Numerically well formulated index-1 DAEs

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

Abstract

For index-1 DAEs with properly stated leading term, we characterize dissipative and contractive flows and study how the qualitative properties of the DAE solutions are reflected by numerical approximations. The best situation occurs when the discretization and the decoupling procedure commute. It turns out out that this is the case if the relevant part of the inherent regular ODE has a constant state space. Different kinds of reformulations are studied to obtain numerically well formulated systems. Those reformulations might be expensive, hence, in order to avoid them, criteria ensuring the given DAE to be numerically equivalent to a numerically well formulated representation are proved.

Key words: differential algebraic equations, numerical integration methods, global stability

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

1 Introduction

An important topic in the development of numerical methods for ordinary differential equations (ODEs) is the study of the numerical solution over large intervals when the system has a specific dynamics. For example, algebraically stable methods are known to be B-stable for contractive problems, i.e., the contractivity property is preserved without any stepsize restriction (e.g. [12]).
For differential algebraic equations (DAEs), different authors have studied the qualitative behaviour of the solutions and their numerical counterparts. In particular, in [5], index-1 DAEs

$$A(x(t),t)x'(t) + b(x(t),t) = 0 \tag{1.1}$$

are considered and the concept of B-stability is extended. Algebraically stable, stiffly accurate methods are proved to be B-stable provided that the leading coefficient $A(x,t)$ has a constant nullspace. However, if this nullspace actually varies, strong additional stepsize restrictions may appear even with algebraically stable methods ([7]). For some DAEs, a certain reformulation of the problem may avoid those stepsize restrictions ([4]). For linear index-2 DAEs that have a constant leading nullspace, contractivity

[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

and B-stability notions are introduced in [6] and [10] respectively. It is shown that contractivity is preserved without any stepsize restriction supposed that certain further subspaces associated with the DAE are constant. Again, subspace movements may cause stepsize restrictions. First results concerning nonlinear index-2 DAEs are given in [11].

In order to find out the source of these stepsize restrictions, we should regard how the standard methods for regular ODEs have been adapted to solve DAEs of the form (1.1). Usually, the derivative $x'(t)$ is approximated as if (1.1) were a regular implicit ODE, but this is not the case. Proceeding in this way, we use the numerical derivatives not only for the components that have a dynamical behaviour, but also for those which are obtained from algebraic relationships. In some sense, problem (1.1) has not yet been formulated properly. In [1, 8], it is proposed to figure out DAEs as equations

$$A(x(t),t)(D(t)x(t))' + b(x(t),t) = 0, \tag{1.2}$$

with coefficients A and D well matched in a certain sense, i.e., with a properly stated leading term that catches precisely the derivatives that are actually involved. In [8] it is shown that the dynamics of an index-1 problem (1.2) is governed by an inherent regular ODE (IRODE) that is uniquely determined by the problem data. The solution can be decoupled into a dynamic part, a solution of this IRODE, and an algebraic part, obtained from an algebraic equation. In order to reproduce numerically the problem dynamics, the numerical solution should also have the same decoupling. In particular, although the IRODE is not explicitly available in practice, inside, the numerical method should integrate this IRODE.

The IRODE

$$u'(t) = \varphi(u(t),t) \tag{1.3}$$

is relevant for our problem along a possibly time-varying invariant subspace $U(t)$. Suppose $P_U(t) = P_U(t)^2$ to be a certain projector along $U(t)$ (i.e. $\ker P_U(t) = U(t)$), the initial condition $u(t_0) \in U(t_0)$ implies $u(t) \in U(t)$ for all t, and it holds that

$$P_U'(t)u(t) + P_U(t)\varphi(u(t),t) = 0. \tag{1.4}$$

Therefore, we should ensure by means of appropriate conditions that $u_0 \in U(t_0)$ implies $u_n \in U(t_n)$ for all n, where u_n is the numerical solution for (1.3) generated by a Runge-Kutta method or a BDF. In the index-1 case, a very transparent condition can be realized, namely, $U(t)$ should not depend on t, i.e., it should be constant. In this case, we can choose a constant additional projector $\tilde{P}_U$ such that $\ker \tilde{P}_U = \ker P_U(t) = U(t)$. Then $P_U(t)\tilde{P}_U = P_U(t)$, $\tilde{P}_U P_U(t) = \tilde{P}_U$ are valid, and thus $\tilde{P}_U P_U'(t) = 0$. Multiplying (1.4) by $\tilde{P}_U$ yields

$$\tilde{P}_U\varphi(u(t),t) = 0.$$

Therefore, given the numerical solution of (1.3) with any Runge-Kutta method,

$$u_{n+1} = u_n + h\sum_{j=1}^{s} b_j\varphi(U_{nj}, t_{nj}) ,$$

we obtain $\tilde{P}_U u_{n+1} = \tilde{P}_U u_n$ immediately, i.e., $u_0 \in U(t_0)$ implies $u_1 \in U(t_1)$ and so on. The same is true for the BDF with starting values from this subspace.
Hence, from the numerical point of view, DAEs with a properly stated leading term and an IRODE that has a relevant constant invariant subspace are nice problems to be solved. Actually, in [8] the subspace im $D(t)$ is pointed out to be the relevant invariant subspace for index-1 DAEs (1.2), and problems with constant im $D(t)$ are said to be numerically well formulated index-1 DAEs. Now, the good index-1 situation studied in [5] appears to be a special case by realizing that (1.1) is numerically equivalent to

$$A(x(t), t)(Px(t))' + b(x(t), t) = 0$$

if $A(x,t)$ has a constant nullspace and $P = P^2$, $\ker P = \ker A(x,t)$. Here we put $D = P$, $\tilde{P}_U = I - P$. Moreover, the reformulation of index-1 equations

$$A(t)x'(t) + b(x(t), t) = 0$$

as

$$(A(t)x(t))' + b(x(t), t) - A'(t)x(t) = 0 \tag{1.5}$$

used in [4] is a close approach. If $A(t)$ has a constant image, with a constant projector R onto im $A(t)$, equation (1.5) will be trivially rewritten as

$$R(A(t)x(t))' + b(x(t), t) - RA'(t)x(t) = 0 ,$$

but this will again be a numerically well formulated DAE of type (1.2). This explains the positive results obtained for constant im $A(t)$.

In the present paper we study the numerical solution of general index-1 DAEs (1.2) with properly stated leading term over long intervals when the DAE has a specific dynamics.

In Section 2 we summarize some results from [8] concerning properly stated leading terms and numerically well formulated DAEs. In Section 3 we characterize dissipative and contractive flows induced by the DAE (1.2), and study how the qualitative properties of the DAE solutions are reflected by their numerical approximations. It turns out that our problems should be given as numerically well formulated ones. However, mostly DAEs are given in the form (1.1), or as (1.2) with time-varying im $D(t)$. How to obtain appropriate reformulations, of course with additional expense, is discussed in Section 4. Sometimes there is no need for reformulations since the original DAE is - although itself not numerically well formulated – numerically equivalent to a version that is numerically well formulated. These cases are also studied in Section 4.

Section 5 contains a straightforward generalization of the DAE (1.2) that allows rectangular matrices A and D, but also a nonlinearity $d(x(t), t)$ instead of $D(t)x(t)$. Furthermore, it is shown how the DAEs resulting from circuit simulation by the modified nodal analysis fit into this form. The Appendix contains two basic linear algebra assertions used frequently.

2 DAEs with properly stated leading term and numerically well formulated DAEs

In [1, 8], DAEs with properly formulated leading terms are introduced as equations

$$A(x(t), t)(D(t)x(t))' + b(x(t), t) = 0, \quad t \in \mathcal{I}, \tag{2.1}$$

where the matrix coefficients $A(x, t) \in L(\mathbb{R}^m)$ and $D(t) \in L(\mathbb{R}^m)$ are well matched.

Let the given functions $A(x, t)$, $D(t)$, $b(x, t)$ depend continuously on their arguments and let the continuous partial derivatives $A'_x(x, t), b'_x(x, t)$ exist. For brevity, we sometimes write

$$f((D(t)x(t))', x(t), t) = 0$$

with $f(y, x, t) := A(x, t)y + b(x, t)$.

Definition 2.1 *The leading term of (2.1) is properly stated if*

$$\ker A(x, t) \oplus \text{im } D(t) = \mathbb{R}^m,$$

and if there is a continuously differentiable projector function $R(t) \in L(\mathbb{R}^m)$ such that

$$\ker A(x, t) = \ker R(t), \quad \text{im } D(t) = \text{im } R(t).$$

Then, the matrix coefficients $A(x, t)$ and $D(t)$ are said to be well matched.

Note that the nullspace of $A(x, t)$ is then independent of x. Both subspaces $\ker A(x, t)$ and $\text{im } D(t)$ have constant dimension. It holds that

$$A(x, t) = A(x, t)R(t), \quad D(t) = R(t)D(t), \quad f(y, x, t) = f(R(t)y, x, t).$$

Observe that for each continuous function $x(\cdot)$ that has a continuously differentiable component $D(\cdot)x(\cdot)$ and satisfies (2.1), the relation $x(t) \in \mathcal{M}_0(t)$ with

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

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

$$C_D^1 := \{x(\cdot) \in C : D(\cdot)x(\cdot) \in C^1\}$$

denote the respective function space. Introduce the leading nullspace (see Lemma 6.1, Appendix)

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

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

$$S_0(y,x,t) := \{z \in \mathbb{R}^m : f'_x(y,x,t)z \in \text{im } A(x,t)\}$$

coincides with the tangent space $T_x\mathcal{M}_0(t)$ for $x \in \mathcal{M}_0(t)$ and $y \in$ im $D(t)$ such that $A(x,t)y + b(x,t) = 0$.

Definition 2.2 *The equation (2.1) is an index-1 DAE if $S_0(y,x,t) \cap N_0(t) = \{0\}$, i.e., these subspaces intersect transversally on the definition domain.*

In the index-1 case, the set $\mathcal{M}_0(t)$ is filled by DAE solutions. Each solution can be expressed as ([8])

$$x(t) = D(t)^- u(t) + Q_0(t)w(u(t),t), \tag{2.2}$$

where $u(t) = D(t)x(t)$ satisfies the inherent regular ODE

$$u'(t) = R'(t)u(t) + D(t)w(u(t),t). \tag{2.3}$$

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

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

The function $w(u,t)$ is implicitly given by the equation

$$0 = f(D(t)w, D(t)^- u + Q_0(t)w, t) =: F(w,u,t).$$

This is ensured by the nonsingularity of $F'_w = f'_y D + f'_x Q_0 = AD + f'_x Q_0$, which is given by the index-1 condition ([8, Lemma 2.1]).

