Acta Numerica (1991), *pp.* 141–198

Numerical methods for differential algebraic equations

Roswitha März
Humboldt-Universität
Fachbereich Mathematik
Postfach 1297, D-O-1086 Berlin, Germany
E-mail: ivs@mathematik.hu-berlin.dbp.de

CONTENTS

0. Introduction

Differential algebraic equations (DAE) are special implicit ordinary differential equations (ODE)

$$f(x'(t), x(t), t) = 0, \tag{0.1}$$

where the partial Jacobian $f'_y(y, x, t)$ is singular for all values of its arguments.

These DAEs arise in various fields of applications. The most popular ones are simulation of electrical circuits, chemical reactions subject to invariants, vehicle system dynamics, optimal control of lumped-parameter systems, semi-discretization of partial differential equation systems and singular perturbation problems. For a fairly detailed survey of applications we refer to Brenan *et al.* (1989).

In the last few years, DAEs have developed into a highly topical subject in applied mathematics. There is a rapidly increasing number of contributions devoted to DAEs in the mathematical literature as well as in the fields of mechanical engineering, chemical engineering, system theory, etc. Frequently, other names such as semi-state equations, descriptor systems, singular systems are assigned to DAEs. In 1971 C.W. Gear proposed that DAEs should be handled numerically by backward differentiation formulae

(BDF). Since then, powerful codes which successfully simulate large circuits have been developed. For a long time DAEs were considered to be essentially similar to regular implicit ODEs in general. However, challenged by computation results that could not be brought into line with this supposition (e.g. Sincovec *et al.*, 1981), the mathematical community started investigating DAEs more thoroughly. With their famous paper, C.W. Gear *et al.* (1981) initiated a discussion on DAEs which will surely continue for a long time.

What kind of mathematical objects are DAEs? First of all, they are singular ODEs. Can they really be treated numerically like regular ODEs? Surely not in every case! How can one characterize single classes of problems for which methods that have proved their value for regular ODEs work well in other instances? What is the reason for their not working otherwise? How are appropriate numerical methods to be constructed then? All these questions can only be answered when more is known about the mathematical nature of DAEs.

In 1984 W.C. Rheinboldt began regarding DAEs as differential equations on manifolds. This approach provided useful insights into the geometrical and analytical nature of these equations (e.g. Reich (1990), Rabier and Rheinboldt (1991)).

Assuming sufficient smoothness of all functions involved, the DAE

$$\left.\begin{aligned} u' + g(u,v) &= 0 \\ h(u,v) &= 0 \end{aligned}\right\} \tag{0.2}$$

can be regarded as a vector field on

$$S_1 := \left\{ \begin{bmatrix} u \\ v \end{bmatrix} : h(u,v) = 0 \right\},$$

$$\left.\begin{aligned} u' &= -g(u,v) \\ v' &= h'_v(u,v)^{-1} h'_u(u,v) g(u,v) \end{aligned}\right\}, \quad \begin{bmatrix} u \\ v \end{bmatrix} \in S_1,$$

provided that $h'_v(u,v)$ is nonsingular everywhere. All solutions belong to S_1, and each point of S_1 is passed by a solution.

The DAE

$$\left.\begin{aligned} u' + g(u,v) &= 0 \\ h(u) &= 0 \end{aligned}\right\} \tag{0.3}$$

is more complicated. By differentiating twice and eliminating derivatives it can be checked that this system generates the vector field

$$\left.\begin{aligned} u' &= -g(u,v) \\ v' &= (h'(u)g'_v(u,v))^{-1}\{h''(u)g(u,v) + h'(u)g'_u(u,v)\}g(u,v) \end{aligned}\right\}, \quad (u^T, v^T)^T \in S_2,$$

where now

$$S_2 := \left\{ \begin{bmatrix} u \\ v \end{bmatrix} : h(u) = 0, h'(u)g(u,v) = 0 \right\}$$

represents the state manifold. The nonsingularity of $h'(u)g'_v(u,v)$ has been assumed here.

Analogously, for the DAE

$$\left.\begin{aligned} v' + f(u,v,w) &= 0 \\ u' + g(u,v) &= 0 \\ h(u) &= 0 \end{aligned}\right\} \tag{0.4}$$

one can define a vector field on the manifold

$$\begin{aligned} S_3 \;:=\; & \{(u^T, v^T, w^T)^T : h(u) = 0, h'(u)g(u,v) = 0, \\ & h''(u)g(u,v)g(u,v) + h'(u)(g'_u(u,v)g(u,v) + g'_v(u,v)f(u,v,w) = 0)\} \end{aligned}$$

provided that $h'(u)g'_v(u,v)f'_w(u,v,w)$ remains nonsingular.

In these three cases one speaks of semi-explicit DAEs with index 1, 2 and 3, respectively. The special structure of equations (0.3) and (0.4) is called the Hessenberg form.

If these vector fields were not considered on the specified manifolds $S_i \subset \mathbb{R}^m$, but formally on $\mathbb{R}^m$, then the resulting regular ODEs could be integrated with the usual methods. Even if we start with consistent initial values, we will very swiftly drift away from S_2 and S_3 in (0.3) and (0.4), respectively. Hence, many authors are concerned with the development of very special methods for (0.3) and (0.4), thereby exploiting the geometry of these equations. There are important applications that have this form, e.g. the Euler–Lagrange formulation of constrained mechanical systems leads to the form (0.4).

Under the corresponding assumptions, a state manifold and a vector field can also be assigned to the general DAE (0.1). However, both are only defined implicitly and, in general, not available in practice. This has already been indicated by the simple case of equation (0.4) and S_3. More general approaches for the constructive use of geometry for numerical mathematics are not known to the author.

If we have a closer look at equation (0.2) it becomes obvious that, theoretically, in the neighbourhood of a consistent initial value $(u_0^T, v_0^T)^T \in S_1$ we could investigate the locally decoupled system

$$u' + g(u, \mathcal{S}(u)) = 0, \; v = \mathcal{S}(u) \tag{0.5}$$

with $h(u, \mathcal{S}(u)) = 0$ instead of (0.2). Now it would be advantageous to integrate this regular ODE for the component u numerically and, then, simply to determine $v_j = \mathcal{S}(u_j)$ in each case. With suitable integration methods, this idea can even be realized in practice for general index-1 equations (0.1).

We would like to point out another aspect of the characterization of DAEs, which is fundamental, in particular, to the numerical treatment. For this, we consider the special equation of the form (0.3), which is perturbed by an inhomogeneity,

$$\left.\begin{aligned} u' - v &= 0 \\ u &= p(t) \end{aligned}\right\}. \tag{0.6}$$

Here, the function p has to be differentiated, i.e. $v(t) = p'(t)$ has to be computed. Differentiation is one of the classical examples of ill posed problems. A corresponding inhomogeneous problem of the form (0.4) will require a second differentiation. The greater the number of differentiations, the more strongly ill posed the problems become.

Both (0.5) and (0.6) make clear that a natural approach to the solution is directed to $u \in C^1$, $v \in C$. In many applications one aims at reducing smoothness, which has, unfortunately, not yet been successfully taken into account in the interpretation of DAEs used to represent ODEs on manifolds.

In the present paper we characterize general DAEs (0.1) under possibly minimal smoothness demands, where the characterization aims at the numerical tractability. Since (from the present point of view) all the essentially new numerical difficulties in comparison with regular ODEs have already become for linear equations with variable coefficients, we devote most of our investigations to the analytical characterization and investigation of integration methods for linear equations.

To apply the results to nonlinear equations we slightly modify the standard arguments of discretization theory. The BDFs are studied in detail here because, on the one hand, they can be especially recommended just for DAEs and, on the other hand, they serve, in a certain sense, as model methods.

We want to emphasize that this paper does not aim at providing a survey of all the available results and methods. In particular, we do not enter into the details of the many nice but very special results for (0.3) and (0.4) (for this, see e.g. Hairer *et al.* (1989), Lubich (1990), Potra and Rheinboldt (1991), Simeon *et al.* (1991)). We focus our interest on exposing problems and showing constructive approaches for their solution, where we try to maintain a uniform concept of representation.

Altogether, many problems with respect to DAEs still remain open. An appropriate numerical treatment requires – provided it is to be more than only favourable intention – profound knowledge about the analytical background of this type of equation.

The paper is organized as follows. In Section 1 the reader becomes acquainted with the fact that additional stability conditions and weak instabilities may occur in the integration of linear constant coefficient DAEs. Section 2 is devoted to the analytical and geometrical foundations of gen-

eral DAEs, where those of linear equations with time-dependent coefficients play a special role. In Section 3 the BDFs are discussed in detail, as already mentioned, as a model for constructing methods. Section 4 presents brief outlines on index reduction as well as on boundary value problems.

1. Analysing linear constant coefficient equations

Linear equations

$$Ax'(t) + Bx(t) = q(t) \tag{1.1}$$

with matrix coefficients $A, B \in L(\mathbb{R}^m)$, A singular, are easy to understand when taking into account the close relationship with matrix pencils $\{A, B\}$ (e.g. Gantmacher (1966)). In this section we explain some basic facts on how and for what reasons well-known discretization methods behave when applied to DAEs.

Definition The ordered pair of matrices $\{A, B\}$ forms a *regular matrix pencil* if the polynomial $p(\lambda) := \det(\lambda A + B)$ does not vanish identically. Otherwise, the pencil is called singular.

Weierstrass (cf. Grantmacher (1966)) has shown that a regular pencil $\{A, B\}$ can be transformed into $\{\tilde{A}, \tilde{B}\}$,

$$\left.\begin{aligned} \tilde{A} &:= EAF = \operatorname{diag}(I, J), \\ \tilde{B} &:= EAF = \operatorname{diag}(W, I) \end{aligned}\right\} \tag{1.2}$$

by the use of suitable regular matrices $E, F \in L(\mathbb{R}^m)$. Thereby, $W \in L(\mathbb{R}^k)$, and $J \in L(\mathbb{R}^{m-k})$ is a nilpotent Jordan block matrix with chains

$$\begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}.$$

Definition $\{\tilde{A}, \tilde{B}\}$ given by (1.2) is called the *Kronecker canonical normal form* of the regular pencil $\{A, B\}$. The *index* of a regular pencil is defined to be $\operatorname{ind}(A, B) := \operatorname{ind}(J) :=$ maximal Jordan chain order of J.

An equation of type (1.1) with a singular matrix pencil $\{A, B\}$ is somewhat incomplete. For these equations, the homogeneous initial value problem

$$Ax'(t) + Bx(t) = 0, \quad x(0) = 0$$

has more than countably many different solutions (see Griepentrog and März

(1986)). A typical example is

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

Singular matrix pencils in (1.1) indicate some defect in the modelling.

Here, we are interested in equations (1.1) with regular matrix pencils $\{A, B\}$ only. Using the transformation matrices E, F leading to the Kronecker normal form (cf. (1.2)) we may transform (1.1) equivalently into

$$\tilde{A}\tilde{x}'(t) + \tilde{B}\tilde{x}(t) = \tilde{q}(t), \tag{1.3}$$

where $\tilde{A}$, $\tilde{B}$ are given by (1.2), $\tilde{q}(t) := Eq(t)$, $\tilde{x}(t) := F^{-1}x(t)$. In more detail, (1.3) reads

$$u'(t) + Wu(t) = p(t) \tag{1.4}$$

$$Jv'(t) + v(t) = r(t), \tag{1.5}$$

where u, v and p, r are the related components of $\tilde{x}$ and $\tilde{q}$, respectively. Now, the decoupled system (1.4), (1.5) is said to be the *Kronecker normal form of equation (1.1)*. Moreover the index of equation (1.1) can also be traced back to $\mathrm{ind}(A, B) =: \mu$.

In accordance with the Jordan structure of J, equation (1.5) decouples into parts such as

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & & 1 \\ & & & 0 \end{bmatrix} w'(t) + w(t) = s(t) \tag{1.6}$$

of dimension $\gamma \leq \mu$.

If $\gamma = 1$, then (1.6) simply yields

$$w(t) = s(t).$$

If $\gamma = 2$, then (1.6) represents

$$\left.\begin{aligned} w_2'(t) + w_1(t) &= s_1(t) \\ w_2(t) &= s_2(t) \end{aligned}\right\}, \tag{1.7}$$

which leads to

$$w(t) = \begin{pmatrix} s_1(t) - s_2'(t) \\ s_2(t) \end{pmatrix}.$$

For $\gamma = 3$ we have

$$\left.\begin{aligned} w_2'(t) + w_1(t) &= s_1(t) \\ w_3'(t) + w_2(t) &= s_2(t) \\ w_3(t) &= s_3(t) \end{aligned}\right\}, \tag{1.8}$$

hence

$$w(t) = \begin{bmatrix} s_1(t) - (s_2(t) - s_3'(t))' \\ s_2(t) - s_3'(t) \\ s_3(t) \end{bmatrix}.$$

In general, if μ denotes the index of our equation (1.1), then (1.5) contains at least one part (1.6) of dimension $\gamma = \mu$, and, in consequence, certain components of the right-hand side have to be differentiated $\mu - 1$ times.

Clearly, (1.4) is a regular linear ODE. For all continuous right-hand sides $p(\cdot) : \mathcal{I} \to \mathbb{R}^k$ there is a unique solution $u(\cdot) : \mathcal{I} \to \mathbb{R}^k$ passing through given $(u^0, t_0) \in \mathbb{R}^k \times \mathcal{I}$.

On the other hand, the solution of (1.5) may be expressed as

$$v(t) = \sum_{j=0}^{\mu-1} (-1)^j (J^j r(t))^{(j)}.$$

The initial value $v(t_0)$ is fixed completely, and for solvability we have to assume $r(\cdot) : \mathcal{I} \to \mathbb{R}^{m-k}$ to be as smooth as necessary. From this point of view, for $\mu > 1$, equation (1.5) represents a differentiation problem. It will be pointed out later that this causes numerical difficulties. (Recall the well known fact that differentiation represents an ill posed problem in the continuous function space!)

Clearly, initial value problems for (1.1) only become solvable for *consistent initial values*

$$x(t_0) = F\tilde{x}(t_0) = F \begin{bmatrix} u^0 \\ v(t_0) \end{bmatrix},$$

where $u^0 \in \mathbb{R}^k$ is a free parameter, but $v(t_0)$ is determined as described earlier.

This is the second essential difference from regular ODEs and, when $\mu > 1$, this also entails considerable numerical problems, which have not yet been solved sufficiently.

At this point it should be emphasized again that the canonical normal form is used only to provide an immediate insight into the structure of (1.1). However, we do not think of transforming (1.1) into (1.5), (1.6) in practical computations!

Next we check what will happen when numerical integration methods approved for regular ODEs are applied to the singular ODE (1.1). First we consider the multi-step method

$$\frac{1}{h} A \sum_{j=0}^{s} \alpha_j x_{\ell-j} + B \sum_{j=0}^{s} \beta_j x_{\ell-j} = q(\bar{t}_\ell), \tag{1.9}$$

$$\bar{t}_\ell := \sum_{j=0}^{s} \beta_j t_{\ell-j}, \qquad \alpha_0 \neq 0,$$

where x_i is expected to approximate the true solution value $x(t_i)$. Again we decouple equations (1.9) according to the Kronecker canonical normal form by multiplying (1.9) by E and transforming

$$F^{-1}x_i = \tilde{x}_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix}.$$

This yields

$$\frac{1}{h}\sum_{j=0}^{s} \alpha_j u_{\ell-j} + W \sum_{j=0}^{s} \beta_j u_{\ell-j} = p(\bar{t}_\ell) \tag{1.10}$$

$$J\frac{1}{h}\sum_{j=0}^{s} \alpha_j v_{\ell-j} + \sum_{j=0}^{s} \beta_j v_{\ell-j} = r(\bar{t}_\ell). \tag{1.11}$$

Formula (1.10) represents the given multi-step method applied to the inherent regular ODE within the singular system (1.1). On the other hand, (1.11) may be solved with respect to v_ℓ if the matrix

$$\alpha_0 J + h\beta_0 I$$

is nonsingular, that is for $\beta_0 \neq 0$.

In the index-1 case $J = 0$, and (1.11) simply becomes

$$\sum_{j=0}^{s} \beta_j v_{\ell-j} = r(\bar{t}_\ell). \tag{1.12}$$

In März (1984, 1985) it was pointed out that, for the stability of the difference equation (1.12), it is necessary for the polynomial $\sum_{j=0}^{s} \beta_j \lambda^{s-j}$ to have all its roots within the interior of the complex unit circle. In particular, symmetric schemes (1.12) like, for example, the centred Euler scheme become unstable.

The best way to avoid error accumulations in (1.12) is to choose $\beta_0 = 1$, $\beta_1 = \cdots = \beta_s = 0$, e.g. to use the BDF.

For higher indexes $\mu > 1$ we only discuss the BDF. In index-2 parts such as (1.7) we have

$$\left.\begin{aligned} \frac{1}{h}\sum_{j=0}^{s} \alpha_j w_{2,\ell-j} + w_{1,\ell} &= s_1(t_\ell) \\ w_{2,\ell} &= s_2(t_\ell) \end{aligned}\right\}, \quad \ell \geq s,$$

thus

$$\left.\begin{aligned} w_{1,\ell} &= s_1(t_\ell) - \frac{1}{h}\sum_{j=0}^{s} \alpha_j s_2(t_{\ell-j}) \\ w_{2,\ell} &= s_2(t_\ell) \end{aligned}\right\}, \quad \ell \geq s,$$

if we assume exact starting values $w_{2,i} = s_2(t_i)$, $i = 0, \ldots, s-1$ are available. Inexact starting values as well as round-off errors in the linear equation (1.9) to be solved for x_ℓ give rise to a weak instability, i.e. the errors are amplified by $1/h$. However, fortunately, only the component $w_{1,\ell}$ is affected by this.

Analogously, in index-3 parts, such as e.g. (1.8), we have

$$\left.\begin{aligned} \frac{1}{h}\sum_{j=0}^{s} \alpha_j w_{2,\ell-j} + w_{1,\ell} &= s_1(t_\ell) \\ \frac{1}{h}\sum_{j=0}^{s} \alpha_j w_{3,\ell-j} + w_{2,\ell} &= s_2(t_\ell) \\ w_{3,\ell} &= s_3(t_\ell) \end{aligned}\right\}, \quad \ell \geq s.$$

Using these formulae for $\ell \geq 2s$ together with exact values $w_{3,i} = s_3(t_i)$, $i = 0, \ldots, 2s-1$, $w_{2,i} = s_2(t_i)$, $i = 0, \ldots, s-1$, would lead to

$$\begin{aligned} w_{1,\ell} &= s_1(t_\ell) - \frac{1}{h}\sum_{j=0}^{s} \alpha_j s_2(t_{\ell-j}) + \frac{1}{h^2}\sum_{j=0}^{s}\alpha_j \sum_{i=0}^{s} \alpha_i s_3(t_{\ell-j-i}) \\ w_{2,\ell} &= s_2(t_\ell) - \frac{1}{h}\sum_{j=0}^{s} \alpha_j s_3(t_{\ell-j}) \\ w_{3,\ell} &= s_3(t_\ell). \end{aligned}$$

Of course, in practical computations $2s$ exact starting values are not available, thus the components $w_{1,\ell}$ and $w_{2,\ell}$ will be affected by instabilities of the type $1/h^2$ and $1/h$, respectively. It should be mentioned that these instabilities are due to the differentiations arising in (1.1), and in this sense they are very natural.

