}_0^r$ we have $\operatorname{rank}(\tilde{A}_{2,1}) = 2$ and we obtain $\tilde{T}_{2,1}$, $\tilde{Z}_{2,1}$ and the strangeness-free representation as

$$\tilde{T}_{2,1} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \tilde{Z}_{1,1} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \text{and} \qquad 0 = \begin{bmatrix} \tilde{Z}_{1,1}^T \tilde{F} \\ \tilde{Z}_{2,1}^T \tilde{F}_1 \end{bmatrix} = \begin{bmatrix} \dot{z} - e^{x_1}\dot{x}_1 \\ zx_2 \\ 1 - x_1 \end{bmatrix}.$$

Note that we have used the precise reformulation (1.5) to generate $\tilde{F}$. If we replace the third equation by $\dot{z} - z\dot{x}_1 = 0$, we get the same characteristic values but cannot use the construction for $\tilde{Z}_{2,1}$ as in the previous discussion. Instead, $\tilde{Z}_{2,1}$ is then given by

$$\tilde{Z}_{2,1}^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

The remaining matrices can be chosen as before.

To prove the result for arbitrary strangeness index, we define the derivative array for the nonlinear function g of level ℓ by

$$G_\ell\left(t, y, \dot{y}, \ldots, y^{(\ell)}\right) = \begin{bmatrix} G(t,x,z) \\ \frac{d}{dt}G(t,x,z) \\ \vdots \\ \left(\frac{d}{dt}\right)^\ell G(t,x,z) \end{bmatrix},$$

using $G(t,x,z) = z - g(t,x)$ and define the restricted solution manifold $\tilde{\mathbb{L}}_\ell^r$ of level ℓ by

$$\tilde{\mathbb{L}}_\ell^r = \left\{ (t, y, \dot{y}, \ldots, \dot{y}^{(\ell+1)}) \in \mathbb{R}^{(\ell+2)(n+m)+1} \;\middle|\; \left(t, x, \ldots, x^{(\ell+1)}\right) \in \mathbb{L}_\ell,\ 0 = G_\ell\left(t, y, \ldots, y^{(\ell)}\right) \right\}.$$

Theorem 3.9 *Suppose that* (1.1) *is regular and satisfies Hypothesis 2.1 with characteristic values* μ, a *and* d*, and that* (1.5) *satisfies*

$$\operatorname{rank}\left(\tilde{F}_{k;y,\dot{y},\ldots,y^{(k+1)}}\right) = (k+1)(n+m) \qquad \text{on } \tilde{\mathbb{L}}_k^r \qquad \text{for } k = 0, \ldots, \mu. \tag{3.2}$$

If the constant rank conditions for M_k *and* N_k *hold for all* $k = 0, \ldots, \mu$*, then the strangeness index of* (1.5) *is* $\tilde{\mu} = \mu$ *on* $\tilde{\mathbb{L}}_\mu^r$ *with characteristic values* $\tilde{a} = a$ *and* $\tilde{d} = d + m$*.*

Proof. The proof follows by induction over μ and as in the previous proofs we consider all equalities only on the restricted solution manifolds. Suppose the claim is valid for $\mu - 1$. Then we have (as in the proof of Theorem 3.7) that

$$Z_{2,\mu} = \begin{bmatrix} Z_{2,\mu-1} & Y_\mu \\ 0 & Z_{p_\mu} \end{bmatrix} \qquad \text{and} \qquad \tilde{Z}_{2,\mu} = \begin{bmatrix} \tilde{Z}_{2,\mu-1} & \tilde{Y}_\mu \\ 0 & \tilde{Z}_{p_\mu} \\ 0 & 0 \end{bmatrix}.$$

