DIFFERENTIAL ALGEBRAIC SYSTEMS ANEW

Roswitha März*

Abstract

Linear differential algebraic equations with properly stated leading term are considered via a decoupling into their essential parts. It is shown why for so-called numerically well formulated equations the decoupling and discretizations commute in some sense. In general one cannot expect this commutativity so that additional difficulties like strong stepsize restrictions may arise. Moreover, abstract differential algebraic equations in infinite dimensional Hilbert spaces are discussed. In particular, a linear-quadratic control problem for those equations is proved to be solvable.

Key words: differential algebraic equations, numerical integration methods, linear-quadratic control, abstract differential algebraic equations

AMS subject classification: 65L80, 65L06, 34A09, 49J27, 49N10

1 Introduction

When dealing with standard linear time-varying coefficient DAEs

$$Ex' + Fx = q \tag{1.1}$$

and their adjoint equations

$$(E^*y)' - F^*y = p \tag{1.2}$$

one is confronted with a kind of unsightly dissymmetry. These equations are of different type. In [3], more precise formulations

$$E(P_E x)' + (F - EP_E')x = q, \tag{1.3}$$

$$P_E^*(E^*y)' - (F^* - P_E^{*'}\,E^*)y = p \tag{1.4}$$

*Humboldt University, Institute of Mathematics, D-10099 Berlin, Germany (iam@mathematik.hu-berlin.de)

are used, where P_E denotes a projector function along $\ker E$, say $P_E = E^+E$. These formulations show a little bit more symmetry. However, (1.4) is treatet in [3] via an auxiliary enlarged system that has a correspondig projector inside the derivative like (1.3). This formal uncompleteness makes us to treat general equations of the form ([2])

$$A(Dx)' + Bx = q, \tag{1.5}$$

where A, D and B are continuous matrix functions, and A and D are well matched so that D precisely figures out which derivatives are actually involved. Neither D nor A have to be projectors. The adjoint equation

$$D^*(A^*y)' - B^*y = p \tag{1.6}$$

now fits nicely into this form.
There are remarkable benefits of this new approach:
The inherent regular differential equation of a DAE (1.5) is now uniquely determined (cf. [2] and Section 2). Nice generalizations of fundamental classical results (like the relation $Y^* = X^{-1}$ for fundamental matrices) are obtained, but also further symmetries concerning the characteristic subspaces and the index of (1.5) and (1.6), respectively ([2]). Further, the boundary value problem resulting from linear-quadratic control problems looks now nice and transparent (cf. Section 5).

The nonlinear counterparts of (1.5)

$$A(x(t),t)(D(t)x(t))' + b(x(t),t) = 0 \tag{1.7}$$

and

$$A(x(t),t)(d(x(t),t))' + b(x(t),t) = 0 \tag{1.8}$$

should be very welcome in applications, e. g. in circuit simulation. In [13], [14], those nonlinear equations are studied as well as numerical integration methods. To be more transparent we do not touch at all the nonlinear cases in the present paper.
On the one hand, the new approach (1.5) allows to maintain all theoretical and numerical results concerning (1.1) via (1.3), but also those via the formulation

$$R_E(Ex)' + (F - R_E E')x = q \ , \tag{1.9}$$

where R_E is a projector function along $\operatorname{im} E$.

On the other hand, a generalization to *abstract differential algebraic equations* in infinite dimensional spaces becomes possible (cf. Sections 4, 5). In particular, this seems to be a useful tool for treating so-called *partial differential algebraic* equations.

A really surprising benefit of the new approach concerns numerical integration methods. At this point it should be mentioned that, e. g. in the early paper [18], the formulation (1.5) was used with the seemingly technical condition on im $D(t)$ to be time-invariant. Now we understand that this condition is exactly the one allowing the DAE-decoupling into its essential parts and the discretization of the DAE to commute in the index-1 case (cf. [13]). In the index-2 case, if im $D(t)$ is decomposed into certain two further constant subspaces, the decoupling and the discretization also commute in some sense. Hence, neither phenomena converting e. g. the implicit Euler method within the inherent regular ODE into the explicit one ([11], [10]) nor additional stepsize restrictions may arise. Since this is primarily a property of the DAE itself, we call those equations *numerically well formulated* (cf. Section 3, [13], [14]). Obviously, now one should try to have numerically well formulated equations at the very beginning. If this is not the case, refactorizations $AD = \tilde{A}\tilde{D}$ (or factorizations $E = AD$ for (1.1)) leading to numerically well formulated versions might be useful (cf. Examples 3.1, 3.2, 3.3 below). In particular, the Euler backward method applied to the famous index-2 example given in [20] is well-known to get into great difficulties and to fail. However, in a slight modification (Example 3.1) it works best, and we understand why this happens.

By the way, the results on the contractivity of standard methods applied to (1.1) given (e. g. in [8], [10]) for the case of a constant nullspace $\ker E$, and those of modified methods (e. g. [12]) for the case of a constant im E are now completely clear.

This paper ist organized as follows.
In Section 2 the basic decoupling is described in the nessesary details. With this, we follow the lines of [2], but now the index-1 and index-2 cases are put together such that index-1 appers to be a special case. Furthermore, the matrix functions A and D may be rectangular ones now, i. e., not necessarily quadratic. In Section 3 we deal with the BDF as a prototype of a discretization (the same can be done, with some more effort, with Runge-Kutta methods, cf. [13], [14]), and show the advatages of numerically well formulated DAEs. A first attempt to formulate and to apply abstract differential algebraic equations is presented in Section 4. A linear-quadratic control problem for a DAE (in abstract formulation) is proved to be solvable in Section 5. Thereby, a Hamiltonian property turns out to be a further benefit of the numerically well formulated case.

2 Linear DAEs with properly stated leading term

Consider equations

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{2.1}$$

with matrix coefficients
$A \in C(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^m)),\ D \in C(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^n)),\ B \in C(\mathcal{I}, L(\mathbb{R}^m))$.
The coefficients A and D that determine the leading term of equation (2.1) are assumed to be well matched in the following sense (cf.[2]). Condition $C1$ for the case that $n = m$).

Definition 2.1 *The ordered pair of continuous matrix functions A and D is said to be* **well matched** *if*

$$\text{im } D(t) \oplus \ker A(t) = \mathbb{R}^n, t \in \mathcal{I}, \tag{2.2}$$

and these subspaces are spanned by continuously differentiable bases.
The **leading term is properly stated** *if A and D are well matched.*

Obviously, if A and D are well matched, there is a uniquely determined projector function $R(t) \in L(\mathbb{R}^n),\ R(t)^2 = R(t),\ \text{im } R(t) = \text{im } D(t)$, $\ker R(t) = \ker A(t)$ and $R \in C^1((\mathcal{I}, L(\mathbb{R}^n))$.

Remark 2.1 *The decomposition (2.2) implies that, for all $t \in \mathcal{I}$,*

$$\text{im } A(t)D(t) = \text{im } A(t),\ \ker A(t)D(t) = \ker D(t),\ \ker A(t) \cap \text{ im } D(t) = 0, \tag{2.3}$$

and vice versa. Roughly speaking, D plays the role of an incidence matrix. It figures out the really involved derivatives of the unknown function.

Remark 2.2 *Starting with a standard form equation*

$$E(t)x'(t) + F(t)x(t) = q(t),\ t \in \mathcal{I}, \tag{2.4}$$

we may use any factorization $E = AD$ with well matched A and D, $D \in C^1(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^n))$, to obtain

$$A(t)(D(t)x(t))' + (F(t) - A(t)D'(t))x(t) = q(t),\ t \in \mathcal{I}. \tag{2.5}$$

In particular, the factorizations $E = EP_E$ or $E = R_E E$ can be used if $P_E(t)$ is a projector along the nullspace of $E(t)$, but $R_E(t)$ denotes a projector onto the image of $E(t)$.

Naturally, a solution of equation (2.1) should be a function $x \in C(\mathcal{I}, \mathbb{R}^m)$ that has a continuously differentiable product Dx, and satisfies the equation at all $t \in \mathcal{I}$. Let

$$C^1_D(\mathcal{I}, R^m) := \{x \in C(\mathcal{I}, \mathbb{R}^m) : Dx \in C^1(\mathcal{I}, \mathbb{R}^n)\}$$

denote the respective function space.

Next we introduce certain subspaces and matrices to be used throughout this paper:

$$\left.\begin{array}{l} G_0(t) := A(t)D(t), \quad B_0(t) := B(t), \\ \text{for } i = 0, 1, \\ N_i(t) := \ker G_i(t), \\ Q_i(t) \in L(\mathbb{R}^m),\ Q_i(t)^2 = Q_i(t), \operatorname{im} Q_i(t) = N_i(t), \\ P_i(t) := I - Q_i(t), \\ W_i(t) \in L(\mathbb{R}^m),\ W_i(t)^2 = W_i(t),\ \ker W_i(t) = \operatorname{im} G_i(t), \\ S_i(t) = \ker W_i(t)B_i(t) = \{z \in \mathbb{R}^m : B_i(t)z \in \operatorname{im} G_i(t)\}, \\ G_{i+1}(t) := G_i(t) + B_i(t)Q_i(t), \quad B_{i+1}(t) := B_i(t)P_i(t). \end{array}\right\} \tag{2.6}$$

Further, $D(t)^-$ denotes the reflexive generalized inverse of $D(t)$ with

$$D(t)D(t)^- = R(t) \quad , \quad D(t)^-D(t) = P_0(t). \tag{2.7}$$

Note that $D(t)^-$ is uniquely determined if $P_0(t)$ is given and vice versa.

Definition 2.2 *(cf. [2]): The DAE (2.1) with properly formulated leading term and nontrivial $N_0(t)$ has* **index** $\mu \in \{1, 2\}$ *if $N_j(t) \cap S_j(t)$ has constant dimension ν_{j+1} on $\mathcal{I}$ and $\nu_j > 0$ for $j < \mu$, $\nu_\mu = 0$.*

Remark 2.3 *In terms of the matrices $G_j(t)$, the index μ case is characterized by*

$$\mathit{rank}\ G_{j+1}(t) =: r_{j+1} = m - \nu_{j+1}, \quad r_j < m \text{ for } j < \mu,\ r_\mu = m.$$

Namely, using the auxiliary matrix

$$\mathcal{G}_{j+1} := G_{j+1}(I - G_j^- B_j Q_j)$$

and the reflexive generalized inverse G_j^- determined by
$G_jG_j^- = I - W_j$, $G_j^-G_j = P_j$, we derive

$$N_{j+1}(t) = (I - G_j(t)^- B_j(t)Q_j(t))(N_j(t) \cap S_j(t)), \tag{2.8}$$

but $I - G_j(t)^- B_j(t)Q_j(t)$ always remains nonsingular.

