Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Completions of nonlinear DAE flows based on index reduction techniques and their stabilization

Stephen L. Campbell[a,*], Peter Kunkel[b]

[a] Department of Mathematics, Box 8205, North Carolina State University, Raleigh, NC 27695-8205, USA
[b] Mathematisches Institut, Universität Leipzig, Johannisgasse 26, 04009 Leipzig, Germany

ARTICLE INFO

Article history:
Received 23 March 2009
Received in revised form 26 June 2009

MSC:
65L80
34A09
34A45

Keywords:
Differential algebraic equations
Numerical methods
Stability

ABSTRACT

Differential algebraic equations (DAEs) define a differential equation on a manifold. A number of ways have been developed to numerically solve some classes of DAEs. Motivated by problems in control theory, numerical simulation, and the use of general purpose modeling environments, recent research has considered the embedding of the DAE solutions of a general DAE into the solutions of an ODE where the added dynamics have special properties. This paper both provides new results on the linear time-varying case and considers the important nonlinear case.

1. Introduction

Nonlinear differential algebraic equations (DAEs)

$$F(t,x,\dot{x})=0 \tag{1}$$

arise in a number of application areas [1,2]. Here $F\in C^0(\mathbb{I}\times\mathbb{D}_x\times\mathbb{D}_{\dot{x}},\mathbb{R}^n)$ is sufficiently smooth, $\mathbb{I}=[\underline{t},\overline{t}]\subseteq\mathbb{R}$ is a (compact) interval, and $\mathbb{D}_x,\mathbb{D}_{\dot{x}}\subseteq\mathbb{R}^n$, are open sets. A number of numerical approaches have been developed for (1). Classical methods such as backward differentiation or implicit Runge–Kutta methods require the system to be low index and have special structure. Several approaches have been developed based on what is called the derivative array [3,2,4].

One possible approach, sometimes called index reduction, is to transform (1) into an index one DAE which has the same solutions as (1) but allows for the application of certain classes of integration methods known from the treatment of ordinary differential equations (ODEs). However, we cannot use every numerical method. In particular, it is not possible to apply explicit methods. This is due to the fact that the new DAE contains all algebraic constraints posed by (1) and the numerical method must be able to resolve these constraints.

An ODE whose solutions include the solutions of a DAE is called a completion of that DAE. There are several reasons why having a completion would be beneficial. For one it would allow the use of explicit integrators. It would also permit easy interfacing with many modeling and design software packages which require ODE models. That is, one derives an ODE for the unknown x from (1) and uses this ODE. But every such ODE will have a larger solution space than (1). In particular, the derivation introduces additional dynamical behavior which can be interpreted as a completion of the flow of (1) which is only defined on the set of all (t,x) which satisfy all constraints contained in (1). This additional dynamics is artificial and depends on the way the corresponding ODE is derived from (1). Examples show that already in the case of linear DAEs

[*] Corresponding author.
E-mail address: slc@math.ncsu.edu (S.L. Campbell).

0377-0427/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2009.08.111

with constant coefficients the additional flow can be quite arbitrary [5] and sometimes cause difficulty with the numerical solvers. See also [6] where the so-called least-squares completion is considered.

This leads to the important problem of developing numerical procedures for generating completions of general nonlinear DAEs for which the additional dynamics of the completion are known and have the desired behavior. In the case of linear DAEs with constant coefficients it has been shown in [6] that the least-squares completion yields additional dynamics with all eigenvalues being zero implying that the additional dynamics is at most polynomial. In [6] it has also been pointed out that for this special case one can define a completion in such a way that the additional dynamics is constant. This is achieved by basing the construction of the completion on techniques also used for the index reduction mentioned above. In [7], it has been shown how the additional dynamics of the least-squares completion in the case of linear DAEs with constant coefficients can be stabilized. Finally, representations of least-squares completions in the case of linear DAEs with time-varying coefficients were obtained in [8]. The aim of the present paper is to generalize these results to nonlinear DAEs. Using some earlier results on stabilization of invariants, we suggest a technique useful for both linear time-varying and nonlinear DAEs that allows for the modification of a computed completion in such a way that every solution will approach the set of all points satisfying the DAE constraints. The results presented also provide a corrected reformulation of some claims of [9].

Stabilization of explicit constraints has been studied on several occasions [9,10]. We are concerned here with general DAEs where some or all of the constraints can be implicit. Stabilization is done in this paper without analytically determining any constraints.

Section 2 discusses the linear time-varying case, presents needed notation, and gives formulas for the computed completions. Section 3 sets up the nonlinear notation and gives results on the nonlinear completions. The key Section 4 first discusses stabilization of invariants for ODEs, and then applies these ideas to stabilizing the completions of general DAEs. The results of this section are new contributions to both the linear time-varying and nonlinear cases. Numerical examples are discussed in Section 5.

2. Linear DAEs with variable coefficients

Consider a linear DAE with variable coefficients of the form

$$E(t)\dot{x} = A(t)x + f(t), \tag{2}$$

where $\mathbb{I}$ is an interval and $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ and $f \in C(\mathbb{I}, \mathbb{R}^n)$ are assumed to be sufficiently smooth, i.e., sufficiently often continuously differentiable. It is well known that (2) defines a well-posed problem if and only if it possesses a well-defined differentiation index. In order to characterize this property, we need the so-called derivative array equation

$$M_\ell(t)\dot{z}_\ell = N_\ell(t)x + g_\ell(t) \tag{3}$$

obtained by differentiating (2) ℓ times. Here

$$\begin{aligned}
&(M_\ell)_{i,j} = \binom{i}{j} E^{(i-j)} - \binom{i}{j+1} A^{(i-j-1)}, && i, j = 0, \ldots, \ell,\\
&(N_\ell)_i = A^{(i)}, && i = 0, \ldots, \ell,\\
&(g_\ell)_i = f^{(i)}, && i = 0, \ldots, \ell,\\
&(z_\ell)_j = x^{(j)}, && j = 0, \ldots, \ell,
\end{aligned} \tag{4}$$

with the convention that $\binom{i}{j} = 0$ for $i < 0, j < 0$ or $j > i$. We then require the following hypothesis, see [2, Chapter 3].

Hypothesis 1. There exist integers μ, a, and d such that the derivative arrays M_μ, N_μ associated with the given pair of matrix functions (E, A) has the following properties:

1. For all $t \in \mathbb{I}$ we have $\operatorname{rank}(M_\mu(t)) = (\mu+1)n - a$. This implies that there exists a smooth matrix function Z_2 of size $((\mu+1)n, a)$ and pointwise orthonormalized columns satisfying $Z_2^T M_\mu = 0$.
2. For all $t \in \mathbb{I}$ we have $\operatorname{rank}(\hat{A}_2(t)) = a$, where $\hat{A}_2 = Z_2^T N_\mu$. This implies that there exists a smooth matrix function T_2 of size (n, d), $d = n - a$, and pointwise orthonormalized columns satisfying $\hat{A}_2 T_2 = 0$.
3. For all $t \in \mathbb{I}$ we have $\operatorname{rank}(E(t)T_2(t)) = d$. This implies that there exists a smooth matrix function Z_1 of size (n, d) and pointwise orthonormalized columns satisfying $\operatorname{rank} \hat{E}_1 T_2 = d$ with $\hat{E}_1 = Z_1^T E$.

If (2) satisfies Hypothesis 1 for some μ, then it satisfies Hypothesis 1 for every larger μ with the same values a and d. If (2) satisfies Hypothesis 1 and μ is chosen minimally, then (2) has a well-defined differentiation index $\nu = \mu + 1$ for $a \neq 0$ and $\nu = \mu = 0$ for $a = 0$. Conversely, if (2) has a well-defined differentiation index ν, then it satisfies Hypothesis 1 with a minimal $\mu = \max\{\nu - 1, 0\}$. For details see again [2, Chapter 3].

We have used the formulation in Hypothesis 1 since the matrices Z_2, T_2, Z_1 are used in the numerical procedures. There are simpler alternative assumptions directly in terms of M_ℓ, N_ℓ that can be checked to see that Hypothesis 1 holds [11].

