Applying BDF to quasilinear differential-algebraic equations of index 2: a perturbation analysis

M. Arnold*

Abstract

For a class of quasilinear differential-algebraic equations (DAEs) of index 2 the error propagation in analytical and numerical solution of an initial value problem is studied. Previously obtained error bounds for backward Euler method applied to DAEs of index 2 in Hessenberg form are extended to BDF discretizations of quasilinear DAEs. The large error terms are shown to be concentrated in the algebraic (or "nullspace") components of the solution while the differential components are much more robust against perturbations. For linear DAEs with time varying coefficients the error bounds can be further decreased. The results of the theoretical analysis are illustrated by numerical tests.

Zusammenfassung

In der vorliegenden Arbeit wird für eine Klasse quasilinearer Algebrodifferentialgleichungen (DAEs) vom Index 2 die Fehlerfortpflanzung in analytischer und numerischer Lösung eines Anfangswertproblems untersucht. Fehlerabschätzungen, die in [4] für das implizite Eulerverfahren für Index-2-Systeme in Hessenbergform bewiesen wurden, werden auf BDF–Diskretisierungen quasilinearer DAEs erweitert. Es wird gezeigt, daß der Fehler in den algebraischen (oder Nullraum–) Komponenten der Lösung konzentriert ist, der Einfluß von kleinen Störungen auf die differentiellen Lösungskomponenten ist wesentlich geringer. Im Spezialfall linearer DAEs mit zeitabhängigen Koeffizienten können die Fehlerschranken für die differentiellen Lösungskomponenten noch einmal deutlich verkleinert werden. Numerische Testrechnungen zeigen Auswirkungen des theoretisch untersuchten Phänomens in praktischen Rechnungen.

Mathematics Subject Classification (1991): 65L05
Keywords: Differential algebraic equations, perturbation index, Backward differentiation

1 Introduction

Starting with the work of Gear ([7]) the straightforward integration of DAEs by ODE methods found much interest ([11], [5]). If an initial value problem is discretized directly (i. e. as DAE and without any analytical transformations like e. g. index reduction) then stability and convergence of the discretization method have to be guaranteed. It is well known that – in contrast to ODE theory – the analytical solution of an initial value problem for higher index DAEs does not depend continuously on small perturbations in the system. This motivates the analysis of the error propagation in the solution of initial value problems, both for analytical and numerical solution. Error bounds for the analytical solution are given by the *perturbation index* ([11, p. 478]), for higher index systems they include derivatives of the perturbations. The discrete

*University of Rostock, Department of Mathematics, Postfach, D – 18051 Rostock, Germany,
E-mail: arnold@mathematik.uni-rostock.d400.de

anologue is the weak instability of discretization methods ([12]): Error bounds for the numerical solution of higher index DAEs usually contain a term $\frac{1}{h^{m-1}}\theta$ in addition to the error terms known from ODE theory (discretization error and error of size $\mathcal{O}(\theta)$). Here h denotes the stepsize of integration, m the perturbation index and θ an upper bound for the error in one single step of integration that arises if the method is implemented on a computer (e. g. round-off errors, errors caused by stopping the iterative solution of nonlinear equations). This *additional error term* tends to infinity if the stepsize is decreased. It causes severe difficulties in the integration of higher index DAEs.

A more detailed analysis shows, however, that the additional error term is substantially smaller in some of the solution components. This fact allows to integrate DAEs in Hessenberg form up to index 3 with stepsize control algorithms that are based on those solution components that are robust against perturbations (e. g. [9]).