By the same arguments as in the proof of Theorem 3.7 we have the relation $\tilde{Z}_{p_\mu} = Z_{p_\mu}$ and hence $\tilde{a} = a$. The constant rank assumptions (3.2) ensure the existence of a matrix $V = \begin{bmatrix} V_1^T & V_2^T \end{bmatrix}$ such that its columns span a subspace of kernel$(\tilde{A}_{2,\mu})$ and at the same time

$$g_x(t,x)V_1 - V_2$$

is nonsingular. The matrix $\tilde{T}_{2,\mu}$ of size $n + m \times d + m$ given by

$$\tilde{T}_{2,\mu} = \begin{bmatrix} T_2 & V_1 \\ g_x(t,x)T_2 & V_2 \end{bmatrix}$$

has full row rank and satisfies $\tilde{A}_{2,\mu}\tilde{T}_{2,\mu} = 0$. The argument for the third part of Hypothesis 2.1 follows along the lines of the proof of Theorem 3.3. □

4 Some observations on the numerical properties

In this section we illustrate some numerical properties of the polynomial reformulation approach. First and obvious, the state dimension is increased. If the system arises from the spatial discretization of a partial differential equation, nonlinear terms are typically present for each component, giving raise to numerous additional equations.

Example 4.1 The two dimensional sine-Gordon equation, see e. g. [3], used for example to model Josephson junctions, is given by

$$\frac{\partial^2 u}{\partial t^2} + \rho \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \zeta^2} - \phi(\xi, \zeta)\sin(u) \qquad \text{in } \Omega \times \mathbb{I}, \tag{4.1}$$

with $u = u(\xi, \zeta, t)$, $\rho \geq 0$ is the *dissipative* term, ϕ is the Josephson current density and $\Omega = \{(\xi, \zeta) \in \mathbb{R}^2 \mid \|(\xi, \zeta)\|_\infty \leq 1\}$ is a square domain. Applying the method of lines with finite differences to (4.1) yields an ODE of the form

$$D^2 x(t) + \rho D x(t) = Ax(t) - G(x(t)), \tag{4.2}$$

where $x \in \mathbb{R}^{(N+1)^2}$ is the spatial approximation of u, D is a differential operator, A the discretized Laplacian and the nonlinear term G contains $\sin(x_i)$ for $i = 1, \ldots, (N+1)^2$. As already discussed in Example 2.4, we require two additional state variables to rewrite a sine function in polynomial form (due to the cosine arising in the derivative), hence a polynomial reformulation of (4.2) increases the state space dimension by $2(N+1)^2$.

The second immediate observation is that the polynomial reformulation may require symbolic manipulation of the system equations, which may not be available.

To analyze the impact of the polynomial reformulation on the numerical integration we consider initial value problems of the form

$$0 = F(t, x, \dot{x}, z, \dot{z}), \qquad x(t_0) = x_0,\ z(t_0) = z_0 = g(t_0, x_0) \tag{4.3}$$

in the interval $\mathbb{I} = [t_0, t_f] \subseteq \mathbb{R}$. We introduce the set of gridpoints $\{t_i\}_{i=0}^N$ satisfying $t_0 < t_1 < \ldots < t_N = t_f$ with stepsizes $h_i := t_{i+1} - t_i$ for $i = 0, \ldots, N-1$, and denote by $\mathfrak{X}_i, \mathfrak{Z}_i$ approximations to the solution $x(t_i), z(t_i)$, respectively. Let us consider the implicit Euler method applied to the strangeness-free problem, which is given by the implicit iteration

$$0 = F\left(t_{i+1}, \mathfrak{X}_{i+1}, \frac{\mathfrak{X}_{i+1} - \mathfrak{X}_i}{h_i}, \mathfrak{Z}_{i+1}, \frac{\mathfrak{Z}_{i+1} - \mathfrak{Z}_i}{h_i}\right),$$

which in each step can be solved using the Newton method.