All of the above properties are invariant under (global) equivalence transformations defined by

$$\tilde{E} = PEQ, \qquad \tilde{A} = PAQ - PE\dot{Q}, \qquad \tilde{x} = Q^{-1}x, \qquad \tilde{f} = Pf, \tag{5}$$

where $P \in C(\mathbb{I}, \mathbb{R}^{n,n})$ and $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ are pointwise nonsingular. In particular, the derivative arrays M_ℓ, N_ℓ belonging to E, A and $\tilde{M}_\ell, \tilde{N}_\ell$ belonging to $\tilde{E}, \tilde{A}$ transform according to

$$\tilde{M}_\ell = \Pi_\ell M_\ell \Theta_\ell, \qquad \tilde{N}_\ell = \Pi_\ell M_\ell Q - \Pi_\ell M_\ell \Psi_\ell, \qquad \tilde{g}_\ell = \Pi_\ell g_\ell \tag{6}$$

with

$$\begin{aligned}
(\Pi_\ell)_{i,j} &= \binom{i}{j} P^{(i-j)}, & i,j &= 0,\dots,\ell,\\
(\Theta_\ell)_{i,j} &= \binom{i+1}{j+1} Q^{(i-j)}, & i,j &= 0,\dots,\ell,\\
(\Psi_\ell)_i &= Q^{(i+1)}, & i &= 0,\dots,\ell.
\end{aligned} \tag{7}$$

The matrix functions Z_1 and Z_2 of Hypothesis 1 define a so-called reduced DAE

$$\begin{bmatrix} \hat{E}_1(t) \\ 0 \end{bmatrix} \dot{x} = \begin{bmatrix} \hat{A}_1(t) \\ \hat{A}_2(t) \end{bmatrix} x + \begin{bmatrix} \hat{f}_1(t) \\ \hat{f}_2(t) \end{bmatrix} \tag{8}$$

with

$$\begin{aligned}
\hat{E}_1 = Z_1^{\mathrm{T}} E, \quad \hat{A}_1 &= Z_1^{\mathrm{T}} A, & \hat{f}_1 &= Z_1^{\mathrm{T}} f,\\
\hat{A}_2 &= Z_2^{\mathrm{T}} N_\mu, & \hat{f}_2 &= Z_2^{\mathrm{T}} g_\mu,
\end{aligned}$$

which is known to possess the same solutions as the original DAE (2). Moreover, (8) satisfies Hypothesis 1 with $\mu = 0$ and the same values a and d as the original DAE. In particular, the matrix functions $Z_1^{\mathrm{T}} E$ and $Z_2^{\mathrm{T}} N_\mu$ form a pointwise nonsingular matrix function. In view of (8) a natural completion of the given DAE (2) would therefore be the ODE

$$\begin{bmatrix} \hat{E}_1(t) \\ -\hat{A}_2(t) \end{bmatrix} \dot{x} = \begin{bmatrix} \hat{A}_1(t) \\ \dot{\hat{A}}_2(t) \end{bmatrix} x + \begin{bmatrix} \hat{f}_1(t) \\ \dot{\hat{f}}_2(t) \end{bmatrix}. \tag{9}$$

In the context of completions the case $\nu = 0$, where the DAE is actually an ODE, is of no interest. In what follows, we therefore only consider the case $\nu \geq 1$ implying that $\nu = \mu + 1$.

2.1. Completions and the derivative arrays

The completion (9) is derived by differentiating a part of the computed (8). The first question concerning the completion (9) is whether it can be derived directly from the derivative array equations (3). To simplify the following discussion, we sometimes work with formally infinite derivative arrays M, N, and g, defined according to (4) by dropping the limit ℓ. Similarly, we use formally infinite transformations Π, Θ, and Ψ according to (7). Note that these formally infinite matrix functions introduce no difficulties since we will actually consider only finite parts of them.

Lemma 2. *Let the (infinite) shift matrix S and the projection V be given by*

$$S = \begin{bmatrix} 0 & & & \\ I_n & 0 & & \\ & I_n & 0 & \\ & & \ddots & \ddots \end{bmatrix}, \qquad V = \begin{bmatrix} I_n \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

and let I denote the (infinite) identity matrix. Then we have the relations

$$\begin{aligned}
&\text{(a)} \quad S^{\mathrm{T}} S = I, \qquad S S^{\mathrm{T}} + V V^{\mathrm{T}} = I,\\
&\text{(b)} \quad S^{\mathrm{T}} M = M S^{\mathrm{T}} + \dot{M} - N V^{\mathrm{T}},\\
&\text{(c)} \quad S^{\mathrm{T}} \Pi = \Pi S^{\mathrm{T}} + \dot{\Pi}.
\end{aligned} \tag{10}$$

Proof. The relations in (10)(a) are trivial. Observing that $S_{i,j} = \delta_{i,j+1} I_n$ with the Kronecker symbol $\delta_{i,j}$ we find that

$$\begin{aligned}
(MS^{\mathrm{T}} + \dot{M} - NV^{\mathrm{T}})_{i,j} &= \sum_{k\geq 0} M_{i,k} S_{j,k} + \dot{M}_{i,j} - N_i \delta_{j,0}\\
&= \sum_{k\geq 0} \left[\binom{i}{k} E^{(i-k)} - \binom{i}{k+1} A^{(i-k-1)} \right] \delta_{j,k+1} + \left[\binom{i}{j} E^{(i-j+1)} - \binom{i}{j+1} A^{(i-j)} \right] - A^{(i)} \delta_{j,0}\\
&= \sum_{k\geq 0} \left[\binom{i}{k} E^{(i-k+1)} - \binom{i}{k+1} A^{(i-k)} \right] \delta_{j,k} + \left[\binom{i}{j} E^{(i-j+1)} - \binom{i}{j+1} A^{(i-j)} \right]\\
&= \left[\binom{i}{j-1} E^{(i-j+1)} - \binom{i}{j} A^{(i-j)} \right] + \left[\binom{i}{j} E^{(i-j+1)} - \binom{i}{j+1} A^{(i-j)} \right],
\end{aligned}$$

which coincides with

$$\begin{aligned}(S^{\mathrm{T}}M)_{i,j} &= \sum_{k\geq 0} S_{k,i}M_{k,j}\\ &= \sum_{k\geq 0}\delta_{k,i+1}\left[\binom{k}{j}E^{(k-j)}-\binom{k}{j+1}A^{(k-j-1)}\right]\\ &= \left[\binom{i+1}{j}E^{(i-j+1)}-\binom{i+1}{j+1}A^{(i-j)}\right].\end{aligned}$$

Similarly, we find that

$$\begin{aligned}(\Pi S^{\mathrm{T}}+\dot{\Pi})_{i,j} &= \sum_{k\geq 0}\Pi_{i,k}\delta_{j,k+1}+\dot{\Pi}_{i,j}\\ &= \sum_{k\geq 0}\binom{i}{k}P^{(i-k)}\delta_{j,k+1}+\binom{i}{j}P^{(i-j+1)}=\binom{i}{j-1}P^{(i-j+1)}+\binom{i}{j}P^{(i-j+1)},\end{aligned}$$

which coincides with

$$(S^{\mathrm{T}}\Pi)_{i,j}=\sum_{k\geq 0}S_{k,i}\Pi_{k,j}\delta_{k,i+1}\binom{k}{j}P^{(k-j)}=\binom{i+1}{j}P^{(i-j+1)}. \quad \square$$

Extending Z_1 and Z_2 to infinite functions by adding zero blocks and using the same notation for these infinite functions, we can write

$$\begin{aligned}\hat{E}_1 = Z_1^{\mathrm{T}}MV,\quad \hat{A}_1 &= Z_1^{\mathrm{T}}N,\quad \hat{f}_1 = Z_1^{\mathrm{T}}g,\\ \hat{A}_2 &= Z_2^{\mathrm{T}}N,\quad \hat{f}_2 = Z_2^{\mathrm{T}}g\end{aligned} \tag{11}$$

for the coefficients in (8). The property $Z_2^{\mathrm{T}}M=0$ together with (10) then implies

$$\begin{aligned}(\dot{Z}_2^{\mathrm{T}}+Z_2^{\mathrm{T}}S^{\mathrm{T}})M &= \dot{Z}_2^{\mathrm{T}}M+Z_2^{\mathrm{T}}S^{\mathrm{T}}M=\dot{Z}_2^{\mathrm{T}}M+Z_2^{\mathrm{T}}(MS^{\mathrm{T}}+\dot{M}-NV^{\mathrm{T}})\\ &= \dot{Z}_2^{\mathrm{T}}M+Z_2^{\mathrm{T}}\dot{M}-Z_2^{\mathrm{T}}NV^{\mathrm{T}}\\ &= \frac{\mathrm{d}}{\mathrm{d}t}(Z_2^{\mathrm{T}}M)-Z_2^{\mathrm{T}}NV^{\mathrm{T}}=-Z_2^{\mathrm{T}}NV^{\mathrm{T}}.\end{aligned}$$

Setting $Z_3=\dot{Z}_2+SZ_2$, we first observe that Z_3 possibly has a nonvanishing νth block due to the involved shift. We therefore actually work with the finite part M_ν of M. Setting simply $M=M_\nu$ in the following with Z_1 and Z_2 of Hypothesis 1 completed with a zero block, we define $Z=[\,Z_1\quad Z_2\quad Z_3\quad Z_4\,]$ in such a way that Z becomes a square matrix function. Consider now the (finite) matrix function

$$Z^{\mathrm{T}}MV=\begin{bmatrix}Z_1^{\mathrm{T}}MV\\ 0\\ -Z_2^{\mathrm{T}}N\\ Z_4^{\mathrm{T}}MV\end{bmatrix}.$$

Since $Z_1^{\mathrm{T}}MV$ and $Z_2^{\mathrm{T}}N$ constitute a pointwise nonsingular matrix function, the first part of Z consisting of Z_1, Z_2, and Z_3 has pointwise full column rank. Hence, we can choose a smooth Z_4 in such a way that Z is pointwise nonsingular.

