ON ASYMPTOTICS IN CASE OF LINEAR INDEX-2 DIFFERENTIAL-ALGEBRAIC EQUATIONS

MICHAEL HANKE[†], EBROUL IZQUIERDO MACANA[‡], AND ROSWITHA MÄRZ[§]

Abstract. Asymptotic properties of solutions of general linear differential-algebraic equations (DAE's) and those of their numerical counterparts are discussed. New results on the asymptotic stability in the sense of Lyapunov as well as on contractive index-2 DAE's are given. The behaviour of BDF, IRK, and PIRK applied to such systems is investigated. In particular, we clarify the significance of certain subspaces closely related to the geometry of the DAE. Asymptotic properties like A-stability and L-stability are shown to be preserved if these subspaces are constant. Moreover, algebraically stable IRK(DAE) are B-stable under this condition. The general results are specialized to the case of index-2 Hessenberg systems.

Key words. Differential-algebraic equation, stability, asymptotic properties, Runge-Kutta method, backward differentiation formulas

AMS subject classifications. 65L06, 34D20

1. Introduction. The present paper is devoted to the study of asymptotic properties of solutions of differential-algebraic equations (DAE's) on infinite intervals and those of their numerical counterparts in integration methods. It is rather surprising that, in spite of numerous papers on numerical integration, there are very few results in this respect.

For index-1 DAE's, asymptotic properties on infinite intervals have been investigated by Griepentrog and März [4]. Among other things, the notion of contractivity and that of B-stability were generalized to the case of DAE's and criteria for total stability were formulated. Algebraically stable IRK(DAE) were shown to be B-stable for index-1 DAE's, too, provided that the nullspace N of the leading Jacobian was constant. If this nullspace rotates, stability properties may change.

In this paper, we study general linear index-2 DAE's

$$A(t)x'(t) + B(t)x(t) = q(t), \quad t \in [t_0, \infty), \tag{1.1}$$

exclusively, where the nullspace $N := \ker A(t)$ is assumed to be independent of t. $A(t)$ and $B(t)$ are assumed to be continuous in t. Equation (1.1) is not assumed to be in Hessenberg form and the coefficients $A(t)$ and $B(t)$ need not commute. Recall that Hessenberg index-2 DAE's have the special form

$$\begin{array}{rcl} x_1'(t) + B_{11}(t)x_1(t) + B_{12}(t)x_2(t) & = & q_1(t) \\ B_{21}(t)x_1(t) & = & q_2(t). \end{array} \tag{1.2}$$

This corresponds to the special coefficient matrices in (1.1)

$$A(t) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, B(t) = \begin{bmatrix} B_{11}(t) & B_{12}(t) \\ B_{21}(t) & 0 \end{bmatrix}.$$

[†]Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstr. 39, D–10117 Berlin, Germany (Email `hanke@wias-berlin.de`)

[‡]Heinrich-Hertz-Institut für Nachrichtentechnik, Einsteinufer 37, D–10587 Berlin, Germany (Email `ebroul@hhi.de`)

[§]Humboldt-Universität zu Berlin, Institut für Mathematik, D–10099 Berlin, Germany (Email `iam@mathematik.hu-berlin.de`)

Moreover, it corresponds to a trivially constant nullspace N, since $A(t)$ itself does not vary with t.

Presenting statements on the linear case we hope, as in the case of regular ordinary differential equations (ODE's), that it will be possible to carry over some properties to nonlinear DAE's via linearization.

As far as we know, in case of index-2 DAE's stability analyses of integration methods on infinite intervals have only been presented for linear systems (see März and Tischendorf [13], Wensch, Weiner, and Strehmel [16]). The latter paper is restricted to special Hessenberg form systems and relies on the so-called *essentially underlying ODE* introduced in Ascher and Petzold [2] for these special systems. We consider this case in Section 3 and describe the close relation between the *inherent regular ODE* which we will take up from [8] in Section 2, and the essentially underlying ODE in detail.

Although the above mentioned paper [2] is not concerned with asymptotic stability on infinite intervals, it contains an observation that is highly interesting for us: Among other things, Ascher and Petzold point out that the backward Euler method applied to (1.2) may yield rather an explicit Euler formula for the essentially underlying ODE, and they discuss the influence of the blocks B_{12}, B_{21} on this phenomenon. We will show that not the derivatives of $B_{12}(t), B_{21}(t)$, but the derivatives of the projector matrix $H(t) := B_{12}(t)(B_{21}(t)B_{12}(t))^{-1}B_{21}(t)$ constitute the essential term, i.e., the rotation velocity of the subspace described by $H(t)$ is the decisive feature.

Our paper is aimed at explaining the importance of additional subspaces for answering questions concerning the asymptotic behaviour of integration methods. Hence, besides introducing the necessary fundamentals, Section 2 provides new results on the asymptotic stability of DAE solutions in the sense of Lyapunov as well as on contractive DAE's.

In Section 4, BDF, IRK, and PIRK are investigated in detail. Asymptotic properties like A-stability and L-stability are shown to be preserved if a certain subspace $\operatorname{im} PP_1(t)$ is constant, i.e., if it does not rotate. Moreover, we show that an algebraically stable IRK(DAE) is B-stable under these conditions.

Section 5 illustrates our results by means of examples.

For convenience of the reader, the short appendix provides the basic linear algebra facts once more.

2. Linear continuous coefficient index-2 equations. Consider the linear equation

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

with continuous coefficients. Assume the nullspace of $A(t) \in L(\mathbb{R}^m)$ to be independent of t and let

$$N := \ker A(t) \subset \mathbb{R}^m.$$

Furthermore, set

$$S(t) := \{z \in \mathbb{R}^m : B(t)z \in \operatorname{im} A(t)\} \subset \mathbb{R}^m.$$

Obviously, $S(t)$ is a subspace of $\mathbb{R}^m$ which contains the solutions of the homogeneous form of the DAE (2.1). Note that the condition

$$S(t) \oplus N = \mathbb{R}^m, \quad t \in J, \tag{2.2}$$

characterizes the class of index-1 DAE's (see Appendix for related facts from linear algebra). Equation (2.2) implies that the matrix

$$G_1(t) := A(t) + B(t)Q \tag{2.3}$$

is nonsingular for all $t \in J$, where $Q \in L(\mathbb{R}^m)$ denotes any projector onto N. Let $P := I - Q$.

Higher index DAE's are characterized by nontrivial intersections $S(t) \cap N$ or equivalently by singular matrices $G_1(t)$.

Definition. The DAE (2.1) is said to be *index-2 tractable* if the following two conditions

$$\dim(S(t) \cap N) = \text{const} > 0, \tag{2.4}$$

$$S_1(t) \oplus N_1(t) = \mathbb{R}^m, \quad t \in J \tag{2.5}$$

hold, where

$$N_1(t) := \ker G_1(t),$$
$$S_1(t) := \{z \in \mathbb{R}^m : B(t)Pz \in \operatorname{im} G_1(t)\}.$$

In the following, let $Q_1(t)$ denote the projector onto $N_1(t)$ along $S_1(t)$, and $P_1(t) := I - Q_1(t)$. Due to the decomposition (2.5), $Q_1(t)$ is uniquely defined.

Remarks.

1. It holds that $\dim N_1(t) = \dim(N \cap S(t))$.
2. Due to Lemma A.1, (2.4) and (2.5) imply that the matrix

$$G_2(t) := G_1(t) + B(t)PQ_1(t) \tag{2.6}$$

is nonsingular. But $G_1(t)$ is singular, independently of how Q is chosen [8].

3. Applying Lemma A.1 once more we obtain the identities

$$Q_1 = Q_1 G_2^{-1} BP, \quad Q_1 Q = 0. \tag{2.7}$$

4. Each DAE (2.1) having global index 2 is index-2 tractable with a continuously differentiable Q_1 [8]. Hence, assuming Q_1 to belong to the class C^1 in the sequel is not restrictive.

