Numerically well formulated index-2 DAEs

Inmaculada Higueras[1]   Roswitha März[2]   Caren Tischendorf[2]

Abstract

For linear index-2 DAEs with properly stated leading term we characterize contractive and dissipative flows. We study under which conditions the qualitative properties of the DAE solutions are reflected by the numerical approximations. This is the case if the discretisation and the decoupling processes commute. Commutativity is achieved if two subspaces associated with the index-2 DAE are constant; in this case we say that the index-2 DAE is numerically well formulated. If both subspaces are time dependent, the problem should be reformulated; in order to avoid numerically equivalent reformulations, a criterion is given. If only one of these subspaces is constant, the problem is in some sense close to a numerically well formulated one and thus depending on the problem context, no reformulations are needed. Contractive and dissipative flows induced by the DAE are characterized and results on qualitative properties of the numerical solution are given.

Key words: differential algebraic equations, numerical integration methods, global stability, BDF, Runge-Kutta

AMS subject classification: 65L80, 65L06, 34A09

1 Introduction

In this paper we consider linear differential algebraic equations (DAEs) of the form

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{1.1}$$

where the leading terms $A \in C(\mathcal{I}, L(\mathbb{R}^\ell, \mathbb{R}^m))$, $D \in C(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^\ell))$ are well matched together and $B \in C(\mathcal{I}, L(\mathbb{R}^m))$. Here, D figures out the really involved derivatives of the unkonwn function. This formulation considered in [2, 11, 12] is often more natural from the point of view of applications than the standard formulation

$$E(t)(x(t))' + F(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{1.2}$$

used in previous papers (e.g. [7, 3, 9]). Often it is straightforward to pass from (1.2) to (1.1). For example, if $E(t)$ is continuously differentiable, we can rewrite (1.2) as

$$R_E(t)(E(t)x(t))' + (F(t) - R_E(t)E(t)')x(t) = q(t), \quad t \in \mathcal{I}, \tag{1.3}$$

[1]Universidad Pública de Navarra, Departamento de Matemática e Informática, Pamplona, Spain, higueras@unavarra.es

[2]Humboldt-University Berlin, Institute of Mathematics, Germany, iam@mathematik.hu-berlin.de, caren@mathematik.hu-berlin.de

with $R_E(t)$ a projector onto im $E(t)$.
In some applications, formulation (1.1) and not (1.2) is the one that arise in a natural way. For example, charge oriented MNA for circuits with linear capacitances and linear inductances leads to very large-dimensional problems of the form

$$A(C(t)A^T x(t))' + f(x,t) = 0,$$

where the rectangular matrix A represents the incidence matrix and $C(t)$ is a positive definite diagonal matrix (see [5, 12, 8]).
Different ODE methods have been adapted to approximate the solution of (1.1) ([11, 14]) and (1.2) (see e.g. [3]) and various convergence results are available in the literature. However, our interest in this paper is not the convergence of the numerical scheme, but somehow the unexpected and non-desirable stepsize restrictions when, for instance, A-stable methods are used. Such situations may occur if the numerical method does not integrate the regular ODE that governs the dynamics of the DAE in a lower-dimensional manifold. In this case, the implicit method actually behaves like an explicit one and stepsize restrictions by stability reasons have to be taken into account. This problem was already reported in [1] were it was observed that the implicit Euler method was transformed into an explicit method when it was applied to solve some problems. It is also well known that the implicit Euler method applied to the index-2 DAE [7]

$$\begin{pmatrix} 0 & 0 \\ 1 & \eta t \end{pmatrix} x' + \begin{pmatrix} 1 & \eta t \\ 0 & 1+\eta \end{pmatrix} x = \begin{pmatrix} g(t) \\ 0 \end{pmatrix}$$

is unstable for $\eta < -0.5$.
Numerical methods for nonlinear index-1 DAEs with well matched leading terms were studied in [11]. It was proved that if the subspace im $D(t)$ is constant, then the numerical method integrates an ODE uniquely associated with the problem, the inherent regular ODE. As a consequence, many results from the theory of numerical methods for ODEs can be transferred for these systems. For the index-1 case, if im $D(t)$ is constant, the DAE is said to be numerically well formulated. In [12] a detailed study on reformulations and the behaviour of the solution over large intervals was carried out. Recall that the results obtained cover the positive results obtained in [9] and [6] for (1.2) in the case that the subspace $\ker E(t)$ or im $E(t)$ is constant.
Asymptotic properties of solutions of index-2 DAEs of the form (1.2) with constant $\ker E(t)$ were studied in [10] and [15]. In [10] it was proved that if certain subspaces associated with the index-2 DAE, namely $S_1(t)$ and $N_1(t)$, are constant, then the ODE methods behave, from the stability point of view, in a similar way as if they were applied to an ODE. In other words, an implicit method always behaves like an implicit one and it is never transformed into an explicit one. Similar results were obtained in [15] for homogeneous DAEs under the assumption that the subspace $S_1(t)$ is constant. As the following example shows, methods may fail when these conditions do not hold.

Example 1.1 *Consider the Hessenberg index-2 DAE*

$$E(t)x'(t) + F(t)x(t) = 0 \tag{1.4}$$

with

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

where $\lambda, \eta \in \mathbb{R}$ *are constant. If* $x_0 \in \mathbb{R}^3$ *is a consistent initial value at* $t = 0$, *(i.e.,* $x_1^0 + x_2^0 = 0$, $x_3^0 + x_2^0 = 0$*), the solution of the DAE is*

$$x_1(t) = x_1^0\, e^{-\lambda t}, \qquad x_2(t) = (\eta t - 1)\, x_1(t), \qquad x_3(t) = -x_2(t).$$

The first component $x_1(t) = x_1^0 e^{-\lambda t}$ *of the exact solution is just the exponential function. For different values of* η *Figure 1 shows the first component of the numerical solutions calculated using the different integration methods.*

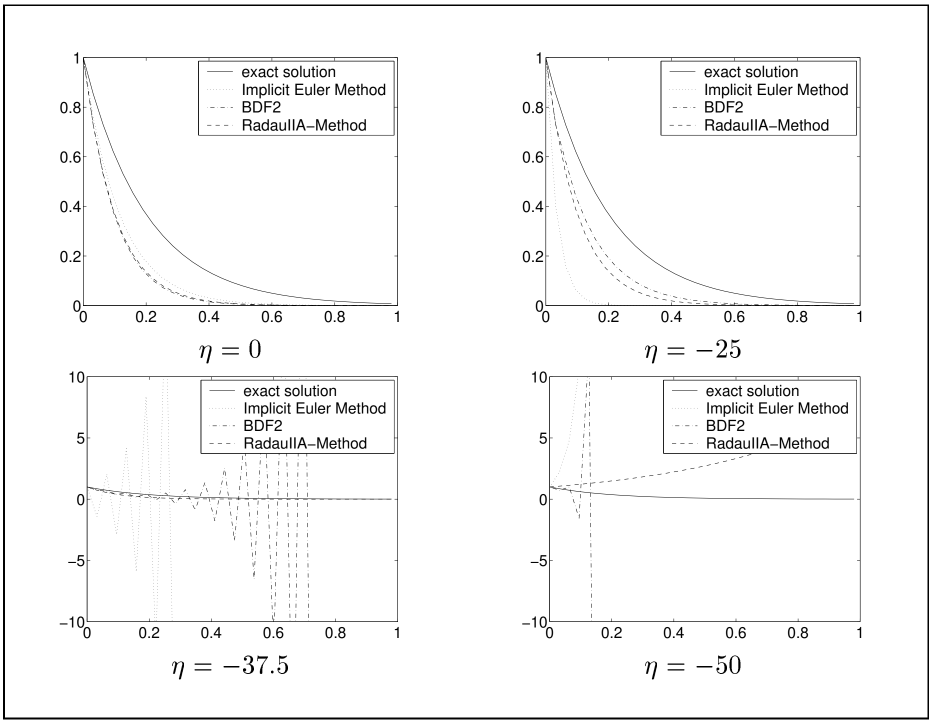

Figure 1: Numerical solutions (first component) of (1.4) for $\lambda = 10$ and $h = 10^{-1.5}$ and the consistent initial value $x^0 = (1, -1, 1)^T$.

Recall that the problem we try to analyze is not the convergence on finite intervals but stepsize restrictions because of stability reasons. As the methods

are convergent, if the stepsize were small enough, such effects would not be present.
In general, models from circuit simulation do not have both subspaces, $S_1(t)$ and $N_1(t)$, constant and thus the results in [10] are not applicable.
Numerical methods for index-2 DAEs of the form (1.1) were studied in [14]. Given (1.2), if the subspace $\ker E(t)$ is constant, then it is numerically equivalent to a DAE of the form (1.1) with $A(t) = E(t)$ and $D(t) = P_A$, where P_A denotes a constant projector onto $\ker E(t)$. Thus the results in [14] cover the previous ones. For index-1 DAEs the condition im $D(t)$ to be constant was enough to have numerically well formulated problems. Observe that as the Example 1.1 shows, some extra conditions are needed for index-2 DAEs.
Basically, what was done in [11, 12] for the index-1 case, was to study the commutativity between the discretisation and the decoupling process and give conditions so that this commutativity is achieved. This task seems to be harder for the index-2 case as long as we are confronted with the hidden constraints that contain derivatives. The best we can expect from the numerical method is that the derivatives involved in the solution are properly approximated, i.e., with the same numerical scheme we are using. Although the problem is hard, it can be tackled as this paper will show.

The paper is organized as follows. In Section 2, we review the analysis for linear index-2 DAEs and introduce the concepts and notation used in the next sections. By means of projectors, the system is decoupled and the inherent regular ODE - uniquely determined by the problem data - as well as the algebraic equations involving the hidden constrains, can be described. This decoupling is a theoretical tool for giving an insight into the DAE structure. In practice it is not explicitly available, but by means of it we are able to study the commutativity between the decoupling and the discretisation process. It turns out that when certain subspaces associated with the DAE, namely $D(t)S_1(t)$ and $D(t)N_1(t)$ are constant, commutativity is obtained, and therefore the numerical method actually integrates the inherent regular ODE. The derivatives in the hidden constraints are also approximated by the numerical method and this is the best situation we may have. For that reason, we refer to these problems as numerically well formulated ones. This is done in Section 3.
In practice we may be confronted with DAEs that are not numerically well formulated. Then one can reformulate the problem. How this can be achieved is studied in Section 4. Furthermore, we investigate under what condition the numerical solution of the original DAE is, in some sense, close to the one of a numerically well formulated DAE. This will be the case if only one of the subspaces $D(t)S_1(t)$ or $D(t)N_1(t)$ is constant and im $D(t)$ is constant. This result is quite relevant from the point of view of applications since, for example in circuit simulation, we actually obtain DAEs where, under some modelling conditions, im $D(t)$ and $D(t)N_1(t)$ are constant subspaces. In this case, in spite of not dealing with numerically well formulated DAEs, there will be no stepsize restrictions due to stability issues, and simply an error term will be present in the Q_0x component. As this error is not propagated, if it is allowable in the

application context, no further reformulations are needed.
In Section 5 we characterize contractive and dissipative flows induced by the DAE (1.1), and study how the qualitative properties of the DAE solutions are reflected by the numerical method.
The Appendix A treats convergence for DAEs with properly formulated leading terms. Appendix B collects technical details as well as auxiliary rules used in the paper.