Solving now the transformed derivative array equation

$$Z^{\mathrm{T}}M\dot{z}=Z^{\mathrm{T}}(Nx+g)$$

with $z=z_\nu$ by means of the Moore–Penrose pseudoinverse [12] of $Z^{\mathrm{T}}M$ yields

$$\dot{z}=(Z^{\mathrm{T}}M)^{+}Z^{\mathrm{T}}(Nx+g).$$

Because of $\operatorname{corank}(M)=a$ corresponding to $Z_2^{\mathrm{T}}M=0$, the above least-squares solution is given by the least-squares solution of

$$\begin{bmatrix}Z_1^{\mathrm{T}}M\\ -Z_2^{\mathrm{T}}NV^{\mathrm{T}}\\ Z_4^{\mathrm{T}}M\end{bmatrix}\dot{z}=\begin{bmatrix}Z_1^{\mathrm{T}}\\ \dot{Z}_2^{\mathrm{T}}+Z_2^{\mathrm{T}}S^{\mathrm{T}}\\ Z_4^{\mathrm{T}}\end{bmatrix}(Nx+g). \tag{12}$$

Since

$$\begin{bmatrix}Z_1^{\mathrm{T}}M\\ -Z_2^{\mathrm{T}}NV^{\mathrm{T}}\end{bmatrix}=\begin{bmatrix}Z_1^{\mathrm{T}}MV\\ -Z_2^{\mathrm{T}}N\end{bmatrix}V^{\mathrm{T}},$$

where the matrix function multiplying V^{T} is pointwise nonsingular, every solution of (12) satisfies

$$\begin{bmatrix} Z_1^{\mathrm{T}} M V \\ -Z_2^{\mathrm{T}} N \end{bmatrix} \dot{x} = \begin{bmatrix} Z_1^{\mathrm{T}} \\ \dot{Z}_2^{\mathrm{T}} + Z_2^{\mathrm{T}} S^{\mathrm{T}} \end{bmatrix} (Nx + g). \tag{13}$$

But this is just the desired completion (9). Hence, (9) can be interpreted as the completion

$$\dot{x} = V^{\mathrm{T}} M^{-} (Nx + g) \tag{14}$$

that is given by the generalized inverse M^{-} of M defined by

$$M^{-} = (Z^{\mathrm{T}} M)^{+} Z^{\mathrm{T}}. \tag{15}$$

Lemma 3. *The (pointwise) generalized inverse M^{-} of M defined by (15) with a (pointwise) nonsingular matrix function Z is (pointwise) an inner and outer inverse of M, that is,*

$$MM^{-}M = M, \qquad M^{-}MM^{-} = M^{-}, \tag{16}$$

and $Z^{\mathrm{T}}(MM^{-})Z^{-\mathrm{T}}$ as well as $M^{-}M$ are (pointwise) symmetric.

Proof. In the following we make use of the properties of $(Z^{\mathrm{T}}M)^{+}$. First, we have $M^{-}M = (Z^{\mathrm{T}}M)^{+}(Z^{\mathrm{T}}M)$, which is symmetric. Second, we have $MM^{-} = M(Z^{\mathrm{T}}M)^{+}Z^{\mathrm{T}}$ or $Z^{\mathrm{T}}(MM^{-})Z^{-\mathrm{T}} = (Z^{\mathrm{T}}M)(Z^{\mathrm{T}}M)^{+}$, which is symmetric. Third, we have $Z^{\mathrm{T}}MM^{-}M = (Z^{\mathrm{T}}M)(Z^{\mathrm{T}}M)^{+}Z^{\mathrm{T}}M = Z^{\mathrm{T}}M$ implying $MM^{-}M = M$. Fourth, we have $M^{-}MM^{-} = (Z^{\mathrm{T}}M)^{+}(Z^{\mathrm{T}}M)(Z^{\mathrm{T}}M)^{+}Z^{\mathrm{T}} = (Z^{\mathrm{T}}M)^{+}Z^{\mathrm{T}} = M^{-}$. □

2.2. Completions and equivalence transformations

In this section, we investigate how the completion behaves under (global) equivalence transformations (5). In particular, what is the relationship between the completions of equivalent systems. As already mentioned, the transformed DAE $\tilde{E}(t)\dot{\tilde{x}} = \tilde{A}(t)x + \tilde{f}(t)$ satisfies Hypothesis 1 if the original DAE does, with the same values μ, a, and d. The corresponding matrix functions $\tilde{Z}_2$, $\tilde{T}_2$, and $\tilde{Z}_1$ are related according to

$$\begin{aligned} &\text{(a)} \quad \tilde{Z}_2 = \Pi^{-\mathrm{T}} Z_2 U_{Z_2}, \\ &\text{(b)} \quad \tilde{T}_2 = Q^{-1} T_2 U_{T_2}, \\ &\text{(c)} \quad \tilde{Z}_1 = \Pi^{-\mathrm{T}} Z_1 U_{Z_1}, \end{aligned} \tag{17}$$

where $U_{Z_2}, U_{T_2}, U_{Z_1}$ are smooth, pointwise nonsingular matrix functions of appropriate size describing the necessary re-orthonormalization and the specific choice of a smooth orthonormal basis. Again, it does not matter whether Z_1, Z_2 are related to M_μ, M_ν, or the infinite matrix function M. We therefore also omit a subscript of Π. The completion of the transformed problem is then given by

$$\begin{bmatrix} \tilde{Z}_1^{\mathrm{T}} \tilde{M} V \\ -\tilde{Z}_2^{\mathrm{T}} \tilde{N} \end{bmatrix} \dot{\tilde{x}} = \begin{bmatrix} \tilde{Z}_1^{\mathrm{T}} \\ \dot{\tilde{Z}}_2^{\mathrm{T}} + \tilde{Z}_2^{\mathrm{T}} S^{\mathrm{T}} \end{bmatrix} (\tilde{N}\tilde{x} + \tilde{g}), \tag{18}$$

where the $\dot{\tilde{Z}}_2$ is computed from $\tilde{Z}_3 = \dot{\tilde{Z}}_2 + S\tilde{Z}_2$.

In terms of the original data, the first block equation reads

$$U_{Z_1}^{-1} Z_1^{\mathrm{T}} \Pi^{-1} \Pi M \Theta V \frac{\mathrm{d}}{\mathrm{d}t}(Q^{-1}x) = U_{Z_1}^{-1} Z_1^{\mathrm{T}} \Pi^{-1}((\Pi N Q - \Pi M \Psi) Q^{-1} x + \Pi g)$$

or, utilizing the special block structure of the involved matrix functions,

$$Z_1^{\mathrm{T}} M V Q (Q^{-1}\dot{x} - Q^{-1}\dot{Q}Q^{-1}x) = Z_1^{\mathrm{T}}(Nx - MV\dot{Q}Q^{-1}x + g),$$

which reduces to

$$Z_1^{\mathrm{T}} M V \dot{x} = Z_1^{\mathrm{T}}(Nx + g), \tag{19}$$

which has the same form. However, the second block equation reads

$$\begin{aligned} -U_{Z_2}^{-1} Z_2^{\mathrm{T}} \Pi^{-1} (\Pi N Q - \Pi M \Psi) \frac{\mathrm{d}}{\mathrm{d}t}(Q^{-1}x) = {} & \left(\frac{\mathrm{d}}{\mathrm{d}t}(U_{Z_2}^{-1}) Z_2^{\mathrm{T}} \Pi^{-1} + U_{Z_2}^{-1} \dot{Z}_2^{\mathrm{T}} \Pi^{-1} - U_{Z_2}^{-1} Z_2^{\mathrm{T}} \Pi^{-1} \dot{\Pi} \Pi^{-1} + U_{Z_2}^{-1} Z_2^{\mathrm{T}} \Pi^{-1} S^{\mathrm{T}} \right) \\ & \times ((\Pi N Q - \Pi M \Psi) Q^{-1} x + \Pi g) \end{aligned}$$