Remark 2.4 *In an analogous way, considering a somehow more complex contruction of the matrices B_j we could continue the chain (2.6) for $i > 1$ and define also an index higher than two. However, we dispense with that here in favour of more transparency, and since in the sections below we fully focus on the lower index case.*

Turn, for a moment, to the well understood case of Hessenberg size-two DAEs

$$\left.\begin{array}{rlcl} x_1'(t)+ & B_{11}(t)x_1(t)+ \; B_{12}(t)x_2(t) & = & q_1(t) \\ & B_{21}(t)x_1(t) & = & q_2(t) \end{array}\right\}, \tag{2.9}$$

$B_{21}(t)B_{12}(t) \in L(\mathbb{R}^{n_2})$ nonsingular.
Letting $m = n_1 + n_2, n = n_1$,

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; D = [I \;\; 0], \; D^- = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; R = I$$

we may rewrite (2.9) as a DAE (2.1) with properly stated leading term. Then we have

$$N_0 = O_{n_1} \times \mathbb{R}^{n_2}, \; S_0(t) = \ker B_{21}(t) \times \mathbb{R}^{n_2},$$

$$N_0 \cap S_0(t) = N_0, \; \dim N_0 \cap S_0(t) = n_2,$$

$$Q_0 = \begin{pmatrix} 0 & \\ & I \end{pmatrix}, G_1(t) = \begin{bmatrix} I & B_{12}(t) \\ 0 & 0 \end{bmatrix}, \quad Q_1(t) = \begin{bmatrix} H(t) & 0 \\ -F(t) & 0 \end{bmatrix},$$

$$F := (B_{21}B_{12})^{-1}B_{21}, \quad H := B_{12}F.$$

Furthermore $S_1(t) = S_0(t), \quad N_1(t) \cap S_1(t) = 0$.
$H(t) \in L(\mathbb{R}^{n_1})$ is the projection onto $\operatorname{im} B_{12}(t)$ along $\ker B_{21}(t)$. In our context,

$$DP_1D^- = I - H, \quad DQ_1D^- = H$$

are projectors that realize the decomposition

$$DS_1(t) \oplus DN_1(t) = \operatorname{im} D = \mathbb{R}^{n_1} \tag{2.10}$$

(we have $\ker A = 0$ here). The second equation in (2.9) gives simply

$$H(t)x_1(t) = B_{12}(t)(B_{21}(t)B_{12}(t))^{-1}q_2(t), \tag{2.11}$$

while the first one leads to

$$x_2(t) = F(t)(q_1(t) - x_1'(t) - B_{11}(t)x_1(t)) \tag{2.12}$$

and

$$(I - H(t))x_1' + (I - H(t))B_{11}(t)x_1(t) = (I - H(t))q_1(t). \tag{2.13}$$

In order to extract a regular explicit ODE with respect to the unknown component $u := DP_1x = (I-H)x_1$ from (2.13) one has to differentiate $I-H$ and to replace $(I-H)x_1' = u' - (I-H)'x_1$. Then as an inherent regular ODE, the equation

$$\begin{aligned} u' - (I-H)'u \quad &+ (I-H)B_{11}u = (I-H)q_1 \\ &+ \{(I-H)' - (I-H)B_{11}\}B_{12}(B_{21}B_{12})^{-1}q_2 \end{aligned} \tag{2.14}$$

results. The solutions of (2.9) are given by the formula

$$\begin{aligned} x &= D^-u + D^-DQ_1x + Q_0x \\ &= \begin{pmatrix} I \\ 0 \end{pmatrix} u + \begin{pmatrix} Hx_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_2 \end{pmatrix}. \end{aligned}$$

Note that there is no need to differentiate the second equation of (2.9) or any part of the coefficients – except for the projection H and the term $B_{12}(B_{21}B_{12})^{-1}q_2 = DQ_1G_2^{-1}q$. Thus we obtain solvability for all

$$q \in C(\mathcal{I}, \mathbb{R}^m) \text{ with } DQ_1G_2^{-1}q \in C^1(\mathcal{I}, \mathbb{R}^m)$$

supposed H belongs to the class C^1.

In the following, we will derive an anlogous result for general equations (2.1), where H and $I-H$ correspond to DQ_1D^- and DP_1D^-.

Lemma 2.1 *Given an index-2 DAE (2.1).*
Let $Q_1(t)$ denote the projector onto $N_1(t)$ along $S_1(t), t \in \mathcal{I}$.
Then, for all $t \in \mathcal{I}$, the decomposition

$$D(t)S_1(t) \oplus D(t)N_1(t) \oplus \ker A(t) = \mathbb{R}^n \tag{2.15}$$

holds true, and

$$D(t)P_1(t)D(t)^-, \; D(t)Q_1(t)D(t)^-, \quad I - R(t) \tag{2.16}$$

are the uniquely determined projectors that realize this decomposition.
If, additionally, $D(t)S_1(t)$ and $D(t)N_1(t)$ are spanned by continuously differentiable on $\mathcal{I}$ functions, then

$$DP_1D^-, \; DQ_1D^- \in C^1(\mathcal{I}, L(\mathbb{R}^n)).$$

Proof.
This assertion can be verified in the same way as [2], Lemma 2.5 and Lemma 2.6 were proved for the case of $m = n$.

□

In contrast to the index-1 case, im D is once more smoothly decomposed into the further subspaces DS_1 and DN_1 for index-2 DAEs. Let us mention, that if we think of index-3, we have to decompose DS_1 again and so on.

Formally, we put the index-1 and index-2 cases together for brevity, as we have trivially

$$Q_1(t) = 0, \quad P_1(t) = I, \quad G_2(t) = G_1(t), \quad t \in \mathcal{I},$$

for index-1 DAEs. Hence, $G_2(t)$ is nonsingular on $\mathcal{I}$ for index-1 and index-2 DAEs, i. e.,

$$N_1(t) \oplus S_1(t) = \mathbb{R}^m.$$

In the following, $Q_1(t)$ always denotes the projector onto $N_1(t)$ along $S_1(t)$. Recall from [8], Appendix A the respresentation

$$Q_1(t) = Q_1(t)G_2(t)^{-1}B(t)P_0(t). \tag{2.17}$$

Relation (2.17) leads to the properties

$$Q_1Q_0 = 0, \quad (P_0P_1)^2 = P_0P_1, \quad (P_0Q_1)^2 = P_0Q_1. \tag{2.18}$$

Further, due to the chain construction we compute

$$\begin{aligned} G_2^{-1}AD &= G_2^{-1}G_2P_1P_0 = P_1P_0, \\ G_2^{-1}B &= G_2^{-1}B(P_0P_1 + P_0Q_1 + Q_0) = G_2^{-1}BP_0P_1 + Q_1 + Q_0. \end{aligned}$$

Consequently, scaling (2.1) by G_2^{-1} leads to

$$P_1P_0D^-(Dx)' + G_2^{-1}BD^-DP_1x + Q_1x + Q_0x = G_2^{-1}q. \tag{2.19}$$

Multiplying (2.19) by DP_1, DQ_1 and Q_0P_1 we decouple this equation into the following three ones:

$$\begin{aligned} DP_1D^-(Dx)' + DP_1G_2^{-1}BD^-DP_1x &= DP_1G_2^{-1}q, & (2.20)\\ DQ_1x &= DQ_1G_2^{-1}q, & (2.21)\\ -Q_0Q_1D^-(Dx)' + Q_0P_1G_2^{-1}BD^-DP_1x + Q_0x &= Q_0P_1G_2^{-1}q\,. & (2.22) \end{aligned}$$

Now, given any solution $x \in C_D^1(\mathcal{I}, \mathbb{R}^m)$, we express

$$\begin{aligned} DP_1D^-(Dx)' &= (DP_1x)' - (DP_1D^-)'Dx, \\ Q_0Q_1D^-(Dx)' &= Q_0Q_1P_0P_1D^-(Dx)' = Q_0Q_1D^-DQ_1D^-(Dx)' \\ &= Q_0Q_1D^-(DQ_1x)' - Q_0Q_1D^-(DQ_1D^-)'DP_1x\,. \end{aligned}$$

By this, we find the solution representation

$$\begin{aligned} x &= D^-Dx + Q_0x = D^-DP_1x + D^-DQ_1x + Q_0x \\ &= KD^-u + D^-DQ_1G_2^{-1}q + Q_0P_1G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1}q)', \end{aligned} \tag{2.23}$$

where $u = DP_1x$ satisfies the inherent regular ODE

$$u' - (DP_1D^-)'(u + DQ_1G_2^{-1}q) + DP_1G_2^{-1}BD^-u = DP_1G_2^{-1}q \tag{2.24}$$

and $K := I - Q_0Q_1D^-(DQ_1D^-)'D - Q_0P_1G_2^{-1}BP_0$ is a nonsingular matrix function.
Clearly, this generalizes the well-known representation described above for Hessenberg size-two forms (2.9). In the index-1 case (2.23), (2.24) simplify to

$$x = KD^-u + Q_0G_1^{-1}q, \quad K = I - Q_0G_1^{-1}B, \tag{2.25}$$

$$u' - R'u + DG_1^{-1}BD^-u = DG_1^{-1}q. \tag{2.26}$$

Of course, the regular explicit ODE (2.24) can be formed without assuming the existence of a solution of the DAE just by giving the coefficients A, D, B and the right-hand side q.

Definition 2.3 *Given a DAE (2.1) of index $\mu \in \{1, 2\}$. The explicit ODE (2.24) is said to be the* **inherent regular ODE** *of the DAE.*

Lemma 2.2 *Given an index-μ DAE (2.1), $\mu \in \{1, 2\}$. Let DP_1D^-, DQ_1D^- be continuously differentiable.*

(i) Then, the subspaces $D(t)S_1(t)$ and $D(t)N_1(t)$ as well as the inherent regular ODE are uniquely determined by the problem data.