The idea behind this decoupling is the solution decomposition

$$x(t) = P_0(t)x(t) + Q_0(t)x(t) = D(t)^- u(t) + Q_0(t)x(t),$$

but also the collection of the continuous terms $Q_0(t)x(t) + D(t)^- u'(t) =: \omega(t)$ such that $D(t)\omega(t) = R(t)u'(t)$, $Q_0(t)\omega(t) = Q_0(t)x(t)$ and

$$\begin{aligned} 0 &= f((D(t)x(t))', x(t), t) = f(R(t)u'(t), D(t)^- u(t) + Q_0(t)x(t), t) \\ &= f(D(t)\omega(t), D(t)^- u(t) + Q_0(t)\omega(t), t), \end{aligned}$$

i.e., $\omega(t) = w(u(t),t)$, or, equivalently, $Q_0(t)\omega(t) = Q_0(t)w(u(t),t)$ and

$$R(t)u'(t) = D(t)w(u(t),t)\,. \tag{2.4}$$

The inherent regular ODE (2.3) has im $D(t)$ as a time-varying invariant subspace. This ODE is uniquely determined by the problem data. It does not depend of the choice of the projector Q_0 ([8, Remark 2.6]).
Let us stress that the decoupling of the DAE (2.1) into (2.2), (2.3) is rather a theoretical tool for giving an insight into the DAE structure. In practice, this decoupling is not explicitly available. Nevertheless, it would be fine to make sure whether the inherent regular ODE (2.3) will be numerically integrated by an appropriate method if we apply a numerical integration method to the original DAE (2.1). It would be best if the decoupling and the discretization commuted.

Given an s-stage Runge-Kutta method with coefficients $(\mathcal{A}, b^T)$, where $\mathcal{A}$ is nonsingular, $\mathcal{A}^{-1} =: (\alpha_{ij})_{ij=1}^{s}$, and the last row of $\mathcal{A}$ coincides with b^T, the numerical approximation x_n for the real solution value $x(t_n)$ is determined by

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

where the internal stages $X_{ni},\ i = 1, \ldots, s$, are obtained by solving the system

$$f([DX]'_{ni}, X_{ni}, t_{ni}) = 0, \quad i = 1, \ldots, s, \tag{2.5}$$

with the internal derivatives defined as

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

In this way, we have $x_n = X_{ns} \in \mathcal{M}_0(t_n)$. Using a k-step BDF, the approximation x_n to $x(t_n)$ is given by

$$f([Dx]'_n,\ x_n, t_n) = 0, \tag{2.6}$$

where

$$[Dx]'_n := \frac{1}{h} \sum_{i=0}^{k} \alpha_i D_{n-i} x_{n-i}.$$

Obviously, it holds that $x_n \in \mathcal{M}_0(t_n)$. For more clarity, we consider only constant stepsizes, but we do not need this restriction for the results formulated in this paper.

As pointed out in [8], some problems with discretizations are caused by a time-varying subspace im $D(t)$. One can see this by applying the decoupling procedure in parallel to the DAE (2.1) itself and to the discretizations (2.5) or (2.6). However, it may happen that e.g. the implicit Euler method is converted into the explicit Euler method for the inherent regular ODE (2.3) inside. In this case, either strong additional stepsize restrictions must be imposed or wrong asymptotics may result. Recall that this does not concern convergence for $h \to 0$ on compact intervals.

Supposed that the subspace im $D(t)$ does not depend on t, with a constant projector $R_D \in L(\mathbb{R}^m)$ onto im $D(t)$ we obtain

$$\begin{aligned} R(t)(D(t)x(t))' &= R(t)(R_D D(t)x(t))' = R(t)R_D(D(t)x(t))' \\ &= R_D(D(t)x(t))' = (D(t)x(t))', \end{aligned}$$

in (2.4), and

$$R_{ni}[DX]'_{ni} = R_{ni}R_D[DX]'_{ni} = R_D[DX]'_{ni} = [DX]'_{ni}$$

and

$$R_n[Dx]'_n = R_n R_D[Dx]'_n = R_D[Dx]'_n = [Dx]'_n,$$

respectively, in (2.5) and (2.6). Thus, the solution component $D(t)x(t)$ satisfies the regular ODE

$$(Dx)'(t) = D(t)w(D(t)x(t), t), \tag{2.7}$$

i.e., the term involving the derivative $R'(t)$ in the inherent regular ODE (2.3) disappears. The same holds true for the discretizations ([8]).

Proposition 2.3 *([8]) Applying stiffly accurate Runge-Kutta methods or BDFs to an index-1 DAE (2.1) with properly stated leading term and a constant subspace* im $D(t)$ *yields*

$$x_n = D_n^- D_n x_n + Q_{0n} w(D_n x_n, t_n) \in \mathcal{M}_0(t_n),$$

and $D_n x_n$ is nothing else but the numerical solution of the regular ODE (2.7) by the same integration method.

It turns out that, due to a constant im $D(t)$, we are actually integrating the inherent regular ODE numerically inside such that the decoupling procedure and the discretization commute. This is the best we can expect.

Definition 2.4 *An index-1 DAE (2.1) with well matched coefficients $A(x,t)$ and $D(t)$ and constant subspace* im $D(t)$ *is said to be numerically well formulated.*

In other words, for numerically well formulated DAEs, discretization and decoupling procedures commute due to the fact that the relevant inherent dynamical part has a constant state space, namely im $D(t)$.

Let us mention that, in case of a constant nullspace $\ker A(x,t)$ but time-varying im $D(t)$, we could immediately rewrite the DAE (2.1) as

$$A(x(t), t)(\tilde{D}(t)x(t))' + b(x(t), t) = 0 \tag{2.8}$$

be means of $\tilde{D}(t) := P_A D(t)$ with any constant projector $P_A \in L(\mathbb{R}^m)$ along $\ker A(x,t)$. Since numerical approximations are invariant under this trivial manipulation, let us agree to speak of numerically well formulated index-1 DAEs (2.1) even if actually just (2.8) is numerically well formulated, i.e., the nullspace $\ker A(x,t)$ is constant but $\operatorname{im} D(t)$ varies with t. Obviously, (2.8) has both characteristic subspaces, $\ker A(x,t)$ and $\operatorname{im} \tilde{D}(t)$ constant.
Similarly, if in (2.1), $\ker A(x,t) = \ker R(t)$ varies with t but $\operatorname{im} D(t)$ does not, letting $\tilde{A}(x,t) := A(x,t)R_D$ with any constant projector $R_D \in L(\mathbb{R}^m)$ onto $\operatorname{im} D(t)$, we could rearrange things easily and obtain

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

with both subspaces being constant. However, there is no need at all to realize these manipulations in practice. This is why we rely just on time-invariant subspaces $\operatorname{im} D(t)$, which seems to be more convenient.

When modelling by DAEs, one should try to have numerically well formulated ones at the beginning. Obtaining DAEs with properly stated leading terms might be a practical problem. If the DAE is in the standard formulation, it might be one possibility to make factorizations of the leading term. For example, given the DAE

$$E(x(t),t)x'(t) + g(x(t),t) = 0\,,$$

we may factorize $E(x,t) = A(x,t)D(t)$ with $D(t)$ continuously differentiable, obtaining

$$A(x(t),t)(D(t)x(t))' + g(x(t),t) - A(x(t),t)D'(t)x(t) = 0\,. \tag{2.9}$$

If

$$\ker A(x,t) \quad \oplus \quad \operatorname{im} D(t) = \mathbb{R}^m, \quad \ker A(x,t) = \ker R(t), \quad \operatorname{im} D(t) = \operatorname{im} R(t),$$

where $R = R^2$ is continuously differentiable, then the leading term in (2.9) is well matched. As the factorization is not unique, it is important to investigate if this process may change the index and lead to different solution spaces.
Thus, we consider another factorization of the leading term, $E(x,t) = \tilde{A}(x,t)\tilde{D}(t)$ with $\tilde{D}(t)$ continuously differentiable, obtaining

$$\tilde{A}(x(t),t)(\tilde{D}(t)x(t))' + g(x(t),t) - \tilde{A}(x(t),t)\tilde{D}'(t)x(t) = 0\,. \tag{2.10}$$

where

$$\ker \tilde{A}(x,t) \quad \oplus \quad \operatorname{im} \tilde{D}(t) = \mathbb{R}^m, \quad \ker \tilde{A}(x,t) = \ker \tilde{R}(t), \quad \operatorname{im} \tilde{D}(t) = \operatorname{im} \tilde{R}(t),$$

and $\tilde{R} = \tilde{R}^2$ is continuously differentiable. In this case, also (2.10) is properly formulated.
We begin studying the characteristic subspaces of (2.9) and (2.10). As

$$A(x,t)D(t) = \tilde{A}(x,t)\tilde{D}(t)\,, \tag{2.11}$$

the leading nullspaces $N_0(t) = \ker A(x,t)D(t)$ and $\tilde{N}_0(t) = \ker \tilde{A}(x,t)\tilde{D}(t)$ coincide, but also $\mathcal{M}_0(t) = \tilde{\mathcal{M}}_0(t)$. Consider the further characteristic subspaces

$$S_0(y,x,t) = \{z \in \mathbb{R}^m : g'_x(x,t)z + (A(x,t)y)'_x z - (A(x,t)D'(t)x)'_x z \in \operatorname{im} A(x,t)\}$$

and the respective $\tilde{S}_0(y,x,t)$. If the subspace $\operatorname{im} A(x,t) = \operatorname{im} A(x,t)D(t) = \operatorname{im} \tilde{A}(x,t)$ does not vary with x, we simply have

$$S_0(y,x,t) = \{z \in \mathbb{R}^m : g'_x(x,t)z \in \operatorname{im} A(x,t)\} = \tilde{S}_0(y,x,t).$$

If $\operatorname{im} A(x,t)$ varies with x, things become technically more complicated. Since applications apparently lead just to constant spaces $\operatorname{im} A(x,t)$, we do without considering the solution-dependent case.

Proposition 2.5 *Given two DAEs, (2.9) and (2.10), with properly stated leading terms and $D, \tilde{D}$ continuously differentiable,* $\operatorname{im} A(x,t)$ *independent of x. Let (2.11) hold true.*

(i) Then (2.9) and (2.10) have identical characteristic subspaces $N_0(t)$ and $S_0(x,t)$, and, in particular, they are index 1 DAEs at the same time.

