NUMERICAL ANALYSIS AND MATHEMATICAL MODELLING
BANACH CENTER PUBLICATIONS, VOLUME 24
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1990

HIGHER-INDEX DIFFERENTIAL-ALGEBRAIC EQUATIONS: ANALYSIS AND NUMERICAL TREATMENT

ROSWITHA MÄRZ

Section of Mathematics, Humboldt University, Berlin, German Democratic Republic

1. Introduction

We are interested in singular ordinary differential equations (ODE's)

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

where the partial Jacobian $f_y'(y, x, t)$ is everywhere singular but has constant rank. Moreover, we suppose the nullspace of $f_y'(y, x, t)$ to be independent of (y, x). Those uniformly singular ODE's are called *differential-algebraic equations* (DAE's). They originate from actual applications in different areas. In particular, dynamical systems subjected to constraints are described by DAE's (cf. [12]).

In the present paper, so-called index-k-tractable DAE's, $k \in \{2, 3\}$ ("higher index" DAE's), are considered. In Section 2, solvability statements for linear index-k-tractable DAE's are given. The linear map representing a linear initial value problem with appropriately formulated initial conditions is injective but has no bounded inverse in its natural setting.

Those problems are called *ill-posed*. In particular, even the constant-stepsize BDF applied to a linear constant-coefficient DAE of this kind becomes unstable (cf. [8]).

On the other hand, the BDF is reported to work well also for some higher index DAE's when using a careful special error control ([2], [6], [12], [19]). Up to now, for linear constant coefficient DAE's and for certain semi-explicit nonlinear index-2 or index-3 DAE's, the BDF is also proved to be convergent ([2], [5], [6], [12]). Clearly, this does not contradict the instability since the Theorem of Lax does not apply to ill-posed problems. Nevertheless, some confusion arose out of these facts.

In Section 3 and 5 the variable-order variable-step BDF applied to the quasilinear index-2-tractable DAE

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

is proved to have a weak instability only. Instead of a uniform stability bound S for stability, we now obtain a bound $\underline{h}^{-1} S$.

Thereby, $A(t)$ is assumed to have a constant nullspace, and a certain subspace is supposed to be smooth. Further, using detailed error estimations, the BDF is shown to be convergent for (1.2) and, moreover, to have the same order as in case of regular ODE's. Firstly, these facts are proved for linear DAE's (Section 3). Then, by commuting linearization and discretization, the mentioned results are obtained also for nonlinear DAE's.

The weak instability causes the BDF to be very sensitive with respect to the error control in the index-2 case. But, fortunately, implementations using an appropriate control of the defects in the nonlinear equations to be solved per step are reported to work reasonably ([6], [12]). However, the situation for index-3 DAE's becomes worse. At present, the best way of the practical treatment of an index-3 DAE seems to turn to an equivalent index-2 DAE by a so-called reduction step (which is an analytical technique) and then apply a BDF.

An alternative way to treat higher-index DAE's numerically appears in outlines by regularization methods and further special methods for solving ill-posed problems. Investigations in this concern are started recently. Section 4 informs on some interesting results.

2. Solvability of linear initial value problems

Let $\mathcal{N}$ denote the set of all ordered pairs $\{A, B\}$ of continuous matrix functions $A, B: [t_0, T] \to L(\boldsymbol{R}^m)$ the first of which has a smooth nullspace, i.e.

$$N(t) := \ker(A(t)) = \operatorname{span}\{n_1(t), \ldots, n_{m-r}(t)\}, \quad t \in [t_0, T], \tag{2.1}$$

$n_1, \ldots, n_{m-r} \in C^1([t_0, T], \boldsymbol{R}^m)$, where r denotes the constant rank of $A(t)$.

Write shortly $C := C([t_0, T], \boldsymbol{R}^m)$, $C^1 := C^1([t_0, T], \boldsymbol{R}^m)$.

Assuming $\{A, B\} \in \mathcal{N}$, we consider the linear equation

$$Ax' + Bx = q. \tag{2.2}$$

Using continuously differentiable projection functions $Q, P: [t_0, T] \to L(\boldsymbol{R}^m)$ so that

$$Q(t)^2 \equiv Q(t), \quad \operatorname{im}(Q(t)) \equiv N(t), \quad P(t) := I - Q(t), \tag{2.3}$$

we may reformulate (2.2) as

$$A((Px)' - P'x) + Bx = q \tag{2.4}$$

and further relate the map $\mathfrak{A}: C_N^1 \to C$,

$$C_N^1 := \{x \in C: Px \in C^1\}, \tag{2.5}$$

$$\mathfrak{A}x := A(Px)' + (B - AP')x, \quad x \in C_N^1, \tag{2.6}$$

to equation (2.2). The natural norm on C_N^1 is

$$\|x\| := \|x\|_\infty + \|(Px)'\|_\infty, \quad x \in C_N^1.$$

Note that $\{C_N^1, \|\ \|\}$ is a Banach space, and $\mathfrak{A}$ becomes bounded.

DEFINITION. $\mathfrak{A}$ is called *tractable* if

$$\dim(\ker(\mathfrak{A})) < \infty. \tag{2.7}$$

Clearly, tractability means that the solutions of the homogeneous equation $Ax' + Bx = 0$ form a finite-dimensional function space so that it becomes possible to select a unique solution by a finite number of initial conditions.

Up to now, the question whether $\ker(\mathfrak{A})$ has finite dimension is answered only partially. For constant coefficients A, B, we have (2.7) if and only if the matrix pencil $\lambda A + B$ is regular, i.e. $\det(\lambda A + B) \not\equiv 0$ ([1], cf. also [8]). Further, in the case of sufficiently smooth coefficients A, B, some results are obtained via reduction methods (cf. [7], [5]). (2.7) is also true for all DAE's (2.2) having a global index in the sense of Gear and Petzold ([5]). Finally, index-k-tractability, $k \in \{1, 2, 3\}$ implies (2.7) provided $\{A, B\} \in \mathcal{N}$ and some canonical subspace varies smoothly ([13], [14]).

It should be mentioned that pre-tractability ([16]) is necessary but not sufficient for tractability. For a conjecture how to characterize tractability in terms of the coefficients $\{A, B\}$ we refer to [15].

We follow the idea to rely on certain canonical subspaces (cf. [8], [13], [14]). For a given pair $\{A, B\} \in \mathcal{N}$, we are going to use the matrix functions

$$G_1 := A + BQ, \tag{2.8}$$

$$A_1 := G_1 - AP'Q, \tag{2.9}$$

$$G_2 := A_1 + BPQ_1, \tag{2.10}$$

$$A_2 := G_2 - A_1(PP_1)'PQ_1, \tag{2.11}$$

$$G_3 := A_2 + BPP_1Q_2, \tag{2.12}$$

as well as the subspaces

$$N(t) := \ker(A(t)), \tag{2.13}$$

$$S(t) := \{z \in \mathbf{R}^m : B(t)z \in \mathrm{im}(A(t))\}, \tag{2.14}$$

$$N_1(t) := \ker(A_1(t)), \tag{2.15}$$

$$S_1(t) := \{z \in \mathbf{R}^m : B(t)P(t)z \in \mathrm{im}(A_1(t))\}, \tag{2.16}$$

$$N_2(t) := \ker(A_2(t)), \tag{2.17}$$

$$S_2(t) := \{z \in \mathbf{R}^m : B(t)P(t)P_1(t)z \in \mathrm{im}(A_2(t))\}; \tag{2.18}$$

where $t \in [t_0, T]$. Thereby, $Q_1(t)$, $Q_2(t)$ denote projections onto $N_1(t)$ and $N_2(t)$, respectively, further

$$P_1(t) := I - Q_1(t), \quad P_2(t) := I - Q_2(t).$$

Using A_2, we always suppose PP_1 to be continuously differentiable. The subspaces (2.13) – (2.18) are called *canonical subspaces* of the DAE (2.2).

DEFINITION ([8], [13], [14]). Assume $\{A, B\} \in \mathcal{N}$. The DAE (2.2) is called
1. *transferable* (or *index-1-tractable*) if $G_1(t)$ remains nonsingular for all t,
2. *index-2-tractable* if $G_1(t)$ is singular but $G_2(t)$ nonsingular for all t,
3. *index-3-tractable* if both $G_1(t)$, $G_2(t)$ are singular but $G_3(t)$ remains nonsingular on $[t_0, T]$.

The point of index-k-tractability is to characterize special classes of tractable maps $\mathfrak{A}$ respectively tractable DAE's (2.2), and to generalize the class of DAE's having the global index k (cf. [8], [13], [14]). In the case of $k > 1$ we speak of *higher-index* equations.