For systems of index 2 in *Hessenberg form*

$$\begin{aligned} y'(t) &= \varphi(y(t), z(t)) \\ 0 &= \gamma(y(t)) \end{aligned} \tag{1}$$

with sufficiently differentiable functions φ and γ that satisfy the *index-2 condition*

$$[\gamma_y \varphi_z](y, z) \quad \text{non-singular} \tag{2}$$

the error of backward Euler method is analysed in great detail in [4]. In the present paper we extend these results to BDF discretizations for a class of quasilinear DAEs of index 2

$$A(t)x'(t) + g(t, x(t)) = 0 \ , \tag{3}$$

that includes both systems (1) in Hessenberg form and linear DAEs with time varying coefficients that are tractable with index 2 and have a constant nullspace $N := \ker A(t)$ ([12]). BDF up to order 6 were proved to be well-defined ("feasible") and convergent for this class of DAEs in [17]. The main difference of our perturbation analysis and the one presented in [17] are smaller error bounds for those solution components that are the most important ones in variable stepsize codes (in (1): for the differential components y). Examples and a numerical test illustrate that these smaller error bounds reflect the influence of perturbations on the numerical solution correctly for *all* solution components and that they can not be further decreased in general.

The paper is organized as follows: in Section 2 the error of 1st and 2nd order BDF applied to a simple test problem is analysed. This gives some motivation for the structure of error bounds for analytical solution (Section 4) and for numerical solution (Section 5) of initial value problems for (3). The class of DAEs that we are concerned with is defined in Section 3. The basic tool of the proofs is the decoupling of the solution in various components. The (very) technical details of this procedure are given separately in Section 6.

2 Example

The perturbation index analysis for systems of index 2 in Hessenberg form ([11, p. 480f]) suggests that problems have to be expected if there are small errors in the algebraic part " $\gamma(y) = 0$ " of (1). To illustrate practical implications we integrated the non-autonomous system of index 2

$$\begin{aligned} y_1'(t) &= 4y_1^2 - y_2^2 + z \\ y_2'(t) &= (y_1 y_2 + 2z^2)z \\ 0 &= 4y_1 + y_2 - 6\sin t \end{aligned}$$