Example 4.2 Consider the initial value problem

$$\begin{aligned} \dot{x}_1 &= \omega \sin(\omega x_2), & x_1(t_0) &= x_{1,0}, \\ \dot{x}_2 &= 1, & x_2(t_0) &= x_{2,0}, \end{aligned} \tag{4.4}$$

with solution $x_2(t) = t + x_{2,0}$ and $x_1(t) = -\cos(\omega(t + x_{2,0})) + x_{1,0} + \cos(\omega x_{2,0})$. A polynomial reformulation is given by

$$\begin{aligned} \dot{x}_1 &= \omega z_1, & x_1(t_0) &= x_{1,0}, \\ \dot{x}_2 &= 1, & x_2(t_0) &= x_{2,0}, \\ \dot{z}_1 &= \omega z_2, & z_1(t_0) &= \sin(\omega x_{2,0}), \\ \dot{z}_2 &= -\omega z_1, & z_2(t_0) &= \cos(\omega x_{2,0}). \end{aligned} \tag{4.5}$$

Clearly, both systems are ODEs and hence strangeness-free. The numerical solution and absolute error of the x_1 component of (4.4) and (4.5) for constant stepsize $h = 0.02$ is depicted in Figure 4.1. The numerical approximation of (4.4) is as expected, with a maximal error $\|x(t_i) - \mathfrak{X}_i\| < 0.06$. In contrast to this, the numerical solution of the polynomial reformulation is given by

$$\begin{bmatrix} \mathfrak{X}_{i+1,1} \\ \mathfrak{X}_{i+1,2} \\ \mathfrak{Z}_{i+1,1} \\ \mathfrak{Z}_{i+1,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \frac{h_i\omega}{1+h_i^2\omega^2} & \frac{h_i^2\omega^2}{1+h_i^2\omega^2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1+h_i^2\omega^2} & \frac{h_i\omega}{1+h_i^2\omega^2} \\ 0 & 0 & -\frac{h_i\omega}{1+h_i^2\omega^2} & \frac{1}{1+h_i^2\omega^2} \end{bmatrix} \begin{bmatrix} \mathfrak{X}_{i,1} \\ h_i + \mathfrak{X}_{i,2} \\ \mathfrak{Z}_{i,1} \\ \mathfrak{Z}_{i,2}. \end{bmatrix}$$

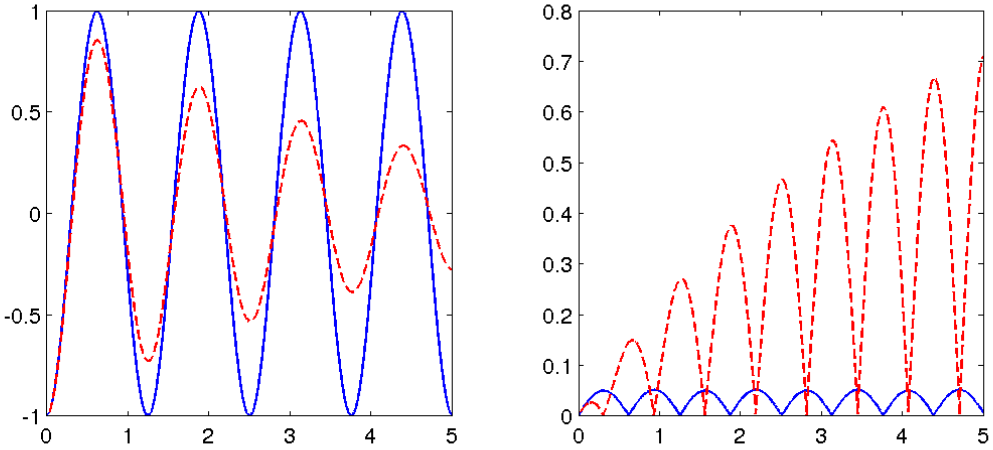

Figure 4.1: Numerical solution (left) and absolute error (right) of (4.4) (blue, solid) and (4.5) (red, dashed) for $\mathbb{I} = [0, 5]$, $\omega = 5$, $N = 251$ and constant stepsize $h = 0.02$.

In particular, we have

$$\mathfrak{Z}_{i+1,1}^2 + \mathfrak{Z}_{i+1,2}^2 = \frac{1}{1 + h_i^2\omega^2} \left(\mathfrak{Z}_{i,1}^2 + \mathfrak{Z}_{i,2}^2\right) < \mathfrak{Z}_{i,1}^2 + \mathfrak{Z}_{i,2}^2 \tag{4.6}$$

implying that the numerical approximation $\mathfrak{X}_{i+1,1}$ decays to zero (in agreement with Figure 4.1).