or, with the help of (10)(c) and $Z_3M = -Z_2^{\mathrm{T}}NV^{\mathrm{T}}$,

$$\begin{aligned} -Z_2^{\mathrm{T}}NQ(Q^{-1}\dot{x} - Q^{-1}\dot{Q}Q^{-1}x) &= (\dot{Z}_2^{\mathrm{T}} + Z_2^{\mathrm{T}}S^{\mathrm{T}} - \dot{U}_{Z_2}U_{Z_2}^{-1}Z_2^{\mathrm{T}})((NQ - M\Psi)Q^{-1}x + g) \\ &= (\dot{Z}_2^{\mathrm{T}} + Z_2^{\mathrm{T}}S^{\mathrm{T}} - \dot{U}_{Z_2}U_{Z_2}^{-1}Z_2^{\mathrm{T}})(Nx + g) + Z_2^{\mathrm{T}}NV^{\mathrm{T}}\Psi Q^{-1}x \end{aligned}$$

which reduces to

$$-Z_2^{\mathrm{T}}N\dot{x} = (\dot{Z}_2^{\mathrm{T}} + Z_2^{\mathrm{T}}S^{\mathrm{T}} - \dot{U}_{Z_2}U_{Z_2}^{-1}Z_2^{\mathrm{T}})(Nx + g). \tag{20}$$

Note that (20) is different from the bottom equation in (18). Hence, in general the completion belonging to the transformed problem is not equivalent to the completion for the original problem we started with. A sufficient condition that we have equivalence is that $\dot{U}_{Z_2} = 0$ which is a restriction on the choice of the (pointwise) orthonormal columns in $\tilde{Z}_2$. It should be observed that this choice has an influence on the additional dynamics introduced in the completion.

2.3. Numerical aspects

When dealing with DAEs numerically, we are faced with the problem that while it is possible, it would in general be too costly to represent Z_1, Z_2 as smooth matrix functions. Rather, we are interested in a procedure which is based on the determination of suitable values of Z_1, Z_2 at some given point. Setting $P = Q = I_n$ in the previous section implying that $\Pi = \Theta = I$ and $\Psi = 0$, the matrix functions $\tilde{Z}_1, \tilde{Z}_2$ can be interpreted as representing the specific choice of orthonormal bases in a numerical procedure when we allow U_{Z_1}, U_{Z_2} to be non-smooth. Note that in the present case U_{Z_1}, U_{Z_2} describe transformations between orthonormal bases and are thus pointwise orthogonal. It is known that such a non-smooth selection does not lead to any problems when we integrate the DAE (8) having non-smooth coefficient functions since in standard discretization methods the transformations U_{Z_1}, U_{Z_2} simply cancel out.

If we want to integrate the completion (13), we must fix a suitable $\dot{\tilde{Z}}_2$ at a given point. For this, we may assume that we have already determined a suitable Z_2. In the context of a numerical method it is also important that the differentiation involved in $\dot{\tilde{Z}}_2$ is not performed by numerical differentiation but based on the use of differentiated data.

Following [2, Chapter 3], we know that we can choose Z_2 in such a way that

$$\begin{bmatrix} T_1'^{\mathrm{T}}M^{\mathrm{T}} \\ Z_2^{\mathrm{T}} \end{bmatrix} \dot{Z}_2 = -\begin{bmatrix} T_1'^{\mathrm{T}}\dot{M}^{\mathrm{T}} \\ 0 \end{bmatrix} Z_2, \tag{21}$$

where the columns of T_1' form a suitable orthonormal basis of cokernel(M). This also guarantees that the leading matrix function is pointwise nonsingular. Fixing $\tilde{T}_1' = T_1'U_{T_1'}$ with a non-smooth, pointwise orthogonal $U_{T_1'}$, we consider now $\dot{\tilde{Z}}_2$ as the solution of

$$\begin{bmatrix} \tilde{T}_1'^{\mathrm{T}}M^{\mathrm{T}} \\ \tilde{Z}_2^{\mathrm{T}} \end{bmatrix} \dot{\tilde{Z}}_2 = -\begin{bmatrix} \tilde{T}_1'^{\mathrm{T}}\dot{M}^{\mathrm{T}} \\ 0 \end{bmatrix} \tilde{Z}_2. \tag{22}$$

Inserting the relations $\tilde{Z}_2 = Z_2U_{Z_2}$ and $\tilde{T}_1' = T_1'U_{T_1'}$ gives

$$\begin{bmatrix} U_{T_1'}^{\mathrm{T}}T_1'^{\mathrm{T}}M^{\mathrm{T}} \\ U_{Z_2}^{\mathrm{T}}Z_2^{\mathrm{T}} \end{bmatrix} \dot{\tilde{Z}}_2 = -\begin{bmatrix} U_{T_1'}^{\mathrm{T}}T_1'^{\mathrm{T}}\dot{M}^{\mathrm{T}} \\ 0 \end{bmatrix} Z_2U_{Z_2}$$

and, therefore, $\dot{\tilde{Z}}_2 = \dot{Z}_2U_{Z_2}$. Thus, the choice of $\dot{\tilde{Z}}_2$ transforms in the same way as $\tilde{Z}_2$ such that U_{Z_2} can be simply removed in (18). In particular, discretizing (18) with the so constructed possibly non-smooth realizations $\tilde{Z}_1, \tilde{Z}_2, \dot{\tilde{Z}}_2$ gives the same numerical solutions as directly discretizing (13).

3. Nonlinear DAEs

In the general case of unstructured nonlinear DAEs (1), the derivative array equations obtained by differentiating (1) ℓ times are given by

$$F_\ell(t, z, \dot{z}, \ldots, z^{(\ell+1)}) = 0, \tag{23}$$

that is,

$$F_\ell(t, z, \dot{z}, \ldots, z^{(\ell+1)}) = \begin{bmatrix} F(t, z, \dot{z}) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}F(t, z, \dot{z}) \\ \vdots \\ \dfrac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}F(t, z, \dot{z}) \end{bmatrix}. \tag{24}$$

The following hypothesis corresponds to Hypothesis 1 in the case of linear DAEs with variable coefficients.

Hypothesis 4. There exist integers μ, a, and d such that $\mathbb{L}_\mu = \{z_\mu \in \mathbb{I}\times\mathbb{R}^n\times\mathbb{R}^n\times\cdots\times\mathbb{R}^n \mid F_\mu(z_\mu) = 0\}$ is not empty and for every point $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ there exists a (sufficiently small) neighborhood in which the following properties hold:

1. We have $\operatorname{rank}(F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}) = (\mu+1)n - a$ on $\mathbb{L}_\mu$. This implies that there exists a smooth full rank matrix function Z_2 of size $((\mu+1)n, a)$ satisfying
$$Z_2^{\mathrm{T}} F_{\mu;\dot{x},\ldots,x^{(\mu+1)}} = 0$$
on $\mathbb{L}_\mu$.
2. We have $\operatorname{rank}(Z_2^{\mathrm{T}} F_{\mu;x}) = a$ on $\mathbb{L}_\mu$. This implies that there exists a smooth full rank matrix function T_2 of size $(n, n-a)$ satisfying
$$Z_2^{\mathrm{T}} F_{\mu;x} T_2 = 0.$$
3. We have $\operatorname{rank}(F_{\dot{x}} T_2) = d = n - a$. This implies that there exists a smooth full rank matrix function Z_1 of size (n, d) satisfying
$$\operatorname{rank} Z_1^{\mathrm{T}} F_{\dot{x}} T_2 = d.$$