2 Analysis of Linear DAEs

Linear DAEs with properly formulated leading terms are introduced in [2, 11] as equations

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

where $B \in C(\mathcal{I}, L(\mathbb{R}^m))$ and the matrix functions $A \in C(\mathcal{I}, L(\mathbb{R}^\ell, \mathbb{R}^m))$, $D \in C(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^\ell))$ are well matched together, i.e.,

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

and there is a projector function $R \in C^1(\mathcal{I}, L(\mathbb{R}^\ell))$ such that

$$\ker A(t) = \ker R(t), \quad \text{im } D(t) = \text{im } R(t) \quad \forall\, t \in \mathcal{I}.$$

Recall (cf. [14, 12]) that (2.2) holds if and only if

$$\ker A(t)D(t) = \ker D(t), \quad \text{im } A(t)D(t) = \text{im } A(t) \quad \forall\, t \in \mathcal{I}.$$

Such a formulation is usually given in applications in a natural way. Note that the assumptions imply both subspaces $\ker A(t)$ and $\text{im } D(t)$ to have constant dimension. Furthermore, the relations

$$A(t) = A(t)R(t), \quad D(t) = R(t)D(t)$$

are satisfied. According to the DAE formulation (2.1) we look for solutions belonging to the function space

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

For each continuous function $x(\cdot)$ having a continuously differentiable component $D(\cdot)x(\cdot)$ and satisfying (2.1), the relation $x(t) \in \mathcal{M}_0(t)$ with

$$\mathcal{M}_0(t) := \{\bar{x} \in \mathbb{R}^m : B(t)\bar{x} - q(t) \in \text{im } A(t)\}$$

holds true, i.e., all solution values have to belong to the set $\mathcal{M}_0(t)$. Introduce the leading nullspace

$$N_0(t) := \ker D(t) = \ker A(t)D(t)$$

and projectors $Q_0(t)$, $P_0(t) \in L(\mathbb{R}^m)$ for $t \in \mathcal{I}$ such that im $Q_0(t) = N_0(t)$ and $P_0(t) := I - Q_0(t)$. The subspace

$$S_0(t) := \{z \in \mathbb{R}^m : B(t)z \in \text{im } A(t)\}$$

coincides with the tangent space $T_x\mathcal{M}_0(t)$ for $x \in \mathcal{M}_0(t)$ if the latter one is defined. Contrarily to the index-1 case, the set $\mathcal{M}_0(t)$ is not filled by DAE solutions for index-2 DAEs. Solutions of index-2 systems have to satisfy certain hidden constraints. We may describe them if we introduce the further characteristic subspaces

$$N_1(t) := \ker G_1(t), \qquad S_1(t) := \{z \in \mathbb{R}^m : B(t)P(t)z \in \text{im } G_1(t)\},$$

where $G_1(t) := A(t)D(t) + B(t)Q(t)$. Remember (cf. [12]) that the DAE (2.1) has index 1 if and only if $G_1(t)$ is nonsingular for all $t \in \mathcal{I}$.

Definition 2.1 *The equation (2.1) is an index-2 DAE if $G_1(t)$ is singular with constant rank and $N_1(t) \cap S_1(t) = \{0\}$ on $\mathcal{I}$.*

Remark: Definition 2.1 is independent of the choice of the projector function $Q(t)$. For any other projector function $\tilde{Q}(t)$ onto $\ker A(t)D(t)$, we introduce the corresponding $\tilde{G}_1(t) := A(t)D(t) + B(t)\tilde{Q}(t)$. Then, we obtain

$$\tilde{G}_1(t) = G_1(t)(P(t) + \tilde{Q}(t)) \quad \text{and} \quad G_1(t) = \tilde{G}_1(t)(\tilde{P}(t) + Q(t)),$$

which implies for the corresponding spaces $\tilde{N}_1(t)$ and $\tilde{S}_1(t)$ that

$$\tilde{N}_1(t) \cap \tilde{S}_1(t) = (\tilde{P} + Q)(N_1(t) \cap S_1(t)).$$

Obviously, $\tilde{N}_1(t) \cap \tilde{S}_1(t)$ is trivial if and only if $N_1(t) \cap S_1(t)$ is trivial. $\square$

Given a DAE in the standard formulation

$$\bar{A}x' + \bar{B}x = q \tag{2.3}$$

we may factorize $\bar{A}(t) = A(t)D(t)$ with D(t) continuously differentiable, and reformulate (2.3) as

$$A(Dx)' + Bx = q \tag{2.4}$$

with $B(t) = \bar{B}(t) - A(t)D(t)'$. If the leading coefficient matrices A and D are well matched, then equation (2.3) has the tractability index 2 (see e.g. [13]) if and only if equation (2.4) has the index 2 (see Definition 2.1). This follows from the following theorem.

Theorem 2.2 *The characteristic subspaces N_0, S_0, N_1 and S_1 of the formulation (2.4) coincide with the corresponding subspaces $\bar{N}_0$, $\bar{S}_0$, $\bar{N}_1$ and $\bar{S}_1$ of the formulation (2.3).*

Proof: Since A and D are well matched together, we have the properties

$$\ker AD = \ker D \quad \text{and} \quad \mathrm{im}\ AD = \mathrm{im}\ A.$$

This leads to

$$\begin{aligned}\bar{N}_0 &= \ker \bar{A} = \ker AD = \ker D = N_0,\\ \bar{S}_0 &= \{z \in R^m : \ \bar{B}z \in \mathrm{im}\ \bar{A}\} = \{z \in R^m : \ (B + AD')z \in \mathrm{im}\ AD\}\\ &= \{z \in R^m : \ Bz \in \mathrm{im}\ A\} = S_0.\end{aligned}$$

Since $\ker \bar{A} = \ker AD$, we may choose $\bar{Q}$ as the same projector function as Q. Regarding $D' - DP' = D'P$, we find $(\bar{B} - \bar{A}\bar{P}')\bar{Q} = BQ$. Consequently,

$$\bar{A}_1 = \bar{A} + (\bar{B} - \bar{A}\bar{P}')\bar{Q} = AD + BQ = G_1.$$

This implies immediately that $\bar{N}_1 = \ker \bar{A}_1 = \ker G_1 = N_1$. Finally,

$$\bar{S}_1 = \{z \in R^m : \ (\bar{B} - \bar{A}\bar{P}')\bar{P}z \in \mathrm{im}\ \bar{A}_1\} = \{z \in R^m : \ BPz \in \mathrm{im}\ G_1\} = S_1.$$

□

Corollary 2.3 *The index as well as the characteristic subspaces N_0, S_0, N_1 and S_1 are independent of the choice of the reformulation (2.4) with well matched matrix functions A and D.*

For index-2 DAEs, the spaces $N_1(t)$ and $S_1(t)$ span the whole space R^m (see e.g. [13]), with other words

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

Therefore there is a projector $Q_1(t) \in L(\mathbb{R}^m)$ onto $N_1(t)$ along $S_1(t)$. Then the matrix

$$G_2(t) := G_1(t) + B(t)P(t)Q_1(t)$$

is nonsingular and the relation $Q_1(t) = Q_1(t)G_2^{-1}(t)B(t)P_0(t)$ holds true. This implies immediately $Q_1(t)Q_0(t) = 0$ to be satisfied. Introduce $P_1(t) := I - Q_1(t)$ and the reflexive generalized inverse $D(t)^-$ of $D(t)$ with the properties

$$D(t)D(t)^- = R(t), \quad D(t)^-D(t) = P_0(t).$$

The inverse of $G_2(t)$ represents a suitable scaling for separating the inherent regular ODE of (2.1) from the constraints, since it satisfies the following properties:

$$G_2^{-1}A = P_1D^-, \quad G_2^{-1}BQ_0 = Q_0, \quad G_2^{-1}BP_0Q_1 = Q_1.$$

For a better reading, we have dropped the argument t. Consequently, (2.1) is equivalent to the system

$$P_1D^-(Dx)' + G_2^{-1}BP_0P_1x + Q_0x + Q_1x = G_2^{-1}q. \tag{2.5}$$

If we regard the relation $D^-(DP_1) + Q_0P_1 + Q_1 = I$, it is clear that (2.5) is equivalent to the system

$$\begin{aligned} DP_1D^-(Dx)' + DP_1G_2^{-1}BP_0P_1x &= DP_1G_2^{-1}q, &(2.6)\\ Q_0P_1D^-(Dx)' + Q_0P_1G_2^{-1}BP_0P_1x + Q_0x &= Q_0P_1G_2^{-1}q, &(2.7)\\ Q_1x &= Q_1G_2^{-1}q. &(2.8)\end{aligned}$$

Assume the spaces $D(t)S_1(t)$ and $D(t)N_1(t)$ to be spanned by continuously differentiable functions on $\mathcal{I}$. Then, the projector DP_1D^- is continuously differentiable on $\mathcal{I}$ and we may rewrite the system (2.6)-(2.8) as

$$\begin{aligned} (DP_1x)' - (DP_1D^-)'(DP_1x) + DP_1G_2^{-1}BD^-(DP_1x)&\\ - (DP_1D^-)'(DQ_1x) &= DP_1G_2^{-1}q, &(2.9)\\ -Q_0Q_1D^-(DQ_1x)' + Q_0x - Q_0Q_1D^-(DP_1D^-)'(DP_1x)&\\ + Q_0P_1G_2^{-1}BD^-(DP_1x) &= Q_0P_1G_2^{-1}q, &(2.10)\\ Q_1x &= Q_1G_2^{-1}q. &(2.11)\end{aligned}$$

Equation (2.11) provides a simple algebraic expression for the component Q_1x. Taking this into account, equation (2.9) may be written as a regular ODE for the component $u := DP_1x$:

$$u' - (DP_1D^-)'u + DP_1G_2^{-1}BD^-u = DP_1G_2^{-1}q + (DP_1D^-)'DQ_1G_2^{-1}q. \quad (2.12)$$

A few technical considerations show (see [14]) that the ODE (2.12) has $D(t)S_1(t)$ as an invariant subspace and that all terms of (2.12) do not depend on the choice of the projector Q_0. Consequently, equation (2.12) represents the inherent regular ODE (IRODE) of (2.1), which is uniquely determined by the problem data. Finally, equation (2.10) provides an expression for the component Q_0x that involves the hidden constraints of the DAE (2.1). All solutions of (2.1) are given by

$$\begin{aligned} x &= D^-(DP_1x) + D^-D(Q_1x) + Q_0x\\ &= D^-u + D^-DQ_1G_2^{-1}q + Q_0P_1G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1}q)'\\ &\quad + Q_0Q_1D^-(DP_1D^-)'u - Q_0P_1G_2^{-1}BD^-u,\end{aligned}$$

where u satisfies the inherent regular ODE (2.12). Regarding the relation $u = DP_1D^-u$, we obtain the solution representation

$$x = KP_0P_1D^-u + (Q_0P_1 + P_0Q_1)G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1}q)' \quad (2.13)$$

where $K := I + Q_0Q_1D^-(DP_1D^-)'D - Q_0P_1G_2^{-1}BP_0$ is a nonsingular matrix function. Obviously, the set

$$\begin{aligned} \mathcal{M}_1(t) \;:= \{z: \; z = (KP_0P_1D^-)(t)u + ((Q_0P_1 + P_0Q_1)G_2^{-1}q)(t)&\\ + (Q_0Q_1D^-)(t)(DQ_1G_2^{-1}q)'(t), \quad u \in (DS_1)(t)\}&\end{aligned}$$

represents the constraints (inclusively the hidden ones) of the DAE (2.1).

Let us stress that the decoupling (2.9)-(2.11) of the DAE (2.1) is rather a theoretical tool for giving an insight into the DAE structure. In practice, this decoupling is not explicitly available. However, the consideration of the inner structure of the DAE will allow us to answer the question whether the inherent regular ODE (2.12) is numerically integrated by an appropriate method if we apply a numerical integration method directly to the original DAE (2.1). One could think that this is always the case. However, in numerous papers (e.g. [1, 10, 4, 15]) it was observed that the implicit Euler method applied to certain DAEs behaves like an explicit method (from the stability point of view). One is suddenly confronted with additional stepsize restrictions. It turns out that the implicit Euler method applied directly to such DAEs leads to an explicit Euler method for the corresponding inherent ODEs. In the next section, we will explain how DAEs and their discretisation should be formulated in order to avoid such effects.

3 Numerical Integration Methods and Numerically Well Formulated Index-2 DAEs

Here, we want to investigate under which conditions the numerical solution of DAEs (2.1) obtained by BDF methods and IRK(DAE) methods (stiffly accurate Runge-Kutta methods with a nonsingular coefficient matrix $\mathcal{A}$) shows the same properties as known for regular ODEs. To answer this question, we will study under which conditions discretisation and decoupling commute and, correspondingly, when the diagram in Figure 2 is commutative. As decoupling we

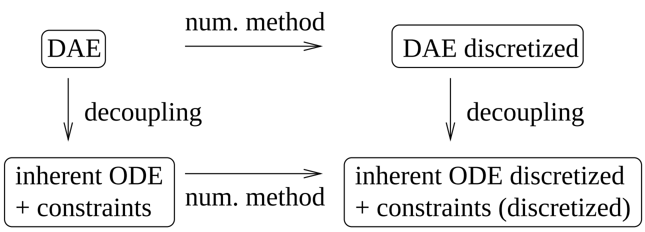

Figure 2: Decoupling and discretisation diagram.

consider the procedure to derive the system (2.6)-(2.8) from the DAE (2.1). Equation (2.9) represents the inherent ODE and the equations (2.10)-(2.11) describe the constraints. We say that we apply a numerical method to the decoupled DAE if the numerical method is applied to the inherent ODE and the constraints are considered in the discrete time points. Naturally, the derivative part $(DQ_1x)' = (DQ_1G_2^{-1}q)'$ involved in the hidden constraints (cf. (2.10)) is considered to be discretized in the same way as the numerical method is for the derivatives of the inherent ODE.