Now, let us turn shortly to implicit Runge–Kutta methods for (1.1). Given the Runge–Kutta tableau

$$\begin{array}{c|c} \mathrm{c} & \mathbf{A} \\ \hline & \beta^T \end{array}$$ we have to solve the system

$$AX_i' + B(x_{\ell-1} + h\sum_{j=1}^{s} \alpha_{ij} X_j') = q(t_{\ell-1} + c_i h) \qquad i = 1, \ldots, s, \tag{1.13}$$

and then to compute

$$x_\ell = x_{\ell-1} + h\sum_{j=1}^{s} \beta_j X_j'. \tag{1.14}$$

Again we use the transformation to the Kronecker normal form. This gives

$$U_i' + W(u_{\ell-1} + h\sum_{j=1}^{s} \alpha_{ij} U_j') = p(t_{\ell-1} + c_i h), \qquad i = 1, \ldots, s, \tag{1.15}$$

$$u_\ell = u_{\ell-1} + h\sum_{j=1}^{s} \beta_j U_j', \tag{1.16}$$

$$JV_i' + v_{\ell-1} + h\sum_{j=1}^{s} \alpha_{ij} V_j' = r(t_{\ell-1} + c_i h), \qquad i = 1, \ldots, s, \tag{1.17}$$

$$v_\ell = v_{\ell-1} + h\sum_{j=1}^{s} \beta_j V_j'. \tag{1.18}$$

Clearly, (1.15) and (1.16) are nothing else but the given Runge–Kutta method applied to the regular inherent ODE (1.4). This part does not cause any new difficulties.

Equations (1.17) and (1.18) decouple further according to the Jordan chains in J (cf. (1.6)). For index-1 chains ($\gamma = 1$) we simply have

$$w_{\ell-1} + h\sum_{j=1}^{s} \alpha_{ij} W_j' = s(t_{\ell-1} + c_i h), \qquad i = 1, \ldots, s, \tag{1.19}$$

$$w_\ell = w_{\ell-1} + h\sum_{j=1}^{s} \beta_j W_j'. \tag{1.20}$$

Now it becomes clear that we have to use a nonsingular Runge–Kutta matrix $\mathbf{A}$ to make the system (1.19) solvable with respect to $W_1', \ldots, W_s'$. Denoting the elements of $\mathbf{A}^{-1}$ by $\hat{\alpha}_{ij}$, we obtain

$$\begin{aligned} w_\ell &= w_{\ell-1} + \sum_{j=1}^{s} \beta_j \sum_{k=1}^{s} \hat{\alpha}_{jk}(r(t_{\ell-1} + c_k h) - w_{\ell-1}) \\ &= \varrho w_{\ell-1} + \sum_{j=1}^{s} \beta_j \sum_{k=1}^{s} \hat{\alpha}_{jk} r(t_{\ell-1} + c_k h) \end{aligned}$$

with

$$\varrho = 1 - \beta^T \mathbf{A}^{-1}(1, \ldots, 1)^T. \tag{1.21}$$

Recall that w_ℓ should approximate $w(t_\ell) = r(t_\ell)$. Obviously, $|\varrho| > 1$ would yield an unstable scheme. Choosing $\beta_j = \alpha_{sj}$, $j = 1, \ldots, s$, in the Runge–Kutta tableau we obtain $\varrho = 0$ and $w_\ell = r(t_{\ell-1} + c_s h)$.

Thus, the so-called IRK (DAE) (cf. Petzold (1986), Griepentrog and März

(1986)), i.e. s-stage Runge–Kutta methods, with

$$\beta_j = \alpha_{sj}, \qquad j = 1, \ldots, s, \qquad c_s = 1, \qquad \mathbf{A} \text{ nonsingular}, \tag{1.22}$$

appears to be an appropriate tool for handling index-1 equations (1.1).

Next we investigate what happens with such a method if (1.5) contains an index-2 block (1.7). As earlier, we compute

$$w_{2,\ell} = s_2(t_{\ell-1} + h).$$

$$\begin{aligned} w_{1,\ell} &= s_1(t_{\ell-1} + h) - \frac{1}{h}\sum_{k=1}^{s} \hat{\alpha}_{s,k} s_2(t_{\ell-1} + c_k h) + \frac{1}{h}\sum_{k=1}^{s} \hat{\alpha}_{s,k} w_{2,\ell-1} \\ &= s_1(t_{\ell-1} + h) - \sum_{k=1}^{s} \hat{\alpha}_{s,k} \frac{1}{h}(s_2(t_{\ell-1} + c_k h) - s_2(t_{\ell-1})) \end{aligned} \tag{1.23}$$

assuming the starting value $w_{2,\ell-1}$ to be consistent, i.e. $w_{2,\ell-1} = s_2(t_{\ell-1})$.

If the Runge–Kutta method has an inner order of consistency ≥ 1 we know the condition

$$\sum_{k=1}^{s} \hat{\alpha}_{sk} c_k = 1$$

is satisfied. Thus, (1.23) with a consistent starting value actually provides an approximation of $w_1(t_\ell) = s(t_\ell) - s_2'(t_\ell)$. However, we do not usually have consistent starting values, and the errors are unstably amplified by $1/h$.

Let us summarize what has been pointed out in this section:

1 Singular systems (1.1) of index μ are mixed regular differential equations (1.4) and equations (1.5) including $\mu - 1$ differentiations.
2 Consistent initial values are not easy to compute in practice.
3 Integration methods handle the inherent regular ODE (1.4) as expected.
4 To avoid singular coefficient matrices in the linear systems to be solved per integration step we should use implicit multi-step methods ($\beta_0 \neq 0$) and nonsingular Runge–Kutta matrices $\mathbf{A}$. Moreover, there have to be additional conditions to ensure stability in the related index-1 parts.
5 Errors in the starting values are amplified by $h^{1-\mu}$ in the best case, but only the components v_j are affected.

The decoupled system (1.4), (1.5) and also (1.10), (1.11) respectively (1.15)–(1.18) lead us to the idea that it would be nice to allow different approaches for the parts (1.4) and (1.5), respectively, say a possibly explicit higher order method for the regular ODE (1.4) and a BDF for (1.5).

Of course, this should be done without knowing the canonical normal form. Furthermore, we regard the linear constant coefficient equation (1.1) as the simplest model with which to give some hints as to how to proceed with more general equations.

2. Characterizing DAEs

2.1. Linear equations with variable coefficients

Consider the linear equation

$$A(t)x'(t) + B(t)x(t) = q(t), \tag{2.1}$$

where $A(\cdot), B(\cdot) : \mathcal{I} \to L(\mathbb{R}^m)$ are continuous matrix functions on the interval $\mathcal{I} \subseteq \mathbb{R}$, and $A(t)$, $t \in \mathcal{I}$, is singular.

The first classification of these singular ODEs was given by C. W. Gear and L. Petzold (1984).

Definition (2.1) is said to be a *global index* μ DAE if there exist regular matrix functions $E \in C(\mathcal{I}, L(\mathbb{R}^m))$, $F \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ so that multiplying (2.1) by $E(t)$ and transforming $F(t)^{-1}x(t) = \tilde{x}(t)$ leads to the decoupled system

$$\begin{bmatrix} I & 0 \\ 0 & J \end{bmatrix} \tilde{x}'(t) + \begin{bmatrix} W(t) & 0 \\ 0 & I \end{bmatrix} \tilde{x}(t) = E(t)q(t), \tag{2.2}$$

where J is a constant nilpotent Jordan block matrix, $\mathrm{ind}(J) = \mu$.

Unfortunately, except for some interesting case studies, this *Kronecker canonical normal form* (2.2) as well as the transforms E, F are not available. Moreover, no way is known for relating this form to nonlinear equations. This is why we are looking for another way to characterize (2.1).

Denote by $N(t) := \ker A(t)$ the null space of $A(t)$, $t \in \mathcal{I}$, and assume this null space to be smooth, i.e. that there exists a matrix function $Q \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ which projects $\mathbb{R}^m$ onto $N(t)$ for each $t \in \mathcal{I}$ (that is $Q(t)^2 = Q(t)$, $\mathrm{im}Q(t) = N(t)$).

If DAE (2.1) has a global index μ, then e.g. $Q(t) = F(t)\mathrm{diag}(0, Q_J)F(t)^{-1}$ represents such a projector function, where Q_J denotes a projector onto $\ker J$. In particular, for global index-1 equations (2.1), we simply have $J = 0$, hence $Q(t) = F(t)\mathrm{diag}(0, I)F(t)^{-1}$.

In the following we let Q denote any such a projector function, and we also use $P(t) := I - Q(t)$, $t \in \mathcal{I}$.

Since $A(t)Q(t) \equiv 0$, we may insert $A(t) \equiv A(t)P(t)$ into (2.1), and rewrite it as

$$A(t)\{(Px)'(t) - P'(t)x(t)\} + B(t)x(t) = q(t)$$

or

$$A(t)(Px)'(t) + (B(t) - A(t)P'(t))x(t) = q(t). \tag{2.3}$$

This makes clear that, in general, we should not ask for C^1 solutions of (2.3) and (2.1), respectively, but for solutions belonging to the function space

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

Example $A(t) = \mathrm{diag}(I, 0)$ immediately leads to

$$\left.\begin{aligned} x_1'(t) + B_{11}(t)x_1(t) + B_{12}(t)x_2(t) &= q_1(t) \\ B_{21}(t)x_1(t) + B_{22}(t)x_2(t) &= q_2(t) \end{aligned}\right\}, \tag{2.4}$$

which is called a *semi-explicit DAE.*

Obviously, it is neither necessary nor useful that $x_2 \in C^1$!

Next we reformulate (2.3) to

$$A(Px)' + (B - AP')(Px + Qx) = q,$$

and then to

$$\{A + (B - AP')Q\}(P(Px)' + Qx) + (B - AP')Px = q. \tag{2.5}$$

Denote $A_1 := A + B_0Q$, $B_0 := B - AP'$ and ask whether $A_1(t)$ is nonsingular for all $t \in \mathcal{I}$. If it is so, we multiply (2.5) by PA_1^{-1} and QA_1^{-1}, respectively. This yields the system

$$\begin{aligned} (Px)' - P'Px + PA_1^{-1}B_0Px &= PA_1^{-1}q \qquad (2.6) \\ Qx + QA_1^{-1}B_0Px &= QA_1^{-1}q, \qquad (2.7) \end{aligned}$$

which decomposes into a regular explicit ODE for the nonnull space component Px and a simple derivative-free equation for determining the null space component Qx. The inherent ODE

$$u' - P'Pu + PA_1^{-1}B_0u = PA_1^{-1}q \tag{2.8}$$

has the property that solutions starting in $\mathrm{im}P(t_0)$ for some $t_0 \in \mathcal{I}$ remain in $\mathrm{im}P(t)$ for all $t \in \mathcal{I}$, since multiplying (2.8) by Q yields

$$(Qu)' - Q'Qu = 0.$$

Consequently, if for any $q \in C(\mathcal{I}, \mathbb{R}^m)$, $u_0 \in \mathrm{im}P(t_0)$, we denote the solution of (2.8) passing through (u_0, t_0) by $u \in C^1$, we obtain, with

$$\begin{aligned} x &:= u - QA_1^{-1}B_0u + QA_1^{-1}q \qquad (2.9) \\ &= (I - QA_1^{-1}B_0)u + QA_1^{-1}q, \end{aligned}$$

a C_N^1 solution of (2.1).

To be sure to address the initial condition to the respective component, we may state as follows

$$P(t_0)(x(t_0) - x^0) = 0. \tag{2.10}$$

This means that $u(t_0) = P(t_0)x(t_0) = P(t_0)x^0$, i.e. $P(t_0)x^0$ plays the role of u_0. Now $x^0 \in \mathbb{R}^m$ can be chosen arbitrarily. In general, $x(t_0) = x^0$ cannot be expected to hold for the solution $x(\cdot)$ of the initial value problem (IVP) (2.1), (2.10), but

$$x(t_0) = (I - Q(t_0)A_1(t_0)^{-1}B_0(t_0))P(t_0)x^0 + Q(t_0)A_1(t_0)^{-1}q(t_0).$$

Lemma 2.1 Let $A, B, Q \in L(\mathbb{R}^m)$ be given, $N := \ker A \neq \{0\}$, $Q^2 = Q$, $\mathrm{im}Q = N$, $S := \{z \in \mathbb{R}^m : Bz \in \mathrm{im}A\}$.

Then the following three statements are equivalent:

(i) $\mathbb{R}^m = N \oplus S$
(ii) $\mathrm{ind}(A, B) = 1$
(iii) $A + BQ$ is nonsingular.

Moreover, if $G := A + BQ$ is nonsingular, then $G^{-1}BQ = Q$, $G^{-1}A = I - Q$, and $QG^{-1}B$ represents the projection onto N along S.

Proof. The first part is given in Griepentrog and März (1986), Theorem A.13. Here we check the second part only.

Trivially, $G^{-1}BQ = G^{-1}(A + BQ)Q = Q$,

$$G^{-1}A = G^{-1}A(I - Q) = G^{-1}(A + BQ)(I - Q) = I - Q.$$

Then, we have for $Q_s := QG^{-1}B$

$$\begin{aligned} Q_s^2 &= QG^{-1}BQG^{-1}B = QG^{-1}B = Q_s, \\ Q_sQ &= QG^{-1}BQ = Q, \quad \text{i.e.} \quad \mathrm{im}Q_s = \mathrm{im}Q = N, \end{aligned}$$

and $Q_sz = 0$ implies

$$G^{-1}Bz = (I - Q)G^{-1}Bz,$$

thus $Bz = G(I - Q)G^{-1}Bz = AG^{-1}Bz \in \mathrm{im}A$. □

Lemma 2.1 applies to our DAE in the following sense.

In addition to $N(t) =: N_0(t)$ introduce

$$\begin{aligned} S_0(t) &:= \{z \in \mathbb{R}^m : B_0(t)z \in \mathrm{im}\, A(t)\} \\ &= \{z \in \mathbb{R}^m : B(t)z \in \mathrm{im}\, A(t)\}. \end{aligned} \tag{2.11}$$

By Lemma 2.1, our matrix $A_1(t)$ is nonsingular if and only if

$$S_0(t) \oplus N_0(t) = \mathbb{R}^m. \tag{2.12}$$

If (2.12) holds, then

$$Q_s(t) := Q(t)A_1(t)^{-1}B_0(t) \tag{2.13}$$

projects $\mathbb{R}^m$ onto $N_0(t)$ along $S_0(t)$.

Definition The DAE (2.1) is said to be *index-1 tractable* (or *transferable*) if A, B are continuous, $A(t)$ is singular but has a smooth null space, and $A_1(t)$ remains nonsingular for all $t \in \mathcal{I}$.

Theorem 2.2 Let (2.1) be transferable. Then

(i) For all $q \in C(\mathcal{I}, \mathbb{R}^m)$, $x^0 \in \mathbb{R}^m$, the IVP (2.1), (2.10) is uniquely solvable on $C_N^1(\mathcal{I}, \mathbb{R}^m)$.

(ii) $S_0(t_0)$ is the set of all consistent initial values at time $t_0 \in \mathcal{I}$ for the homogeneous equation, all IVPs $Ax' + Bx = 0$, $x(t_0) = x_0 \in S_0(t_0)$ are uniquely solvable.

Proof. It only remains to check the consistency of $x_0 \in S_0(t_0)$. In fact, solving the IVP $Ax' + Bx = 0$, $P(t_0)(x(t_0) - x_0) = 0$, $x_0 \in S_0(t_0)$, we derive

$$x(t_0) = (I - Q_s(t_0))P(t_0)x_0 = (I - Q_s(t_0))x_0 = x_0.$$

□

Remarks

1 The semi-explicit system (2.4) is transferable if $B_{22}(t)$ remains nonsingular. Here we simply have

$$Q = \mathrm{diag}(0, I), \quad A_1 = A + BQ = \begin{bmatrix} I & B_{12} \\ 0 & B_{22} \end{bmatrix}.$$

and, furthermore,

$$Q_s = \begin{bmatrix} 0 & 0 \\ B_{22}^{-1}B_{21} & I \end{bmatrix}.$$

2 Equation (2.2) in Kronecker canonical normal form is transferable if $J + Q_J$ is regular, that is if $\mu = 1$.

3 It may be easily checked whether each DAE (2.1) which has a global index $\mu = 1$ is also transferable, whereby even Q_s is continuously differentiable,

$$Q_s = F\mathrm{diag}(0, I)F^{-1}.$$

Theorem 2.3 Supposed (2.1) is transferable, the system

$$\left.\begin{aligned} A(t_0)y_0 + B(t_0)x_0 &= q(t_0) \\ Q(t_0)y_0 + P(t_0)(x_0 - x^0) &= 0 \end{aligned}\right\} \tag{2.14}$$

is uniquely solvable with respect to x_0, y_0. x_0 is the fully consistent initial value related to (2.1), (2.10), $y_0 = (Px)'(t_0) - P'(t_0)x_0$.

Proof. Rewrite the first equation of (2.14) as

$$A(t_0)\{y_0 + P'(t_0)x_0\} + B_0(t_0)x_0 = q(t_0).$$

Rearrange this as

$$A_1(t_0)\{P(t_0)y_0 + P(t_0)P'(t_0)x_0 + Q(t_0)x_0\} + B_0(t_0)P(t_0)x^0 = q(t_0).$$

Now we decouple into

$$\begin{aligned} P(t_0)y_0 + P(t_0)P'(t_0)x_0 + P(t_0)A_1(t_0)^{-1}B_0(t_0)P(t_0)x^0 \\ = P(t_0)A_1(t_0)^{-1}q(t_0) \end{aligned}$$

$$Q(t_0)x_0 + Q_s(t_0)P(t_0)x^0 = Q(t_0)A_1(t)^{-1}q(t_0)$$

and compare those with (2.6), (2.7), in order to obtain

$$\begin{aligned} x_0 &= x(t_0), \\ y_0 &= (Px)'(t_0) - P'(t_0)x(t_0). \end{aligned}$$

Finally, the matrix

$$\begin{bmatrix} A(t_0) & B(t_0) \\ Q(t_0) & P(t_0) \end{bmatrix} \tag{2.15}$$

is nonsingular since $A(t_0) + B(t_0)Q(t_0)$ is so. □

Now, let us turn to nontransferable DAEs (2.1), that is to those DAEs with a singular matrix $A_1(t)$.