The conditions (2.4), (2.5) imply the decompositions

$$\begin{aligned} \mathbb{R}^m &= N \oplus PS_1(t) \oplus PN_1(t) = N \oplus \operatorname{im} PP_1(t) \oplus \operatorname{im} PQ_1(t), \\ S_1(t) &= N \oplus \operatorname{im} PP_1(t), \end{aligned} \tag{2.8}$$

which are relevant for the index-2 case. Taking this into account we decompose the DAE solution x into

$$x = Qx + PP_1x + PQ_1x =: w + u + Pv. \tag{2.9}$$

Multiplying (2.1) by $PP_1G_2^{-1}$, $QP_1G_2^{-1}$ and $Q_1G_2^{-1}$, respectively, and carrying out a few technical computations, we decouple the index-2 DAE into the system

$$u' - (PP_1)'u + PP_1G_2^{-1}Bu = PP_1G_2^{-1}q + (PP_1)'v, \tag{2.10}$$

$$-(Qv)' + (QQ_1)'(u + Pv) + w + QP_1G_2^{-1}Bu = QP_1G_2^{-1}q, \tag{2.11}$$

$$v = Q_1G_2^{-1}q. \tag{2.12}$$

Equation (2.10) represents the *inherent regular ODE* of the DAE system.

On the other hand, if we consider (2.10) separately from its origin via the decomposition (2.9), we know that $\operatorname{im} PP_1(t)$ is an invariant subspace of this explicit ODE in u. To be more precise: If we have

$$u(t_0) \in \operatorname{im} PP_1(t_0) \tag{2.13}$$

at some $t_0 \in J$, then (2.10) implies $u = PP_1u$. Furthermore, (2.12) and (2.11) lead to $v = Q_1v$ and $w = Qw$, respectively. Thus, solving (2.10) – (2.13) and setting $x := u + Pv + w$, we obtain the solutions of the DAE (2.1).

Inspired by the above decoupling procedure, we state initial conditions for (2.1) as

$$PP_1(t_0)(x(t_0) - x^0) = 0, \quad x^0 \in \mathbb{R}^m \text{ given.} \tag{2.14}$$

This yields

$$u(t_0) = PP_1(t_0)x(t_0) = PP_1(t_0)x^0,$$

but we do not expect $x(t_0) = x^0$ to hold.

Next, we shortly turn to the case of a homogeneous equation (2.1): For $q = 0$ the system (2.10) – (2.12) yields $v = 0$ and

$$x = u + w = (I - (QQ_1)' - QP_1G_2^{-1}B)u = (I - (QQ_1)' - QP_1G_2^{-1}B)PP_1u.$$

The matrix $\Pi(t) := (I - (QQ_1)'(t) - (QP_1G_2^{-1}B)(t))PP_1(t)$ is also a projector, and $\ker \Pi(t) = \ker PP_1(t)$. $\Pi(t)$ is said to be the *canonical projector* for the index-2 case.

Now, the following assertion is easily proved by means of the decoupling explained above.

THEOREM 2.1. *Let* (2.1) *be index-*2 *tractable with continuously differentiable* Q_1. *Then it holds:*

(i) *The initial value problems* (2.1), (2.14) *are uniquely solvable in*

$$C_N^1(J, \mathbb{R}^m) := \{x \in C(J, \mathbb{R}^m) : Px \in C^1(J, \mathbb{R}^m)\},$$

provided that $q \in C(J, \mathbb{R}^m)$, $Q_1G_2^{-1}q \in C^1(J, \mathbb{R}^m)$.

(ii) *If* $x(.)$ *solves the homogeneous equation, then it holds that*

$$x(t) \in M(t) := \operatorname{im} \Pi(t) \subset S(t), \quad t \in J.$$

(iii) *Through each* $x_* \in M(t_*)$ *there passes exactly one solution of the homogeneous equation at time* $t_* \in J$. *The solution space* $M(t)$ *is a proper subspace of* $S(t)$ *and*

$$\dim M(t) = m - \dim N - \dim N \cap S(t).$$

REMARKS.

1. The inherent regular ODE (2.10) is determined by the complete coefficient matrix $PP_1G_2^{-1}B - (PP_1)'$, but not only by its first term $PP_1G_2^{-1}B$. If $PP_1(t)$ varies rapidly with t, the second term $(PP_1)'$ may be the dominant one. This should also be taken into account when considering the asymptotic behaviour of solutions of (2.1).

2. In general, the linear DAE (2.1) appears to be much simpler if the relevant subspaces N, N_1, S_1 and the two projectors Q, Q_1 are constant. In that case (2.10) – (2.12) simplifies to

$$u' + PP_1G_2^{-1}Bu = PP_1G_2^{-1}q, \tag{2.15}$$

$$-(Qv)' + w + QP_1G_2^{-1}Bu = QP_1G_2^{-1}q, \tag{2.16}$$

$$v = Q_1G_2^{-1}q. \tag{2.17}$$

3. The value $x^0 \in \mathbb{R}^m$ involved in the initial condition (2.14) is not expected to be a consistent initial value. What we have is $PP_1(t_0)x(t_0) = PP_1(t_0)x^0$, but not $x_0 := x(t_0) = x^0$. As shown above, a consistent initial value for the homogeneous equation always belongs to $M(t_0)$, which is precisely the set of consistent initial values then.

4. If the product PP_1 is time invariant, we have $(QQ_1)'PP_1 = (QQ_1PP_1)' = 0$, hence

$$\Pi = (I - QP_1G_2^{-1}B)PP_1.$$

Note that $(QP_1G_2^{-1}B)(t)$ is also a projector onto $\ker A(t)$. It should be mentioned that the solution space $M(t)$ remains time-invariant provided that both projectors PP_1 and $QP_1G_2^{-1}B$ are constant.

Now we turn to the asymptotic behaviour of the solutions of the homogeneous equation. Considering the decoupled system (2.10) – (2.12) once more, we see that the component $u = PP_1x$ represents the dynamic one. Supposed the canonical projector $\Pi(t)$ remains bounded on the whole interval $J = [t_0, \infty)$, the asymptotic behaviour of the solution

$$x(t) = \Pi(t)u(t)$$

is completely determined by that of its component $u(t)$. Clearly, if u solves a constant coefficient regular ODE, we may characterize asymptotics by means of the corresponding eigenvalues. This is what we try to realize for the DAE in the following theorem.

Theorem 2.2. *Let* (2.1) *be index-*2 *tractable,* Q_1 *be of class* C^1, $(PP_1)' = 0$. *Let* $PP_1G_2^{-1}B$ *be constant,* $r := \operatorname{rank} PP_1$.

(i) *Then the pencil* $\lambda A(t) + B(t)$ *has the eigenvalues* $\lambda_1, \dots, \lambda_r$, *uniformly for* $t \in J$.

(ii) $\operatorname{Re} \lambda_i < 0$, $\quad i = 1, \dots, r$, *implies each homogeneous equation solution to tend to zero as* $t \to \infty$, *provided that the projector* $\Pi(t)$ *remains uniformly bounded.*

Proof. Due to our assumptions, the inherent regular ODE has the constant coefficient $PP_1G_2^{-1}B$. On the other hand, the nontrivial eigenvalues of $-PP_1G_2^{-1}B$ (that is, eigenvalues that do not correspond to $\ker PP_1$) are exactly the pencil eigenvalues of $\lambda A + B$ (cf. [10]).

Let $U(\cdot)$ denote the fundamental solution matrix of $u' + PP_1G_2^{-1}Bu = 0$ with $U(t_0) = I$. Taking the solution representation

$$x(t) = \Pi(t)U(t)PP_1(t_0)x^0$$

into account, the assertion follows right away. □

Roughly speaking, the assumptions that PP_1 and $PP_1G_2^{-1}B$ have to be constant mean that there is a constant coefficient inherent regular ODE and a possible time dependence of the system may be caused by (time dependent) couplings only.

Next, what about contractivity in case of index-2 DAE's? In the regular ODE theory, contractivity is well-known to permit very attractive asymptotic properties of numerical integration methods. Corresponding results are obtained for index-1 DAE's in [4] by means of an appropriate contractivity notion. In particular, this notion says that a linear index-1 DAE (2.1) is contractive if there are a constant $c > 0$ and a positive-definite matrix S such that the inequality

$$\langle y, Px\rangle_S \le -c|Px|_S \tag{2.18}$$

holds true for all $y, x \in \mathbb{R}^m$ with

$$A(t)y + B(t)x = 0, \quad Qy = 0.$$

Here, we have used the scalar product $\langle z, v\rangle_S := \langle Sz, v\rangle$ and the norm $|z|_S := \langle z, z\rangle_S^{1/2}$.