Again, alternative characterizations exist [11,13], but the preceding formulas fit our numerical procedures better. We may assume that μ is chosen minimally and set $\nu = \mu + 1$ as in Section 2. For convenience, we use the shorthand notation $y = (\dot{x}, \ldots, x^{(\mu+1)})$. Given $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ we set

$$\hat{Z}_1 = Z_1(t_0, x_0, y_0), \qquad \hat{Z}_2 = Z_2(t_0, x_0, y_0).$$

Moreover, due to Hypothesis 4 we can choose a $\hat{T}_1$ such that

$$\begin{bmatrix} F_{\mu;y}(t_0, x_0, y_0) & \hat{Z}_2 \\ \hat{T}_1^{\mathrm{T}} & 0 \end{bmatrix}$$

is nonsingular. Defining

$$H(t, x, y, w) = \begin{bmatrix} F_\mu(t, x, y) + \hat{Z}_2 w \\ \hat{T}_1^{\mathrm{T}}(y - y_0) \end{bmatrix}$$

we immediately see that $H(t_0, x_0, y_0, 0) = 0$ and that $H_{y,w}(t_0, x_0, y_0, 0)$ is nonsingular. Hence, the implicit function theorem shows that the equation $H(t, x, y, w) = 0$ can locally be solved for y, w, say according to

$$y = K(t, x), \qquad w = L(t, x).$$

Obviously, every (t, x) with $L(t, x) = 0$ satisfies $F_\mu(t, x, K(t, x)) = 0$ and hence x is consistent at point t. But also the converse holds, i.e., if x is consistent at point t then (t, x) satisfies $L(t, x) = 0$. See [2, Section 7.2] for more details. It follows that the relation $L(t, x) = 0$ constitutes all constraints imposed by the given DAE. Moreover, the problem

$$\hat{Z}_1^{\mathrm{T}} F(t, x, \dot{x}) = 0, \tag{25a}$$

$$L(t, x) = 0 \tag{25b}$$

is an index reduced DAE belonging to the original DAE. In particular, locally it possesses the same solutions as the original DAE (1) but is index one. Note that we are able to evaluate L (and K) numerically by means of Newton's method applied to $H(t, x, y, w) = 0$. Simply differentiating the constraints, a possible completion is then implicitly defined by

$$\hat{Z}_1^{\mathrm{T}} F(t, x, \dot{x}) = 0, \tag{26a}$$

$$L_t(t, x) + L_x(t, x)\dot{x} = 0. \tag{26b}$$

The involved derivatives can be obtained numerically due to the implicit function theorem by solving the linear system

$$\begin{bmatrix} F_{\mu;y}(t, x, K(t, x)) & \hat{Z}_2 \\ \hat{T}_1^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} K_t(t, x) & K_x(t, x) \\ L_t(t, x) & L_x(t, x) \end{bmatrix} = -\begin{bmatrix} F_{\mu;t}(t, x, K(t, x)) & F_{\mu;x}(t, x, K(t, x)) \\ 0 & 0 \end{bmatrix}. \tag{27}$$

In a numerical realization of this approach one can combine the systems $H(t, x, y, w) = 0$, (26) and (27) in the form

$$H(t, x, y, w) = 0, \tag{28a}$$

$$\hat{Z}_1^{\mathrm{T}} F(t, x, \dot{x}) = 0, \tag{28b}$$

$$L_1 + L_2\dot{x} = 0, \tag{28c}$$

$$\begin{bmatrix} F_{\mu;y}(t, x, y) & \hat{Z}_2 \\ \hat{T}_1^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} * & * \\ L_1 & L_2 \end{bmatrix} = -\begin{bmatrix} F_{\mu;t}(t, x, y) & F_{\mu;x}(t, x, y) \\ 0 & 0 \end{bmatrix}, \tag{28d}$$

where the last relation (28d) can be used to eliminate the unknowns L_1 and L_2. Note that utilizing in this way the structure of the nonlinear system (28) the computational costs compared with the standard integration of the DAE as by the general purpose code GENDA [14,15] are only slightly increased due to the additional bordering given by (28c).

Standard integration has to discretize (28b), say by BDF, and then to solve this together with (28a) and $w = 0$ by a nonlinear equation solver. An efficient implementation of such a solver can be based on the solution of linear systems with coefficient matrices as in (28d). Thus, the only additional part which comes into play because of the completion are the a equations of (28c) which can be decoupled by first dealing with (28a) and (28b). The overall additional costs of solving (28) compared with standard integration therefore consists of the decoupling of (28c) and of the determination of the Jacobian belonging to (28c), say by numerical differentiation.

4. First integrals and stabilization

When constructing a completion, we lose the information that the solution of the original DAE must satisfy the algebraic constraints. In the completions presented in the previous sections we rather have the property that the original constraints describe invariants or first integrals of the constructed ODEs. In this section we therefore consider ODEs which possess first integrals and study modifications of them to insure that solutions which do not satisfy a given first integral with a prescribed value at least yield values of the first integral which tend to the prescribed value exponentially. Moreover, solutions with the prescribed value of the first integral should not be altered. We then apply the results to the presented completions.

Consider the ODE

$$\dot{x} = f(t, x) \tag{29}$$

with $f : \mathbb{I} \times \mathbb{D} \to \mathbb{R}^n$, $\mathbb{D} \subseteq \mathbb{R}^n$ open, and let $I : \mathbb{I} \times \mathbb{D} \to \mathbb{R}^m$, $m \le n$, be a first integral of (29), that is,

$$I_t(t, x) + I_x(t, x)f(t, x) = 0 \quad \text{for all } (t, x) \in \mathbb{I} \times \mathbb{D}. \tag{30}$$

This property immediately implies that $I(t, x)$ stays constant along every solution of (29). If I_0 is the given prescribed value of the first integral, we want to modify (29) in such a way that $g : \mathbb{I} \times \mathbb{D} \to \mathbb{R}^m$ defined by

$$g(t, x) = I(t, x) - I_0 \tag{31}$$

tends to zero exponentially along every solution of (29) at least when $g(t, x)$ is already sufficiently small. That is, we want to construct a stabilization of the constraint $g(t, x) = 0$. As was done for completions, this kind of stabilization should be defined in a smooth way and it should be easy to carry out numerically.

4.1. Stabilization by Gauß–Newton flows

The stabilization we are going to present will be based on a generalized Gauß–Newton flow for the nonlinear equation $g(t, x) = 0$. In particular, it involves a certain class of generalized inverses of the Jacobian $g_x(t, x)$ which is still assumed to be of full row rank. A Gauß–Newton flow can be viewed as a continuous Newton method for solving $g = 0$.

In order to introduce the class under consideration let $A \in \mathbb{R}^{m,n}$ be a matrix with full row rank and let $R \in \mathbb{R}^{n,n}$ be nonsingular. We introduce R both as a design parameter and also because earlier work on the linear case showed that some different appearing completions could be viewed as coming from different choices of generalized inverses.

Lemma 5. *The matrix $A^- \in \mathbb{R}^{n,m}$ defined by $A^- = R^{-1}(AR^{-1})^+$ is the unique generalized inverse satisfying*

$$AA^-A = A, \qquad A^-AA^- = A^-, \qquad AA^- = I_m, \qquad RA^-AR^{-1} \text{ symmetric}. \tag{32}$$

Proof. It is easy to see that $A^- = R^{-1}(AR^{-1})^+$ satisfies the four properties stated in (32). On the other hand, setting $B = AR^{-1}$ and $B^+ = RA^-$ the properties (32) imply $BB^+B = B$, $B^+BB^+ = B^+$, and BB^+ and B^+B are symmetric. Hence, B^+ is the Moore–Penrose pseudoinverse of B. □

For later use we need an appropriate form of the property that $A^-Ax = x$ holds for a vector $x \in \mathbb{R}^n$. That is, x is in the range of the projection A^-A. Inserting the definition of A^- into A^-A we get the condition $(AR^{-1})^+(AR^{-1})Rx = Rx$ or $Rx \in \text{cokernel}(AR^{-1})$. That is,

$$T^\mathrm{T}R^\mathrm{T}Rx = 0,$$

where the columns of T span kernel(A).

A Gauß–Newton flow for the nonlinear equation $g(t, x) = 0$ based on the given class of generalized inverses of Lemma 5 is given by

$$\dot{x} = -g_x(t, x)^-g(t, x), \tag{33}$$

where $g_x(t, x)^- = R(t, x)^{-1}(g_x(t, x)R(t, x)^{-1})^+$ with an appropriately chosen smooth, pointwise nonsingular matrix function R. The smoothness of R together with the smoothness and constant rank of g give the smoothness of the

Moore–Penrose pseudoinverse and hence guarantee the smoothness of $g(t,x)^-$. Note that the standard Gauß–Newton flow is included as a special case for the choice $R(t,x) = I_n$ in which case $g(t,x)^- = g(t,x)^+$.

We now combine the Gauß–Newton flow (33) with the original flow (29) in the form

$$\dot{x} = f(t,x) - Cg_x(t,x)^- g(t,x), \tag{34}$$

where $C \in \mathbb{R}^{n,n}$ is a coupling matrix which we are still free to choose to suit our purposes. Note that we have two parameters at our disposal, R and C. It is obvious that a solution of (29) which satisfies the constraint $g(t,x) = 0$ also solves (34). The question now is how a solution $\tilde{x}$ of (34) behaves when $g(t,\tilde{x}(t))$ is sufficiently small in norm.

