Analyzing the stability behaviour of DAE solutions and their approximations

Roswitha März
and
Antonio R. Rodríguez Santiesteban

Humboldt-University Berlin
Institute of Mathematics
Unter den Linden 6
D-10099 Berlin, Germany

Contents

1 Introduction

Usually, lower-index differential-algebraic equations (DAEs)

$$f(x'(t), x(t), t) = 0 \tag{1.1}$$

are integrated by numerical algorithms that have been developed originally for regular ordinary differential equations (ODEs). There are numerous papers justifying this by convergence proofs, order results etc. (cf. [1], [2], [3]). In particular, the backward differentiation formula (BDF)

$$f\left(\frac{1}{h}\sum_{j=0}^{k}\alpha_j x_{n-j}, x_n, t_n\right) = 0, \tag{1.2}$$

but also implicit Runge-Kutta methods (IRK)

$$x_n = x_{n-1} + h\sum_{j=1}^{s}\beta_j X'_{nj}, \tag{1.3 a}$$

$$f\left(X'_{ni}, x_{n-1} + h\sum_{j=1}^{s}\alpha_{ij}X'_{nj}, t_{n-1} + c_i h\right) = 0, \quad i = 1, \ldots, s, \tag{1.3 b}$$

are used in applications with really great success. Step by step each method provides numerical approximations x_n of the true solution values $x(t_n), t_n = t_{n-1} + h$.

The convergence results mentioned above concern the behaviour of the global error on a compact interval, say $[t_0, T]$, as the stepsize of the discretization $t_0 < t_1 < \cdots < t_n = T$ tends to zero.
However, what do we know about the error $x(t_n) - x_n$ if the stepsize h is constant and $n \to \infty$, that is, $t_n \to \infty$? When integrating regular ODEs numerically, one tries to carefully match the dynamics of the numerical algorithm with the dynamical behaviour of the true solution. How could this be done in case of DAEs?

Classical linear stability theory is concerned with the analysis of approximating the so-called scalar linear test equation

$$z'(t) = \lambda z(t), \quad \mathrm{Re}\lambda < 0. \tag{1.4}$$

Basic notions like the region of absolute stability, error growth function etc., rely on (1.4) and apply, via similarity transforms, to constant coefficient regular ODEs

$$x'(t) - Wx(t) = 0. \tag{1.5}$$

Considering DAEs, the respective constant coefficient system

$$Ax'(t) + Bx(t) = 0 \tag{1.6}$$

has a singular leading coefficient matrix A, but a regular matrix pencil $\{A, B\}$, i. e., $\det(\lambda A + B) \not\equiv 0$. There are nonsingular matrices E, F that transform (1.6) into its so-called Kronecker normal form

$$\begin{pmatrix} I & \\ & J \end{pmatrix} \tilde{x}'(t) + \begin{pmatrix} -W & \\ & I \end{pmatrix} \tilde{x}(t) = 0, \tag{1.7}$$

where $EAF = \begin{pmatrix} I & \\ & J \end{pmatrix}$, $EBF = \begin{pmatrix} -W & \\ & I \end{pmatrix}$, $\tilde{x} = F^{-1}x$, and J is a nilpotent block, $\operatorname{ind}(J) = \operatorname{ind}\{A, B\}$.
Obviously, the solution of the homogeneous equation (1.6) consists of its "dynamical part" only, i. e.,

$$x(t) = F\tilde{x}(t) = F \begin{pmatrix} \tilde{u}(t) \\ 0 \end{pmatrix},$$

where $\tilde{u}(t)$ solves the regular ODE $\tilde{u}'(t) = W\tilde{u}(t)$.
On the other hand, we may apply the same transformations E, F to decouple the BDF applied to (1.6), that is,

$$A\frac{1}{h}\sum_{j=0}^{k} \alpha_j x_{n-j} + Bx_n = 0$$

to

$$\begin{pmatrix} I & \\ & J \end{pmatrix} \frac{1}{h}\sum_{j=0}^{k} \alpha_j \tilde{x}_{n-j} + \begin{pmatrix} -W & \\ & I \end{pmatrix} \tilde{x}_n = 0.$$

Starting with consistent values $x_0, \cdots, x_{k-1}$ (such that $F^{-1}x_j = \begin{pmatrix} \tilde{u}_j \\ 0 \end{pmatrix}$, $j = 0, \cdots, k$) we obtain

$$x_n = F\tilde{x}_n = F \begin{pmatrix} \tilde{u}_n \\ 0 \end{pmatrix},$$

where $\tilde{u}_n$ is given by the BDF applied to the regular ODE $\tilde{u}'(t) = W\tilde{u}(t)$, i. e.,

$$\frac{1}{h}\sum_{j=0}^{k} \alpha_j \tilde{u}_{n-j} = W\tilde{u}_n.$$

In the same way we may proceed with Runge-Kutta methods. Obviously, the rich world of classical stability theory applies to constant coefficient DAEs (1.6) via transformation into Kronecker normal form plus similarity transform.
However, unfortunately, the homogeneous coefficient DAE (1.6) does not play a similar role in the DAE analysis as equation (1.5) does in the regular ODE-theory. DAEs represent a much more complex class of problems. The constant coefficient

case (1.6) is only a very poor model, which does not reflect important geometric features of DAEs at all.

Positive results concerning long term integrations of index-1 DAEs are reported in [4]. There, the leading nullspace is supposed to remain invariant. Moreover, in [4] it is stressed that varying subspaces may have its effect.
In [5], the following small example was discussed to show the bad effect of a rotating nullspace. The linear index-1 DAE

$$\begin{pmatrix} \delta-1 & \delta t \\ 0 & 0 \end{pmatrix} x'(t) + \mu \begin{pmatrix} \delta-1 & \delta t \\ \delta-1 & \delta t - 1 \end{pmatrix} x(t) = 0 \tag{1.8}$$

with real parameters $\mu \neq 0,\ \delta \neq 1$ has the solutions

$$\begin{aligned} x_1(t) &= (\delta-1)^{-1}(1-\delta t)x_2(t), \\ x_2(t) &= \exp((\delta-\mu)t)x_2(0). \end{aligned}$$

The nullspace of the leading coefficient matrix is

$$N(t) := \left\{ z \in \mathbb{R}^2 : (\delta-1)z_1 + \delta t z_2 = 0 \right\}.$$

For $\delta \neq 0$, this nullspace varies with t. Applying the backward Euler method to (1.8) yields

$$\begin{aligned} x_{n,1} &= (\delta-1)^{-1}(1-\delta t_n)x_{n,2}, \\ x_{n,2} &= \tfrac{1+h\delta}{1+h\mu} x_{n-1,2}. \end{aligned}$$

Obviously, it holds that $x_n \to 0$ $(n \to \infty)$ if and only if $|1+h\delta| < |1+h\mu|$. For $h\mu = -1$ the method is not feasible at all. Further, there is a large region on the $(h\delta, h\mu)$-plane where the true solution $x(t_n)$ vanishes asymptotically, but the numerical approximation grows unboundedly.
Note that the spectrum of the pencil $\{A(t), B(t)\}$ is time-invariant, namely

$$\sigma(A(t), B(t)) := \{\lambda \in \mathbb{C} : \det(\lambda A(t) + B(t)) = 0\} = \{-\mu\}.$$

If we put $\delta = 0$, we obtain again a constant coefficient DAE and everything is fine as expected before.

One could think that the strange asymptotic behaviour is due to the rotations of the nullspace of the leading coefficient $A(t)$ or the leading partial Jacobian $f'_{x'}(x', x, t)$ in (1.1). Since this matrix has a constant nullspace or is constant itself in most applications, we can perhaps consider equation (1.8) to be just an academic example. However, similar phenomena of a "wrong numerical dynamical behaviour" may arise even in the case of a constant leading coefficient matrix. Such phenomena are discussed in [6], where linear index-2 DAEs are considered.

Roughly speaking, in [6] it is shown that the numerical approximations fit the dynamical behaviour of the true solution well, supposed both the characteristic subspaces $N_1(t)$ and $S_1(t)$ do not vary with t in fact.
The aim of the present paper is to show that the same invariance condition for $S_1(t)$ will do. This weaker sufficient condition ensures also an appropriate reflection of the exponential decay. Since e.g. in circuit simulation $S_1(t)$ is often invariant but $N_1(t)$ is not, our new approach is of particular interest for those applications.
This paper is organized as follows. In Section 2 the related analytical background is given. Useful analytical decoupling and reduction techniques are discussed. As a by product, Theorem 2.2 provides new solvability results with respect to weaker smoothness demands. Also contractivity is now discussed under those weaker smoothness conditions.
In Section 3 the analytical decoupling and reduction techniques are applied to BDF and IRK methods, to prove that the time invariance of the subspace $S_1(t)$ will do in fact.
Finally we demonstrate by two characteristic examples that are even in Hessenberg form, how a rotating subspace $S_1(t)$ effects the dynamic reflection.

2 Analysis of linear index-2 DAE's

2.1 Fundamentals

Consider the linear equation

$$A(t)x'(t) + B(t)x(t) = q(t), \ t \in J := [t_0, \infty), \tag{2.1}$$

where the coefficients are continuous and the matrix $A(t) \in L(\mathbb{R}^m)$ is singular for all $t \in J$, but has constant rank r. Introduce the basic subspaces

$$\begin{aligned} N(t) :=& \ \ker A(t), \\ S(t) :=& \ \{z \in \mathbb{R}^m : B(t)z \in \text{ im } A(t)\}. \end{aligned}$$

Obviously, each solution of the homogeneous equation satisfies $x(t) \in S(t), t \in J$.

In the following, we assume $N(t)$ to be spanned by $m-r$ continuously differentiable base functions. Then, there is *a* C^1 matrix function $Q : J \to L(\mathbb{R}^m)$ that projects $\mathbb{R}^m$ pointwise onto $N(t)$, i. e., $Q(t)^2 = Q(t)$, im $Q(t) = N(t), t \in J$.
In what follows, Q denotes such *a* C^1 projector function onto N, and $P := I - Q$.
Recall (e.g. [2]) the equation

$$A(t)\{(Px)'(t) - P'(t)x(t)\} + B(t)x(t) = q(t), \quad t \in J, \tag{2.2}$$

to be the precise formulation of (2.1). We are looking for continuous functions $x(\cdot) : J \to I\!R^m$ that have continuously differentiable parts $(Px)(\cdot)$, and that satisfy (2.2) pointwisely. Denote this function space by

$$C^1_N := \left\{ y \in C(J, I\!R^m) : Py \in C^1(J, I\!R^m) \right\}.$$

Next, we introduce further subspaces which are relevant for index-2 DAEs [2], namely

$$\begin{aligned} N_1(t) &:= \ker A_1(t), \\ S_1(t) &:= \{ z \in I\!R^m : B(t)P(t)z \in \text{ im } A_1(t) \}, \\ A_1(t) &:= A(t) + (B(t) - A(t)P'(t))Q(t), \quad t \in J. \end{aligned}$$

Definition: *The DAE (2.1) is said to be index-2 tractable if* $\dim(N(t) \cap S(t)) = \nu > 0,\ N_1(t) \cap S_1(t) = \{0\}, \quad t \in J.$

Recall that index-2 tractability generalizes the case of Kronecker index 2. Let $Q_1(t)$ denote the projector onto $N_1(t)$ along $S_1(t), t \in J$.