(ii) $D(t)S_1(t)$ is a time-varying invariant subspace of the inherent regular ODE, i. e., if a solution belongs to this subspace at a certain point, it runs within this subspace all the time.

(iii) If $D(t)S_1(t)$ and $D(t)N_1(t)$ do not vary with time t, then, solving the IVP for (2.24) with the initial condition $u(t_*) \in D(t_*)S_1(t_*)$ yields the same solution as solving this IVP for

$$u' + DP_1G_2^{-1}BD^-u = DP_1G_2^{-1}q\,.$$

Proof.

(i) (cf. [2]) When constructing the ODE (2.24) the only arbitrariness is the choice of the projector P_0 and the corresponding D^-. So we start with two different P_0 and $\tilde{P}_0$ and the corresponding D^- and $\tilde{D}^-$. Recall that then $DD^- = R = D\tilde{D}^-, D^-D = P_0, \tilde{D}^-D = \tilde{P}_0$. Compute $\tilde{G}_1 = G_0 + B\tilde{Q}_0 = G_0 + BQ_0\tilde{Q}_0 = G_1(P_0 + \tilde{Q}_0)$,
$(P_0 + \tilde{Q}_0)(\tilde{P}_0 + Q_0) = I,\ \tilde{N}_1 = (\tilde{P}_0 + Q_0)N_1,\ \tilde{S}_1 = S_1,\ \tilde{Q}_1 := (\tilde{P}_0 + Q_0)Q_1(P_0 + \tilde{Q}_0) = (\tilde{P}_0 + Q_0)Q_1, \operatorname{im}\tilde{Q}_1 = \tilde{N}_1, \ker\tilde{Q}_1 = S_1,\ D\tilde{Q}_1\tilde{D}^- = DQ_1\tilde{D}^- = DQ_1\tilde{D}^-D\tilde{D}^- = DQ_1\tilde{D}^-DD^- = DQ_1D^-,\ D\tilde{P}_1\tilde{D}^- = D\tilde{D}^- - D\tilde{Q}_1\tilde{D}^- = R - DQ_1D^- = DP_1D^-$.

Consequently, since the projectors DP_1D^-, DQ_1D^- do not depend on the choice of P_0, their images do not so, either.

Further, due to
$\tilde{G}_2 = G_2(P_0+\tilde{Q}_0+Q_0\tilde{P}_0Q_1)$, $(P_0+\tilde{Q}_0+Q_0\tilde{P}_0Q_1)(\tilde{P}_0+Q_0-Q_0\tilde{P}_0Q_1) = I$, we have
$D\tilde{G}_2^{-1} = D(\tilde{P}_0 + Q_0 - Q_0\tilde{P}_0Q_1)G_2^{-1} = DG_2^{-1}$,
$D\tilde{P}_0\tilde{G}_2^{-1} = D\tilde{P}_1\tilde{D}^-D\tilde{G}_2^{-1} = DP_1D^-DG_2^{-1} = DP_1G_2^{-1}$,
$D\tilde{P}_1\tilde{G}_2^{-1}B\tilde{D}^- = D\tilde{P}_1\tilde{G}_2^{-1}B\tilde{D}^-D\tilde{D}^- = D\tilde{P}_1\tilde{G}_2^{-1}B\tilde{D}^-DD^- =$
$D\tilde{P}_1\tilde{G}_2^{-1}B\tilde{P}_0D^- = D\tilde{P}_1G_2^{-1}BD^- = DP_1G_2^{-1}BD^-$.

(ii) Given any solution $\tilde{u} \in C^1(\mathcal{I}, \mathbb{R}^m)$ of (2.24) with $\tilde{u}(t_0) \in D(t_0)S_1(t_0)$ for some $t_0 \in \mathcal{I}$. Insert $\tilde{u}$ into (2.24) and multiply the resulting identity by $(I-DP_1D^-)$. This gives $(I-DP_1D^-)\tilde{u}'-(I-DP_1D^-)(DP_1D^-)'\tilde{u} = 0$, hence, for $\tilde{v} := (I - DP_1D^-)\tilde{u}$, $\tilde{v}' - (I - DP_1D^-)'\tilde{v} = 0, \tilde{v}(t_0) = 0$. $\tilde{v}(t)$ vanishes identically as the solution of a homogeneous regular initial value problem.

(iii) (cf. [13]) If $D(t)S_1(t)$ and $D(t)N_1(t)$ are constant, we may use constant projectors $U, V \in L(\mathbb{R}^n)$ onto these subspaces. Then it holds that
$(DP_1D^-)'U = (DP_1D^-U)' = (U)' = 0$,
$(DP_1D^-)'V = (DP_1D^-V)' = 0, (DQ_1D^-)'U = (DQ_1D^-U)' = 0$,
$(DQ_1D^-)'V = (DQ_1D^-V)' = (V)' = 0$.
In the consequence the term $(DP_1D^-)'(u + DQ_1G_2^{-1}q)$ disappears.

□

In comparison with different notions used in the standard DAE theory (underlying ODE, essentially underlying ODE, inherent regular ODE), where certain transforms resp. projectors are not completely determined by the problem data, the decoupling described above is fully given by the problem. This is a real advantage. If the nullspace of $A(t)$ is trivial, i. e., im $D(t)$ has dimension n as it is the case for the Hessenberg size-two DAE (2.9), we can also speak of an inherent state space system (in minimal coordinates), supposed $D(t)S_1(t)$ does not vary with t. Then, the inherent regular ODE (2.24) is essentially nothing else but an explicit ODE located in the constant state space DS_1. However, in the index-2 case $D(t)S_1(t)$ and $D(t)N_1(t)$ (cf. $H(t)$ in (2.9)) vary with time in general.

Theorem 2.3 *Given a DAE (2.1) with index μ, $\mu \in \{1, 2\}$, and $DP_1D^-, DQ_1D^- \in C^1(\mathcal{I}, L(\mathbb{R}^m))$.*

(i) For any

$$q \in C^1_{DQ_1G_2^{-1}}(\mathcal{I}, \mathbb{R}^m) = \{w \in C(\mathcal{I}, \mathbb{R}^m) : DQ_1G_2^{-1}w \in C^1(\mathcal{I}, \mathbb{R}^m)\},$$

equation (2.1) is solvable on $C^1_D(\mathcal{I}, \mathbb{R}^m)$.

(ii) The initial value problem for (2.1) with the initial condition

$$D(t_0)P_1(t_0)(x(t_0)-x^0) = 0 \quad , \quad x^0 \in \mathbb{R}^m, \tag{2.27}$$

and $t_0 \in \mathcal{I}$, $q \in C^1_{DQ_1G_2^{-1}}(\mathcal{I}, \mathbb{R}^m)$ *has a unique solution* $x \in C^1_D(\mathcal{I}, \mathbb{R}^m)$.

(iii) Supposed $\mathcal{I}$ *is compact, the inequality*

$$\|x\|_\infty + \|(Dx)'\|_\infty \le \mathcal{K}(|D(t_0)P_1(t_0)x^0| + \|q\|_\infty + \|(DQ_1G_2^{-1}q)'\|_\infty)$$

becomes true, i. e., the DAE has perturbation index μ.

(iv) The subspace $S_{\mathrm{ind}\mu}(t) := K(t)D(t)^- D(t)S_1(t) = \operatorname{im} K(t)P_0(t)P_1(t)$ *is the geometric solution space of the homogeneous equation (2.1) with* $q = 0$, *i. e.,* $S_{\mathrm{ind}\mu}(t)$ *is filled by solutions and all solutions run within this subspace* $S_{\mathrm{ind}\mu}(t) = S_0(t)$.

Proof. (cf. [2]):

(i) is a simple consequence of (ii).

(ii) Solve the regular ODE (2.24) with the initial condition $u(t_0) = D(t_0)P_1(t_0)x^0 \in D(t_0)S_1(t_0)$. Using the solution $u \in C(\mathcal{I}, \mathbb{R}^n)$ we construct the function $x \in C(\mathcal{I}, \mathbb{R}^m)$ according to formula (2.23). Obviously, $Dx = DD^-u + DQ_1G_2^{-1}q = u + DQ_1G_2^{-1}q$ is continuously differentiable, thus $x \in C^1_D(\mathcal{I}, \mathbb{R}^m)$. Then we check that x satisfies (2.1) indeed.
The homogeneous IVP ($q = 0$, $D(t_0)P_1(t_0)x^0 = 0$) has only the trivial solution.