Recall ([8], [13], [14]) that $G_i(t)$ and $A_i(t)$ are always singular or nonsingular simultaneously. This fact is useful for testing index-k-tractability in practice. Moreover, recall also that transferability means

$$N(t) \oplus S(t) = \boldsymbol{R}^m, \quad t \in [t_0, T], \tag{2.19}$$

while index-2-tractability is equivalent to

$$N(t) \cap S(t) \neq \{0\}, \quad N_1(t) \oplus S_1(t) = \boldsymbol{R}^m, \quad t \in [t_0, T]. \tag{2.20}$$

Finally, index-3-tractability is equivalent to the subspace-properties

$$N(t) \cap S(t) \neq \{0\}, \quad N_1(t) \cap S_1(t) \neq \{0\}, \quad N_2(t) \oplus S_2(t) = \boldsymbol{R}^m, \quad t \in [t_0, T]. \tag{2.21}$$

It is well known that the formulation of appropriate initial conditions for index-k-tractable DAE's depends on the DAE itself. Here, we use the conditions

$$\mathcal{L}x := \left\{ \begin{array}{ll} P(t_0)x(t_0) & \text{for } k = 1, \\ P(t_0)P_1(t_0)x(t_0) & \text{for } k = 2, \\ P(t_0)P_1(t_0)P_2(t_0)x(t_0) & \text{for } k = 3 \end{array} \right\} = b. \tag{2.22}$$

Denote, for index-k-tractable DAE's,

$$M := \left\{ \begin{array}{ll} \operatorname{im}(P(t_0)) & \text{if } k = 1, \\ \operatorname{im}(P(t_0)P_1(t_0)) & \text{if } k = 2, \\ \operatorname{im}(P(t_0)P_1(t_0)P_2(t_0)) & \text{if } k = 3. \end{array} \right. \tag{2.23}$$

Clearly, the linear map

$$\mathcal{L} := (\mathfrak{A}, \mathcal{L}): C_N^1 \to C \times M \tag{2.24}$$

is bounded. The equation $\mathscr{L}x = (q, b)$ represents the initial value problem (2.2), (2.22). For the numerical treatment of such an IVP, it should be known whether this problem is well-posed in Hadamard's sense, i.e. whether $\mathscr{L}$ has a bounded inverse. Unfortunately, $\mathscr{L}$ is a homeomorphism only in the case of transferable (= index-1-tractable) DAE's. This is why we should differentiate the transferable DAE's from the nontransferable ones basically.

THEOREM 2.1. *Suppose that* (2.2) *is transferable, the IVP* (2.2), (2.22) *is uniquely solvable on* C_N^1 *for all* $q \in C$, $b \in M$.

Then, it holds that $\operatorname{im}(\mathfrak{A}) = C$, $\dim(\ker(\mathfrak{A})) = r$, *and* $\mathscr{L}$ *becomes a homeomorphism.*

([2], Theorem 1.2.20).

THEOREM 2.2. *Let* (2.2) *be index-2-tractable and* $Q_1(t)$ *be chosen to project onto* $N_1(t)$ *along* $S_1(t)$. *Assume* $Q_1 \in C^1([t_0, T], \mathbf{R}^m)$, *additionally.*

Then, the IVP (2.2), (2.22) *is uniquely solvable on* C_N^1 *for all* $q \in \mathfrak{M}$, $b \in M$, *where*

$$\mathfrak{M} := \{p \in C: Q_1 G_2^{-1} p \in C^1\}. \tag{2.25}$$

It holds that $\operatorname{im}(\mathfrak{A}) = \mathfrak{M}$,

$$\dim(\ker(\mathfrak{A})) = \operatorname{rank}(P(t_0) P_1(t_0)) < r.$$

$\mathfrak{M}$ *is a proper nonclosed subset within* C, *and* $\mathscr{L}$ *has no bounded inverse.*

([13], Theorem 2.4).

EXAMPLE. For the semi-explicit linear index-2 DAE

$$\begin{bmatrix} I & \\ & 0 \end{bmatrix} x' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix} x = q$$

we compute

$$Q = \begin{bmatrix} 0 & \\ & I \end{bmatrix}, \quad Q_1 = \begin{bmatrix} I & B_{12} \\ 0 & 0 \end{bmatrix},$$

$$Q_1 = \begin{bmatrix} B_{12}(B_{21} B_{12})^{-1} B_{21} & 0 \\ -(B_{21} B_{12})^{-1} & B_{21} \quad 0 \end{bmatrix}.$$

Note that the projector function $B_{12}(B_{21} B_{12})^{-1} B_{21}$ is assumed to be smooth in [6], [12] also. ∎

Recall from [13] that, under the assumptions of Theorem 2.2, the IVP (2.2), (2.22) is equivalent to the form

$$\begin{aligned} x &= (I - QP_1 G_2^{-1} B - (QQ_1)')y + PQ_1 G_2^{-1} q + QP_1 G_2^{-1} q \\ &\quad -(QQ_1)' PQ_1 G_2^{-1} q + (QQ_1 G_2^{-1} q)', \\ y' &= (PP_1)' y - PP_1 G_2^{-1} By + (PP_1)' PQ_1 G_2^{-1} q + PP_1 G_2^{-1} q, \\ y(t_0) &= b. \end{aligned} \tag{2.26}$$

THEOREM 2.3. *Suppose that* (2.2) *is index-3-tractable, and moreover, that* $Q_1(t)$, $Q_2(t)$ *are chosen in such a way that*

$$Q_1(t)\,Q(t) = 0 \tag{2.27}$$

holds and $Q_2(t)$ *projects onto* $N_2(t)$ *along* $S_2(t)$. *Additionally, assume*

$$Q_1, Q_2, \tilde{Q}_1 \in C^1([t_0, T], L(\boldsymbol{R}^m)),$$

$$Q_1 Q_2 \in C^2([t_0, T], L(\boldsymbol{R}^m)),$$

where $Q_1(t)$ *denotes the projection onto* $N_1(t)$ *along*

$$\{z \in \boldsymbol{R}^m: (B(t) - A_1(t)(PP_1)'(t))\,P(t)z \in \operatorname{im}(G_3(t)P_1(t))\}.$$

Then, the IVP (2.2), (2.22) *is uniquely solvable on* C_N^1 *for all* $q \in \mathfrak{M}$, $b \in M$, *where*

$$\mathfrak{M} := \{p \in C: Q_2 G_3^{-1} p \in C^1, \tag{2.28}$$

$$Q_1 P_2 G_3^{-1} p + Q_1 Q_2 (PP_1 Q_2 G_3^{-1} q)' \in C^1\}.$$

It holds that $\operatorname{im}(\mathfrak{A}) = \mathfrak{M}$

$$\dim(\ker(\mathfrak{A})) = \operatorname{rank}(P(t_0)P_1(t_0)P_2(t_0)) < r.$$

$\mathfrak{M}$ *is a proper nonclosed subset within* C, *and* $\mathscr{L}$ *has no bounded inverse.*

([14], Theorem 2.4).

In both cases of Theorems 2.2 and 2.3, the range $\operatorname{im}(\mathscr{L}) = \mathfrak{M} \times M$ is nonclosed within $C \times M$. Further, $\mathscr{L}$ is injective, but its inverse $\mathscr{L}^{-1}$ acting from $C \times M$ onto C_N^1 is no more continuous. Those problems in which the continuous dependence of the solutions on the right-hand sides is missing are called *ill-posed* or *essentially ill-posed* in Tikhonov's sense.

Surely, we could turn to maps $\mathscr{L}: C_N^1 \to \mathfrak{M} \times M$ and use an appropriate stronger norm $\|\cdot\|_{\mathfrak{M}}$ on $\mathfrak{M}$ to have a Banach space $\{\mathfrak{M}, \|\cdot\|_{\mathfrak{M}}\}$ and a homeomorphism $\mathscr{L}$ as well (cf. [17]). However, this approach seems to make no sense in view of the numerical treatment we are interested in, and also in view of possible linearizations of nonlinear problems.

It should also be mentioned, that $\mathscr{L}$ itself becomes unbounded in the setting $\mathscr{L}: C \to C \times M$, while $\mathscr{L}^{-1}: C \times M \to C$ also remains unbounded. This is easily proved considering (2.26) and the related decoupling for the index 3 case in [14].

3. On the BDF applied to linear index-2-tractable DAE's

The first question we should deal with is about feasibility. A variable-order variable-stepsize BDF applied to (2.2) is simply

$$A(t_j)\frac{1}{h_j}\sum_{i=0}^{k} a_{ij} x_{j-i} + B(t_j)x_j = q(t_j), \tag{3.1}$$

thus, per integration step, a linear system with the coefficient matrix

$$F_j := A(t_j)\frac{a_{j_0}}{h_j} + B(t_j) \tag{3.2}$$

has to be solved.

LEMMA 3.1. *If* (2.2) *is index-2-tractable and*

$$I - Q(t)Q_1(t)P'(t) \tag{3.3}$$

remains nonsingular for all t, *where* $Q_1(t)$ *denotes the canonical projection onto* $N_1(t)$ *along* $S_1(t)$, *then* F_j *is nonsingular for sufficiently small* h_j.