For more clarity, we will consider only constant stepsizes, but we do not need this restriction for the results formulated in this paper. The BDF method applied to the index-2 DAE (2.1) can be written as

$$A_n[Dx]'_n + B_n x_n = q_n, \tag{3.1}$$

with $A_n := A(t_n)$, $B_n := B(t_n)$, $q_n := q(t_n)$, and

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

IRK(DAE) methods have a nonsingular coefficient matrix function $\mathcal{A}$, the last row of $\mathcal{A}$ coincides with b^T and $c_s = 1$. Consequently, an IRK(DAE) method is given by the system

$$A_{ni}[DX]'_{ni} + B_{ni}X_{ni} = q_{ni}, \qquad i = 1, ..., s \tag{3.2}$$

together with

$$x_n = x_{n-1} + h\sum_{i=1}^{s} b_i X_{ni},$$

with $A_{ni} := A(t_{ni})$, $B_{ni} := B(t_{ni})$, $q_{ni} := q(t_{ni})$, $\mathcal{A}^{-1} =: (\hat\alpha_{ij})$, $t_{ni} := t_{n-1} + c_i h$,

$$[DX]'_{ni} := \frac{1}{h}\sum_{j=1}^{s} \hat\alpha_{ij}(D_{nj}X_{nj} - D_{n-1}x_{n-1}), \qquad i = 1, ..., s.$$

Obviously, the solution x_n coincides with the last stage solution X_{ns}. Consequently, the BDF solution x_n as well as the Runge-Kutta solution x_n belong to the manifold $\mathcal{M}_0(t_n)$, i.e., BDF methods and IRK(DAE) methods are distinguished by the fact that the numerical solution automatically satisfies the index-1 constraints.
We cannot expect that the numerical solution x_n satisfies the hidden constraints, i.e. that it belongs to $\mathcal{M}_1(t_n)$. Therefore, we consider the corresponding "numerical" constrained set

$$\begin{aligned}\mathcal{M}_{1n} \quad := \{z: \quad & z = (KP_0P_1D^-)(t_n)u + ((Q_0P_1 + P_0Q_1)G_2^{-1}q)(t_n) \\ & + (Q_0Q_1D^-)(t_n)[DQ_1G_2^{-1}q]'_n, \quad u \in DS_1(t_n)\}\end{aligned}$$

with

$$[DQ_1G_2^{-1}q]'_n := \begin{cases} \frac{1}{h}\sum_{j=0}^{k} \alpha_j (DQ_1G_2^{-1}q)(t_{n-j}) & \text{BDF} \\ \frac{1}{h}\sum_{j=1}^{s} \hat\alpha_{sj}\left((DQ_1G_2^{-1}q)(t_{nj}) - (DQ_1G_2^{-1}q)(t_{n-1})\right) & \text{IRK}\end{cases}$$

and we require for a commutative diagram in Figure 2 that the numerical solution x_n of (3.1) or (3.2) belongs to $\mathcal{M}_{1n}$.

Applying the same decoupling technique as in Section 2 to the BDF equation (3.1) we obtain the equivalent system

$$\begin{aligned} (DP_1D^-)_n[Dx]'_n + (DP_1G_2^{-1}BP_0P_1)_nx_n &= (DP_1G_2^{-1}q)_n, & (3.3)\\ (Q_0P_1D^-)_n[Dx]'_n + (Q_0P_1G_2^{-1}BP_0P_1)_nx_n + Q_{0n}x_n &= (Q_0P_1G_2^{-1}q)_n, & (3.4)\\ Q_{1n}x_n &= (Q_1G_2^{-1}q)_n. & (3.5)\end{aligned}$$

If one replaces the index n by the index ni and x by X in the system (3.3)-(3.5) one obtains the decoupled system of the IRK(DAE) equation (3.2).

For shorter expressions, let us introduce the notations

$$[Mx]'_n := \frac{1}{h}\sum_{j=0}^{k} \alpha_j M(t_{n-j})x_{n-j}$$

in the BDF case and

$$[MX]'_{ni} := \frac{1}{h}\sum_{j=1}^{s} \hat{\alpha}_{ij}(M(t_{nj})X_{nj} - M(t_{n-1})x_{n-1})$$

in the IRK(DAE) case for any matrix function $M(t)$ defined on $\mathcal{I}$. Then, we may rewrite the system (3.3)-(3.5) as

$$\begin{aligned} [DP_1x]'_n - (I-(DP_1D^-)_n)[DP_1x]'_n + (DP_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n)&\\ + (DP_1D^-)_n[DQ_1x]'_n &= (DP_1G_2^{-1}q)_n, & (3.6)\\ -(Q_0Q_1D^-)_n[DQ_1x]'_n + Q_{0n}x_n - (Q_0Q_1D^-)_n[DP_1x]'_n&\\ + (Q_0P_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n) &= (Q_0P_1G_2^{-1}q)_n, & (3.7)\\ Q_{1n}x_n &= (Q_1G_2^{-1}q)_n. & (3.8)\end{aligned}$$

Again, an analogous decoupling for IRK(DAE) methods is given by replacing the index n by the index ni and x by X in the system (3.6)-(3.8).
We want to stress here that we have an error propagation only in equation (3.6). Obviously, there is no error propagation in (3.8) and, consequently, none in (3.7).

Remark: As explained in Appendix A, the decoupled system (3.6)-(3.8) provides immediately the convergence of BDF and IRK(DAE) methods applied to (2.1) on compact $\mathcal{I}$ if the stepsize h tends to zero. □

The system (3.6)-(3.8) represents the result of the decoupling after discretising the DAE (2.1). On the other hand, if we decouple first (see (2.9)-(2.11)) and discretize afterwards as described in Figure 2, then we obtain

$$\begin{aligned} [DP_1x]'_n - (DP_1D^-)'_n(D_nP_{1n}x_n) + (DP_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n)&\\ - (DP_1D^-)'_n(D_nQ_{1n}x_n) &= (DP_1G_2^{-1}q)_n, & (3.9)\\ -(Q_0Q_1D^-)_n[DQ_1x]'_n + Q_{0n}x_n - (Q_0Q_1D^-)_n(DP_1D^-)'_n(D_nP_{1n}x_n)&\\ + (Q_0P_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n) &= (Q_0P_1G_2^{-1}q)_n, & (3.10)\\ Q_{1n}x_n &= (Q_1G_2^{-1}q)_n. & (3.11)\end{aligned}$$

for BDF methods. Analogously, we get the corresponding system for IRK(DAE) methods by replacing the index n by the index ni and x by X.

It is clear that the system (3.6)-(3.8) coincides with the system (3.9)-(3.11) if

$$
\begin{aligned}
(I - (DP_1D^-)_n)[DP_1x]'_n &= (DP_1D^-)'_n(D_nP_{1n}x_n), && (3.12)\\
(DP_1D^-)_n[DQ_1x]'_n &= -\,(DP_1D^-)'_n(D_nQ_{1n}x_n) && (3.13)\\
(Q_0Q_1D^-)_n[DP_1x]'_n &= -\,(Q_0Q_1D^-)_n(DP_1D^-)'_n(D_nP_{1n}x_n) && (3.14)
\end{aligned}
$$

The next lemma presents practical criteria for the conditions (3.12)-(3.14) to be satisfied.

Lemma 3.1 *The following conditions are satisfied.*

(i) If the space $DS_1 = \mathrm{im}\ DP_1D^-$ *is constant, then*

$$
\begin{aligned}
(I - (DP_1D^-)_n)[DP_1x]'_n &= (DP_1D^-)'_n(D_nP_{1n}x_n) = 0,\\
(Q_0Q_1D^-)_n[DP_1x]'_n &= -\,(Q_0Q_1D^-)_n(DP_1D^-)'_n(D_nP_{1n}x_n) = 0.
\end{aligned}
$$

(ii) If the space $DN_1 = \mathrm{im}\ DQ_1D^-$ *is constant, then*

$$(DP_1D^-)_n[DQ_1x]'_n = -\,(DP_1D^-)'_n(D_nQ_{1n}x_n) = 0.$$

The assertions follow directly from Lemma B.1 part (iii) and (iv). Lemma 3.1 implies immediately that the diagram in Figure 2 commutes supposed DS_1 and DN_1 are constant. This leads us to the following definition.

Definition 3.2 *The DAE index-2 (2.1) is said to be numerically well formulated if the spaces* $DS_1 = \mathrm{im}\ DP_1D^-$ *and* $DN_1 = \mathrm{im}\ DQ_1D^-$ *are constant.*

Remark: If the DAE (2.1) has only index 1, then we have $N_1 = 0$. This implies that the definition harmonizes with the definition of numerically well formulated index-1 DAEs in [12], where im D is required to be constant. □

4 Numerically Well Formulated Refactorizations

In practice we may be confronted with DAEs which are not numerically well formulated. In this section, we show how one can reformulate DAEs such that they are numerically well formulated. Furthermore, we investigate under which conditions a numerical method applied to the original DAE produces a solution in a small neighbourhood of the solution of a numerical well-formulated refactorization.

Example 4.1 *In [7] it is shown that the BDF method applied to the example*

$$\begin{pmatrix} 0 & 0 \\ 1 & \eta t \end{pmatrix} x' + \begin{pmatrix} 1 & \eta t \\ 0 & 1+\eta \end{pmatrix} x = \begin{pmatrix} g(t) \\ 0 \end{pmatrix} \qquad (4.1)$$

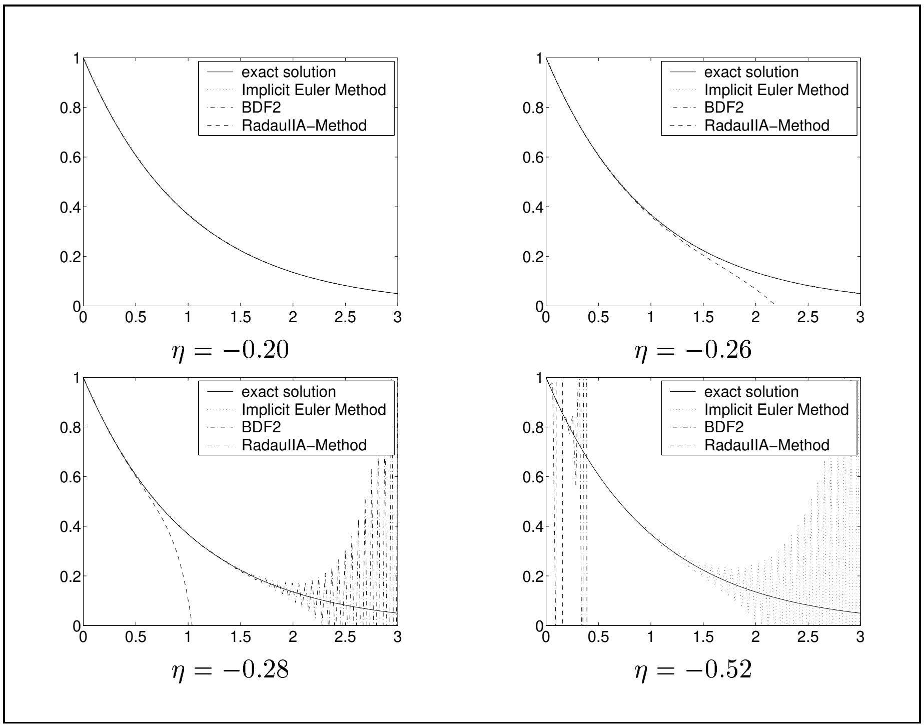

Figure 3: Numerical solutions (second component) of (4.1) for $h = 10^{-1.5}$ and different values of η. $x^0 = (1,1)^T$ was used as consistent initial value.

fails completely for $\eta = -1$ and is exponentially unstable for all the other parameter values $\eta < -0.5$. See the results in Figure 3 for $g(t) = e^{-t}$, which implies the exact solution to be $x_1(t) = (1-\eta t)e^{-t}$, $x_2(t) = e^{-t}$.
But, simple reformulations are numerically well formulated and lead, consequently, to a correct BDF solution. For example, if we reformulate (4.1) as

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \left(\begin{pmatrix} 1 & \eta t \end{pmatrix} x \right)' + \begin{pmatrix} 1 & \eta t \\ 0 & 1 \end{pmatrix} x = \begin{pmatrix} g(t) \\ 0 \end{pmatrix},$$

then we have $DN_1 = \mathbb{R}$ and $DS_1 = 0$. That means, solving this reformulation yields correct numerical solutions (see Figure 4).

We consider again DAEs of the form

$$A(t)(D(t)x)' + B(t)x(t) = q(t), \tag{4.2}$$

which have well matched A and D but are not numerically well formulated. In [12], it was already discussed how to obtain reformulations of (4.2) with im D constant. Let us assume here that im D is already constant.