Due to [4], Lemma A.13, the matrix

$$G_2(t) := A_1(t) + B(t)P(t)Q_1(t)$$

remains nonsingular now and it holds that $Q_1(t) = Q_1(t)G_2(t)^{-1}B(t)P(t)$ on J. Since we may represent

$$\ker A_1(t) = (I - P(t)A(t)^+(B(t) - A(t)P'(t))Q(t))\,(N(t) \cap S(t)),$$

we know $N_1(t)$ to have the same constant dimension ν as $N(t) \cap S(t)$.

In the following, we drop the argument t if possible.
By straightforward calculations we prove the relations

$$G_2^{-1}A = P_1P, \quad G_2^{-1}B = G_2^{-1}BPP_1 + Q_1 + Q + P_1PP'Q.$$

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

$$P_1P((Px)' - P'x) + G_2^{-1}BPP_1x + Q_1x + Qx + P_1PP'Qx = G_2^{-1}q. \tag{2.3}$$

Multiplying (2.3) by $PP_1,\ QP_1$ and Q_1, respectively, and carrying out simple computations we obtain the decoupled system

$$\begin{aligned} PP_1(Px)' + PP_1G_2^{-1}BPP_1x &= PP_1G_2^{-1}q, && (2.4) \\ -QQ_1(Px)' + QP_1G_2^{-1}BPP_1x + Qx &= QP_1G_2^{-1}q, && (2.5) \\ Q_1x &= Q_1G_2^{-1}q. && (2.6) \end{aligned}$$

The System (2.4), (2.5), (2.6) provides the basic idea of what the solutions of the DAE (2.1) look like in case of PP_1, PQ_1 and $PQ_1G_2^{-1}q$ being C^1. Then (2.4), (2.5) may be rewritten as

$$(PP_1x)' - (PP_1)'PP_1x + PP_1G_2^{-1}BPP_1x = PP_1G_2^{-1}q + (PP_1)'PQ_1G_2^{-1}q \tag{2.4'}$$

$$\begin{aligned} Qx = -(QP_1G_2^{-1}B + QQ_1(PQ_1)')PP_1x \;&+\; QQ_1(PQ_1G_2^{-1}q)' + QP_1G_2^{-1}q \\ &-\; QQ_1(PQ_1)'PQ_1G_2^{-1}q \end{aligned} \tag{2.5'}$$

and the solution is composed by $x = PP_1x + PQ_1x + Qx$, where the component PP_1x solves the inherent regular ODE (2.4'), Qx and PQ_1x are given by (2.5') and (2.6), respectively. (cf. [2]).

In the next section, we will generalize the solvability results given in [8] in the sense that we will do with lower smoothness. The new inherent regular ODE, which uses a different projector instead of PP_1, permits to obtain new insights into the asymptotic stability behaviour of numerical approximations (Section 3).
Note that, by construction, $P(t)P_1(t)$ projects onto $P(t)S_1(t)$ along $N(t) \oplus N_1(t)$, i. e., it is related to the decomposition

$$I\!R^m = P(t)S_1(t) \oplus N(t) \oplus N_1(t). \tag{2.7}$$

2.2 Solvability and decoupling

Now we also use the orthoprojector $V(t)$, which projects along $S_1(t)$. Then, $I - V(t)$ is the orthoprojector onto $S_1(t)$. Because of $N(t) \subset S_1(t)$, it holds that $V(t)Q(t) = 0$, and

$$\Pi(t) := P(t)(I - V(t))$$

is a projector, too. Obviously, we have im $\Pi(t) = P(t)S_1(t)$. Since the two projectors $\Pi(t)$ and $P(t)P_1(t)$ have the same image space $P(t)S_1(t)$, it holds that $PP_1\Pi = \Pi, \Pi PP_1 = PP_1$.
Additionally, we denote the orthoprojector onto im $A_1(t)$ by $I - W(t)$.
As we will realize below, the projection W extricates exactly that derivative free part of equations that has to be differentiated once when reducing the index.

The following example of a Hessenberg form DAE is to make the meaning of the different projectors more transparent.

Example: *Given a Hessenberg form DAE of size 2 that contains n_1 equations with derivatives and n_2 derivative-free ones*

$$\left.\begin{array}{rll} x_1' + & B_{11}x_1 + B_{12}x_2 & = q_1 \\ & B_{21}x_1 & = q_2 \end{array}\right\}, \quad B_{21}B_{12} \textit{ nonsingular.}$$

Here we have a constant nullspace N, $N = \{z : z_1 = 0\}$, further

$$Q = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \quad A_1 = \begin{pmatrix} I & B_{12} \\ 0 & 0 \end{pmatrix}, \quad W = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$$

$$S(t) = S_1(t) = \{z : B_{21}(t)z_1 = 0\}, \ S(t) \cap N = N,$$

$$N_1(t) = \{z : z_1 + B_{12}(t)z_2 = 0\},$$

$$N_1(t) \cap S_1(t) = \{z : z_1 + B_{12}(t)z_2 = 0, \ B_{21}(t)z_1 = 0\} = \{0\}$$

due to the nonsingularity of $B_{21}B_{12}$.
The projector PP_1 is given by

$$PP_1 = \begin{pmatrix} I - L & 0 \\ 0 & 0 \end{pmatrix},$$

where $L = B_{12}(B_{21}B_{12})^{-1}B_{21}$ is also a projector ($L(t)$ projects $\mathbb{R}^{n_1}$ onto im $B_{12}(t)$ along $\ker B_{21}(t)$). The orthoprojector along $S_1(t)$ is now

$$\begin{aligned} V(t) &= \begin{pmatrix} B_{21}(t)^+B_{21}(t) & 0 \\ 0 & 0 \end{pmatrix}, \\ \textit{and} \quad \Pi(t) &= \begin{pmatrix} I - B_{21}(t)^+B_{21}(t) & 0 \\ 0 & 0 \end{pmatrix}, \\ \textit{im}\ \Pi(t) &= P(t)S_1(t) = \{z : B_{21}(t)z_1 = 0, \ z_2 = 0\} \\ &= \ker B_{21}(t) \times \{0\}. \end{aligned}$$

It is well-known that, in Hessenberg systems, all derivative-free equation should be differentiated when deriving a solution via a reduction step.
Consequently, it seems to be more natural to assume $B_{21} \in C^1$ (hence, $V, \Pi \in C^1$) instead of $L \in C^1$ (i. e., $PP_1 \in C^1$).
□

Next, we collect some nice properties of our projectors and matrices to be used below:

1) im $A_1 = \ker W$, im $A_1 = \ker G_2PQ_1G_2^{-1}$, hence there are two projectors W and $G_2PQ_1G_2^{-1}$ along im A_1. Therefore

$$W = WG_2PQ_1G_2^{-1}, \quad G_2PQ_1G_2^{-1} = G_2PQ_1G_2^{-1}W.$$

2) From $WB = WG_2PQ_1G_2^{-1}B = WG_2PQ_1G_2^{-1}BPQ_1 = WBPQ_1$ and $PQ_1G_2^{-1}WB = PQ_1$ it follows that

$$\ker WB = \ker PQ_1 = \ker Q_1 = S_1, \qquad \text{im } WB = \text{ im } W.$$

3) $V = (WB)^+WB,\ W = (I - A_1A_1^+) = (PQ_1G_2^{-1})^+PQ_1G_2^{-1}$,
$WB = (PQ_1G_2^{-1})^+PQ_1.$

4) $V = VPQ_1,\ PQ_1 = PQ_1V,\ V = VP,\ V = VQ_1,$
$VG_2^{-1} = (WB)^+WBG_2^{-1} = (WB)^+WBPQ_1G_2^{-1}$
$= (WB^+WG_2PQ_1G_2^{-1}BPQ_1G_2^{-1} = (WB)^+WG_2PQ_1G_2^{-1} = (WB)^+W.$

The third property indicates that, for continuously differentiable $PQ_1G_2^{-1}$ and PQ_1, we also have continuously differentiable W and V. However, as we could realize by means of the special Hessenberg form DAE above, the opposite is not true. In this sense, Theorem 2.2 below generalizes the related results of [8].

To get a better insight we reconsider the decoupled form (2.4), (2.5), (2.6) of the DAE (2.1) and try to reformulate it in such a way that (2.4) returns into a regular ODE for the solution component Πx. Now, we do not assume PP_1, PQ_1 to be from C^1 as we did in order to obtain (2.4'), (2.5'), but we allow only P, Π and V to be continuously differentiable now.
Equation (2.6) yields immediately

$$Vx = (WB)^+Wq. \tag{2.8}$$

To realize what equation (2.4) looks like in more detail we derive

$$\begin{aligned} PP_1(Px)' &= PP_1(\Pi + PV)(Px)' = \Pi(Px)' + PP_1V(Px)' \\ &= (\Pi x)' - \Pi' Px + PP_1(Vx)' - PP_1V'Px. \end{aligned}$$

Consequently, (2.4) can be rewritten as

$$\begin{aligned} (\Pi x)' - \Pi'(\Pi x+ &\ PVx) - PP_1V'(\Pi x + PVx) + PP_1(Vx)' \\ &+ PP_1G_2^{-1}BPP_1(\Pi x + PVx) = PP_1G_2^{-1}q. \end{aligned} \tag{2.9}$$

Analogously, with $QQ_1(Px)' = QQ_1V(Px)' = QQ_1(Vx)' - QQ_1V'(\Pi x + PVx)$, equation (2.5) can be reformulated:

$$-QQ_1(Vx)' + QQ_1V'(\Pi x + PVx) + QP_1G_2^{-1}BPP_1(\Pi x + PVx) + Qx = QP_1G_2^{-1}q. \tag{2.10}$$

If $x \in C^1_N$ solves the DAE (2.1), then $(WB)^+Wq = Vx = VPx$ belongs to C^1 since V and Px do so. We are allowed to replace Vx and $(Vx)'$ in (2.9) and (2.10) by means of (2.8). This gives rise to consider the ODE

$$\begin{aligned} u' - \Pi' u \ & -PP_1V'u + PP_1G_2^{-1}BPP_1u \\ &= \Pi' P(WB)^+Wq + PP_1V'P(WB)^+Wq - PP_1G_2^{-1}BPP_1(WB)^+Wq \\ &\quad -PP_1((WB)^+Wq)' + PP_1G_2^{-1}q, \end{aligned} \tag{2.11}$$

and to call it an inherent regular ODE. If we multiply equation (2.11) by $(I - \Pi)$, we obtain

$$(I - \Pi)u' - (I - \Pi)\Pi' u = 0,$$

hence

$$((I - \Pi)u)' + \Pi'(I - \Pi)u = 0.$$

It results that if $u \in C^1$ solves (2.11), then the function $\omega = (I - \Pi)u$ satisfies the homogeneous regular ODE $\omega' + \Pi'\omega = 0$. If, additionally, $\omega(t_*) = (I - \Pi(t_*))u(t_*) = 0$, then $\omega(t)$ vanishes identically, i. e., $u(t) = \Pi(t)u(t),\ t \in J$. This proves the following assertion.