(ii) The solution spaces coincide, i.e., $C^1_D = C^1_{\tilde{D}}$. Each solution of (2.9) satisfies (2.10) and vice versa.

Proof: It remains to show (ii). Since $\ker D(t) = \ker \tilde{D}(t) = N_0(t)$ and $D, \tilde{D} \in C^1$, we may choose $\tilde{P}_0 = P_0 \in C^1, D^-, \tilde{D}^- \in C^1$. Because of $\tilde{D} = \tilde{D}P_0 = \tilde{D}D^-D$ and $D = DP_0 = D\tilde{D}^-\tilde{D}$, the functions Dx and $\tilde{D}x$ are C^1 at the same time, i.e., $C^1_D = C^1_{\tilde{D}}$. Furthermore, for $x \in C^1_D$, it holds that

$$\begin{aligned}
\tilde{A}(\tilde{D}x)' - \tilde{A}\tilde{D}'x &= \tilde{A}(\tilde{D}D^-Dx)' - \tilde{A}(\tilde{D}D^-D)'x = \tilde{A}\tilde{D}D^-(Dx)' - \tilde{A}\tilde{D}D^-D'x \\
&= ADD^-(Dx)' - ADD^-D'x = A(Dx)' - AD'x,
\end{aligned}$$

which completes the proof. $\square$

3 Contractive and dissipative flow

In this paper we are mainly interested in the proper reflection of the qualitative properties of the DAE solutions by their numerical approximations. For regular ODEs, it is well known how numerical methods behave for contractive and dissipative problems [12]. In particular, for contractive ODEs, algebraically stable Runge-Kutta methods maintain the dynamics without any stepsize restriction; for dissipative ODEs, no stepsize restriction is needed with the backward Euler method, but it is needed for multistage algebraically stable methods.

Given the DAE (2.1), we study what contractivity and dissipativity mean. In the case of contractive flows, we would like to avoid stepsize restrictions caused by asymptotic stability problems as we are used to do in the regular ODE case. Similarly, in the case of dissipative flows, essentially, we would also like to have the results of the ODE case.

3.1 Contractivity

One property of regular ODEs that is often discussed is contractivity. Let us clarify first what contractivity means for index-1 DAEs.
Given two arbitrary solutions $x(\cdot), \bar{x}(\cdot) \in C^1_D([t_0,\infty),\mathbb{R}^m)$ of the well formulated index 1 DAE (2.1), we may make use of the solution representation (2.2), (2.3). This yields ([8])

$$\begin{aligned} x(t)-\bar{x}(t) &= D(t)^-(u(t)-\bar{u}(t)) + Q_0(t)(w(u(t),t)-w(\bar{u}(t),t)) \\ &= \int_0^1 \{D(t)^- + Q_0(t)w'_u(\tau u(t)+(1-\tau)\bar{u}(t)),t)\}\, d\tau\, (u(t)-\bar{u}(t)). \end{aligned}$$

If we denote the canonical projector onto $S_0(y,x,t)$ along $N_0(t)$ by $P_{\text{can}}(y,x,t)$, it holds that ([8, Remark 2.5 & Lemma 2.1])

$$D(t)^- + Q_0(t)w'_u(u,t) = P_{\text{can}}(Dw(u,t), D(t)^-u + Q_0(t)w(u,t),t)D(t)^-$$

and thus

$$x(t)-\bar{x}(t) = \int_0^1 P_{\text{can}}(\eta_{(u(t),\bar{u}(t),t)}(\tau))\, d\tau\, D(t)^-\,(u(t)-\bar{u}(t)) \tag{3.1}$$

with

$$\begin{aligned} \eta_{(u(t),\bar{u}(t),t)}(\tau) \; := \; & (D(t)w(\tau u(t)+(1-\tau)\bar{u}(t),t), \\ & D(t)^-(\tau u(t)+(1-\tau)\bar{u}(t)) + Q_0(t)w(\tau u(t)+(1-\tau)\bar{u}(t),t),t). \end{aligned}$$

The representation (3.1) makes clear that the flow $x(t)-\bar{x}(t)$ is mainly driven by $u(t)-\bar{u}(t) = D(t)x(t)-D(t)\bar{x}(t)$, but there is also an affect of the DAE geometry via the projector P_{can}. Recall once more that im $D(t)$ is an invariant subspace of the regular ODE (2.3), and that our solution components lie in this subspace, i.e., $u(t)=R(t)u(t)$, $\bar{u}(t)=R(t)\bar{u}(t)$. Taking an inner product on $\mathbb{R}^m$ we may compute

$$\begin{aligned} \tfrac{1}{2}\tfrac{d}{dt}|u(t)-\bar{u}(t)|^2 &= \langle u'(t)-\bar{u}'(t), u(t)-\bar{u}(t)\rangle \\ &= \langle R(t)(u'(t)-\bar{u}'(t)) + R'(t)(u(t)-\bar{u}(t)), u(t)-\bar{u}(t)\rangle\,. \end{aligned}$$

If we had an inequality of the form

$$\frac{1}{2}\frac{d}{dt}|u(t)-\bar{u}(t)|^2 \le -\beta|u(t)-\bar{u}(t)|^2$$

with $\beta \ge 0$, we would obtain contractivity

$$|u(t)-\bar{u}(t)| \le e^{-\beta(t-t_0)}|u(t_0)-\bar{u}(t_0)|, \quad t\ge t_0\,.$$

Therefore, we should suppose the inherent regular ODE to be contractive in the usual sense on im $D(t)$, i.e., the inequality

$$\langle D(t)(w(u,t)-w(\bar{u},t)) + R'(t)(u-\bar{u}), u-\bar{u}\rangle \le -\beta|u-\bar{u}|^2 \tag{3.2}$$

should be given for all $u,\bar{u}\in \text{im } D(t)$, $t\ge t_0$. The following definition takes up this idea, but in terms of the original DAE.

Definition 3.1 *The well formulated index-1 DAE (2.1) is said to be contractive (strongly contractive) on* $[t_0, \infty)$ *if there is an inner product on* $\mathbb{R}^m$ *and a constant* $\beta \geq 0$ $(\beta > 0)$ *such that the inequality*

$$\langle z - \bar{z}, D(t)(x - \bar{x})\rangle + \langle R'(t)D(t)(x - \bar{x}), D(t)(x - \bar{x})\rangle \leq -\beta |D(t)(x - \bar{x})|^2 \tag{3.3}$$

holds true for all $x, \bar{x} \in \mathcal{M}_0(t)$, $z, \bar{z} \in \operatorname{im} D(t)$, $f(z, x, t) = f(\bar{z}, \bar{x}, t) = 0$, $t \geq t_0$.

We have now he following result.

Proposition 3.2 *Given a well formulated index 1 DAE (2.1)*

(i) If (2.1) is contractive, then, for any two solutions $x(\cdot), \bar{x}(\cdot) \in C^1_D([t_0, \infty), \mathbb{R}^m)$, *it holds that*

$$|D(t)x(t) - D(t)\bar{x}(t)| \leq e^{-\beta(t-t_0)} \left|D(t_0)(x(t_0) - \bar{x}(t_0))\right|, \quad t \geq t_0,$$

and

$$|x(t) - \bar{x}(t)| \leq K_{(x,\bar{x})}(t)|D(t)x(t) - D(t)\bar{x}(t)|, \quad t \geq t_0,$$

where $K_{(x,\bar{x})}(t) := \max\limits_{\tau \in [0,1]} \left\| P_{\mathrm{can}}(\eta_{(D(t)x(t), D(t)\bar{x}(t), t)}(\tau)) D(t)^- \right\|$.

(ii) If (2.1) is contractive, then the inherent regular ODE (2.3) is contractive on its basic invariant subspace $\operatorname{im} D(t)$ *and vice versa.*

Proof: It remains to prove (ii). Let (3.3) be given. For $t \geq t_0$, $u, \bar{u} \in \operatorname{im} D(t)$ we determine

$$\begin{aligned} x &= Q_0(t)w(u, t) + D(t)^- u \in \mathcal{M}_0(t)\,, & \bar{x} &= Q_0(t)w(\bar{u}, t) + D(t)^- \bar{u} \in \mathcal{M}_0(t)\,, \\ z &= D(t)w(u, t) \in \operatorname{im} D(t)\,, & \bar{z} &= D(t)w(\bar{u}, t) \in \operatorname{im} D(t)\,. \end{aligned}$$

This yields $D(t)x = u$, $D(t)\bar{x} = \bar{u}$, $f(z, x, t) = 0$, $f(\bar{z}, \bar{x}, t) = 0$ by construction of the function w ([8]). Now (3.3) gives

$$\langle D(t)(w(u, t) - w(\bar{u}, t)), u - \bar{u}\rangle + \langle R'(t)(u - \bar{u}), u - \bar{u}\rangle \leq -\beta |u - \bar{u}|^2,$$

i.e., the inherent regular ODE is contractive on $\operatorname{im} D(t)$, where the same inner product and constant β as in (3.3) may be used. Conversely, if (3.2) is given, we take $t \geq t_0$, $x, \bar{x} \in \mathcal{M}_0(t)$, $z, \bar{z} \in \operatorname{im} D(t)$ with $f(z, x, t) = 0$, $f(\bar{z}, \bar{x}, t) = 0$ and introduce

$$u := D(t)x, \quad \bar{u} := D(t)\bar{x}, \quad \bar{\omega} := D(t)^- \bar{z} + Q_0(t)\bar{x}, \quad \omega := D(t)^- z + Q_0(t)x$$

such that $\omega = w(u, t)$ and $\bar{\omega} = w(\bar{u}, t)$ become true ([8]). Then, (3.2) leads immediately to (3.3). $\square$

Remarks:

1) Note that the term $P_{\text{can}}D^-$ in the definition of the bound $K_{(x,\bar{x})}(t)$ does not depend on the choice of P_0. Namely, for two different projectors $P_0, \tilde{P}_0$ and the corresponding generalized inverses $D^-, \tilde{D}^-$ we may derive

$$\begin{aligned} P_{\text{can}}D^- &= P_{\text{can}}D^-DD^- = P_{\text{can}}D^-R = P_{\text{can}}D^-D\tilde{D}^- \\ &= P_{\text{can}}P_0\tilde{D}^- = P_{\text{can}}\tilde{D}^-. \end{aligned}$$

2) It makes no sense to demand contractivity of the inherent regular ODE on the whole $\mathbb{R}^m$.