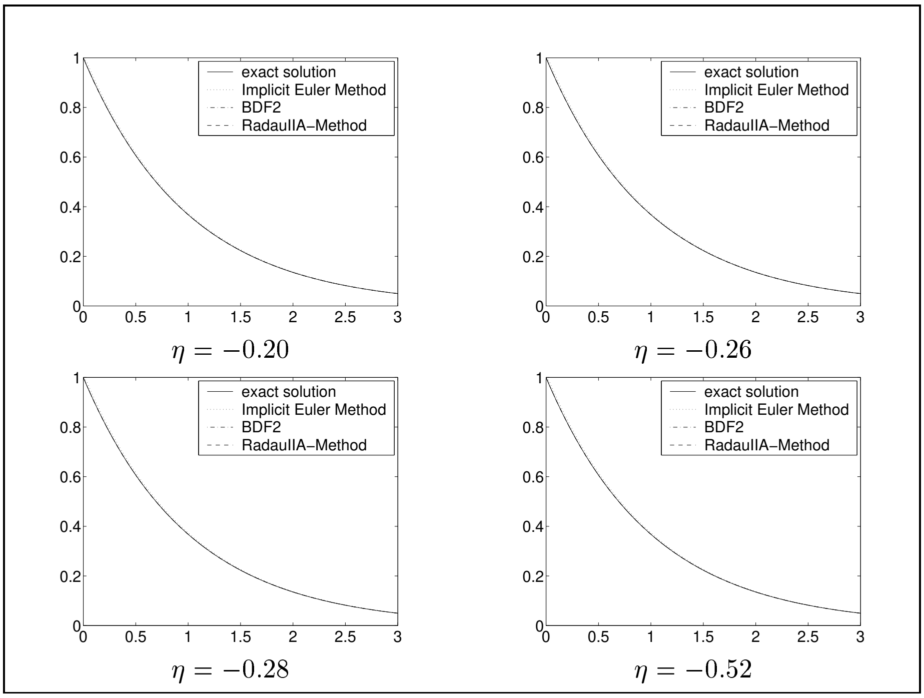

Figure 4: Numerical solutions (second component) of (4.1) for $h = 10^{-1.5}$ and different values of η. $x^0 = (1,1)^T$ was used as consistent initial value.

For any nonsingular matrix K, we may reformulate (4.2) as

$$\tilde{A}(t)(\tilde{D}(t)x)' + \tilde{B}(t)x(t) = q(t) \tag{4.3}$$

with the matrix coefficients $\tilde{A}(t) = A(t)K^{-1}(t)$, $\tilde{D}(t) = K(t)D(t)$ and $\tilde{B}(t) = B(t) - A(t)K^{-1}(t)K'(t)D(t)$ or $\tilde{A}(t) = A(t)K(t)$, $\tilde{D}(t) = K(t)^{-1}D(t)$ and $\tilde{B}(t) = B(t) + A(t)K'(t)K(t)^{-1}D(t)$. Remember that (4.3) has well matched $\tilde{A}(t)$ and $\tilde{D}(t)$ since $A(t)$ and $D(t)$ are well matched (cf. [12]). Furthermore, Theorem 2.2 implies that (4.3) has index μ if and only if (4.2) has index μ (for $\mu = 1$ and $\mu = 2$).

Such kind of reformulations may lead to numerically well formulated problems. Consider again the Example 1.1 presented in the introduction. We may reformulate it as

$$\begin{pmatrix} 1 & 0 \\ \eta t - 1 & 1 \\ 0 & 0 \end{pmatrix} \left(\begin{pmatrix} 1 & 0 & 0 \\ 1-\eta t & 1 & 0 \end{pmatrix} x \right)' + \begin{pmatrix} \lambda & -1 & -1 \\ \eta t(1-\eta t) & \lambda & -\eta t \\ 1-\eta t & 1 & 0 \end{pmatrix} x = 0$$

using $\tilde{A}(t) = A(t)K^{-1}(t)$ and $\tilde{D}(t) = K(t)D(t)$ with

$$K = \begin{pmatrix} 1 & 0 \\ 1-\eta t & 1 \end{pmatrix}.$$

In this way, the spaces

$$\tilde{D}S_1 = \{z \in \mathbb{R}^2 :\ z_2 = 0\}, \qquad \tilde{D}N_1 = \{z \in \mathbb{R}^2 :\ z_1 = z_2\}$$

are constant and the numerical solution has the expected asymptotic behaviour for all parameter values η (see Figure 5). Note that the successful reformulation is no longer a Hessenberg system but a semi-implicit DAE.

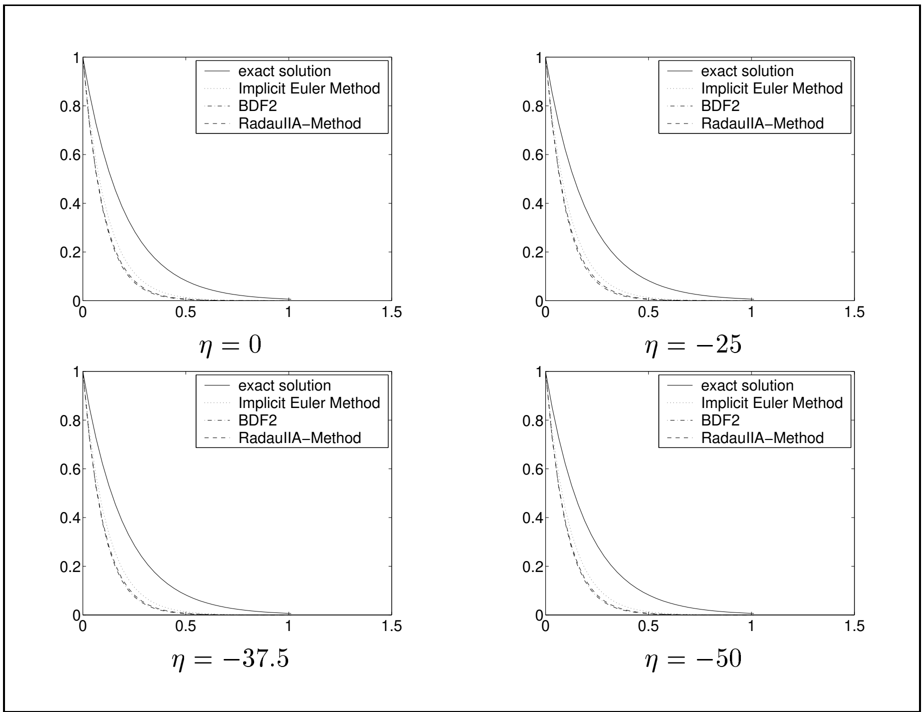

Figure 5: Numerical solutions (first component) of (1.4) for $\lambda = 10$ and $h = 10^{-1.5}$ and the consistent initial value $x^0 = (1, -1, 1)^T$.

Regarding the fact that (cf. [14])

$$DN_1 \oplus DS_1 \oplus \ker A = \mathbb{R}^\ell,$$

we find a suitable reformulation if the bases of the subspaces DS_1, DN_1 and $\ker A$ are known. If

$$\begin{aligned} DN_1 &= \text{span}\,\{d_1, ..., d_{m-r_1}\}, \\ DS_1 &= \text{span}\,\{d_{m-r_1+1}, ..., d_{r_0}\}, \\ \ker A &= \text{span}\,\{d_{r_0+1}, ..., d_\ell\} \end{aligned}$$

with $d_i \in C^1(\mathcal{I}, \mathbb{R}^\ell)$ for $i = 1, ..., \ell$, $r_0 = \text{rank}G_0 = \text{rank}D$ and $r_1 = \text{rank}G_1$, then the particular matrix $K = (d_1 ... d_\ell)$ leads to numerically well formulated

DAEs with

$$\tilde{S}_1 = 0 \times \mathbb{R}^{r_0+r_1-m} \times 0, \qquad \tilde{D}N_1 = \mathbb{R}^{m-r_1} \times 0 \times 0.$$

This reformulation corresponds to a coordinate transformation in $\mathbb{R}^\ell$.
Observe that the reformulations (4.3) require the computation of $\tilde{A}K'D$ or $AK'\tilde{D}$ and therefore they may be expensive. It may happen that the reformulated problem gives the same numerical solution as the original problem, i.e. both them are numerically equivalent. For instance if

$$\begin{aligned} K(t_*)D(t) &= K(t)D(t) \quad \text{for all } t_*, t \in \mathcal{I} \quad \text{or} \qquad (4.4)\\ A(t)K(t_*) &= A(t)K(t) \quad \text{for all } t_*, t \in \mathcal{I}, \qquad (4.5)\end{aligned}$$

then (4.2) is numerically equivalent to (4.3). In this case the effort to compute the reformulated problem is completely useless.
For index-2 DAEs, it holds that

$$D(t)S_1(t) \oplus D(t)N_1(t) = \text{im } D(t).$$

Numerically well formulated index-2 DAEs satisfy that DS_1 and DN_1 are constant subspaces, and thus, trivially, $D(t)S_1(t) \oplus D(t_*)N_1(t_*) = \text{im } D(t)$ for all $t, t_* \in \mathcal{I}$ holds. As the following result shows, this condition also holds if the DAE is numerically equivalent to a numerically well formulated one.

Proposition 4.2 *Given (4.2) and the reformulated problem (4.3), if $\tilde{D}S_1$ or $\tilde{D}N_1$ are constant,* im D *is constant and either (4.4) or (4.5) holds, then*

$$D(t)S_1(t) \oplus D(t_*)N_1(t_*) = \text{im } D(t) \quad \textit{for all } t, t_* \in \mathcal{I} \qquad (4.6)$$

Proof: We assume first that (4.4) holds true. Recall that this condition is associated with the factorization $A(t) = \tilde{A}(t)K(t)$ and $\tilde{D}(t) = K(t)D(t)$. We begin checking that

$$D(t)S_1(t) \cap D(t_*)N_1(t_*) = \{0\}.$$

For $x \in D(t)S_1(t) \cap D(t_*)N_1(t_*)$ we have $x = D(t)P_1(t)w = D(t_*)Q_1(t_*)v$ for some $w, v \in \mathbb{R}^m$. Thus for any $\bar{t}$,

$$K(\bar{t})x = K(\bar{t})D(t)P_1(t)w = K(t)D(t)P_1(t)w = \tilde{D}(t)P_1(t)w,$$

and

$$K(\bar{t})x = K(\bar{t})D(t_*)Q_1(t_*)v = K(t_*)D(t_*)Q_1(t_*)v = \tilde{D}(t_*)Q_1(t_*)v,$$

hold true, i.e., $K(\bar{t})x = \tilde{D}(t)P_1(t)w = \tilde{D}(t_*)Q_1(t_*)v$. If $\tilde{D}S_1$ is constant or $\tilde{D}N_1$ is constant, then we obtain $K(\bar{t})x = 0$. Consequently, $x = 0$.
We assume now that (4.5) holds true. Recall that in this case the factorisation is given by $D(t) = K(t)\tilde{D}(t)$, and $\tilde{A}(t) = A(t)K(t)$. To prove that $D(t)S_1(t) \cap$

$D(t_*)N_1(t_*) = \{0\}$, we take $x \in D(t)S_1(t) \cap D(t_*)N_1(t_*)$, i.e., $x = D(t)P_1(t)w = D(t_*)Q_1(t_*)v$ for some $w,\, v \in \mathbb{R}^m$. Thus, for any $\bar{t}$,

$$A(\bar{t})x = A(\bar{t})K(t)\tilde{D}(t)P_1(t)w = \tilde{A}(\bar{t})\tilde{D}(t)P_1(t)w,$$

and also

$$A(\bar{t})x = A(\bar{t})K(t_*)\tilde{D}(t_*)Q_1(t_*)v = \tilde{A}(\bar{t})\tilde{D}(t_*)Q_1(t_*)v$$

hold true. Multiplying these expressions by $K(\bar{t})^{-1}D(\bar{t})G_2(\bar{t})^{-1}$ we obtain (with $\tilde{D}^- = D^-K$)

$$\tilde{D}(\bar{t})P_1(\bar{t})\tilde{D}(\bar{t})^-\tilde{D}(t)P_1(t)w = \tilde{D}(\bar{t})P_1(\bar{t})\tilde{D}(\bar{t})^-\tilde{D}(t_*)Q_1(t_*)v.$$

If $\tilde{D}S_1$ is constant, we take $\bar{t} = t_*$. If $\tilde{D}N_1$ is constant, we take $\bar{t} = t$ to obtain $\tilde{D}(t)P_1(t)w = 0$ and thus $x = 0$.
The fact that im D is constant implies $D(t)S_1(t) + D(t_*)N_1(t_*) \subseteq \text{im } D$, and as both subspaces have the same dimension we obtain the desired result. □

Given the DAE (2.3) in standard form we may factorize $\bar{A}(t) = A(t)D(t)$ to reformulate it with well matched leading matrices. If

$$A(t)D(t_*) = A(t)D(t) \quad \forall t,\ t_* \in \mathcal{I}, \tag{4.7}$$

then (2.3) and (2.4) are numerically equivalent and thus there is no need to realize the reformulation in practice. Recall that for index-2 DAEs the equality $S_1(t) \oplus N_1(t) = \mathbb{R}^m$ holds true. The following result gives a criterion to avoid unnecessary reformulations.