Lemma 2.1 *The subspace im $\Pi(t) \subset I\!R^m$ represents an invariant subspace of the regular inherent ODE (2.11).*

As far as the homogeneous case $q = 0$ is concerned, we know from (2.8) that each solution of $Ax' + Bx = 0$ has a trivial component $Vx = 0$. Using (2.9), (2.10) and taking into account the inherent regular ODE we may express each solution as $x = \Pi x + Qx = \Pi_{can}u$, where u solves (2.11) as well as the initial condition $u(t_o) \in$ im $\Pi(t_0)$, and

$$\Pi_{can} := KPP_1,\ K := (I - QP_1G_2^{-1}BPP_1 - QQ_1V'PP_1).$$

The matrix function K is nonsingular, thus $\ker \Pi_{can}(t) = \ker P(t)P_1(t),\ t \in J$. It is easily checked that Π_{can} is also a projector. Π_{can} is said to be the canonical projector for the index-2 case. It projects onto the solution space of the homogeneous DAE along $N \oplus N_1$, as we will see below.

Theorem 2.2 *Given an index-2 DAE (2.1) with continuously differentiable subspaces N, S_1 and im A_1. Additionally, let WB be C^1 and $q \in \{p \in C : Wp \in C^1\}$.*

(i) Then, the IVP for (2.1) with the initial condition $\Pi(t_0)(x(t_0) - x^0) = 0$, $x^0 \in I\!R^m$, has exactly one C^1_N-solution.

(ii) Exactly one solution of the homogeneous equation at t_0 passes through each $x_0 \in$ im $\Pi_{can}(t_0)$.

Proof:

(i) Denote by $u \in C^1$ the solution of the regular ODE (2.12) that satisfies the initial condition $u(t_0) = \Pi(t_0)x^0$. Then we compose the function $x = u+v+w$,
$v := (WB)^+ Wq \in C^1$,
$w := QP_1G_2^{-1}q + QQ_1v' - QQ_1V'(u+Pv) - QP_1G_2^{-1}BPP_1(u+Pv)$.
Due to this construction we have $v = Vv,\ w = Qw,\ Px = u + Pv \in C^1$,
$x \in C,\ \Pi(t_0)x(t_0) = u(t_0) = \Pi(t_0)x^0$.
Finally, straightforward checking shows that x satisfies the DAE (2.1), indeed.

(ii) Solve the special IVP for $q = 0$ and $x^0 := x_0 = \Pi_{can}(t_0)x_0$. Its solution is $x = \Pi_{can}u$, and we have, in particular,

$$
\begin{aligned}
x(t_0) &= \Pi_{can}(t_0)u(t_0) = \Pi_{can}(t_0)\Pi(t_0)x_0 \\
&= \Pi_{can}(t_0)\Pi(t_0)\Pi_{can}(t_0)x_0 = \Pi_{can}(t_0)\Pi(t_0)(PP_1)(t_0)x_0 \\
&= \Pi_{can}(t_0)(PP_1)(t_0)x_0 \\
&= \Pi_{can}(t_0)x_0 = x_0.
\end{aligned}
$$

□

Corollary: *On each compact interval $[t_0, T]$, the perturbation index of an index-2 tractable DAE is two.*

Proof: For each $T > t_0$, there is a constant K_T such that the estimation

$$\|x\|_T \le K_T \left\{ |\Pi(t_0)x^0| + \|q\|_T + \|(Vq)'\|_T \right\} \tag{2.12}$$

holds for all IVP solutions with $x^0 \in I\!R^m,\ q \in C,\ Vq \in C^1$, where
$\|y\|_T := \max\{|y(t)| : t \in [t_0, T]\}$.

□

Example: *For the Hessenberg form DAE*

$$
\left.
\begin{array}{rlcl}
x_1' + & B_{11}x_1 + B_{12}x_2 & = & q_1 \\
& B_{21}x_1 & = & q_2
\end{array}
\right\}
$$

the conditions of Theorem(2.2) are satisfied if $q_1 \in C,\ q_2 \in C^1$, and $B_{21} \in C^1$. The initial condition reads in detail

$$0 = \Pi(t_0)(x(t_0) - x^0) = \begin{pmatrix} (I - B_{21}(t_0)^+ B_{21}(t_0))(x_1(t_0) - x_1^0) \\ 0 \end{pmatrix}.$$

2.3 Index reduction by differentiation

Considering the DAE (2.1) we observe immediately that each solution has always to proceed in the set

$$
\begin{aligned}
\mathcal{M}(t) := \ & \{z \in I\!R^m : B(t)z - q(t) \in \text{ im } A(t)\} = \\
& \{z \in I\!R^m : (I - W(t))(B(t)z - q(t)) \in \text{ im } A(t), W(t)(B(t)z - q(t)) = 0\},
\end{aligned}
$$

which is called a constraint manifold. Under the conditions of Theorem 2.2, the part of equations described by

$$WBx = Wq$$

can be differentiated to obtain

$$WBx' + (WB)'x = (Wq)'.$$

Here, WBx' stands for $WB\{(Px)' - P'x\}$. On the other hand, due to (2.1) we may express $Px' := (Px)' - P'x = PA^+(q - Bx)$. In consequence, the relation

$$WBA^+(q - Bx) + W(WB)'x = W(Wq)'$$

has to be satisfied additionally, i. e., the solution has also to proceed in the so-called hidden constraint manifold

$$\begin{aligned}\mathcal{H}(t) := \ &\{z \in I\!R^m : (-W(t)B(t)A(t)^+B(t) + W(WB)'(t))\, z \\ &= W(Wq)'(t) - W(t)B(t)A(t)^+q(t)\}\,.\end{aligned}$$

As we will see below, the set

$$\mathcal{M}_1(t) := \mathcal{M}(t) \cap \mathcal{H}(t)$$

is characteristic of the DAE (2.1). $\mathcal{M}_1$ contains all solutions (for given q), but it is filled by those solutions. One could call $\mathcal{M}_1(t)$ the state manifold. For the case of a homogeneous equation, $S(t) = \mathcal{M}(t)$ is trivially given. Moreover, there is a closed relationship between our subspace $S_1(t)$ and the state manifold $\mathcal{M}_1(t)$.

Lemma 2.3 *For a homogeneous index-2 tractable DAE (2.1) with continuously differentiable N, S_1, im A_1, WB, it holds that*

$$\mathcal{M}_1(t)|_{q=0} = \left\{z \in S_1(t) : Q(t)z = -((QP_1G_2^{-1}B + QQ_1G_2^{-1}(WB)')PP_1)(t)z\right\},$$

and

$$P(t)\mathcal{M}_1(t)|_{q=0} = P(t)S_1(t). \tag{2.13}$$

Proof:
The second relation is a simple consequence of the representation of $\mathcal{M}_1(t)$, which we want to realize now. We drop the argument t.
$z \in \mathcal{M}_1$ means by definition:
$-WBA^+Bz + W(WB)'z = 0,\ Bz + Aw = 0$ for a certain $w = Pw$.
This is equivalent to
$WBw + W(WB)'z = 0,\ w = -PA^+Bz, \quad PQ_1z = 0,$
$Qz - QQ_1w + QP_1G_2^{-1}BPP_1z + QP_1PP'Qz = 0,$
and to

$PQ_1z = 0,\ w = -PA^+Bz,\ 0 = WBw + W(WB)'PP_1z + W(WB)'Qz =$
$WBw + W(WB)'PP_1z + WBP'Qz,$
$Qz - QQ_1w - QQ_1P'Qz + QP_1G_2^{-1}BPP_1z = 0.$

Because of $QQ_1G_2^{-1}WB = QQ_1G_2^{-1}B = QQ_1$, $QQ_1G_2^{-1}W = QQ_1G_2^{-1}$, we find $z \in \mathcal{M}_1$ to be characterized by

$$PQ_1z = 0,\ w = -PA^+Bz,$$

$$QQ_1w + QQ_1P'Qz + QQ_1G_2^{-1}(WB)'PP_1z = 0$$

$$\begin{aligned} Qz &= QQ_1w + QQ_1P'Qz - QP_1G_2^{-1}BPP_1z \\ &= -QQ_1P'Qz + QQ_1P'Qz - QQ_1G_2^{-1}(WB)'PP_1z - QP_1G_2^{-1}BPP_1z \\ \square \quad &= -QQ_1G_2^{-1}(WB)'PP_1z - QP_1G_2^{-1}BPP_1z \end{aligned}$$

Note that the linear space $\mathcal{M}_1(t)|_{q=0}$ represents the tangent space for $\mathcal{M}_1(t)$. Of course, this is much more important for nonlinear problems. Lemma 2.3 shows $P(t)S_1(t)$ to be a kind of practical substitute of the tangent space. It should be stressed once more that we are looking for C_N^1 solutions and that there is no natural need for Q-components of the elements of the tangent spaces.

Next, rewrite the DAE (2.1) as

$$Ax' + (I - W)(Bx - q) + W(Bx - q) = 0$$

and replace the part $W(Bx - q) = 0$ by its differentiated form

$$W\{WBx' + (WB)'x - (Wq)'\} = 0$$

to compose the new DAE

$$(A + WB)x' + ((I - W)B + W(WB)')x = (I - W)q + W(Wq)'. \tag{2.14}$$

Denote $\tilde{A} := A + WB, \tilde{B} := (I - W)B + W(WB)'$ and consider (2.14) in more detail. Since $(A(t) + W(t)B(t))z = 0$ decomposes into $Az = 0$, $WBz = 0$, we have $\ker \tilde{A}(t) \equiv \ker A(t) = N(t)$. In consequence, the solutions of (2.15) belong to the same class C_N^1 as the solutions of (2.1). Further, we have

$$\begin{aligned} \tilde{S}(t) &= \{z \in I\!R^m : \ \tilde{B}(t)z \in \ \operatorname{im} \tilde{A}(t)\} \\ &= \{z \in I\!R^m : \ (I - W(t))B(t)z \in \ \operatorname{im} A(t), \\ &\qquad W(t)(WB)'(t)z = W(t)B(t)A(t)^+(I - W(t))B(t)z\}, \end{aligned}$$

$$\begin{aligned} \widetilde{\mathcal{M}}(t) = \{&z \in I\!R^m : (I - W(t)(B(t)z - q(t)) \in \ \operatorname{im} A(t), \\ &W(t)(WB)'(t)z - W(t)(Wq)'(t) = W(t)B(t)A(t)^+(I - W(t))(B(t)z - q(t))\}, \end{aligned}$$

$$\widetilde{\mathcal{M}}(t) \cap \{z \in I\!R^m : W(t)(B(t)z - q(t)) = 0\} = \mathcal{M}_1(t).$$

Theorem 2.4 *Given the conditions of Theorem 2.2.*

(i) Then equation (2.14) is index-1 tractable.