An immediate idea is to use the error indicator $\|g(t_i, \mathfrak{X}_i) - \mathfrak{Z}_i\|$ to implement a stepsize control. In Example 4.2 this will lead to very small stepsizes h_i satisfying $(1 + h_i^2\omega^2)^{-1} = 1$ in machine precision (compare with (4.6)). Employing the DAE context, one could also enforce the invariant by adding the constraint

$$1 = z_1^2 + z_2^2 \tag{4.7}$$

to (4.5). The system (4.5), (4.7) then becomes overdetermined and one could solve it as an overdetermined system or add a Lagrange multiplier λ for the extra constraint, yielding the system

$$0 = \begin{bmatrix} \dot{x}_1 - \omega z_1 \\ \dot{x}_2 - 1 \\ \dot{z}_1 - \omega z_2 + \lambda(z_1^2 + z_2^2 - 1) \\ \dot{z}_2 + \omega z_1 + \lambda(z_1^2 + z_2^2 - 1) \\ z_1^2 + z_2^2 - 1 \end{bmatrix},$$

which then requires another index reduction.

Remark 4.3 The polynomial reformulation procedure for system (4.4) adds the eigenvalues $\pm i\omega$ to the system. If the time integrator is not stable for purely imaginary eigenvalues, as for example the explicit Euler method, the integration of the reformulated system is likely to fail. In this case the use of structure preserving or geometric integrators [12] will allow to preserve structural properties of the system lost in the reformulation process.

A structure preserving integrator may require additional computational cost and is not necessarily required for the complete system. Rewriting the reformulated system (1.5) as

$$0 = \tilde{F}(t, y, \dot{y}) = \begin{bmatrix} F(t, x, \dot{x}, z) \\ G(t, x, \dot{x}, z, \dot{z}) \end{bmatrix} \tag{4.8}$$

with $G(t, x, \dot{x}, z, \dot{z}) = \dot{z}(t) - g_x(x, t)\dot{x} - g_t(t, x)$, we can use different integrators for F and G and solve for $(\mathfrak{X}_{i+1}, \mathfrak{Z}_{i+1})$ either simultaneously, or iteratively with a dynamic iteration scheme [20], which is well understood for coupled ODEs.

Example 4.4 (Example 4.2 continued) We use (4.8) and solve F with the implicit Euler and for G, we employ the two stage Lobatto IIIA method also known as (implicit) trapezoidal rule. This scheme preserves the invariant (4.7). Using the same settings as in Example 4.2 this approach yields an approximation with absolute error similar to the approximation of (4.4) with the implicit Euler method, as is illustrated in Figure 4.2.

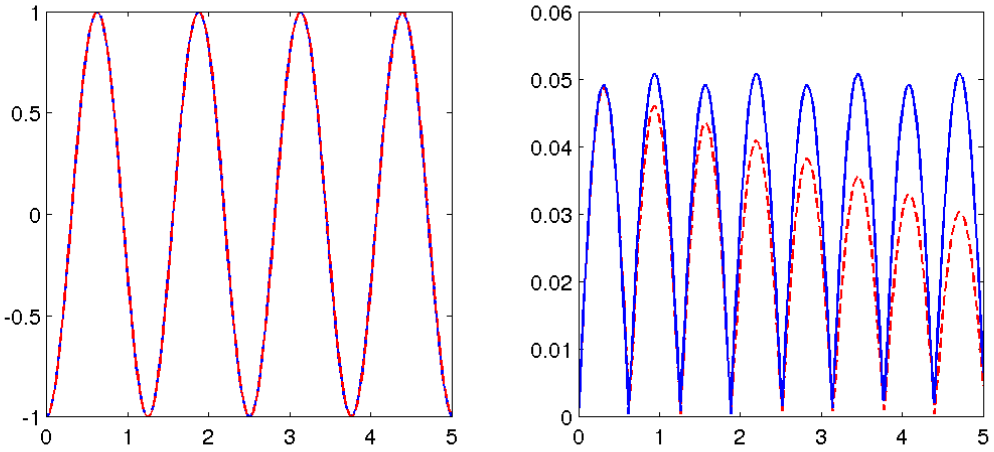

Figure 4.2: Numerical solution (left) and absolute error (right) of (4.4), with the implicit Euler (blue, solid), and (4.5), with the combined Euler, LobattoIIIA approach (red, dashed).