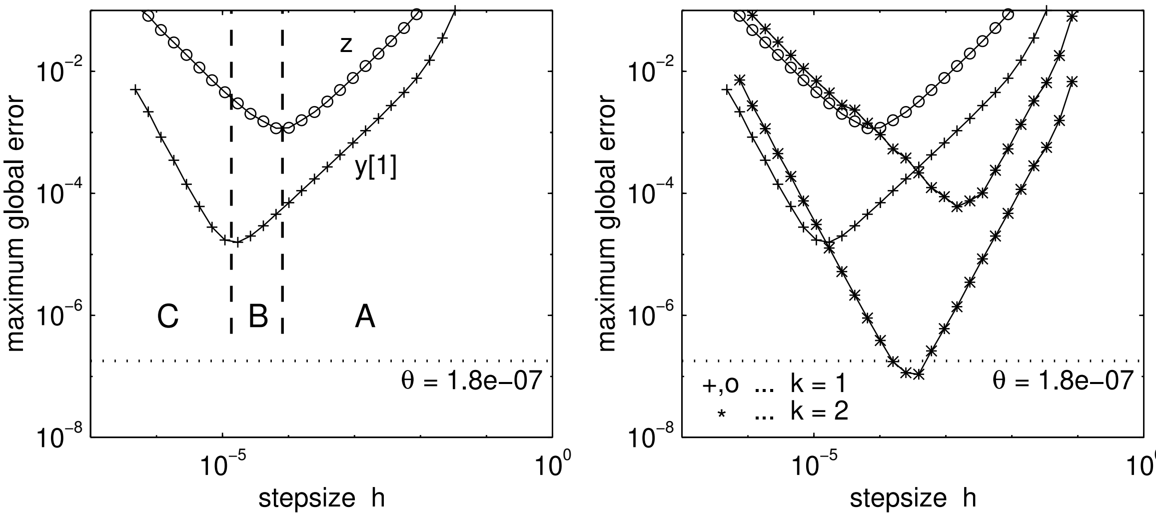


Figure 1: Global error of BDF applied to an index-2 DAE

with $t \in [0.5, 1]$ by 1st and 2nd order BDF. Initial values were taken from the analytical solution $y_1(t) = \sin t$, $y_2(t) = 2 \sin t$, $z(t) = \cos t$. Computations were performed with fixed stepsize in `IEEE extended` arithmetic (10 Bytes, i. e. machine epsilon $< 10^{-19}$). The integration was perturbed evaluating the time dependent term $\sin t$ in `single` precision, only (4 Bytes, the error in the algebraic constraint is bounded by $\theta = 1.8_{10}\text{-}7 \approx 6 \cdot \epsilon_{\text{single}}$). The left plot of Figure 1 shows the maximum global error for the differential component y_1 ("+") and the algebraic component z ("∘") versus stepsize h for backward Euler method (i. e. 1st order BDF). The plot was splitted in 3 characteristic regions. In region A, i. e. for larger stepsizes, the discretization error (that is $\mathcal{O}(h)$ for all solution components) dominates, slope +1 of the graphs in double logarithmic scale corresponds to order 1 of convergence of backward Euler. If the stepsize is decreased (region B) the error in y_1 is further decreased with $\mathcal{O}(h)$ but the error in z now *increases* since the additional error term becomes larger and is now the dominating term in the global error. Finally, for very small stepsizes (region C) the additional error term dominates in *all* solution components, the errors grow rapidly for decreasing stepsize h. The slopes of the graph are -2 for y and -1 for z, corresponding to error terms const_y/h^2 and const_z/h , respectively. The error grows faster in the differential components but nevertheless the errors in the algebraic component are much larger than in the differential ones. This is in very good agreement with theoretical investigations in [4] that predict additional error terms $\mathcal{O}(\frac{1}{h^2}\theta^2)$ for y and $\mathcal{O}(\frac{1}{h}\theta)$ for z. In the proofs of convergence for one-step methods and BDF (e. g. [5], [9]) there are severe restrictions on the size of θ, e. g. $\theta = \mathcal{O}(h^2)$. That is why standard theory explains the results in region A, but *not* in regions B and C.

To illustrate that this additional error term is typical for index-2 systems we repeated the test for 2nd order BDF. The right plot of Figure 1 shows again the maximum global error but now for 1st order ("+" and "∘" as before) and 2nd order ("*") BDF. The discretization error for the 2nd order method is much smaller than for backward Euler, $\mathcal{O}(h^2)$ corresponds to slope +2 of the graphs. The additional error terms are of the same size for both discretization methods. As it could be expected the minima of the errors are smaller for the 2nd order method. Note, however, that even in this very simple example errors of the 2nd order method start to increase at the fairly large stepsize $h = 2.0_{10}\text{-}3$.

Variable stepsize codes suppose, of course, that stepsize reductions reduce the error. So, they will work in situations like region A and can be adapted to work in region B. This is done in RADAU5 by scaling the error in the algebraic components z by h ([9]) and in DASSL by ignoring the errors in components z ([15]). Any variable stepsize code will fail if the stepsize is once as small as in region C of Figure 1.

3 A class of quasilinear DAEs of index 2

Error propagation in higher index DAEs is studied splitting the solution vector in various components. The way of splitting depends strongly on the structure of the DAE (e. g. differential and algebraic components in systems (1) in Hessenberg form). März and Tischendorf ([14]) present an approach based on the *inherent* structure of quasilinear DAEs (3) of index 2 instead of outer characteristics like e. g. "Hessenberg form". They put much emphasis on low smoothness assumptions w. r. t. the explicitly time dependent parts of (3) to cover as much examples from practical applications as possible. So (3) is *not* transferred to an autonomous DAE (as it is done by most other authors adding the trivial equation $t' = 1$). In the present paper we analyse discretizations for a class of DAEs (3) that is defined in [14] and includes both systems (1) in Hessenberg form and linear DAEs with $\ker A(t) = \text{const}$ that are tractable with index 2. In this section we introduce basic notations, for a very detailed discussion of the DAEs under consideration we refer to [14].