(iii) and (iv) are due to the construction. □

Remark 2.5 *One could relax the constant rank conditions in the Definitions 2.1 and 2.2 to characterize a class of DAEs (2.1) that have index* μ, $\mu \le 2$, *but undergo certain nondangerous index changes. If the interval* $\mathcal{I}$ *consists of a number of subintervals where all constant rank conditions are satisfied, we realize the decoupling on all these subintervals simulaneously and ask then for a continuously differentiable extension of* DP_1D^- *on the whole interval. If it exists, we may take advantage of (2.23), (2.24) again. However, we have to consider that some terms may have discontinuities now.*

Example 2.1 *The system* $\alpha x_2' + x_1 = q_1$, $x_2 = q_2$ *has the solution* $x_1 = q_1 - \alpha q_2'$, $x_2 = q_2$ *independently whether the coefficient function* $\alpha \in C(\mathcal{I}, \mathbb{R})$ *vanishes. If we put this system in the form (2.1), we choose* $B(t) = I$, *and further, if* $\alpha(t)$ *does not vanish,*

$$A = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}, \; D = (0\ 1), \; P_0 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \; D^- = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; R = 1,$$

otherwise

$$A = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ D = (0 \ 0), \ P_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \ D^- = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ R = 0\,.$$

Then, on subintervals with $\alpha(t) \neq 0$, we have

$G_1 = \begin{pmatrix} 1 & \alpha \\ 0 & 0 \end{pmatrix}, \qquad G_2 = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$*, i. e., the index is two there, further*

$G_2^{-1} = \begin{pmatrix} 1 & -\alpha \\ 0 & 1 \end{pmatrix}, \ P_1 = \begin{pmatrix} 1 & \alpha \\ 0 & 0 \end{pmatrix}, \ Q_1 = \begin{pmatrix} 0 & -\alpha \\ 0 & 1 \end{pmatrix},$

$DP_1D^- = 0$, $DQ_1D^- = 1$, $DP_1G_2^{-1} = 0$, thus $u = DP_1x = 0$. On subintervals and points where $\alpha(t)$ is zero we have $G_1 = B = I$. Trivially the problem has index 1 on those intervals. Then $G_2 = G_1$, $P_1 = I$, $DP_1D^- = 0$, $DP_1G_2^{-1} = 0$, thus $u = DP_1x = 0$.
In both cases we have $DP_1D^- = 0$ and (cf. (2.23)

$$\begin{aligned} x &= D^-DQ_1G_2^{-1}q + Q_0P_1G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1})', \\ x &= (P_0Q_1 + Q_0P_1)G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1}q)' \\ &= \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} q + \begin{pmatrix} -\alpha \\ 0 \end{pmatrix} q_2' = q - \begin{pmatrix} \alpha q_2' \\ 0 \end{pmatrix}. \end{aligned}$$

Remark 2.6 *In [2] we aimed at a theory to treat both an original DAE and its adjoint equation in the same way . The adjoint equation of*

$$A(Dx)' + Bx = 0 \tag{2.28}$$

is now

$$D^*(A^*y)' - B^*y = 0. \tag{2.29}$$

Both equations have properly stated leading terms at the same time. Moreover, they have index μ, $\mu \in \{1,2\}$, simultaneously. Nice relations between the fundamental solution matrices are shown in [2].

Remark 2.7 *The index notion given by Definition 2.2 resp. Remark 2.3 is approved for creating a practical index monitor ([6]).*

3 Numerical integration methods and numerically well formulated DAEs

Given a DAE (2.1) with a properly formulated leading term and having index μ, $\mu \in \{1,2\}$. The natural modification of the k-step BDF applied to (2.1) is

$$A(t_n)\frac{1}{h}\sum_{j=0}^{k} \alpha_j D(t_{n-j})x_{n-j} + B(t_n)x_n = q(t_n) \tag{3.1}$$

to provide an approximation x_n to $x(t_n)$ at the current step $t_n = t_{n-1} + h$. Let us use shorter denotations $A_n := A(t_n)$ and so on for all given coefficient functions and resulting projectors.
Denote the numerical derivative by

$$[Dx]'_n := \frac{1}{h}\sum_{j=0}^{k} \alpha_j D(t_{n-j})x_{n-j}. \tag{3.2}$$

With (3.1) resp.

$$A_n[Dx]'_n + B_n x_n = q_n \tag{3.3}$$

we proceed in the same way as we did, in Section 2, with the DAE (2.1) itself, i. e. we scale by G_{2n}^{-1} and then use the decoupling projectors. The resulting system analogous to system (2.20)-(2.22) is, with $u_n := D_n P_{1n} x_n$, $v_n := D_n Q_{1n} x_n$, $w_n = Q_{0n} x_n$,

$$D_n P_{1n} D_n^-[Dx]'_n \quad + \quad D_n P_{1n} G_{2n}^{-1} B_n D_n^- u_n = D_n P_{1n} G_{2n}^{-1} q_n, \tag{3.4}$$

$$v_n = D_n Q_{1n} G_{2n}^{-1} q_n, \tag{3.5}$$

$$-Q_{0n} Q_{1n} D_n^-[Dx]'_n \quad + \quad Q_{0n} P_{1n} G_{2n}^{-1} B_n D_n^- u_n + w_n = Q_{0n} P_{1n} G_{2n}^{-1} q_n. \tag{3.6}$$

Denoting

$$[u]'_n := \frac{1}{h}\sum_{j=0}^{k} \alpha_j u_{n-j} \quad , \quad [v]'_n := \frac{1}{h}\sum_{j=0}^{k} \alpha_j v_{n-j} \ .$$

we derive

$$\begin{aligned} D_n P_{1n} D_n^-[Dx]'_n \quad &= \quad [u]'_n - \tfrac{1}{h}\sum_{j=0}^{k} \alpha_j \\ &\qquad (D_{n-j} P_{1n-j} D_{n-j}^- - D_n P_{1n} D_n^-) D_{n-j} x_{n-j} \\ &=: \quad [u]'_n - \sum_{j=0}^{k} \mathcal{A}_{nj} D_{n-j} x_{n-j}, \\ \text{where} \qquad \mathcal{A}_{nj} \quad &:= \quad -j\alpha_j \int_0^1 (DP_1D^-)'(t_n - sjh)ds, \end{aligned} \tag{3.7}$$

further

$$Q_{0n} Q_{1n} D_n^-[Dx]'_n \quad = \quad Q_{0n} Q_{1n} D_n^- D_n Q_{1n} D_n^-[Dx]'_n,$$

$$D_n Q_{1n} D_n^-[Dx]'_n \quad = \quad [v]'_n - \sum_{j=0}^{k} \mathcal{B}_{nj} D_{n-j} x_{n-j},$$

where $\mathcal{B}_{nj} := -j\alpha_j \int_0^1 (DQ_1D^-)'(t_n - sjh)ds$, so that

$$Q_{0n} Q_{1n} D_n^-[Dx]'_n = Q_{0n} Q_{1n} D_n^- \left\{ [v]'_n - \sum_{j=0}^{k} \mathcal{B}_{nj} D_{n-j} x_{n-j} \right\}. \tag{3.8}$$

Insert (3.7) into (3.4) to obtain the recursion formula for u_n:

$$[u]'_n - \sum_{j=0}^{k} \mathcal{A}_{nj}(u_{n-j} + D_{n-j}Q_{1n-j}G_{2n-j}^{-1}q_{n-j}) \quad +D_nP_{1n}G_{2n}^{-1}B_nD_n^-u_n \\ = D_nP_{1n}G_{2n}^{-1}q_n \tag{3.9}$$

Then, by (3.5), (3.6), (3.8), we compute the representation of x_n in the form

$$x_n = \quad \tilde{K}_nD_n^-u_n - Q_0Q_1D_n^- \sum_{j=0}^{k} \mathcal{B}_{nj}(u_{n-j} + D_{n-j}Q_{1n-j}G_{2n-j}^{-1}q_{n-j}) \\ +Q_{0n}Q_{1n}D_n^- \left[DQ_1G_2^{-1}q\right]'_n + (Q_{0n}P_{1n} + P_{0n}Q_{1n})G_{2n}^{-1}q, \tag{3.10}$$

$$\text{with} \quad \left[DQ_1G_2^{-1}q\right]'_n := \tfrac{1}{h}\sum_{j=0}^{k} \alpha_jD_{n-j}Q_{1n-j}G_{2n-j}^{-1}q_{n-j},$$

$$\tilde{K}(t) := K(t){+}Q_0(t)Q_1(t)D(t)^-(DQ_1D^-)'(t)D(t).$$

On the other hand, if we decoupled the DAE (2.1) first, applied the same BDF to the inherent regular ODE (2.24) and then used formula (2.23) to compute x_n, we would obtain

$$[u]'_n \quad -(DP_1D^-)'_n(u_n + D_nQ_{1n}G_{2n}^{-1}q_n) \\ +D_nP_{1n}G_{2n}^{-1}B_nD_n^-u_n = D_nP_1G_{2n}^{-1}q_n, \tag{3.11}$$

$$x_n = \quad \tilde{K}_nD_n^-u_n - Q_{0n}Q_{1n}D_n^-(DQ_1D^-)'_n(u_n + D_nQ_{1n}G_{2n}^{-1}q_n) \\ +Q_{0n}Q_{1n}D_n^-(DQ_1G_2^{-1}q)'_n + (Q_{0n}P_{1n} + P_{0n}Q_{1n})G_{2n}^{-1}q_n\,. \tag{3.12}$$

Comparing (3.11), (3.12) and (3.9), (3.10) we observe that, except for the terms

$$[DQ_1G_2^{-1}q]'_n \text{ resp. } (DQ_1G_2^{-1}q)'_n := (DQ_1G_2^{-1}q)'(t_n),$$

these formulas coincide, supposed the subspaces $D(t)S_1(t), D(t)N_1(t)$ do not vary with t (cf. Lemma 2.2 above). Namely, in (3.11) and (3.12), we have then

$$(DP_1D^-)'_n(u_n + D_nQ_{1n}G_{2n}^{-1}q_n) = (DP_1D^-)'_n(Uu_n + VD_nQ_{1n}G_{2n}^{-1}q_n \\ = (DP_1D^-U)'_nu_n + (DP_1D^-V)'_nD_nQ_{1n}G_{2n}^{-1}q_n = 0, \\ (DQ_1D^-)'_n(u_n + D_nQ_{1n}G_{2n}^{-1}q_n = 0\ ,$$

but in (3.9), (3.10) we use $u_{n-j} = Uu_{n-j}, v_{n-j} = Vv_{n-j}$

$$\mathcal{B}_{nj}U = -j\alpha_j \int_0^1 (DQ_1D^-U)'(t_n - sjh)ds = 0,$$

and, analogously, $\mathcal{A}_{nj}U = 0,\ \ \mathcal{A}_{nj}V = 0,\ \ \mathcal{B}_{nj}V = 0.$

Definition 3.1 *We will say that a discretization method and the DAE-decoupling commute, if for the approximation x_n obtained by applying the method to the original DAE, and the further approximation $\hat{x}_n$ obtained by applying the method to the inherent regular ODE and then using formula (2.23), it holds that*

$$x_n - \hat{x}_n = Q_{0n} Q_{1n} D_n^- \left\{ [DQ_1 G_2^{-1} q]'_n - (DQ_1 G_2^{-1} q)'(t_n) \right\} \tag{3.13}$$

and $[DQ_1 G_2^{-1} q]'_n$ denotes a numerical approximation of $(DQ_1 G_2^{-1} q)'(t_n)$.

Let us stress that in the index-1 case, commutativity means simply $x_n = \hat{x}_n$ because of $Q_1 = 0$.