3) In essence, $K_{(\cdot)}(t)$ is a bound of the canonical projector $P_{\text{can}}(\cdot,\cdot,t)$ in the neighbourhood of the two solutions at time t. Clearly, for orthogonal subspaces $S(y,x,t)$ and $N(t)$ we obtain $\|P_{\text{can}}(\cdot,\cdot,t)\| = 1$, but in general we have to expect also large values.
If $P_{\text{can}}(y,x,t)D(t)^-$ is globally bounded by a constant K, Proposition 3.2 implies

$$|x(t) - \bar{x}(t)| \le K\ e^{-\beta(t-t_0)}|D(t_0)(x(t_0) - \bar{x}(t_0))|,\ t \ge t_0. \qquad \square$$

For the class of Runge-Kutta methods considered in this paper we have the following statement.

Proposition 3.3 *Given a numerically well formulated and contractive index-1 DAE (2.1), we apply an algebraically stable Runge-Kutta method with starting values* $x_0, \bar{x}_0 \in \mathcal{M}_0(t_0)$. *Then*

$$\begin{aligned} |D_n x_n - D_n\bar{x}_n| &\le |D_{n-1}x_{n-1} - D_{n-1}\bar{x}_{n-1}|, \\ |x_n - \bar{x}_n| &\le K_{(D_n x_n, D_n\bar{x}_n, t_n)}(t_n)|D_n x_n - D_n\bar{x}_n|,\ n \ge 1\,. \end{aligned}$$

Proof: By Proposition 3.2, the inherent regular ODE is contractive on im $D(t)$, but this subspace is now constant. Since, in this case, discretization and decoupling commute, we may apply standard arguments to obtain the B-stability inequality for the components $D_n(x_n - \bar{x}_n)$. The second inequality is due to $x_n, \bar{x}_n \in \mathcal{M}_0(t_n)$ and an analogue of the representation (3.1). $\square$

3.2 Dissipativity

A further interesting qualitative property in the case of regular ODEs is dissipativity, where an absorbing set sucks up all solutions. Let us clarify what this means for index-1 DAEs (2.1).
The geometric solution set is now $\mathcal{M}_0(t)$, which may depend on time, i.e., all solutions at time $t \ge t_0$ remain in $\mathcal{M}_0(t)$ and $\mathcal{M}_0(t)$ is completely filled by the solutions of (2.1) ([8]). We denote by $x(t; t_*, x_*)$ the solution of the DAE (2.1) passing at time t_* trough $x_* \in \mathcal{M}_0(t_*)$. A set $\mathcal{B}(t) \subset \mathcal{M}_0(t)$, $t \ge t_0$, is called a *positively invariant set* of the DAE (2.1) if $x_* \in \mathcal{B}(t_*)$ implies $x(t; t_*, x_*) \in \mathcal{B}(t)$ for all $t > t_*$.

Definition 3.4 *A positively invariant set* $\mathcal{B}(t), t \geq t_0$*, is called an absorbing set of the DAE (2.1) if, for any* $t_* \in [t_0, \infty)$ *and any bounded set* $E \subset \mathcal{M}_0(t_*)$*, there is a time* $t_{(E,t_*)} \geq t_*$ *such that* $x_* \in E$ *implies* $x(t, t_*, x_*) \in \mathcal{B}(t)$ *for* $t \geq t_{(E,t_*)}$.

Definition 3.5 *The DAE (2.1) is said to be dissipative if it has a bounded absorbing set.*

In a similar way as we have done it for contractivity, we can give a condition ensuring the IRODE to be dissipative, and thus we derive that the DAE is dissipative.

Proposition 3.6 *Given an index 1 DAE (2.1) with a properly stated leading term, let the inequality*

$$\langle z, D(t)x\rangle + \langle R'(t)D(t)x, D(t)x\rangle \leq \alpha - \beta|D(t)x|^2 \tag{3.4}$$

be satisfied for all $x \in \mathcal{M}_0(t)$, $z \in \mathrm{im}\ D(t)$, $f(z, x, t) = 0$, $t \geq t_0$, *with* $\alpha \geq 0$, $\beta > 0$ *constant.*

(i) Then the inherent regular ODE (2.3) is dissipative on $\mathrm{im}\ D(t)$ *where, for any* $\varepsilon > 0$,

$$\mathcal{B}_{\mathrm{IRODE}}\ (t) := \left\{ v \in \mathrm{im}\ D(t) : |v|^2 \leq \frac{\alpha}{\beta} + \varepsilon \right\}$$

is an absorbing set.

(ii) If, additionally, there are global bounds K, γ *so that*

$$\|P_{\mathrm{can}}(y, x, t)D(t)^-\| \leq K, \quad |Q_0(t)\omega(0, t)| \leq \gamma,$$

then (2.1) is dissipative with the absorbing set

$$\mathcal{B}(t) = \left\{ x \in \mathcal{M}_0(t) : |x| \leq K\left(\frac{\alpha}{\beta} + \varepsilon\right)^{1/2} + \gamma \right\}, \quad \varepsilon > 0.$$

Proof:

(i) Recall that the function $w(u, t)$ in (2.3) is constructed so that

$$f(D(t)w(u, t), D(t)^- u + Q_0(t)w(u, t), t) = 0\,.$$

For $t \geq t_0$, $u \in D(t)\mathcal{M}_0(t)$ we denote $z := D(t)w(u, t)$, $x := D(t)^- u + Q_0(t)w(u, t)$. Then we have that $f(z, x, t) = 0$, and hence, by (3.4),

$$\langle D(t)w(u, t), u\rangle + \langle R'(t)u, u\rangle \leq \alpha - \beta|u|^2. \tag{3.5}$$

For any solution $u(t)$ of the inherent regular ODE (2.3) that belongs to im $D(t)$, (3.5) yields

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}|u(t)|^2 &= \langle u'(t), u(t)\rangle = \langle R'(t)u(t) + D(t)w(u(t),t), u(t)\rangle \\ &\le \alpha - \beta |u(t)|^2.\end{aligned}$$

Therefore, for any $t_* \ge t_0$, $u_* \in D(t_*)\mathcal{M}_0(t_*)$ we may solve the corresponding IVP. The IVP solution satisfies the inequality

$$|u(t)|^2 \le \frac{\alpha}{\beta} + e^{-2\beta(t-t_*)}\left\{|u_*|^2 - \frac{\alpha}{\beta}\right\},\ t \ge t_*,$$

and thus

$$|u(t)| \le \max\left\{|u_*|, \left(\frac{\alpha}{\beta}\right)^{1/2}\right\},\ t \ge t_*,$$

which shows $\mathcal{B}_{\text{IRODE}}(t)$ to be a positively invariant set. Next, given any bounded set $E_u \subset D(t_*)\mathcal{M}(t_*)$, we denote $r := \sup\{|v| : v \in E_u\}$. For all $u_* \in E_u$, the IVP solutions satisfy

$$|u(t)| \le \frac{\alpha}{\beta} + e^{-2\beta(t-t_*)}\left\{r^2 - \frac{\alpha}{\beta}\right\},\ t \ge t_*.$$

Choosing $\bar{t} = \bar{t}(E_u, t_*)$ so that $e^{-2\beta(E-t_*)}\left\{r^2 - \frac{\alpha}{\beta}\right\} \le \varepsilon$, we obtain $|u(t)| \le \frac{\alpha}{\beta} + \varepsilon$ for all $u_* \in E_u$ and $t \ge \bar{t}$, i.e., $\mathcal{B}_{\text{IRODE}}(t)$ absorbs the solutions indeed.

(ii) Consider a bounded set $E \subset \mathcal{M}_0(t_*)$, $t_* \ge t_0$ arbitrarily fixed, $E_u := D(t_*)E$. For any $x_* \in E$, we represent the IVP solution $x(t) = x(t; t_*, x_*)$ as

$$x(t) = D(t)^- u(t) + Q_0(t)w(u(t),t),$$

where $u(t) = D(t)x(t)$ satisfies $u(t_*) = u_* = D(t_*)x_* \in E_u$ and the inherent regular ODE (2.3). Due to (i), it holds that $|u(t)| \le \sqrt{\frac{\alpha}{\beta} + \varepsilon}$ for all $t \ge \tilde{t}$ uniformly for all $u_* \in E_u$. Consequently,

$$\begin{aligned}|x(t)| &= |D(t)^- u(t) + Q_0(t)w(u(t),t) - Q_0(t)w(0,t) + Q_0(t)w(0,t)| \\ &\le |\int_0^1 \{I - Q_0(t)w'_u(\tau u(t),t)\}\, ds\, D(t)^- u(t)| + |Q_0(t)w(0,t)| \\ &\le K|u(t)| + \gamma \quad \le \quad K\left(\frac{\alpha}{\beta} + \varepsilon\right)^{1/2} + \gamma \qquad \text{for } t \ge \bar{t}. \qquad \square\end{aligned}$$

Remark: In the linear case of $f(y,x,t) \equiv A(t)y + B(t)x - q(t)$, the equation

$$f(D(t)w, D(t)^{-}u + Q_0(t)w, t) = 0$$

means

$$w = -A_1(t)^{-1}\{B(t)D(t)^{-}u - q(t)\}, \quad \text{where } A_1 := AD + BQ_0.$$

Therefore, we have

$$\begin{aligned} w(u,t) &= -A_1(t)^{-1}\{B(t)D(t)^{-}u - q(t)\}, \\ w(0,t) &= A_1(t)^{-1}q(t), \\ Q_0(t)w(0,t) &= Q_0(t)A_1(t)^{-1}q(t). \end{aligned}$$

It may be checked that $Q_0(t)A_1(t)^{-1}$ is independent of the choice of $Q_0(t)$. □

Again, if the DAE (2.1) is numerically well formulated, most of the results about the numerical solution of dissipative regular ODEs hold for the DAE (2.1), too. For example, in [12] it is shown that the backward Euler method reflects the dissipativity behaviour without any stepsize restriction, whereas general algebraically stable Runge-Kutta methods reflect the dissipative flow under certain stepsize restrictions.
We give the result for the backward Euler method here. Some other results of [12] could be adopted in a similar way.