Proposition 4.3 *Given the DAE (2.3) and the reformulated DAE (2.4), if DS_1 or DN_1 are constant and (4.7) is satisfied, then*

$$S_1(t) \oplus N_1(t_*) = \mathbb{R}^m \quad \textit{for all } t, t_* \in \mathcal{I}. \tag{4.8}$$

Proof: Observe that (4.7) implies that $\ker D(t)$ is constant. To prove the result, we simply have to check that $N_1(t_*) \cap S_1(t) = \{0\}$. For $x \in S_1(t) \cap N_1(t_*)$ we have $x = P_1(t)x = Q_1(t_*)x$. Thus for any $\bar{t}$ it holds

$$\bar{A}(\bar{t})x = \bar{A}(\bar{t})P_1(t)x = A(\bar{t})D(\bar{t})P_1(t)x = A(\bar{t})D(t)P_1(t)x$$

and

$$\bar{A}(\bar{t})x = \bar{A}(\bar{t})Q_1(t_*)x = A(\bar{t})D(\bar{t})Q_1(t_*)x = A(\bar{t})D(t_*)Q_1(t_*)x.$$

Now, multiplying the above expressions by $D(\bar{t})G_2(\bar{t})^{-1}$ we obtain

$$D(\bar{t})P_1(\bar{t})D(\bar{t})^-D(t)P_1(t)x = D(\bar{t})P_1(\bar{t})D(\bar{t})^-D(t_*)Q_1(t_*)x.$$

If DS_1 is constant, then we take $\bar{t} = t_*$. If DN_1 is constant, then we take $\bar{t} = t$ in order to obtain $D(t)P_1(t)x = 0$, i.e., $x \in \ker D(t)$. As $Q_1 = Q_1P_0$ and $\ker D(t)$ is constant, we get $x = Q_1(t_*)x = Q_1(t_*)P_0(t_*)x = 0$. □

Consequently, if a DAE is not numerically well formulated, we should check condition (4.6) before realizing reformulations. If (4.6) holds, we should take into consideration the numerical equivalence to a numerically well formulated DAE, i.e., the DAE only seems to be not well formulated.

If, on the contrary, condition (4.6) is not satisfied, then we can conclude that the reformulations proposed by factorising the leading terms will never lead to a numerically well formulated DAE. However, the numerical solution of (4.2) may only slightly differ from the numerical solution of a numerically well formulated DAE. This is the case if only one of the spaces DS_1 and DN_1 is constant. We will explain this as follows.

We consider special reformulations with a continuously differentiable involution H $(H^2 = I)$

$$A(t)H(t)(H(t)D(t)x)' + (B(t) - A(t)H(t)H'(t)D(t))x(t) = q(t). \qquad (4.9)$$

These are numerically well formulated in the following cases.

Lemma 4.4 *If* im D *and* DN_1 *of (4.2) are constant, then the reformulation (4.9) is numerically well formulated for* $H = I - DQ_1D^- - P_{DN_1}$ *with a constant projector* P_{DN_1} *onto* DN_1.

Proof: Denote again $\tilde{A} := AH$, $\tilde{D} := HD$ and $\tilde{B} := B - AHH'D$ and recall (see Theorem 2.2) that $\tilde{N}_1 = N_1$ and $\tilde{S}_1 = S_1$. Finally,

$$\tilde{D}\tilde{N}_1 = HDN_1 = \text{im } HDQ_1 = \text{im } DQ_1$$

is constant and

$$\tilde{D}\tilde{S}_1 = HDS_1 = \text{im } HDP_1 = \text{im } (DP_1 - P_{DN_1}DP_1) = \text{im } (I - P_{DN_1})D$$

is constant since im D is constant. □

Lemma 4.5 *If* im D *and* DS_1 *of (4.2) are constant, then the reformulation (4.9) is numerically well formulated for the involution* $H = I - DP_1D^- - P_{DS_1}$ *with a constant projector* P_{DS_1} *onto* DS_1.

Proof: Denote $\tilde{A} := AH$, $\tilde{D} := HD$ and $\tilde{B} := B - AHH'D$. From Theorem 2.2 we know that $\tilde{N}_1 = N_1$ and $\tilde{S}_1 = S_1$. Consequently,

$$\tilde{D}\tilde{S}_1 = HDS_1 = \text{im } HDP_1 = \text{im } DP_1$$

is constant and

$$\tilde{D}\tilde{N}_1 = HDN_1 = \text{im } HDQ_1 = \text{im } (DQ_1 - P_{DS_1}DQ_1) = \text{im } (I - P_{DS_1})D$$

is constant since im D is constant. □

Now it is of interest how the numerical solution of (4.2) differs from the numerical solution of the numerically well formulated DAE (4.9). For simplicity, consider the BDF method only. For IRK(DAE) methods, the statements hold analogously. Let x_n be the BDF solution of (4.2) and $\tilde{x}_n$ be the BDF solution of (4.9). Using the same notation as in Section 3 we obtain

$$A_n[Dx]'_n + B_n x_n = q_n \tag{4.10}$$

and

$$(AH)_n[HD\tilde{x}]'_n + (B - AHH'D)_n\tilde{x}_n = q_n. \tag{4.11}$$

Subtracting one equation from the other yields

$$A_n[Dx]'_n - (AH)_n[HD\tilde{x}]'_n + (AHH'D)_n\tilde{x}_n + B_n(x_n - \tilde{x}_n) = 0.$$

This is equivalent to

$$\begin{aligned} A_n[D(x-\tilde{x})]'_n + B_n(x_n - \tilde{x}_n) \quad &= \\ &(AH)_n[HD\tilde{x}]'_n - A_n[D\tilde{x}]'_n - (AHH'D)_n\tilde{x}_n. \end{aligned}$$

Scaling this equation by $G_2^{-1}(t_n)$ and regarding $(G_2^{-1}A)_n = (P_1D^-)_n$ (see Section 2) we obtain

$$\begin{aligned} (P_1D^-)_n[D(x-\tilde{x})]'_n + (G_2^{-1}B)_n(x_n - \tilde{x}_n) \quad &= \\ (P_1D^-H)_n[HD\tilde{x}]'_n - (P_1D^-)_n[D\tilde{x}]'_n - (P_1D^-HH'D)_n\tilde{x}_n& \end{aligned} \tag{4.12}$$

Multiplying equation (4.12) by Q_1 and taking into account that $Q_1 = Q_1G_2^{-1}B$ we obtain that

$$Q_{1n}x_n = Q_{1n}\tilde{x}_n. \tag{4.13}$$

For further analysis, we consider the two cases DS_1 constant and DN_1 constant separately.

Theorem 4.6 *Let DN_1 and* im D *be constant. Then, the BDF and RK solutions x_n of (4.2) and $\tilde{x}_n$ of (4.9) with $H = I - DQ_1D^- - P_{DN_1}$, for a constant projector P_{DN_1} onto DN_1, satisfy the relation*

$$x_n - \tilde{x}_n = (Q_0P_1D^-)_n \left((DP_1D^-)'_n(D\tilde{x}_n - D\tilde{x}(t_n)) + (DP_1\tilde{x})'_n - [DP_1\tilde{x}]'_n)\right)$$

if the initial values satisfy $D_ix_i = D_i\tilde{x}_i$ for $i = 0, ..., k-1$.

Remarks:

1. Theorem 4.6 implies immediately $D_nx_n = D_n\tilde{x}_n$ to be true.
2. Theorem 4.6 is of great interest from the point of view of circuit simulation. As already described in [12], circuit systems modelled by the charge oriented modified nodal analysis (MNA) are usually of the form

   $$A(D(t)x(t))' + B(t)x(t) = q(t),$$

   with well matched matrix functions A and $D(t)$. Here, x describes the vector of charges, fluxes, nodal potentials and branch currents of voltage defining network elements. Furthermore, the structural analysis of these systems in [5] makes clear that im D as well as DN_1 are constant, which guarantees that the numerical solution by BDF and IRK(DAE) methods reflects the inner stability behaviour properly. □

Proof: Taking into consideration Lemma B.2, we find

$$\begin{aligned}(DP_1D^-H)_n[HD\tilde{x}]'_n &= (DP_1D^-)_n[Dx]'_n, \\ DP_1D^-HH'D &= 0.\end{aligned}$$

Multiplying (4.12) by D and regarding (4.13), we obtain

$$(DP_1D^-)_n[Dx - D\tilde{x}]'_n + (DP_1G_2^{-1}BD^-)_n(D_nx_n - D_n\tilde{x}_n) = 0$$

Since im D is supposed to be constant, by multiplication by $(I - P_{DN_1})$ yields that

$$\begin{aligned}&[(I - P_{DN_1})(Dx - D\tilde{x})]'_n + \\ &\quad (I - P_{DN_1})(DP_1G_2^{-1}BD^-)_n(I - P_{DN_1})D_n(x_n - \tilde{x}_n) \;=\; 0. \quad (4.14)\end{aligned}$$

This represents nothing else but the BDF method applied to a homogeneous explicit ODE for $(I - P_{DN_1})D(x - \tilde{x})$. Since the starting values satisfy $D_ix_i - D_i\tilde{x}_i = 0$ for $i = 0, ..., k$, the solution of (4.14) is identical to zero, i.e.,

$$(I - P_{DN_1})_nD_nx_n = (I - P_{DN_1})_nD_n\tilde{x}_n.$$

Multiplication by DP_1D^- implies that $(DP_1)_nx_n = (DP_1)_n\tilde{x}_n$. Regarding again (4.13), we obtain $D_nx_n = D_n\tilde{x}_n$. Lemma B.2 yields

$$\begin{aligned}(Q_0P_1D^-H)_n[HD\tilde{x}]'_n &= (Q_0P_1D^-)_n[DQ_1\tilde{x}]'_n, \\ (Q_0P_1D^-HH'D)_n &= (Q_0P_1D^-)_n(DQ_1D^-)'_nD_n.\end{aligned}$$

Multiplying (4.12) by Q_{0n}, we obtain finally

$$Q_{0n}(x_n - \tilde{x}_n) = (Q_0P_1D^-)_n(-[DP_1\tilde{x}]'_n + (DP_1D^-)'_nD_n\tilde{x}_n).$$

Since $D_nx_n - D_n\tilde{x}_n = 0$, we have $x_n - \tilde{x}_n = Q_{0n}(x_n - \tilde{x}_n)$. Furthermore, if we denote by $\tilde{x}(t)$ the exact solution of the numerically well formulated DAE 4.2, then

$$x_n - \tilde{x}_n = (Q_0P_1D^-)_n\left((DP_1D^-)'_n(D\tilde{x}_n - D\tilde{x}(t_n)) + (DP_1\tilde{x})'_n - [DP_1\tilde{x}]'_n)\right).$$

□

Example 4.7 *Consider the linear circuit given in Figure 6. The modified nodal analysis leads directly to the system*

$$\begin{pmatrix}1 & -1\\ 0 & 1\\ 0 & 0\end{pmatrix}\begin{pmatrix}C_1(t)e_1(t)\\ -C_2e_1(t) + C_2e_2(t)\end{pmatrix}' + \begin{pmatrix}G & 0 & 0\\ 0 & 0 & -1\\ 0 & 1 & 0\end{pmatrix}\begin{pmatrix}e_1(t)\\ e_2(t)\\ j(t)\end{pmatrix} = \begin{pmatrix}0\\ 0\\ v(t)\end{pmatrix}. \tag{4.15}$$

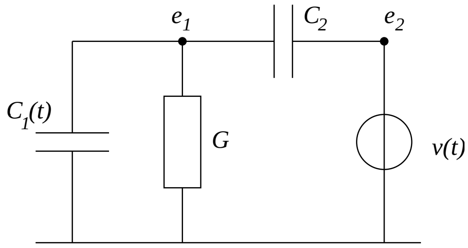

Figure 6: Linear circuit with a time dependent capacitance.

Choosing $C_1(t) = 1 + 0.25\,(\sin(t) + \cos(t))$, $C_2 = 1$, $G = 2$ *and the input voltage* $v(t) = 4\,\sin(t) + 0.25\,\sin(2t)$ *yields the solution*

$$\begin{aligned} e_1(t) &= \sin(t) + \cos(t), \\ e_2(t) &= 4\sin(t) + 0.25\sin(2t), \\ j(t) &= 3\cos(t) + 0.5\cos(2t) + \sin(t), \end{aligned}$$

for the consistent initial value $e_1(0) = 1$, $e_2(0) = 0$ *and* $j(0) = 3.5$. *Regarding*

$$D(t) = \begin{pmatrix} C_1(t) & 0 & 0 \\ -C_2 & C_2 & 0 \end{pmatrix}$$

we obtain $DN_1 = \{z \in \mathbb{R}^2 :\ z_1 = z_2\}$ *to be constant, whereas the space*

$$DS_1 = \{z \in \mathbb{R}^2 :\ C_2 z_1 + C_1(t) z_2 = 0\}$$

varies with time. Figure 7 shows that the numerical solution approximates the exact solution very well in accordance with Theorem 4.6.