Introduce new subspaces

$$\begin{aligned} N_1(t) &:= \ker A_1(t) \\ S_1(t) &:= \{z \in \mathbb{R}^m : B(t)P(t)z \in \mathrm{im}A_1(t)\} \\ &= \{z \in \mathbb{R}^m : B_0(t)P(t)z \in \mathrm{im}A_1(t)\}. \end{aligned} \tag{2.16}$$

and now assume that

$$N_1(t) \oplus S_1(t) = \mathbb{R}^m, \quad t \in \mathcal{I}$$

holds. Choose $Q_1(t)$ to be the projector onto $N_1(t)$ along $S_1(t)$, and let Q_1 be continuously differentiable. Note that for $B_1 := (B_0 - A_1(PP_1)')P$

$$S_1(t) = \{z \in \mathbb{R}^m : B_1(t)z \in \mathrm{im}(A_1(t))\}$$

holds. By Lemma 2.1, the matrix

$$A_2(t) := A_1(t) + B_1(t)Q_1(t), \quad t \in \mathcal{I}$$

becomes nonsingular and, finally,

$$Q_1(t) = Q_1(t)A_2(t)^{-1}B_1(t),$$

which implies that

$$Q_1(t)Q(t) = 0, \quad t \in \mathcal{I}, \tag{2.17}$$

is true. As a consequence, the products $P(t)P_1(t)$, $P(t)Q_1(t)$ are also projectors. Hence, it makes sense to look for a decomposition $x = PP_1x + PQ_1x + Qx$ of the solution. To this end, rewrite (2.1) again (cf. (2.5)) as

$$A_1(P(Px)' + Qx) + B_0Px = q,$$

then as

$$A_1\{(PP_1x)' + PQ_1(Px)' + Qx\} + (B_0 - A_1(PP_1)')Px = q$$

and finally as

$$A_2\{P_1(PP_1x)' + P_1PQ_1(Px)' + P_1Qx + Q_1x\} + B_1PP_1x = q. \tag{2.18}$$

Multiplying (2.18) by $PP_1A_2^{-1}$, $QP_1A_2^{-1}$ and $Q_1A_2^{-1}$, respectively, and performing some technical calculations we obtain the system

$$(PP_1x)' - (PP_1)'PP_1x + PP_1A_2^{-1}B_1PP_1x = PP_1A_2^{-1}q \tag{2.19}$$

$$\begin{aligned} -(QQ_1x)' + Qx &= QP_1A_2^{-1}q - (QQ_1)'PQ_1x \\ &\quad -\{(QQ_1 - QP_1)' + QP_1A_2^{-1}B_1\}PP_1x \qquad (2.20) \\ Q_1x &= Q_1A_2^{-1}q. \qquad (2.21) \end{aligned}$$

Clearly, (2.19) represents a regular ODE for the component PP_1x, (2.21) simply determines Q_1x, but to obtain the null space component Qx we have to insert $Q_1x = Q_1A_2^{-1}q$ into the term $(QQ_1x)'$, i.e. we have to differentiate $QQ_1A_2^{-1}q$ once.

Multiplying the ODE

$$u' - (PP_1)'u + PP_1A_2^{-1}B_1u = PP_1A_2^{-1}q \tag{2.22}$$

by $I - PP_1$ leads to $((I - PP_1)u)' + (PP_1)'(I - PP_1)u = 0$. Therefore, $u(t_0) \in \operatorname{im} P(t_0)P_1(t_0)$ implies $u(t) \in \operatorname{im} P(t)P_1(t)$ for all $t \in \mathcal{I}$.

Example Consider the semi-explicit DAE (2.4) with $B_{22}(t) \equiv 0$ and assume that $B_{21}(t)B_{12}(t)$ is nonsingular. We have

$$A_1(t) = \begin{bmatrix} I & B_{12}(t) \\ 0 & 0 \end{bmatrix},$$

$$S_1(t) = \left\{ \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^m : B_{21}(t)u = 0 \right\}.$$

Now $\binom{u}{v} \in N_1(t) \cap S_1(t)$ implies $u = -B_{12}(t)v$, $B_{21}(t)u = 0$, that is $v = 0$, $u = 0$. Then, compute

$$\begin{aligned} Q_1 &= \begin{bmatrix} B_{12}(B_{21}B_{12})^{-1}B_{21} & 0 \\ -(B_{21}B_{12})^{-1}B_{21} & 0 \end{bmatrix}, \\ PP_1 &= \begin{bmatrix} I - B_{12}(B_{21}B_{12})^{-1}B_{21} & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

It should be mentioned that this kind of equation is often discussed, and it is said to be an index-2 DAE in Hessenberg form. The simplest system of this type is (cf. (1.7))

$$\left.\begin{aligned} x_1' + x_2 &= q_1 \\ x_1 &= q_2 \end{aligned}\right\}.$$

Let us turn back to the general equation (2.1). If $A_2(t)$ is also singular,

we proceed analogously using new subspaces and projectors. More precisely, for given $A, B \in L(\mathcal{I}, L(\mathbb{R}^m))$, we define the chain of matrix functions

$$\begin{aligned} A_0 &:= A, \quad B_0 := B - AP', \\ A_{i+1} &:= A_i + B_i Q_i, \\ B_{i+1} &:= (B_i - A_{i+1}(P_0 P_1 \cdots P_{i+1})')P_i, \quad i \geq 0, \end{aligned} \tag{2.23}$$

where $P_j = I - Q_j$, and $Q_j(t)$ projects onto $N_j(t) := \ker A_j(t)$, $t \in \mathcal{I}$, $j \geq 0$. Introduce further

$$\begin{aligned} S_j(t) &:= \{z \in \mathbb{R}^m : B_j(t)z \in \mathrm{im} A_j(t)\} \\ &= \{z \in \mathbb{R}^m : B_{j-1}(t)P_{j-1}(t)z \in \mathrm{im} A_j(t)\}, \quad j \geq 1. \end{aligned}$$

Definition The ordered pair $\{A, B\}$ of continuous matrix functions (and also the DAE (2.1)) is said to be *index-μ tractable* if all matrices $A_j(t)$, $t \in \mathcal{I}$, $j = 0, \ldots, \mu - 1$, within the chain (2.23) are singular with smooth null spaces, and $A_\mu(t)$ remains nonsingular on $\mathcal{I}$.

Theorem 2.4 If the DAE (2.1) has the global index μ, then this DAE is also index-μ tractable.

Proof. We refer to Hansen (1990), where this assertion is verified by means of a very complicated induction. □

Theorem 2.5 Let $\{A, B\}$ be index-μ tractable. Then the IVP (2.1),

$$P_0(t_0) \ldots P_{\mu-1}(t_0)(x(t_0) - x^0) = 0 \tag{2.24}$$

is uniquely solvable on $C_N^1(\mathcal{I}, \mathbb{R}^m)$ for any given $x^0 \in \mathbb{R}^m$ and sufficiently smooth right-hand sides q, in particular for all $q \in C^{\mu-1}(\mathcal{I}, \mathbb{R}^m)$.

Proof. The assertion follows from the previous explanations for $\mu = 1$ and $\mu = 2$. It is proved in März (1989) for $\mu = 3$, and for $\mu > 3$ in Griepentrog and März (1989) and Hansen (1989). □

Remark The solution of an index-μ-tractable DAE, $\mu > 1$ decomposes in the following way:

$$x = P_0 \ldots P_{\mu-1}x + P_0 \ldots P_{\mu-2}Q_{\mu-1}x + \cdots + P_0 Q_1 x + Q_0 x.$$

Thereby $P_0 \ldots P_{\mu-1}x \in C^1$ solves the inherent regular ODE,

$$P_0 \ldots P_{\mu-2}Q_{\mu-1}x \in C^1$$

is given by the 'algebraic' part. The components $P_0 \ldots P_{\mu-j}Q_{\mu-j+1}x \in C^1$ include derivatives of order $j-2$ for $j = 3, \ldots, \mu$ and, finally, $Q_0 x \in C$ includes a $(\mu - 1)$ derivative.

When investigating discretizations we are often interested in compact intervals $\mathcal{I}$, and in the properties of the maps representing our IVPs and

BVPs. Let $\mathcal{I} := [t_0, T]$, and $\{A, B\}$ be index-μ tractable. Let $\Pi_\mu := P_0(t_0) \cdots P_{\mu-1}(t_0)$, $M_\mu := \mathrm{im}(\Pi_\mu) \subseteq \mathbb{R}^m$. Then introduce the linear map

$$\mathcal{L} : C_N^1(\mathcal{I}, \mathbb{R}^m) \to C(\mathcal{I}, \mathbb{R}^m) \times M_\mu =: C \times M_\mu$$

by defining

$$\mathcal{L}x := (A(Px)' + B_0 x, \Pi_\mu x(t_0)), \quad x \in C_N^1(\mathcal{I}, \mathbb{R}^m). \tag{2.25}$$

The function space $C_N^1(\mathcal{I}, \mathbb{R}^m)$ completed with its natural norm

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

becomes a Banach space. Note that the topology of this space is independent of the choice of projector function.

The map $\mathcal{L}$ is bounded, but does a bounded inverse exist?

Theorem 2.6 Let $\{A, B\}$ be index-μ tractable, $\mathcal{I} = [t_0, T]$. Then

(i) it holds that

$$\|x\| \le K \left\{ \sum_{j=0}^{\mu-1} \|q^{(j)}\|_\infty + |\Pi_\mu x(t_0)| \right\} \tag{2.26}$$

for all solutions x corresponding to sources $q \in C^{\mu-1}(\mathcal{I}, \mathbb{R}^m)$;

(ii) the map $\mathcal{L}$ is injective;

(iii) $\mathcal{L}$ is surjective for $\mu = 1$, but for $\mu > 1$ $\mathrm{im}(\mathcal{L})$ becomes a nonclosed proper subset within $C \times M_\mu$.

Proof. The first assertion is obvious for $\mu = 1$ and $\mu = 2$ (cf. (2.6), (2.7) respectively (2.19)–(2.21)). In general, it can be verified by decoupling the DAE (Griepentrog and März 1989, Hansen 1989).

The injectivity of $\mathcal{L}$ is given by Theorem 2.5. Moreover, for $\mu = 1$, Theorem 2.2 provides solvability for all continuous right-hand sides q, i.e. $\mathrm{im}(\mathcal{L}) = C \times M_\mu$.

In the higher index cases, that is for $\mu \ge 2$, we have to assume that certain components of q are continuously differentiable for solvability. However, the set of these functions is not closed in the continuous function space, but it is a nonclosed proper subset. □

Remarks

1 Inequality (2.26) is somewhat liberal. It could be stated more strictly but would take immense technical effort. To do this by means of the decoupling technique, those parts of q which have to be differentiated have to be described precisely. In particular for $\mu = 2$, the system (2.19)–(2.21) makes this transparent. There we have

$$\mathrm{im}(\mathcal{L}) = \{q \in C : Q_1 A_2^{-1} q \in C^1\} \times M_2,$$

and

$$\|x\| \le \tilde{K}\{\|q\|_\infty + \|(Q_1 A_2^{-1} q)'\|_\infty + |\Pi_2 x(t_0)|\}. \tag{2.27}$$

2 The inequalities

$$\|x\|_\infty \le \bar{K} \left\{ \sum_{j=1}^{\mu-1} \|q^{(j)}\|_\infty + |x(t_0)| \right\} \tag{2.28}$$

are used in Hairer *et al.* (1989) to define the so-called perturbation index. In our framework, (2.28) as well as its sharper form (2.26) appear as secondary effects.

Corollary 2.7 If $\mu = 1$, then the inverse of $\mathcal{L}$ is bounded, and $\mathcal{L}$ is a homeomorphism. If $\mu > 1$, then the inverse of $\mathcal{L}$ becomes unbounded.

Proof. Since $\mathcal{L}$ is acting in Banach spaces, this assertion follows immediately from Theorem 2.6. □

In other words, higher index DAEs ($\mu > 1$) become *ill posed* in Tichonov's sense in the given setting, i.e. the solutions do not depend continuously on the inputs. This has bad consequences for the numerical treatment. The unboundedness of $\mathcal{L}^{-1}$ makes the related discretized maps unstable.

At this point it should be recalled that the explanations in Section 1 concerning integration methods confirm the expected instability. On the other hand, in certain cases they cause us to be optimistic as they are only weak instabilities and we are to be able to handle them.

We conclude this section by emphasizing once more that the described decoupling of (2.1) should be understood as an appropriate technique for analysing large classes of DAEs and the precise behaviour of numerical methods.

It should also be possible to compute the projectors $Q_j(t)$ and matrices $A_i(t)$ at certain points t in order to formulate the initial conditions and organize a numerical index-testing. However, in general the decoupling technique is not aimed at representing a numerical method.

2.2. DAEs as vector fields on manifolds

The most frequently used notion of an index of a general nonlinear DAE

$$f(x', x, t) = 0 \tag{2.29}$$

is the *differentiation index*, which goes back to the work of S.L. Campbell on linear DAEs with smooth coefficients (e.g. Campbell (1987), Brenan *et al.* (1989)).

Assuming f and the respective solutions to be smooth enough we form

the system

$$\left.\begin{aligned} f(x',x,t) &= 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}f(x',x,t) &= \frac{\partial}{\partial x'}f(x',x,t)x'' + \cdots = 0 \\ &\vdots \\ \frac{\mathrm{d}^{\mu}}{\mathrm{d}t^{\mu}}f(x',x,t) &= \frac{\partial}{\partial x'}f(x',x,t)x^{(\mu+1)} + \cdots = 0 \end{aligned}\right\} \tag{2.30}$$

by differentiating μ times. Consider (2.30) as a system in separate dependent variables $x', x'', \ldots, x^{(\mu+1)}$ with x, t as independent variables.

Definition The DAE (2.29) has the *differentiation index* μ if there exists an integer μ such that system (2.30) can be solved for $x' = H(x,t)$, H continuously differentiable, and μ is the smallest integer having this property.

We do not recommend carrying out this procedure in order to obtain the underlying regular ODE $x' = H(x,t)$ in practice. This ODE does not give a good reflection of the qualitative behaviour of the original equation.

Example (Führer and Leimkuhler, 1989) The inherent regular ODE of the DAE

$$\left.\begin{aligned} x_1' - x_2 + ax_1^2 &= 0 \\ x_2 - ax_1^2 &= 0 \end{aligned}\right\} \tag{2.31}$$

is $x_1' = 0$, and the origin represents a stable equilibrium (all solutions are stationary here). By differentiating once we formally obtain the system

$$\left.\begin{aligned} x_1' - x_2 + ax_1^2 &= 0 \\ x_2 - ax_1^2 &= 0 \\ x_1'' - x_2' + 2ax_1x_1' &= 0 \\ x_2' - 2ax_1x_1' &= 0 \end{aligned}\right\},$$

which leads to

$$\left.\begin{aligned} x_1' &= x_2 - ax_1^2 \\ x_2' &= 2ax_1(x_2 - ax_1^2) \end{aligned}\right\}, \tag{2.32}$$

but now the origin is no longer stable.

System (2.30) suggests the idea of defining a compound function or a derivative array

$$F_{\mu}(\bar{y}_{\mu},x,t) := \begin{bmatrix} f(y_1,x,t) \\ \dfrac{\partial}{\partial x'}f(y_1,x,t)y_2 + \cdots \\ \vdots \\ \dfrac{\partial}{\partial x'}f(y_1,x,t)y_{\mu+1} + \cdots \end{bmatrix}, \tag{2.33}$$

where $\bar{y}_\mu := (y_1^T, \ldots, y_{\mu+1}^T)^T \in \mathbb{R}^{(\mu+1)m}$.

If we assume the Jacobian $H_\mu(\bar{y}_\mu, x, t) := \partial F_\mu(\bar{y}_\mu, x, t)/\partial \bar{y}_\mu$ has constant rank, we can form the constraint manifold of order μ

$$S_\mu := \{(x,t) \in \mathbb{R}^m \times \mathbb{R} : F_\mu(\bar{y}_\mu, x, t) = 0 \text{ for a } \bar{y}_\mu \in \mathbb{R}^{(\mu+1)m}\}$$

as well as

$$\begin{aligned} M_\mu(x,t) &:= \{\bar{y}_\mu \in \mathbb{R}^{(\mu+1)m} : F_\mu(\bar{y}_\mu, x, t) = 0\}, \\ M_\mu^1(x,t) &:= \{y_1 \in \mathbb{R}^m : \bar{y}_\mu \in M_\mu(x,t)\} \text{ for } (x,t) \in S_\mu. \end{aligned}$$

Definition (Griepentrog, 1991) $f \in C^{\mu+1}(\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m)$ is called an *index-μ mapping* if S_μ is nonempty and $M_\mu(t,x)$ is a singleton for all $(t,x) \in S_\mu$, and if μ is the smallest integer with these properties.

Clearly, DAE (2.29) has the differentiation index μ if f is an index-μ mapping. However, now it becomes transparent that this DAE represents a vector field defined on S_μ, namely

$$v(x,t) := y_1 \in M_\mu^1(x,t) \quad \text{for } (x,t) \in S_\mu.$$

By definition, $f(v(x,t), x, t) = 0$ holds for $(x,t) \in S_\mu$. The solution of each IVP

$$x'(t) = v(x(t), t), \quad (x(t_0), t_0) \in S_\mu \tag{2.34}$$

evolves in the manifold and solves the DAE. More precisely, the following assertion is proved in Griepentrog (1991).

Theorem 2.8 If (2.29) is an index-μ DAE, then all solutions proceed in the differentiable constraint manifold S_μ. A vector field $v(x,t)$ is defined on S_μ and has the following properties:

(i) v is continuously differentiable; and
(ii) the solutions of (2.29) are identical with the solutions of the IVPs (2.34).

Remark Griepentrog (1991) describes both the manifold S_μ and the vector field v in detail by means of the rank theorem. These investigations are closely related to the differential-geometric concepts of regular DAEs in Reich (1990) and Rabier and Rheinboldt (1991). However, these studies are still in an early phase, but they are very promising and are aimed at making the results of differential geometry applicable to numerical treatment.

Return shortly to the trivial example (2.31). Now, it appears to be an index-1 equation, whereby

$$S_1 := \{(x,t) \in \mathbb{R}^2 \times \mathbb{R} : x_2 - a x_1^2 = 0\},$$

and $M_1^1(x,t) = \{0\}$ for all $(x,t) \in S_1$.

We would like to direct attention to an essential detail of this index definition as well as of Theorem 2.8, namely the condition that the Jacobian of the

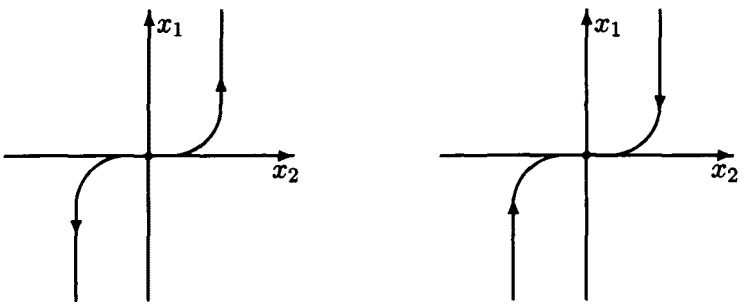
compound function (2.33) has constant rank. If this property is lost, different singularities may arise, as is illustrated by the next two easy examples, which model certain *RC* circuits (Chua and Deng, 1989).

Examples Consider the DAEs

$$x_1' - x_2 = 0, \quad x_1 - x_2^3 = 0 \tag{2.35}$$

and

$$x_1' + x_2 = 0, \quad x_1 - x_2^3 = 0. \tag{2.36}$$



In both cases the Jacobian $H_1(\bar{y}_1, x, t) = \partial F_1(\bar{y}_1, x, t)/\partial \bar{y}_1$ has constant rank 3 for $x_2 \neq 0$, but it suffers from a rank deficiency at $x_2 = 0$. In any case, the origin becomes a stationary solution. Besides the trivial solution, (2.35) has the solution $x_1(t) = (\frac{2}{3}t)^{3/2}$, $x_2(t) = (\frac{2}{3}t)^{1/2}$, which starts at the origin. On the other hand, (2.36) has no solutions coming out of the origin, but $x_1(t) = (1 - \frac{2}{3}t)^{3/2}$, $x_2(t) = (1 - \frac{2}{3}t)^{1/2}$ starts at $(1, 1)$ for $t = 0$, and ends, for $t = \frac{3}{2}$, at the origin.