Proposition 3.7 *Under the conditions of Proposition 3.6, and if, moreover, (2.1) is numerically well formulated, the backward Euler method reflects the dissipativity behaviour properly without any stepsize restriction. The absorbing sets are the same as described in Proposition 3.6.*

Proof: As im $D(t)$ is constant, discretization and decoupling commute. If we use the respective result (e.g. [12, Theorem 5.5.3]) for regular ODEs and match the components as we did in Proposition 3.6, we obtain the desired result. □

4 Numerically equivalent representations

In this section we deal with DAEs

$$A(t)(D(t)x(t))' + g(x(t),t) = 0, \quad t \in \mathcal{I}, \tag{4.1}$$

which have a properly stated leading term, but which are not numerically well formulated, i.e., neither im $D(t)$ nor $\ker A(t)$ are constant. For this kind of problems we cannot ensure the numerical solution to follow the dynamics of the inherent regular ODE. In this situation, we can try to reformulate the problem. For example, we may decompose $A(t) = \tilde{A}(t)K(t)$ with $K(t)$ continuously differentiable, and write $\tilde{D}(t) = K(t)D(t)$ to transform (4.1) into

$$\tilde{A}(t)(\tilde{D}(t)x(t))' - \tilde{A}(t)K'(t)D(t)x(t) + g(x(t),t) = 0, \tag{4.2}$$

or we may decompose $D(t) = K(t)\tilde{D}(t)$ and write $\tilde{A}(t) = A(t)K(t)$ to obtain

$$\tilde{A}(t)(\tilde{D}(t)x(t))' - A(t)K'(t)\tilde{D}(t)x(t) + g(x(t),t) = 0. \tag{4.3}$$

If $\ker \tilde{A}(t) \oplus \operatorname{im} \tilde{D}(t) = \mathbb{R}^m$ and either $\operatorname{im} \tilde{D}(t)$ or $\ker \tilde{A}(t)$ is constant, then (4.2) and (4.3), respectively, are numerically well formulated.
Of course, to compute the extra term $\tilde{A}K'D$ or $AK'\tilde{D}$ during the numerical integration might be not so nice. So we will ask for conditions causing one of those terms to disappear. As we will see later, applying the BDF or a stiffly accurate IRK method to the original DAE (4.1) under these conditions yields the same approximations x_n as if applying this method to the reformulation (4.2) or (4.3), i.e., these equations are numerically equivalent to (4.1). If it turns out that (4.2) or (4.3) are numerically well formulated, then the integration method applied to the original DAE (which was not numerically well formulated) actually generates values corresponding to a numerically well formulated representation of (4.1).

Recall (cf. Proposition 2.5) that the characteristic subspaces $N_0(t)$ and $S_0(x,t)$ do not at all depend on the special factorization determining the leading term. Moreover, if we start with a standard formulation ([5], [2]),

$$E(t)x'(t) + f(x(t),t) = 0$$

and factorize $E(t) = A(t)D(t)$ to obtain

$$A(t)(D(t)x(t))' + f(x(t),t) - A(t)D'(t)x(t) = 0 \tag{4.4}$$

we will obtain the subspaces

$$\begin{aligned} N_0(t) &= \ker A(t)D(t) = \ker E(t), \\ S_0(x,t) &= \{z \in \mathbb{R}^m : f'_x(x,t)z - A(t)D'(t)z \in \operatorname{im} A(t)D(t)\} \\ &= \{z \in \mathbb{R}^m : f'_x(x,t)z \in \operatorname{im} E(t)\} \end{aligned}$$

independently of the chosen factorization.

Let us further mention that, for homogeneous linear index-1 DAEs, the index-1 condition $N_0(t) \oplus S_0(t) = \mathbb{R}^m$ allows to make use of the canonical projector $P_{\text{can}}(t)$, which projects onto $S_0(t)$ along $N_0(t)$ (cf. Section 3). Since $S_0(t)$ represents the geometrical solution space of homogeneous linear index-1 DAEs, it holds for all solutions that $x(t) = P_{\text{can}}(t)x(t)$. In the consequence, the index-1 equation

$$E(t)x'(t) + F(t)x(t) = 0 \tag{4.5}$$

trivially may be understood as

$$E(t)(P_{\text{can}}(t)x(t))' + F(t)x(t) = 0 \tag{4.6}$$

(i.e., the term $EP'_{\mathrm{can}}x = EP'_{\mathrm{can}}P_{\mathrm{can}}x$ disappears), provided that P_{can} belongs to the class C^1. This leading term is properly stated.

To illustrate the situations that may occur, we give two examples of homogeneous index-1 DAEs with time-dependent subspaces $N(t)$ and $S_0(t)$. Recall that although they are supposed to be as in (4.5), they are numerically equivalent to (4.6) and should be understood in this way.

Example 4.1 *([7]): The DAE*

$$\begin{pmatrix} \delta-1 & \delta t \\ 0 & 0 \end{pmatrix} x'(t) + \mu \begin{pmatrix} \delta-1 & \delta t \\ \delta-1 & \delta t-1 \end{pmatrix} x(t) = 0 \tag{4.7}$$

has index-1 for $\delta \neq 1, \mu \neq 0$. Its solution is

$$x_1(t) = \frac{1-\delta t}{\delta-1} x_2(t), \quad x_2(t) = e^{(\delta-\mu)t} x_2(0).$$

Both subspaces,

$$N_0(t) = \left\{ z \in \mathbb{R}^2 : z_1 = \frac{\delta t}{1-\delta} z_2 \right\}$$

and

$$S_0(t) = \left\{ z \in \mathbb{R}^2 : z_1 = -\frac{\delta t-1}{\delta-1} z_2 \right\},$$

vary with t, i.e., neither $N_0(t)$ nor $\mathrm{im}\, P_{\mathrm{can}}(t) = S_0(t)$ are constant. We cannot ensure a qualitatively good behaviour of the numerical solution without extra stepsize restriction. In particular, starting the implicit Euler method with consistent initial values gives

$$x_{1,n+1} = \frac{\delta t_{n+1}-1}{\delta-1} x_{2,n+1}, \quad x_{2,n+1} = \frac{1+\delta h}{1+h\mu} x_{2,n},$$

and thus, for some values of the parameters δ and μ, namely for $\delta < \mu$, we are confronted with strong stepsize restrictions to obtain $|1+\delta h| < |1+h\mu|$.

In Figure 1 we show the global error of the numerical solution at $t = 10$ with the backward Euler method for $\delta = -100$ and μ varying from -95 to -45. The stepsize h varies from 10^{-1} to 10^{-3}. Observe that for stepsizes h greater than 10^{-2} the numerical solution explodes.

On the other hand, with

$$A(t) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad D(t) = \begin{pmatrix} \delta-1 & \delta t \\ 0 & 0 \end{pmatrix}$$

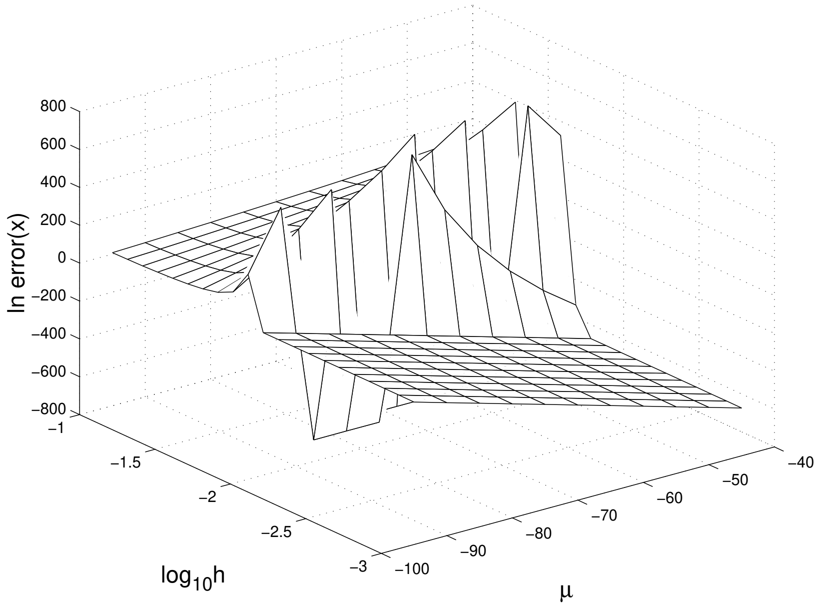

Figure 1: Numerical solution of Example 4.1, which is not numerically well formulated (see Equation 4.7).

and with

$$A(t) = \begin{pmatrix} \delta-1 & \delta t \\ 0 & 0 \end{pmatrix}, \quad D(t) = \begin{pmatrix} 1 & \frac{\delta}{\delta-1}t \\ 0 & 0 \end{pmatrix}$$

we obtain properly stated leading parts with constant im $D(t)$*. However, now the additional terms* $AD'x$ *do not vanish, i.e., we actually have reformulations of the problem. For the first one, we obtain*

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \left(\begin{pmatrix} \delta-1 & \delta t \\ 0 & 0 \end{pmatrix} x(t) \right)' + \mu \begin{pmatrix} \delta-1 & \delta t - \frac{\delta}{\mu} \\ \delta-1 & \delta \end{pmatrix} x(t) = 0 \tag{4.8}$$

In both cases, the implicit Euler method leads to

$$x_{1,n+1} = \frac{\delta t_{n+1} - 1}{\delta - 1} x_{2,n+1}, \quad x_{2,n+1} = \frac{1}{1 - h(\delta - \mu)} x_{2,n}.$$

Obviously, these reformulations improve the numerical solution. Note that the modified approaches discussed in [4] and [5] give the reformulation (4.4) for this example.