Proof. Write shortly $F = A/\tau + B$, $\tau := h_j/a_{j_0}$, and drop the arguments t_j as well as the index j. The equation $Fz = y$ is equivalent to

$$(A + BQ - AP'Q)\left(\frac{1}{\tau}Pz + Qz\right) + BPz = y - AP'Qz,$$

further to

$$(A_1 + BPQ_1)\left(\frac{1}{\tau}P_1Pz + P_1Qz + Q_1z\right) + BPP_1z = y - AP'Qz$$

or

$$\frac{1}{\tau}P_1Pz + P_1Qz + Q_1z + G_2^{-1}BPP_1z = G_2^{-1}y - G_2^{-1}AP'Qz. \tag{3.4}$$

Now we may use the identity $Q_1 = Q_1G_2^{-1}BP$ (cf. [13]) and, moreover, the relations (cf. [5])

$$G_2^{-1}AP'Q = G_2^{-1}A_1PP'Q = P_1PP'Q, \qquad PP_1P = PP_1,$$
$$PP_1Q = 0, \qquad QP_1P = -QQ_1, \qquad QP_1Q = Q, \qquad QP_1PP'Q = -QQ_1P'Q.$$

Multiplying (3.4) by τPP_1, QP_1 and Q_1, we obtain the system

$$PP_1z + \tau PP_1G_2^{-1}BPP_1z + \tau PP_1P'Qz = \tau PP_1G_2^{-1}y,$$
$$-\frac{1}{\tau}QQ_1z + Qz + QP_1G_2^{-1}BPP_1z - QQ_1P'Qz = QP_1G_2^{-1}y,$$
$$Q_1z = Q_1G_2^{-1}y.$$

Hence, we have

$$\begin{aligned}(I + \tau PP_1G_2^{-1}B)PP_1z + \tau PP_1P'Qz &= \tau PP_1G_2^{-1}y,\\ QP_1G_2^{-1}BPP_1z + (I - QQ_1P')Qz &= QP_1G_2^{-1}y + \frac{1}{\tau}QQ_1G_2^{-1}y\end{aligned} \tag{3.5}$$

to determine the components PP_1z, Qz. For small $\tau > 0$, the matrix $H := I + \tau PP_1G_2^{-1}B$ is nonsingular, and also the Kronecker complement in (3.5)

$$K := I - QQ_1P' - \tau QP_1G_2^{-1}BH^{-1}PP_1P'$$

is so. Finally, solving (3.5) with respect to $PP_1 z$, Qz, we obtain

$$(3.6)\quad z = \{(I-\tau H^{-1} PP_1 P')K^{-1}\left(QP_1+\frac{1}{\tau}QQ_1-\tau QP_1 G_2^{-1} BH^{-1} PP_1\right) + PQ_1+\tau H^{-1} PP_1\} G_2^{-1} y. \blacksquare$$

EXAMPLE ([5]).

$$A(t)=\begin{bmatrix}0 & 0\\ 1 & \eta t\end{bmatrix},\quad B(t)=\begin{bmatrix}1 & \eta t\\ 0 & 1+\eta\end{bmatrix},\quad \eta\in\boldsymbol{R}.$$

The related DAE (2.2) is index-2-tractable. We compute

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

$$Q_1(t)=\begin{bmatrix}1+\eta t & \eta t(1+\eta t)\\ -1 & -\eta t\end{bmatrix},\quad G_2(t)=\begin{bmatrix}1 & \eta t\\ 1 & 1+\eta t\end{bmatrix},$$

$$P(t)P_1(t)=0,\quad I-Q(t)Q_1(t)P'(t)=\begin{bmatrix}1 & -\eta^2 t\\ 0 & 1+\eta\end{bmatrix}.$$

Thus the matrix (3.3) is nonsingular for $\eta\neq -1$ but singular otherwise. Moreover, it is easy to check that F_j becomes singular for $\eta=-1$. $\blacksquare$

Note that (3.6) gives an explicit expression for F_j^{-1}. In particular, in the constant-nullspace case we have $P'=0$, $K=I$,

$$(3.8)\quad F_j^{-1}=\left\{QP_1+PQ_1+\frac{a_{j0}}{h_j}QQ_1+\frac{h_j}{a_{j0}}(I-QP_1 G_2^{-1} B)H^{-1} PP_1\right\}G_2^{-1}\Big|_{t_j}.$$

The BDF's are well known to become unstable when they are applied to higher-index DAE's. This fact is true even in the case of constant-coefficient DAE's and constant-stepsize BDF's (cf. [8], [16]).

In the following we use grids

$$(3.9)\qquad \pi: t_0 < t_1 < \ldots < t^n = T$$

belonging to a given gridclass Π. The smallest gridclass we are interested in is the set Π_{equ} of all equidistant grids (3.9). Assume always

$$\Pi_{\text{equ}} \subseteq \Pi.$$

Denote by h and $\underline{h}$ the maximal and minimal stepsize of (3.9), respectively.

Introduce the linear map $\mathscr{L}_\pi$: $\boldsymbol{R}^{m(n+1)} \to \boldsymbol{R}^{m(n+1)}$,

$$\mathscr{L}_\pi = \begin{bmatrix} I & & & \\ \vdots & & I & \\ \frac{a_{kk}}{h_k} A_k & \cdots & \frac{a_{k1}}{h_k} A_k & F_k \\ & \ddots & & \ddots \\ & & \frac{a_{nk}}{h_n} A_n & \cdots \frac{a_{n1}}{h_n} A_n \quad F_n \end{bmatrix}. \tag{3.10}$$

Clearly, $\mathscr{L}_\pi$ is related to the BDF (3.1). $\mathscr{L}_\pi$ is nonsingular if the assumptions of Lemma 3.1 are satisfied. However, there is no uniform bound for $\|\mathscr{L}_\pi^{-1}\|_\infty$ for all $\pi \in \Pi$, i.e. the BDF is unstable. In particular, the nontrivial term $\frac{1}{h_j} Q(t_j) Q_1(t_j)$ of F_j^{-1} (cf. (3.6) resp. (3.8)) grows unboundedly if h_j tends to zero. The instability of the BDF is a reflection of the unboundedness of the map

$$\mathscr{L}^{-1}: C \times M \to C.$$

Nevertheless, for some special classes of index-2-tractable DAE's the convergence of the BDF is proved ([2], [6], [12]), and the BDF is reported to work well, when a special error control is realized (e.g. [19]).

To make both these facts more transparent we show in the following that in those special cases the instability is only weak. More precisely, we show that there is a bound S so that the inequality

$$\|\mathscr{L}_\pi^{-1}\|_\infty \leqslant \underline{h}^{-1} S, \quad \pi \in \Pi(\eta_1, \eta_2, h_{\max}) \tag{3.11}$$

is true, where $\Pi(\eta_1, \eta_2, h_{\max})$ denotes the set of all grids (3.9) with

$$\eta_1 \leqslant \frac{h_j}{h_{j-1}} \leqslant \eta_2, \quad h \leqslant h_{\max}. \tag{3.12}$$

Let us agree to use only $k \leqslant 6$ and $\Pi(\eta_1, \eta_2, h_{\max})$ such that the BDF applied to an explicit ODE is stable.

Now, take an arbitrary grid $\pi \in \Pi(\eta_1, \eta_2, h_{\max})$, arbitrary $w_j \in \boldsymbol{R}^m$, $j = 0, \ldots, n$,

$$A_j \frac{1}{h_j} \sum_{i=1}^{k} a_{ij} z_{j-i} + B_j z_j = w_j, \tag{3.13}$$

$$j = k, \ldots, n, \; z_j = w_j, \; j = 0, \ldots, k-1,$$

that is we consider the equation $\mathscr{L}_\pi z = w$. Assume $P' = 0$ and decouple (3.13) in the same way as we proved Lemma 3.1. This yields

$$P_1(t_j) P \frac{1}{h_j} \sum_{i=0}^{k} a_{ij} z_{j-i} + P_1(t_j) Q z_j + P_1(t_j) G_2(t_j)^{-1} B_j P P_1(t_j) z_j = P_1(t_j) G_2(t_j)^{-1} w_j, \tag{3.14}$$

$$Q_1(t_j) z_j = Q_1(t_j) G_2(t_j)^{-1} w_j \tag{3.15}$$

Multiplying (3.14) by P and Q, respectively, we derive

$$\frac{1}{h_j} PP_1(t_j) \sum_{i=0}^{k} a_{ij} z_{j-i} + PP_1(t_j) G_2(t_j)^{-1} B_j PP_1(t_j) z_j = PP_1(t_j) G_2(t_j)^{-1} w_j, \tag{3.16}$$

$$-\frac{1}{h_j} QQ_1(t_j) \sum_{i=0}^{k} a_{ij} z_{j-i} + Qz_j + QP_1(t_j) G_2(t_j)^{-1} B_j PP_1(t_j) z_j = QP_1(t_j) G_2(t_j)^{-1} w_j, \tag{3.17}$$

Inserting the relation

$$\begin{aligned} &\frac{1}{h_j} PP_1(t_j) \sum_{i=0}^{k} a_{ji} z_{j-i} \\ &\quad = \frac{1}{h_j} \sum_{i=0}^{k} a_{ji} PP_1(t_{j-i}) z_{j-i} \\ &\qquad + \sum_{i=0}^{k} a_{ji} P \frac{1}{h_j} \big(P_1(t_j) - P_1(t_{j-i})\big)\big(PP_1(t_{j-i}) z_{j-i} + PQ_1(t_{j-i}) z_{j-i}\big) \end{aligned}$$

into (3.16) and using (3.15) we verify the inequalities

$$\max_{0 \leqslant j \leqslant n} |PP_1(t_j) z_j| \leqslant S_1 \max_{0 \leqslant j \leqslant n} |w_j|,$$

$$\max_{0 \leqslant j \leqslant n} |Q_1(t_j) z_j| \leqslant S_2 \max_{0 \leqslant j \leqslant n} |w_j|.$$