2.3. Many open questions are left

What do the differentiation index and the tractability index have to do with each other. At first glance seemingly nothing. Let us consider the matter in the case of homogeneous linear DAEs in Kronecker canonical normal form (1.4), (1.5). In this case we obtain

$$F_1(y_1, y_2, x, t) = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & J & 0 & 0 \\ W & 0 & I & 0 \\ 0 & I & 0 & J \end{bmatrix} \begin{bmatrix} u' \\ v' \\ u'' \\ v'' \end{bmatrix} + \begin{bmatrix} W & 0 \\ 0 & I \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix},$$

where

$$y_1 =: \begin{bmatrix} u' \\ v' \end{bmatrix} \qquad y_2 =: \begin{bmatrix} u'' \\ v'' \end{bmatrix} \qquad x =: \begin{bmatrix} u \\ v \end{bmatrix},$$

$$S_1 := \{(x, t) \in \mathbb{R}^m \times \mathbb{R} : v \in \mathrm{im} J\},$$

and $M_1^1(x,t)$ is a singleton if and only if

$$Jv' = 0, \quad v' + Jv'' = 0$$

implies $v' = 0$, that is $J = 0$, ind$J = 1$.

In general, for linear DAEs (2.1), the different index notations are related to different smoothness requirements with respect to the coefficients A, B, q; however, they are identical in essence.

Theorem 2.9 Each linear DAE (2.1) with a differentiation index μ_{D} also has a global Kronecker normal form index $\mu_{\mathrm{K}} = \mu_{\mathrm{D}}$. Each DAE (2.1) having a global index μ_{K} is tractable with index $\mu_{\mathrm{T}} = \mu_{\mathrm{K}}$.

For the technically expensive proof we refer to Hansen (1990), Griepentrog and März (1989) and Griepentrog (1991).

For linear equations, the concept of index-μ tractability seems to be the most general one. But, how is the index-μ tractability to be defined for nonlinear DAEs? First, this concept is based upon another notion of solution, which can also be reasonably applied to nonlinear equations (2.29) under certain assumptions.

Assumption 2.10 Let the function $f \in C(\mathcal{G}, \mathbb{R}^m)$, where $\mathcal{G} = \mathbb{R}^m \times \mathcal{D} \times \mathcal{I} \subseteq \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}$ is an open set and $f_y'(y,x,t), f_x'(y,x,t) \in L(\mathbb{R}^m)$ exist for all $(y,x,t) \in \mathcal{G}$, and $f_y', f_x' \in C(\mathcal{G}, L(\mathbb{R}^m))$.

Suppose that the null space of $f_y'(y,x,t)$ is independent of (y,x), i.e.

$$N(t) := \ker f_y'(y,x,t), \quad (y,x,t) \in \mathcal{G}. \tag{2.37}$$

Let $N(t)$ be smooth in t. Let $Q \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ denote the corresponding projector function onto N and set $P := I - Q$.

In most applications we know, the null space $N(t)$ is kept constant.

Due to Assumption 2.10, the identity

$$\begin{aligned} & f(y,x,t) - f(P(t)y,x,t) \\ &= \int_0^1 f_y'(sy + (1-s)P(t)y, x, t)Q(t)ds = 0, \qquad (y,x,t) \in \mathcal{G} \end{aligned}$$

becomes true. Consequently, (2.29) may be rewritten as

$$f((Px)'(t) - P'(t)x(t), x(t), t) = 0, \tag{2.38}$$

hence the function space to which the solutions of (2.29) should belong again appears to be

$$C_N^1(\mathcal{I}_0, \mathbb{R}^m) := \{x \in C(\mathcal{I}_0, \mathbb{R}^m) : Px \in C^1(\mathcal{I}_0, \mathbb{R}^m)\},$$

where $\mathcal{I}_0 \subseteq \mathcal{I}$ is a certain interval.

This space seems to be very natural. If $x_* \in C_N^1(\mathcal{I}_0, \mathbb{R}^m)$ is any given

function whose trajectory remains in $\mathcal{G}$, then the equation linearized along $x_*(t)$ has continuous coefficients

$$\begin{aligned} A_*(t) &= f_y'(\zeta(t)), \quad B_*(t) = f_x'(\zeta(t)), \\ \zeta(t) &:= ((Px_*)'(t) - P'(t)x_*(t), x_*(t), t) \in \mathcal{G}, \end{aligned}$$

and the null space of $A_*(t)$ is again $N(t)$.

It should also be mentioned that the given nontrivial solutions of the examples (2.35) and (2.36) do not belong to

$$C^1([0,\infty), \mathbb{R}^2),$$

but to

$$C_N^1([0,\infty), \mathbb{R}^2).$$

Does it make sense to define the notion of index-μ tractability via linearization?

Definition Suppose (2.29) satisfies Assumption 2.10, and $x_* \in C_N^1(\mathcal{I}_0, \mathbb{R}^m)$ is given, $\mathcal{T}_* := \{\zeta(t) : t \in \mathcal{I}_0\} \subset \mathcal{G}$. The DAE (2.29) is said to be *transferable or index-1 tractable* around x_* if the pair $\{A_*, B_*\}$ is index-1 tractable.

Lemma 2.11 $\{A_*, B_*\}$ becomes index-1 tractable if and only if the matrix

$$G(y,x,t) := f_y'(y,x,t) + f_x'(y,x,t)Q(t) \tag{2.39}$$

remains nonsingular for all (y,x,t) from a neighbourhood $\mathcal{N} \subseteq \mathcal{G}$ of $\mathcal{T}_*$.

Proof. Let $\{A_*, B_*\}$ be index-1 tractable, that means that $A_1 := A_* + (B_* - A_*P')Q$ is nonsingular, then $G_* := A_* + B_*Q = A_1 + A_*P'Q = A_1(I + PP'Q)$ is also nonsingular. Next, $G(y,x,t)$ becomes nonsingular for all $(y,x,t) \in \mathcal{T}_*$, because of

$$G_*(t) = G(\zeta(t)), \quad t \in \mathcal{I}.$$

Since G depends continuously on its arguments, there is a neighbourhood $\mathcal{N}$ of $\mathcal{T}_*$ where $G(y,x,t)$ remains nonsingular. Now the assertion is evident. □

Theorem 2.12 Let $x_* \in C_N^1([t_0,T], \mathbb{R}^m)$ solve the DAE (2.29). Let (2.29) be transferable around x_*. Then, for any given $q \in C([t_0,T], \mathbb{R}^m)$, $x^0 \in \mathbb{R}^m$ the IVP

$$\left.\begin{aligned} f(x'(t), x(t), t) &= q(t) \\ P(t_0)(x(t_0) - x^0) &= 0 \end{aligned}\right\} \tag{2.40}$$

is uniquely solvable on $C_N^1([t_0,T], \mathbb{R}^m)$, provided that $\|q\|_\infty$ and $|P(t_0)(x^0 - x_*(t_0))|$ are sufficiently small.

Moreover,

$$\|x - x_*\| \le K\{\|q\|_\infty + |P(t_0)(x^0 - x_*(t_0))|\}$$

is valid with constant K.

Proof. Denote shortly $C_N^1 := C_N^1([t_0,T],\mathbb{R}^m)$, $C := C([t_0,T],\mathbb{R}^m)$. For $x \in C_N^1$, $q \in C$, $\beta \in \mathrm{im}(P(t_0))$ with

$$\|x - x_*\| \le \varrho, \quad \|q\|_\infty \le \varrho, \quad |\beta - \beta_*| \le \varrho,$$
$$\beta_* := P(t_0)x_*(t_0), \quad \varrho > 0 \quad \text{sufficiently small,}$$

we define the map $\mathcal{F}$ by

$$\mathcal{F}(x,q,\beta) := (f((Px)'(\cdot) - P'(\cdot)x(\cdot), x(\cdot), \cdot) - q(\cdot), P(t_0)x(t_0) - \beta).$$

$\mathcal{F}$ maps a ball within $C_N^1 \times C \times \mathrm{im}P(t_0)$ into $C \times \mathrm{im}P(t_0)$, it is continuously differentiable and, in particular for $z \in C_N^1$,

$$\mathcal{F}'_x(x_*,0,\beta_*)z = (A_*((Pz)' - P'z) + B_*z, P(t_0)z(t_0)).$$

holds. Trivially, $\mathcal{F}(x_*,0,\beta_*) = 0$. Due to Corollary 2.7 (cf. also Theorem 2.2), $\mathcal{F}'_x(x_*,0,\beta_*)$ is a homeomorphism, hence it remains to apply the Implicit Function Theorem. □

Remarks

1 To obtain the fully consistent initial value $x_0 := x(t_0)$ related to the IVP (2.40) the system

$$\left.\begin{aligned} f(y_0,x_0,t_0) - q(t_0) &= 0 \\ P(t_0)(x_0 - x^0) + Q(t_0)y_0 &= 0 \end{aligned}\right\}$$

will be helpful (cf. Theorem 2.3). The Jacobian of this system is nonsingular because of the index-1 requirement.

2 Let $(y_0,x_0,t_0) \in \mathcal{G}$ be given, and let $G(y_0,x_0,t_0)$ be nonsingular. Rewrite $f(y,x,t) = f(w,u + Q(t)w,t) =: \tilde{f}(w,u,t)$ where new variables $w = P(t)y + Q(t)x$, $u = P(t)x$ are introduced.

Clearly, since $\tilde{f}(w_0,u_0,t_0) = 0$, and $\tilde{f}'_w(w_0,u_0,t_0) = G(y_0,x_0,t_0)$ is nonsingular, due to the Implicit Function Theorem there exists a continuous function $w(u,t)$ with continuous partial Jacobian $w'_u(u,t)$, satisfying $\tilde{f}(w,u,t) = 0$. Then, it is easy to check that (cf. (2.9))

$$x(t) = u(t) + Q(t)w(u(t),t) \tag{2.41}$$

represents a solution of $f(x',x,t) = 0$ passing through (x_0,t_0), whereby u denotes the solution of the inherent regular IVP

$$\begin{aligned} u'(t) - P'(t)u(t) &= P(t)(I + P'(t))w(u(t),t), \\ u(t_0) &= P(t_0)x_0. \end{aligned} \tag{2.42}$$

3 Supposing that $G(y,x,t)$ remains nonsingular for all $(y,x,t) \in \mathcal{G}$, we immediately know that

$$S_1 := \{(x,t) \in \mathcal{D} \times \mathcal{I} : f(y,x,t) = 0 \text{ for a } y \in \mathbb{R}^m\}$$

is the manifold of consistent initial values. Comparing this with the matters related to understanding DAEs as vector fields on manifolds we feel the smoothness demanded there to be difficult to realise.

4 In our examples (2.35), (2.36) we compute $G(y,x,t)$ to be equal to

$$\begin{bmatrix} 1 & -1 \\ 0 & -3x_2^2 \end{bmatrix} \quad \text{respectively} \quad \begin{bmatrix} 1 & 1 \\ 0 & -3x_2^2 \end{bmatrix}.$$

This shows that the transferability matrix may be used as a tool for detecting singularities numerically.

Unfortunately, the situation becomes much more complicated for higher indexes. According to Theorem 2.6 and Corollary 2.7 linear DAEs with an index-μ-tractable coefficient pair $\{A_*, B_*\}$, $\mu > 1$, result in ill posed IVPs, i.e. they have discontinuous inverse mappings in the given topologies. Now, the standard arguments used in Theorem 2.12 no longer apply because the derivative $\mathcal{F}_x'(x_*, 0, \beta_*)$ does not have a continuous inverse.

By means of the following example we want to elucidate that Theorem 2.12 cannot be saved for the index-2 case even if $P(t_0)P_1(t_0)$ is appropriately used instead of $P(t_0)$ (cf. Theorem 2.5), and if only q from C^1 is admissible.

Example (Chua and Deng, 1989; März, 1991) Consider the system

$$x_1' = x_3^2, \quad x_2' = -x_3, \quad x_3^3 + x_2 x_3 + x_1 = 0. \tag{2.43}$$

$x_*(t) = (2(\frac{t}{6}+1)^3, -3(\frac{t}{6}+1)^2, \frac{t}{6}+1)^T$ solves this DAE, and

$$A_*(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_*(t) = \begin{bmatrix} 0 & 0 & -\frac{1}{3}t - 2 \\ 0 & 0 & 1 \\ 1 & \frac{1}{6}t + 1 & 0 \end{bmatrix}$$

form an index-2-tractable pair $\{A_*, B_*\}$. However, e.g.,

$$x(t) = (2+t, -3-t, 1)^T$$

represents another solution, and $x_*(0) = x(0)$ holds, i.e. certain bifurcation phenomena arise. The respective matrix (2.39)

$$G(y,x,t) = \begin{bmatrix} 1 & 0 & -2x_3 \\ 0 & 1 & 1 \\ 0 & 0 & x_2 + 3x_3^2 \end{bmatrix}$$

is nonsingular for $x_2 + 3x_3^2 \neq 0$. Thus, equation (2.43) represents an index-1 DAE everywhere, where $x_2 + 3x_3^2 \neq 0$ holds. The whole thing should be understood as an index-1 DAE with a singularity at $x_2 + 3x_3^2 = 0$.

We might possibly overcome these problems by making the following definition: The nonlinear DAE (2.29) is called index-μ tractable around x_* if for *all* $x \in \{\tilde{x} \in C_N^1 : \|\tilde{x} - x_*\| < \varrho\}$, ϱ sufficiently small, the respective pairs $\{A, B\}$ are index-μ tractable. However, how can this be checked?

Finally, we are also interested in conditions that can be treated numerically, such as those provided by Lemma 2.11 for instance. So far, statements have only been successfully made for index-2 equations (cf. Lemma 3.5) and for special index-3 equations (e.g. März, 1989). In these cases it has also been successfully proved that the differentiation index and the tractability index coincide identically except for smoothness.

On the other hand, the approach of considering higher index DAEs as differential equations on manifolds seems to be easier to grasp and handle. In particular, this is true for DAEs with a special structure, e.g. those of Hessenberg form. In this respect, interesting results are to be expected. However, a uniform analysis of DAEs with natural smoothness is still out of sight.

3. Numerical integration methods

3.1. General remarks on the BDF

The integration method used most frequently for regular as well as for singular implicit equations

$$f(x'(t), x(t), x(t)) = 0 \tag{3.1}$$

is the BDF. It is well known that there are powerful codes like DASSL (cf. Brenan *et al.* (1989)) which treat large classes of DAEs well.

On the other hand, the following example shows that BDFs may fail even in very simple cases. Thus, in this section we try to clarify the related problems together with possible ways out.

Example The DAE

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & \eta t & 1 \\ 0 & 0 & 0 \end{bmatrix} x'(t) + \begin{bmatrix} 1 & 0 & 0 \\ 0 & \eta + 1 & 0 \\ 0 & \eta t & 1 \end{bmatrix} x(t) = q(t) \tag{3.2}$$

has global index-3 for all parameter values $\eta \in \mathbb{R}$. The leading coefficient matrix also has constant null space and constant image space. Table 1 shows results generated by BDFs with different constant step-sizes h and consistent starting values for parameter values $\eta = 0, -0.5$ and 2.0, respectively.

The exact solution is

$$x_1(t) = \mathrm{e}^{-t} \sin t, \quad x_2(t) = \mathrm{e}^{-2t} \sin t, \quad x_3(t) = \mathrm{e}^{-t} \cos t,$$

and $[0, 0.1]$ is the integration interval. The absolute errors at $t = 0.1$ arising in the components of the solution belonging to the null space $\ker A(t)$ and the other one are given separately.

Note that for $\eta = 0$ we have a linear constant coefficient equation as discussed in Section 1. Obviously, the null space component is particularly

Table 1

	$\eta = -0.5$		$\eta = 0$		$\eta = 2.0$	
$h \approx$	Px	Qx	Px	Qx	Px	Qx
BDF_2						
2.5-2	3.-2	3.+0	3.-4	4.-3	1.-4	5.-2
3.1-3	3.+7	2.+10	5.-6	6.-5	1.-5	2.-5
7.8-4	1.+43	3.+46	3.-7	4.-6	1.-6	2.-6
3.9-4	—	—	8.-8	1.-6	2.-7	5.-7
1.9-4	—	—	2.-8	2.-7	6.-8	1.-7
9.7-5	—	—	5.-9	9.-9	1.-8	3.-8
BDF_3						
2.5-2	—	—	1.-5	2.-2	1.-3	8.-2
3.1-3	—	—	3.-8	1.-7	3.-4	1.-2
7.8-4	—	—	4.-10	2.-9	1.-1	1.+2
3.9-4	—	—	5.-11	1.-8	3.+3	4.+6
1.9-4	—	—	5.-12	3.-8	6.+12	2.+16
9.7-5	—	—	9.-13	1.-7	1.+32	1.+36
BDF_6						
3.1-3	—	—	2.-13	3.-9	—	—
7.8-4	—	—	5.-13	2.-8	—	—
3.9-4	—	—	4.-12	3.-8	—	—
1.9-4	—	—	2.-12	2.-7	—	—
9.7-5	—	—	4.-12	4.-6	—	—

($\sigma = 1.\text{-}16$)

affected by round-off errors. Furthermore, order expectations do not become true in practice.

Before we investigate the BDF applied to DAEs (3.1) we describe this class of DAEs in more detail. Assume DAE (3.1) satisfies Assumption 2.10 and, in particular,

$$N(t) := \ker f_y'(y, x, t), \quad (y, x, t) \in \mathcal{G}, \tag{3.3}$$

Let $Q \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ denote the corresponding projector function onto N, $P := I - Q$. Recall that Assumption 2.10 allows equation (3.1) to be rewritten as

$$f((Px)'(t) - P'(t)x(t), x(t), t) = 0, \tag{3.4}$$

thus the function space to which the solutions of (3.4) should belong again appears to be

$$C_N^1(\mathcal{I}_0, \mathbb{R}^m) := \{x \in C(\mathcal{I}_0, \mathbb{R}^m) : Px \in C^1(\mathcal{I}_0, \mathbb{R}^m)\},$$

where $\mathcal{I}_0 \subseteq \mathcal{I}$.

We now ask how integration methods approximate solutions of (3.4) and (3.1). For this purpose, we assume here that solutions exist, say on the interval $\mathcal{I}_0 = [t_0, T]$. However, it should be mentioned once more that a comprehensive analysis of nonlinear DAEs is only in its infancy. Very interesting problems remain to be solved. In particular, solvability is closely related to the description of the set of consistent initial values.

Assume $x_* \in C_N^1 := C_N^1([t_0, T], \mathbb{R}^m)$ solves DAE (3.1). Let $\mathcal{B}(x_*, \varrho_0) \subseteq C_N^1$ denote a small ball around x_* within C_N^1 such that $x(t) \in \mathcal{D}$ for $t \in [t_0, T]$, and for all $x \in \mathcal{B}(x_*, \varrho_0)$. Introduce the map

$$\mathcal{F} : \mathcal{B}(x_*, \varrho_0) \subseteq C_N^1 \to C := C([t_0, T])$$

by means of

$$\begin{aligned}(\mathcal{F}x)(t) := f((Px)'(t) - P'(t)x(t), x(t), t),\\ t \in [t_0, T],\ x \in \mathcal{B}(x_*, \varrho_0).\end{aligned} \tag{3.5}$$

The map is continuously differentiable; and its Frechet derivative at x_* is given by

$$\begin{aligned}(\mathcal{F}'(x_*)z)(t) := A_*(t)((Pz)'(t) - P'(t)z(t)) + B_*(t)z(t),\\ t \in [t_0, T],\ z \in C_N^1,\end{aligned} \tag{3.6}$$

where

$$\begin{aligned}A_*(t) &:= f_y'(\zeta(t)),\ B_*(t) := f_x'(\zeta(t)),\\ \zeta(t) &:= ((Px_*)'(t) - P'(t)x_*(t), x_*(t), t).\end{aligned}$$

Let the interval $[t_0, T]$ be partitioned by

$$\pi : t_0 < t_1 < \cdots < t_N = T.$$

Denote by h, $\underline{h}$ the maximal and minimal step-sizes of π, respectively, and $h_j := t_j - t_{j-1}$. Given starting values $x_0, \ldots, x_{s-1}$, we apply the variable step-size BDF to (3.1), i.e.

$$f\left(\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_{j-i}, x_j, t_j\right) = 0, \quad j = s, \ldots, N, \tag{3.7}$$

expecting x_j to become an approximation of the true solution value $x_*(t_j)$.