Clearly, this reminds us of the one-sided Lipschitz condition used for contractivity in the regular ODE case (i.e. $A(t) = I$, $P = I$ in (2.1)). In the latter case we have, with $y = -B(t)x$,

$$\langle B(t)x, x\rangle_S \le -c|x|_S.$$

However, things are more difficult for index-2 DAE's. First, considering the decoupled system (2.10) – (2.12) again, we observe that each solution of the homogeneous DAE (2.1) satisfies the identities

$$\begin{aligned}
&Q_1(t)x(t) = 0,\\
&y(t) := (Px)'(t) = (PP_1x)'(t) + (PQ_1x)'(t) = (PP_1x)'(t),\\
&Qy(t) = 0, \quad Q_1(t)y(t) = Q_1(t)(PP_1x)'(t) = -Q_1'(t)PP_1(t)x(t).
\end{aligned}$$

Now, inspired by the notion of contractivity given for the index-1 case in [4], we state the following definition.

Definition. The index-2 tractable DAE (2.1) is called *contractive* if the following holds: There is a constant $c > 0$ and a symmetric positive-definite matrix S such that

$$A(t)y + B(t)x = 0, \quad Qy = 0, \quad Q_1(t)y = -Q_1'(t)PP_1(t)x, \quad x, y \in \mathbb{R}^m \tag{2.19}$$

imply

$$\langle y, Px\rangle_S \le -c|Px|_S^2. \tag{2.20}$$

As usually, with this notion of contractivity, too, we aim at an inequality

$$|Px(t)|_S \le e^{-c(t-t_0)}|Px(t_0)|_S$$

for all solutions of the homogeneous DAE, that shows the component $Px = PP_1x$ to decrease in that norm. The following theorem will show: If the canonical projector $\Pi(t)$ is uniformly bounded, then the complete solution $x(t)$ decreases.

Theorem 2.3. *Let* (2.1) *be index-*2 *tractable,* Q_1 *belong to* C^1*,* $\Pi(t)$ *be uniformly bounded on* J *and* (2.1) *be contractive. Then, it holds for each solution of the homogeneous equation that*

$$|x(t)|_S \le \gamma e^{-c(t-t_0)}|Px(t_0)|_S, \quad t \ge t_0, \tag{2.21}$$

where γ *is a bound of* $|\Pi(t)|_S$.

Proof. We have $Q_1(t)x(t) = 0$, further $x(t) = \Pi(t)PP_1(t)x(t) = \Pi(t)Px(t)$. □

Not surprisingly, we obtain

COROLLARY 2.4. *Let* (2.1) *be index-2 tractable with continuously differentiable* Q_1 *and uniformly bounded* $\Pi(t)$. *If the condition*

$$\langle u, \{(PP_1)'(t) - (PP_1G_2^{-1}B)(t)\}u\rangle_S \leq -c|u|_S^2, \tag{2.22}$$

is satisfied for all $u \in \operatorname{im} PP_1(t)$, $t \in J$, *the estimate* (2.21) *is valid.*

Proof. It may be checked immediately that (2.19) and (2.22) lead to (2.20), i.e., (2.22) implies contractivity. □

Note that there is no need for assuming (2.22) for all $u \in \mathbb{R}^m$. For the assertion of Corollary 2.4 to become true, it is sufficient that (2.22) holds for all $u \in \operatorname{im} PP_1(t)$, only.

Inequality (2.22) looks like the usual contractivity condition for the regular ODE (2.10), i.e., the inherent regular ODE of (2.1). The only difference is that the values u are taken from the subspace $\operatorname{im} PP_1(t)$ instead of all of $\mathbb{R}^m$. Roughly speaking, one has: The DAE (2.1) is contractive if the inherent regular ODE (2.10) is contrctive on the subspace $\operatorname{im} PP_1(t)$.

As a direct consequence of other results on stability ([15], e.g.) one can deduce counterparts for linear index-2 DAE's, e.g. the well-known Poincaré-Lyapunov Theorem.

3. Specification of the projector framework for index-2 Hessenberg-form DAE's. Most authors restrict their interest to so-called *Hessenberg-form equations*, i.e., to systems

$$\begin{array}{rcl} x_1' + B_{11}x_1 + B_{12}x_2 & = & q_1 \\ B_{21}x_1 & = & q_2 \end{array}, \tag{3.1}$$

where $x = (x_1^T, x_2^T)^T$, $x_1 \in \mathbb{R}^{m_1}$, $x_2 \in \mathbb{R}^{m_2}$, $m = m_1 + m_2$. In our context this corresponds to

$$A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & B_{12} \\ 0 & 0 \end{bmatrix},$$

$S(t) = S_1(t) = \{(z_1^T, z_2^T)^T \in \mathbb{R}^m : B_{21}(t)z_1 = 0\}$.

Obviously, $z \in S_1(t) \cap N_1(t)$ implies $z = 0$ if and only if $B_{21}(t)B_{12}(t)$ is nonsingular, which is the well-known Hessenberg-form index-2 condition. Under this condition the block

$$B_{12}(t)(B_{21}(t)B_{12}(t))^{-1}B_{21}(t) =: H(t)$$

is also a projector. It projects onto $\operatorname{im} B_{12}(t)$ along $\ker B_{21}(t)$.

Denote $E := B_{12}(B_{21}B_{12})^{-1}$ and $F := (B_{21}B_{12})^{-1}B_{21}$. It holds

$$Q_1 = \begin{bmatrix} H & 0 \\ -F & 0 \end{bmatrix}, \quad PP_1 = \begin{bmatrix} I - H & 0 \\ 0 & 0 \end{bmatrix}, \quad PQ_1 = \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.2}$$

Furthermore, one has

$$PP_1G_2^{-1}B = \begin{bmatrix} (I - H)B_{11}(I - H) & 0 \\ 0 & 0 \end{bmatrix}.$$

The canonical projector Π is

$$\Pi = \begin{bmatrix} I-H & 0 \\ -FB_{11}(I-H)+F'(I-H) & 0 \end{bmatrix}.$$

Recall that $M(t) = \operatorname{im}\Pi(t) \subset \mathbb{R}^m$ is precisely the solution space of the homogeneous form of (3.1). It is time dependent if the projector $H(t)$ is. However, $M(t)$ may also rotate with t even if $H(t)$ is independent of time. Note that PP_1 is easier to compute than Π.

Furthermore, the nontrivial part (i.e., dropping the zero rows) of the inherent regular ODE (2.10) reads now as

$$u_1' + H'u_1 + (I-H)B_{11}u_1 = (I-H)q_1 - (H' + (I-H)B_{11})Eq_2, \tag{3.3}$$

where $u_1 := (I-H)x_1$. Let us emphasize once more that quickly varying subspaces may cause the term H' to dominate within this regular ODE. Clearly, $H'u_1$ corresponds to the term $(PP_1)'u$ in (2.10).

Theorems 2.1 and 2.2 apply immediately. In particular, we obtain: Suppose $H(t)$ and $(I-H(t))B_{11}(t)(I-H(t))$ are time-invariant. Then the eigenvalues of $(I-H(t))B_{11}(t)(I-H(t))$ determine the asymptotic behaviour of the solution.

Now, let us turn to the discussion of aspects of contractivity. For index-2 Hessenberg-form DAE's (3.1), relation (2.20) applies to the first components only, i.e.

$$\langle y_1, x_1\rangle_{S_{11}} \le -c|x_1|^2_{S_{11}}$$

should be satisfied if (cf. (2.19))

$$\begin{aligned} y_1 + B_{11}(t)x_1 + B_{12}(t)x_2 &= 0, \\ B_{21}(t)x_1 \qquad\qquad &= 0, \\ y_2 = 0,\ \ H(t)y_1 = -H'(t)(I-H(t))x_1&. \end{aligned}$$

Moreover, (2.22) simplifies to

$$\langle u_1, \{-H'(t) - (I-H(t))B_{11}(t)(I-H(t))\}u_1\rangle_{S_{11}} \le -c|u_1|^2_{S_{11}}, \quad \text{for all } u_1 \in \operatorname{im}(I-H(t)), \quad t \in J.$$

Again we see that the constant-subspace case $H'(t) \equiv 0$ becomes much easier.

It should be stressed that the above decoupling as well as the inherent ODE are stated in the original coordinates. In particular, the subspace $M(t) \subset \mathbb{R}^m$ is precisely the one that contains the solutions of the original DAE. No coordinate transformation is applied and only a decomposition into characteristic components is employed.