Hence, we have also

$$\max_{0 \leqslant j \leqslant n} |Pz_j| \leqslant S_3 \max_{0 \leqslant j \leqslant n} |w_j|. \tag{3.18}$$

On the other hand, (3.17) yields, for $j = k, \ldots, n$,

$$\begin{aligned} Qz_j &= \frac{1}{h_j} QQ_1(t_j) \sum_{i=0}^{k} a_{ji} \{PP_1(t_{j-i}) + PQ_1(t_{j-i})\} z_{j-i} \\ &\quad - QP_1(t_j) G_2(t_j)^{-1} B_j PP_1(t_j) z_j + QP_1(t_j) G_2(t_j)^{-1} w_j \\ &= \frac{1}{h_j} QQ_1(t_j) \sum_{i=0}^{k} a_{ji} Q_1(t_{j-i}) \tilde{w}_{j-i} \\ &\quad + \sum_{i=0}^{k} a_{ji} Q \frac{1}{h_j} \big(Q_1(t_j) - Q_1(t_{j-i})\big) PP_1(t_{j-i}) z_{j-i} \\ &\quad - QP_1(t_j) G_2(t_j)^{-1} B_j PP_1(t_j) z_j + QP_1(t_j) \tilde{w}_j, \end{aligned} \tag{3.19}$$

where

$$\tilde{w}_j := \begin{cases} w_j & \text{for } j = 0, \ldots, k-1, \\ G_2(t_j)^{-1} w_j & \text{for } j = k, \ldots, n. \end{cases}$$

Now, we have

$$|Qz_j| \leqslant \frac{1}{h_j}\Big|QQ_1(t_j)\sum_{i=0}^{k} a_{ji}\,Q_1(t_{j-i})\,\tilde{w}_{j-i}\Big| + S_4 \max_{0\leqslant l\leqslant n} |w_l|. \tag{3.20}$$

(3.18), (3.20) imply (3.11). Moreover, for $z_j = x(t_j) - x_j$, $j = 0, \ldots, n$, where $x(t_j)$ denotes the value of the exact solution at t_j, and x_j is the exact numerical solution generated by the BDF using given starting values $x_0, \ldots, x_{k-1}$, the corresponding values w_j are

$$w_j = x(t_j) - x_j \quad \text{for } j = 0, \ldots, k-1,$$

$$\begin{aligned} w_j = \tau_j &:= \frac{1}{h_j} A_j \sum_{i=0}^{k} a_{ji}\,x(t_{j-i}) + B_j x(t_j) - q(t_j) \\ &= \frac{1}{h_j} A_j \sum_{i=0}^{k} a_{ji}\,x(t_{j-i}) - A_j (Px)'(t_j), \end{aligned} \tag{3.21}$$

for $j = k, \ldots, n$. Since $\tau_j \in \operatorname{im}(A_j)$ we may write

$$w_j = \tau_j = A_j A_j^{+} \tau_j, \quad j = k, \ldots, n,$$

further

$$\begin{aligned} \tilde{w}_j &= G_2(t_j)^{-1} A_j A_j^{+} \tau_j \\ &= G_2(t_j)^{-1} A_1(t_j) P A_j^{+} \tau_j = P_1(t_j) P A_j^{+} \tau_j, \quad j = k, \ldots, n \end{aligned} \tag{3.22}$$

Surely, for exact starting values $x_j = x(t_j)$, $j = 0, \ldots, k-1$, (3.20) yields simply

$$|Q(x(t_j) - x_j)| \leqslant S_4 \max_{k\leqslant l\leqslant n} |\tau_l|, \quad j \geqslant k,$$

further

$$\max_{0\leqslant j\leqslant n} |x(t_j) - x_j| \leqslant S_5 \max_{k\leqslant l\leqslant n} |\tau_l|. \tag{3.23}$$

Finally, we collect our results in

THEOREM 3.2. *Let the assumptions of Theorem* 2.2 *be satisfied and, further,* $P' = 0$. *Let the variable-order variable-stepsize BDF* (3.1) *be stable on* $\Pi(\eta_1, \eta_2, h_{\max})$ *for explicit ODE's. Then this BDF applied to the DAE* (2.2) *is weakly unstable and satisfies* (3.11). *Moreover, this BDF is convergent and has the same order as for explicit ODE's.*

Theorem 3.2 generalizes the related convergence results in [5], [6], [12] obtained for semi-explicit DAE's. For the values $\tilde{x}_k, \ldots, \tilde{x}_n \in \boldsymbol{R}^m$ generated by the perturbed equation

$$F_j \tilde{x}_j = q(t_j) - \frac{1}{h_j} A_j \sum_{i=1}^{k} a_{ji}\,\tilde{x}_{j-i} + \delta_j, \quad j = k, \ldots, n, \tag{3.24}$$

using starting values $\tilde{x}_0, \ldots, \tilde{x}_{k-1} \in \boldsymbol{R}^m$, we obtain the error estimation

$$\max_{0 \leqslant j \leqslant n} |x(t_j) - \tilde{x}_j| \leqslant S_6 \max\Big\{ \max_{0 \leqslant j \leqslant k-1} |x(t_j) - \tilde{x}_j|, \max_{k \leqslant j \leqslant n} |\tau_j - \delta_j| \Big\} + \max_{k \leqslant j \leqslant n} \frac{1}{h_j} \Big| QQ_1(t_j) \sum_{i=0}^{k} a_{ji} Q_1(t_{j-i}) \tilde{\delta}_{j-i} \Big|, \tag{3.25}$$

where

$$\tilde{\delta}_j := \begin{cases} -G_2(t_j)^{-1} \delta_j & \text{for } j = k, \ldots, n, \\ x(t_j) - \tilde{x}_j & \text{for } j = 0, \ldots, k-1, \end{cases}$$

and δ_j represents defects in the linear equations due to roundoff errors. Thus, the components $Q_1(t_j) \tilde{\delta}_j / h_j$ must be kept small by an appropriate error control.

From (3.10) we know immediately that

$$\|\mathscr{L}_{\Pi}\|_{\infty} \leqslant \underline{h}^{-1} K. \tag{3.26}$$

This yields

$$\|\mathscr{L}_{\Pi}\|_{\infty} \|\mathscr{L}_{\Pi}^{-1}\|_{\infty} \leqslant \underline{h}^{-2} KS. \tag{3.27}$$

Similarly, the matrices F_j of the linear systems to be solved per integration step have the condition numbers

$$\|F_j\| \, \|F_j^{-1}\| \sim h_j^{-2}$$

since QQ_1 is always nontrivial for index-2-tractable DAE's.

In comparison with this, we recall that

$$\|\mathscr{L}_{\Pi}\|_{\infty} \leqslant \underline{h}^{-1} K_0, \quad \|\mathscr{L}_{\Pi}^{-1}\|_{\infty} \leqslant S_0,$$
$$\|F_j\| \, \|F_j^{-1}\| \leqslant \text{const}$$

is true for explicit ODE's. In the case of transferable DAE's we have (cf. [8]) also

$$\|\mathscr{L}_{\Pi}\|_{\infty} \leqslant \underline{h}^{-1} K_1, \quad \|\mathscr{L}_{\Pi}^{-1}\|_{\infty} \leqslant S_1,$$

but $\|F_j\| \, \|F_j^{-1}\| \sim h_j^{-1}$.

Thus, it is not surprising that integration methods applied to index-2-DAE's depend more sensitively on the error control. However, using a careful error control, integration methods may work well (cf. [6], [12]) also for a class of index-2-DAE's.

However, for index-3-DAE's the situation becomes worse. At the best we may expect

$$\|\mathscr{L}_{\Pi}^{-1}\|_{\infty} \leqslant \underline{h}^{-2} K \quad \text{and} \quad \|F_j\| \, \|F_j^{-1}\| \sim h_j^{-3}.$$

A detailed analysis of BDF's applied to index-3-DAE's is in preparation.

4. Some remarks on regularization methods

Since the IVP (2.2), (2.22) is essentially ill-posed in Tikhonov's sense, if (2.2) is index-2-tractable, numerical methods for ill-posed problems are of interest also for those nontransferable DAE's. We are going to deal with two regularizations which are very closely connected with singular pertubations.

The linear equation (2.2) is approximated by

$$(A+\varepsilon B)x_\varepsilon'+Bx_\varepsilon = q \tag{4.1}$$

using the so-called *pencil-regularization* (e.g. [1], [3]). An alternative way is the use of the regularization

$$(A+\varepsilon BP)x_\varepsilon'+(B+\varepsilon BPP')x_\varepsilon = q \tag{4.2}$$

proposed in [18].