We shall use the implicit function theorem. For this consider

$$H(t,x,\tilde{x}) = \begin{bmatrix} g(t,x) \\ T(t,x)^{\mathrm{T}} R(t,x)^{\mathrm{T}} R(t,x)(\tilde{x} - x) \end{bmatrix},$$

where the columns of $T(t,x)$ form a smoothly parameterized (orthonormal) basis of kernel($g_x(t,x)$). We then have that $H(t,x,x) = 0$ for all $(t,x) \in g^{-1}(\{0\})$. Note that

$$H_x(t,x,x) = \begin{bmatrix} g_x(t,x) \\ -T(t,x)^{\mathrm{T}} R(t,x)^{\mathrm{T}} R(t,x) \end{bmatrix}. \tag{35}$$

We need (35) to be nonsingular. Multiply (35) on the right by $[\,T' \quad T\,]$ where T' is such that $[\,T' \quad T\,]$ is pointwise orthogonal. Then (35) becomes

$$\begin{bmatrix} g_x(t,x)T'(t,x) & 0 \\ * & -T(t,x)^{\mathrm{T}} R(t,x)^{\mathrm{T}} R(t,x) T(t,x) \end{bmatrix}. \tag{36}$$

The matrix in (36) is nonsingular since $g_x(t,x)T'(t,x)$ is nonsingular due to the full row rank of $g_x(t,x)$ and since RT is full column rank. Hence, the equation $H(t,x,\tilde{x}) = 0$ can be locally solved for x in terms of $(t,\tilde{x})$, say according to $x = S(t,\tilde{x})$. The solution $\tilde{x}$ of (34) therefore defines locally a function x by

$$x(t) = S(t,\tilde{x}(t)). \tag{37}$$

By construction, the function x satisfies

$$g(t,x(t)) = 0, \tag{38a}$$

$$g_t(t,x(t)) + g_x(t,x(t))\dot{x}(t) = 0, \tag{38b}$$

$$g_x(t,x(t))^- g_x(t,x(t))(\tilde{x}(t) - x(t)) = \tilde{x}(t) - x(t). \tag{38c}$$

Note that at this point $x(t)$ from (37) is not a solution of a specific differential equation. It is just a point on the solution manifold of the DAE at time t. Since g is related to a first integral according to (31), the property (30) implies

$$g_t(t,x(t)) + g_x(t,x(t))f(t,x(t)) = 0 \tag{39}$$

and thus $g_x(t,x(t))\dot{x}(t) = g_x(t,x(t))f(t,x(t))$ from (38b). We are now able to evaluate (omitting obvious arguments of the functions)

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{x} - x\|_2^2 &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|g_x(t,x)^- g_x(t,x)(\tilde{x} - x)\|_2^2 \\ &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}[(\tilde{x} - x)^{\mathrm{T}}(g_x(t,x)^- g_x(t,x))^{\mathrm{T}}(g_x(t,x)^- g_x(t,x))(\tilde{x} - x)] \\ &= (\tilde{x} - x)^{\mathrm{T}} g_x(t,x)^- g_x(t,x)(\dot{\tilde{x}} - \dot{x}) + (\tilde{x} - x)^{\mathrm{T}}\frac{\mathrm{d}}{\mathrm{d}t}[g_x(t,x)^- g_x(t,x)](\tilde{x} - x). \end{aligned} \tag{40}$$

For the first term of (40) we get

$$\begin{aligned} (\tilde{x} - x)^{\mathrm{T}} g_x(t,x)^- g_x(t,x)(\dot{\tilde{x}} - \dot{x}) &= (\tilde{x} - x)^{\mathrm{T}} g_x(t,x)^- g_x(t,x)(f(t,\tilde{x}) - Cg_x(t,\tilde{x})^- g(t,\tilde{x}) - f(t,x)) \\ &= (\tilde{x} - x)^{\mathrm{T}} g_x(t,x)^- g_x(t,x) f_x(t,x)(\tilde{x} - x) \\ &\quad - (\tilde{x} - x)^{\mathrm{T}} g_x(t,x)^- g_x(t,x) C g_x(t,x)^- g_x(t,x)(\tilde{x} - x) + (\tilde{x} - x)^{\mathrm{T}} r(t,x,\tilde{x})(\tilde{x} - x) \end{aligned}$$

by Taylor expansion, where we used the fact in the last step that $g(t,x) = 0$ holds. The remainder term is bounded according to $r(t,x,\tilde{x}) = \mathcal{O}(\|\tilde{x} - x\|_2)$. Let L be a constant such that

$$\left\| g_x(t,x)^- g_x(t,x) f_x(t,x) + \frac{\mathrm{d}}{\mathrm{d}t}[g_x(t,x)^- g_x(t,x)] + r(t,x,\tilde{x}) \right\|_2 \le L. \tag{41}$$

If $R = I$, then $\|g_x(t,x)^- g_x(t,x)\|_2 = 1$. Under the assumption that

$$v^{\mathrm{T}} g_x(t,x)^- g_x(t,x) C g_x(t,x)^- g_x(t,x) v \ge \lambda \|v\|_2^2 \tag{42}$$

for all $v \in \operatorname{range}(g_x(t,x)^- g_x(t,x))$, it then follows that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\tilde{x}-x\|_2^2 \le (L-\lambda)\|\tilde{x}-x\|_2^2. \tag{43}$$

Note that given L, then (42) scales linearly in C, λ. Thus by increasing C we can get $L-\lambda < 0$.

Theorem 6. *Suppose that L is given by (41). Pick $\lambda > L$. Choose the coupling matrix $C \in \mathbb{R}^{n,n}$ in such a way that (42) holds. Then for a solution $\tilde{x}$ of (34) with $g(t,\tilde{x}(t))$ sufficiently small in norm we have*

$$\|\tilde{x}(t) - x(t)\|_2 \le M\mathrm{e}^{(L-\lambda)t} \tag{44}$$

with an appropriate constant $M > 0$ and $x(t)$ satisfies $g(t, x(t)) = 0$.

Proof. The claim follows directly from (43) by application of Gronwall's lemma, see, e.g., [16]. □

Remark 7. The property (42) always holds for the choice $C = \lambda I_n$. In the case of $g_x(t,x)^- = g_x(t,x)^+$ the condition (42) is equivalent to

$$v^{\mathrm{T}}(g_x(t,x)^+ g_x(t,x))^{\mathrm{T}} C (g_x(t,x)^+ g_x(t,x)) v \ge \lambda \|v\|_2^2,$$

which holds for every $v \in \operatorname{range}(g_x(t,x)^+ g_x(t,x))$ if $C \in \mathbb{R}^{n,n}$ is an arbitrary symmetric positive definite matrix with λ being its smallest eigenvalue.

Remark 8. The claim of Theorem 6 remains valid if g only represents an invariant according to

$$g_t(t,x) + g_x(t,x)f(t,x) = 0 \quad \text{for all } (t,x) \text{ with } g(t,x) = 0,$$

which is sufficient to guarantee (39).

Example 9. Consider the ODE

$$\dot{x}_1 = 1, \qquad \dot{x}_2 = x_2 \tag{45}$$

together with

$$g(x_1,x_2) = x_2\mathrm{e}^{-x_1}.$$

The solution manifold of $g(x) = 0$ is just $\{(x_1,0)^{\mathrm{T}} \mid x_1 \in \mathbb{R}\}$. Because of

$$g_x(x)f(x) = \begin{bmatrix} -x_2\mathrm{e}^{-x_1} & \mathrm{e}^{-x_1} \end{bmatrix} \begin{bmatrix} 1 \\ x_2 \end{bmatrix} = 0,$$

the function g represents a first integral of (45). With the choice

$$g_x(x)^- = \begin{bmatrix} 0 \\ \mathrm{e}^{x_1} \end{bmatrix}$$

and $C = \lambda I$ in (34), we get that (45) becomes the ODE

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 \\ x_2 \end{bmatrix} - \lambda \begin{bmatrix} 0 \\ \mathrm{e}^{x_1} \end{bmatrix} x_2\mathrm{e}^{-x_1} = \begin{bmatrix} 1 \\ (1-\lambda)x_2 \end{bmatrix}.$$

Obviously, the solution manifold of $g(x) = 0$ is stable only for $\lambda > 1$. This contradicts Propositions 2.1 and Proposition 2.2 in [9], where it is claimed that $\lambda > 0$ already guarantees stability. The problem is that $g(x)$ tending to zero does not imply that x approaches the solution manifold of $g(x) = 0$. For example, g goes to zero for $\dot{x}_1 = 1$, $\dot{x}_2 = \frac{1}{2}x_2$ but x_2 does not go to zero. Moreover, arguing with the help of a Lyapunov function requires the solution manifold to be bounded, see [17]. This is reflected in the presence of our bound L.

4.2. Application to DAEs

The completions of Sections 2 and 3 have the property that they introduce a trivial dynamics with respect to the original constraints in the sense that the computed constraints for the DAE represent a first integral of the completion. In particular, the completions fit into the framework of Section 4.1 when we take I as the part of the index reduced DAE that describes the constraints and set $I_0 = 0$.