In Figure 2 we show the global error of the numerical solution at $t = 10$ *with the backward Euler method for* $\delta = -100$ *and* μ *varying from* -95 *to* -45*. The stepsize* h *varies from* 10^{-1} *to* 10^{-3}*.*

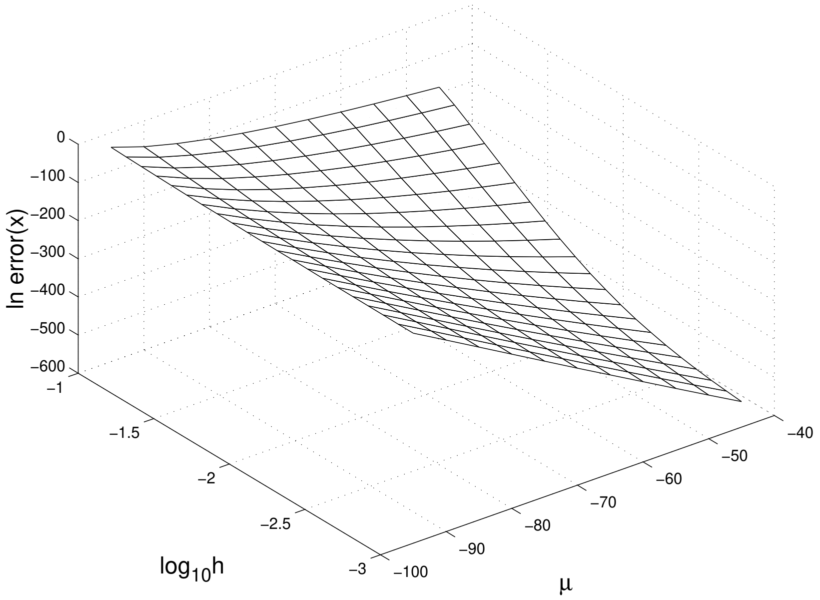

Figure 2: Numerical solution of Example4.1, which is numerically well formulated (see Equation 4.8).

Observe that

$$\begin{aligned}\ker E(t_*) \cap \operatorname{im} P_{\text{can}}(t) &= N_0(t_*) \cap S_0(t) = \\ &= \left\{ z \in \mathbb{R}^2 : z_1 = \tfrac{\delta t_*}{1-\delta} z_2,\ (\delta(t_* - t) + 1) z_2 = 0 \right\}.\end{aligned}$$

If $\delta = 0$*, i.e., if (4.7) has constant coefficients,* $N_0(t_*) \cap S_0(t) = \{0\}$*. However, for* $\delta \neq 0$*, the intersection* $N_0(t_*) \cap S_0(t)$ *is nontrivial for* $t_* - t = -\frac{1}{\delta}$*.*

Example 4.2 *The index-1 DAE*

$$\begin{pmatrix} 1 & p(t) & p(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} x'(t) + F(t)x(t) = 0 \tag{4.9}$$

with

$$\begin{aligned}&F(t) = (f_{ij}(t)),\\ &f_{21}(t) = q(t)(f_{22}(t) - f_{23}(t)), f_{31} = q(t)(f_{32}(t) - f_{33}(t)),\\ &f_{22}(t)f_{33}(t) - f_{23}(t)f_{32}(t) \neq 0,\end{aligned}$$

has the solution

$$x_2(t) = -q(t)x_1(t), \quad x_3(t) = q(t)x_1(t)$$

and $x_1(t)$ *solves the regular ODE*

$$x_1'(t) + (f_{11}(t) - f_{12}(t)q(t) + f_{13}(t)q(t))x_1(t) = 0.$$

The subspaces

$$N_0(t) = \{z \in \mathbb{R}^3 : z_1 + p(t)(z_2 + z_3) = 0\},$$
$$S_0(t) = \{z \in \mathbb{R}^3 : z_2 = -z_3, z_3 = q(t)z_1\}$$

actually vary with time t. Consequently (4.9) or, equivalently, (4.6), are not numerically well formulated. Since $E(t)$ has a constant image we may again use the factorization $E = AD$,

$$A(t) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad D(t) = \begin{pmatrix} 1 & p(t) & p(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and turn to the reformulation (4.4) i.e.,

$$A(t)(D(t)x(t))' + (F(t) - A(t)D'(t))x(t) = 0.$$

Now, the implicit Euler method yields

$$\begin{aligned} x_{2,n+1} &= -q_{n+1}x_{1,n+1}, \ x_{3,n+1} = q_{n+1}x_{1,n+1}, \\ x_{1,n+1} &= x_{1,n} - h(f_{11} - f_{12}q + f_{13}q)_{n+1}x_{1,n+1}. \end{aligned}$$

However, applying the implicit Euler method to the original equation (4.9) yields, surprisingly, exactly the same result as with the above reformulation. We conclude that (4.9) only seems to be not numerically well formulated. In fact, it is numerically equivalent to a numerically well formulated version.
In this case, the expensive reformulation (4.4), where the terms $AD'x = AD'P_{\mathrm{can}}x$ must be computed, is completely useless.

In the following, we show why the relevant terms vanish in Example 4.2. More generally, we consider the reformulations (4.2) and (4.3) to analyze whether they are simply a numerically equivalent representation of (4.1). First we study conditions ensuring (4.2) and (4.3) to have properly stated leading terms.

Lemma 4.3 *Given A, D with*

$$\ker A \oplus \mathrm{im}\ D = \mathbb{R}^m. \tag{4.10}$$

(i) Factorize $A = \tilde{A}K$ and put $\tilde{D} = KD$ such that $AD = \tilde{A}\tilde{D}$. If $\mathrm{im}\ A = \mathrm{im}\ \tilde{A}$, *then* $\ker \tilde{A} \oplus \mathrm{im}\ \tilde{D} = \mathbb{R}^m$.

(ii) Factorize $D = K\tilde{D}$ and put $\tilde{A} = AK$ such that $AD = \tilde{A}\tilde{D}$. If $\ker D = \ker \tilde{D}$, *then* $\ker \tilde{A} \oplus \mathrm{im}\ \tilde{D} = \mathbb{R}^m$.

Proof:

(i) Proving that $\ker \tilde{A}\tilde{D} = \ker \tilde{D}$, $\mathrm{im}\ \tilde{A}\tilde{D} = \mathrm{im}\ \tilde{A}$ and $\ker \tilde{A} \cap \mathrm{im}\ \tilde{D} = \{0\}$, we obtain the required result by Lemma 6.1 (Appendix). Observe that

$$\mathrm{im}\ \tilde{A}\tilde{D} = \mathrm{im}\ AD = \mathrm{im}\ A = \mathrm{im}\ \tilde{A}.$$

Next we verify the relation $\ker \tilde{A} \cap \mathrm{im}\ \tilde{D} = \{0\}$. Let $z \in \ker \tilde{A} \cap \mathrm{im}\ \tilde{D}$, that means, $\tilde{A}z = 0$, $z = \tilde{D}w$ for some w. Thus, $0 = \tilde{A}\tilde{D}w = ADw$, in other words, $Dw \in \mathrm{im}\ D \cap \ker A = \{0\}$, hence $Dw = 0$, further $z = \tilde{D}w = KDw = 0$. It remains to show that $\ker \tilde{A}\tilde{D} = \ker \tilde{D}$ is true. Trivially, we have $\ker \tilde{D} \subseteq \ker \tilde{A}\tilde{D}$. Let $z \in \ker \tilde{A}\tilde{D}$, i.e., $0 = \tilde{A}\tilde{D}z = ADz$. Because of $Dz \in \mathrm{im}\ D \cap \ker A = \{0\}$ it holds that $Dz = 0$, therefore $\tilde{D}z = KDz = 0$ and, finally, $\ker \tilde{A}\tilde{D} = \ker \tilde{D}$.

(ii) Consider $\tilde{A}^* = K^*A^*, D^* = \tilde{D}^*K^*$. Taking into account that (4.10) is equivalent to $\ker D^* \oplus \mathrm{im}\ A^* = \mathbb{R}^m$ (see Lemma 6.2, Appendix), we derive from

$$\mathrm{im}\ D^* = (\ker D)^\perp = (\ker \tilde{D})^\perp = \mathrm{im}\ \tilde{D}^*$$

that $\ker \tilde{D}^* \oplus \mathrm{im}\ \tilde{A}^* = \mathbb{R}^m$. But this is equivalent to $\ker \tilde{A} \oplus \mathrm{im}\ \tilde{D} = \mathbb{R}^m$. □

Without further proof we may state the following assertion.

Proposition 4.4 *Given the DAE (4.1) with properly formulated leading term.*

(i) Factorize the matrices $A(t) = \tilde{A}(t)K(t)$ with continuously differentiable K and put $\tilde{D}(t) = K(t)D(t)$ such that $A(t)D(t) = \tilde{A}(t)\tilde{D}(t)$. If $\mathrm{im}\ \tilde{A}(t) = \mathrm{im}\ A(t)$, then (4.2) has also a properly formulated leading term.

(ii) Factorize the matrices $D(t) = K(t)\tilde{D}(t)$ with continuously differentiable K and put $\tilde{A}(t) = A(t)K(t)$ such that $A(t)D(t) = \tilde{A}(t)\tilde{D}(t)$. If $\ker \tilde{D}(t) = \ker D(t)$, then (4.3) has a properly stated leading term.

Given a DAE (4.1) with properly stated leading term, let us discuss some special cases of reformulations by means of factorizations:

Case 1: For any continuously differentiable nonsingular matrix $K(t)$, we may take $\tilde{A}(t) = A(t)K(t)^{-1}$, $\tilde{D}(t) = K(t)D(t)$. It holds trivially that $\mathrm{im}\ \tilde{A}(t) = \mathrm{im}\ A(t)$, $\ker \tilde{D}(t) = \ker D(t)$. Hence, (4.2) and (4.3) have a properly stated leading term.

Case 2: Factorize $A(t) = A(t)P_A(t)$ with $P_A(t)$ a continuously differentiable projector along $\ker A(t)$. Put $\tilde{A}(t) = A(t)$ and $\tilde{D}(t) = P_A(t)D(t)$, which corresponds to $K(t) = P_A(t)$ in (4.2). Due to $\mathrm{im}\ \tilde{A}(t) = \mathrm{im}\ A(t)$, (4.2) has a properly stated leading term.

Case 3: Factorize $A(t) = R_A(t)A(t)$ with $R_A(t)$ a projector onto $\mathrm{im}\ A(t)$. Assume $A(t)$ to be continuously differentiable. Letting $\tilde{A}(t) = R_A(t)$, $\tilde{D}(t) = A(t)D(t)$, i.e., $K(t) = A(t)$ in (4.2), because of $\mathrm{im}\ \tilde{A}(t) = \mathrm{im}\ A(t)$, the reformulation (4.2) has a properly stated leading term.

Case 4: Factorize $D(t) = D(t)P_0(t)$ with $P_0(t)$ a projector along $\ker D(t)$ and assume $D(t)$ to be continuously differentiable. In this case $\tilde{D}(t) = P_0(t)$, $\tilde{A}(t) = A(t)D(t)$ and $K(t) = D(t)$ in (4.3) as well as $\ker \tilde{D}(t) = \ker D(t)$, thus (4.3) has a properly stated leading term.