The analysis is centered around a linearization of DAE (3). We suppose that the initial value problem $x(t_0) = x_0$ for (3) has a solution $x(t)$, $(t \in [t_0, t_e])$ and that this solution is sufficiently differentiable. The leading coefficient $A(t)$ is supposed to be continuous on $[t_0, t_e]$ and to have a constant (i. e. not time depending) nullspace $N := \ker A(t)$. So DAEs of index 2 that can not be solved by ODE methods are excluded (e. g. the example of [8], see also [11, Exercise VI.6.2]). Q denotes any projector from $\mathbb{R}^{n_x}$ onto N, the components Qx correspond to the algebraic components z in (1) and are called *nullspace components*. Then $P := I - Q$ is again a projector and because of $A(t) = A(t)P$ the DAE (3) is equivalent to

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

This motivates to investigate in the perturbation analysis not only continuously differentiable functions but all functions of

$$\mathcal{C}_N^1[t_0, t_e] := \{\, x \in \mathcal{C}[t_0, t_e] \;:\; Px \in \mathcal{C}^1[t_0, t_e] \,\} \tag{4}$$

(i. e. the nullspace components Qx are not supposed to be differentiable). We consider functions $g\,:\,[t_0, t_e] \times \mathcal{D} \to \mathbb{R}^{n_x}$ that are continuous and have continuous partial derivatives g_x. Here $\mathcal{D} \subset \mathbb{R}^{n_x}$ is open with $\{\, x(t) \,:\, t \in [t_0, t_e] \,\} \subset \mathcal{D}$ and $Px \in \mathcal{D}$ for all $x \in \mathcal{D}$.

With these assumptions (3) can be linearized in $x(t)$, $(t \in [t_0, t_e])$, the (time varying) coefficients in the linearized DAE are $A(t)$ and

$$B(t) := g_x(t, x(t)) \quad , \tag{5}$$

respectively. The linearized DAE is supposed to be *tractable with index 2*, i. e.

$$S(t) \cap N \neq \emptyset \quad , \quad \dim N_1(t) = \text{const} \quad , \quad S_1(t) \oplus N_1(t) = \mathbb{R}^{n_x} \quad , \quad (t \in [t_0, t_e]) \quad . \tag{6}$$

Here we used the notations

$$\begin{aligned} S(t) &:= \{\, \xi \in \mathbb{R}^{n_x} \;:\; B(t)\xi \in \operatorname{im} A(t) \,\} \quad , \quad A_1(t) := A(t) + B(t)Q \quad , \\ S_1(t) &:= \{\, \xi \in \mathbb{R}^{n_x} \;:\; B(t)P\xi \in \operatorname{im} A_1(t) \,\} \quad , \quad N_1(t) := \ker A_1(t) \quad . \end{aligned}$$

In the same way in that vectors $x \in \mathbb{R}^{n_x}$ are splitted into Px and $Qx \in N = \ker A(t)$ ("differential" and "algebraic" or "nullspace" components) the differential components Px are splitted again: $Px = PP_1(t)x + PQ_1(t)x$ with $Q_1(t)x \in N_1(t) = \ker A_1(t)$. In this equation $Q_1(t)$ denotes the projector of $\mathbb{R}^{n_x}$ onto $N_1(t)$ along $S_1(t)$, i. e.

$$Q_1(t)Q \equiv 0 \ , \ \ (t \in [t_0, t_e]) \tag{7}$$

since $Q\xi \in S_1(t)$, $(\xi \in \mathbb{R}^{n_x})$. It is verified straightforwardly that $P_1(t) := I - Q_1(t)$, $PP_1(t)$ and $PQ_1(t)$ are projectors with $\mathbb{R}^{n_x} = \operatorname{im} PP_1(t) \oplus \operatorname{im} PQ_1(t) \oplus \operatorname{im} Q$ and

$$PP_1(t)P = PP_1(t) \quad , \quad QP_1(t)P = -QQ_1(t) \ . \tag{8}$$

The matrix

$$G_2(t) := A_1(t) + [BPQ_1](t) = A(t) + B(t)Q + [BPQ_1](t)$$

is non-singular in $[t_0, t_e]$, (see Section 6). Because of $QP = 0$ and $[Q_1P_1](t) \equiv 0$ the coefficient $A(t)$ can be splitted into

$$\begin{aligned} A(t) &= A(t)P = (A(t) + B(t)Q)P = A_1(t)P \ , \\ A(t) &= A_1(t)P = A_1(t)P_1(t)P = G_2(t)P_1(t)P \ . \end{aligned} \tag{9}$$