In the case of linear DAEs with variable coefficients the constraint equation is $Z_2^{\mathrm{T}}(Nx+g) = 0$. Adding the corresponding (standard) Gauß–Newton flow to (14), we obtain the stabilized completion

$$\dot{x} = V^{\mathrm{T}}M^-(Nx+g) - C(Z_2^{\mathrm{T}}N)^+ Z_2^{\mathrm{T}}(Nx+g). \tag{46}$$

In the nonlinear case, let (26) imply $\dot{x} = f(t,x)$. The constraint equation is here given by $L(t,x) = 0$. Adding the corresponding (standard) Gauß–Newton flow to (26), we obtain the stabilized completion

$$\dot{x} = f(t,x) - CL_x(t,x)^+ L(t,x). \tag{47}$$

Theorem 6 is applicable in both cases.

In the special case that the index reduced problem is an index one semi-explicit DAE one can also proceed in a slightly different way. Let

$$\dot{x}_1 = f(t,x_1,x_2), \qquad 0 = g(t,x_1,x_2) \tag{48}$$

be a semi-explicit DAE of index $\nu = 1$, i.e., let the constraint equation $g(t,x_1,x_2) = 0$ be solvable and let $g_{x_2}(t,x_1,x_2)$ be invertible on the corresponding solution set. The DAE (48) then satisfies Hypothesis 4 with $\mu = 0$, the quantities d and a being the sizes of x_1 and x_2. The completion described in Section 3 is given by

$$\dot{x}_1 = f(t,x_1,x_2), \tag{49a}$$

$$\dot{x}_2 = -g_{x_2}(t,x_1,x_2)^{-1}(g_t(t,x_1,x_2) + g_{x_1}(t,x_1,x_2)f(t,x_1,x_2)). \tag{49b}$$

The stabilization (47) modifies both right-hand sides in (49). If the original first equation should be preserved in a stabilized version, one can proceed as follows. Defining

$$g_x(t,x_1,x_2)^- = \begin{bmatrix} 0 \\ g_{x_2}(t,x_1,x_2)^{-1} \end{bmatrix} \tag{50}$$

we get a generalized inverse of $g_x(t,x_1,x_2)$ which is covered by Lemma 5. With this choice we obtain (omitting arguments)

$$g_x^- g_x = \begin{bmatrix} 0 & 0 \\ g_{x_2}^{-1} g_{x_1} & I \end{bmatrix}$$

and $v \in \operatorname{range}(g_x^- g_x)$ if and only if $v^{\mathrm{T}} = [\,0 \quad w^{\mathrm{T}}\,]$ with arbitrary vector w. Taking

$$C = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{C} \end{bmatrix}$$

and observing $v^{\mathrm{T}} v = w^{\mathrm{T}} w$, we see that (42) holds if and only if $w^{\mathrm{T}} \tilde{C} w \geq \lambda \|w\|_2^2$. Hence Theorem 6 is applicable if we choose $\tilde{C}$ symmetric and positive definite with λ being its smallest eigenvalue. Due to the block structure of C and g_x^-, the stabilizing Gauß–Newton flow only affects the (49).

Remark 10. Stabilized differentiation, or Baumgarte stabilization [10], for (48) yields $\dot{x}_1 = f(t,x_1,x_2)$ together with

$$g_t(t,x_1,x_2) + g_{x_1}(t,x_1,x_2)\dot{x}_1 + g_{x_2}(t,x_1,x_2)\dot{x}_2 = -\lambda g(t,x_1,x_2).$$

Solving for $\dot{x}_2$ gives the ODE

$$\dot{x}_1 = f(t,x_1,x_2), \tag{51a}$$

$$\dot{x}_2 = -g_{x_2}(t,x_1,x_2)^{-1}(g_t(t,x_1,x_2) + g_{x_1}(t,x_1,x_2)f(t,x_1,x_2)) - \lambda g_{x_2}(t,x_1,x_2)^{-1} g(t,x_1,x_2). \tag{51b}$$

Comparing (51) with (34), we see that (51) just corresponds to the choice (50) of the generalized inverse together with the choice $\tilde{C} = \lambda I$ in the coupling matrix C.

Example 11. As in [6] we consider the semi-explicit DAE

$$\dot{x}_1 = \beta x_1, \qquad 0 = \mathrm{e}^{\alpha t}(x_1 - x_2)$$

of index $\nu = 1$. Due to its structure the completion presented in this paper coincides with the least-squares completion of [6] and is given by

$$\dot{x}_1 = \beta x_1, \qquad \dot{x}_2 = (\alpha + \beta)x_1 - \alpha x_2.$$

Note that this completion is a linear ODE with constant coefficients. The eigenvalues are $\{\beta, -\alpha\}$ indicating that the additional dynamics described by $-\alpha$ may be stable or unstable. Since $g(t,x) = \mathrm{e}^{\alpha t}(x_1 - x_2)$, we have $g_x(t,x) = \mathrm{e}^{\alpha t}[\,1 \quad -1\,]$ and hence

$$g_x(t,x)^+ = \frac{1}{2}\mathrm{e}^{-\alpha t}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad g_x(t,x)^- = \mathrm{e}^{-\alpha t}\begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Using the standard Gauß–Newton flow with $C = \lambda I$ the stabilized ODE has the form

$$\dot{x}_1 = \beta x_1 - \frac{1}{2}\lambda(x_1 - x_2), \qquad \dot{x}_2 = (\alpha + \beta)x_1 - \alpha x_2 + \frac{1}{2}\lambda(x_1 - x_2)$$

with eigenvalues $\{\beta, -(\alpha+\lambda)\}$. Using the Gauß–Newton flow on the basis of g_x^- with $\tilde{C} = \lambda I$ on the other hand gives the stabilized ODE

$$\dot{x}_1 = \beta x_1, \qquad \dot{x}_2 = (\alpha+\beta)x_1 - \alpha x_2 + \lambda(x_1 - x_2)$$

which also has eigenvalues $\{\beta, -(\alpha+\lambda)\}$. Both systems only differ in the eigenvector that belongs to the second eigenvalue. Obviously we must choose $\lambda > -\alpha$ to obtain stable systems.

5. Numerical experiments

Since the stabilized ODEs in Example 11 are linear with constant coefficients it is at once clear how they behave under numerical discretization. For our numerical experiments we therefore choose the nonlinear DAE

$$\dot{x}_1 = x_4, \qquad \dot{x}_4 = 2x_1x_7, \tag{52a}$$

$$\dot{x}_2 = x_5, \qquad \dot{x}_5 = 2x_2x_7, \tag{52b}$$

$$\dot{x}_3 = x_6, \qquad \dot{x}_6 = -1 - x_7, \tag{52c}$$

$$0 = x_1^2 + x_2^2 - x_3 \tag{52d}$$

due to [13]. It describes the three-dimensional motion of a mass point restricted to an upright parabolic bowl under gravity. It is known to satisfy Hypothesis 4 with $\mu = 2$, $d = 4$, and $a = 3$. As described earlier, a stabilized completion may be computed directly from the derivative array of (52). However, it is instructive to exploit the structure of example (52). Differentiating the constraint and eliminating the differentiated variables with the help of the other equations of the DAE yields the hidden constraint

$$0 = 2x_1x_4 + 2x_2x_5 - x_6. \tag{53}$$

Differentiating once more and eliminating derivatives gives

$$0 = 2x_4^2 + 4x_1^2x_7 + 2x_5^2 + 4x_2^2x_7 + x_7 + 1. \tag{54}$$

Observing that the constraints (52d), (53) and (54) can be solved for (x_3, x_6, x_7), an equivalent index reduced DAE is given by

$$\dot{x}_1 = x_4, \qquad \dot{x}_4 = 2x_1x_7, \tag{55a}$$

$$\dot{x}_2 = x_5, \qquad \dot{x}_5 = 2x_2x_7, \tag{55b}$$

$$0 = x_1^2 + x_2^2 - x_3, \tag{55c}$$

$$0 = 2x_1x_4 + 2x_2x_5 - x_6, \tag{55d}$$

$$0 = 2x_4^2 + 4x_1^2x_7 + 2x_5^2 + 4x_2^2x_7 + x_7 + 1. \tag{55e}$$

The completion of Section 3 is then obtained by replacing the constraints $g(x) = 0$, where

$$g(x) = \begin{bmatrix} x_1^2 + x_2^2 - x_3 \\ 2x_1x_4 + 2x_2x_5 - x_6 \\ 2x_4^2 + 4x_1^2x_7 + 2x_5^2 + 4x_2^2x_7 + x_7 + 1 \end{bmatrix},$$

by

$$g_x(x)\dot{x} = 0 \tag{56}$$

and then solving (56) along with (55a) and (55b) for $\dot{x}$. A straightforward computation yields

$$\begin{aligned} &\dot{x}_1 = x_4, \qquad \dot{x}_4 = 2x_1x_7, \\ &\dot{x}_2 = x_5, \qquad \dot{x}_5 = 2x_2x_7, \\ &\dot{x}_3 = 2x_1x_4 + 2x_2x_5, \\ &\dot{x}_6 = 2x_4^2 + 4x_1^2x_7 + 2x_5^2 + 4x_2^2x_7, \\ &\dot{x}_7 = -16x_7(x_1x_4 + x_2x_5)/(1 + 4x_1^2 + 4x_2^2). \end{aligned}$$