(ii) Exactly one solution at $t_* \in J$ passes through each $x_* \in \widetilde{\mathcal{M}}(t_*)$.

(iii) $\mathcal{M}_1(t)$ and $\mathcal{M}(t)$ form invariant subsets for (2.14).

Proof:

(i) We check the nonsingularity of the matrix $\tilde{G}_1(t) \in L(\mathbb{R}^m), \tilde{G}_1 := A + WB + ((I-W)B + W(WB)')Q$. Again, we drop the t-argument. $\tilde{G}_1 z = 0$ yields $Az + (I-W)BQz = 0, \quad WBz = -W(WB)'Qz = -WBP'Qz$, hence $PQ_1 z = -PQ_1P'Qz, \ (A + BQ)z = 0$.
Because of $(A+BQ)(I-PP'Q) = A_1, \ A_1(I+PP'Q) = A+BQ$, the element $(I+PP'Q)z = \tilde{z}$ belongs to $\ker A_1 = N_1$, i. e., $\tilde{z} = Q_1\tilde{z}$. On the other hand, $PQ_1 z + PQ_1P'Qz = 0$ implies $PQ_1(I+PP'Q)z = 0$, i. e., $PQ_1\tilde{z} = 0$ and finally $\tilde{z} = 0, \ z = 0$.

(ii) Due to the index-1 property, this assertion is now given by the index-1 results, e. g. in [2].

(iii) Let x denote a solution of the index-1 tractable DAE (2.14) and $x(t_*) \in \mathcal{M}(t_*)$. Then, (2.14) gives

$$Ax' + (I-W)Bx = (I-W)q, \quad WBx' + W(WB)'x = W(Wq)'.$$

Consider the function $\alpha := W(Bx - q)$. Derive

$$\begin{aligned} \alpha' &= (WB)'x + WBx' - (Wq)' \\ &= (WB)'x - W(WB)'x + W(Wq)' - (Wq)' \\ &= (I-W)(WB)'x - (I-W)(Wq)' \\ &= -(I-W)'(WBx - Wq) = W'\alpha. \end{aligned}$$

Since $x(t_*) \in \mathcal{M}(t_*)$ implies $\alpha(t_*) = W(t_*)(B(t_*)x(t_*) - q(t_*)) = 0$, the function α vanishes identically, hence $Ax' + Bx = q$ is satisfied.
Therefore, $x_*(t) \in \mathcal{M}(t_*)$ implies $x(t) \in \mathcal{M}(t)$, but also $x(t) \in \mathcal{M}_1(t)$ for all $t \in J$.
□

Corollary: *The index-2 tractable DAE has differentiation index two.*

Example: *For the Hessenberg system*

$$\left.\begin{array}{rcrcrcl} x_1' & + & B_{11}x_1 & + & B_{12}x_2 & = & q_1 \\ & & B_{21}x_1 & & & = & q_2 \end{array}\right\}$$

we have now

$$\begin{aligned}
\mathcal{M}(t) &= \{z : B_{21}(t)z_1 - q_2(t) = 0\} \\
\mathcal{H}(t) &= \{z : -B_{21}(t)B_{11}(t)z_1 - B_{21}(t)B_{12}(t)z_2 + \\
&\quad +B_{21}'(t)z_1 = q_2'(t) - B_{21}(t)q_1(t)\}\,, \\
\mathcal{M}_1(t) &= \{z : B_{21}(t)z_1 = q_2(t),\ (B_{21}(t)B_{12}(t))z_2 = \\
&\quad B_{21}'(t)z_1 - B_{21}(t)B_{11}(t)z_1 - q_2'(t) + B_{21}(t)q_1(t)\} \\
\tilde{\mathcal{M}}(t) &= \{z : B_{21}'(t)z_1 - q_2'(t) = B_{21}(t)(B_{11}(t)z_1 + B_{12}(t)z_2 - q_1(t))\}\,, \\
\tilde{\mathcal{M}}(t) &\cap \{z : B_{21}(t)z_1 = q_2(t)\} = \mathcal{M}_1(t).
\end{aligned}$$

The index-reduced equation (2.15) is, as expected,

$$\left.\begin{array}{rcrcrcl} x_1' & + & B_{11}x_1 & + & B_{12}x_2 & = & q_1 \\ & & B_{21}x_1' & + & B_{21}'x_1 & = & q_2' \end{array}\right\}$$

2.4 Contractivity

Now we turn to the asymptotic behaviour of the solutions of the homogeneous equations

$$Ax' + Bx = 0. \tag{2.15}$$

As we know, the solutions may be represented by

$$x = KPP_1u = \Pi_{can}u,$$

where u solves the inherent regular ODE

$$u' = \Pi'u + PP_1V'u - PP_1G_2^{-1}BPP_1u, \tag{2.16}$$

$$u(t_0) \in \operatorname{im}\Pi(t_0).$$

Recall once more that $\operatorname{im}\Pi(t)$ is an invariant subspace of the ODE (2.16), i. e.,

$$u(t) = \Pi(t)u(t),\ t \in J.$$

Obviously, the stability behaviour of x is mainly governed by the dynamics of the inherent regular ODE, that is, by u. Hence, it would be nice to formulate criteria on how the flow of regular ODE behaves within an given invariant subspace.

The matrix coefficient of the inherent regular ODE (2.16) has to be expected to become singular. In particular, for constant subspaces N and S_1, it is simply the matrix $PP_1G_2^{-1}BPP_1$, which has at least the nullspace $N \oplus N_1$ of dimension $m-r+\nu \geq 2$. The standard approaches to estimate how the solutions grow usually relate to all solutions of (2.16), included those solutions that do not start in im $\Pi(t_0)$. No exponential decay can be realized in this way. The standard techniques are somewhat coarse. Hence, we try to improve them by relating all things to the invariant subspaces.

For a given subspace $U \subset \mathbb{R}^m$ we introduce the matrix-semi-norm

$$\|G\|^U := \max\left\{\frac{|Gz|}{|z|} : z \neq 0, z \in U\right\}$$

such that $|Gz| \leq \|G\|^U |z|$ holds for all $z \in U$.

Then, we define the logarithmic matrix norm relative to the subspace U by

$$\mu^U(G) := \lim_{h\to 0} \frac{1}{h}(\|I + hG\|^U - 1). \tag{2.17}$$

This logarithmic norm relative to U is well-defined and has similar properties as the standard version of Dahlquist for $U = \mathbb{R}^m$ (cf.([9]).In fact this generalization is in fact very natural and straightforward. On the other hand, it is worth mentioning that $\mu^U(G)$ appears to be a special case of the logarithmic norm proposed in [10] for matrix pencils.

Lemma 2.5 *Given a regular linear ODE*

$$u'(t) + M(t)u(t) = 0,\ t \in J = [t_0, \infty),$$

which has the invariant subspace $U(t) \subset \mathbb{R}^m,\ t \in J$. *Let the function*

$$\gamma(t) := \int_{t_0}^{t} \mu^{U(s)}(-M(s))ds,\ t \in J,$$

be well-defined, i. e., the integrals do exist.
Then, for all ODE solutions starting with initial values $u(t_0) \in U(t_0)$*, the inequality*

$$|u(t)| \leq e^{\gamma(t)}|u(t_0)|,\ t \geq t_0,$$

is valid.

Proof:
This proof follows the lines of the standard theory [9], too. Given a solution with $u(t_0) \in U(t_0)$ we have, $u(t) \in U(t)$, for all $t \in [t_0, \infty)$.
Denote $m(t) := |u(t)|,\ t \in [t_0, \infty)$, and derive

$$\begin{aligned} m(t+h) &= |u(t) + hu'(t) + o(h)| = |(I - hM(t))u(t) + o(h)| \\ &\leq \|(I - hM(t)\|^{U(t)}|u(t)| + o(h), \end{aligned}$$

thus

$$\frac{1}{h}(m(t+h) - m(t)) \leq \frac{1}{h}(\|I - hM(t)\|^{U(t)} - 1)m(t) - o(h^0).$$

Letting $h \to 0$, we obtain

$$D_+m(t) \leq \mu^{U(t)}(-M(t))m(t),\ t \in [t_0, \infty),$$

where $D_+m(t)$ denotes the respective Dini derivative. By Peano's Lemma it follows immediately that

$$m(t) \leq e^{\gamma(t)}m(t_0).$$

□

Theorem 2.6 *Let (2.15) be index-2 tractable with N, S_1 and WB from C_1. Then, the estimation*

$$|x(t)| \leq \|\Pi_{can}(t)\|e^{\gamma(t)}|\Pi(t_0)x(t_0)|,\ t \geq t_0, \tag{2.18}$$

with $\gamma(t) := \int_{t_0}^{t} \mu^{(PS_1)(\tau)}(-M(\tau))d\tau,\ t \geq t_0$, holds true for each solution, provided that $\gamma(t)$ is well-defined. Here,

$$M := -\Pi' - PP_1V' + PP_1G_2^{-1}BPP_1$$

is the coefficient matrix of the inherent regular ODE (2.16).