With (4.1) one aims for a regular ODE. This goal is not reached in general. For instance, for the special DAE given in (3.7) with $\eta = -1$, the regularized equation (4.1) is again an index-2-tractable DAE. On the other hand, by (4.1), transferable DAE's are approximated by regular stiff ODE's, i.e. (4.1) yields "parasitic" boundary layers (cf. [11]). This is why (4.2) is proposed in [18].

If $P' = 0$, both (4.1) and (4.2) may be applied. The initial condition for index-2-tractable DAE's is (cf. (2.22))

$$P(t_0)P_1(t_0)x(t_0) = b. \tag{4.3}$$

For (4.2), we may use the additional condition (cf. (2.26))

$$P(t_0)Q_1(t_0)x(t_0) = P(t_0)Q_1(t_0)G_2(t_0)^{-1}q(t_0). \tag{4.4}$$

Additionally, when using (4.1), we have to determine $Q(t_0)x(t_0)$ besides (4.3), (4.4). However, the formula given by (2.26), i.e.

$$\begin{aligned} Q(t_0)x(t_0) = {} & Q(t_0)P_1(t_0)G_2(t_0)^{-1}\{q(t_0)-B(t_0)b\}+(QQ_1G_2^{-1}q)'(t_0) \\ & -(QQ_1)'(t_0)\{b+P(t_0)Q_1(t_0)G_2(t_0)^{-1}q(t_0)\}, \end{aligned} \tag{4.5}$$

is not reasonable for the practical use. Thus, we must compute an approximation of the consistent initial value for (4.1).

THEOREM 4.1. *Let the assumptions of Theorem* 2.2 *be satisfied,* $q\in\mathfrak{M}$, $b\in M$.

Then (4.2) *is transferable for sufficiently small* $\varepsilon > 0$.

Let x, $x_\varepsilon\in C_N^1$ *denote the solutions of* (2.2), (4.3) *and* (4.2), (4.3), (4.4), *respectively.*

Then, $\|x_\varepsilon - x\|_{H_N^1}\to 0$ $(\varepsilon\to 0)$, *further,* $\|Px_\varepsilon - Px\|_{L_2} = O(\varepsilon)$.

Moreover, if PQ_1, $PP_1\in C^2$, $PQ_1G_2^{-1}q\in C^2$, *then*

$$\|x_\varepsilon - x\|_{H_N^1} = O(\varepsilon^{1/2}).$$

The proof is given in [10] for a certain special case. The general case is considered in [9].

EXAMPLE. The differentation problem $x_2' - x_1 = 0$, $x_2 = q_2$ is approximated by (cf. (4.1))

$$\begin{aligned} -\varepsilon x_1' + x_2' - x_1 &= 0, \\ \varepsilon x_2' + x_2 &= q_2, \\ x_2(t_0) = q_2(t_0), \quad x_1(t_0) &\approx q_2'(t_0), \end{aligned} \tag{4.6}$$

and by (cf. (4.2))

$$\begin{aligned} x_2' - x_1 &= 0, \\ \varepsilon x_2' + x_2 &= q_2, \\ x_2(t_0) &= q_2(t_0), \end{aligned} \tag{4.7}$$

respectively.

It should be mentioned that further assertions concerning the convergence of (4.1) and (4.2), respectively, are given in [1], [3], [9], [11].

5. Nonlinear DAE's

Let us start this section formulating some smoothness conditions for the general nonlinear DAE

$$f(x'(t), x(t), t) = 0. \tag{5.1}$$

ASSUMPTION (A): $f \in C(\mathcal{G}, \boldsymbol{R}^m)$, $\mathcal{G} := \boldsymbol{R}^m \times \boldsymbol{R}^m \times [t_0, T]$. Let $f_y'(y, x, t)$, $f_x'(y, x, t)$ exist for all $(y, x, t) \in \mathcal{G}$, and let $f_y', f_x' \in C(\mathcal{G}, L(\boldsymbol{R}^m))$. Let the nullspace of $f_y'(y, x, t)$ be independent of (y, x). Denote

$$N(t) := \equiv \ker(f_y'(y, x, t)). \tag{5.2}$$

Let $N(\cdot)$ be smooth, and denote by Q, P the corresponding projection functions onto $N(\cdot)$ and its complement. ∎

The nonlinear map $\mathfrak{A}: C_N^1 \to C$,

$$(\mathfrak{A}x)(t) := \equiv f((Px)'(t) - P'(t)x(t), x(t), t), \tag{5.3}$$

is defined on the whole space C_N^1, and is Fréchet-differentiable there. For each given $x_* \in C_N^1$, the linearized map represents a linear DAE with the coefficients

$$\{f_y'(M_*(\cdot)), f_x'(M_*(\cdot))\} \in \mathcal{M}, \tag{5.4}$$

where

$$M_*(t) := ((Px_*)'(t) - P'(t)x_*(t), x_*(t), t).$$

DEFINITION. Let (A) be satisfied, $k \in \{2, 3\}$, $x_* \in C_N^1$. Then (5.1) is called *transferable around* x_* and *index-k-tractable at* x_*, if the linear DAE having the coefficients (5.4) is transferable and index-k-tractable, respectively.

To test whether index-k-tractability is given say on the ball $\{x_* \in C_N^1 : \|x_* - x_0\| \leqslant \varrho\}$, $x_0 \in C_N^1$ fixed, we may use the following matrices (cf. [13], [14]):

$$G_1(y, x, t) := f_y'(y, x, t) + f_x'(y, x, t) Q(t), \tag{5.5}$$

$$A_1(y, x, t) := G_1(y, x, t) - f_y'(y, x, t) P'(t) Q(t), \tag{5.6}$$

$$G_2(y, x, t) := A_1(y, x, t) + f_x'(y, x, t) P(t) Q_1(x, t), \tag{5.7}$$

$$A_2(y, x, t) := G_2(y, x, t) - A_1(y, x, t) K(y, x, t) P(t) Q_1(x, t); \tag{5.8}$$

$$G_3(y, x, t) := A_2(y, x, t) + f_x'(y, x, t) P(t) P_1(x, t) Q_2(y, x, t). \tag{5.9}$$

Moreover, the nullspace of $A_1(y, x, t)$ is assumed to be independent of y and $Q(t)x$. Denote

$$\ker\big(A_1(y, x, t)\big) \equiv : N_1(x, t) \equiv N_1\big(P(t)x, t\big). \tag{5.10}$$

$Q_1(x, t)$ is defined to be a projection onto $N_1(x, t)$, further

$$P_1(x, t) := I - Q_1(x, t).$$

Using $A_2(y, x, t)$, we assume additionally that

$$\Pi(x, t) := P(t) P_1(x, t) = P(t) P_1\big(P(t)x, t\big)$$

depends continuously differentiable on (x, t). Finally,

$$K(y, x, t) := \Pi_x'(x, t)\big(y + P'(t)x\big) + \Pi_t'(x, t). \tag{5.11}$$

For index-2-tractability, we have to check the singularity of $G_1(y, x, t)$ as well as the nonsingularity of $G_2(y, x, t)$ on the related neighbourhood of the trajectory of x_0 within $\mathscr{G}$. For index-3-tractability, both $G_1(y, x, t)$, $G_2(y, x, t)$ have to be singular but $G_3(y, x, t)$ has to become nonsingular (cf. [13] Theorem 3.1, [14], Theorem 4.1).

Note that semi-explicit DAE's having the global index $k \in \{2, 3\}$ (cf. [5], [12]) are shown to be index-k-tractable ([13], [14]).

Now, let us turn to the semi-explicit system

$$u'(t) - v(t) = 0, \tag{5.12}$$

$$v'(t) + g\big(u(t), v(t), t\big) + h_u'\big(u(t), t\big)^T w(t) = 0, \tag{5.13}$$

$$h\big(u(t), t\big) = 0, \tag{5.14}$$

which describes, e.g., mechanical motions subjected to constraints (cf. [6], [12]). Assume that $h_u'(u, t) =: H(u, t)$ has full rank, i.e. the constraints (5.14) are linearly independent, $\ker\big(H(u, t)^T\big) = \{0\}$. Putting (5.12)–(5.14) in the general form (5.1), we obtain, with $x = (u, v, w)$.

$$f_y'(y, x, t) = \begin{bmatrix} I & & \\ & I & \\ & & 0 \end{bmatrix}, \quad f_x'(y, x, t) = \begin{bmatrix} 0 & -I & 0 \\ g_u' + h_{uu}''^T w & g_v' & H^T \\ H & 0 & 0 \end{bmatrix},$$

$$A_1(y, x, t) = \begin{bmatrix} I & & \\ & I & H(u, t)^T \\ & & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & & \\ & 0 & \\ & & I \end{bmatrix},$$

Clearly, the nullspace of $A_1(y, x, t)$ is independent of y and $Qx = w$. Compute

$$Q_1 = \begin{bmatrix} 0 & & \\ & H^T(HH^T)^{-1}H & \\ & -(HH^T)^{-1}H & 0 \end{bmatrix}, \quad PP_1 = \begin{bmatrix} I & & \\ & I - H^T(HH^T)^{-1}H & \\ & & 0 \end{bmatrix}$$

Now, the assumption for $\Pi := PP_1$ to be smooth appears to be a smoothness condition for the nullspace of $H(u, t)$. Since $H(u, t)$ has constant rank, a sufficient condition for this nullspace to be smooth is that $H(u, t)$ itself is continuously differentiable. Then, also P_1 becomes continuously differentiable.