Thus the analytical solution $x(t)$ satisfies the algebraic constraint

$$[Q_1G_2^{-1}](t)g(t, x(t)) = 0 \ , \ \ (t \in [t_0, t_e]) \tag{10}$$

(multiply DAE (3) by $[Q_1G_2^{-1}](t)$). Furthermore we have

$$[G_2^{-1}B](t)Q = Q \tag{11}$$

since $G_2(t)Q = A(t)Q + B(t)Q^2 + [BPQ_1](t)Q = B(t)Q$.

In contrast to *linear* DAEs (3) conditions (6) are *not* sufficient to get tractability with index 2 in the *nonlinear* case. Tractability with index 2 can be ensured (in a neighbourhood of $x(t)$) if the *structural condition*

$$s(t, \xi) := [Q_1G_2^{-1}](t)g(t, \xi) = [Q_1G_2^{-1}](t)g(t, P\xi) \ , \ \ (\ t \in [t_0, t_e] \ , \ \xi \in \mathcal{D} \ ) \tag{12}$$

is satisfied, i. e. if $[Q_1G_2^{-1}](t)g(t, \xi)$ does not depend on the nullspace components $Q\xi$, (see [14, Section 3] for the discussion of singularities in an example of [6] that does not satisfy (12)). Condition (12) yields to (see Section 6)

$$Q_1(t) = [Q_1G_2^{-1}](t)g_x(t, Px(t))P \ . \tag{13}$$

Throughout the present paper we suppose that

- the linearization of (3) in $x(t)$ is tractable with index 2 (i. e. (6) is satisfied),
- the projector $Q_1(t)$ is continuously differentiable,
- the structural condition (12) is satisfied,
- function $s(t, x)$ has continuous partial derivatives w. r. t. t and x (see (12)) and
- the functions $g_x(t, .)$, $s_t(t, .)$ and $s_x(t, .)$ are lipschitzian w. r. t. x with Lipschitz constants that are independent of t.

Remark 1

a) With similar assumptions [14] prove the (unique) solvability of initial value problems $PP_1(t_0)(\xi(t_0)-\xi_0)=0$ for (3).

b) Straightforward computation verifies that the structural condition (12) is satisfied for linear DAEs that are tractable with index 2 and for DAEs (1) in Hessenberg form. Thus the class of DAEs that we consider includes systems of index 2 in Hessenberg form (provided that φ and γ are sufficiently differentiable and satisfy the index-2 condition (2)). Here Px corresponds to the differential components y and Qx to the algebraic components z.

c) März proves (in a more general framework) that coupling of additional *index-0 equations* ($h^{(0)}(t,x(t),z(t))=0$, $h^{(0)}_z$ non-singular) and *index-1 equations* ($h^{(1)}(t,x(t),z(t),z'(t))=0$, $h^{(1)}_{z'}$ non-singular) to DAE (3) with (12) results in an (enlarged) DAE that satisfies the structural condition again ([13]).