Proof:
Applying Lemma 2.5 to (2.16), which has the invariant subspace $\operatorname{im} \Pi(t) = P(t)S_1(t)$, yields $|u(t)| \leq e^{\gamma(t)}|u(t_0)|,\ t \geq t_0$, provided that $u(t_0) \in \operatorname{im} \Pi(t_0)$.
The DAE solutions have the representation $x(t) = \Pi_{can}(t)u(t)$, and $u(t_0) = \Pi(t_0)x(t_0)$.
□

Corollary: *If there is a* $\beta > 0$ *such that* $\mu^{(PS_1)(t)}(-M(t)) \leq -\beta, t \geq t_0$*, all DAE solutions satisfy the inequality*

$$|x(t)| \leq \|\Pi_{can}(t)\| e^{-\beta(t-t_0)} |\Pi(t_0)x(t_0|, \; t \geq t_0. \tag{2.19}$$

Hence, if $\Pi_{can}(t)$ *grows "moderately" like polynomials or is bounded by* $\|\Pi_{can}(t)\| \leq Ce^{\alpha t}, \; t \geq t_0$*, with* $\alpha < \beta$*, then the solutions decrease exponentially.*

To realize whether numerical approximations reflect the stability behaviour of the true solution well, one often uses the notion of contractivity and so-called one-sided Lipschitz conditions (cf. ([9]). This approach applies also to DAEs provided that we relate the things again to our invariant subspaces.
Standard numerical integration methods applied to index-2 DAE are expected to work well if the basic nullspace N is constant. However, it is well-known that these methods may fail in case this nullspace rotates with time. This is why we suppose $P' = 0$ now.
If the vector norm used to define the logarithmic norm is related to an inner product, it holds that

$$\langle Gz, z \rangle \leq \mu^U(G)|z|^2 \quad \text{for all } z \in U.$$

Definition: *An index-2 tractable DAE with constant* P *is said to be contractive if there are an inner product* $\langle , \rangle$ *and a constant* $\beta > 0$ *such that the inequality*

$$\langle y, Px \rangle \leq -\beta |Px|^2 \tag{2.20}$$

is valid for all $y, x \in I\!R^m, \; t \in [t_0, \infty)$*, with*

$$A(t)y + B(t)x = 0, \quad Qy = 0, \quad V(t)y = V(t)\Pi'(t)x. \tag{2.21}$$

This definition is given in [6] for the case of a continuously differentiable Q_1 and a real scalar product. If $Q_1 \in C^1$, we may make use of

$$Q_1(t)y = Q_1(t)V(t)y = Q_1(t)\Pi'(t)x = -Q_1'(t)\Pi(t)x = -Q_1'(t)PP_1(t)x$$

and arrive at the same expression as used in [6] instead of (2.21).
By decoupling $A(t)y + B(t)x = 0$, we find that (2.21) leads to

$$y + M(t)\Pi(t)x = 0, \; V(t)x = 0, \; Px = \Pi(t)x.$$

Therefore, (2.20) reads then

$$\langle -M(t)\Pi(t)x, \; \Pi(t)x \rangle \leq -\beta |\Pi(t)x|^2,$$

which means a contractivity condition for the inherent regular ODE relative to the invariant subspace $\operatorname{im} \Pi(t) = PS_1(t)$.

3 Analyzing numerical integration methods for linear index-2 DAEs

In this section we derive conditions for preserving stability properties of numerical methods when solving index-2 linear DAEs. Let the conditions of Theorem 2.2 be fulfilled in the following. As in [6] we assume $N(t)$ to be constant, otherwise it is well known that those methods will fail. We shall consider the discrete version of the decoupling of the last section for two of the most important families of methods. An analogous analysis was made in [6] using the standard approach mentioned in Section 2.1. In that paper they obtained as a sufficient condition for preserving the stability properties that both subspaces N_1 and S_1 have to be time invariant. By the discrete decoupling with the new approach described in Section 2.2, we shall get a weaker condition as in [6]. Namely, the time invariance of S_1 will do.

3.1 BDF

Consider the homogeneous system

$$A(t)x'(t) + B(t)x(t) = 0\ , \qquad t \in J := [t_0, \infty). \tag{3.1}$$

For homogeneous index-2 tractable DAEs (3.1), the solutions are given by the expression

$$x = (I - QP_1G_2^{-1}BPP_1 - QQ_1V'PP_1)PP_1u = \Pi_{can}u, \tag{3.2}$$

where u solves the regular ODE

$$u' - \Pi'u - PP_1V'u + PP_1G_2^{-1}BPP_1u = 0, \tag{3.3}$$

with $u(t_0) \in Im\Pi(t_0)$. Suppose $P' = 0$.
If the projector V is time invariant, or equivalently, $\Pi = P(I - V)$ is so, this solution representation simplifies to

$$x = (I - QP_1G_2^{-1}BPP_1)PP_1u, \tag{3.4}$$

$$u' + PP_1G_2^{-1}BPP_1u = 0. \tag{3.5}$$

Given the step size $h > 0$, $t_i = t_0 + ih$, $i \in I\!N$. The BDF applied to (3.1) reads

$$A_i \sum_{j=0}^{k} \alpha_j x_{i-j} + hB_i x_i = 0, \qquad i \geq k, \tag{3.6}$$

where the starting values $x_0, ..., x_{k-1}$ are supposed to be known.
First, we multiply by $((WB)^+W)_i$ and obtain

$$V_i x_i = 0, \qquad i \geq k, \tag{3.7}$$

which means that this component is correctly computed. Let us now consider what happens with the other components. Assuming that, the previous values $x_{i-1}, \ldots, x_{i-k}$ as well as x_i, also fulfil (3.7), we have $x_i = \Pi_i x_i + Q x_i$, $\forall i$, and, keeping in mind that $P' = 0$, the equation (3.6) can be rewritten as

$$A_i \sum_{j=0}^{k} \alpha_j (\Pi x)_{i-j} + h B_i (\Pi x)_i + h B_i (Q x)_i = 0, \qquad i \geq k.$$

Multiplying by $G_{2,i}^{-1}$ we obtain

$$P_{1,i} P \sum_{j=0}^{k} \alpha_j (\Pi x)_{i-j} + h (G_2^{-1} B P P_1)_i (\Pi x)_i + h Q x_i = 0. \tag{3.8}$$

Denote briefly $H := P P_1 G_2^{-1} B P P_1$ and $u_j = \Pi_j x_j$.
Multiplying (3.8) by $P P_{1,i}$ implies

$$P P_{1,i} \sum_{j=0}^{k} \alpha_j u_{i-j} + h H_i u_i = 0,$$

which is equivalent to

$$\sum_{j=0}^{k} \alpha_j u_{i-j} + h H_i u_i + P P_{1,i} \sum_{j=0}^{k} \alpha_j u_{i-j} - \sum_{j=0}^{k} \alpha_j u_{i-j} = 0,$$

$$\sum_{j=0}^{k} \alpha_j u_{i-j} + h H_i u_i + \sum_{j=1}^{k} \alpha_j (\Pi_i - \Pi_{i-j}) u_{i-j} + \sum_{j=1}^{k} \alpha_j [P P_{1,i} - \Pi_i] u_{i-j} = 0,$$

$$\sum_{j=0}^{k} \alpha_j u_{i-j} + h H_i u_i + \sum_{j=1}^{k} \alpha_j (\Pi_i - \Pi_{i-j}) u_{i-j} + P P_{1,i} \sum_{j=1}^{k} \alpha_j (V_i - V_{i-j}) u_{i-j} = 0 \tag{3.9}$$

This is the discrete analogue of (3.3).
For the Q-component we multiply (3.8) by Q, thus obtaining

$$h Q x_i - Q Q_{1,i} \sum_{j=0}^{k} \alpha_j u_{i-j} + h (Q P_1 G_2^{-1} B P P_1)_i u_i = 0,$$

$$Q x_i = Q Q_{1,i} \frac{1}{h} \sum_{j=0}^{k} \alpha_j u_{i-j} - (Q P_1 G_2^{-1} B P P_1)_i u_i. \tag{3.10}$$

Now, regarding (3.9) and (3.10) we observe that, if the projection Π is constant, then the BDF discretization of (3.1) coincides with the corresponding method applied to (3.5) and formula (3.4). Namely, we have $x_i = \Pi_{can}(t_i) \Pi_i x_i = \Pi_{can}(t_i) u_i$.

Theorem 3.1 *Let (3.1) be index-2 tractable with continuously differentiable W and WB and $P' = 0$. Suppose the starting values to satisfy $x_i \in \mathcal{M}(t_i), i = 0, \ldots, k-1$. Then the BDF (3.6) applied to (3.1) generates exactly the same BDF method applied to the inherent ODE (3.3) iff the projection V (the subspace S_1) is constant.*

Remark:

1) The numerical approximation $x_i = \Pi_{can}(t_i)u_i$ reflects the asymptotic behaviour of the true solution $x(t_i) = \Pi_{can}(t_i)u(t_i)$ nicely if the BDF works well for the inherent regular ODE (3.3).

2) Let $S_1(t)$ be time-invariant. Then, applying the BDF to (3.1) and decoupling this is exactly the same as decoupling (3.1) first and then applying the BDF to the inherent regular ODE. In this sense, the BDF-discretization and the decoupling commute.

3) In circuit simulation, the DAEs obtained by the classical modified nodal analysis fulfil the condition $V' = 0$ (cf. [7]).

In the previous analysis we have checked whether the BDF scheme is transmitted to the inherent regular ODE (3.3). Analogously, we could also check if the BDF method is transmitted to the reduced index equation (2.14). This would be very helpful, because in the constant null space case we know that the BDF preserves it asymptotical stability properties (cf. [4]).
Looking at (3.6) we immediately see that $x_i \in \mathcal{M}(t_i)$. Splitting (3.6) with the projections W_i and $I - W_i$ yields, $\forall i \geq k$,

$$A_i \sum_{j=0}^{k} \alpha_j x_{i-j} + h(I - W_i)(Bx)_i = 0, \tag{3.11}$$

$$(WBx)_i = 0. \tag{3.12}$$

However, we cannot hope for a reproduction of the hidden constraint, in fact, this will be a key condition for the BDF methods to preserve their asymptotical stability properties. To see this we rewrite (3.11) as

$$(A + WB)_i \sum_{j=0}^{k} \alpha_j x_{i-j} + h(I - W_i)(Bx)_i - (WB)_i \sum_{j=0}^{k} \alpha_j x_{i-j} = 0.$$

Here we see that, if we had the relation

$$hW_i(WB)'_i x_i = -(WB)_i \sum_{j=0}^{k} \alpha_j x_{i-j},$$

then we would really obtain the BDF discretization of the index-1 problem (2.14). The above condition can also be written as

$$W_i\left\{(WB)_i\frac{1}{h}\sum_{j=0}^{k}\alpha_j x_{i-j}+(WB)'_i x_i\right\}=0, \tag{3.13}$$

which is nothing but the BDF discretization of the differentiated constraint. Furthermore, (3.6) provides

$$P\frac{1}{h}\sum_{j=0}^{k}\alpha_j x_{i-j}=-P(A^+Bx)_i,$$

so (3.13) can also be written as

$$W_i\left\{-(WB)_i(A^+Bx)_i+(WB)'_i x_i\right\}=0, \qquad \text{i.e., } x_i\in\mathcal{H}(t_i).$$

Theorem 3.2 *Let (3.1) be index-2 tractable with continuously differentiable W and WB and $P'=0$. Then a BDF method applied to (3.1) generates exactly the same BDF method applied to the index-1 DAE (2.14) if the condition (3.13) is fulfilled.*

Remarks:

1) The projection $(I-Q\widetilde{A}_1^{-1}\widetilde{B})P$ plays the role of Π_{can} for an index-1 tractable DAE; the solution of (2.14) satisfies $x=(I-Q\widetilde{A}_1^{-1}\widetilde{B})Px$.