Further we compute $Q_1 Q = 0$,

$$G_2 = \begin{bmatrix} I & -T & \\ & I + g_v' T & H^T \\ & & 0 \end{bmatrix},$$

$$T := H^T(HH^T)^{-1}H, \quad S := I - T,$$

$$A_2 = \begin{bmatrix} I & & -T & \\ & I + (g_v' + T_u'\beta + T_t')T & & H^T \\ & & & 0 \end{bmatrix},$$

$$f_x' PP_1 = \begin{bmatrix} 0 & -I & 0 \\ g_u' + h_{uu}''^T w & g_v' & 0 \\ H & 0 & 0 \end{bmatrix}$$

Finally,

$$A_2(y, x, t)z = 0, \quad f_x'(y, x, t)P(t)P_1(x, t)z \in \operatorname{im}(A_2(y, x, t)), \quad z = \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

imply

$$a = Tb, \quad b + (\ldots)Tb + H^T c = 0, \quad Ha = 0,$$

thus $a = Ta$, $b = Tb$, further $a = 0$, $b = 0$, $H^T c = 0$, hence $c = 0$. In consequence, (5.12)–(5.14) is index-3-tractable everywhere.

Following the idea of [6], Theorem 1.1, where systems (5.12)–(5.14) with autonomous constraints are considered, we turn from (5.12)–(5.14) to the equivalent system

$$\begin{aligned} u'(t) - v(t) + h_u'(u(t), t)^T z(t) &= 0, \\ v'(t) + g(u(t), v(t), t) + h_u'(u(t), t)^T w(t) &= 0, \\ h(u(t), t) &= 0, \\ h_u'(u(t), t)v(t) + h_t'(u(t), t) &= 0. \end{aligned} \tag{5.15}$$

Surely, if $u_*, v_* \in C^1$, $w_* \in C$ form a solution of (5.12)–(5.14) then $(u_*, v_*, w_*, \mathcal{O})$ satisfies (5.15). Conversely, for each solution (u_*, v_*, w_*, z_*) of (5.15) with $z_* = \mathcal{O}$ the first components (u_*, v_*, w_*) solve (5.12)–(5.14). The point of (5.15) is that any solution has a trivial fourth component (cf. [6]). Namely, we have $h'_v v + h'_t = 0$, $u' = v - h'^T_u z$. Differentiating the constraint in (5.15) yields

$$h'_u u' + h'_t = 0,$$

hence

$$h'_u v + h'_t - h'_u h'^T_u z = 0,$$

and finally $z = 0$.

The DAE (5.15) is index-2-tractable. To show this we put this system in our general form (5.1) and compute the related matrices. We have

$$f'_y = \begin{bmatrix} I & & & \\ & I & & \\ & & 0 & \\ & & & 0 \end{bmatrix}, \quad f'_x = \begin{bmatrix} h''^T_{uu} z & -I & 0 & H^T \\ g'_u + h''^T_{uu} w & g'_v & H^T & 0 \\ H & 0 & 0 & 0 \\ h''_{uu} v + h''_{tu} & H & 0 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} I & & H^T \\ & I\ H^T & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} H^T(HH^T)^{-1}H & & \\ & H^T(HH^T)^{-1}H & \\ & -(HH^T)^{-1}H\ 0 & \\ -(HH^T)^{-1}H & & 0 \end{bmatrix}.$$

Now,

$$\tilde{z} \in \ker(A_1(y, x, t)), \quad f'_x(y, x, t) P\tilde{z} \in \operatorname{im}(A_1(y, x, t)), \quad \tilde{z} = [a^T b^T c^T d^T]^T,$$

imply

$$a + H^T d = 0, \quad b + H^T c = 0, \quad Ha = 0, \quad (\ldots)a + Hb = 0.$$

This yields $a = 0$, $d = 0$, $b = 0$, $c = 0$, i.e. the matrix $G_2 = A_1 + f'_x PQ_1$ remains nonsingular. Hence, (5.15) is index-2-tractable.

In [6], [12], the BDF's applied to index-2 systems are reported to work well. But the BDF applied to index-3 systems becomes much more unreliable. A reduction step, for instance from (5.12)–(5.14) to (5.15), is recommended in [6] (cf. also [7]).

Now, consider the variable-order variable-step BDF applied to a nonlinear index-2-tractable DAE (5.1). Denote by $x_* \in C^1_N$ the solution of (5.1) to be approximated. Let starting values $x_0, \ldots, x_{k-1} \in \boldsymbol{R}^m$ be given. The BDF is now

$$f\left(\frac{1}{h_j}\sum_{i=0}^{k} a_{ji} x_{j-i}, x_j, t_j\right) = 0, \quad j = k, \ldots, n. \tag{5.16}$$

Recall that convergence is proved for certain semi-explicit systems ([2], [6], [12]), in particular, for (5.15).

The related map acting in $\boldsymbol{R}^{m(n+1)}$ is

$$\mathscr{F}_{\Pi} z := \begin{bmatrix} z_0 - x_0 \\ z_{k-1} - x_{k-1} \\ f\left(\frac{1}{h_k}\sum_{i=0}^{k} a_{ki} z_{k-i},\ z_k,\ t_k\right) \\ f\left(\frac{1}{h_n}\sum_{i=0}^{k} a_{ni} z_{n-i},\ z_n,\ t_n\right) \end{bmatrix}, \qquad z =: \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_n \end{bmatrix} \in \boldsymbol{R}^{m(n+1)}.$$

Surely, $\mathscr{F}_{\Pi}$ is continuously differentiable provided assumption (A) is satisfied. More precisely, the Jacobian taken at

$$z^* = \begin{bmatrix} x_*(t_0) \\ \vdots \\ x_*(t_n) \end{bmatrix}$$

is

$$\mathscr{F}'_{\Pi}(z^*) = \begin{bmatrix} I & & \\ \cdots & \cdots & \cdots \\ & I & \\ \frac{a_{kk}}{h_k}\tilde{A}_k & \cdots \frac{a_{k1}}{h_k}\tilde{A}_k & \tilde{F}_k \\ \cdots & \cdots & \cdots \\ \frac{a_{nk}}{h_n}\tilde{A}_n & \cdots \frac{a_{n1}}{h_n}\tilde{A}_n & \tilde{F}_n \end{bmatrix} \tag{5.17}$$

where

$$\tilde{A}_j := f'_y(\eta_j), \qquad \tilde{F}_j := \frac{a_{j0}}{h_j} f'_y(\eta_j) + f'_x(\eta_j),$$

$$\eta_j := \left(\frac{1}{h}\sum_{i=0}^{k} a_{ji} x_*(t_{j-i}),\ x_*(t_j),\ t_j\right) \in \mathscr{G}.$$

Besides the nonlinear DAE (5.1) we consider the linear DAE

$$Az' + Bz = q \tag{5.18}$$

with the coefficients $\{A, B\} = \{f'_y(M_*(\cdot)), f'_x(M_*(\cdot))\} \in \mathscr{N}$ (cf. (5.4)). This linear equation is index-2-tractable. Apply the BDF under consideration to the linear DAE (5.18) also. For $P' = 0$, the matrix $\mathscr{L}_{\Pi}$ (cf. (3.10)) which is related to the BDF applied to (5.18) is nonsingular. Supposed the related projection $Q_1(t)$ is smooth, the inequalities (3.11), (3.18) and (3.20) hold for $\mathscr{L}_{\Pi} z = w$.

Now, we turn to a more restricted class of nonlinear DAE's, namely to quasilinear ones of the form

$$A(t)x'(t) + g(x(t), t) = 0 \tag{5.19}$$

where the nullspace of $A(t)$ is assumed to be constant, and $P' = 0$. Moreover, let $g'_x(x, t)$ be Lipschitz with respect to x. Then, for $j = k, \ldots, n$, it holds that

$$A_j - \tilde{A}_j = f'_y(M_*(t_j)) - f'_y(\eta_j) = 0,$$
$$F_j - \tilde{F}_j = f'_x(M_*(t_j)) - f'_x(\eta_j) = 0.$$

This yields

$$\mathscr{L}_{\Pi} = \mathscr{F}'_{\Pi}(z^*), \tag{5.20}$$

further, for

$$\tilde{\mathscr{F}}_{\Pi}(z, \bar{z}) := \int_0^1 \mathscr{F}'_{\Pi}(sz + (1-s)\bar{z})\, ds,$$

$$\tilde{\mathscr{F}}_{\Pi}(z, \bar{z}) - \mathscr{F}'_{\Pi}(z^*) = \operatorname{diag}(0, \ldots, 0, \alpha_k, \ldots, \alpha_n),$$

where

$$\alpha_j := \int_0^1 (g'_x(sz_j + (1-s)\bar{z}_j, t_j) - g'_x(z_j^*, t_j))\, ds.$$

Now, for

$$z, \bar{z} \in \mathscr{B}_{\Pi}(z^*, \varrho) := \{\tilde{z} \in \boldsymbol{R}^{m(n+1)} : \|\tilde{z} - z^*\|_\infty \leqslant \varrho\},$$

we derive

$$\|\alpha_j\|_\infty \leqslant \varrho L, \quad j = k, \ldots, n,$$

thus

$$\|\tilde{\mathscr{F}}_{\Pi}(z, \bar{z}) - \mathscr{F}'_{\Pi}(z^*)\|_\infty \leqslant \varrho L. \tag{5.21}$$

The relations (5.21), (5.20), (3.11) imply, by the Banach Lemma, the nonsingularity of $\tilde{\mathscr{F}}_{\Pi}(z, \bar{z})$ as well as the inequality

$$\|\tilde{\mathscr{F}}_{\Pi}(z, \bar{z})^{-1}\|_\infty = \frac{\underline{h}^{-1} S}{1 - \underline{h}^{-1} S \varrho L}, \tag{5.22}$$

if $\varrho > 0$ is chosen so that

$$\underline{h}^{-1} S L \varrho < 1 \tag{5.23}$$

is valid.