Introduce the map

$$\mathcal{F}_\pi z := \begin{bmatrix} z_0 - x_0 \\ z_{s-1} - x_{s-1} \\ f\left(\frac{1}{h_s}\sum_{i=0}^{s}\alpha_{si}z_{s-i}, z_s, t_s\right) \\ \vdots \\ f\left(\frac{1}{h_N}\sum_{i=0}^{s}\alpha_{Ni}z_{N-i}, z_N, t_N\right) \end{bmatrix},$$

$$z \in \mathbb{R}^{m(N+1)}, \quad |z_j - x_*(t_j)| < \varrho_0, \quad j = 0, 1, \ldots, N,$$

which represents the discretized map corresponding to the BDF. $\mathcal{F}_\pi$ acts within $\mathbb{R}^{m(N+1)}$. Then denote

$$x_\pi^* = \begin{bmatrix} x_*(t_0) \\ \vdots \\ x_*(t_N) \end{bmatrix} \in \mathbb{R}^{m(N+1)}$$

and compute the Jacobian

$$\mathcal{F}_\pi'(x_\pi^*) = \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ \frac{\alpha_{ss}}{h_s}A_s^* & \cdots & \frac{\alpha_{s1}}{h_s}A_s^* & F_s^* & & \\ & \ddots & & & \ddots & \\ & & \frac{a_{Ns}}{h_N}A_N^* & \cdots & \frac{a_{N1}}{h_N}A_N^* & F_N^* \end{bmatrix} \in L(\mathbb{R}^{m(N+1)}),$$

whereby

$$A_j^* := f_y'(\eta_j), \qquad F_j^* := \frac{\alpha_{jo}}{h_j}A_j^* + f_x'(\eta_j),$$

and

$$\eta_j := \left(P(t_j)\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_*(t_{j-i}), x_*(t_j), t_j\right).$$

Complete $\mathbb{R}^{m(N+1)}$ with respect to the norms

$$\|z\|_\infty := \max\{|z_i| : i = 0, 1, \ldots, N\},$$

$$\|z\|_\pi := \|z\|_\infty + \max\left\{\left|\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}P(t_{j-i})z_{j-i}\right| : j = s, \ldots, N\right\},$$

which are consistent with the norms of C and C_N^1, respectively. We then

use the matrix norms

$$\begin{aligned} \|G\|_\pi &:= \max\{\|Gz\|_\pi : \|z\|_\infty = 1\}, \\ \|G\|_\pi &:= \max\{\|Gz\|_\infty : \|z\|_\pi = 1\}, \qquad G \in L(\mathbb{R}^{m(N+1)}). \end{aligned}$$

Let $\mathcal{B}_\pi(x_\pi^*, \varrho) := \{z \in \mathbb{R}^{m(N+1)} : \|z - x_\pi^*\|_\pi < \varrho\}$.

In the following we use grids π belonging to a given grid class Π, e.g. the class of locally uniform grids with given constants c_1, c_2, $h_{\max}$, such that $c_1 h_{j-1} \le h_j \le c_2 h_{j-1}$, $h \le h_{\max}$, for all j and all $\pi \in \Pi$. The smallest grid class in which we are interested is the set Π_{equ} of all sufficiently fine equidistant grids; however we always assume $\Pi_{\text{equ}} \subseteq \Pi$.

Definition The BDF (3.7) is stable for (3.1) on grid class Π if there exist constants $S > 0$, $\varrho > 0$ such that for arbitrary $\pi \in \Pi$ the inequality

$$\|z - \bar{z}\|_\pi \le S\|\mathcal{F}_\pi z - \mathcal{F}_\pi \bar{z}\|_\infty \tag{3.8}$$

is satisfied for all $z, \bar{z} \in \mathcal{B}_\pi(x_\pi^*, \varrho)$.

Definition The BDF (3.7) is weakly unstable for (3.1) on Π if the inequality

$$\|z - \bar{z}\|_\pi \le S\underline{h}^{-\gamma}\|\mathcal{F}_\pi z - \mathcal{F}_\pi \bar{z}\|_\infty \tag{3.9}$$

is valid for all $z, \bar{z} \in \mathcal{B}_\pi(x_\pi^*, \varrho_\pi)$, $\pi \in \Pi$, where $S > 0$, $\gamma > 0$ are constants but $\varrho_\pi > 0$ may depend on the chosen grid π; γ is said to be the order of instability.

Then, introduce the *local discretization error* $\tau_\pi := \mathcal{F}_\pi x_\pi^*$. Clearly, its first components τ_j, $j = 0, \dots, s-1$, represent the errors in the starting values, but

$$\tau_j = f\left(\frac{1}{h_j}\sum_{i=0}^{s} \alpha_{ji} x_*(t_{j-i}), x_*(t_j), t_j\right) \qquad \text{for } j = s, \dots, N.$$

Surely, the point of interest is the so-called *global error*

$$\varepsilon_\pi := x_\pi^* - x_\pi$$

where $x_\pi \in \mathbb{R}^{m(N+1)}$ consists of the components $x_0, x_1, \dots, x_N \in \mathbb{R}^m$. Note that $\varepsilon_j = \tau_j$ for $j = 0, 1, \dots, s-1$.

Recall some standard arguments from discretization theory (e.g. Keller (1975)), which we apply and modify, appropriately.

1 First of all, if the x_j, $j \ge s$ in (3.7) exist and $x_\pi \in \mathcal{B}_\pi(x_\pi^*, \varrho)$, then stability implies the error estimate

$$\|\varepsilon_\pi\|_\pi \le S\|\tau_\pi\|_\infty,$$

hence

$$\max_{j \ge 0} |x_*(t_j) - x_j| \le S\{\max_{j \le s-1} |x_*(t_j) - x_j| + \max_{j \ge s} |\tau_j|\}. \tag{3.10}$$

2 Then, it is sufficient for stability that there exists a uniform bound S_1,

$$\|(\mathcal{F}'_\pi(x^*_\pi))^{-1}\|_\pi \le S_1, \quad \pi \in \Pi. \tag{3.11}$$

3 Assuming stability, for sufficiently small $h_{\max}$ and ϱ, the equation

$$\mathcal{F}_\pi z = 0$$

has exactly one solution x_π on $\mathcal{B}_\pi(x^*_\pi, \varrho)$, which can be computed by the Newton method.

4 (3.11) can be proved by permuting linearization and discretization, and using the Banach lemma.

Under which conditions do these standard arguments remain valid when the BDF is applied to DAEs?

In Section 1 we have learnt that, in higher index cases, some weak instabilities should be expected. Is it possible to carry over these standard arguments then? How can weak instabilities be distinguished?

3.2. On the BDF applied to linear DAEs

In any case, the behaviour of the BDF applied to linear DAEs plays a crucial role. This is why we investigate this question in more detail. It should not be surprising that stability and instability, respectively, depend on the index of the DAE. In the following we will point out that certain time-dependent subspaces are also responsible for exponential instabilities.

Let us turn to the special case when the BDF is applied to linear DAEs, that is

$$A(t_j)\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_{j-i} + B(t_j)x_j = q(t_j), \qquad j \ge s, \tag{3.12}$$

where starting values $x_0, \ldots, x_{s-1}$ are given. For the local error τ_ℓ we now derive

$$\begin{aligned}\tau_j &= A(t_j)\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_*(t_{j-i}) + B(t_j)x_*(t_j) - q(t_j)\\ &= A(t_j)\left\{\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_*(t_{j-i}) - (Px_*)'(t_j) + P'(t_j)x_*(t)\right\};\end{aligned} \tag{3.13}$$

consequently, the local error belongs to a subspace,

$$\tau_j \in \mathrm{im}A(t_j), \quad j \ge s,$$

which is characteristic of DAEs, and we will make use of this later.

To obtain x_j for $j \ge s$, we have to solve the linear system

$$F_j x_j = A(t_j)\frac{1}{h_j}\sum_{i=1}^{s}\alpha_{ji}x_{j-i} + q(t_j), \tag{3.14}$$

with the coefficient matrix

$$F_j := \frac{1}{h_j}\alpha_{jo}A(t_j) + B(t_j). \tag{3.15}$$

Is this matrix F_j nonsingular? Clearly, nonsingularity is necessary for the BDF to work well. Note that for $\eta = -1$ in (3.2), all related coefficient matrices F_j are singular. However, as example (3.2) also shows, even if the BDF becomes formally feasible, i.e. if all matrices F_j are nonsingular, the BDF may fail.

Equation $\mathcal{F}_\pi x = 0$, which represents the BDF, now has the special form $\mathcal{F}_\pi x = \mathcal{L}_\pi x - q_\pi$, where

$$\mathcal{L}_\pi := \begin{bmatrix} I & & & & & \\ & \ddots & & & & \\ & & I & & & \\ \frac{\alpha_{ss}}{h_s}A_s & \cdots & \frac{\alpha_{s1}}{h_s}A_s & F_s & & \\ & \ddots & & \ddots & \ddots & \\ & & \frac{\alpha_{Ns}}{h_N}A_N & \cdots & \frac{\alpha_{N1}}{h_N}A_N & F_N \end{bmatrix} \in L(\mathbb{R}^{m(N+1)}), \tag{3.16}$$

$A_j := A(t_j)$, $j \geq 0$, $(q_\pi)_j := x_j$ for $j = 0, \ldots, s-1$, $(q_\pi)_j = q(t_j)$ for $j \geq s$.

Clearly, $\mathcal{L}_\pi$ is nonsingular if $F_s, \ldots, F_N$ are. Moreover, in the case of linear DAEs, the stability inequality (3.8) simplifies to

$$\|\mathcal{L}_\pi^{-1}\|_\pi \leq S. \tag{3.17}$$

It should be mentioned that $\mathcal{L}_\pi$ is to be understood as a discretized map $\mathcal{L}$ given in (2.25). By Corollary 2.7, we may expect $\mathcal{L}_\pi$ to be stable for index-1-tractable DAEs only.

Next we are going to prove the characteristic inequalities of $\mathcal{L}_\pi^{-1}$.

Case 1: Assume (2.1) to be transferable (index-1 tractable).
Recall from Section 2 that $A_1 := A + (B - AP')Q$ becomes a nonsingular matrix function. In addition, by Lemma 2.1, $G := A + BQ$ also remains nonsingular; and furthermore $G = A_1 + AP'Q = A_1(I + PP'Q)$. First of all, it is easy to check that F_j is nonsingular if

$$I + \alpha_{jo}\frac{1}{h_j}P(t_j)G(t_j)^{-1}B(t_j) =: H_j$$

is, i.e. at least for small h_j. Then we compute

$$F_j^{-1} = \left\{Q(t_j) + \frac{1}{\alpha_{jo}}h_j P_s(t_j)H_j^{-1}P(t_j)\right\}G(t_j)^{-1}, \tag{3.18}$$

$\mathrm{cond}(F_j) \sim h_j^{-1}$.

Since $F_s, \ldots, F_N$ are nonsingular, so is $\mathcal{L}_\pi$. Next, by decoupling $\mathcal{L}_\pi z = w$ in a similar way as (2.1) in Section 2, we generate a uniform bound for $\mathcal{L}_\pi$ on an appropriate grid class Π. $\mathcal{L}_\pi z = w$ means in detail that $z_j = w_j$, $j = 0, \ldots, s-1$, and

$$A(t_j)\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}z_{j-i} + B(t_j)z_j = w_j, \qquad j = s, \ldots, N. \tag{3.19}$$

Denote $u_j := P(t_j)z_j$, $v_j := Q(t_j)z_j$. Multiply (3.19) by $P(t_j)A_1(t_j)^{-1}$ and $Q(t_j)A_1(t_j)^{-1}$, respectively. This yields

$$\left.\begin{aligned} \frac{1}{h_j}\sum_{i=o}^{s}\alpha_{ji}P(t_j)(u_{j-i}+v_{j-i}) + (PA_1^{-1}B)(t_j)u_j &= (PA_1^{-1})(t_j)w_j \\ v_j + Q_s(t_j)u_j &= (QA_1^{-1})(t_j)w_j. \end{aligned}\right\} \tag{3.20}$$

Clearly, if the projector function P is constant, then this formula fits the system (2.6), (2.7) very well. In particular, the first equation simplifies to

$$\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}Pu_{j-i} + (PA_1^{-1}B)(t_j)u_j = (PA_1^{-1})(t_j)w_j.$$

For $w = q_\pi$, this is in fact the same expression we would obtain by applying the BDF to the regular ODE inherent in (2.1)

$$u' + PA_1^{-1}Bu = PA_1^{-1}q.$$

If P' does not vanish, there arises some additional feedback between the components in (3.20). Because

$$\begin{aligned} P(t_j)(u_{j-i}+v_{j-i}) &= u_{j-i} + (P(t_j) - P(t_{j-i}))(u_{j-i}+v_{j-i}) \\ &= u_{j-i} + \int_0^1 P'(t_{j-i} + s(t_j - t_{j-i}))\mathrm{d}s(t_j - t_{j-i})(u_{j-i}+v_{j-i}) \end{aligned}$$

we are able to rearrange the first equation in (3.20) to

$$\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}u_{j-i} + \sum_{i=1}^{s}D_{ji}(u_{j-i}+v_{j-i}) + (PA_1^{-1}B)(t_j)u_j = (PA_1^{-1})(t_j)w_j,$$

where the coefficient matrices are uniformly bounded.

Theorem 3.1 Let the given variable step-size BDF applied to a regular explicit ODE be stable on the grid class Π. Let the DAE (2.1) be index-1 tractable. Then, there is a bound S such that $\mathcal{L}_\pi$ is bijective, and

$$\|\mathcal{L}_\pi^{-1}\|_\infty \le \|\mathcal{L}_\pi^{-1}\|_\pi \le S \quad \text{for all } \pi \in \Pi. \tag{3.21}$$

Proof. By standard arguments we easily obtain

$$\max_{j\ge s}|u_j| \le S_1 \max_{j\ge 0}|w_j|,$$

therefore

$$\max_{j\geq s}|v_j| \leq S_2 \max_{j\geq 0}|w_j|$$

and

$$\max_{j\geq s}|\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}u_{j-i}| \leq S_3 \max_{j\geq 0}|w_j|.$$

□

Note that Theorem 3.1 implies the error estimation

$$\max_{j\geq s}|x_*(t_j)-x_j| \leq S\left\{\max_{j\leq s-1}|x_*(t_j)-x_j| + \max_{j\geq s}|\tau_j|\right\}, \tag{3.22}$$

which is well known in the case of regular ODEs.

Case 2: Assume (2.1) to be index-2-tractable.
We begin this part by quoting the nice linear index-2 DAE from Gear and Petzold (1984), which was constructed to illustrate the instability of Euler's backward rule.

Example The DAE

$$\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix} x'(t) + \begin{bmatrix} 1 & \eta t \\ 0 & 1+\eta \end{bmatrix} x(t) = q(t) \tag{3.23}$$

has the global index-2 for all parameter values $\eta \in \mathbb{R}$. Compute here (cf. (2.19)–(2.21))

$$\begin{aligned} Q(t) &= \begin{bmatrix} 0 & -\eta t \\ 0 & 1 \end{bmatrix}, \quad A_1(t) = \begin{bmatrix} 0 & 0 \\ 1 & 1+\eta t \end{bmatrix} \\ Q_1(t) &= \begin{bmatrix} 1+\eta t & \eta t(1+\eta t) \\ -1 & -\eta t \end{bmatrix}, \quad P(t)P_1(t) = 0. \end{aligned}$$

The backward Euler rule applied to (3.23) gives for $\eta \neq -1$

$$\begin{aligned} x_{1,j} &= q_1(t_j) - \eta t_j x_{2,j}, \\ x_{2,j} &= \frac{\eta}{1+\eta}x_{2,j-1} + \frac{1}{1+\eta}\left\{q_2(t_j) - \frac{1}{h_j}(q_1(t_j) - q_1(t_{j-1}))\right\}, \end{aligned}$$

but the exact solution is

$$\begin{aligned} x_1(t) &= q_1(t) - \eta t x_2(t), \\ x_2(t) &= q_2(t) - q_1'(t). \end{aligned}$$

Careful further investigation will reveal that the backward Euler rule for this problem is weakly unstable but convergent if $\eta > -0.5$, and exponentially unstable for all $\eta < -0.5$, $\eta \neq -1$. For $\eta = -1$ the backward Euler rule does not work at all.

In the following we exclude such situations where the behaviour of a numerical method depends essentially on parameter values, all of which belong to the same category, by restricting the class of DAEs (2.1) to those with constant $\ker A(t)$ and $P' = 0$, respectively.

Let us turn back to the BDF applied to (2.1), that is to formula (3.12). The first problem to be solved is the nonsingularity of F_j given by (3.15). The question may be answered by the use of the decoupling technique described in Section 2. Supposing that $P' = 0$, and

$$\tilde{H}(t_j) := I + h_j \frac{1}{\alpha_{jo}} (PP_1 G_2^{-1} B)(t_j)$$

becomes nonsingular (which happens at least for small h_j), the matrix F_j will also be nonsingular, and

$$F_j^{-1} = \left(\left\{ QP_1 + P_1 Q_1 + \alpha_{jo} \frac{1}{h_j} QQ_1 + h_j \frac{1}{\alpha_{jo}} (I - QP_1 G_2^{-1} B) \tilde{H}^{-1} PP_1 \right\} G_2^{-1} \right)(t_j) \tag{3.24}$$

$\mathrm{cond}(F_j) \sim h_j^{-2}$.

Expression (3.24) is evaluated in März (1990, Lemma 3.1). Thereby, $G_2 := A_1 + B_0 P Q_1$ is used instead of A_2 in Section 2. Due to Lemma 2.1 (cf. (2.16)), both A_2 and G_2 are nonsingular simultaneously. More precisely, $A_2 = G_2(I - P_1(PP_1)'Q_1)$, $(I - P_1(PP_1)'Q_1)^{-1} = (I + P_1(PP_1)'Q_1)$ are valid.

It should be mentioned that the term QQ_1 within (3.24) does not vanish principally as a matter of index-2 tractability. This is true independently of possible special structural forms of the DAE itself. However, if the DAE has a special form, e.g. Hessenberg form, then, employing the special structure, we can look for an appropriate scaling of F_j.