Theorem 3.1 *Given an index-μ DAE (2.1), $\mu \in \{1, 2\}$, which has constant characteristic subspaces DS_1 and DN_1. Then the BDF-discretization and the decoupling commute.*

Remark 3.1 *Recall once more that one has $Q_1 = 0$ for the index-1 case. Trivially, DS_1 and DN_1 are constant iff* $\mathrm{im}\, D = DS_1$ *is so. Since $DS_1 \oplus DN_1 = \mathrm{im}\, D$, for index-2 problems,* $\mathrm{im}\, D$ *being constant is a necessary condition for DS_1 and DN_1 to be constant.*

Remark 3.2 *Here, we have chosen the BDF only for brevity. Of course, similar results can be proved for Runge-Kutta methods. Also, nonlinear DAEs may be considered ([14], [15]).*

Remark 3.3 *In particular, Theorem 3.1 generalizes the respective results obtained for DAEs with constant leading nullspace (cf. (1.3) with $P_E' = 0$) e. g. in [8] for $\mu = 1$ and in [10] for $\mu = 2$. On the other hand, Theorem 3.1 covers the situation studied in [12], i. e., an equation (1.1) rewritten as $(Ex)' + (F - E')x = q$, where* $\mathrm{im}\, E$ *is time-invariant.*

Remark 3.4 *It is a remarkable benefit of commutativity that we can now prove nice assertions on contractive or dissipative flows and the respective discretized versions ([14]). Having modified the respective notions (contractivity and dissipativity inequalities, adsorbing sets etc.) in such a way that the standard ODE notions apply to the inherent regular ODE on its invariant subspace DS_1, we may use the well-known techniques and results approved in the regular ODE case.*
In particular ([14]), a stiffly accurate and algebraically stable Runge-Kutta method applied to a contractive index-1 DAE yields

$$|D_n x_n - D_n \bar{x}_n| \le |D_{n-1} x_{n-1} - D_n \bar{x}_{n-1}|,$$
$$|x_n - \bar{x}_n| \le \gamma |D_n x_n - D_n \bar{x}_n|, \ n \ge 1,$$

for each two sequences of approximations starting with consistent $x_0, \bar{x}_0$. Thereby, γ is a bound of the canonical version of D^- corresponding to the (canonical) projector onto S_0 along N_0.
It is also shown in [14] that the Euler backward method reflects the dissipativity behaviour properly without any stepsize restriction.

Example 3.1 *Rewrite the equation (cf. [20])*

$$E(t)x'(t)+F(t)x(t) = q(t) \tag{3.14}$$

with

$$E(t) = \begin{pmatrix} 0 & 0 \\ 1 & \eta t \end{pmatrix}, \quad F(t) = \begin{pmatrix} 1 & \eta t \\ 0 & 1+\eta \end{pmatrix}, \quad q(t) = \begin{pmatrix} g(t) \\ 0 \end{pmatrix},$$

as

$$R_E(t)(E(t)x(t))'+(F(t)-R_E(t)E'(t))x(t) = q(t), \tag{3.15}$$

or as

$$E(t)(P_E(t)x(t))'+(F(t)-E(t)P_E'(t))x(t) = q(t), \tag{3.16}$$

where

$$R_E(t) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad P_E(t) = \begin{pmatrix} 1 & \eta t \\ 0 & 0 \end{pmatrix}.$$

Both reformulated versions (3.15) and (3.16) have properly formulated leading terms, $R_E E = E P_E = E$. *Compute* $Q_0 = \begin{pmatrix} 0 & -\eta t \\ 0 & 1 \end{pmatrix}$,
$S_0(t) = \{z \in \mathbb{R}^2 : F(t)z \in \mathrm{im}\ E(t)\} = \{z : z_1 + \eta t z_2 = 0\}$, $S_0(t) \cap N_0(t) = N_0(t)$,
$S_1(t) = \{z : F(t)P_0(t)z \in \mathrm{im}\ (E(t) + F(t)Q_0(t))\} = \{z : z_1 + \eta t z_2 = 0\}$,
$N_1(t) = \{z : z_1 + (1+\eta t)z_2 = 0\} \quad N_1(t) \cap S_1(t) = 0$,
$P_E(t)S_1(t) = 0, \quad P_E(t)N_1(t) = \{w \in \mathbb{R}^2 : w_2 = 0\}$ *is constant,*
$E(t)S_1(t) = 0, \quad E(t)N_1(t) = \{w \in \mathbb{R}^2 : w_1 = 0\}$ *is constant.*

Now, we observe that we have index-2 DAEs with constant subspaces as supposed in Theorem 3.1. Thus, for both versions (3.15) and (3.16), the BDF and the decoupling commute.
In both cases, the Euler backward yields simply

$$x_{n,1} = -\eta t_n x_{n,2} + g_n, \ x_{n,2} = -\frac{1}{h}(g_n - g_{n-1}).$$

Recall that the Euler backward directly applied to (3.14) gets into great difficulties for $|\eta| > |1+\eta|$ *(e. g. [4], [20]).*

□

Example 3.2 *The standard form DAE* $E(t)x'(t) + F(t)x(t) = 0$ *with*

$$E(t) = \begin{pmatrix} \delta - 1 & \delta t \\ 0 & 0 \end{pmatrix}, \; F(t) = \sigma \begin{pmatrix} \delta - 1 & \delta t \\ \delta - 1 & \delta t - 1 \end{pmatrix} \; x(t) = 0$$

has index-1 on $\mathcal{I} = \mathbb{R}$ *for arbitrary values* $\sigma \neq 0, \; \delta \neq 1$. *The solution is*

$$x_1(t) = \frac{1 - \delta t}{\delta - 1} \, x_2(t) \; , x_2(t) = e^{(\delta - \sigma)t} x_{2,0}.$$

For $\delta < \sigma$, *one would expect the backward Euler method to generate a sequence* $\{x_{n,2}\}_n$ *such that* $x_{n,2} \longrightarrow 0 (n \to \infty)$ *without any stepsize restriction. However, the formula*

$$E_n \frac{1}{h}(x_n - x_{n-1}) + F_n x_n = 0$$

reads in detail

$$x_{n,1} = \frac{1}{\delta - 1}(1 - \delta t_n) x_{n,2} \; , \quad x_{n,2} = \frac{1 + \delta h}{1 - \sigma h} x_{n-1,2}.$$

While for the constant coefficient case $\delta = 0$ *it works fine, for* $\delta \neq 0$, *the numerical solution may explode close to* $\sigma h \approx -1$. *To realize* $|1 + \delta h| < |1 - \sigma h|$ *one has to accept strong extra stepsize restriction.*
On the other hand, if we apply the Euler method to the reformulation

$$R_E(Ex)' + (F - R_E E')x = 0,$$

$$R_E = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \; \textit{we arrive at}$$

$$x_{n,1} = \frac{1}{\delta - 1}(1 - \delta t_n) x_{n,2}, \quad x_{n,2} = \frac{1}{1 - (\delta - \sigma)h} x_{n-1,2}$$

and things work well. No stepsize restrictions are caused by stability.

□

Example 3.3 *Consider the seemingly harmless Hessenberg index-2 DAE*

$$Ex'(t) + F(t)x(t) = 0$$

with

$$E = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 0 \end{pmatrix}, \quad F(t) = \begin{pmatrix} \lambda & -1 & -1 \\ \eta t(1 - \eta t) - \eta & \lambda & -\eta t \\ 1 - \eta t & 1 & 0 \end{pmatrix}.$$

The leading term may be rewritten in a proper form by choosing (cf. (2.9))

$$\tilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \tilde{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \tilde{A}\tilde{D} = E.$$

With $\tilde{B} := F$, we have type (2.1).

The subspaces

$$N_1(t) = \{x \in \mathbb{R}^3 : x_1 - x_3 = 0\,, \quad x_2 - \eta t x_3 = 0\},$$
$$S_1(t) = \{x \in \mathbb{R}^3 : x_2 - (\eta t - 1)x_1 = 0\}$$

but also

$$\begin{aligned} \tilde{D}N_1(t) &= \{z \in \mathbb{R}^2 : z_2 - (\eta t - 1)z_1 = 0\}, \\ \tilde{D}S_1(t) &= \{z \in \mathbb{R}^2 : z_2 - \eta t z_1\} \end{aligned}$$

move with time. We refer to [10] for the sobering effect of the numerical tests. Recall that this problem was introduced in [10] to demonstrate that just a Hessenberg form DAE may get into diffeculties with extra stepsize restrictions.
Observe, on the other hand, that there are more possibilities in factorizing E. If we find a further well-matched factorization $E = AD$ such that $DN_1(t), DS_1(t)$ appear to be time-invariant, we are done since Theorem 3.1 applies. Such a much more comfortable factorization is given by $\tilde{A}T(t)^{-1}T(t)\tilde{D}$,

$$A(t) = \begin{bmatrix} 1 & 0 \\ \eta t - 1 & 1 \\ 0 & 0 \end{bmatrix}, \; D(t) = \begin{bmatrix} 1 & 0 & 0 \\ 1 - \eta t & 1 & 0 \end{bmatrix}, \; T(t) \begin{bmatrix} 1 & 0 & 0 \\ 1 - \eta t & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The reformulated DAE is $A(Dx)' + Bx = 0$ with $B := E - AT'\tilde{D}$, and the subspaces

$$DS_1 = \{z \in \mathbb{R}^2 : z_2 = 0\}, \; DN_1 = \{z \in \mathbb{R}^2 : z_1 = z_2\}$$

are constant such that Theorem 3.1 applies in fact. □

As we could see above, the problem whether discretization and decoupling commute depends primarily on how the DAE is formulated.

Definition 3.2 *A DAE (2.1) with index 1 is said to be* **numerically well formulated** *if* im $D(t)$ *is time-invariant (cf. [13]). A DAE (2.1) with index-2 is said to be* **numerically well formulated** *if $D(t)S_1(t)$ and $D(t)N_1)$ are constant.*

This new possibility to rearrange subspaces for better numerical properties is a surprise. One should think further on how to exploit this idea best.

Fortunately, e. g. in circuit simulation, the relevant subspaces could be shown to be constant for a large class of problems ([7], [5]), hence there is no need for reformulations.
Special questions concerning necessary reformulations will be discussed in [14], [15].