Theorem 4.8 *Let* DS_1 *and* $\operatorname{im} D$ *be constant. Then, the BDF and RK solutions* x_n *of (4.2) and* $\tilde{x}_n$ *of (4.9) with* $H = I - DP_1D^- - P_{DS_1}$, *for a constant projector* P_{DS_1} *onto* DS_1, *satisfy the equations*

$$\begin{aligned} [DP_1x - DP_1\tilde{x}]'_n + (DP_1G_2^{-1}BD^-)_n((D_nP_{1n}x_n - D_nP_{1n}\tilde{x}_n)& \\ = \ (DP_1D^-)_n((DQ_1G_2^{-1}q)'_n - [DQ_1G_2^{-1}q]'_n)& \qquad (4.16) \\ Q_{0n}x_n - Q_{0n}\tilde{x}_n \ = \ -(Q_0P_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n - D_nP_{1n}\tilde{x}_n)& \\ Q_{1n}x_n - Q_{1n}\tilde{x}_n \ = \ 0& \end{aligned}$$

Remarks:

1. For BDF methods of order k, the small perturbation on the right-hand side of (4.16) is of order $\mathcal{O}(h^k)$. For IRK(DAE) methods of stage s, it is of order $\mathcal{O}(h^s)$.

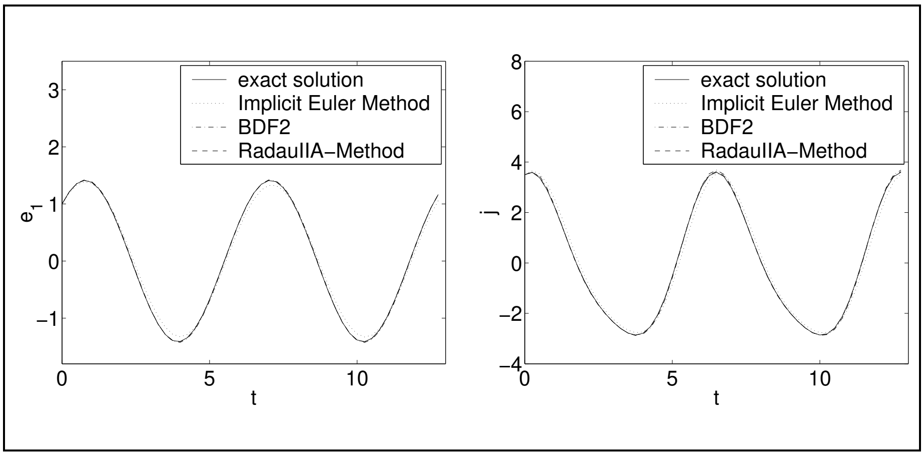

Figure 7: Numerical solutions e_1 and j (index-2 component) of (4.15) with the stepsize $h = 0.2$.

2. Since $I = D^-DP_1 + D^-DQ_1 + Q_0$, the difference between the BDF solutions x_n of (4.2) and $\tilde{x}_n$ of (4.9) is completely described by Theorem 4.8. Furthermore, it implies that $x_n = \tilde{x}_n$ if $DQ_1G_2^{-1}q \equiv 0$ and $(DP_1)_n(x_n - \tilde{x}_n) = 0$ for $n = 0, ..., k-1$.

3. Theorem 4.8 reflects the results for systems with $D = P_0$ and S_1 constant as considered in [15].

Proof: From Lemma B.3 we obtain

$$\begin{aligned} (DP_1D^-H)_n[HD\tilde{x}]'_n &= (DP_1D^-)_n[DP_1\tilde{x}]'_n, \\ (DP_1D^-HH'D)_n &= (DP_1D^-)'_n(DQ_1)_n. \end{aligned}$$

Multiplying (4.12) by D and regarding (4.13), we obtain

$$\begin{aligned} (DP_1D^-)_n[D(x-\tilde{x})]'_n &+ (DP_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n - D_nP_{1n}\tilde{x}_n) \\ &= (DP_1D^-)_n(-[DQ_1\tilde{x}]'_n - (DP_1D^-)'_n(DQ_1)_n\tilde{x}_n) \end{aligned}$$

If we multiply (4.11) by $(Q_1G_2^{-1})_n$ we find that $(DQ_1)_n\tilde{x}_n = (DQ_1G_2^{-1}q)_n$. Taking into consideration Lemma B.3 again, we obtain

$$\begin{aligned} [DP_1x - DP_1\tilde{x}]'_n &+ (DP_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n - D_nP_{1n}\tilde{x}_n) \\ &= (DP_1D^-)_n((DQ_1G_2^{-1}q)'_n - [DQ_1G_2^{-1}q]'_n) \end{aligned}$$

On the other hand, if we multiply (4.12) by Q_0 and regard (see Lemma B.3)

$$\begin{aligned} (Q_0P_1D^-H)_n[HD\tilde{x}]'_n - (Q_0P_1D^-)_n[D\tilde{x}]'_n &= 0, \\ (Q_0P_1D^-HH'D)_n &= 0, \end{aligned}$$

then we obtain

$$Q_{0n}x_n - Q_{0n}\tilde{x}_n + (Q_0P_1G_2^{-1}BD^-)_n(D_nP_{1n}x_n - D_nP_{1n}\tilde{x}_n) = 0.$$

Together with (4.13) all assertions have been proved. □

5 Contractive and Dissipative Flow

Now we consider the index-2 DAE (2.1) on the infinite interval $\mathcal{I} = [0, \infty)$. Recall that each solution $x \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ has the form (cf. (2.13))

$$x = P_{\text{can}}D^-u + (P_0Q_1 + Q_0P_1)G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1}q)', \tag{5.1}$$

with the component $u = DP_1x \in C^1(\mathcal{I}, \mathbb{R}^\ell)$ satisfying the IRODE

$$u' - (DP_1D^-)'u + DP_1G_2^{-1}BD^-u = DP_1G_2^{-1}q + (DP_1D^-)'DQ_1G_2^{-1}q, \tag{5.2}$$

$P_{\text{can}} := KP_0P_1$ and $K = I + Q_0Q_1D^-(DP_1D^-)'D - Q_0P_1G_2^{-1}BP_0$. For all solutions it holds that

$$x(t) \in \mathcal{M}_1(t), \;\; u(t) \in D(t)S_1(t), \;\; t \in \mathcal{I}.$$

Hence, the total DAE flow is restricted to the possibly time-varying constraint set $\mathcal{M}_1(t) \in \mathbb{R}^m$, which is, on the other hand, completely occupied by DAE solutions. There is further an inner flow determined by the IRODE on its invariant subspace $D(t)S_1(t) = \text{im } D(t)P_1(t)D(t)^-$. This inner flow mainly governs the DAE flow. Obviously, by the representation (5.1), the canonical projector P_{can} plays its special role when packaging the inner flow for obtaining the total one.
Let us stress once more that the derivative $(Dx)'$ involved in the DAE (2.1) consists now of two different parts. The component $(DP_1x)'$ is in fact related to the inner vector field, while $(DQ_1x)' = (DQ_1G_2^{-1}q)'$ results from the inherent differentiation.
The expression

$$(DP_1x)' = R(Dx)' + R'Dz - (DQ_1G_2^{-1}q)'$$

indicates how the inner vector field is related to the total one. Our next lemma describes this total DAE vector field. It shows also the different components to be mixed as a consequence of time-dependences.

Lemma 5.1 *Given a DAE (2.1) with properly stated leading term. Let $t \in \mathcal{I}$ be fixed.*

(i) For each $x \in \mathcal{M}_0(t)$ there is a uniquely determined $z \in \mathbb{R}^\ell$ such that

$$z = R(t)z, \quad A(t)z + B(t)x = q(t). \tag{5.3}$$

(ii) If the DAE (2.1) has index 2, and x is taken from $\mathcal{M}_1(t)$, then (5.3) leads to

$$\begin{aligned} z &= -(DP_1G_2^{-1}BD^-)(t)u + (DP_1G_2^{-1}q)(t) \\ &\quad +(DQ_1D^-(DP_1D^-)')(t)u + (DQ_1D^-(DQ_1G_2^{-1}q)')(t) \end{aligned} \tag{5.4}$$

with $u := D(t)P_1(t)x$.

Proof:

(i) For $x \in \mathcal{M}_0(t)$, by the definition of $\mathcal{M}_0(t)$, there is a $\tilde{z} \in \mathbb{R}^n$ so that $A(t)\tilde{z} + B(t)x = q(t)$ is fulfilled. Due to $A(t) = A(t)R(t)$, the relations (5.3) are satisfied by $z := R(t)\tilde{z}$. If there are two elements z_1, z_2 satisfying (5.3), it results that $A(t)(z_1 - z_2) = 0$, thus $R(t)(z_1 - z_2) = 0$, i.e., $z_1 = z_2$.

(ii) In the index-2 case, $x \in \mathcal{M}_1(t)$ implies (the argument t is dropped now) $Q_1x = Q_1G_2^{-1}q$ and

$$\begin{aligned} Q_0x &= -Q_0P_1G_2^{-1}BD^-u + Q_0Q_1D^-(DP_1D^-)'u \\ &\quad + Q_0P_1G_2^{-1}q + Q_0Q_1D^-(DQ_1G_2^{-1}q)'. \end{aligned}$$

Using these relations, regarding (i) and inserting (5.4) into (5.3) yields the assertion.

□

For linear systems, contractivity is simply related to homogeneous equations. Letting $q = 0$, the solutions given by (5.1) are

$$x(t) = P_{\mathrm{can}}(t)D(t)^-u(t). \tag{5.5}$$

Definition 5.2 *The DAE (2.1) is said to be contractive or to have a contractive inner flow if there are an inner product $< \cdot, \cdot >$ and a constant $\beta \geq 0$ such that the inequality*

$$\langle z + R'(t)D(t)x, D(t)x\rangle \leq -\beta|D(t)x|^2 \tag{5.6}$$

is satisfied for all $x \in \mathbb{R}^m$, $z \in \mathbb{R}^\ell$ with

$$x \in \mathcal{M}_1(t), \quad z = R(t)z, \quad A(t)z + B(t)x = 0, \tag{5.7}$$

and all $t \geq 0$.

Remark: Applying Lemma 5.1 we know (5.7) to determine $z = -DP_1G_2^{-1}BD^-u + DQ_1D^-(DP_1D^-)'u$ with $u := DP_1x$. Further, $Q_1x = 0$, thus $Dx = DP_1x = u \in DS_1$ are valid. Derive further that

$$\begin{aligned} z + R'Dx &= -DP_1G_2^{-1}BD^-u + DQ_1D^-(DP_1D^-)'u + R'Dx \\ &= -DP_1G_2^{-1}BD^-u - (DP_1D^-)'u + (DQ_1D^- + DP_1D^-)'Dx \\ &= -DP_1G_2^{-1}BD^-u + (DP_1D^-)'u, \end{aligned}$$

i.e., the contractivity inequality (5.6), (5.7) is actually the contractivity inequality for the IRODE on the subspace DS_1, namely

$$\langle -(DP_1G_2^{-1}BD^-)(t)u + (DP_1D^-)'(t)u, u\rangle \leq -\beta|u|^2$$

for $u \in (DS_1)(t)$. □

Consequently, contractivity means in fact the contractivity of the inner flow on the invariant subspace DS_1.
For the special case of $\ell = m$, and supposing that $D = R, R' = 0$, this contractivity notion, i.e., Definition 5.2, is given in [10].
Since we operate in different spaces $\mathbb{R}^\ell$ and $\mathbb{R}^m$, we have to consider the contractivity norm on $\mathbb{R}^\ell$, but also a norm on $\mathbb{R}^m$. We denote both, the $\mathbb{R}^\ell$-norm and the $\mathbb{R}^m$-norm, by $|.|$, further the induced matrix norms by $\|.\|$.
Additionally, for a given subspace $L \subset \mathbb{R}^\ell$ resp. $L \subset \mathbb{R}^m$, we use the matrix semi-norm

$$\|M\|_L := \max_{\substack{z\in L\\|z|=1}} |Mz|, \quad M \in L(\mathbb{R}^\ell, \mathbb{R}^m) \text{ resp. } M \in L(\mathbb{R}^m, \mathbb{R}^\ell).$$

Now, by standard arguments the following can be proved.