2) Condition (3.13) means that the numerical approximation for x must lie on the hidden constraint manifold in every integration step. This is more difficult to check than the condition $V'=0$. The BDF solution of (3.1) at the point t_i is $x_i=(\Pi x)_i+Qx_i$. Inserting this expression in the left part of (3.13) we obtain

$$W_i\left\{(WB)_i\frac{1}{h}\sum_{j=0}^{k}\alpha_j(\Pi x+Qx)_{i-j}+(WB)'_i(\Pi x+Qx)_i\right\}.$$

Then, assuming V to be constant and taking into account that $WB\Pi=WBQ=0$, $(WB)'\Pi=(WB\Pi)'=0$, $(WB)'Q=(WBQ)'=0$, one concludes that (3.13) is fulfilled, indeed.

3.2 Runge-Kutta methods

The implicit Runge-Kutta methods can be realized for (3.1) in the following way [11]. Given an approximation x_{l-1} of the solution at t_{l-1}, a new approximation x_l at $t_l = t_{l-1} + h$ is obtained from

$$x_l = x_{l-1} + h \sum_{i=1}^{s} b_i X'_{li}, \tag{3.14}$$

where X'_{li} are defined by

$$A_{li} X'_{li} + B_{li} X_{li} = 0,\ i = 1, \ldots, s, \tag{3.15}$$

$$t_{li} = t_{l-1} + c_i h$$

and the internal stages are given by

$$X_{li} = x_{l-1} + h \sum_{j=1}^{s} a_{ij} X'_{lj},\ i = 1, \ldots, s. \tag{3.16}$$

The coefficients a_{ij}, b_i, c_i determine the IRK method, and s is the number of stages. We define the matrix $\mathcal{A} := (a_{ij})_{i,j=1}^{s}$ and the vectors $b := (b_1, \ldots, b_s)^T$, $c := (c_1, \ldots, c_s)^T$. A condition for $X'_{l1}, \ldots, X'_{ls}$ to be uniquely defined by (3.15)-(3.16) is the non-singularity of $\mathcal{A}$, [12], which we shall assume in the following. Denoting $\widehat{\mathcal{A}} := \mathcal{A}^{-1} = (\hat{a}_{ij})_{i,j=1}^{s}$ and $\rho := 1 - \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij}$, (3.14)-(3.16) is equivalent to

$$x_l = \rho x_{l-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} X_{lj}, \tag{3.17}$$

$$A_{li} \sum_{j=1}^{s} \hat{a}_{ij} (X_{lj} - x_{l-1}) + h B_{li} X_{li} = 0,\ i = 1, \ldots, s. \tag{3.18}$$

Looking at (3.18) we observe that the internal stages do not depend on Qx_{l-1}. Further, it holds that $X_{lj} \in \mathcal{M}(t_{lj})$.
The special class of IRK methods (IRK(DAE) [4]) with coefficients

$$b_i = a_{si}, \quad i = 1, \ldots, s, \quad c_s = 1, \tag{3.19}$$

is shown to stand out from all IRK methods in view of its applicability to DAEs. Since $\rho = 0$ in this case, the new value $x_l = X_{ls}$ always belongs to the constraint manifold $\mathcal{M}(t_l)$.
For index-2 Hessenberg equations, the fact that $x_l \in \mathcal{M}(t_l)$ simplifies to

$$B_{21} x_{1,l} = q_{2,l}.$$

In general, if (3.19) is not fulfilled, then we have $\rho \neq 0$, and x_l does not belong to $\mathcal{M}(t_l)$ any more. Since this behaviour is a source of instability (for $h \to 0$), Ascher and Petzold [13] propose another version for the application of IRK methods to index-2 Hessenberg systems (cf. the discussion of Hessenberg systems in § 2.2), the so-called Projected IRK methods (PIRK). Actually, after realizing the standard internal stage computation, the recursion (3.17) for the Hessenberg system is now replaced by

$$\hat{x}_{1,l} = \rho \hat{x}_{1,l-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} \widehat{X}_{1,lj} + B_{12}(t_l)\lambda_l \tag{3.20}$$

and λ_l is determined by

$$B_{21}(t_l)\hat{x}_{1,l} = q_{2,l}. \tag{3.21}$$

If we multiply (3.20) by $I - L_l$ (L is defined in § 2.2), λ_l can be eliminated:

$$(I - L_l)\hat{x}_{1,l} = \rho(I - L_l)\hat{x}_{1,l-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} (I - L_l)\widehat{X}_{1,lj}. \tag{3.22}$$

On the other hand, (3.21) is equivalent to

$$L_l \hat{x}_{1,l} = B_{12}(t_l)(B_{21}(t_l)B_{12}(t_l))^{-1} q_{2,l}. \tag{3.23}$$

It should be mentioned that, for IRK(DAE), the projected version is exactly the same as the original one, since (3.19) implies $\lambda_l = 0$ in (3.20), (3.21). An immediate generalization of PIRK methods to fully implicit linear index-2 systems is suggested in [6]:

$$PP_{1,l}\hat{x}_l = \rho PP_{1,l}\hat{x}_{l-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} PP_{1,l}\widehat{X}_{lj}, \tag{3.24}$$

$$Q_{1,l}\hat{x}_l = Q_{1,l}A_{2,l}^{-1} q_l. \tag{3.25}$$

Since the internal stages do not depend on $Q\hat{x}_{l-1}$, there is no need to compute $Q\hat{x}_l$ in this place.
Now return to the standard IRK methods (3.17)-(3.18) for a homogeneous equation. First, we multiply (3.18) by $((WB)^{+}W)_{li}$, and obtain for all internal stages

$$(VX)_{li} = 0, \qquad i = 1, \ldots, s, \tag{3.26}$$

which means that the V-component is correctly computed for all internal stages. This is a consequence of the fact that $X_{li} \in \mathcal{M}(t_{li})$. Next, multiplying (3.17) by V_l results in

$$(Vx)_l = \rho V_l x_{l-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} V_l X_{lj},$$

$$(Vx)_l = \rho(Vx)_{l-1} + \rho(V_l - V_{l-1})x_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij}(V_l - V_{lj})X_{lj}, \tag{3.27}$$

which shows that an IRK scheme does not generate an approximation x_l in $\mathcal{M}(t_l)$ in general. Let us see what happens with the Π-component. Multiplying (3.17) by Π_l we obtain

$$(\Pi x)_l = \rho \Pi_l x_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} \Pi_l X_{lj},$$

$$(\Pi x)_l = \rho(\Pi x)_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij}(\Pi X)_{lj} + \rho(\Pi_l - \Pi_{l-1})x_{l-1}$$
$$+ \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij}(\Pi_l - \Pi_{lj})X_{lj}, \tag{3.28}$$

and for (3.18), assuming that $x_{l-1} \in \mathcal{M}(t_{l-1})$, we can write

$$A_{li} \sum_{j=1}^{s} \hat{a}_{ij}[(\Pi X)_{lj} - (\Pi x)_{l-1}] + h(BX)_{li} = 0.$$

Multiplying by $G_{2,li}^{-1}$ and using the shorter denotations $u_j := \Pi_j x_j$ $U_{lj} := \Pi_{lj} X_{lj}$ we obtain

$$P_{1,li} P \sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] + h(G_2^{-1}BU)_{li} + hQX_{li} = 0. \tag{3.29}$$

Then, multiplying by $PP_{1,li}$ leads to

$$PP_{1,li} \textstyle\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] + hH_{li}U_{li} = 0,$$

$$\textstyle\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] + hH_{li}U_{li} - \sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}]$$

$$+PP_{1,li} \textstyle\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] = 0,$$

$$\textstyle\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] - \sum_{j=1}^{s} \hat{a}_{ij}(\Pi_{lj} - \Pi_{li})U_{lj} + \sum_{j=1}^{s} \hat{a}_{ij}[PP_{1,li} - \Pi_{li}]U_{lj}$$

$$+hH_{li}U_{li} + \textstyle\sum_{j=1}^{s} \hat{a}_{ij} PQ_{1,li} u_{l-1} = 0,$$

$$\textstyle\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] + hH_{li}U_{li} - \sum_{j=1}^{s} \hat{a}_{ij}(\Pi_{lj} - \Pi_{li})U_{lj} + \sum_{j=1}^{s} \hat{a}_{ij} PP_{1,li} V_{li} U_{lj}$$

$$+ \textstyle\sum_{j=1}^{s} \hat{a}_{ij} PQ_{1,li} u_{l-1} = 0,$$

$$\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] + hH_{li}U_{li} - \sum_{j=1}^{s} \hat{a}_{ij}(\Pi_{lj} - \Pi_{li})U_{lj} - PP_{1,li}\sum_{j=1}^{s} \hat{a}_{ij}(V_{lj} - V_{li})U_{lj}$$

$$+ \sum_{j=1}^{s} \hat{a}_{ij}PQ_{1,li}[\Pi_{l-1} - \Pi_{li}]u_{l-1} = 0. \tag{3.30}$$

Finally, for the Q-component we multiply (3.17) by Q and obtain

$$Qx_l = \rho Qx_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\hat{a}_{ij}QX_{lj}, \tag{3.31}$$

and (3.29) multiplied by Q gives

$$-QQ_{1,li}\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] + h(QP_1G_2^{-1}BU)_{li} + hQX_{li} = 0,$$

$$QX_{li} = QQ_{1,li}\frac{1}{h}\sum_{j=1}^{s} \hat{a}_{ij}[U_{lj} - u_{l-1}] - (QP_1G_2^{-1}BU)_{li}. \tag{3.32}$$

After these algebraic manipulations we can extract the desired conclusions. If $S_1(t)$ is time-invariant ($V' = 0$), then (3.27), (3.26), (3.28), (3.30), (3.31) and (3.32) can be reduced to

$$(Vx)_l = \rho(Vx)_{l-1}, \tag{3.33}$$

$$(\Pi x)_l = \rho(\Pi x)_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\hat{a}_{ij}(\Pi X)_{lj}, \tag{3.34}$$

$$(Qx)_l = \rho(Qx)_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\hat{a}_{ij}(QX)_{lj}, \tag{3.35}$$

$$(VX)_{li} = 0, \quad i = 1, \dots, s \tag{3.36}$$

$$\sum_{j=1}^{s} \hat{a}_{ij}[(\Pi X)_{lj} - (\Pi x)_{l-1}] + hH_{li}(\Pi X)_{li} = 0, \quad i = 1, \dots, s \tag{3.37}$$

and

$$(QX)_{li} = -(QP_1G_2^{-1}B\Pi X)_{li}, \quad i = 1, \dots, s \tag{3.38}$$

In (3.33) we have left the term $\rho(PVx)_{l-1}$ to emphasize that an “undesirable” recursion occurs in the PV-component. An even worse situation occurs in the Q-component, here the expression (3.35) does not correspond to (3.4) any more, which may cause instabilities in the computation. What really is true is the following:

Theorem 3.3 *Suppose (3.1) to be index-2 tractable with continuous differentiable W, WB and $P' = 0$. Then, an IRK method applied to (3.1) generates exactly the same IRK method applied to the inherent ODE (3.3) iff the projection V (the subspace S_1) is constant and the initial value x_0 belongs to $\mathcal{M}(t_0)$.*