4 Abstract DAEs

Nowadays, e. g. in circuit simulation and simulation of multibody dynamics, there is a remarkable interest in complex systems that consist of coupled systems of DAEs and PDEs (cf. [21], [9]). There are also certain proposals for treating so-called PDAEs (= partial differential algebraic equations, e. g. [17]). Thereby, one of the questions to be considered is how to formulate initial and boundary conditions in an appropriate way.
In the following we try to show the usefulness of the decoupling procedure given in Section 2 for the finite dimensional case, now in an abstract modification. Being able to understand those abstract DAEs, a more systematical constructive and numerical treatment can be started on.

In this section we deal with abstract DAEs

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{4.1}$$

where $A(t) : Z \to Y,\ B(t) : X \to Y,\ D(t) : X \to Z$ are linear operators acting in the real Hilbert spaces X, Y, Z.
For all $t \in \mathcal{I}$, let $A(t)$ and $D(t)$ be bounded and normally solvable. Moreover, let $A(.)$ and $D(.)$ depend continuously (in the norm sense) on t. As bounded maps, $A(t)$ and $D(t)$ have nullspaces $\ker A(t)$ and $\ker D(t)$, respectively, which are closed linear manifolds, i. e., subspaces. Due to the normal solvability, $\mathrm{im}\ A(t)$ and $\mathrm{im}\ D(t)$ are subspaces (e. g. [1]), too.
For all $t \in \mathcal{I}$, let the operator $B(t)$ be defined on a dense subset $\mathcal{D}_B \subseteq X$, and let $B(.)x \in C(\mathcal{I}, Y)$ for all $x \in \mathcal{D}_B$.

A continuous path $x \in C(\mathcal{I}, X)$ with $x(t) \in \mathcal{D}_B,\ t \in \mathcal{I}$, is called a solution of (4.1) if $Dx \in C^1(\mathcal{I}, Z)$ and equation (4.1) is satisfied pointwise.

Definition 4.1 *The* **leading term** *of (4.1)* **is properly stated** *if the operators $A(t)$ and $D(t), t \in \mathcal{I}$, are* **well matched** *in the following sense:*

(i) $\mathrm{im}\ A(t)D(t) = \mathrm{im}\ A(t), \ker A(t)D(t) = \ker D(t)$,
$\ker A(t) \oplus \mathrm{im}\ D(t) = Z,\ t \in \mathcal{I}$.

(ii) The projector $R(t) \in L_b(Z)$ that realizes this decomposition of Z (i. e., $R(t)^2 = R(t),\ \mathrm{im}\ D(t) = \mathrm{im}\ R(t),\ \ker R(t) = \ker A(t)$) depends continuously differentiably on t.

Remark 4.1 *Since $R(t)$ depends smoothly on t, the subspaces* $\mathrm{im}\ D(t)$ *are isomorphic for different values of t. For the same reason, the nullspaces* $\ker A(t)$ *are also isomorphic.*

Let $L_b(X)$ denote the linear space of linear bounded maps on X etc. For bounded maps, we alway assume their definition region to be the respective whole space.
Let clM denote the closure of the set M in the respective space.

Next we generalize the matrix- and subspace-chain given in (2.6) by introducing the following further linear maps, linear manifolds and subspaces for all $t \in \mathcal{I}$:

$G_0(t) = A(t)D(t), \quad N_0(t) = \ker G_0(t) = cl \ \ker G_0(t),$
$Q_0(t) \in L_b(X), \quad Q_0(t)^2 = Q_0(t), \ \operatorname{im} Q_0(t) = N_0(t),$
$W_0(t) \in L_b(Y), \quad W_0(t)^2 = W_0(t), \quad \ker W_0(t) = \operatorname{im} G_0(t) = cl \ \operatorname{im} G_0(t),$
$P_0(t) := I - Q_0(t),$
$S_0(t) = \ker W_0(t)B(t),$
$G_1(t) = G_0(t) + B(t)Q_0(t), \quad N_1(t) = cl \ \ker G_1(t),$
$Q_1(t) \in L_b(X), \quad Q_1(t)^2 = Q_1(t), \operatorname{im} Q_1(t) = N_1(t),$
$W_1(t) \in L_b(Y), \quad W_1(t)^2 = W_1(t), \quad \ker W_1(t) = cl \operatorname{im} G_1(t),$
$P_1(t) := I - Q_1(t),$
$S_1(t) := \ker(W_1(t)B(t)P_0(t)),$
$G_2(t) := G_1(t) + B(t)P_0(t)Q_1(t).$

By construction, the operators $G_1(t)$ and $G_2(t)$ are at least densely defined in X. In our context, operator products (in particular those with projectors) are often defined on a larger region by trivial reasons. We will always use the maximal trivial extensions.

Let $D(t)^- : Z \to X$ denote the reflexive generalized inverse of $D(t)$ such that $D(t)D(t)^- = R(t), \ D(t)^- D(t) = P_0(t), \ t \in \mathcal{I}.$

Definition 4.2 *Equation (4.1) with a properly stated leading term is said to be*

(i) *an* **abstract index-1 DAE** *if, for all* $t \in \mathcal{I}, \ \dim(\operatorname{im} W_0(t)) > 0$ *and* $G_1(t)$ *is injective and densely solvable,*

(ii) *an* **abstract index-2 DAE** *if, for all* $t \in \mathcal{I}$, $\dim(\operatorname{im} W_i(t)) > 0, \ i = 0, 1, \ W_1(t)$ *depends continuously on* t*, and* $G_2(t)$ *is injective and densely solvable.*

If also $B(t) : X \to Y$ is a bounded map that is continuous with respect to t in norm sense, then the assertions in Section 2, in particular Theorem 2.3, can be modified to hold true for the abstract DAE in a straightforward way. The more challenging problem is an only densely defined $B(t)$. Hence, we are going to study three different cases of this type.

Case 1: A coupled system of a PDE and Fredholm integral equations (on a special application of this type I was kindly informed by Hermann Brunner). Given a linear Fredholm integral operator $F : L_2(\Omega)^s \to L_2(\Omega)^s$, $\|F\| < 1$, a linear differential operator $K : \dot{H}^1(\Omega) \to L_2(\Omega), Kw := -\Delta w + cw$ for

$w \in \dot{H}^1(\Omega)$, with $c \geq 0$, and linear bounded coupling operators $L : L_2(\Omega)^s \to L_2(\Omega)$, $\quad E : L_2(\Omega) \to L_2(\Omega)^s$. The system to be considered is

$$\left.\begin{array}{rcl} x_1'(t) + Kx_1(t) + L\, x_2(t) &=& q_1(t), \\ E\, x_1(t) + (I-F)x_2(t) &=& q_2(t),\ t \in [0,1] \end{array}\right\}. \tag{4.2}$$

Using the corresponding matrix representations for A, D, B, with $X := L_2(\Omega)^s \times L_2(\Omega)$, $Y := X$, $Z := L_2(\Omega)$, we rewrite (4.2) in the form (4.1). Namely, we have

$$A = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ D = (1\ 0),\ P_0 = AD = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ D^- \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ R = 1$$

and $B = \begin{bmatrix} K & L \\ E & I-F \end{bmatrix}$. By construction, B is defined on $\mathcal{D}_B := \dot{H}^1 \times L_2(\Omega)^s$.
Clearly, it holds that $N_0 = 0 \times L_2(\Omega)^s$,

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},\ W_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},\ W_0 B = \begin{bmatrix} 0 & L \\ 0 & I-F \end{bmatrix},\ G_1 = \begin{bmatrix} 1 & L \\ 0 & I-F \end{bmatrix}.$$

$W_0 B$ and G_1 are defined on X (as trivial extensions of bounded maps).
G_1 is a bijection such that this abstract DAE has index 1.
With $G_1^{-1} = \begin{bmatrix} 1 & -L(I-F)^{-1} \\ 0 & (I-F^1 \end{bmatrix}$ we find $DG_1^{-1}BD^- = K - L(I-F)^{-1}E$ defined on $\dot{H}^1$. Each solution of the DAE is given by the expression

$$\begin{array}{rcl} x(t) &=& D^- u(t) + Q_0 x(t), \\ Q_0 x(t) &=& Q_0 G_1^{-1} q(t) - Q_0 G_1^{-1} B D^- u(t), \end{array}$$

where $u(t)$ is a solution of the abstract regular differential equation

$$u'(t) + DG_1^{-1}BD^- u(t) = DG_1^{-1} q(t). \tag{4.3}$$

Obviously, one has to state an appropriate initial condition for (4.3), i. e., $u(0) = D(0)x(0) = D(0)x^0$, $x^0 \in \mathcal{D}_B$.

Case 2: A special linear constant coefficient PDAE discussed in [17].
Consider the PDAE

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{bmatrix} w_t(y,t) + \begin{bmatrix} -1 & 0 & 0 \\ 0 & -b & 0 \\ 0 & 0 & 0 \end{bmatrix} w_{yy}(y,t) + \begin{bmatrix} 0 & c_1 & 0 \\ 0 & 0 & c_2 \\ 0 & c_3 & 0 \end{bmatrix} w(y,t) = q(y,t), \tag{4.4}$$

$y \in \Omega := (-1,1), \quad t \in \mathcal{I} = (0,T]$.

Suppose $a, b > 0,\ c_i \neq 0,\ i = 1,2,3$.
How should we formulate boundary and initial conditions?