Case 5: Factorize $D(t) = R_D(t)D(t)$ with a continuously differentiable projector $R_D(t)$ onto $\mathrm{im}\, D(t)$ and put $\tilde{D}(t) = D(t)$, $K(t) = R_D(t)$, $\tilde{A}(t) = A(t)R_D(t)$. Trivially, $\ker \tilde{D}(t) = \ker D(t)$, and the leading term of (4.3) is properly stated.

Now we are in the position to study under which conditions the original DAE (4.1) is numerically equivalent to (4.2) or (4.3).

Proposition 4.5 *Given an index-1 DAE (4.1) with well matched coefficients A and D.*

(i) Factorize $A(t) = \tilde{A}(t)K(t)$ and put $\tilde{D}(t) = K(t)D(t)$ in such a way that $A(t)D(t) = \tilde{A}(t)\tilde{D}(t)$, $\mathrm{im}\, A(t) = \mathrm{im}\, \tilde{A}(t)$, and K is continuously differentiable. If, additionally,

$$K(t_*)D(t) = K(t)D(t) \text{ for all } \quad t_*, t \in \mathcal{I}, \tag{4.11}$$

then (4.1) is numerically equivalent to

$$\tilde{A}(t)(\tilde{D}(t)x(t))' + g(x(t), t) = 0. \tag{4.12}$$

In case of $\ker K(t) = \ker A(t)$, $\tilde{D}(t)$ has a time-invariant image, i.e., (4.12) is numerically well formulated.

(ii) Factorize $D(t) = K(t)\tilde{D}(t)$ and put $\tilde{A}(t) = A(t)K(t)$ in such a way that $A(t)D(t) = \tilde{A}(t)\tilde{D}(t)$, $\ker \tilde{D}(t) = \ker D(t)$, and K is continuously differentiable. If, additionally,

$$A(t)K(t_*) = A(t)K(t) \text{ for all } \quad t_*, t \in \mathcal{I}, \tag{4.13}$$

then (4.1) is numerically equivalent to (4.12). In case of $\mathrm{im}\, K(t) = \mathrm{im}\, D(t)$, $\tilde{A}(t)$ has a time-invariant nullspace, i.e., (4.12) is numerically well formulated.

Proof:

(i) By Proposition 4.4, $\tilde{A}$ and $\tilde{D}$ are well matched. Due to condition (4.11), the term $K'(t)D(t)$ vanishes identically and (4.2) is exactly the same as (4.12).
For Runge-Kutta methods, (4.11) implies

$$A_{ni}[DX]'_{ni} = \tilde{A}_{ni}K_{ni}[DX]'_{ni} = \tilde{A}_{ni}[KDX]'_{ni} = \tilde{A}_{ni}[\tilde{D}X]'_{ni}.$$

Similarly, for BDFs, $A_n[Dx]'_n = \tilde{A}_n[\tilde{D}x]'_n$ holds true. Thus, (4.1) is numerically equivalent to (4.12). If, additionally, $\ker K(t) = \ker A(t)$, then we conclude from $\ker K(t) \oplus \mathrm{im}\, D(t) = \mathbb{R}^m$ that $\mathrm{im}\, \tilde{D}(t) = \mathrm{im}\, K(t)D(t) = \mathrm{im}\, K(t)$. Further, we have $\mathrm{im}\, K(t) = \mathrm{im}\, K(t)D(t) = \mathrm{im}\, K(t_*)D(t) \subseteq \mathrm{im}\, K(t_*)$. Since $K(t)$ has constant rank, $\mathrm{im}\, K(t)$ is time-invariant, and consequently also $\mathrm{im}\, \tilde{D}(t)$.

(ii) Again, by Proposition 4.4, $\tilde{A}$ and $\tilde{D}$ are well matched. Now, condition (4.13) leads to $AK' = 0$ such that (4.3) and (4.12) coincide. For Runge-Kutta methods, (4.13) implies

$$A_{ni}[DX]'_{ni} = A_{ni}[K\tilde{D}X]'_{ni} = A_{ni}K_{ni}[\tilde{D}X]'_{ni} = \tilde{A}_{ni}[\tilde{D}X]'_{ni}.$$

Similarly, for BDFs, it holds that $A_n[Dx]'_n = \tilde{A}_n[\tilde{D}x]'_n$. Hence, (4.1) and (4.12) are numerically equivalent.
In case of $\operatorname{im} D(t) = \operatorname{im} K(t)$, due to $\ker A(t) \oplus \operatorname{im} K(t) = \mathbb{R}^m$ we easily derive that $\ker \tilde{A}(t) = \ker A(t)K(t) = \ker K(t)$. Further,

$$\ker K(t) = \ker A(t)K(t) = \ker A(t)K(t_*) \supseteq \ker K(t_*).$$

Because of the constant rank of $K(t)$, it holds that $\ker K(t) = \ker K(t_*)$, in other words, $\ker \tilde{A}(t) = \ker \tilde{A}(t_*)$ has to be true. But then (4.12) is numerically well formulated. □

In the special cases 2–5 discussed above, the assumptions in Proposition 4.5, except for (4.11) and (4.13), are given by construction. So we arrive at the following.

Corollary 4.6 *If any of the following conditions holds, then (4.1) is numerically equivalent to a numerically well formulated DAE:*

(i) There is a continuously differentiable projector $P_A(t)$ along $\ker A(t)$ such that

$$P_A(t_*)D(t) = P_A(t)D(t), \quad t_*, t \in \mathcal{I}. \tag{4.14}$$

(ii) $A(t)$ is continuously differentiable and it holds that

$$A(t_*)D(t) = A(t)D(t), \quad t_* t \in \mathcal{I}. \tag{4.15}$$

(iii) $D(t)$ is continuously differentiable and it holds that

$$A(t)D(t_*) = A(t)D(t), \quad t_*, t \in \mathcal{I}. \tag{4.16}$$

(iv) There is a continuously differentiable projector $R_D(t)$ onto $\operatorname{im} D(t)$ such that

$$A(t)R_D(t_*) = A(t)R_D(t), \quad t_*, t \in \mathcal{I}. \tag{4.17}$$

Example 4.7 *Continue the discussion of the DAE (4.9) resp. (4.6) in Example 4.2. In fact, we have a DAE of the form*

$$A(t)(D(t)x(t))' + B(t)x(t) = 0$$

with

$$A(t) = \begin{pmatrix} 1 & p(t) & p(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; D(t) = P_{\text{can}}(t) = \begin{pmatrix} 1 & p(t) & p(t) \\ -q(t) & -p(t)q(t) & -p(t)q(t) \\ q(t) & p(t)q(t) & p(t)q(t) \end{pmatrix}.$$

Observe that

$$A(t_*)D(t) = A(t)D(t) \text{ for all } \quad t_*, t,$$

i.e.,(4.9) is numerically equivalent to a numerically well formulated DAE. This confirms what was mentioned in Example 4.2. Observe further that

$$\ker A(t_*) \oplus \text{im } D(t) = \mathbb{R}^m \qquad \text{is true for all } t_*, t.$$

Unfortunately, the conditions in Proposition 4.5 are not easy to verify in general. We may be lucky to find decompositions satisfying (4.11) or (4.13). However, the fact that we do not succeed in finding any, does not mean that they do not exist. Consequently, some criteria ensuring that such factorizations do not exist would also be helpful.

Lemma 4.8 *Given an index-1 DAE with well matched coefficients A and D.*

(i) $\ker K(t) \oplus \text{im } D(t) = \mathbb{R}^m$ *and (4.11) imply*

$$\ker K(t_*) \oplus \text{im } D(t) = \mathbb{R}^m \qquad \text{for all } \ t, t_* \in \mathcal{I}.$$

(ii) $\ker A(t) \oplus \text{im } K(t) = \mathbb{R}^m$ *and (4.13) imply*

$$\ker A(t) \oplus \text{im } K(t_*) = \mathbb{R}^m \qquad \text{for all } \ t, t_* \in \mathcal{I}.$$

Proof:

(i) Recall that, due to condition (4.11), im $K(t)$ is constant. We show the relations $\ker K(t_*) \cap \text{im } D(t) = \{0\}$, $\ker K(t_*)D(t) = \ker D(t)$, $\text{im } K(t_*)D(t) = \text{im } K(t_*)$ to be true. Then, by Lemma 6.1 (Appendix) the desired assertion results.
Let $x \in \ker K(t_*) \cap \text{im } D(t)$, i.e., $K(t_*)x = 0$, $x = D(t)w$ for some w. Due to condition (4.11), it holds that

$$K(t_*)x = K(t_*)D(t)w = K(t)D(t)w = K(t)x.$$

This leads to $K(t)x = 0$, thus $x \in \ker K(t) \cap \text{im } D(t) = \{0\}$. Finally,

$$\ker K(t_*)D(t) = \ker K(t)D(t) = \ker D(t)$$

and

$$\text{im } K(t_*)D(t) = \text{im } K(t)D(t) = \text{im } K(t) = \text{im } K(t_*).$$

(ii) The conditions imposed imply for the dual spaces

$$\ker K(t)^* \oplus \text{im } A(t)^* = \mathbb{R}^m,$$

$$K(t_*)^*A(t)^* = K(t)^*A(t)^* \quad \text{for all } t, t_* \in \mathcal{I}.$$

Applying part (i) we obtain $\ker K(t_*)^* \oplus \text{im } A(t)^* = \mathbb{R}^m$, and, equivalently, $\ker A(t) \oplus \text{im } K(t_*) = \mathbb{R}^m$. □

Note 4.9 *Let (4.1) have a properly stated leading term. If all conditions of either Proposition 4.5 (i) or Proposition 4.5 (ii) are satisfied, then*

$$\ker A(t) \oplus \operatorname{im} D(t_*) = \mathbb{R}^m \quad \text{for all } t, t_* \in \mathcal{I}. \tag{4.18}$$

Proof: The conditions of Proposition 4.5 (i) (respectively (ii)) immediately imply the conditions of Lemma 4.8 (i) (respectively (ii)) to be satisfied. Consequently,

$$\ker A(t_*) \oplus \operatorname{im} D(t) = \mathbb{R}^m. \qquad \square$$

For numerically well formulated DAEs (4.1), either im $D(t)$ or $\ker A(t)$ are constant, and (4.18) holds trivially. For DAEs with time-dependent im $D(t)$ and $\ker A(t)$, (4.18) is a necessary condition for the DAE (4.1) to be numerically equivalent to a numerically well formulated DAE. Therefore, if (4.18) is satisfied, before making expensive reformulations, we should consider the possibility that the given DAE is actually numerically well formulated, but we simply have a bad matrix representation of the problem as it was the case in Example 4.2.