Ascher and Petzold [2] use a different approach to decouple characteristic parts of linear index-2 Hessenberg systems: They use a coordinate change $x = Tz$ such that

$$\begin{aligned} x_1 &= Sz_1 + Ez_3 \\ x_2 &= -(B_{21}B_{12})^{-1}(-B_{21}' + B_{21}B_{11})Sz_1 + (B_{21}B_{12})^{-1}z_2 \end{aligned}$$

and

$$\begin{aligned} z_1 &= Rx_1 \\ z_2 &= (-B_{21}' + B_{21}B_{11})SRx_1 + B_{21}B_{12}x_2 \\ z_3 &= B_{21}x_1 \end{aligned}$$

(cf. also [16]). In [2] the matrices R and S are constructed in the following way. Let $m_1 > m_2$. First, a matrix R with $m_1 - m_2$ linearly independent rows is chosen so that

$$RB_{12} = 0, \quad \text{i.e.} \quad R = R(I - H)$$

is satisfied. As a consequence, the $m_1 \times m_1$ block

$$\begin{bmatrix} R \\ B_{21} \end{bmatrix}$$

is nonsingular. Choosing S in such a way that $RS = I$, $SR = I - H$, $S = (I - H)S$ hold true, we have $\begin{bmatrix} R \\ B_{21} \end{bmatrix}^{-1} = [S \;\; E]$. The relation $\begin{bmatrix} z_1 \\ z_3 \end{bmatrix} = \begin{bmatrix} R \\ B_{21} \end{bmatrix} x_1$ exhibits the main idea of that transformation. $R(t)$ and $S(t)$ are assumed to be smooth. Carrying out a few straightforward computations one obtains a regular ODE for the component $z_1 = Rx_1$, namely

$$z_1' = (R' - RB_{11})Sz_1 + Rq_1 + (R' - RB_{11})Eq_2. \tag{3.4}$$

Equation (3.4) is said to be the *essentially underlying ODE (EUODE)* of the DAE (3.1).

What does the EUODE have in common with the inherent regular ODE? What is the difference?

Multiplying the EUODE (3.4) by S and taking into account that

$$Sz_1 = SRx_1 = (I - H)x_1 = u_1$$

is given, we obtain (3.3). On the other hand, scaling the inherent regular ODE (3.3) by R leads back to the EUODE (3.4) because

$$Ru_1 = R(I - H)x_1 = R(I - H)Sz_1 = z_1.$$

Thus, the EUODE turns out to be nothing else but a scaled version of the inherent regular ODE and vice versa. Due to

$$R(I - H)S = RS = I$$

the $(m_1 - m_2)$-dimensional subspace $\operatorname{im}(I - H(t)) = \ker B_{21}(t) \subset \mathbb{R}^{m_1}$ is uniformly traced back to $\mathbb{R}^{m_1 - m_2}$. Thus, the EUODE has the advantage to be written in the minimal coordinate space $\mathbb{R}^{m_1 - m_2}$. Unfortunately, the matrices R and S are not uniquely determined. Consequently, the EUODE is strongly affected by the choice of R, S. Note that once an R is chosen, we may multiply by any regular $K \in L(\mathbb{R}^{m_1 - m_2})$ to obtain another one by $\tilde{R} := KR$.

From this point of view, the inherent regular ODE (3.3) seems to be more natural, since all its terms are uniquely determined by the original data. $u_1 = (I - H)x_1$ is a direct component of the original variable x_1, but the ODE (3.3) lives in the higher-dimensional space $\mathbb{R}^{m_1}$, and $\operatorname{im}(I - H(t)) = \ker B_{21}(t)$ represents an invariant subspace.

Ascher and Petzold [2] observed that the Euler backward method applied to the DAE (3.1) may behave like an explicit Euler method. Choose $q = 0$, $B_{11} = 0$ in (3.1), which simplifies the EUODE to

$$z_1' = R'Sz_1. \tag{3.5}$$

Via the transform $x = Tz$, the Euler backward formula applied to this special DAE (3.1) yields

$$\frac{1}{h}(z_{1,n+1} - z_{1,n}) = \int_0^1 R'(t_n + sh)\mathrm{d}s(S(t_n)z_{1,n} - E(t_n)z_{3,n}). \tag{3.6}$$

If additionally $B_{12}(t)$ and $R'(t)$ do not vary with t, (3.6) simplifies to

$$\frac{1}{h}(z_{1,n+1} - z_{1,n}) = R'(t_n)S(t_n)z_{1,n}.$$

This is the explicit Euler formula for (3.5). Clearly this phenomenon is closely related to time-varying blocks $R(t)$ and $S(t)$ of the coordinate transformation. Let us mention again that this behaviour depends on the choice of R and S.

In the following section we show that the behaviour of the characteristic subspace im $(I - H(t))$ resp. im $PP_1(t)$ in the general case is decisive for understanding what really happens.

4. Asymptotic stability of integration methods. A number of widely used notions for the characterization of asymptotic properties of integration methods for *explicit* ODE's relies on the complex scalar test equation

$$z' = \lambda z. \tag{4.1}$$

The asymptotic behaviour of a numerical method applied to (4.1) characterizes the asymptotics in the case of linear constant coefficient systems

$$x' = -Bx. \tag{4.2}$$

Here, the role of λ is replaced by the eigenvalues of $-B$. The justification for restricting the consideration to (4.2) is given by Lyapunov's theory: The linearization of a nonlinear autonomous explicit system at a stationary point provides criteria for the asymptotic behaviour of solutions. In essence, the same is true for index-1 and -2 DAE's [12]. Therefore, we are led to the constant coefficient DAE

$$Ax'(t) + Bx(t) = 0 \tag{4.3}$$

with regular matrix pencil $\lambda A + B$. This equation can be transformed into the *Kronecker canonical form*

$$\begin{bmatrix} I & 0 \\ 0 & J_0 \end{bmatrix} \begin{bmatrix} y' \\ z' \end{bmatrix} + \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = 0. \tag{4.4}$$

J_0 is a nilpotent matrix ($J_0^k = 0$ for some integer k). Since discretization and transformation to (4.4) commute for many methods, the numerical solution for z vanishes identically, whereas y is discretized like an explicit system. Hence, numerical methods applied to constant coefficient linear DAE's trivially preserve their asymptotic stability properties that are based on the test equation (4.1) (e.g. A-, $A(\alpha)$-, L-stability). Thus, at first glance, one could expect the well-known concepts of asymptotics in the numerical integration of explicit ODE's to be sufficient for DAE's, too. However, as described in Sections 2 and 3, DAE's have a more difficult structure than explicit ODE's, even in view of numerical integration. Roughly speaking, we should not expect the numerical methods to match the subspace structure exactly if those subspaces rotate. The scalar test equation (4.1) turns out to be an inappropriate model in case of DAE's.

Similar results about B-stability are more difficult to obtain. It is well-known that so-called *algebraically stable* Runge-Kutta methods are B-stable [6, p. 193] for explicit systems. In [4, p.129] a similar result is shown to be true for index-1 DAE's provided that (i) the nullspace $N(A(t))$ of $A(t)$ does *not* depend on t, and (ii) the Runge-Kutta method is a so-called IRK(DAE) (a stiffly accurate method [6, p. 45]). There are simple linear examples showing that the backward Euler method loses its B-stability if (i) is not valid.

We recall the notion of B-stability for DAE's having a constant leading nullspace:

Definition [4]. The one-step method $x_{j+1} = \varphi(x_j, t_j, h_j)$ is called B-stable if for each contractive DAE the inequalities

$$|Px^{(1)}_{j+1} - Px^{(2)}_{j+1}|_S \le |Px^{(1)}_j - Px^{(2)}_j|_S,$$

and

$$|Qx^{(1)}_{j+1} - Qx^{(2)}_{j+1}|_S \le K|Px^{(1)}_{j+1} - Px^{(2)}_{j+1}|_S, \quad j \ge 0,$$

are satisfied. Here, $K > 0$ is a constant and $x^{(1)}_0$, $x^{(2)}_0$ are arbitrary consistent initial values.

4.1. BDF applied to linear index-2 DAE's. The k-step BDF applied to (2.1) reads as

$$A(t_\ell)\frac{1}{h}\sum_{j=0}^{k}\alpha_j x_{\ell-j} + B(t_\ell)x_\ell = q(t_\ell), \quad \ell \ge k. \tag{4.5}$$

At each step, equation (4.5) provides an approximation x_ℓ of the exact solution value $x(t_\ell)$, $t_\ell := t_0 + \ell h$. Recall that the nullspace of $A(t)$ is assumed to be constant.