5 Conclusion

The polynomial representation procedure introduced in [9] has been extended to a class of nonlinear differential-algebraic systems with well-defined strangeness index μ. It has been shown that the strangeness index is preserved during the polynomial representation. In particular, if the original system can be rewritten as an ODE, then the reformulated system is also an ODE. To avoid ill-conditioning, one should in general first derive the strangeness-free reformulation and then apply the polynomial representation procedure. The performed numerical study has shown that the time integration might require a different ODE solver to preserve additional invariants introduced by the polynomial reformulation.

References

[1] A. Bellen and M. Zennaro. *Numerical methods for delay differential equations*. Clarendon Press, 2003.

[2] P. Benner and T. Breiten. Interpolation-base $\mathcal{H}_2$-Model Reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.*, 33(3):859–885, 2012.

[3] A. G. Bratsos. The solution of the two-dimensional sine-gordon equation using the method of lines. *J. Comput. Appl. Math.*, 206:251–277, 2007.

[4] T. Breiten and T. Damm. Krylov subspace methods for model order reduction of bilinear control systems. *Syst. Control Lett.*, 59(8):443–450, 2010.

[5] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18:1101–1115, 1987.

[6] S. L. Campbell and C. D. Meyer. *Generalized Inverses of Linear Transformations*. Pitman, San Francisco, CA, 1979.

[7] T. Carleman. Application de la thorie des quations intgrales linaires aux systems d'quations diffrentielles non linaires. *Acta Math.*, 59:63–87, 1932.

[8] L. Dieci and T. Eirola. On smooth decompositions of matrices. *SIAM J. Matrix Anal. Appl.*, 20(3):800–819, 1999.

[9] C. Gu. QLMOR: a projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *Comput. Des. Integr. Circuits Syst.*, 30(9):1307–1320, 2011.

[10] N. Guglielmi and E. Hairer. Computing breaking points in implicit delay differential equations. *Adv. Comput. Math.*, 29(3):229–247, 2008.

[11] P. Ha, V. Mehrmann, and A. Steinbrecher. Analysis of Linear Variable Coefficient Delay Differential-Algebraic Equations. *J. Dyn. Differ. Equations*, 26(3):1–26, 2014.

[12] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics. Springer-Verlag, 2006.

[13] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, Germany, 2nd edition, 1996.

[14] P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *J. Comput. Appl. Math.*, 56:225–259, 1994.

[15] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations. Analysis and Numerical Solution*. European Mathematical Society, 2006.

[16] P. Kunkel, V. Mehrmann, and I. Seufer. GENDA: A software package for the numerical solution of general nonlinear differential-algebraic equations. Preprint 730–2002, Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, FRG, 2002.

[17] V. Mehrmann. Index concepts for differential-algebraic equations. Preprint 03–2012, Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, FRG, 2012. To appear in Encycl. Appl. Math.

[18] V. Mehrmann and C. Shi. Transformation of high order linear differential-algebraic systems to first order. *Numer. Algorithms*, 42:281–307, 2006.

[19] V. Mehrmann and L. Wunderlich. Hybrid systems of differential-algebraic equations - Analysis and numerical solution. *J. Process Control*, 19(8):1218–1228, 2009.

[20] U. Miekkala and O. Nevanlinna. Convergence of dynamic iteration methods for initial value problems. *SIAM J. Sci. Statist. Comput.*, 8(4):459–482, 1987.

[21] M. P. Quéré and G. Villard. An algorithm for the reduction of linear DAE. In A. H. M. Levelt, editor, *Proceedings of the 1995 International Symposium on Symbolic and Algebraic Computation (ISSAC'95)*, pages 223–231. ACM Press, New York, NY, 1995.

[22] W. J. Rugh. *Nonlinear System Theory, The Volterra/Wiener Approach.* The Johns Hopkins University Press, Baltimore and London, 1981.