Observing that

$$g_x(x) = \begin{bmatrix} 2x_1 & 2x_2 & -1 & 0 & 0 & 0 & 0 \\ 2x_4 & 2x_5 & 0 & 2x_1 & 2x_2 & -1 & 0 \\ 8x_1x_7 & 8x_2x_7 & 0 & 4x_4 & 4x_5 & 0 & 1 + 4x_1^2 + 4x_2^2 \end{bmatrix}$$

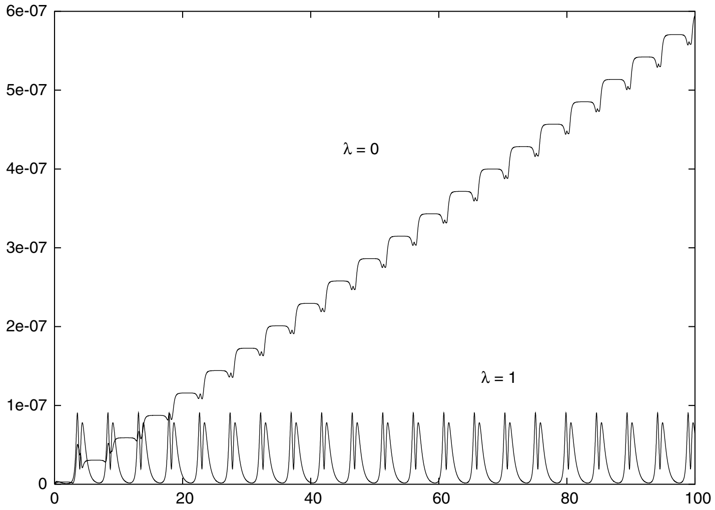

Fig. 1. Norm of constraint residual for $\lambda = 0$ and $\lambda = 1$ using $g_x(t, x)^+$.

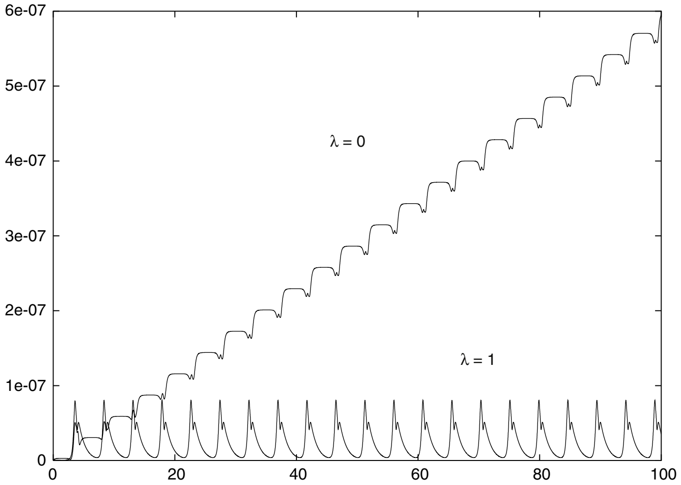

Fig. 2. Norm of constraint residual for $\lambda = 0$ and $\lambda = 1$ using $g_x(t, x)^-$.

natural stabilizations are given by $g_x(x)^+$ and by

$$g_x(x)^- = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/(1+4x_1^2+4x_2^2) \end{bmatrix}^{\mathrm{T}}.$$

Both completions have been solved with various choices of λ in $C = \lambda I$ and $\tilde{C} = \lambda I$ using the classical Runge–Kutta method with fixed stepsize. Figs. 1 and 2 show the behavior of $\|g(t, x)\|_2$ for $\lambda = 0$, i.e., for the un-stabilized completion, and for $\lambda = 1$ when we use $g_x(t, x)^+$ and $g_x(t, x)^-$, respectively. Recall that using $g_x(t, x)^-$ corresponds to Baumgarte stabilization. The constant stepsize was $h = 0.01$. One can easily observe the drift effect for $\lambda = 0$, which disappears in the stabilized version. It should be mentioned that it is clear that the discretization becomes unstable for too large values of $h\lambda$.

6. Conclusions

We have examined the problem of constructing computable completions of vector fields for nonlinear DAEs. This was done by using a family of weighted Gauß–Newton flows. By varying the weights R, C in the definition of the flow we were able to unify the development of some of the previously presented completions for the linear time-varying case. For the nonlinear DAEs we do not require the explicit representation of constraints in the original DAE. Our constraint characterization is based on pointwise numerical calculations and thus applies to general solvable nonlinear DAEs. Our results also correct some earlier results in the literature.

Many software modeling packages assume an ODE model. This paper lays the groundwork for having that ODE model given by a call to a subroutine that generates the right-hand side of a stabilized completion of the DAE of interest. We note that most of the needed numerical subroutines used to generate the submatrices needed in the completion have already been written and tested in such codes as GENDA [14,15] where they are used for a different purpose.

Acknowledgements

We thank an anonymous referee for detailed comments and suggestions that helped to improve the readability of the paper.

The first author's work has been supported in part by the National Science Foundation under DMS-0101802 and ECS-0620986.

References

[1] K.E. Brenan, S.L. Campbell, L.R. Petzold, Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations, SIAM, Philadelphia, 1996.
[2] P. Kunkel, V. Mehrmann, Differential-Algebraic Equations, Analysis and Numerical Solution, EMS Publishing House, Zürich, Switzerland, 2006.
[3] C. Arevalo, S.L. Campbell, M. Selva, Unitary partitioning in general constraint preserving DAE integrators, Math. Comput. Modelling 40 (2004) 1273–1284.
[4] E. Moore, S.L. Campbell, Constraint preserving integrators for general nonlinear higher index DAEs, Numer. Math. 69 (1995) 383–399.
[5] S.L. Campbell, Uniqueness of completions for linear time varying differential algebraic equations, Linear Algebra Appl. 161 (1992) 55–67.
[6] I. Okay, S.L. Campbell, P. Kunkel, The additional dynamics of least squares completions for linear differential algebraic equations, Linear Algebra Appl. 425 (2007) 471–485.
[7] I. Okay, S. L. Campbell, P. Kunkel, Completions of implicitly defined vector fields and their applications, in: Proc. 18th International Symposium on Mathematical Theory of Networks and Systems, MTNS 08, Blacksburg, Virginia, 2008.
[8] I. Okay, S.L. Campbell, P. Kunkel, Completions of implicitly defined linear time varying vector fields, Linear. Algebra Appl. 431 (2009) 1422–1438.
[9] U.M. Ascher, H. Chin, S. Reich, Stabilization of DAEs and invariant manifolds, Numer. Math. 67 (1994) 131–149.
[10] J. Baumgarte, Stabilization of constraints and integrals of motion in dynamical systems, Comput. Methods Appl. Mech. Engrg. 1 (1972) 1–16.
[11] S.L. Campbell, E. Griepentrog, Solvability of general differential algebraic equations, SIAM J. Sci. Comput. 16 (1995) 257–270.
[12] S.L. Campbell, C.D. Meyer Jr., Generalized Inverses of Linear Transformations, SIAM, Philadelphia, 2008.
[13] W.C. Rheinboldt, Differential-algebraic systems as differential equations on manifolds, Math. Comp. 43 (1984) 473–482.
[14] P. Kunkel, V. Mehrmann, Home page for general nonlinear differential algebraic equation solver, http://www.math.tu-berlin.de/numerik/mt/NumMat/Software/GENDA/info.shtml.
[15] P. Kunkel, V. Mehrmann, I. Seufer, GENDA: A software package for the numerical solution of general nonlinear differential-algebraic equations, Institut für Mathematik, TU Berlin Technical Report 730, Berlin, Germany, 2002.
[16] D. Hinrichsen, A.J. Pritchard, Mathematical Systems Theory I. Modelling, State Space Analysis, Stability and Robustness, Springer-Verlag, New York, NY, 2005.
[17] J.P. LaSalle, Recent advances in Liapunov stability theory, SIAM Rev. 6 (1964) 1–11.