Supposed (2.1) is index-2 tractable, we may decouple (4.5) and (2.1) simultaneously (cf. (2.10) – (2.12)), which yields

$$\frac{1}{h}\sum_{j=0}^{k}\alpha_j u_{\ell-j} + \frac{1}{h}\sum_{j=1}^{k}\alpha_j(PP_{1,\ell} - PP_{1,\ell-j})u_{\ell-j} + PP_{1,\ell}G^{-1}_{2,\ell}B_\ell u_\ell$$

$$= PP_{1,\ell}G^{-1}_{2,\ell}q_\ell - \frac{1}{h}\sum_{j=1}^{k}\alpha_j(PP_{1,\ell} - PP_{1,\ell-j})v_{\ell-j}, \tag{4.6}$$

$$-\frac{1}{h}\sum_{j=0}^{k}\alpha_j Qv_{\ell-j} - \frac{1}{h}\sum_{j=1}^{k}\alpha_j(QQ_{1,\ell} - QQ_{1,\ell-j})(u_{\ell-j} + Pv_{\ell-j}) + w_\ell + QP_{1,\ell}G^{-1}_{2,\ell}B_\ell u_\ell$$

$$= QP_{1,\ell}G^{-1}_{2,\ell}q_\ell, \tag{4.7}$$

$$v_\ell = Q_{1,\ell}G^{-1}_{2,\ell}q_\ell, \tag{4.8}$$

where we have used the above decomposition again, i.e.,

$$x_\ell = PP_{1,\ell}x_\ell + PQ_{1,\ell}x_\ell + Qx_\ell =: u_\ell + Pv_\ell + w_\ell. \tag{4.9}$$

In particular, if the inhomogeneity q vanishes identically, then the Q_1-components $v(t_\ell)$ and v_ℓ are both zero, and one has

$$\frac{1}{h}\sum_{j=0}^{k}\alpha_j u_{\ell-j} + \frac{1}{h}\sum_{j=1}^{k}\alpha_j P(P_{1,\ell} - P_{1,\ell-j})u_{\ell-j} + PP_{1,\ell}G_{2,\ell}^{-1}B_\ell u_\ell = 0 \tag{4.10}$$

for approximation of

$$u' - PP_1'u + PP_1G_2^{-1}Bu = 0 \tag{4.11}$$

and

$$w_\ell = -QP_{1,\ell}G_{2,\ell}^{-1}B_\ell u_\ell + \frac{1}{h}\sum_{j=1}^{k}\alpha_j Q(Q_{1,\ell} - Q_{1,\ell-j})u_{\ell-j}$$

for approximation of

$$w = -QP_1G_2^{-1}Bu + QQ_1'u.$$

The following proposition is an immediate consequence.

PROPOSITION 4.1. *Let* (2.1) *be index-*2 *tractable with continuously differentiable* Q_1. *Then the BDF method applied to* (2.1) *generates exactly the same BDF method applied to the inherent regular ODE* (4.11) *if and only if the projector* $PP_1(t)$ *does not vary with* t. *For a constant projector* PP_1, *the BDF methods retain their asymptotic stability properties for index-*2 *DAE's provided the canonical projector* $\Pi(t)$ *remains uniformly bounded.*

On the other hand, varying subspaces may cause the term PP_1' to dominate the inherent regular ODE itself. For instance, the backward Euler method provides then

$$\frac{1}{h}(u_\ell - u_{\ell-1}) - \frac{1}{h}P(P_{1,\ell} - P_{1,\ell-1})u_{\ell-1} + PP_{1,\ell}G_{2,\ell}^{-1}B_\ell u_\ell = 0,$$

which shows that $u(t_\ell) - u_\ell \to 0$ ($\ell \to \infty$) may or may not happen. As it was mentioned in Section 3, Ascher and Petzold [2] have observed this phenomenon in case of linear index-2 Hessenberg systems (3.1) (cf. also Section 3). However, this is not surprising since we cannot expect any discretization method to follow the subspaces precisely without profound information on the inner structure of the DAE.

Naturally, similar arguments apply to Runge-Kutta methods, too.

4.2. Implicit Runge-Kutta methods and their projected counterparts applied to linear index-2 DAE's. According to the originally conceived method for the numerical solution of ordinary differential equations, an implicit Runge-Kutta (IRK) method can be realized for the DAE (2.1) in the following way [14]: Given an approximation $x_{\ell-1}$ of the solution of (2.1) at $t_{\ell-1}$, a new approximation x_ℓ at $t_\ell = t_{\ell-1} + h$ is obtained via

$$x_\ell = x_{\ell-1} + h\sum_{i=1}^{s} b_i X_{\ell i}', \tag{4.12}$$

where $X'_{\ell i}$ is defined by

$$A(t_{\ell i})X'_{\ell i} + B(t_{\ell i})X_{\ell i} = q(t_{\ell i}), \quad i = 1, \dots, s, \tag{4.13}$$
$$t_{\ell i} = t_{\ell-1} + c_i h$$

and the internal stages are given by

$$X_{\ell i} = x_{\ell-1} + h \sum_{j=1}^{s} a_{ij} X'_{\ell j}, \quad i = 1, \dots, s. \tag{4.14}$$

The coefficients a_{ij}, b_i, c_i determine the IRK method, and s represents the number of stages. Assume the matrix $A := (a_{ij})_{i,j=1}^{s}$ to be nonsingular and denote its inverse by $(\hat{a}_{ij})_{i,j=1}^{s}$. Let $\varrho := 1 - \sum\limits_{i=1}^{s} \sum\limits_{j=1}^{s} b_i \hat{a}_{ij}$.

Equations (4.12) – (4.14) are equivalent to

$$x_\ell = \varrho x_{\ell-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} X_{\ell j}, \tag{4.15}$$

$$A(t_{\ell i}) \sum_{j=1}^{s} \hat{a}_{ij}(X_{\ell j} - x_{\ell-1}) + hB(t_{\ell i})X_{\ell i} = hq(t_{\ell i}),\ i = 1, \dots, s. \tag{4.16}$$

Looking at (4.16) we observe that the internal stages do not depend on $Qx_{\ell-1}$.

The special class of IRK methods (IRK(DAE)) with coefficients

$$b_i = a_{si}, \quad i = 1, \dots, s, \quad c_s = 1 \tag{4.17}$$

is shown to stand out from all IRK methods in view of their applicability to DAE's [4]. Since $\varrho = 0$ in that case, the new value $x_\ell = X_{\ell s}$ always belongs to the obvious constraint manifold

$$\mathcal{N}(t_\ell) := \{z \in \mathbb{R}^m : B(t_\ell)z - q(t_\ell) \in \operatorname{im} A(t_\ell)\}.$$

Therefore we have

$$B(t_\ell)x_\ell - q(t_\ell) \in \operatorname{im} A(t_\ell). \tag{4.18}$$

For Hessenberg equations (3.1), relation (4.18) simplifies to

$$B_{21}(t_\ell)x_{1,\ell} = q_2(t_\ell). \tag{4.19}$$

In general, if (4.17) is not fulfilled, then we have $\varrho \neq 0$, and (4.18) resp. (4.19) are no longer true. Since this behaviour is a source of instability (for $h \to 0$), Ascher and Petzold [1] propose another version for the application of IRK methods to index-2 Hessenberg DAE's (2.18), the so-called Projected IRK (PIRK). Actually, after realizing the standard internal stage computation, the recursion (4.15) is now replaced by

$$\hat{x}_{1,\ell} = \varrho \hat{x}_{1,\ell-1} + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \hat{a}_{ij} \hat{X}_{1,\ell j} + B_{12}(t_\ell)\lambda_\ell \tag{4.20}$$

and λ_ℓ is determined by

$$B_{21}(t_\ell)\hat{x}_{1,\ell} = q_2(t_\ell). \tag{4.21}$$

If we multiply (4.20) by $I - H(t_\ell)$, λ_ℓ can be eliminated:

$$(I - H(t_\ell))\hat{x}_{1,\ell} = \varrho(I - H(t_\ell))\hat{x}_{1,\ell-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\hat{a}_{ij}(I - H(t_\ell))\hat{X}_{1,\ell j}. \tag{4.22}$$

On the other hand, (4.21) is equivalent to

$$H(t_\ell)\hat{x}_{1,\ell} = B_{12}(t_\ell)\left(B_{21}(t_\ell)B_{12}(t_\ell)\right)^{-1} q_2(t_\ell). \tag{4.23}$$

It should be mentioned that for IRK(DAE) the projected version is exactly the same as the original one, since (4.17) implies $\lambda_\ell = 0$ in (4.20), (4.21).