Next we decompose the system $\mathcal{L}_\pi z = w$ (cf. (3.16), (3.19)) to gain information about $\mathcal{L}_\pi^{-1}$, once again using the projector technique. Multiplying (3.19) by $(PP_1 G_2^{-1})(t_j)$, $(QP_1 G_2^{-1})(t_j)$ and $(Q_1 G_2^{-1})(t_j)$, respectively, we derive

$$\frac{1}{h_j} PP_1(t_j) \sum_{i=0}^{s} \alpha_{ji} z_{j-i} \quad + \quad PP_1(t_j) G_2(t_j)^{-1} B(t_j) PP_1(t_j) z_j = PP_1(t_j) G_2(t_j)^{-1} w_j, \tag{3.25}$$

$$-\frac{1}{h_j} QQ_1(t_j) \sum_{i=0}^{s} \alpha_{ji} z_{j-i} + Qz_j \quad + \quad QP_1(t_j) G_2(t_j)^{-1} B(t_j) PP_1(t_j) z_j = QP_1(t_j) G_2(t_j)^{-1} w_j, \tag{3.26}$$

$$Q_1(t_j) z_j = Q_1(t_j) G_2(t_j)^{-1} w_j, \tag{3.27}$$

for $j \geq s$. Recall that $z_j = w_j$ for $j = 0, \ldots, s-1$. Note that we also use P,

Q here as constant projector matrices. Inserting

$$\frac{1}{h_j}PP_1(t_j)\sum_{i=0}^{s}\alpha_{ji}z_{j-i} = \frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}PP_1(t_{j-i})z_{j-i}+$$

$$+ \sum_{i=0}^{s}\alpha_{ji}P\frac{1}{h_j}(P_1(t_j)-P_1(t_{j-i}))\{PP_1(t_{j-i})z_{j-i}+PQ_1(t_{j-i})z_{j-i}\}$$

in equation (3.25), and taking into account (3.27), we are able to prove the inequality

$$\max_{j\geq 0}|PP_1(t_j)z_j| \leq S_1 \max_{j\geq 0}|w_j|$$

by standard arguments. Trivially,

$$\max_{j\geq 0}|Q_1(t_j)z_j| \leq S_2 \max_{j\geq 0}|w_j|$$

also becomes true due to (3.27). Moreover, (3.26), (3.27) yield

$$\begin{aligned} Qz_j &= \frac{1}{h_j}QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}\{PP_1(t_{j-i})+PQ_1(t_{j-i})\}z_{j-i} \\ &\quad -QP_1(t_j)G_2(t_j)^{-1}B(t_j)PP_1(t_j)z_j + QP_1(t_j)G_2(t_j)^{-1}w_j \\ &= \frac{1}{h_j}QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}Q_1(t_{j-i})G_2(t_{j-i})^{-1}\tilde{w}_{j-i} \qquad (3.28) \\ &\quad +\sum_{i=0}^{s}\alpha_{ji}Q\frac{1}{h_j}(Q_1(t_j)-Q_1(t_{j-i}))PP_1(t_{j-i})z_{j-i} \\ &\quad -QP_1(t_j)G_2(t_j)^{-1}B(t_j)PP_1(t_j)z_j + QP_1(t_j)G_2(t_j)^{-1}w_j, \end{aligned}$$

for $j \geq s$, where we introduce, for more convenience,

$$\begin{aligned} \tilde{w}_j &:= w_j && \text{for } j \geq s, \\ \tilde{w}_j &:= G_2(t_j)w_j && \text{for } j \leq s-1. \end{aligned} \qquad (3.29)$$

Now, we can estimate

$$|Qz_j| \leq \frac{1}{h_j}|QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}Q_1(t_{j-i})G_2(t_{j-i})^{-1}\tilde{w}_{j-i}| + S_3\max_{j\geq 0}|w_j|. \qquad (3.30)$$

Since (3.19) immediately implies

$$\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}Pz_{j-i} = PA(t_j)^{+}\{-B(t_j)z_j+w_j\}, \qquad j \geq s, \qquad (3.31)$$

it follows that

$$\|z\|_\pi \leq S_4\left\{\max_{j\geq 0}|w_j| + \max_{j\geq s}\frac{1}{h_j}\left|QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}Q_1(t_{j-i})G_2(t_{j-i})^{-1}\tilde{w}_{j-i}\right|\right\} \qquad (3.32)$$

becomes valid.

Consider now expression (3.14) again. Solving this equation in practice, instead of the values x_j, only certain $\tilde{x}_j$ satisfying

$$F_j\tilde{x}_j = -A(t_j)\frac{1}{h_j}\sum_{i=1}^{s}\alpha_{ji}\tilde{x}_{j-i} + q(t_j) + \delta_j, \qquad j \geq s, \tag{3.33}$$

where $\tilde{x}_j := x_j$, $j = 0, \ldots, s-1$, are generated. The δ_j represent round-off errors (but also errors that arise later when solving nonlinear equations).

Then, if we put $z_j = x_*(t_j) - \tilde{x}_j$, $j \geq 0$, $w_j = \tau_j - \delta_j$ results for $j \geq s$, and $w_j = x_*(t_j) - x_j$ for the starting phase $j = 0, \ldots, s-1$. Because $\tau_j \in \mathrm{im}A(t_j)$ (cf. (3.13)) we obtain for all $k \geq s$

$$\begin{aligned} Q_1(t_k)G_2(t_k)^{-1}\tilde{w}_k &= Q_1(t_k)G_2(t_k)^{-1}w_k \\ &= Q_1(t_k)G_2(t_k)^{-1}(\tau_k - \delta_k) \\ &= Q_1(t_k)G_2(t_k)^{-1}(A(t_k)A(t_k)^{+}\tau_k - \delta_k). \end{aligned}$$

However, on the other hand,

$$\begin{aligned} G_2^{-1}A &= G_2^{-1}(A + (B - AP')Q)P = G_2^{-1}A_1P \\ &= G_2^{-1}(A_1 + BPQ_1)P_1P = P_1P \end{aligned}$$

holds, thus $Q_1G_2^{-1}A = Q_1P_1P = 0$.

Consequently,

$$Q_1(t_k)G_2(t_k)^{-1}\tilde{w}_k = -Q_1(t_k)G_2(t_k)^{-1}\delta_k \tag{3.34}$$

for $k \geq s$, which appears to be characteristic of these DAEs.

Finally, collect this result in

Theorem 3.2 Let the given BDF applied to regular explicit ODEs become stable on the grid class Π. Let the DAE (2.1) be index-2 tractable, and, additionally, let $P' = 0$.

(i) Then $\mathcal{L}_\pi$ is bijective, and

$$\|\mathcal{L}_\pi^{-1}\|_\infty \leq \|\mathcal{L}_\pi^{-1}\|_\pi \leq S\underline{h}^{-1}, \qquad \pi \in \Pi, \tag{3.35}$$

is true with a certain constant $S > 0$.

(ii) The following precise error estimates hold:

$$\max_{j\geq s}|P(x_*(t_j) - \tilde{x}_j| \leq S_P\{\max_{j\leq s-1}|P(x_*(t_j) - x_j)| + \max_{j\geq s}|\tau_j - \delta_j|\} \tag{3.36}$$

and

$$\begin{aligned} &|Q(x_*(t_j) - \tilde{x}_j)| \leq \\ &\leq S_Q\left\{\max_{j\leq s-1}|P(x_*(t_j) - x_j)| + \max_{j\geq s}|\tau_j - \delta_j|\right\} + \end{aligned}$$

$$+\frac{1}{h_j}\left|QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}Q_1(t_{j-i})G_2(t_{j-i})^{-1}\tilde{\delta}_{j-i}\right|, \qquad (3.37)$$

where $\tilde{\delta}_j := G_2(t_j)(x_*(t_j) - x_j)$, $j = 0, \ldots, s-1$, reflect the errors in the starting values, and

$$\tilde{\delta}_j := \delta_j \quad \text{for } j \geq s.$$

Remarks

1 If exact values $x_j = x_*(t_j)$, $j = 0, \ldots, s-1$, are used in the starting phase, (3.36), (3.37) immediately imply, for $\delta_j = 0$, $j \geq s$,

$$\max_{j\geq 0}|x_*(t_j) - x_j| \leq \bar{S}\max_{j\geq s}|\tau_j|, \qquad (3.38)$$

hence the BDF converges formally with the expected order. However, practical computations cannot be managed in such a way that all $\tilde{\delta}_j$ vanish in reality.

2 Expression (3.28) shows that, for small h_j, Qz_j behaves in fact mainly as

$$\frac{1}{h_j}QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}Q_1(t_{j-i})G_2(t_{j-i})^{-1}\tilde{w}_{j-i}.$$

In this sense, (3.37) and (3.35) cannot be improved. We really have to deal with a weak instability. Fortunately, this instability does not affect the nonnull space components at all (cf. (3.36)).

Case 3: Assume (2.1) to be index-3 tractable.
First recall our example (3.2) to illustrate that a restriction to the class of DAEs with constant null space $\ker A(t)$ will not do. It may be checked that, in (3.2), $(PQ_1Q_2)'(t)$ does not vanish identically. This seems to be the crucial point in the index-3 case. As in the previous parts we consider $\mathcal{L}_\pi$ given by (3.16).

Recall that (cf. (3.16), (3.33))

$$\mathcal{L}_\pi x_\pi = q_\pi := \begin{bmatrix} x_0 \\ \vdots \\ x_{s-1} \\ q(t_s) \\ \vdots \\ q(t_N) \end{bmatrix}, \quad \mathcal{L}_\pi \tilde{x}_\pi - q_\pi = \delta_\pi := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \delta_s \\ \vdots \\ \delta_N \end{bmatrix}$$

$$\mathcal{L}_\pi x_\pi^* - q_\pi = \tau_\pi := \begin{bmatrix} x_*(t_0) - x_0 \\ \vdots \\ x_*(t_{s-1}) - x_{s-1} \\ \tau_s \\ \vdots \\ \tau_N \end{bmatrix},$$

thus $\mathcal{L}_\pi(x_\pi^* - \tilde{x}_\pi) = \tau_\pi - \delta_\pi$.

By the use of this projector technique we decompose (3.19) respectively $\mathcal{L}_\pi z = w$. Now we multiply (3.19) by $PP_1P_2A_3^{-1}$, $QP_1P_2A_3^{-1}$, $Q_1P_2A_3^{-1}$ and $Q_2A_3^{-1}$, respectively. We omit these straigthforward but very extensive evaluations here and mention only that we now have to insert expressions given by the BDF into each other twice to approximate first and second derivatives. This is why the BDF on the whole becomes active just for $j \geq 2s$. The first s steps have to be analysed separately. Let us formulate the results:

Theorem 3.3 Let the given BDF applied to regular explicit ODE become stable on the grid class Π. Let the DAE (2.1) be index-3 tractable and, in addition, let $P' = 0$, $(PQ_1Q_2)' = 0$.

(i) Then $\mathcal{L}_\pi$ is bijective,

$$\|\mathcal{L}_\pi^{-1}\|_\infty \leq \|\mathcal{L}_\pi^{-1}\|_\pi \leq S\underline{h}^{-2}, \qquad \pi \in \Pi, \tag{3.39}$$

holds with a certain constant S.

(ii) The following detailed error estimates become true with

$$\omega_\pi := \max_{j \geq s} |\tau_j - \tilde{\delta}_j| + \max_{j \leq s-1} |x_*(t_j) - x_j|, \quad \text{for } j \geq 2s:$$

$$|PP_1(t_j)(x_*(t_j) - \tilde{x}_j)| \leq S_1\omega_\pi, \tag{3.40}$$

$$|PQ_1(t_j)(x_*(t_j) - \tilde{x}_j)| \leq S_2\omega_\pi \quad + \quad \frac{1}{h_j}\left|\sum_{i=0}^{s} \alpha_{ji} PQ_1Q_2A_3(t_{j-i})^{-1}\tilde{\delta}_{j-i}\right| \tag{3.41}$$

and

$$\begin{aligned} |Q(x_*(t_j) - \tilde{x}_j| \leq\ & S_3\omega_\pi + \\ & + \frac{1}{h_j}\left|\sum_{i=0}^{s} \alpha_{ji}(QQ_1P_2A_3^{-1})(t_{j-i})(\tau_{j-i} - \tilde{\delta}_{j-i})\right| \\ & + \gamma\frac{1}{h_j}\left|\sum_{i=0}^{s} \alpha_{ji}\frac{1}{h_{j-i}}\sum_{k=0}^{s} \alpha_{j-i-k} PQ_1Q_2A_3(t_{j-i-k})^{-1}\tilde{\delta}_{j-i-k}\right| \end{aligned} \tag{3.42}$$

Thereby, $\gamma := \max |QQ_1(t)|$, and

$$\begin{aligned} \tilde{\delta}_j &:= \delta_j \quad \text{for } j \geq s, \\ \tilde{\delta}_j &:= A_3(t_j)(x_*(t_j) - x_j) \quad \text{for } j \leq s-1. \end{aligned}$$

Remarks

1 Now, the worst error sensitivity has order h_j^{-2}. It is again somewhat local and belongs to the null space component only.

2 Putting $\tilde{\delta}_j = 0$ in (3.40)–(3.42), i.e. using exact starting values and computing without any round-off error, (3.40), (3.41) provide for $j \geq 2s$

$$|P(x_*(t_j) - x_j)| \leq S_4 \max_{j \geq s} |\tau_j|, \tag{3.43}$$

but

$$\begin{aligned} |Q(x_*(t_j) - x_j)| \quad &\leq \quad S_3 \max_{i \geq s} |\tau_j| + \\ &+ \quad \frac{1}{h_j} |\sum_{i=0}^{s} \alpha_{ji} (QQ_1P_2A_3^{-1})(t_{j-i})\tau_{j-i}|. \end{aligned} \tag{3.44}$$

The last term in (3.44) is also troublesome. It reflects the new quality of the problem of index-3-tractable DAEs.

In constant step-size computations, the local error is smooth if the solution $x_*(t)$ itself is smooth enough. In this case, we again have $|x_*(t_j) - x_j| = \mathcal{O}(h^s)$, $j \geq 2s$. However, step-size changes, and similarly the first s steps raise difficulties. In particular, the variable step backward Euler method does not converge since

$$\begin{aligned} &\frac{1}{h_j}((QQ_1P_2A_3^{-1})(t_j)\tau_j - (QQ_1P_2A_3^{-1})(t_{j-1})\tau_{j-1} \approx \\ &\approx -\frac{1}{2h_j}(QQ_1Q_2)(t_j)(Px_*)''(t_j)(h_j - h_{j-1}). \end{aligned}$$

Unfortunately, the backward Euler method also fails to provide accurate starting values.

3.3. On the BDF applied to nonlinear DAEs

Now, having provided information on $\mathcal{L}_\pi$ we continue the investigation of nonlinear DAEs started previously. In the following, we understand $\mathcal{L}_\pi$ to be related to the equation linearized in the solution x_*, i.e. (cf. (3.6))

$$\mathcal{L}_\pi = \mathcal{F}'(x_*)_\pi.$$

In other words, $\mathcal{L}_\pi$ represents the discretization of the linearization. On the other hand, $\mathcal{F}'_\pi(x_\pi^*)$ is the derivative of the discretized map $\mathcal{F}_\pi$ at x_π^*.

Compute

$$\mathcal{L}_\pi - \mathcal{F}'_\pi(x^*_\pi) = \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ \frac{\alpha_{ss}}{h_s}D_s & \cdots & \frac{\alpha_{s1}}{h_s}D_s & K_s & & \\ & \ddots & & \ddots & \ddots & \\ & & \frac{\alpha_{Ns}}{h_N}D_N & & \frac{\alpha_{N1}}{h_N}D_N & K_N \end{bmatrix}, \quad (3.45)$$

where

$$\begin{aligned} D_j &:= A_*(t_j) - A^*_j = f'_y(\zeta(t_j)) - f'_y(\eta_j), \\ K_j &:= \frac{\alpha_{jo}}{h_j}A_*(t_j) + B_*(t_j) - \frac{\alpha_{jo}}{h_j}A^*_j - f'_x(\eta_j) \\ &= \frac{\alpha_{jo}}{h_j}(f'_y(\zeta(t_j) - f'_y(\eta_j)) + f'_x(\zeta(t_j) - f'_x(\eta_j)), \\ \zeta(t) &:= ((Px_*)'(t) - P'(t)x_*(t), x_*(t), t), \\ \eta_j &:= (P(t_j)\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x^*_{j-i}, x^*_j, t_j), \quad x^*_\ell := x_*(t_\ell). \end{aligned}$$

Further, we have for $j \geq s$

$$((\mathcal{L}_\pi - \mathcal{F}'_\pi(x^*))z)_j = D_j\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}z_{j-i} + (f'_x(\zeta(t_j)) - f'_x(\eta_j))z_j,$$

and $D_j = D_jP(t_j)$,

$$\begin{aligned} D_j\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}z_{j-i} &= D_j\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}P(t_{j-i})z_{j-i} + \\ &\quad + D_j\sum_{i=0}^{s}\alpha_{ji}\frac{1}{h_j}(P(t_j) - P(t_{j-i}))z_{j-i}. \end{aligned}$$

Consequently

$$\|(\mathcal{L}_\pi - \mathcal{F}'_\pi(x^*_\pi))z\|_\infty \leq \gamma_\pi\|z\|_\pi \quad (3.46)$$

with a grid-dependent value

$$\gamma_\pi := c_0 \max_{j\geq s}\|D_j\| + \max_{j\geq s}\|f'_x(\zeta(t_j) - f'_x(\eta_j)\|,$$

where c_0 denotes a certain constant. Clearly, γ_π becomes small for refined grids.

Note that for all quasi-linear DAEs of the form

$$A(t)x'(t) + g(x(t), t) = 0 \quad (3.47)$$

we just obtain

$$\mathcal{L}_\pi = \mathcal{F}'_\pi(x^*_\pi). \tag{3.48}$$

In other words, the BDF discretization and linearization commute asymptotically in general, but for special DAEs (3.47) they commute exactly.

Next, supposing $\mathcal{F}'_\pi(x^*_\pi)$ to be bijective, we turn to the question as to whether the nonlinear equation $\mathcal{F}_\pi z = 0$ is solvable. To this end, we introduce the equivalent fixed point problem $E_\pi z = z$, where the map E_π acts in $\mathbb{R}^{m(N+1)}$,

$$E_\pi z := z - \mathcal{F}'_\pi(x^*_\pi)^{-1}\mathcal{F}_\pi z, \quad z \in \mathcal{B}_\pi(x^*_\pi, \varrho_0).$$

As usual when we mean to apply Banach's Fixed Point Theorem, we state

$$E_\pi z - E_\pi \bar{z} = \mathcal{F}'_\pi(x^*_\pi)^{-1} \int_0^1 \{\mathcal{F}'_\pi(x^*_\pi) - \mathcal{F}'_\pi(sz + (1-s)\bar{z})\}\mathrm{d}s(z - \bar{z}) \tag{3.49}$$

and

$$\begin{aligned} E_\pi z - x^*_\pi &= z - x^*_\pi - \mathcal{F}'_\pi(x^*_\pi)^{-1}(\mathcal{F}_\pi z - \mathcal{F}_\pi x^*_\pi + \mathcal{F}_\pi x^*_\pi) \\ &= \mathcal{F}'_\pi(x^*_\pi)^{-1} \int_0^1 \{\mathcal{F}'_\pi(x^*_\pi) - \mathcal{F}'_\pi(sz + (1-s)x^*_\pi)\}\mathrm{d}s(z - x^*_\pi) - \\ &\quad - \mathcal{F}'_\pi(x^*_\pi)^{-1}\mathcal{F}_\pi x^*_\pi, \end{aligned} \tag{3.50}$$

for $z, \bar{z} \in \mathcal{B}_\pi(x^*_\pi, \varrho_0)$.