Theorem 5.3 *Let the index-2 DAE (2.1) be contractive. Then, for any two solutions $x, \bar{x} \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ and their components $u = DP_1x$, $\bar{u} = DP_1\bar{x}$, the inequalities*

$$|u(t) - \bar{u}(t)| \leq e^{-\beta(t-t_*)}|u(t_*) - \bar{u}(t_*)|, \tag{5.8}$$

$$|x(t) - \bar{x}(t)| \leq \|P_{\mathrm{can}}(t)D(t)^-\|_{(DS_1)(t)}|u(t) - \bar{u}(t)|, \tag{5.9}$$

and

$$|x(t) - \bar{x}(t)| \leq \\ e^{-\beta(t-t_*)}\|P_{\mathrm{can}}(t)D(t)^-\|_{(DS_1)(t)}\|D(t_*)P_1(t_*)\|_{\mathrm{im}\ P_{\mathrm{can}}(t)}|x(t_*) - \bar{x}(t_*)| \tag{5.10}$$

are satisfied for all $t \geq t_* \geq 0$.

As the next theorem will show, algebraically stable IRK(DAE) reflect the contractivity of the inner and total flows well, provided the DAE is numerically well formulated.

Theorem 5.4 *Let the index-2 DAE (2.1) be contractive and have constant subspaces $D(t)N_1(t)$, $D(t)S_1(t)$. Then an algebraically stable IRK(DAE) applied to (2.1), starting with $x_0, \bar{x}_0 \in \mathcal{M}_1(t_0)$, yields*

$$\begin{aligned} |u_n - \bar{u}_n| &\leq |u_{n-1} - \bar{u}_{n-1}|, \\ |x_n - \bar{x}_n| &\leq \|P_{\mathrm{can}}(t_n)D(t_n)^-\|_{(DS_1)(t_n)}|u_n - \bar{u}_n| \end{aligned}$$

and

$$|x_n - \bar{x}_n| \leq \|(P_{\mathrm{can}}D^-)(t_n)\|_{(DS_1)(t_n)}\|(DP_1)(t_{n-1})\|_{\mathrm{im}\ P_{\mathrm{can}}(t_{n-1})}|x_{n-1} - \bar{x}_{n-1}|$$

for $n \geq 0$, without any stepsize restriction.

Proof: The decoupling and the discretisation commute. Hence, the given IRK acts on the contractive IRODE. □

Concerning dissipativity, the situation is slightly more complicated.

Definition 5.5

(i) *The DAE (2.1) is called dissipative if there is a bounded positively invariant set $\mathfrak{B}(t) \subset \mathcal{M}_1(t), \quad t \geq 0$, so that, for any $t_* \geq 0$ and any bounded set $E \subset \mathcal{M}_1(t_*)$ there is a $t_{(E,t_*)} \geq t_*$ such that $x_* \in E$ implies $x(t, t_*, x_*) \in \mathfrak{B}(t)$ for all $t \geq t_{(E,t_*)}$.*
The possibly time-varying set $\mathfrak{B}(t), t \geq 0$, which sucks up the solutions, is said to be an absorbing set.

(ii) *The DAE (2.1) has a dissipative inner flow if the IRODE is dissipative on DS_1, i.e., if there is a bounded positively invariant set $\mathfrak{B}(t)_{\mathrm{IRODE}} \subset D(t)S_1(t), \; t \geq 0$, such that, for any $t_* \geq 0$ and any bounded set $E \subset D(t_*)S_1(t_*)$, there exists a $t_{(E,t_*)} \geq t_*$, and $u_* \in E$ implies $u(t, t_*, u_*) \in \mathfrak{B}(t)_{\mathrm{IRODE}}$ for all $t \geq t_{(E,t_*)}$.*

First, let us consider an index-2 DAE (2.1) that has a right-hand side q satisfying

$$DQ_1G_2^{-1}q = 0. \tag{5.11}$$

For those equations, the existence of an inner product $< .,. >$ as well as of constants $\bar{\alpha} \geq 0, \; \bar{\beta} > 0$ such that

$$x \in \mathcal{M}_1(t), \; z = R(t)z, \; A(t)z + B(t)x = q(t), \;\; t \geq 0, \tag{5.12}$$

implies the inequality

$$\langle z + R'(t)D(t)x, D(t)x \rangle \leq \bar{\alpha} - \bar{\beta}|D(t)x|^2, \tag{5.13}$$

leads (via Lemma 5.1) to the dissipativity inequality for the inner flow, i.e.,

$$\langle -(DP_1G_2^{-1}BD^- - (DP_1D^-)')(t)u + (DP_1G_2^{-1}q)(t), u \rangle \leq \bar{\alpha} - \bar{\beta}|u|^2 \tag{5.14}$$

holds for all $u \in D(t)S_1(t), \; t \geq 0$, and vice versa.

By analogous arguments as used in [12, §3.2] we obtain the following assertion.

Theorem 5.6 *Let an index-2 DAE (2.1) with (5.11) be given. Let (5.12) imply (5.13)*

(i) *Then the inner flow is dissipative with absorbing sets*

$$\mathfrak{B}(t)_{\mathrm{IRODE}} = \left\{ v \in D(t)S_1(t) : |v|^2 \leq \frac{\bar{\alpha}}{\bar{\beta}} + \varepsilon \right\}, \quad \varepsilon > 0.$$

(ii) If there are uniform bounds

$$\|P_{\mathrm{can}}(t)D(t)^-\|_{D(t)S_1(t)} \le \kappa_1, \ \ |(Q_0P_1G_2^{-1}q)(t)| \le \kappa_2, \ \ t \ge 0,$$

the total DAE-flow is dissipative with absorbing sets

$$\mathfrak{B}(t) = \big\{y \in \mathcal{M}_1(t) : |y| \le \kappa_1\big(\frac{\bar{\alpha}}{\bar{\beta}} + \varepsilon\big)^{\frac{1}{2}} + \kappa_2\big\}.$$

(iii) If, additionally, the subspaces $D(t)N_1(t)$ and $D(t)S_1(t)$ remain constant, then the backward Euler method reflects both the inner and the total dissipativity behaviour well without any stepsize restriction. The absorbing sets of the numerical flow are the same as described in (i) and (ii).

In the general case, if (5.11) is not given, the relations (5.12) lead to nontrivial components $D(t)Q_1(t)x = D(t)Q_1(t)G_2(t)^{-1}q(t)$. By shifting the unknown function

$$x(t) - P_0(t)Q_1(t)G_2(t)^{-1}q(t) =: \tilde{x}(t)$$

we may turn to the new index-2 DAE

$$A(t)(D(t)\tilde{x}(t))' + B(t)\tilde{x}(t) = \tilde{q}(t) \tag{5.15}$$

with $\tilde{q} := q - A(DQ_1G_2^{-1}q)' - BP_0Q_1G_2^{-1}q$, so that $DQ_1G_2^{-1}\tilde{q} = 0$ is satisfied. Note that it holds that $(DP_1x)(t) = (DP_1\tilde{x})(t)$, $(Q_0x)(t) = (Q_0\tilde{x})(t)$. The DAEs (2.1) and (5.15) have a common IRODE, thus their inner flows coincide. If this inner flow is dissipative with absorbing sets

$$\mathfrak{B}(t)_{\mathrm{IRODE}} = \big\{v \in D(t)S_1(t) : |v|^2 \le \frac{\bar{\alpha}}{\bar{\beta}} + \varepsilon\big\}, \ \ \varepsilon > 0, \tag{5.16}$$

the total flow of (2.1) is also dissipative, supposed that there are the above uniform bounds κ_1, κ_2 as well as

$$|(DQ_1G_2^{-1}q)(t)| \le \kappa_3, \ \ |(Q_0Q_1D^-(DQ_1G_2^{-1}q)')(t)| \le \kappa_4, \ \ t \ge 0. \tag{5.17}$$

Then, the DAE (2.1) has the absorbing sets

$$\mathfrak{B}(t) = \big\{y \in \mathcal{M}_1(t) : |y| \le \kappa_1\big(\frac{\bar{\alpha}}{\bar{\beta}} + \varepsilon\big)^{\frac{1}{2}} + \kappa_2 + \kappa_3 + \kappa_4\big\}, \varepsilon > 0. \tag{5.18}$$

What concerns the numerical flow determined by the implicit Euler method, for achieving commutativity of decoupling and discretisation, it is now assumed that the subspaces $D(t)S_1(t)$ and $D(t)N_1(t)$ do not vary with t.
Since, in the general case, the implicit Euler method generates values $x_n \in \mathcal{M}_{1\,\mathrm{num}}(t_n)$, i.e., the total numerical flow is located on $\mathcal{M}_{1\,\mathrm{num}}(t_n)$, when describing the absorbing sets of the total numerical flow one has to take into account a bound

$$|(Q_0Q_1D^-)(t_n)\{[DQ_1G_2^{-1}q]'_n - (DQ_1G_2^{-1}q)'_n\}| \le \tau, \ \ n \ge 1, \tag{5.19}$$

which, for small $\tau > 0$, may lead to a stepsize restriction.
Sometimes it might be easier to check contractivity than dissipativity inequalities.

Theorem 5.7 *Let a strictly contractive ((5.6) with $\beta > 0$) index-2 DAE (2.1) be given and let the function $p := DP_1G_2^{-1}q + (DP_1D^-)'DQ_1G_2^{-1}q$ be uniformly bounded, i.e.,*

$$|p(t)| \leq \gamma, \quad t \geq 0.$$

(i) Then the inner flow is dissipative with absorbing sets (5.16), where $\bar{\alpha} := \frac{\gamma^2}{2\beta}$, $\bar{\beta} := \frac{\beta}{2}$.

(ii) If the above uniform bounds κ_1, κ_2, κ_3, κ_4 exist, then the DAE flow is dissipative with absorbing sets (5.17).

(iii) If, additionally, the subspaces $D(t)S_1(t)$ and $D(t)N_1(t)$ do not depend on t, the implicit Euler method reflects the dissipativity behaviour of the inner flow well without any stepsize restriction and with the same absorbing sets. For the numerical total flow, there are absorbing sets

$$\mathfrak{B}(t_n)_{\text{num}} = \big\{y \in \mathcal{M}_{1\text{num}}(t_n) : |y| \leq \kappa_1\big(\frac{\bar{\alpha}}{\bar{\beta}} + \varepsilon\big)^{\frac{1}{2}} + \kappa_2 + \kappa_3 + \kappa_4 + \tau\big\}.$$

Proof: It remains to show the first assertion. Let (5.6) be given with $\beta > 0$. Then, the IRODE is contractive with the same inner product and $\beta > 0$. Denote shortly $-DP_1G_2^{-1}BD^- + (DP_1D^-)' =: M$ so that the IRODE reads $u' = Mu + p$. For each $u \in D(t)S_1(t)$, due to the contractivity, it holds that

$$\begin{aligned}
\langle M(t)u + p(t), u\rangle &= \langle M(t)u, u\rangle + \langle p(t), u\rangle \\
&\leq -\beta|u|^2 + |p(t)|\,|u| \leq -\beta|u|^2 + \gamma|u| \\
&\leq -\tfrac{\beta}{2}|u|^2 + \gamma|u| - \tfrac{\beta}{2}|u|^2.
\end{aligned}$$

With $\bar{\alpha} := \frac{\gamma^2}{2\beta}$ we conclude

$$\begin{aligned}
\gamma|u| - \tfrac{\beta}{2}|u|^2 - \bar{\alpha} &= -\tfrac{\beta}{2}\big(|u|^2 - 2\tfrac{\gamma}{\beta}|u| + \tfrac{\gamma^2}{\beta^2}\big) \\
&= -\tfrac{\beta}{2}\big(|u| - \tfrac{\gamma}{\beta}\big)^2 \leq 0,
\end{aligned}$$

hence, $\gamma|u| = \frac{\beta}{2}|u|^2 \leq \bar{\alpha}$, i.e., the IRODE satisfies a dissipativity inequality. □

6 Conclusions

For a successful and effective numerical integration of DAEs it is not only important to choose an appropriate numerical method but also an appropriate formulation of the DAE. This is necessary in order to guarantee that the inner flow is approximated properly, i.e., that the numerical method retains its stability behaviour as known for explicit ODEs. Otherwise, one is confronted with artificial stepsize restrictions since implicit methods may behave like explicit

ones, or, even worse, the method may fail completely.
For an appropriate DAE formulation, one should exploit the modelling procedure, first of all in such a way that the modelled system is described by DAEs with a properly formulated leading term (see (2.1)). Such a representation is often naturally given (e.g. in circuit simulation, multibody dynamics). Furthermore, for a numerical well formulated index-2 DAE, the spaces DN_1 and DS_1 have to be constant. If only DN_1 is constant, the inner flow is still reflected properly but the index-2 components of the numerical solution are influenced by additional perturbations as described in Theorem 4.6. If only DS_1 is constant, the inner flow is slightly perturbed as described in Theorem 4.8.
If both spaces DN_1 and DS_1 vary with time, one should check condition (4.6). If it is satisfied, then the current problem is numerically equivalent to a numerically well formulated one, i.e., its formulation is appropriate. However, if condition (4.6) is not satisfied, then one should review the modelling procedure and reformulate the problem to a numerically well formulated one.