Rewrite (4.4) in abstract form with $x(t) := w(\cdot, t)$.
Choose $X = L_2(\Omega) \times L_2(\Omega) \times L_2(\Omega),\ Y = X,\ Z = L_2(\Omega) \times L_2(\Omega)$ and use again matrix representaions for our coefficients, i. e.,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & a \\ 0 & 0 \end{bmatrix},\ D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ AD = \begin{bmatrix} 1 & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{bmatrix},\ P_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad Bx = \begin{bmatrix} -\Delta & c_1 & 0 \\ 0 & -b\Delta & c_2 \\ 0 & c_3 & 0 \end{bmatrix}.$$

Δ is the Laplacian. Let us start with

$$\mathcal{D}_B := C^2(\Omega) \times C^2(\Omega) \times L_2(\Omega). \tag{4.5}$$

The operator

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & a & c_2 \\ 0 & 0 & 0 \end{bmatrix}$$

is defined on X and bounded.
Obviously, $\ker G_1 = N_1 = \{x \in X : x_1 = 0,\ ax_2 + c_2 x_3 = 0\}$ is a nontrivial closed subspaces and im $G_1 = L_2(\Omega) \times L_2(\Omega) \times 0$ is closed.
With $W_1 = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix}, \quad W_1 B P_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & c_3 & 0 \end{bmatrix}$ is defined on $X, S_1 = \{x \in X : x_2 = 0\} = \ker W_1 B P_0$ is a subspace, and obviously $S_1 \cap N_1 = 0$.
$Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{a}{c_3} & 0 \end{bmatrix}$ represents a projector onto N_1 along S_1, thus $X = S_1 = S_1 \oplus N_1$. The map

$$G_2 = \begin{bmatrix} 1 & c_1 & 0 \\ 0 & a - b\Delta & c_2 \\ 0 & c_3 & 0 \end{bmatrix}$$

is defined on $L_2(\Omega) \times C^2(\Omega) \times L_2(\Omega)$, further
im $G_2 = L_2(\Omega) \times L_2(\Omega) \times C^2(\Omega), cl$im $G_2 = Y$, i. e., G_2 is densely solvable. The injectivity of G_2 is immediately checked, therefore, this DAE has index 2.
The inherent regular differential equation is now

$$u'(t) + DP_1 G_2^{-1} B D^- u(t) = DP_1 G_2^{-1} q(t) \tag{4.6}$$

with $u(t) = DP_1x(t) = \begin{pmatrix} x_1(t) \\ 0 \end{pmatrix}$ and

$$G_2^{-1} = \begin{bmatrix} 1 & 0 & -\frac{c_1}{c_3} \\ 0 & 0 & \frac{1}{c_3} \\ 0 & \frac{1}{c_2} & -\frac{1}{c_2c_3}(a-b\Delta) \end{bmatrix}, \text{ defined on } L_2(\Omega)\times L_2(\Omega)\times C^2(\Omega)\ ,$$

$$DP_1G_2^{-1}BD^- = \begin{bmatrix} -\Delta & 0 \\ 0 & 0 \end{bmatrix} \text{ defined on } C^2(\Omega)\times L_2(\Omega),$$

i. e., (4.6) is in fact nothing else but

$$x_1'(t) - \Delta x_1(t) = q_1(t) - \frac{c_1}{c_3}\, q_3(t)\ . \tag{4.7}$$

Now it becomes clear that, for (4.6), the initial condition $DP_1(x(0) - x^0) = 0,\ x^0 \in \mathcal{D}_B$, as well as boundary conditions should be given. We take homogeneous Dirichlet conditions and put them into the definition region of B, i. e., we restart our procedure with

$$\mathcal{D}_B := \dot{C}^2(\Omega)\times C^2(\Omega)\times L_2(\Omega)$$

instead of (4.5) or, in spite of more general solvability, with

$$\mathcal{D}_B := \dot{H}^1(\Omega)\times H^1(\Omega)\times L_2(\Omega). \tag{4.8}$$

Using (4.8), G_2 is defined on $L_2(\Omega)\times H^1(\Omega)\times L_2(\Omega)$, but G_2^{-1} on $L_2(\Omega)\times L_2(\Omega)\times H^1(\Omega)$. Consequently, admissible right-hand sides are $q(t)$ with $q_1(t), q_2(t) \in L_2(\Omega),\ q_3(t) \in H^1(\Omega)$.

Computing

$$Q_0P_1G_2^{-1}BD^- = 0,\ Q_0Q_1D^- = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & -\frac{a}{c_2} \end{bmatrix},$$

$$DQ_0P_1G_2^{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{c_2} & \frac{b}{c_2c_3}\Delta \end{bmatrix}, \quad Q_1G_2^{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{c_3} \end{bmatrix}$$

we find the solution representation (cf. (2.23)

$$\begin{aligned} x(t) &= D^-u(t) + (P_0Q_1G_2^{-1} + Q_0P_1G_2^{-1})q(t) \\ &\quad + Q_0Q_1D^-(DQ_1G_2^{-1}q(t))' \\ &= \begin{pmatrix} x_1(t) \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{c_3} \\ 0 & \frac{1}{c_2} & \frac{b}{c_2c_3}\Delta \end{pmatrix} q(t) + \begin{pmatrix} 0 \\ 0 \\ -\frac{a}{c_2c_3}q_3'(t) \end{pmatrix} \end{aligned}$$

i. e., $x_2(t) = \frac{1}{c_3}q_3(t), x_3(t) = \frac{1}{c_2}q_2(t) + \frac{b}{c_2c_3}\Delta q(t) - \frac{a}{c_2c_3}q_3'(t)$,

while $x_1(t)$ solves (4.7). No further initial or boundary conditions should be given. With (4.8) we obtain the unique solvability of the IVP $A(Dx(t))' + Bx(t) = q(t)$,
$DP_1(x(0) - x^0) = 0$ for $x^0 \in \mathcal{D}_B$ and

$$q \in C(\mathcal{I}, \mathcal{D}_{G_2^{-1}} \subseteq Y), DQ_1G_2^{-1} \in C^1(\mathcal{I}, Y).$$

Case 3: A PDAE and a DAE coupled by a restriction operator.
Consider the system

$$\begin{aligned} \tilde{u}_t(y,t) - \;& \tilde{u}_{yy}(y,t) + c\tilde{u}(y,t) = f(y,t)\ , \quad y \in \Omega, t \geq 0, \\ & \tilde{u}|_{\partial\Omega} = 0, \end{aligned} \tag{4.9}$$

$$\tilde{A}(t)(\tilde{D}(t)\tilde{x}(t))' + \tilde{B}(t)\tilde{x}(t) + \text{ Restrik}(t)\tilde{u}(\cdot,t) = g(t),\ t \geq 0. \tag{4.10}$$

Assume the linear restriction map $\text{Restrik}(t) : H^1(\Omega) \to \mathbb{R}^m$ to be bounded and to depend continuously on t.

Rewrite the system (4.9), (4.10) as an abstract DAE for $x(t) = \begin{pmatrix} \tilde{u}(\cdot,t) \\ \tilde{x}(t) \end{pmatrix}$ and choose $X = Y = L_2(\Omega) \times \mathbb{R}^m,\ Z = L_2(\Omega) \times \mathbb{R}^n$,

$$A := \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A} \end{bmatrix},\ D = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{D} \end{bmatrix},\ AD = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A}\tilde{D} \end{bmatrix},\ Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{Q}_0 \end{bmatrix},$$

$$D^- = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{D}^- \end{bmatrix},\ R = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{R} \end{bmatrix},\ B = \begin{bmatrix} -\Delta + c & 0 \\ \text{Restrik} & \tilde{B} \end{bmatrix},$$

where $\mathcal{D}_B := \dot{H}^1(\Omega) \times \mathbb{R}^m$.

$G_1 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{G}_1 \end{bmatrix}$ is defined on X,

$N_1 = \ker G_1 = O \times \tilde{N}_1,\ \text{im } G_1 = L_2(\Omega) \times \text{im } \tilde{G}_1$,

$S_1 = \{x \in X : \text{Restrik } x_1 + \tilde{B}\tilde{P}x_2 \in \text{im } \tilde{G}_1\}$

Supposing that the operator $\text{Restrik}(t)$ maps into $\text{im}\tilde{G}_1(t)$ it holds that

$S_1 = \{x \in X : \tilde{B}\tilde{P}_0x_2 \in \text{im } \tilde{G}_1\} = \dot{H}^1(\Omega) \times \tilde{S}_1$. Then,

$Q_1 = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{Q}_1 \end{bmatrix}$ is the projector onto N_1 along S_1, $G_2 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{G}_2 \end{bmatrix}$.

In the general case the projector Q_1 onto N_1 along S_1 is more difficult to construct. Obviously, we have $N_1 \cap S_1 = O \times (\tilde{N}_1 \cap \tilde{S}_1)$.
G_2 is a bijection if $\tilde{G}_2$ is so.

Consequently, the coupled system (4.9) (4.10) interpreted as an abstract DAE has the same index as the DAE (4.10).

5 Linear-quadratic control problems

Now, the quadratic cost functional

$$\begin{aligned} J(u) := &\ \langle x(T), Vx(T)\rangle \\ &+\tfrac{1}{2}\int_0^T \{\langle x(t), W(t)x(t)\rangle + 2\langle x(t), S(t)u(t)\rangle + \langle u(t), R(t)u(t)\rangle\}dt \end{aligned} \tag{5.1}$$

is to be minimized on solutions of the DAE

$$A(t)(D(t)x(t))' = C(t)x(t)+B(t)u(t),\ t \in [t_0, T], \tag{5.2}$$

which satisfy the initial condition

$$D(0)(x(0)-x^0) = 0\,. \tag{5.3}$$

Here, $T > 0$ is fixed, $x^0 \in X$.
X, Y, Z, U denote real Hilbert spaces,
$A(t) \in L_b(Z, Y), D(t) \in L_b(X, Z),\ C(t) \in L_b(X, Y),\ B(t) \in L_b(U, Y)$,
$W(t) \in L_b(X),\ R(t) \in L_b(U),\ S(t) \in L_b(U, X)$ for all $t \in \mathcal{I} := [0, T]$.
Let all these operators depend continuously in the norm sense on t.
Assume further:

(i) $W(t) = W(t)^*,\ R(t) = R(t)^*,\ V = V^*,$ im $V \subseteq$ im $D(T)^*$.

(ii) V and the maps $\begin{bmatrix} W(t) & S(t) \\ S(t)^* & R(t) \end{bmatrix}$, $t \in \mathcal{I}$, are positive semidefinite.

(iii) $A(t)$ and $D(t)$ are normally solvable and well matched for $t \in \mathcal{I}$.

Admissible controls are those functions $u \in C(\mathcal{I}, U)$ for which a solution $x \in C^1_D(\mathcal{I}, X)$ of the IVP (5.2), (5.3) exists.