Given a constant $\alpha < 1$, we choose $\varepsilon = \varepsilon(\pi)$ such that

$$\varepsilon \|\mathcal{F}'_\pi(x^*_\pi)^{-1}\|_\pi \le \alpha < 1. \tag{3.51}$$

Moreover, since $\mathcal{F}'_\pi$ is continuous, there exists a $\varrho = \varrho(\varepsilon(\pi)) > 0$ so that

$$\|\mathcal{F}'_\pi(x^*_\pi) - \mathcal{F}'_\pi(y)\|_\pi \le \varepsilon \quad \text{for all } y \in \overline{\mathcal{B}_\pi(x^*_\pi, \varrho)}.$$

Hence, for $z, \bar{z} \in \overline{\mathcal{B}_\pi(x^*_\pi, \varrho)}$ (3.49), (3.51) provide

$$\|E_\pi z - E_\pi \bar{z}\|_\pi \le \alpha \|z - \bar{z}\|_\pi \tag{3.52}$$

$$\|E_\pi z - x^*_\pi\|_\pi \le \alpha \|z - x^*_\pi\|_\pi + \|\mathcal{F}'_\pi(x^*_\pi)^{-1}\tau_\pi\|_\pi. \tag{3.53}$$

If we are sure to manage the inequality

$$\|\mathcal{F}'_\pi(x^*_\pi)^{-1}\tau_\pi\|_\pi \le (1-\alpha)\varrho, \tag{3.54}$$

we know the map E_π to have a unique fixed point on $\overline{\mathcal{B}(x^*_\pi, \varrho)}$. However, keep in mind that ε and ϱ, may both depend on the grid π.

The same arguments apply to the perturbed equation

$$\mathcal{F}_\pi z = \delta_\pi. \tag{3.55}$$

If we suppose the inequality

$$\|\mathcal{F}'_\pi(x^*_\pi)^{-1}(\tau_\pi - \delta_\pi)\|_\pi \le (1-\alpha)\varrho$$

can be satisfied, equation (3.55) is uniquely solvable on $\overline{\mathcal{B}_\pi(x^*_\pi, \varrho)}$, and, for its solution $\tilde{x}_\pi$, the error estimate

$$\|x^*_\pi - \tilde{x}_\pi\|_\pi \le \frac{1}{1-\alpha}\|\mathcal{F}'_\pi(x^*_\pi)^{-1}(\tau_\pi - \delta_\pi)\|_\pi \tag{3.56}$$

is valid.

The BDF is said to be *feasible* in this case, i.e. if the nonlinear equations to be solved per step are locally uniquely solvable. Then, the Newton method may be applied, where (3.56) suggests how accurate the defects δ_j should be.

In the following $\Pi_0 \subseteq \Pi$ always denotes a grid class where the maximal step-sizes of all grids are sufficiently small.

Theorem 3.4 Let the given BDF applied to regular explicit ODEs be stable on the grid class Π. Let the DAE (3.1) satisfy Assumption 2.10, and let $x_* \in C^1_N$ solve this DAE. In addition, let $\{A_*, B_*\}$ be index-1 tractable. Then the BDF is feasible and stable on $\Pi_0 \subseteq \Pi$. The convergence order is the same as in case of regular ODEs.

Proof. By Theorem 3.1, there is a uniform bound S such that $\|\mathcal{L}_\pi^{-1}\|_\pi \le S$ for all $\pi \in \Pi$. Choose sufficiently fine grids (cf. (3.46)) so that

$$\gamma_\pi S < 1, \qquad \pi \in \Pi_0.$$

Hence $\|\mathcal{L}_\pi - \mathcal{F}'_\pi(x^*)\|_\pi \le \gamma_\pi$, $\|\mathcal{L}_\pi^{-1}\|_\pi \le S$ imply the bijectivity of $\mathcal{F}'_\pi(x^*)$ as well as

$$\|\mathcal{F}'_\pi(x^*_\pi)^{-1}\|_\pi \le \frac{S}{1-\gamma_\pi S} =: S_1.$$

Consequently, in (3.51) we need uniform $\varepsilon = \alpha/S_1$ and δ, respectively, for all $\pi \in \Pi_0$.

Moreover, (3.54) is easy to satisfy by choosing refined grids and sufficiently accurate starting values such that

$$\|\tau_\pi\|_\infty \le \frac{1}{S_1}(1-\alpha)\varrho.$$

Moreover, for $z, \bar{z} \in \overline{\mathcal{B}_\pi(x^*_\pi, \varrho)}$ the matrix $\int_0^1 \mathcal{F}'_\pi(sz + (1-s)z)\mathrm{d}s =: \mathcal{F}'_\pi[z, \bar{z}]$ is also nonsingular since

$$\|\mathcal{F}'_\pi(x^*) - \mathcal{F}'_\pi[z, \bar{z}]\|_\pi \le \varepsilon, \qquad \|\mathcal{F}'_\pi[z, \bar{z}]^{-1}\|_\pi \le \frac{S_1}{(1-\alpha)} =: S_2.$$

Hence, $z - \bar{z} = \mathcal{F}'_\pi[z, \bar{z}]^{-1}(\mathcal{F}_\pi z - \mathcal{F}_\pi \bar{z})$ implies stability immediately. Then, (3.10) (or (3.56)) provides convergence. □

Remark Clearly, the nonlinear equations to be solved per step are locally uniquely solvable, and the Newton method can be applied. The same is true for the perturbed equations (cf. (3.56)).

In the case of an index-2-tractable matrix coefficient pair $\{A_*, B_*\}$ the situation becomes worse. Theorem 3.2 only provides for $\|\mathcal{L}_\pi^{-1}\|_\pi \le S\underline{h}^{-1}$ for $\pi \in \Pi$. How does this affect nonlinear DAEs?

Lemma 3.5 Let Assumption 2.10 be fulfilled, and let $x_* \in C_N^1$ be given. In addition, let $P' = 0$. Then, let $A_1(y,x,t) := f_y'(y,x,t) + f_x'(y,x,t)Q$ be singular for all (y,x,t) belonging to a neighbourhood $\mathcal{N}$ of the trajectory $\mathcal{T}_*$ of x_* within $\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}$, but with constant rank there. Moreover, let $A_1(\zeta(t))$ have a smooth null space. Furthermore let

$$\ker A_1(y,x,t) \cap S_1(y,x,t) = \{0\}, \tag{3.57}$$

$$S_1(y,x,t) := \{z \in \mathbb{R}^m : f_x'(y,x,t)Pz \in \mathrm{im} A_1(y,x,t)\},$$

for all $(y,x,t) \in \mathcal{T}$. Then the DAE linearized in x_* is index-2 tractable.

Proof. We have $A_*(t) := f_y'(\zeta(t))$, $B_*(t) := f_x'(\zeta(t))$, furthermore $A_{*;1}(t) = A_*(t) + B_*(t)Q = A_1(\zeta(t))$, $S_{*,1}(t) := S_1(\zeta(t))$, $\ker(A_{*,1}(t)) \cap S_{*,1}(t) = \{0\}$. Since the null space of $A_1(\zeta(t))$ is assumed to depend continuously differentiably on t we are done. □

Lemma 3.6 Let Assumption 2.10 be valid and let $x_* \in C_N^1$ solve the DAE (3.1). In addition, let $\mathrm{im} f_y'(y,x,t)$ be independent of y, i.e.

$$\mathrm{im}(f_y'(y,x,t,)) =: R(x,t).$$

Then, for the local errors τ_j generated by the BDF (3.7), the implication

$$\tau_j \in R(x_*(t_j), t_j), \quad j \ge s, \tag{3.58}$$

becomes true.

Proof. Denote shortly

$$\mu_j := P(t_j)\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_*(t_{j-i}), \qquad \lambda_j := (Px_*)'(t_j) - P'(t_j)x_*(t_j).$$

Derive

$$\begin{aligned}\tau_j &:= f(\eta_j) = f(\mu_j, x_*(t_j), t_j)\\ &= f(\mu_j, x_*(t_j), t_j) - f(\lambda_j, x_*(t_j), t_j)\\ &= \int_0^1 f_y'(s\mu_j + (1-s)\lambda_j, x_*(t_j), t_j)\mathrm{d}s(\mu_j - \lambda_j),\end{aligned}$$

thus (3.58) is valid. □

Theorem 3.7 Let Π be such a grid class where the quotient of the maximal and minimal step-sizes of any $\pi \in \Pi$ is bounded by a global constant K, i.e.

$$h \cdot \underline{h}^{-1} \leq K, \qquad \pi \in \Pi.$$

Let the given BDF applied to regular explicit ODEs become stable on Π. Let DAE (3.1) satisfy all assumptions of Lemma 3.5 as well as Lemma 3.6; furthermore, let the partial Jacobians f'_y, f'_x be Lipschitz with respect to (y, x). In addition, let

$$\left| (Px_*)'(t_j) - \frac{1}{h_j} \sum_{i=0}^{s} \alpha_{ji} Px_*(t_{j-i}) \right| \leq c_0 h^p, \tag{3.59}$$

be valid for all $j \geq s$, $\pi \in \Pi$ and certain constants $c_0 > 0$, $p > 1$. Then the BDF is feasible and weakly instable on Π.

If we suppose the Q_1-components of the starting values to have the order of accuracy $p+1$ and the other ones order p, then the convergence order is p.

Proof. Because of the Lipschitz continuity of f'_y and f'_x, $\mathcal{F}'_\pi$ also becomes Lipschitz continuous. In particular

$$\|\mathcal{F}'_\pi(x^*_\pi) - \mathcal{F}'_\pi(y)\|_\pi \leq L \|x^*_\pi - y\|_\pi \tag{3.60}$$

is valid. Moreover, in (3.46) we may estimate (using the notation from the proof of Lemma 3.6)

$$\gamma_\pi \leq c_1 \max_{j \geq s} |\mu_j - \lambda_j| \leq c_2 h^p.$$

Since (cf. (3.46)) $\|\mathcal{L}_\pi - \mathcal{F}'_\pi(x^*_\pi)\|_\pi \leq c_2 h^p$ and, due to Theorem 3.2, $\|\mathcal{L}_\pi^{-1}\|_\pi \leq S\underline{h}^{-1}$, we may refine the grids in such a way that

$$c_2 S \underline{h}^{-1} h^p \leq c_2 S K h^{p-1} < 1.$$

Consequently, $\mathcal{F}'_\pi(x^*_\pi)$ becomes nonsingular, and

$$\|\mathcal{F}'_\pi(x^*_\pi)^{-1}\|_\pi \leq S\underline{h}^{-1}(1 - c_2 S K h^{p-1})^{-1}.$$

Next, by Lemma 3.6,

$$\tau_j \in R(x_*(t_j), t_j) = \operatorname{im} A_*(t_j), \quad j \geq s.$$

Taking into account (3.32), this implies

$$\begin{aligned} \|\mathcal{L}_\pi^{-1}\tau_\pi\|_\pi \;\leq\; & \tilde{S} \Bigg\{ \max_{j \geq o} |\tau_j| + \\ & + \max_{s \leq j \leq 2s-1} \left| \frac{1}{h_j} Q Q_1(t_j) \sum_{i=j+1-s}^{s} \alpha_{ji} Q_1(t_{j-i})(x_*(t_{j-i}) - x_{j-i}) \right| \Bigg\}. \end{aligned} \tag{3.61}$$

Choosing the starting values to be as accurate as necessary for

$$\begin{aligned} \frac{1}{h_j}|Q_1(t_j)(x_*(t_j) - x_j)| &\leq c_3 h^p, \\ |x_*(t_j) - x_j| &\leq c_3 h^p, \quad j = 0, \ldots, s-1, \end{aligned} \tag{3.62}$$

to be satisfied, we obtain

$$\|\mathcal{L}_\pi^{-1}\tau_\pi\|_\pi \leq c_4 h^p.$$

Then, because

$$\mathcal{F}'_\pi(x_\pi^*)^{-1} = (I - \mathcal{L}_\pi^{-1}(\mathcal{L}_\pi - \mathcal{F}'_\pi(x_\pi^*)))^{-1}\mathcal{L}_\pi^{-1} \tag{3.63}$$

the inequality

$$\begin{aligned} \|\mathcal{F}'_\pi(x_\pi^*)^{-1}\tau_\pi\|_\pi &\leq \|(I - \mathcal{L}_\pi^{-1}(\mathcal{L}_\pi - \mathcal{F}'_\pi(x_\pi^*)))^{-1}\| \, \|\mathcal{L}_\pi^{-1}\tau_\pi\|_\pi \\ &\leq (1 - c_2 S K h^{p-1})^{-1} c_4 h^p \end{aligned}$$

becomes true.

Next we show that both (3.51) and (3.54) may be satisfied. Given $0 < \alpha < 1$, choose $\varepsilon = \alpha(1/S)\underline{h}(1 - c_2 S K h^{p-1})$ to make E_π contractive. Since $\mathcal{F}'_\pi$ fulfils the Lipschitz condition (3.60), we may choose the related $\varrho(\varepsilon(\pi)) =: \varrho$ as $\varrho = 1/L\varepsilon$.

Finally, condition (3.54) becomes valid if

$$(1 - c_2 S K h^{p-1})^{-1} c_4 h^p \leq (1 - \alpha)\varrho = (1 - \alpha)\frac{1}{L}\alpha\frac{1}{S}\underline{h}(1 - c_2 S K h^{p-1})$$

is satisfied, or equivalently

$$c_4 h^{p-1} \leq \alpha(1 - \alpha)\frac{1}{LS}(1 - c_2 S K h^{p-1})^2,$$

but this can be managed by refining the grids.

By the same arguments as in Theorem 3.4, we derive

$$\|\mathcal{F}'_\pi[z, \bar{z}]^{-1}\|_\pi \leq \frac{1}{1 - \alpha} S \underline{h}^{-1}(1 - c_2 S K h^{p-1})^{-1},$$

and hence the BDF becomes weakly unstable. □

Remarks

1 From (3.56), (3.63) we conclude the error estimate

$$\|x_\pi^* - \tilde{x}_\pi\|_\pi \leq S_1 \|\mathcal{L}_\pi^{-1}(\tau_\pi - \delta_\pi)\|_\pi.$$

Taking (3.32) into consideration we are recommended to compute the Q_1 components of the starting values with a higher order accuracy than the remaining components (cf. also (3.61)). Moreover, the defects δ_j in the nonlinear equations should also be kept smaller in those components which do not belong to $\mathrm{im}(f'_y(\ldots, x_*(t_j), t_j) = R(x_*(t_j), t_j)$. This can be realized more easily if this subspace is kept constant.

2 Clearly, $Px_* \in C^s$ implies $p = s$ in (3.59).

3 Theorem 3.7 does not apply to the backward Euler method. It is not yet clear whether the condition $p > 1$ is a technical one for that large class of index-2 DAEs considered. However, Theorem 3.2 is also valid for the backward Euler method. The detailed error estimates (3.36), (3.37) show that the weak instabilities only affect certain components, and, moreover, only act locally. Inequality (3.35) does not comprise this situation precisely, but it represents a crude upper bound.

Using the detailed information given by (3.37), (3.36) and (3.32) for investigating nonlinear equations requires much technical effort. For special DAEs in Hessenberg form (0.3) this is done in Gear *et al.* (1985), Lötstedt and Petzold (1986) and Brenan and Engquist (1988). The statements of Theorem 3.2 remain valid for these nonlinear DAEs. In particular, the backward Euler is proved to converge.

Analogously to Theorem 3.7, an assertion concerning index-3-tractable DAEs could be proved using Theorem 3.3. In Lötstedt and Petzold (1986) and Brenan and Engquist (1988), a careful detailed decoupling of nonlinear index-3 DAEs in Hessenberg form (0.4) is carried out to obtain results similar to those we have proved for the linear case (cf. (3.40)–(3.42)).

While we are optimistic about overcoming the practical problems arising in large classes of index-2 equations, like error estimation and step-size control, the difficulties concerning the index-3 case seem to be more intractable. As is shown by (3.42), the Q_2 components of the starting values should now have order h^{p+2}, if the local error τ_j has order p. Moreover, the defects δ_j of the nonlinear equations to be solved per integration step should be kept small enough in the respective subspaces. Furthermore, remember that providing sufficiently accurate initial and starting values now becomes difficult. The nonlinear systems to be solved are ill conditioned, namely $\mathrm{cond}(F_i)$ behaves like h_j^{-3}.

For special nonlinear index-3 DAEs describing constrained mechanical motion, BDF codes are reported to work (e.g. Petzold and Lötstedt (1986), Führer (1988)) if the critical components are omitted from the error control, and only the PP_1 component (cf. (3.40)) is controlled. This may be applied if computing these PP_1 components only will do for practical reasons.

For a fairly detailed discussion of software for DAEs we refer to Brenan *et al.* (1989) and Hairer and Wanner (1991).

3.4. Further integration methods

First of all, it should be mentioned that these results, which have been proved for variable step-size BDFs, also apply to variable order variable step-size BDFs in the same way.

For one-step methods, there is a natural extension to fully implicit DAEs

(3.1), namely

$$f\left(\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_{j-i}, \sum_{i=0}^{s}\beta_{ji}x_{j-i}, \bar{t}_j\right) = 0, \qquad t_j := \sum_{i=0}^{s}\beta_{ji}t_{j-i}. \tag{3.64}$$

The consistency conditions are the same as those for the regular ODE case, but extra stability requirements are needed to ensure stability even in index-1 DAEs (cf. (1.12)). We do not recommend this method since it did not work well in experiments.

If the leading coefficient matrix has a constant range $\operatorname{im} f'_y(y,x,t) =: R$, and $S \in L(\mathbb{R}^m)$ denotes a projector onto R, $T := I - S$, we may formulate (März, 1985) a projected version of (3.64) as follows:

$$Sf\left(\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_{j-i}, \sum_{i=0}^{s}\beta_{ji}x_{j-i}, \bar{t}_j\right) + Tf(0, x_j, t_j) = 0. \tag{3.65}$$

Applied to semi-linear DAEs

$$u' + g(u,v,t) = 0, \quad h(u,v,t) = 0 \tag{3.66}$$

this simply means

$$\left.\begin{aligned} \frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}u_{j-i} + g\left(\sum_{i=0}^{s}\beta_{ji}u_{j-i}, \sum_{i=0}^{s}\beta_{ji}v_{j-i}, \bar{t}_j\right) = 0 \\ h(u_j, v_j, t_j) = 0 \end{aligned}\right\}. \tag{3.67}$$

Moreover, linear multi-step methods may be formulated as

$$\left.\begin{aligned} P(t_j)\left\{\frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}x_{j-i} - \sum_{i=0}^{s}\beta_{ji}y_{j-i}\right\} - y_j = 0 \\ f(y_j, x_j, t_j) = 0 \end{aligned}\right\}. \tag{3.68}$$

Method (3.68) is motivated by the equivalent formulation of (3.1)

$$\left.\begin{aligned} P(t)\{x'(t) - y(t)\} - y(t) = 0 \\ f(y(t), x(t), t) = 0 \end{aligned}\right\}. \tag{3.69}$$

When applied to the semi-explicit system (3.66), this linear multi-step method leads to

$$\left.\begin{aligned} \frac{1}{h_j}\sum_{i=0}^{s}\alpha_{ji}u_{j-i} - \sum_{i=0}^{s}\beta_{ji}g((u_{j-i}, v_{j-i}, t_{j-i}) = 0 \\ h(u_j, v_j, t_j) = 0 \end{aligned}\right\}. \tag{3.70}$$

All these methods are considered on general nonequidistant partitions π : $t_0 < t_1 < \cdots < t_N = T$. Our notation does not only allow for variable step-size, but also for formulae of different order and type, which is the common situation in many ODE codes.