Here, the IRK(DAE) methods show their good properties, $\rho = 0$ in this case, and

$$(Vx)_l = (VX)_{ls} = 0, \quad (Qx)_l = (QX)_{ls} = -(QP_1 G_2^{-1} B\Pi X)_{ls};$$

thus, there occurs no recursion, neither in the PV-component nor in the Q-component, and these components are computed correctly. Hence, for this type of methods we can state:

Theorem 3.4 *For a constant projection V, an IRK(DAE) method applied to the index-2 tractable DAE (3.1) yields $x_l = \Pi_{can}(t_l)u_l$, $x_l \in \mathcal{M}_1(t_l)$.*

Next, let us briefly come back to the PIRK methods (3.24), (3.25). Our decoupling for the internal stages X_{li} holds true also in this case for the $\widehat{X}_{li}$ values. For a homogeneous systems under the assumptions of Theorem 3.3 we obtain

$$PP_{1,l}\hat{x}_l = \rho PP_{1,l}\hat{x}_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} PP_{1,l}\widehat{X}_{lj},$$
$$Q_{1,l}\hat{x}_l = 0.$$

The second equation is equivalent to

$$V\hat{x}_l = 0, \tag{3.39}$$

inserting $\hat{x}_l = \Pi\hat{x}_l + Q\hat{x}_l$, $\hat{x}_{l-1} = \Pi\hat{x}_{l-1} + Q\hat{x}_{l-1}$ and $\widehat{X}_{lj} = \Pi\widehat{X}_{l,j} + Q\widehat{X}_{l,j}$ in the equation for the PP_1-component and taking into account that Π and Q are constant provides

$$(\Pi\hat{x})_l = \rho(\Pi\hat{x})_{l-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} (\Pi\hat{X})_{lj}, \tag{3.40}$$

which is identical with (3.34). Hence, the application of a PIRK scheme solves the problem with the "undesirable" iteration in the $V(Q_1)$-component, but, as the Q-component is the same as in a "standard" IRK method, we still have the recursion (3.35) for this component. Consequently, a result like Theorem 3.4 is not possible, but the analogue of Theorem 3.3 remains true for PIRK methods.
In the same way as it was done for BDF methods, we shall analyze IRK methods by means of the index reduction technique.
Considering again the IRK scheme (3.17), (3.18); we split (3.18) with the projections $I - W_{li}$ and W_{li} and obtain

$$A_{li}\sum_{j=1}^{s} \hat{a}_{ij}(X_{lj} - x_{l-1}) + h(I - W_{li})B_{li}X_{li} = 0, \ i = 1, \ldots, s \tag{3.41}$$

$$(WBX)_{li} = 0, \ i = 1, \ldots, s. \tag{3.42}$$

However, the hidden restriction need not be fulfilled in general, neither for the stages X_{li} nor for x_l.
Transforming (3.41) appropriately we obtain

$$\begin{aligned}(A+(WB))_{li}\textstyle\sum_{j=1}^{s}\hat{a}_{ij}(X_{lj}-x_{l-1}) \quad & +h(I-W_{li})(BX)_{li} \\ & -(WB)_{li}\textstyle\sum_{j=1}^{s}\hat{a}_{ij}(X_{lj}-x_{l-1})=0.\end{aligned}$$

If we had the relation

$$h(W(WB)'X)_{li} = -(WB)_{li}\sum_{j=1}^{s}\hat{a}_{ij}(X_{lj}-x_{l-1}), \quad i=1,\ldots,s,$$

then we would arrive at the same method applied to the index-1 problem (2.14) with constant null-space N. In other words, similar to the BDF-methods, we obtain

$$W_{li}\left\{(WB)_{li}\frac{1}{h}\sum_{j=1}^{s}\hat{a}_{ij}(X_{lj}-x_{l-1})+(WB)'_{li}X_{li}\right\}=0, \tag{3.43}$$

as condition that all stages must fulfil the discretization of the differentiated constraint.

In [4] it is shown that an IRK (DAE) applied to an index-1 DAE with constant leading nullspace generates the same values as the application of the same IRK to the inherent regular ODE, and $x_l \in \widetilde{\mathcal{M}}(t_l)$.

Theorem 3.5 *Let (3.1) be index-2 tractable with continuously differentiable W, WB and $P'=0$. Then, an IRK(DAE) method applied to (3.1) generates exactly the same IRK for the index-1 DAE (2.14) if condition (3.43) is fulfilled.*

In the same way as we did for the BDF methods, we can show that (3.43) is fulfilled if $V'=0$ and $x_{l-1} \in \mathcal{M}(t_{l-1})$.

In Section 2.4 we have extended the notion of contractivity to linear index-2 tractable DAEs. The reason of this concept in regular ODE theory is to generalize the A-stability notion related to the test equation

$$x' = \lambda x,$$

by that of B-stability for a contractive equation (cf. [9]). An assertion of the type: "algebraically stable Runge-Kutta methods are B-stable", was shown to be true in [4] for index-1 tractable DAEs provided that (i) the null space $N(t)$ is time-invariant, and (ii) the Runge-Kutta method is an IRK(DAE). In [6] such a kind of assertion was proved for linear index-2 tractable DAEs under the additional conditions that Q_1 be differentiable, $PP_1'=0$ and $\|\Pi_{can}(t)\| \leq K$, on $[t_0,\infty)$.

Definition: *The one-step method* $x_{j+1} = \phi(x_j, t_j, h_j)$ *is called B-stable if, for each contractive DAE, the inequalities*

$$|Px_{j+1}^{(1)} - Px_{j+1}^{(2)}|_s \le |Px_j^{(1)} - Px_j^{(2)}|_s,$$

and

$$|Qx_{j+1}^{(1)} - Qx_{j+1}^{(2)}|_s \le K|Px_{j+1}^{(1)} - Px_{j+1}^{(2)}|_s, \qquad j \ge 0,$$

are satisfied. Here, $K > 0$ *is a constant,* $x_0^{(1)}$, $x_0^{(2)}$ *are arbitrary consistent initial values and* $|\cdot|_s$ *denotes a suitable norm.* Following the same technique as in [6] we can improve the result of that work by using the projectors W and V instead of $Q_1G_2^{-1}, Q_1$.

Theorem 3.6 *Let (3.1) be index-2 tractable with continuously differentiable* W, WB; $P' = 0$, $V' = 0$ *and* $\|\Pi_{can}(t)\|$ *bounded on* $[t_0, \infty)$. *Then, each algebraically stable IRK(DAE) applied to (3.1) is B-stable.*

4 Two illustrating examples

In this paper we do not present any "real life" problem as an application, but we know that the condition $V' = 0$, which played an important role in our results, is often fulfilled in linear circuit simulation [7]. However, we would like to illustrate our theory by simpler academic examples.

Example 4.1 *The first example is well known [6],*

$$\begin{pmatrix} x_1' \\ x_2' \\ 0 \end{pmatrix} + \begin{pmatrix} \lambda & -1 & -1 \\ \eta t(1-\eta t) - \eta & \lambda & -\eta t \\ 1 - \eta t & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0, \; t \ge 0.$$

It is a Hessenberg system with $B_{21}B_{12} = 1$ *and, thus, index-2 tractable for all* t. *The general exact solution is given by*

$$\begin{aligned} x_1(t) &= x_1(0)e^{-\lambda t}, \\ x_2(t) &= (\eta t - 1)x_1(t), \\ x_3(t) &= -(\eta t - 1)x_1(t), \end{aligned}$$

which, evidently, is exponentially asymptotically stable for $\lambda > 0$.
If we take

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

then it results that

$$A_1(t) = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -\eta t \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$S_1(t) = \{z \in I\!R^3 : (1-\eta t)z_1 + z_2 = 0\}.$$

After computing the canonical projection Q_1, the projection onto $Ker(A_1)$ along S_1, we have

$$P_1(t) = \begin{pmatrix} \eta t & -1 & 0 \\ \eta t(\eta t - 1) & 1-\eta t & 0 \\ \eta t - 1 & -1 & 1 \end{pmatrix}, \qquad PP_1(t) = \begin{pmatrix} \eta t & -1 & 0 \\ \eta t(\eta t - 1) & 1-\eta t & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In [6], as already mentioned, the authors showed their stability results under the condition $PP_1' = 0$, and by means of this example they illustrated that, if this condition fails, even when using methods with good stability properties, the numerical discretizations may show an asymptotical behaviour that is different from that of the exact solution. Note that in this example we have $PP_1' = 0$ for $\eta = 0$ only. One integration step with the implicit Euler method consists in solving the following linear system

$$\begin{pmatrix} 1+h\lambda & -h & -h \\ h(\eta t_{i+1}(1-\eta t_{i+1}) - \eta) & 1+h\lambda & -h\eta t_{i+1} \\ 1-\eta t_{i+1} & 1 & 0 \end{pmatrix} \begin{pmatrix} x_{1,i+1} \\ x_{2,i+1} \\ x_{3,i+1} \end{pmatrix} = \begin{pmatrix} x_{1,i} \\ x_{2,i} \\ 0 \end{pmatrix},$$

which is invertible for $1 + h(\lambda + \eta) \neq 0$. The third equation yields

$$x_{2,i+1} = (\eta t_{i+1} - 1)\; x_{1,i+1},$$

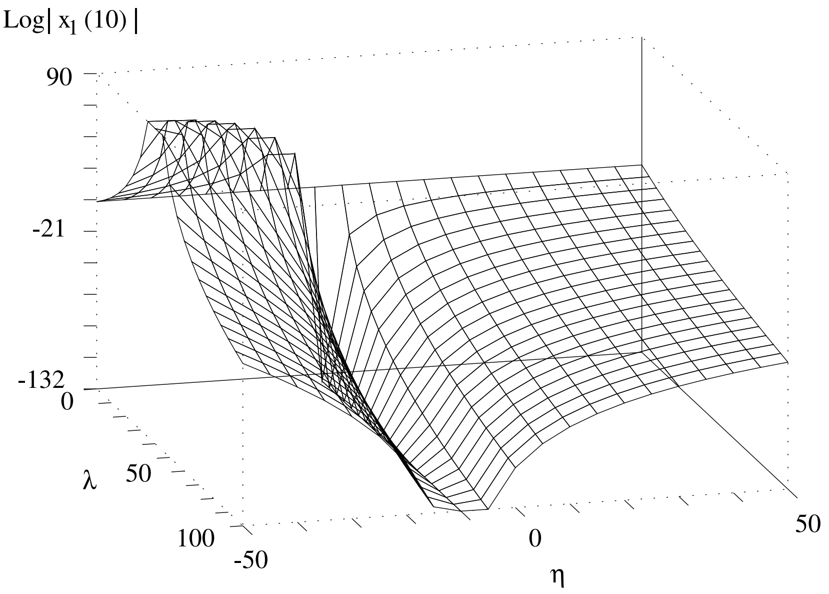
and then, solving the linear system, we obtain the expression

$$x_{1,i+1} = \frac{1+h\eta}{1+h(\lambda+\eta)}\; x_{1,i}$$

for $x_{1,i+1}$.
In order to have a decaying sequence for x_1, the following condition is needed

$$\left| \frac{1+h\eta}{1+h(\lambda+\eta)} \right| < 1.$$

This can be violated for $\eta < 0$, while the exact solution decays for all $\lambda > 0$. We have computed the numerical solution with this method and stepsize $h = 0.1$ on $[0, 10]$. The following figure shows, in a logarithmic scale, the absolute value of the first component at $t = 10$ for different values of λ and η.