Consider the boundary value problem

$$\left.\begin{aligned} A(Dx)' &= Cx + Bu \\ D^*(A^*\Psi)' &= -C^*\Psi + Wx + Su \\ 0 &= S^*x - B^*\Psi + Ru \end{aligned}\right\}, \tag{5.4}$$

$$D(0)(x(0)-x^0) = 0, \quad D(T)^*A(T)^*\Psi(T) = Vx(T) \quad . \tag{5.5}$$

The system (5.4) is a DAE with a properly statet leading term. $D(t)^*$ and $A(t)^*$ are normally solvable and well matched at the same time as $A(t)$ and $D(t)$ are so.

Theorem 5.1 *If the triple $x_* \in C^1_D(\mathcal{I},X), \Psi_* \in C^1_{A^*}(\mathcal{I},Y), u_* \in C(\mathcal{I},U)$ solves the BVP (5.4), (5.5), then u_* is an optimal control for the problem (5.1)-(5.3).*

Proof. By straightforward calculations we prove that the expression

$$\begin{aligned}\mathcal{A} := \;& \langle x(T)-x_*(T), Vx_*(T)\rangle + \int_0^T \{\langle x(t)-x_*(t), Wx_*(t)\rangle + \\ &+\langle x(t)-x_*(t), S(t)u_*(t)\rangle + \langle u(t)-u_*(t), S^*(t)x_*(t)\rangle \\ &+\langle u(t)-u_*(t), R(t)u_*(t)\rangle\}dt\end{aligned}$$

vanishes. Then the assertion follows immediately from

$$\begin{aligned}&J(u)-J(u_* = \mathcal{A} + \tfrac{1}{2}\langle x(T)-x_*(T), V(x(T)-x_*(T))\rangle \\ &+\tfrac{1}{2}\int_0^T \left\langle \begin{pmatrix} x(t)-x_*(t) \\ u(t)-u_*(t)\end{pmatrix}, \begin{pmatrix} W(t) & S(t) \\ S(t)^* & R(t)\end{pmatrix} \begin{pmatrix} x(t)-x_*(t) \\ u(t)-u_*(t)\end{pmatrix}\right\rangle dt\end{aligned}$$

□

In Theorem 5.1, no assumption on the index of the DAE (5.2) is made. However, for obtaining a theoretically and practically solvable BVP (5.4), (5.5), additional conditions ensuring the index-1 property of (5.4) should be satisfied.

Let Q_0, W_0 denote the orthoprojectors onto $\ker AD$ and along $\operatorname{im} AD$, respectively (cf. Section 4).
Denote

$$\mathcal{F} := \begin{bmatrix} W_0CQ_0 & 0 & W_0B \\ Q_0^*WQ_0 & -Q_0^*C^*W_0^* & Q_0^*S \\ -S^*Q_0 & B^*W_0^* & -R \end{bmatrix}. \tag{5.6}$$

Lemma 5.2 *The DAE (5.4) has index 1 iff, for $t \in \mathcal{I}$,*
$\mathcal{F}(t) : \ker D(t) \times \ker A(t)^* \times U \to \ker A(t)^* \times \ker D(t) \times U$ is bijective.

Proof. Put $\tilde{X} = X \times Y \times U, \quad \tilde{Y} = Y \times X \times U, \quad \tilde{Z} = Z \times Z.$
The respective maps for (5.4) rewritten as $\tilde{A}(\tilde{D}\tilde{X})' + \tilde{B}\tilde{x} = 0$, are

$$\tilde{G}_0 = \begin{bmatrix} AD & 0 & 0 \\ 0 & D^*A^* & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tilde{Q}_0 = \begin{bmatrix} Q_0 & 0 & 0 \\ 0 & W_0^* & 0 \\ 0 & 0 & I \end{bmatrix}, \tilde{B} = -\begin{bmatrix} C & 0 & B \\ W & -C^* & S \\ -S^* & B^* & -R \end{bmatrix},$$

$$\tilde{W}_0 = \begin{bmatrix} W_0 & 0 & 0 \\ 0 & Q_0^* & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{\mathcal{G}}_1 = \tilde{G}_0 + \tilde{W}_0\tilde{B}\tilde{Q}_0, \quad \tilde{G}_1 = \tilde{G}_0 + \tilde{B}\tilde{Q}_0.$$

$\tilde{\mathcal{G}}_1(t) : \tilde{X} \to \tilde{Y}$ is a bijection iff $\mathcal{F}(t) = -\tilde{W}_0(t)\tilde{B}(t)\tilde{Q}_0(t)$ acts bijectively on $\ker D(t) \times \ker A(t)^* \times U$ onto $\ker A(t)^* \times \ker D(t) \times U$.

Further, $\tilde{G}_1(t)$ is bijective at the same time as $\mathcal{G}_1(t)$ is so (cf. Remark 2.3).

□

Remark 5.1 *In [16], a linear-quadratic control problem of a similar form is studied. Instead of the DAE (5.2) the equation*

$$(A(t)x(t))' = C(t)x(t) + B(t)u(t)$$

is considered. Because of the somehow incomplete leading term, the resulting system corresponding to (5.4) looks less fine and transparent.

Remark 5.2 *Since the operator $\mathcal{F}$ has the same structure as its counterpart in [16], all sufficient conditions for bijectivity proved in [16] hold true also here.*

Lemma 5.3 *Let, for $t \in \mathcal{I}$,*

$$\mathcal{F}(t) : \ker(D(t) \times \ker A(t)^* \times U(t) \to \ker A(t)^* \times \ker D(t) \times U(t) \tag{5.7}$$

be bijective. Then the inherent regular differential equation of (5.4) is of the form

$$\begin{pmatrix} Dx \\ A^*\Psi \end{pmatrix}' = \begin{pmatrix} E_1 + R'R & E_2 \\ E_3 & -E_1^* + R^{*\prime} R^* \end{pmatrix} \begin{pmatrix} Dx \\ A^*\Psi \end{pmatrix}, \tag{5.8}$$

where $E_2 = E_2^*$ $E_3 = E_3^*$ and E_2, E_3 are positive semidefinite.

Proof.
This assertion can be proved following the lines of [16], Theorem 1.

□

Obviously, (5.8) is a non-negative Hamiltonian system iff $R'R = 0$, $RR' = 0$, i. e., $R' = 0$, but this is exactly the case if (5.4) is numerically well formulated. For Hamiltonian systems, we know the respective BVPs to be solvable.

Theorem 5.4 *Let $\mathcal{F}(t)$ in (5.7) be bijective, $R'(t) = 0$, $t \in \mathcal{I}$. Then, there is at least one optimal control for the problem (5.1), (5.2), (5.3).*
If Z is finite dimensional, there is exactly one optimal control of (5.1), (5.2), (5.3).

Proof.
It remains to show uniqueness in case of finite dimensional Z, but this can be done in the same way as [16], Theorem 4, is proved.

□

References

[1] N. J. Achieser and J. M. Glasmann: Theorie der linearen Operatoren im Hilbertraum. Akademie-Verlag Berlin 1981.

[2] K. Balla and R. März: A unified approach to linear differential algebraic equations and their adjoint equations. Preprint 2000-18, Humboldt-Universität Berlin, Institut für Mathematik.

[3] K. Balla, R. März: Linear differential algebraic equations of index 1 and their adjoint equations. Results in Mathematics, 37 (2000), 13–35.

[4] K. E. Brenan, S. L. Campbell, L. R.Petzold: Numerical solution of initial value problems in differential algebraic equations. North Holland, Amsterdam, 1989.

[5] D. Estévez Schwarz: Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation. Dissertation, Humboldt-Universität Berlin, Institut für Mathematik, 2000.

[6] D. Estévez Schwarz, U. Feldmann, R. März, S. Sturtzel and C. Tischendorf: Finding beneficial DAE structures in circuit simulation. Preprint 2000-7, Humboldt-Universität Berlin, Institut für Mathematik.

[7] D. Estévez Schwarz and Caren Tischendorf: Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. 28 (2000) 131–162.

[8] E. Griepentrog, R. März: Differential-algebraic equations and their numerical treatment. Teubner, Leipzig, 1986.

[9] M. Günther and P. Rentrop: PDAE-Netzwerkmodelle in der elektrischen Schaltungssimulation. Preprint 99/3. Universität Karlsruhe, Institut für Wiss. Rechnen und Mathem. Modellbildung.

[10] M. Hanke, E. Izquierdo Macana and R. März: On asymptotics in case of linear index-2 differential-algebraic equations. SIAM J. Numer. Anal. 35 (1998) 1326–1346.

[11] M. Hanke and R. März: On asymptotics in case of daes. ZAMM 76 (1996) Suppl. 1, 99–102.

[12] B. Garcia–Celayeta, J. Higueras: Runge-Kutta methods for DAEs. A new approach. J. Computational and Applied Mathematics, 111(1–2) (1999), 49–61.

[13] I. Higueras and R. März: Formulating differential algebraic equations properly. Preprint 2000–20. Humboldt-Universität Berlin, Institut für Mathematik.

[14] I. Higueras, R. März and C. Tischendorf: Numerically well formulated index-1 DAEs. In preparation.

[15] I. Higueras, R. März and C. Tischendorf: Numerically well formulated index-2 DAEs. In preparation.

[16] G. Kurina and R. März: On linear-quadratic optimal control problems for time-varying descriptor systems. Preprint 2000–10. Humboldt-Universität Berlin, Institut für Mathematik.

[17] W. Lucht, K. Strehmel and C. Eichler-Liebenow: Indexes and special discretization methods for linear partial differential algebraic equations. BIT 39 No. 3 (1999), 484–512.

[18] R. März: On initial value problems in differential-algebraic equations and their numerical treatment. Computing 35 (1985) 13–37.

[19] R. März: Numerical methods for differential algebraic equations. Acta Numerica 1992, 141–198.

[20] L. Petzold: Differential/ algebraic equations are not ODEs. SIAM J. Sci. Statis. Comput. 3 (1982) 367–384.

[21] B. Simeon: Numerische Simulation gekoppelter Systeme von partiellen und differential-algebraischen Gleichungen der Mehrkörperdynamik. VDI Verlag Düsseldorf, Reihe 20, Nr. 325, 2000.