On the other hand, if (4.18) is not given, we have to reformulate the problem to obtain a numerically well formulated problem. This is the situation described by Example 4.1. Let us stress that (4.18) is necessary but not sufficient for the DAE (4.1) to have a numerically well formulated representation.

5 Generalizations and an application

In [9], linear equations $A(t)(D(t)x(t))' + B(t)x(t) = q(t)$ are considered with possibly rectangular matrix coefficients $D(t) \in L(\mathbb{R}^m, \mathbb{R}^n)$, $A(t) \in L(\mathbb{R}^n, \mathbb{R}^m)$ within the leading term. This idea may be immediately combined with the suggestion of [8] to deal with more general equations

$$A(x(t), t)(d(x(t), t))' + b(x(t), t) = 0. \tag{5.1}$$

The additional, possibly nonlinear function $d(x, t)$ is assumed to be continuous and to have the continuous partial derivative $d'_x(x, t) =: D(x, t)$.

Definition 5.1 *The leading term of (5.1) is properly stated if* $A(x, t) \in L(\mathbb{R}^n, \mathbb{R}^m)$, $D(x, t) \in L(\mathbb{R}^m, \mathbb{R}^n)$, $\ker A(x, t) \oplus \operatorname{im} D(x, t) = \mathbb{R}^n$ *for all* x, t *from the definition domain, and if there is a projector function* $R(t) \in L(\mathbb{R}^n)$ *continuously differentiable with respect to* t *such that* $\ker R(t) = \ker A(x, t)$, $\operatorname{im} D(x, t) = \operatorname{im} R(t)$ *as well as* $d(x, t) = R(t)d(x, t)$.

Actually, if the leading term is properly stated, the characteristic subspaces $\ker A(x, t)$ and im $D(x, t)$ have constant dimension and do not depend on x.

Definition 5.2 *The DAE (5.1) with properly stated leading term has index 1, if*

$$N_0(x,t) \cap S_0(y,x,t) = \{0\}$$

on the definition domain, where

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

$$S_0(y,x,t) := \{z \in \mathbb{R}^m : \{b'_x(x,t) + (A(x,t)y)'_x\}z \in \operatorname{im} A(x,t)\}.$$

Definition 5.3 *The index 1 DAE (5.1) is numerically well formulated if* $\operatorname{im} D(x,t)$ *is constant.*

Since (5.1) is equivalent to the enlarged system ([8])

$$\left.\begin{aligned} A(x(t),t)(R(t)y(t))' + b(x(t),t) &= 0 \\ y(t) - d(x(t),t) &= 0 \end{aligned}\right\}, \tag{5.2}$$

which is of the form (2.1), all results relying on (2.1) may be immediately reformulated for (5.1) via (5.2).

The formulation (2.1) resp. (5.1) of differential algebraic systems seems to be quite natural also from the point of view of applications. We want to show this for DAEs arising in circuit simulation. Using the modified nodal analysis (MNA), which is one of the most applied modelling techniques in circuit simulation packages, we obtain ([3]) a system of the form

$$\begin{aligned} A_C\left(q(A_C^T e(t),t)\right)' + A_R r\left(A_R^T e(t),t\right) + A_L j_L(t) + A_V j_V(t) \qquad & \\ + A_I i\left(\bar{A}^T e(t), j_L(t), j_V(t), t\right) &= 0, \qquad (5.3)\\ (\Phi(j_L(t),t))' - A_L^T e(t) &= 0, \qquad (5.4)\\ A_V^T e(t) - v\left(\bar{A}^T e(t), j_L(t), j_V(t), t\right) &= 0. \qquad (5.5) \end{aligned}$$

Here, the unknowns are the nodal potential $e(t)$, the currents of inductances $j_L(t)$ and the currents of voltage sources $j_V(t)$. The matrix $\bar{A} = (A_C, A_L, A_R, A_V, A_I)$ represents the incidence matrix (which is constant and has just entries from $\{-1,0,1\}$) describing the branch and node relations. More precisely, A_C corresponds to all capacitive, A_L to all inductive, A_R to all resistive, A_V to all voltage source and A_I to all current source branches. Obviously, with

$$x = \begin{pmatrix} e \\ j_L \\ j_V \end{pmatrix}, \quad A = \begin{pmatrix} A_C & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix}, \quad d(x,t) := \begin{pmatrix} q(A_C^T e, t) \\ \Phi(j_L, t) \end{pmatrix}$$

and

$$b(x,t) := \begin{pmatrix} A_R r\left(A_R^T e, t\right) + A_L j_L + A_V j_V + A_I i\left(\bar{A}^T e, j_L, j_V, t\right) \\ -A_L^T e \\ A_V^T e - v\left(\bar{A}^T e, j_L, j_V, t\right) \end{pmatrix}$$

the system (5.3)-(5.5) can be rewritten as

$$A(d(x(t),t))' + b(x(t),t) = 0. \tag{5.6}$$

If e.g. all capacitances and inductances are linear (possibly time-varying), then

$$q\left(A_C^T e, t\right) = C(t)A_C^T e, \quad \Phi(j_L, t) = L(t)j_L,$$

with positive-definite diagonal matrices $C(t)$ and $L(t)$. In this case, we have

$$D(t) = d'_x(x,t) = \begin{pmatrix} C(t)A_C^T & 0 & 0 \\ 0 & L(t) & 0 \end{pmatrix} = \begin{pmatrix} C(t) & 0 \\ 0 & L(t) \end{pmatrix} A^T$$

and

$$AD(t) = \begin{pmatrix} A_C C(t) A_C^T & 0 & 0 \\ 0 & L(t) & 0 \\ 0 & 0 & 0 \end{pmatrix} = A \begin{pmatrix} C(t) & 0 \\ 0 & L(t) \end{pmatrix} A^T.$$

It holds that $\ker AD(t) = \ker D(t)$, $\operatorname{im} AD(t) = \operatorname{im} A$, $\ker A \cap \operatorname{im} D(t) = \{0\}$. If the entries of $C(t)$ and $L(t)$ are continuously differentiable, then the respective projector function $R(t)$ belongs to the class C^1 and the leading term in (5.6) is properly stated. In the case of nonlinear capacitances and/or inductances, instead of the matrices $C(t)$ and $L(t)$ the respective partial derivatives of the function q and Φ have to be considered.
Observe that the matrix A is always constant, hence the subspace $\ker A$ is constant, which leads to numerically well formulated DAEs in the index-1 case (cf. argumentation following Definition 2.4 on page 7). For index-1 conditions concerning the special structure of (5.3)-(5.5) we refer to [3].

It should be emphasized that whether the leading term in (2.1) or (5.1) is stated properly or not is independent of the index of this DAE. Consequently, those considerations concerning possible reformulations and rearrangements (e.g. Proposition 2.5, Proposition 4.4) hold true also for higher index. However, unfortunately, the condition for the subspace $\operatorname{im} D(t)$ to be constant is necessary but not sufficient for a numerical well-formulation if the index is greater than one. Just in the index-2 case, things are much more difficult. First results for linear index-2 equations are reported in [7], [6], [10] but in [11] for nonlinear equations. In these papers the matrix D is assumed to be a constant projector. We are planning to clarify what happens in more general index-2 cases in a forthcoming paper.

6 Appendix

Lemma 6.1 *Given matrices $D \in L(\mathbb{R}^m, \mathbb{R}^n)$, $A \in L(\mathbb{R}^n, \mathbb{R}^m)$. Then the relation $\ker A \oplus \operatorname{im} D = \mathbb{R}^n$ holds true if*

$$\ker AD = \ker D, \quad \operatorname{im} AD = \operatorname{im} A, \quad \ker A \cap \operatorname{im} D = \{0\}, \tag{6.1}$$

and vice versa.

Proof: If $\ker A \oplus \text{im } D = \mathbb{R}^n$, we may use the projector $R \in L(\mathbb{R}^n)$ onto im D along $\ker A$. Then, $A = AR$, $D = RD$. The relation $\ker A \cap \text{im } D = \{0\}$ holds trivially, further im $AD \subseteq$ im A, $\ker AD \supseteq \ker D$. Taking $z \in \ker AD$, i.e., $Dz \in \text{im } D \cap \ker A = \{0\}$ we may conclude $z \in \ker D$ and, finally, $\ker AD = \ker D$. Taking $z \in$ im A, i.e., $z = Aw = ARw = AD\tilde{w}$ we obtain $z \in$ im AD.
Now, let (6.1) be given and $r := \text{rank}(AD)$. Here we have

$$\dim \ker A = n - \text{rank}A = n - \text{rank}AD = n - r$$

as well as

$$\dim \text{im } D = \text{rank}D = m - \dim \ker D = m - \dim \ker AD = m - (m - r) = r.$$

For dimensional reasons, $\ker A \cap \text{im } D = \{0\}$ implies $\ker A \oplus \text{im } D = \mathbb{R}^n$. □

Lemma 6.2 *Given matrices $D \in L(\mathbb{R}^m, \mathbb{R}^n)$, $A \in L(\mathbb{R}^n, \mathbb{R}^m)$. Then*

$$\ker A \oplus \text{im } D = \mathbb{R}^n$$

holds true if and only if

$$\ker D^* \oplus \text{im } A^* = \mathbb{R}^n.$$

Proof: Supposed that $\ker A \oplus \text{im } D = \mathbb{R}^n$ is true, we use again the projector $R \in L(\mathbb{R}^n)$ with $\ker R = \ker A$, im $R =$ im D. Then, R^* is also a projector, and we obtain

$$\text{im } R^* = \ker R^{\perp} = \ker A^{\perp} = \text{im } A^*, \quad \ker R^* = \text{im } R^{\perp} = \text{im } D^{\perp} = \ker D^*,$$

thus $\ker D^* \oplus \text{im } A^* = \mathbb{R}^m$. □

References

[1] 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.

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

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

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

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

[6] 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.

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

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

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

[10] R. März and A. R. Rodriguez Santiesteban: Analyzing the stability behaviour of DAE solutions and their approximations. Humboldt-Universität Berlin, Institut für Mathematik, Preprint 99-2.

[11] A. R. Rodriguez Santiesteban: Asymptotische Stabilität von Index-2-Algebro-Differentialgleichungen. Dissertation, Humboldt-Universiät Berlin, Institut für Mathematik, 2001.

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