Theorem 3.8 Let the methods considered be stable on the grid class Π for regular ODEs. Let DAE (3.1) satisfy Assumption 2.10, and let $x_* \in C_N^1$ solve this DAE. Furthermore, let $\{A_*, B_*\}$ be index-1 tractable. In addition, let the partial Jacobian $f_y'(y, x, t)$ have a constant null space N when applying the linear multi-step method (3.68), but a constant range R when applying the one-step method (3.65).

Then both methods are feasible and stable on $\Pi_0 \subseteq \Pi$. The order of consistency is the same as for regular ODEs.

Proof. Taking into account that (3.69) is again an index-1-tractable DAE, we apply the same arguments as those used for Theorem 3.4 in both cases. □

Remarks

1 In both methods, the choice $\beta_{jo} = 0$ is allowed. In particular, in (3.70) one can take advantage of such 'explicit' methods.

2 The linear multi-step method is also proved to be stable (by the same arguments) for a time varying null space $N(t)$. However, then a certain order reduction may occur. This is caused by a somewhat inexact realization of the subspace structure of the DAE. More precisely, if x_* is smooth enough, we have

$$\begin{aligned}
\tau_j &= P(t_j)\sum_{i=0}^{s}\left\{\frac{1}{h_j}\alpha_{ji}(Px_*)(t_{j-i}) - \beta_{ji}(Px_*)'(t_{j-i})\right\} \\
&\quad + P(t_j)\sum_{i=0}^{s}\left\{\frac{1}{h_j}\alpha_{ji}Q(t_{j-i})x_*(t_{j-i}) + \beta_{ji}P'(t_{j-i})x_*(t_{j-i})\right\} \\
&= P(t_j)\sum_{i=0}^{s}\left\{\frac{1}{h_j}\alpha_{ji}(Px_*)(t_{j-i}) - \beta_{ji}(Px_*)'(t_{j-i}) + \right. \\
&\quad \left. + \left(\frac{1}{h_j}\alpha_{ji}Q(t_{j-i}) - \beta_{ji}Q'(t_{j-i})\right)x_*(t_{j-i})\right\}.
\end{aligned}$$

Again, the conditions

$$\sum_{i=0}^{s}\alpha_{ji} = 0, \quad \sum_{i=0}^{s}\alpha_{ji}(t_{j-i} - t_j) = h_j\sum_{i=0}^{s}\beta_{ji}$$

turn out to be necessary and sufficient for the consistency at all. However, for order 2 we need two more conditions, the expected one

$$\sum_{i=0}^{s}\{\alpha_{ji}(t_{j-i} - t_j)^2 - 2h_j\beta_{ji}(t_{j-i} - t_j)\} = 0$$

as well as

$$\sum_{i=0}^{s}\{\alpha_{ji}(t_{j-i}-t_j)^2 - h_j\beta_{ji}(t_{j-i}-t_j)\}Q'(t_j) = 0.$$

Hence, e.g. the trapezoidal rule has order 1 only in this case.

3 The methods (3.65) and (3.68) naturally generate values x_j belonging to the state manifold of the DAE, which turns out to be a favourable property.

Among the Runge–Kutta methods

$$\left.\begin{aligned} x_j &= x_{j-1} + h_j \sum_{i=1}^{s} b_i X_i', \\ f(X_i', x_{j-1} + h_j \sum_{k=1}^{s} a_{ik} X_k', t_{j-1} + c_i h_j) &= 0, \\ i &= 1, \ldots, s, \end{aligned}\right\} \tag{3.71}$$

those with the coefficients $b_i = a_{si}$, $i = 1, \ldots, s$, $c_s = 1$ and a nonsingular matrix (a_{ik}) automatically provide values x_j belonging to the state manifold, if the method is applied to an index-1 DAE (3.1) with constant null space (e.g. Griepentrog and März (1986)). Then, the method maintains the order which it has for regular ODEs.

As we have learnt in Section 1, explicit Runge–Kutta methods (those having $a_{ij} = 0$ for $i \leq j$) are not suited for DAEs.

A fairly detailed discussion of general implicit Runge–Kutta methods as well as of extrapolation methods for index-1 DAEs is given in Brenan *et al.* (1989). As already indicated in Section 1, additional stability conditions have to be fulfilled, and one has to put up with order reduction.

A comprehensive exposition of Runge–Kutta methods for index-2 and index-3 DAEs in Hessenberg form (cf. (0.3), (0.4)) one can find in Hairer *et al.* (1989). A good work in which the well-known extrapolation methods, for example, are extended can be found in Deuflhard *et al.* (1987), Lubich (1990) and Hairer *et al.* (1989). All these methods extensively use the special structure of the given Hessenberg form DAEs. In particular, the problems caused by weak instability have been overcome e.g. by special error control in the nonlinear equations and by projections onto the given manifolds, respectively.

The projected implicit Runge–Kutta methods (Ascher and Petzold, 1990) also use these ideas.

4. Brief remarks on related problems

4.1. Index reduction

From the point of view of computational tractability, it is desirable for the DAE to have an index which is as small as possible. The procedure for

determining the differentiation index described in Section 2.2 is an index reduction method (cf. Griepentrog (1991)), in fact. However, in Section 2.2 it was mentioned that the attained system does not reflect the stability behaviour of the original DAE well.

A different method for reducing the index of a DAE is presented in Mrziglod (1987) and Čistjakov (1982). Instead of replacing constraints by differential equations, in their method suitable differential equations are deleted. This method works for linear DAEs, but it is not clear to what kind of nonlinear DAEs it may be applicable.

A very useful idea for reducing the index is proposed in Gear *et al.* (1985) for the special index-3 Hessenberg system

$$u'(t) - v(t) = 0, \tag{4.1}$$

$$v'(t) + g(u(t), v(t), t) + h'_u(u(t), t)^T w(t) = 0, \tag{4.2}$$

$$h(u(t), t) = 0, \tag{4.3}$$

which results from the Euler–Lagrange formulation of a constrained mechanical system. As mentioned earlier, the system with the differentiated constraint

$$h'_u(u(t), t)v(t) + h'_t(u(t), t) = 0 \tag{4.4}$$

instead of (4.3) would cause the numerical solution to drift away from the constraint manifold. Note that the system (4.1), (4.2), (4.4) has index 2. To stabilize the obtained index-2 system, an additional Lagrange multiplyer z is introduced, and (4.3) is summed again. The resulting system

$$\left.\begin{aligned} u'(t) - v(t) + h'_u(u(t), t)^T z(t) &= 0 \\ v'(t) + g(u(t), v(t), t) + h'_u(u(t), t)^T w(t) &= 0 \\ h'_u(u(t), t)v(t) + h'_t(u(t), t) &= 0 \\ h(u(t), t) &= 0 \end{aligned}\right\} \tag{4.5}$$

is index-2 tractable. Is it easy to check that any solution of (4.5) has a trivial component z. Note that Führer and Leimkuhler (1990) took advantage of this fact to create a skilful special BDF modification to the Euler–Lagrange equations.

In Section 2 we pointed out that higher index DAEs lead to ill posed IVPs in the naturally given topologies (cf. Corollary 2.7). Hence, we may treat them as such, i.e. use some regularization procedure. At first glance this approach might appear a heavy gun, which is true insofar as standard regularization techniques (Tikhonov regularization, least-squares collocation) are concerned. However, different special parametrizations may be created, which are closely connected with the structure of the DAEs and the source of their ill posedness. As usual, the regularized equations represent singularly perturbed index-1-tractable DAEs and ODEs, respectively.

For instance, the index-2 DAE

$$x_1' - x_2 = 0, \quad x_1 = q$$

may be approximated by the index-1 system

$$x_1' - x_2 = 0, \quad \varepsilon x_1' + x_1 = q, \quad x_1(0) = q(0).$$

For general DAEs $f(x'(t), x(t), t) = 0$ the same regularization method provides

$$f(x'(t), x(t) + \varepsilon P(t)(Px)'(t), t) = 0.$$

We refer to Hanke (1990, 1991) for a comprehensive survey on methods, convergence results, asymptotic expansions etc.

4.2. Boundary value problems

Let us consider the linear equation

$$A(t)x'(t) + Bx(t) = q(t), \quad t \in [t_0, T], \tag{4.6}$$

once again. Now we are interested in a solution of (4.6) that satisfies the boundary condition

$$D_1 x(t_0) + D_2 x(T) = d \tag{4.7}$$

with given matrices $D_1, D_2 \in L(\mathbb{R}^m)$, $d \in M := \mathrm{im}(D_1, D_2)$. According to the discussion of linear IVPs in section 2.1 we determine a fundamental solution matrix $X(\cdot)$ by

$$AX' + BX = 0 \tag{4.8}$$

$$\Pi_\mu(X(t_0) - I) = 0, \tag{4.9}$$

where $\Pi_\mu := P_0(t_0) \dots P_{\mu-1}(t_0)$, and the coefficient matrix pair $\{A, B\}$ is supposed to be index-μ tractable. From Theorem 2.5, the fundamental solution matrix is uniquely determined, the columns of X belong to C_N^1.

Now, (4.8), (4.9) immediately imply that

$$X(t) = X(t)\Pi_\mu \tag{4.10}$$

holds, i.e. $X(t)$ is singular for all t. Moreover, even $\ker X(t) = \ker \Pi_\mu$ is true. However, what about the so-called shooting matrix

$$K := D_1 X(t_0) + D_2 X(T). \tag{4.11}$$

Trivially, K becomes singular, too.

Theorem 4.1 Let $\{A, B\}$ be index-μ tractable and q sufficiently smooth. Then the BVP (4.6), (4.7) is uniquely solvable for each $d \in M$ if and only if

$$\mathrm{im} K = M, \quad \ker K = \ker \Pi_\mu \tag{4.12}$$

are valid.

Proof. By standard arguments, we have to consider the linear system

$$Kz = d - D_2\tilde{x}(T),$$

where $\tilde{x} \in C_N^1$ denotes that solution of (4.6) which satisfies $\Pi_\mu \tilde{x}(t_0) = 0$.

Clearly, $\ker \Pi_\mu \subseteq \ker K$ holds. Furthermore, $Kz = 0$, $\Pi_\mu z \neq 0$ would imply that $X(t)\Pi_\mu z$ has to become a nontrivial solution of the homogeneous BVP. □

Remarks

1 Theorem 4.1 generalizes facts that are well known for regular ODEs ($M = \mathbb{R}^m$) and index-1-tractable DAEs (Griepentrog and März, 1986), respectively.

2 The relations (4.12) mean that the *boundary conditions are stated well*; in particular, the number of linearly independent boundary conditions is rank Π_μ.

3 The *whole* BVP is well posed in the naturally given topologies if and only if $\mu = 1$, and if (4.12) is satisfied (cf. Corollary 2.7).

Linear and nonlinear BVPs in transferable (index-1-tractable) DAEs are well understood. Classical arguments apply for discretizations by finite differences (Griepentrog and März, 1986) and spline-collocation (Degenhardt, 1991; Ascher, 1989), respectively. In particular, it is possible to trace the stability question of the BVP back to that of the IVPs. However, for the latter we refer to the typical explanations in Section 3, which are carried out in the same manner for certain one-step methods, for example in Griepentrog and März (1986). Furthermore, dichotomy is considered in its relationship to the conditioning of the BVP in Lentini and März (1990a,b).

Of course, the singular shooting equation causes numerical difficulties. This is why modified shooting techniques yielding isolatedly solvable nonlinear shooting equations have been proposed (Lamour, 1991a,b). The basic idea is to combine the shooting equation and the equation for calculating consistent initial values (cf. Theorem 2.3). For instance, simple shooting for a linear BVP leads to the system

$$\begin{aligned} A(t_0)y_0 \;+\; & B(t_0)x_0 = q(t_0), \\ Q(t_0)y_0 \;=\; & 0, \qquad Kx_0 = d - D_2\tilde{x}(T). \end{aligned}$$

Surely, Theorem 4.1 suggests to apply shooting methods also to higher index DAEs. This will work, supposed we are able to integrate the IVPs.

We are looking forward to related results for general index-2-tractable DAEs.

REFERENCES

U. Ascher (1989), 'On numerical differential algebraic problems with application to semi-conductor device simulations', *SIAM J. Numer. Anal.* **26**, 517–538.

U. Ascher and L.R. Petzold (1990), *Projected implicit Runge–Kutta methods for differential-algebraic equations*, Preprint, Lawrence Livermore National Laboratory

K.E. Brenan and B.E. Engquist (1988), 'Backward differentiation approximations of nonlinear differential/algebraic systems', and Supplement, *Math. Comput.* **51**, 659–676, S7–S16.

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

S.L. Campbell (1987), 'A general form for solvable linear time varying singular systems of differential equations', *SIAM J. Math. Anal.* **18**, 1101–1115.

L.O. Chua and A. Deng (1989), 'Impasse points. Part I: Numerical aspects', *Int. J. Circuit Theory Applics* **17**, 213–235.

V.F. Čistjakov (1982), 'K metodam rešenija singuljarnych linejnych sistem obyknovennych differencial'nych uravnenij', in *Vyroždennyje Sistemy Obyknovennych Differencial'nych Uravnenij* (Ju.E. Bojarincev, ed.) Nauka Novosibirsk, 37–65.

A. Degenhardt (1991), 'A collocation method for boundary value problems of transferable DAEs', *Numer. Math.*, to appear.

P. Deuflhard, E. Hairer, J. Zugck (1987), 'One-step and extrapolation methods for differential-algebraic systems' *Numer. Math.* **51**, 501–516.

C. Führer (1988), Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen. Theorie, numerische Ansätze und Anwendungen, Dissertation, Techn. Univ. München, Fak. für Mathematik und Informatik.

C. Führer and B. Leimkuhler (1989), Formulation and numerical solution of the equations of constrained mechanical motion, DFVLR-Forschungsbericht 89-08, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt, Oberpfaffenhofen.

C. Führer and B.J. Leimkuhler (1990), A new class of generalized inverses for the solution of discretized Euler–Lagrange equations, in *NATO Advanced Research Workshop on Real-Time Integration Methods for Mechanical System Simulation* (Snowbird, Utah 1989) (E. Haug and R. Deyo, eds.) Springer (Berlin).

F.R. Gantmacher (1966), *Teorija matric*, Nauka (Moskva).

C.W. Gear (1971), 'The simultaneous numerical solution of differential-algebraic equations', *IEEE Trans. Circuit Theory*, **CT-18**, 89–95.

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

C.W. Gear, H.H. Hsu and L. Petzold (1981), Differential-algebraic equations revisited, *Proc. ODE Meeting, Oberwolfach, Germany*, Institut für Geom. und Praktische Mathematik, Technische Hochschule Aachen, Bericht 9, Germany.

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

E. Griepentrog (1991), Index reduction methods for differential-algebraic equations, Preprint 91-12, Humboldt-Univ. Berlin, Fachbereich Mathematik.

E. Griepentrog and R. März (1986), *Differential-Algebraic Equations and their Numerical Treatment (Teubner Texte zur Mathematik 88)* Teubner (Leipzig).

E. Griepentrog and R. März (1989), 'Basic properties of some differential-algebraic equations', *Z. Anal. Anwend.* **8** (1), 25–40.

E. Hairer and G. Wanner (1991), *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems (Springer Series in Computational Mathematics 14)* Springer (Berlin).

E. Hairer, Ch. Lubich and M. Roche (1989), *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods (Lecture Notes in Mathematics 1409)* Springer (Berlin).

M. Hanke (1990) Regularization methods for higher index differential-algebraic equations, Preprint 268, Humboldt-Univ. Berlin, Fachbereich Mathematik.

M. Hanke (1991), 'On the asymptotic representation of a regularization approach to nonlinear semiexplicit higher-index differential-algebraic equations', *IMA J. Appl. Math.* **46**, 225–245.

B. Hansen (1989), Comparing different concepts to treat differential-algebraic equations, Preprint 220, Humboldt-Univ. Berlin, Sektion Mathematik.

B. Hansen (1990), Linear time-varying differential-algebraic equations being tractable with the index k, Preprint 246, Humboldt-Univ. Berlin, Sektion Mathematik.

H.B. Keller (1975), 'Approximation methods for nonlinear problems with application to two-point boundary value problems', *Math. Comput.* **29** (130), 464–474.

H.B. Keller and A.B. White (1975), 'Difference methods for boundary value problems in ordinary differential equations', *SIAM J. Numer. Anal.* **12**, 791–802.

R. Lamour (1991a), 'A well-posed shooting method for transferable DAE's', *Numer. Math.* **59**.

R. Lamour (1991b), Oscillations in differential-algebraic equations, Preprint 272, Humboldt-Universität Berlin, Fachbereich Mathematik.

M. Lentini and R. März (1990a), 'The conditioning of boundary value problems in transferable differential-algebraic equations', *SIAM J. Numer. Anal.* **27**, 1001–1015.

M. Lentini and R. März (1990b),'Conditioning and dichotomy in transferable differential-algebraic equations', *SIAM J. Numer. Anal.* **27**, 1519–1526.

P. Lötstedt and L. Petzold (1986), 'Numerical solution of nonlinear differential equations with algebraic constraints I: Convergence results for backward differentiation formulas', *Math. Comput.* **46**, 491–516.

Ch. Lubich (1990), Extrapolation integrators for constrained multibody systems, Report, Univ. Innsbruck.

R. März (1984), 'On difference and shooting methods for boundary value problems in differential-algebraic equations', *ZAMM* **64** (11), 463–473.

R. März (1985), 'On initial value problems in differential-algebraic equations and their numerical treatment', *Computing* **35**, 13–37.

R. März (1989), 'Some new results concerning index-3 differential-algebraic equations', *J. Math. Anal. Applics* **140** (1), 177–199.

R. März (1990), 'Higher-index differential-algebraic equations: Analysis and numerical treatment', *Banach Center Publ.* **24**, 199–222.

R. März (1991), On quasilinear index 2 differential algebraic equations, Preprint 269, Humboldt-Universität Berlin, Fachbereich Mathematik.

T. Mrziglod (1987), Zur Theorie und Numerischen Realisierung von Lösungsmethoden bei Differentialgleichungen mit angekoppelten algebraischen Gleichungen, Diplomarbeit, Universität zu Köln.

L.R. Petzold (1986), 'Order results for implicit Runge–Kutta methods applied to differential/algebraic systems', *SIAM J. Numer. Anal.* **23**, 837–852.

L. Petzold and P. Lötstedt (1986), 'Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications', *SIAM J. Sci. Stat. Comput.* **7**, 720–733.

F.A. Potra and W.C. Rheinboldt (1991), 'On the numerical solution of Euler–Lagrange equations', *Mech. Struct. Machines* **19** (1).

P.J. Rabier and W.C. Rheinboldt (1991), 'A general existence and uniqueness theory for implicit differential-algebraic equations', *Diff. Int. Eqns* **4** (3), 563–582.

S. Reich (1990), Beitrag zur Theorie der Algebrodifferentialgleichungen, Dissertation (A), Technische Universität Dresden.

W.C. Rheinboldt (1984), 'Differential-algebraic systems as differential equations on manifolds', *Math. Comput.* **43**, 473–482.

B. Simeon, C. Führer and P. Rentrop (1991), 'Differential-algebraic equations in vehicle system dynamics', *Surv. Math. Ind.* **1** (1), 1–37.

R.F. Sincovec, A.M Erisman, E.L. Yip and M.A. Epton (1981), 'Analysis of descriptor systems using numerical algorithms', *IEEE Trans. Aut. Control*, **AC-26**, 139–147.