Considering (4.22) – (4.23) in association with the projector formulae (3.2), an immediate generalization of PIRK methods to fully implicit linear index-2 systems (2.1) is suggested by

$$PP_1(t_\ell)\hat{x}_\ell = \varrho PP_1(t_\ell)\hat{x}_{\ell-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\hat{a}_{ij}PP_1(t_\ell)\hat{X}_{\ell j}, \tag{4.24}$$

$$Q_1(t_\ell)\hat{x}_\ell = Q_1(t_\ell)G_2^{-1}(t_\ell)q(t_\ell). \tag{4.25}$$

Since the internal stages $\hat{X}_{\ell j}$ do not depend on $Q\hat{x}_{\ell-1}$, there is no need to compute $Q\hat{x}_\ell$ at this stage.

Now return to the standard IRK (4.15) – (4.16) and decouple (4.16) in the same way as (2.10) – (2.12). For that, we decompose

$$\begin{aligned} X_{\ell i} &= PP_1(t_{\ell i})X_{\ell i} + PQ_1(t_{\ell i})X_i + QX_{\ell i} =: U_{\ell i} + PV_{\ell i} + W_{\ell i}, \\ x_\ell &= PP_1(t_\ell)x_\ell + PQ_1(t_\ell)x_\ell + Qx_\ell =: u_\ell + Pv_\ell + w_\ell. \end{aligned}$$

A straightforward computation yields

$$\begin{aligned} &\frac{1}{h}\sum_{j=1}^{s}\hat{a}_{ij}(U_{\ell j} - u_{\ell-1}) + P\Big\{\frac{1}{h}\sum_{j=1}^{s}\hat{a}_{ij}(P_1(t_{\ell i}) - P_1(t_{\ell j}))U_{\ell j}\Big\} \\ &- P\big\{\frac{1}{h}\sum_{j=1}^{s}\hat{a}_{ij}(P_1(t_{\ell i}) - P_1(t_{\ell-1}))x_{\ell-1}\Big\} + PP_1G_2^{-1}B(t_{\ell i})U_{\ell i} \\ &\qquad = PP_1G_2^{-1}q(t_{\ell i}) - P\Big\{\frac{1}{h}\sum_{j=1}^{s}(P_1(t_{\ell i}) - P_1(t_{\ell j}))V_{\ell j}\Big\}, \end{aligned} \tag{4.26}$$

$$\begin{aligned} &\frac{1}{h}\sum_{j=1}^{s}\hat{a}_{ij}(QV_{\ell j} - Qv_{\ell-1}) - Q\Big\{\frac{1}{h}\sum_{j=1}^{s}(Q_1(t_{\ell i}) - Q_1(t_{\ell j}))(U_{\ell j} + PV_{\ell j})\Big\} \\ &+ Q\Big\{\frac{1}{h}\sum_{j=1}^{s}(Q_1(t_{\ell i}) - Q_1(t_{\ell-1}))(u_{\ell-1} + Pv_{\ell-1})\Big\} + W_{\ell i} + QP_1G_2^{-1}B(t_{\ell i})U_{\ell i} \\ &\qquad = QP_1G_2^{-1}q(t_{\ell i}), \end{aligned} \tag{4.27}$$

$$V_{\ell i} = Q_1G_2^{-1}q(t_{\ell i}). \tag{4.28}$$

The recursion (4.15) can be decomposed simply by multiplying by the projections:

$$
\begin{aligned}
u_\ell = {}&\varrho u_{\ell-1} + \varrho P(P_1(t_\ell) - P_1(t_{\ell-1}))(u_{\ell-1} + Pv_{\ell-1}) + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} U_{\ell j} \\
&+ P \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} (P_1(t_\ell) - P_1(t_{\ell j})(U_{\ell j} + PV_{\ell j})
\end{aligned} \tag{4.29}
$$

$$
\begin{aligned}
Pv_\ell = {}&\varrho Pv_{\ell-1} + \varrho P(Q_1(t_\ell) - Q_1(t_{\ell-1}))(u_{\ell-1} + Pv_{\ell-1}) + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} PV_{\ell j} \\
&+ P \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} (Q_1(t_\ell) - Q_1(t_{\ell j})(U_{\ell j} + PV_{\ell j})
\end{aligned} \tag{4.30}
$$

$$
w_\ell = \varrho w_{\ell-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} W_{\ell j} \tag{4.31}
$$

Now, consider the homogeneous case, that is we set $q = 0$. If the inhomogeneity q vanish identically, then v does so, too. Moreover, all values $V_{\ell i}$ are equal to zero. However, if $\varrho \neq 0$, this is no longer true for $Q_1 P v_\ell = v_\ell$. This means that, in general, the resulting x_ℓ has a nontrivial component $Q_1(t_\ell)x_\ell$ in contrast to the exact solution that fulfills $Q_1(t_\ell)x(t_\ell) = 0$.

In more detail, (4.26) reduces to

$$
\begin{aligned}
&\frac{1}{h}\sum_{j=1}^{s} \hat{a}_{ij}(U_{\ell j} - u_{\ell-1}) + P\Big\{\frac{1}{h}\sum_{j=1}^{s} \hat{a}_{ij}(P_1(t_{\ell i}) - P_1(t_{\ell j}))U_{\ell j}\Big\} \\
&-P\Big\{\frac{1}{h}\sum_{j=1}^{s}(P_1(t_{\ell i}) - P_1(t_{\ell-1}))x_{\ell-1}\Big\} + PP_1G_2^{-1}B(t_{\ell i})U_{\ell i} = 0,
\end{aligned} \tag{4.32}
$$

which supposedly approximates

$$
u' - PP_1'u + PP_1G_2^{-1}Bu = 0. \tag{4.33}
$$

Moreover, (4.27) yields

$$
\begin{aligned}
W_{\ell i} = {}& -QP_1G_2^{-1}B(t_{\ell i})U_{\ell i} + Q\Big\{\frac{1}{h}\sum_{j=1}^{s}(Q_1(t_{\ell i}) - Q_1(t_{\ell j}))U_{\ell j}\Big\} \\
&-Q\Big\{\frac{1}{h}\sum_{j=1}^{s}(Q_1(t_{\ell i}) - Q_1(t_{\ell-1}))u_{\ell-1}\Big\}
\end{aligned} \tag{4.34}
$$

for approximating

$$
w = -QP_1G_2^{-1}Bu + QQ_1'u.
$$

In the consequence, the following result holds true for IRK methods analogously to Proposition 4.1 for the case of BDF methods:

PROPOSITION 4.2. *Let* (2.1) *be index-*2 *tractable with continuously differentiable* Q_1. *Then the IRK method applied to* (2.1) *generates exactly the same IRK method applied to the inherent regular ODE* (4.30) *if and only if* $PP_1(t)$ *does not vary with* t.

For constant PP_1, the solution

$$x(t) = \Pi(t)u(t) = (I - (QP_1G_2^{-1}B)(t))u(t), \tag{4.35}$$

$$u' + PP_1G_2^{-1}Bu = 0, \tag{4.36}$$

$$u(t_0) \in \operatorname{im} PP_1, \tag{4.37}$$

of the homogeneous equation is approximated at t_ℓ by

$$x_\ell = u_\ell + Pv_\ell + w_\ell, \tag{4.38}$$

$$u_\ell = \varrho u_{\ell-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} U_{\ell j}, \tag{4.39}$$

$$\frac{1}{h}\sum_{j=1}^{s} \hat{\alpha}_{ij}(U_{\ell j} - u_{\ell-1}) + (PP_1G_2^{-1}B)(t_{\ell i})U_{\ell i} = 0,\ i = 1, \ldots, s, \tag{4.40}$$

$$Pv_\ell = \varrho Pv_{\ell-1}, \tag{4.41}$$

$$w_\ell = \varrho w_{\ell-1} - \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \hat{a}_{ij} (QP_1G_2^{-1}B)(t_{\ell j})U_{\ell j}. \tag{4.42}$$

Starting with a consistent initial value x_0 (with $v_0 = 0$), the components v_ℓ vanish step by step, too.