As predicted by the above analysis, we see that for negative values of η, if λ is not big enough, the numerical approximation "explodes". The gap about $\eta = -10$ is due to the fact that we have for this value of η and $h = 10^{-1}$ that $x_{1,i+1} = 0$ for all i.
Now let us have a look at the projection V. According to the expression for P_1 we deduce that $S_1(t)$ is generated by the last two columns of this projector matrix, thus $S_1(t)$ is independent of t and, consequently, also $V(t)$ in case of $\eta = 0$ only. Note that for this example

$$W = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad (WB)(t) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1-\eta t & 1 & 0 \end{pmatrix},$$

hence, for V we obtain the expression

$$V(t) = \frac{1}{1+(1-\eta t)^2} \begin{pmatrix} (1-\eta t)^2 & 1-\eta t & 0 \\ 1-\eta t & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

We compute the matrix $M := PP_1G_2^{-1}B - \Pi' - PP_1V'$ of the inherent ODE (3.3), thus obtaining

$$M(t) = \begin{pmatrix} \eta(\eta+\lambda)t - \frac{\eta}{\eta t(\eta t-2)+2} & -(\lambda+\eta) + \frac{\eta(1-\eta t)}{\eta t(\eta t-2)+2} & 0 \\ \frac{\eta t(-2\lambda+\eta(-3+(\lambda+\eta)t(4+\eta t(\eta t-3))))}{\eta t(\eta t-2)+2} & \lambda - \eta(\lambda+\eta)t + \frac{\eta(2-\eta t)}{\eta t(\eta t-2)+2} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Moreover, it holds that

$$Im(\Pi) = span\{\mathcal{U}\}, \qquad \mathcal{U} := \begin{pmatrix} 1 \\ \eta t - 1 \\ 0 \end{pmatrix}.$$

In Section 2.4, motivated by the notion of the logarithmic norm of a matrix corresponding to a matrix-semi-norm, we have introduced a contractivity notion for linear index-2 tractable DAEs. We can compute the Euclidean logarithmic norm for this problem in the following way:

$$\mu_2^{Im(\Pi)}[-M] = \sup_{\substack{x \neq 0 \\ x \in Im(\Pi)}} \frac{\langle x, -Mx\rangle}{\langle x, x\rangle} = \sup_{u_1 \neq 0} \frac{\langle \mathcal{U}u_1, -M\mathcal{U}u_1\rangle}{\langle \mathcal{U}u_1, \mathcal{U}u_1\rangle},$$

which simplifies to

$$\sup_{u_1 \neq 0} \frac{-(\eta + 2\lambda - 2\lambda\eta t + \eta^2 t(\lambda t - 1))u_1^2}{(1 + (\eta t - 1)^2)u_1^2} = -\lambda + \frac{\eta(\eta t - 1)}{1 + (\eta t - 1)^2}$$

and this expression is bounded by $-\lambda + \frac{|\eta|}{2}$. *So we have contractivity at least for* $\frac{|\eta|}{2} < \lambda$ *in the Euclidean norm. Moreover we can observe that*

$$\lim_{t \to \infty} \mu_2^{Im(\Pi)}[-M](t) = -\lambda.$$

Example 4.2 *An opposite example is the following*

$$\begin{pmatrix} x_1' \\ x_2' \\ 0 \end{pmatrix} + \begin{pmatrix} \lambda & (\eta t - 1)^2 & -(\eta t - 1) \\ \eta t(\eta t - 1) & \lambda & -\eta t \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0, \ t \geq 0,$$

whose exact solution for $x_1(0) = 1$ *is*

$$\begin{aligned} x_1(t) &= e^{-\lambda t}, \\ x_2(t) &= x_1(t), \\ x_3(t) &= (\eta t - 1)x_1(t), \end{aligned}$$

with exponential asymptotical stability for $\lambda > 0$.
Let us compute the relevant matrices and subspaces. Taking, also in this case, P *as in the first example, we obtain for* A_1 *and the canonical projector* P_1 *that*

$$A_1(t) = \begin{pmatrix} 1 & 0 & 1 - \eta t \\ 0 & 1 & -\eta t \\ 0 & 0 & 0 \end{pmatrix}, \qquad P_1(t) = \begin{pmatrix} \eta t & 1 - \eta t & 0 \\ \eta t & 1 - \eta t & 0 \\ 1 & -1 & 1 \end{pmatrix}.$$

Thus, PP_1 is not constant again, but if we look at S_1 we see that

$$S_1 = \left\{ z \in I\!R^3 : z_1 = z_2 \right\} = span \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\},$$

which is time-independent and so is V, too. As in example 4.1 the projector W is given by

$$W = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$WB = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & 0 \end{pmatrix}.$$

The computation of V gives

$$V = \frac{1}{2} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Hence, the assumptions of Theorem 3.4 are fulfilled for the implicit Euler method. An integration step of this method

$$\begin{pmatrix} 1 + h\lambda & h(\eta t_i - 1)^2 & -h(\eta t_i - 1) \\ h\eta t_i(\eta t_i - 1) & 1 + h\lambda & -h\eta t_i \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} x_{1,i} \\ x_{2,i} \\ x_{3,i} \end{pmatrix} = \begin{pmatrix} x_{1,i-1} \\ x_{2,i-1} \\ 0 \end{pmatrix},$$

provides

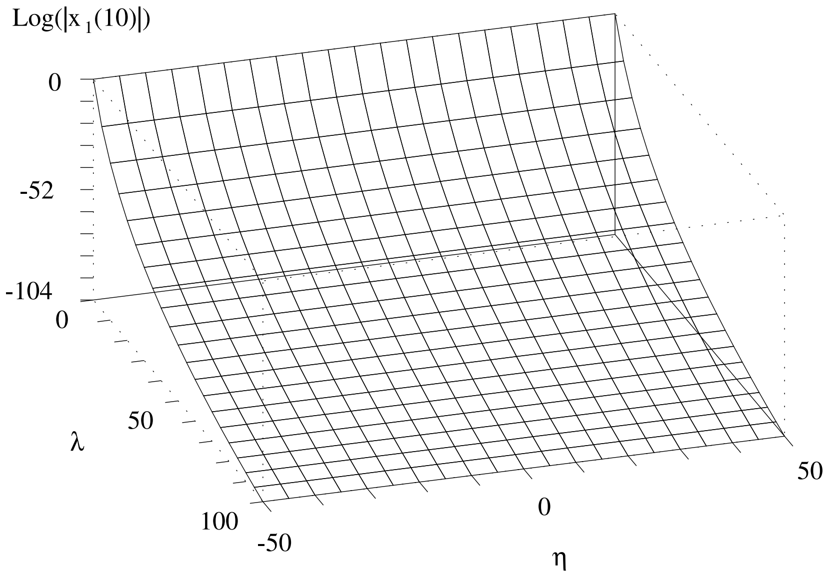
$$x_{1,i} = \frac{1}{1 + h\lambda}\, x_{1,i-1}.$$

For $\lambda > 0$ this fulfils the dissipative condition

$$\left| \frac{1}{1 + h\lambda} \right| < 1.$$

So, in this case the numerical method reflects the asymptotical behaviour of the exact solution. As in the first example the following figure shows the absolute value of x_1 at $t = 10$, for different values of λ and η in a logarithmic scale using the implicit Euler scheme. Again, we have chosen $h = 0.1$.
The coefficient matrix of the inherent ODE (2.9) now reads

$$M(t) = (PP_1 G_2^{-1} B)(t) = \begin{pmatrix} \eta\lambda t & \lambda(1 - \eta t) & 0 \\ \eta\lambda t & \lambda(1 - \eta t) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$



Then, taking into account that

$$\Pi = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \Pi x =: u = \begin{pmatrix} u_1 \\ u_1 \\ 0 \end{pmatrix},$$

equation (2.9) simplifies in essence to

$$u_1' + \lambda u_1 = 0. \tag{3.44}$$

On the other hand, if we compute the Euclidean logarithmic norm of $-M$ *we obtain*

$$\mu_2^{Im(\Pi)}[-M] = \sup_{\substack{x \neq 0 \\ x \in Im(\Pi)}} \frac{\langle x, -Mx\rangle}{\langle x, x\rangle} = \sup_{u_1 \neq 0} \frac{\langle \mathcal{U}u_1, -M\mathcal{U}u_1\rangle}{\langle \mathcal{U}u_1, \mathcal{U}u_1\rangle} = -\lambda,$$

in this example.

References

[1] K. E. Brenan, S. L. Campbell and L. R. Petzold: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. North-Holland (Amsterdam)1989.

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

[3] E. Hairer and G. Wanner: Solving Ordinary Differential Equations. 2nd edition, Springer, Berlin 1993.

[4] E. Griepentrog and R. März: Differential-Algebraic Equations and Their Numerical Treatment. Teubner-Texte zur Mathematik 88, Leipzig 1986.

[5] M. Hanke and R. März: On the asymptotics in the case of differential algebraic equations. Talk given in Oberwolfach, October 1995.

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

[7] D. Estévez Schwarz and C. Tischendorf: Structural analysis for electric circuits and consequences for MNA. Humboldt-Univ. Berlin, Institut für Mathematik, Preprint 98-21, 1998.

[8] R. März: Index-2 Differential-Algebraic Equations. Results in Mathematics 15, 149-171, 1989.

[9] K. Dekker and J. G. Verwer: Stability of Runge-Kutta methods for stiff nonlinear differential equations. CWI Monographs 2, Centre for Mathematics and Computer Science, North-Holland, 1984.

[10] I. Higueras and B. Garcia-Celayeta: Logarithmic norms for matrix pencils. Preprint Nr. 14 (1997), Seccion 1. Departamento de Matemática e informática, Universidad pública de Navarra.

[11] L. R. Petzold: Order results for implicit Runge-Kutta methods applied to differential algebraic systems. SIAM J. Numer. Anal. 23 (1986), pp.837-852.

[12] E. Izquierdo Macana: Numerische Approximation von Algebro-Differentialgleichungen mit Index 2 mittels impliziter Runge-Kutta-Verfahren. Doctoral thesis, Humboldt-Univ., Fachbereich Mathematik, Berlin, 1993.

[13] U. Ascher and L. R. Petzold: Projected implicit Runge-Kutta methods for differential-algebraic equations. SIAM J. Numer. Anal. 28 (1991), pp. 1097-1120.