Introduce the map $E_{\Pi}: \boldsymbol{R}^{m(n+1)} \to \boldsymbol{R}^{m(n+1)}$ by

$$E_{\Pi} z := z - \mathscr{F}'_{\Pi}(z^*)^{-1} \mathscr{F}_{\Pi} z, \quad z \in \boldsymbol{R}^{m(n+1)}.$$

For $z, \bar{z} \in B_{\Pi}(z^*, \varrho)$, we obtain

$$\begin{aligned} \|E_{\Pi} z - E_{\Pi} \bar{z}\|_\infty &= \|z - \bar{z} - \mathscr{F}'_{\Pi}(z^*)(\mathscr{F}_{\Pi} z - \mathscr{F}_{\Pi} \bar{z})\|_\infty \\ &\leqslant \|\mathscr{F}'_{\Pi}(z^*)^{-1}\|_\infty \|\mathscr{F}'_{\Pi}(z^*) - \tilde{\mathscr{F}}_{\Pi}(z, z)\|_\infty \|z - \bar{z}\|_\infty \\ &\leqslant \underline{h}^{-1} S \varrho L \|z - \bar{z}\|_\infty, \end{aligned} \tag{5.24}$$

$$(5.25)\quad \|E_\Pi z - z^*\|_\infty = \|z - z^* - \mathscr{F}'_\Pi(z^*)^{-1}(\mathscr{F}_\Pi z - \mathscr{F}_\Pi z^* + \mathscr{F}_\Pi z^*)\|_\infty$$
$$\leqslant \|\mathscr{F}'_\Pi(z^*)^{-1}\|_\infty \|\mathscr{F}'_\Pi(z^*) - \tilde{\mathscr{F}}_\Pi(z, z^*)\|_\infty \|z - z^*\|_\infty + \|\mathscr{F}'_\Pi(z^*)^{-1}\mathscr{F}_\Pi z^*\|_\infty$$
$$\leqslant \underline{h}^{-1} S\varrho L \|z - z^*\|_\infty + \|\mathscr{F}'_\Pi(z^*)^{-1}\mathscr{F}_\Pi z^*\|_\infty.$$

Denote $\mathscr{F}_\Pi z^* =: \tau$. The values $\tau_j = x_*(t_j) - x_j$, $j = 0, \ldots, k-1$, represent the errors of the starting values while

$$\tau_j = A(t_j)\frac{1}{h_j}\sum_{i=0}^{k} a_{ji}\, x_*(t_{j-i}) + g(x_*(t_j), t_j), \qquad j = k, \ldots, n,$$

are the local discretization errors.

Taking into account that (cf. (3.22), (5.7))

$$\tau_j \in \operatorname{im}(A(t_j)),$$
$$G_2(t_j)^{-1}\tau_j = G_2(t_j)^{-1} A(t_j) A(t_j)^+ \tau_j = P_1(t_j) P A(t_j)^+ \tau_j,$$
$$G_2(t_j) := G_2(M_*(t_j)),$$

hold for $j = k, \ldots, n$, we derive the inequality

$$(5.26)\quad \|\mathscr{F}'_\Pi(z^*)^{-1}\mathscr{F}'_\Pi z^*\|_\infty$$
$$\leqslant S_7 \|\tau\|_\infty + S_8 \max_{j=k,\ldots,2k-1} \frac{1}{h_j}\Big|QQ_1(t_j)\sum_{i=j-k+1}^{k} a_{ji}\, Q_1(t_{j-i})\tau_{j-i}\Big|$$

from (3.18), (3.19), (3.20).

Now, for a fixed value $\alpha \in (0, 1)$ we choose a refined grid $\pi \in \Pi(\eta_1, \eta_2, h_{\max})$ and sufficiently accurate starting values so that

$$(5.27)\qquad \|\mathscr{F}'_\Pi(z^*)\mathscr{F}_\Pi z^*\|_\infty \leqslant (1-\alpha).$$

This is always possible by (5.26). Practically, this means a choice $x_j = x_*(t_j) + O(h_j^{\mu+1})$, $j = 0, \ldots, k-1$, applying the μ-order BDF.

Then, we choose ϱ small enough so that

$$(5.28)\qquad \underline{h}^{-1} SL\varrho < \alpha.$$

In this way, the map E_Π is contractive with $\alpha < 1$ on the closed ball $\mathscr{B}_\Pi(z^*, \varrho)$, and $E_\Pi(\mathscr{B}_\Pi(z^*, \varrho)) \subseteq \mathscr{B}_\Pi(z^*, \varrho)$. Consequently, by Banach's Fixed Point Theorem, the map E_Π has a unique fixed point onto $\mathscr{B}_\Pi(z^*, \varrho)$, i.e. the nonlinear equation

$$\mathscr{F}_\Pi z = 0,$$

representing our integration method has a locally unique solution, i.e. there exist locally unique values $x_k, \ldots, x_n \in \boldsymbol{R}^m$, satisfying (5.16).

Since, for $z, \bar{z} \in \mathscr{B}_\Pi(z^*, \varrho)$,

$$z - \bar{z} = \tilde{\mathscr{F}}_\Pi(z, \bar{z})^{-1}(\mathscr{F}_\Pi z - \mathscr{F}_\Pi \bar{z})$$

holds, the inequality

$$\|z-\bar{z}\|_\infty \leqslant \frac{\underline{h}^{-1}S}{1-\underline{h}^{-1}SQL}\|\mathscr{F}_\Pi z-\mathscr{F}_\Pi \bar{z}\|_\infty \tag{5.29}$$

follows (cf. (5.22)). Thus, (5.16) is shown to be only weakly unstable.

Using (5.25) for the true numerical solution $z = x$, i.e. the fixed point of E_Π, we obtain the error estimation

$$\max_{j=0,\ldots,n} |x_*(t_j)-x_j| \leqslant \frac{1}{1-\alpha}\|\mathscr{F}'_\Pi(z^*)^{-1}\mathscr{F}_\Pi z^*\|_\infty. \tag{5.30}$$

Consequently, if exact starting values are used, i.e. $x_*(t_j) = x_j$, $j = 0, \ldots, k-1$, we have simply

$$\max_{j=0,\ldots,n} |x_*(t_j)-x_j| \leqslant \frac{1}{1-\alpha}\max_{j=k,\ldots,n}|\tau_j| \tag{5.31}$$

that means the variable-order variable-step BDF applied to the quasilinear DAE (5.19) is convergent with the same order as in the case of regular explicit ODE's.

THEOREM 5.1. *Let $x_* \in C_N^1$ be a solution of the quasilinear DAE* (5.19). *Let* (5.19) *be index-2-tractable in the neighbourhood of x_*. Further, let g'_x be Lipschitz there with respect to x, and let $A(t)$ have a constant nullspace.*

Moreover, let the projection $Q_1(t)$ onto the nullspace $N_1(t) := \ker(A_1(t))$ along $S_1(t) := \{z \in \boldsymbol{R}^m\colon g'_x(x_*(t), t)Pz \in \operatorname{im}(A_1(t))\}$ be continuously differentiable, where

$$A_1(t) := A_1(M_*(t)) = A(t)+g'_x(x_*(t), t)Q \tag{5.32}$$

Let the variable-step variable order BDF be stable on $\Pi(\eta_1, \eta_2, h_{\max})$ in the regular ODE-case, $h_{\max}$ sufficiently small. Then this method is only weakly unstable on $\Pi(\eta_1, \eta_2, h_{\max})$. It is convergent with the same order as in the case of regular ODE's.