For IRK(DAE), (4.42) provides

$$w_\ell = -(QP_1G_2^{-1}B)(t_\ell)U_{\ell s}, \quad u_\ell = U_{\ell s},$$
$$x_\ell = (I - (QP_1G_2^{-1}B)(t_\ell))u_\ell,$$

that is, in the case of constant PP_1, the approximation x_ℓ belongs to the solution manifold $M(t_\ell)$ given in Theorem 2.1.

Let us briefly turn to PIRK methods (4.24), (4.25). For homogeneous equations, (4.25) yields $\hat{v}_\ell = Q_1(t_\ell)\hat{x}_\ell = 0$. The decoupled system parts (4.32), (4.34) remain valid also for the "$\hat{}$" values.

Proposition 4.3. *Proposition* 4.2 *is true for PIRK methods, too.*

It should be mentioned that, for constant PP_1, in PIRK methods we have simply

$$\hat{v}_\ell = 0. \tag{4.43}$$

instead of (4.41). Ascher and Petzold [1] have not considered a recursion for the component $Q\hat{x}_\ell$ for Hessenberg systems (2.19). Nevertheless, if one is interested in approximations $Q\hat{x}_\ell$, a recursion like (4.42) will come up again. In that case, the only difference between PIRK and IRK methods is the determination of the Q_1-components ((4.43) versus (4.41)). Note again that PIRK and IRK are identical for IRK(DAE).

Next, concerning B-stability, the following assertion shows the notion of contractivity given in Section 2 to be useful.

Theorem 4.4. *Let* (2.1) *be index-*2 *tractable with continuously differentiable* Q_1, $(PP_1)' = 0$, *and* $\|\Pi(t)\| \le K$, $t \in [t_0, \infty)$. *Then, each algebraically stable IRK(DAE) applied to* (2.1) *is B-stable.*

Proof. Denote $m_{ij} := b_i a_{ij} + b_j a_{ji} - b_j b_i$. Due to the algebraical stability, $m = (m_{ij})_{ij}$ is a positively semi-definite matrix.

Since we deal with linear DAE's only, it remains to show the inequalities $|Px_\ell|_S \le |Px_{\ell-1}|_S$, $|Qx_\ell|_S \le K|Px_\ell|$ for the case of the homogeneous equation (2.1).

With $q = 0$, (4.13) yields

$$Q_1(t_{\ell i})X_{\ell i} = 0, \quad i = 1, \dots, s,$$

therefore $PX_{\ell i} = PP_1(t_\ell)X_{\ell i}$. Additionally, with an IRK (DAE) we also have $Q_1(t_{\ell-1})x_{\ell-1} = 0$. Then, (4.14) implies

$$\begin{aligned} Q_1(t_{\ell i})X'_{\ell i} &= Q_1(t_{\ell i})\frac{1}{h}\sum_{j=1}^{s} \hat{a}_{ij}(X_{\ell j} - x_{\ell-1}) \\ &= Q_1(t_{\ell i})\frac{1}{h}\sum_{j=1}^{s} \hat{a}_{ij}(PP_1(t_{\ell j})X_{\ell j} - PP_1(t_{\ell-1})x_{\ell-1}) = 0, \end{aligned}$$

since $Q_1(t_{\ell i})PP_1(\tilde{t}) = 0$ holds true for all $\tilde{t}$. Hence, using the contractivity (cf. Section 2) we obtain the inequalities

$$\langle PX'_{\ell i}, PX_{\ell i}\rangle_S \le -c|PX_{\ell i}|_S^2, \quad i = 1, \dots, s.$$

Now, following the standard lines, we compute

$$\begin{aligned} &|Px_\ell|_S^2 \\ &= |Px_{\ell-1} + h\sum_{i=1}^{s} b_i PX'_{\ell i}|_S^2 \\ &= |Px_{\ell-1}|_S^2 + 2h\sum_{i=1}^{s} b_i\langle Px_{\ell-1}, PX'_{\ell i}\rangle_S + h^2\sum_{i,j=1}^{s} a_{ij}\langle PX'_{\ell i}, PX'_{\ell j}\rangle_S \\ &= |Px_{\ell-1}|_S^2 + 2h\sum_{i=1}^{s} b_i\langle PX_{\ell i} - h\sum_{j=1}^{s} a_{ij}PX'_{\ell j}, PX'_{\ell i}\rangle_S + h^2\sum_{i,j=1}^{s} a_{ij}\langle PX'_{\ell i}, PX'_{\ell j}\rangle_S \\ &= |Px_{\ell-1}|_S^2 + 2h\sum_{i=1}^{s} b_i\langle PX_{\ell i}, PX'_{\ell i}\rangle_S - h^2\sum_{i,j=1}^{s} m_{ij}\langle PX'_{\ell i}, PX'_{\ell j}\rangle_S \\ &\le |Px_{\ell-1}|_S^2 - 2hc\sum_{i=1}^{s} b_i|PX_{\ell i}|^2 \\ &\le |Px_{\ell-1}|_S^2. \end{aligned}$$

Finally, $x_\ell = \Pi(t_\ell)Px_\ell$ implies

$$|Qx_\ell|_S \le K|Px_\ell|_S. \qquad \square$$

It should be noted that Theorem 4.4 does not apply to PIRK. While the first part, i.e., $|Px_\ell|_S \le |Px_{\ell-1}|$, holds true analogously, the necessary relation for the nullspace component is not given at all for $\varrho \ne 0$.

5. A numerical counterexample. In the previous sections we have seen that BDF and Runge-Kutta methods preserve their stability behaviour if PP_1 is constant. The following example shows that these properties get lost if PP_1 varies with time. Consider the DAE

$$A(t)x'(t) + B(t)x(t) = 0, \quad t \ge 0, \tag{5.1}$$

with

$$A(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B(t) = \begin{bmatrix} \lambda & -\beta & -1 \\ \beta\eta t(1-\eta t) - \eta & \lambda & -\eta t \\ 1-\eta t & 1 & 0 \end{bmatrix},$$

where $\beta, \lambda, \eta \in \mathbb{R}$ are constant. Note that (5.1) is an index-2 Hessenberg system. One easily computes (using $A(t) \equiv P$)

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

such that

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

Compute the projections

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

Taking into account that $q = 0$ in (5.1), the inherent regular ODE (2.10) reads

$$u' + \mathcal{M}(t)u = 0, \tag{5.2}$$

where $u(t) = PP_1 x(t)$ and

$$\begin{aligned} \mathcal{M}(t) &= PP_1 G_2^{-1} B - (PP_1)' \\ &= \begin{bmatrix} \lambda\eta t - \eta + \eta^2 t & -\lambda - \eta & 0 \\ (\lambda+\eta)\eta t(\eta t - 1) + \eta(1 - 2\eta t) & (\lambda+\eta)(1-\eta t) + \eta & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

The solution subspace $M(t)$ (cf. Theorem 2.2) is given by

$$M(t) = \operatorname{im}\Pi(t) = \{x \in \mathbb{R}^3 : (\eta t - 1)x_1 - x_2 = 0,\ \beta x_2 + x_3 = 0\}.$$

Since $\dim \operatorname{im}(PP_1(t)) \equiv 1$, (5.2) subject to consistent initial values (2.13) may be reduced to the scalar ODE,

$$u_1' + \lambda u_1 = 0, \tag{5.3}$$

together with

$$u_2(t) = (\eta t - 1)u_1(t), \quad u_3(t) = 0. \tag{5.4}$$

Consequently, the asymptotic stability of (5.2) is governed by the sign of λ (independently of $\eta \in \mathbb{R}$). The parameter η measures the change of $N_1(t)$. β serves only for mixing the P component with the nullspace component. Now the complete solution of (5.1) can be easily computed using (2.10) – (2.12). If $x^0 \in \mathbb{R}^3$ is a consistent initial value at $t = 0$ (i.e., $x_1^0 + x_2^0 = 0$, $x_3^0 + \beta x_2^0 = 0$), the solution of (5.1) is

$$x(t) = \begin{bmatrix} x_1^0 e^{-\lambda t} \\ x_1^0(\eta t - 1)e^{-\lambda t} \\ x_1^0 \beta(1-\eta t)e^{-\lambda t} \end{bmatrix}.$$

(5.1) was solved using the 5-step BDF (Fig. 5.1) and an algebraically stable 2-stage Runge-Kutta method introduced by Crouzeix (cf. [5, p. 207]) with $\rho \approx -0.73$ (Fig 5.2). The figures show the norm of the numerical solution at the end of the interval $[0, T]$ for different values of η and λ. Note that, for $\eta = 0$, (5.1) represents a constant coefficient system. The results indicate that the asymptotic behaviour of the numerical solution depends not only on the asymptotic stability of the differential equation (5.3) (controlled by λ), but also on the geometry of the problem (controlled by η).