A Convergence

Consider an index-2 DAE (2.1) on the compact interval $\mathcal{I} = [t_0, T]$. Let x_n now denote the values actually generated by the k-step BDF along the partition $t_n = t_n + h$, i.e., x_n satisfies

$$A_n[Dx]'_n + B_n x_n - q_n = \delta_n, \quad n \geq k, \tag{A.1}$$

with an error δ_n resulting from rounding-off. Let $x_0, \ldots, x_{k-1}$ denote the starting values, $x(t_n),\ n \geq 0$, the true solution values to be approximated, and $\varepsilon_n := x(t_n) - x_n, \quad n \geq 0$, the global error. Introduce the local error

$$\tau_n := A_n[(Dx)]'_n + B_n x_n - q_n = A_n\{[(Dx)]'_n - (Dx)'_n\},\ \ n \geq k,$$

with $[(Dx)]'_n := \frac{1}{h}\sum\limits_{j=0}^{k} \alpha_j (Dx)(t_{n-j})$. Notice that $(Q_1 G_2^{-1})_n \tau_n = 0,\ \ n \geq k$.
The resulting global error recursion

$$A_n[D\varepsilon]'_n + B_n \varepsilon_n = \tau_n - \delta_n, \quad n \geq k, \tag{A.2}$$

can be decoupled along the lines of §3. The decoupled system corresponds to (3.6), (3.7), (3.8), where we have to replace x_n by ε_n and $q(t_n)$ by $\tau_n - \delta_n$. With the denotation

$$\delta_n := -G_2^{-1}\varepsilon_n, \quad n = 0, \ldots, k-1,$$

we obtain $[DQ_1x]'_n = -[DQ_1G_2^{-1}\delta]'_n$. Now standard arguments apply to the recursion corresponding to (3.6) for the components $(DP_1)_n\varepsilon_n$. Finally, we derive the error estimation

$$\begin{aligned} \max_{l=k,\ldots,n} |\varepsilon_l| \quad \leq \quad & S\big\{ \max_{l=0,\ldots,k-1} |D_l \varepsilon_l| + \max_{l=k,\ldots,n} |\tau_l \delta_l| \\ & + \max_{l=0,\ldots,n} \big|\tfrac{1}{h}(DQ_1G_2^{-1})_n\delta_n\big| \big\}, \end{aligned} \tag{A.3}$$

which holds true for all $t_n \leq T$ with a stability bound S independent of the stepsize h. Notice that this inequality (A.3) indicates a weak instability caused by the inherent differentiation, which becomes apparent for small stepsizes h.

As a consequence of the error inequality (A.3), if the roundoff errors are dropped and exact starting values are used (i.e., $\delta_n = 0, n \geq 0$), it results that

$$\max_{l\geq 0} |\varepsilon_l| \leq S \max_{l \geq k} |\tau_l|,$$

i.e., the BDF converges for $h \to 0$ with the same order as it does in the regular ODE case.
Similar arguments apply to IRK(DAE).

B Technical Details

Lemma B.1 *The following properties hold:*

(i) *If two projectors U and V have the same image space, then $UV = V$ and $VU = U$. If they have the same kernel, then $UV = U$, $VU = V$.*

(ii) *Given a matrix M and a projector U. If* $\operatorname{im} M = \operatorname{im} U$, *then $UM = M$. If* $\ker M = \ker V$, *then $MV = M$.*

(iii) *Given two smooth matrices $M(t)$ and $U(t)$ such that* $\operatorname{im} U(t)$ *is constant and* $\operatorname{im} U(t) \subseteq \ker M(t)$, *then $M(t)(U(t)x)' = 0$, $M(t)'U(t) = 0$, $M(t)U'(t) = 0$, $M_n[Ux]'_n = 0$, $M_{ni}[Ux]'_{ni} = 0$.*

(iv) *Given a smooth projector $M(t)$ and a matrix $U(t)$ such that* $\operatorname{im} M(t) = \operatorname{im} U(t)$ *and* $\operatorname{im} U(t)$ *is a constant subspace, then $M(t)(U(t)x)' = (U(t)x)'$, $M(t)'U(t) = 0$, $M_n[Ux]'_n = [Ux]'_n$, $M_{ni}[Ux]'_{ni} = [Ux]'_{ni}$.*

Proof: Parts (i) and (ii) are straightforward. Consider part (iii). We can choose a constant projector V onto $\operatorname{im} U(t)$. Thus $VU(t) = U(t)$ and $M(t)(U(t)x)' = M(t)(VU(t)x)' = M(t)V(U(t)x)' = 0$. Similarly one gets $M(t)'U(t) = 0$, $M(t)U(t)' = 0$. Finally, $M_n[Ux]'_n = M_n[VUx]'_n = M_nV[Ux]'_n = 0$, and similarly for $M_{ni}[Ux]'_{ni}$. Part (iv) can be proved analogously by choosing a constant projector V onto $\operatorname{im} M(t)$ satisfying $M(t)V = V$, $VU(t) = U(t)$. □

DAEs and projectors. Choose Q_0 to be a projector along $\ker AD$, and denote $P_0 = I - Q_0$. Define the generalized inverse D^- of D with the properties $DD^- = R$, $D^-D = P_0$.
Let $Q_1(t)$ be a projector onto $N_1(t)$ along $S_1(t)$ and denote $P_1 = I - Q_1$. By construction it holds that $Q_1 = Q_1G_2^{-1}BP_0$. For these projectors, the following properties hold: $D = DP_0$, $DQ_0 = 0$, $D^- = P_0D^-$, $Q_0D^- = 0$, $Q_1Q_0 = 0$, $P_1Q_0 = Q_0$, $Q_0P_1D^- = -Q_0Q_1D^-$, $DP_1Q_0 = 0$, $DP_1P_0 = DP_1$, ${}_1 = Q_1P_0$. Moreover, DP_1D^- and DQ_1D^- are projectors, and $\operatorname{im} DP_1D^- = \operatorname{im} DP_1 = DS_1$, $\operatorname{im} DQ_1D^- = \operatorname{im} DQ_1 = DN_1$.

Lemma B.2 *Let us assume that DN_1 is a constant subspace. We denote $H := I - DQ_1D^- - P_{DN_1}$ with P_{DN_1} a constant projector onto DN_1, then*

$$\begin{aligned} DP_1D^-HH' &= 0, && \text{(B.1)}\\ (DP_1D^-H)_n[HDx]'_n - (DP_1D^-)_n[Dx]'_n &= 0, && \text{(B.2)}\\ Q_0Q_1D^-HH' &= Q_0Q_1D^-(DQ_1D^-)', && \text{(B.3)}\\ (Q_0Q_1D^-H)_n[HDx]'_n &= (Q_0Q_1D^-)_n[DQ1x]'_n. && \text{(B.4)} \end{aligned}$$

Similar expressions hold when we put ni instead of n.

Proof: As $H' = -(DQ_1D^-)'$, im $DQ_1D^- = DN_1$ is a constant subspace and $(DP_1D^-H)(DQ_1D^-) = 0$, we can apply part (iii) of Lemma B.1 to obtain (B.1). As $Q_0Q_1D^-H = -Q_0Q_1D^-P_{DN_1}$, we obtain (B.3). As im $HDQ_1 =$ im DQ_1 is a constant subspace, and $(DP_1D^-H)(HDQ_1) = 0$, part (iii) of Lemma B.1 gives $(DP_1D^-H)_n[HDx]'_n = (DP_1D^-H)_n[HDP_1x]_n$. We simply have to observe that $DP_1D^-H = DP_1D^-$, and $HDP_1 = (I - P_{DN_1})DP_1$ to obtain

$$(DP_1D^-H)_n[HDP_1x]'_n = (DP_1D^-)_n[(I-P_{DN_1})DP_1x]'_n = (DP_1D^-)_n[DP_1x]'_n.$$

Applying again part (iii) of Lemma B.1, part (B.2) follows. As im $HDP_1 =$ im $(I - P_{DN_1})D$ is constant, and $(Q_0P_1D^-H)(HDP_1) = 0$, Lemma B.1 gives

$$(Q_0Q_1D^-H)_n[HDx]'_n = (Q_0Q_1D^-H)_n[HDQ_1x]'_n\,.$$

As $Q_0Q_1D^-H = -Q_0Q_1D^-P_{DN_1}$ and $HDQ_1 = -P_{DN_1}DQ_1 = -DQ_1$, computations similar to the ones above lead to (B.4). $\square$

Lemma B.3 *Let us assume that DS_1 is a constant subspace. We denote $H := I - DP_1D^- - P_{DS_1}$ with P_{DS_1} a constant projector onto DN_1, then*

$$\begin{aligned} DP_1D^-HH'D &= (DP_1D^-)'DQ_1 && \text{(B.5)}\\ (DP_1D^-H)_n[HDx]'_n &= (DP_1D^-)_n[DP_1x]'_n, && \text{(B.6)} \end{aligned}$$

$$\begin{aligned} Q_0P_1D^-HH' &= 0 && \text{(B.7)}\\ (Q_0P_1D^-H)_n[HDx]'_n - (Q_0P_1D^-)_n[Dx]'_n &= 0. && \text{(B.8)} \end{aligned}$$

Similar expressions hold when we put ni instead of n.

Proof: Since $DP_1D^-H = -P_{DS_1}$ and $H' = -(DP_1D^-)'$, the relation (B.5) follows by part (iv) of Lemma B.1. As im $HDQ_1 =$ im $(I - P_{DS_1})D$ is a constant subspace, and $(DP_1D^-H)(HDQ_1) = 0$, applying part (iii) of Lemma B.1 we obtain

$$(DP_1D^-H)_n[HDx]'_n = (DP_1D^-H)_n[HDP_1x]'_n.$$

We simply have to observe that $DP_1D^-H = -DP_1D^-P_{DS_1}$, and $P_{DS_1}HDP_1 = -DP_1$ to obtain (B.6). Since $Q_0Q_1D^-H = Q_0Q_1D^-(I - P_{DS_1})$, part (B.7) follows from

$$Q_0Q_1D^-HH' = Q_0Q_1D^-(I - P_{DS_1})(DP_1D^-)' = 0.$$

Now as im $HDP_1 = DS_1$ is constant, and $(Q_0 Q_1 D^- H)(HDP_1) = 0$, part (iii) of Lemma B.1 gives

$$(Q_0 Q_1 D^- H)_n [HDx]'_n = (Q_0 Q_1 D^- H)_n [HDQ_1 x]'_n = (Q_0 Q_1 D^-)_n [DQ_1 x]'_n.$$

Using part (iii) of Lemma B.1, we obtain

$$(Q_0 Q_1 D^- H)_n [HDx]'_n - (Q_0 Q_1 D^-)_n [Dx]'_n = -(Q_0 Q_1 D^-)_n [DP_1 x]'_n = 0$$

which implies (B.8) if one regards $Q_0 Q_1 D^- = -Q_0 P_1 D^-$. □

References

[1] U.M. Ascher, L.R. Petzold: Stability of computational methods for constrained dynamics systems. SIAM J. SCIC (14) 1993, 95-120.

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

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

[4] E. Eich-Soellner and C. Führer: Numerical Methods in Multibody Dynamics. B.G.Teubner, Stuttgart, 1998.

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

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

[7] C.W. Gear, L.R. Petzold: ODE Methods for the Solution of Differential/Algebraic Systems. SIAM J. Numer. Anal. 21 (1984) 716–728.

[8] M. Günther, P. Rentrop: Numerical simulation of electrical circuits. University of Karlsruhe, IWRMM, Preprint 01/01.

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

[10] M. Hanke, E. Izquierdo Macana and R. März: On asymptotics in case of linear index-2 differential-algebraic equations. SIAM J. Numer. Anal. 4 (35) 1998, 1326–1346.

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

[12] I. Higueras, R. März, and C. Tischendorf: Numerically well formulated index-1 DAEs. Humboldt-Universität Berlin, Institut für Mathematik, Preprint 2001–05.

[13] R. März: Index-2 differential–algebraic equations. Results in Mathematics (15) 1989, 148–171.

[14] R. März: Differential algebraic systems anew. Humboldt-Universität Berlin, Institut für Mathematik, Preprint 2000-21.

[15] R. März and A. Rodrígues Santiesteban: Analyzing the stability behaviour of DAE solutions and their approximations. Humboldt-Univ. Berlin, Institut für Mathematik, Preprint 99-2. To appear in Math. Comp.

[16] A. M. Stuart and A. R. Humphries: Dynamical systems and numerical analysis. Cambridge University Press, Cambridge, UK, 1998.