Moreover, for each grid $\pi \in \Pi(\eta_1, \eta_2, h_{\max})$, for sufficiently accurate starting values $x_0, \ldots, x_{k-1}$, the nonlinear equations (5.16) *are uniquely solvable with respect to x_j on balls $\{u \in \boldsymbol{R}^m\colon |x_*(t_j)-u| \leqslant \varrho\}$, $j = k, \ldots, n$, where ϱ depends on Π,*

$$\lim_{\underline{h}\to 0} \varrho = 0. \tag{5.33}$$

The behaviour of ϱ (cf. e.g. (5.23)) seems not to be only a consequence of the technique used for the proof but reflects as matters stand in the nonlinear equations to be solved. In any case, we are not able to compute the values x_j exactly. Instead of x_j we generate only certain $\tilde{x}_j \in \boldsymbol{R}^m$ satisfying (5.16) approximately, i.e. we have

$$f\left(\frac{1}{h_j}\sum_{i=0}^{k} a_{ji}\,\tilde{x}_{j-i},\ \tilde{x}_j,\ t_j\right)=\delta_j,\quad j=k,\ldots,n, \tag{5.34}$$

$$\tilde{x}_j := x_j,\quad j=0,\ldots,k-1. \tag{5.35}$$

Recall that $\tau_j := x_*(t_j)-x_j = x_*(t_j)-\tilde{x}_j$, $j=0,\ldots,k-1$, denote the errors in the starting values.

THEOREM 5.2. *Let the assumptions of Theorem 5.1 be satisfied. Then, for sufficiently small defects δ_j in (5.34), (5.35), the error estimation*

$$\max_{j\geq k}|x_*(t_j)-\tilde{x}_j| \leqslant S_9 \max\Big\{\max_{j\leqslant k-1}|\tau_j|,\ \max_{j\geq k}|\tau_j-\delta_j|\Big\} + \max_{j\geq k}\left|QQ_1(t_j)\frac{1}{h_j}\sum_{i=0}^{k} a_{ji}\,Q_1(t_{j-i})\,\omega_{j-i}\right| \tag{5.36}$$

holds, where

$$\omega_j := \begin{cases}\tau_j, & j\leqslant k-1,\\ G_2(t_j)^{-1}\,\delta_j, & j\geqslant k,\end{cases}$$

$$G_2(t_j) := A(t_j)+g'_x(x_*(t_j),\,t_j)\big(Q+PQ_1(t_j)\big).$$

Proof. With

$$\tilde{x}=\begin{bmatrix}\tilde{x}_0\\ \vdots\\ \tilde{x}_n\end{bmatrix},\quad z^*=\begin{bmatrix}x_*(t_0)\\ \vdots\\ x_*(t_n)\end{bmatrix},\quad \delta=\begin{bmatrix}0\\ \vdots\\ 0\\ \delta_k\\ \vdots\\ \delta_n\end{bmatrix},\quad \tau=\begin{bmatrix}\tau_0\\ \vdots\\ \tau_n\end{bmatrix}$$

we have $\mathscr{F}_{II}\tilde{x}=\delta$, $\mathscr{F}_{II}z^*=\tau$,

$$\begin{aligned}\tilde{x}-z^* &= \tilde{x}-z^*-\mathscr{F}'_{II}(z^*)^{-1}\{\mathscr{F}_{II}\tilde{x}-\mathscr{F}_{II}z^*\}+\mathscr{F}'_{II}(z^*)^{-1}\{\mathscr{F}_{II}\tilde{x}-\mathscr{F}_{II}z^*\}\\ &= \mathscr{F}'_{II}(z^*)^{-1}\{\mathscr{F}'_{II}(z^*)-\mathscr{F}'_{II}(\tilde{x},\,z^*)\}(\tilde{x}-z^*)+\mathscr{F}'_{II}(z^*)^{-1}(\delta-\tau),\end{aligned}$$

trivially. Hence (cf. (5.21), (3.11), (5.28))

$$\|\tilde{x}-z^*\|_\infty \leqslant \underline{h}^{-1}SQL\|\tilde{x}-z^*\|_\infty+\|\mathscr{F}'_{II}(z^*)^{-1}(\delta-\tau)\|_\infty, \tag{5.37}$$

further

$$\|\tilde{x}-z^*\|_\infty \leqslant \frac{1}{1-\alpha}\|\mathscr{F}'_{II}(z^*)^{-1}(\delta-\tau)\|_\infty. \tag{5.38}$$

Due to (3.18), (3.20), we find

$$\|\mathscr{F}'_{II}(z^*)^{-1}(\delta-\tau)\|_\infty \leqslant S_{10}\|\sigma-\tau\|_\infty+\max_{j=k,\ldots,n}\frac{1}{h_j}\left|QQ_1(t_j)\sum_{i=0}^{k} a_{ji}\,Q_1(t_{j-i})\,\omega_{j-i}\right| \tag{5.39}$$

where

$$\omega_j := \tau_j, \quad j = 0, \ldots, k-1,$$
$$\omega_j := G_2^*(t_j)^{-1}\delta_j. \ \blacksquare$$

EXAMPLE. For the Euler backward method applied to the index-2 system (5.15) we have

$$PQ_1 = \operatorname{diag}(T, T, 0, 0), \quad T := H^+ H, \quad H^+ := H^T(HH^T)^{-1},$$

$$QQ_1(t_j)\frac{1}{h_j}(Q_1(t_j)\omega_j - Q_1(t_{j-1})\omega_{j-i})$$
$$= \begin{cases} QQ_1(t_1)\dfrac{1}{h_1}G_2(t_1)^{-1}\delta_1 - QQ_1(t_1)\dfrac{1}{h_1}Q_1(t_0)(x_*(t_0)-x_0) \\ QQ_1(t_j)\dfrac{1}{h_j}(Q_1(t_j)G_2(t_j)^{-1}\delta_j - Q_1(t_{j-1})G_2(t_{j-1})^{-1}\delta_{j-1} \end{cases} \quad \text{for } j \geqslant 1,$$

further

$$Q_1(t)G_2(t)^{-1}\begin{bmatrix}\bar{u}\\ \bar{v}\\ \bar{w}\\ \bar{z}\end{bmatrix} = \begin{bmatrix} H^+\bar{w} \\ H^+(z-(h''_{uu}v+h''_{tu})H^+\bar{w}) \\ -(HH^T)^{-1}(z-(h''_{uu}v+h''_{tu})H^+\bar{w}) \\ -(HH^T)^{-1}w \end{bmatrix}.$$

Thus, the defects in the third and fourth equation of the system (5.15) should be kept small in comparison with the defects in the first and second equation. This corresponds completely to the considerations in [6], [12] as well as to the practical experience reported e.g. in [6] and [19]. ∎

References

[1] Ju. E. Boyarintsev, *Regulyarnye i singulyarnye sistemy linejnych obyknovennych differencialnych uravnenij*, Nauka (Sib. otd.), Novosibirsk 1980.

[2] K. E. Brenan and B. E. Engguist, *Backward difference approximations of nonlinear differential-algebraic equations*, California, El Segundo, The Aerospace Corporation, 1985 (ATR-85 (9990)-5).

[3] S. L. Campbell, *Regularizations of linear time varying singular systems*, Automatica – J. IFAC 20 (1984) 365 370.

[4] F. R. Gantmakher, *Theory of Matrices*.

[5] C. W. Gear and L. Petzold, *ODE methods for the solution of differential/algebraic systems*, SIAM J. Numer. Anal. 21 (1984), 716-728.

[6] C. W. Gear, B. Leimkuhler and G. K. Gupta, *Automatic integration of Euler-Lagrange equations with constraints*, J. Comput. Appl. Math. 12-13 (1985), 77-90.

[7] E. Griepentrog, *Transformation concepts for differential-algebraic equations*, this volume, 223-231.

[8] E. Griepentrog and R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Leipzig; BSB B. G. Teubner, 1986 (Teubner-Texte zur Mathem. 88).

[9] M. Hanke, *On the regularization of index 2 differential-algebraic equations*, Berlin, Humboldt University, Section of Mathematics, 1987 (Preprint 137).

[10] M. Hanke, R. März and A. Neubauer, *On the regularization of a class of nontransferable differential algebraic equations*, J. Differential Equations, 73 (1988), 119–132.
[11] –, –, –, *On the regularization of linear differential-algebraic equation*, in H. W. Engl and C. W. Groetsch (eds.), *Inverse and Ill-Posed Problems*, Academic Press, New York 1987.
[12] P. Lötstedt and L. Petzold, *Numerical solution of nonlinear differential equations with algebraic constraints*, I. Math. Comp. 46 (1986), 491–516.
[13] R. März, *Index-2 differential-algebraic equations*, Results in Math. 15 (1989), 149–171.
[14] –, *Some new results concerning index-3-differential-algebraic equations*, J. Math. Anal. Appl., 140 (1989), 177–199.
[15] –, *Recent Results on Higher-Index Differential-Algebraic Equations*, Teubner Texte zur Mathem. (Proc. Halle 1987), 104, 104–112.
[16] –, *On the numerical treatment of differential-algebraic equation*, Z. Angew. Math. Mech. 67 (1987), 23–34.
[17] –, *On correctness and numerical treatment of boundary value problem in DAE's*, Zh. Vychisl. Mat. i Mat. Fiz. 26 (1986), 50–64.
[18] –, *On tractability with index 2*, Berlin, Humboldt University, Section of Mathematics, 1987 (Preprint 109).
[19] L. Petzold and P. Lötstedt, *Numerical solution of nonlinear differential equations with algebraic constraints, II, Practical implications*, SIAM J. Sci. Statist. Comput. 7 (1986), 720–733

Presented to the Semester
Numerical Analysis and Mathematical Modelling
February 25 – May 29, 1987