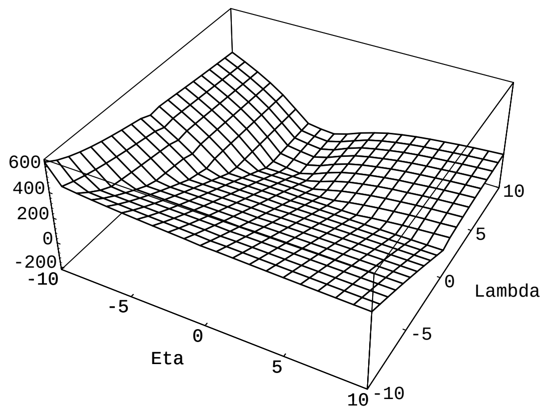

FIG. 5.1. *5-step BDF, $T = 40$, $\beta = 1$, $\lambda = -10, \ldots, +10$, $\eta = -10 \ldots + 10$, $h = 0.1$*

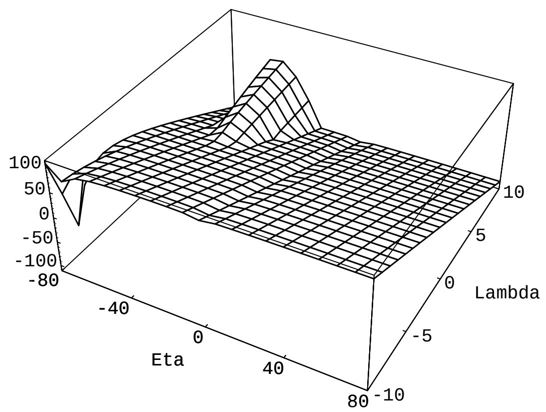

FIG. 5.2. *RKM of Crouzeix, $T = 10$, $\beta = 1$, $\lambda = -10 \ldots + 10$, $\eta = -80 \ldots + 80$, $h = 0.1$*

Appendix: Basic linear algebra lemma. A basic connection between the spaces appearing in the tractability index and the choice of the corresponding projectors is given by the following lemma, which may be directly obtained from Theorem A.13. and Lemma A.14. in [4].

LEMMA A.1. *Let* $\bar{A}, \bar{B}, \bar{Q} \in L(\mathbb{R}^m)$ *be given,* $\bar{Q}^2 = \bar{Q}$, $\mathrm{im}(\bar{Q}) = \ker(\bar{A})$*, i.e., let* $\bar{Q}$ *be a projector onto* $\ker(\bar{A})$*. Denote* $\bar{S} := \{z \in \mathbb{R}^m : \bar{B}z \in \mathrm{im}(\bar{A})\}$*. Then the following conditions are equivalent:*

(i) *The matrix* $\bar{G} := \bar{A} + \bar{B}\bar{Q}$ *is nonsingular.*

(ii) $\mathbb{R}^m = \bar{S} \oplus \ker(\bar{A})$.

(iii) $\bar{S} \cap \ker(\bar{A}) = \{0\}$.

If $\bar{G}$ *is nonsingular, then the relation*

$$\bar{Q}_s = \bar{Q}\bar{G}^{-1}\bar{B}$$

holds for the canonical projector $\bar{Q}_s$ (*canonical means:* $\bar{Q}_s$ *projects* $\mathbb{R}^m$ *onto* $\ker(\bar{A})$ *along* $\bar{S}$).

Proof. (i) $\rightarrow$ (ii) The space $\mathbb{R}^m$ can be described as $\bar{S} + \ker(\bar{A})$, because

$$z = (I - \bar{Q}\bar{G}^{-1}\bar{B})z + \bar{Q}\bar{G}^{-1}\bar{B}z =: z_1 + z_2 \qquad (*)$$

holds for any $z \in \mathbb{R}^m$. z_2 obviously lies in $\ker(\bar{A})$, because $\bar{Q}$ is a projector onto $\ker(\bar{A})$. For z_1 we obtain

$$\bar{B}z_1 = (I - \bar{B}\bar{Q}\bar{G}^{-1})\bar{B}z = \bar{A}\bar{G}^{-1}\bar{B}z \in \mathrm{im}(\bar{A}),$$

i.e., $z_1 \in \bar{S}$.

It remains to show that $\bar{S} \cap \ker(\bar{A}) = \{0\}$. To this end, let $x \in \bar{S} \cap \ker(\bar{A})$. Then $x = \bar{Q}x$ holds and there exists a $z \in \mathbb{R}^m$ such that $\bar{A}z = \bar{B}x = \bar{B}\bar{Q}x$ and $\bar{G}^{-1}\bar{A}z = \bar{G}^{-1}\bar{B}\bar{Q}x$. Consequently, $(I - \bar{Q})z = \bar{Q}x$, so $0 = \bar{Q}x = x$.

(ii) $\rightarrow$ (iii) This holds trivially by definition.

(iii) $\rightarrow$ (i) Let $x \in \mathbb{R}^m$ be chosen such that $\bar{G}x = 0$, i.e., $\bar{B}\bar{Q}x = -\bar{A}x$ and so $\bar{Q}x \in \bar{S}$. On the other hand, $\bar{Q}x$ lies in $\ker(\bar{A})$. Thus, $x \in \ker(\bar{Q})$ holds due to the assumption. That means, $\bar{A}x = 0$, hence $x \in \mathrm{im}(\bar{Q})$. Then $x = 0$ has to be true, and $\bar{G}$ is nonsingular.

Because of the uniqueness of the decomposition $(*)$, the latter assertion follows immediately. ∎

REFERENCES

[1] U. ASCHER AND L. R. PETZOLD, *Projected implicit Runge-Kutta methods for differential-algebraic equations*, SIAM J. Numer. Anal. 28 (1991), pp. 1097–1120.

[2] U. ASCHER AND L. R. PETZOLD, *Stability of computational methods for constrained dynamic systems*, SIAM J. Sci. Comp. 14 (1993), pp. 95–120.

[3] K. E. BRENAN, S. L. CAMPBELL AND L. R. PETZOLD, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, North Holland, New York, Amsterdam, London, 1989.

[4] E. GRIEPENTROG AND R. MÄRZ, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner Texte zur Mathematik 88, Teubner, Leipzig, 1986.

[5] E. HAIRER, S. P. NØRSETT AND G. WANNER, *Solving Ordinary Differential Equations I*, 2nd edition, Springer, Berlin, 1993.

[6] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II*, Springer, Berlin, 1991.

[7] E. IZQUIERDO MACANA, *Numerische Approximation von Algebro-Differentialgleichungen mit Index* 2 *mittels impliziter Runge-Kutta-Verfahren*, Doctoral thesis, Humboldt-Univ., Fachbereich Mathematik, Berlin, 1993.

[8] R. MÄRZ, *Index-2 differential-algebraic equations*, Results in Mathematics 15 (1989), pp. 148–171.
[9] ———, *Higher-index differential-algebraic equations: Analysis and numerical treatment*, Banach Center Publications 24 (1990), pp. 199–222.
[10] ———, *On quasilinear index 2 differential-algebraic equations.* Preprint 269, Humboldt-Univ., Fachbereich Mathematik, Berlin, 1991.
[11] ———, *Numerical methods for differential-algebraic equations*, Acta Numerica (1992), pp. 141–198.
[12] ———, *Practical Lyapunov stability criteria for differential-algebraic equations*, in Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices, R. E. Bank et al., eds., Birkhäuser, Basel, 1994, pp. 73–81.
[13] R. MÄRZ AND C. TISCHENDORF, *Solving more general index-2 differential algebraic equations*, Computers and Mathematics with Applications 28 (1994),pp. 77–105.
[14] L. R. PETZOLD, *Order results for implicit Runge-Kutta methods applied to differential algebraic systems*, SIAM J. Numer. Anal. 23 (1986), pp. 837–852.
[15] F. VERHULST, *Nonlinear Differential Equations and Dynamical Systems*, Springer, Berlin, 1990.
[16] J. WENSCH, R. WEINER AND K. STREHMEL, *Stability investigations for index-2-systems*, Reports on Computer Science and Scientific Computing 1, Universität Halle, Halle, 1994.