d) DAEs of index 2 that do *not* fit the structural condition (12) arise in mechanical multibody systems with contact conditions. The *index-2 formulation* of the model equations (y : position and velocity coordinates, z : Lagrangian multipliers) can be written as

$$\begin{aligned} y'(t) &= \varphi(t,y(t),p(t),z(t)) \\ 0 &= \gamma(y(t),p(t)) \\ 0 &= \sigma(y(t),p(t)) \end{aligned} \tag{14}$$

with (position) coordinates $p(t)$ of the contact point, that are uniquely defined by $y(t)$. Here σ_p is non-singular, i. e. (14) could be transformed into Hessenberg form ($\sigma(y,p)=0$ is solved w. r. t. p and $p=p(y)$ is inserted into φ and γ). In the present form system (14) satisfies the structural condition (12) if σ and the contact condition γ are linear. The contact conditions in model equations from railway system dynamics ([16], [3]) are, however, nonlinear in p and (14) violates the structural condition. Nevertheless all results of our perturbation analysis can be carried over to these equations since the BDF discretization yields to $\sigma(y_n,p_n)=0$, i. e. the analysis of error propagation for (14) is equivalent to the analysis for the corresponding system in Hessenberg form.

4 Perturbations in quasilinear DAEs of index 2

To analyse the error of discretization methods for DAEs of index 2 the error term caused by the *problem* (i. e. by index 2) has to be isolated from error terms introduced by the *discretization*. Following the theory of perturbation index we therefore compare *the analytical solution* $x(t)$, ($t\in[t_0,t_e]$) with functions $\hat{x}(t)$ that do not satisfy (3) but

$$A(t)\hat{x}'(t)+g(t,\hat{x}(t))=\delta(t) \tag{15}$$

with (small) residuals $\delta(t)$. In this section error bounds for $P(\hat{x}-x)$ and for $Q(\hat{x}-x)$ are given, that can be extended to error bounds for the BDF discretization (Section 5).

From Section 2 we expect a large influence of errors in the algebraic constraint. The analogue to the residual θ in the numerical test is here (see (10))

$$\theta(t):=[Q_1G_2^{-1}\delta](t) \ . \tag{16}$$

Furthermore smaller error bounds have to be expected for the solution components Px:

Theorem 1 *There is a constant C_0 such that the estimates*

$$\|P(\hat{x}(t)-x(t))\| \;\leq\; C_0\Big(\Delta_0 + \max_{\tau\in[t_0,t]}\|\int_{t_0}^{\tau} P[P_1G_2^{-1}\delta](\hat{\tau})\,d\hat{\tau}\| + \max_{\tau\in[t_0,t]}\|\theta(\tau)\| + D(t)^2\Big) \quad (17)$$

$$\|Q(\hat{x}(t)-x(t))\| \;\leq\; C_0(\,\Delta_0 + \max_{\tau\in[t_0,t]}\|\delta(\tau)\| + \max_{\tau\in[t_0,t]}\|\theta(\tau)\| + \max_{\tau\in[t_0,t]}\|\theta'(\tau)\|\,) \quad (18)$$

with

$$\Delta_0 := \|PP_1(t_0)(\hat{x}(t_0)-x(t_0))\| \quad , \qquad D(t) := \max_{\tau\in[t_0,t]}\|\delta(\tau)\| + \max_{\tau\in[t_0,t]}\|\theta'(\tau)\|$$

are satisfied for all $t\in[t_0,t_e]$ and for all $\hat{x}\in\mathcal{C}_N^1[t_0,t_e]$ with (15) provided that the expression at the right hand side of (17) is sufficiently small.

Proof The basic idea is to split the error $\hat{x}(t)-x(t)$ w. r. t. the images of projectors $PP_1(t)$, $PQ_1(t)$ and Q (part (i)) and to bound the error component $PP_1(t)(\hat{x}(t)-x(t))$ applying Gronwall's Lemma (part (ii)).

(i) With (9) and (8) multiplication of (3) and (15) by $P[P_1G_2^{-1}](t)$, $[Q_1G_2^{-1}](t)$ and $QG_2^{-1}(t)$, respectively, results in

$$\begin{aligned} PP_1(t)(P\hat{x}'(t)-Px'(t)) + P[P_1G_2^{-1}](t)\big(g(t,\hat{x}(t))-g(t,x(t))\big) &= P[P_1G_2^{-1}\delta](t)\,, && (19)\\ [Q_1G_2^{-1}](t)\big(g(t,\hat{x}(t))-g(t,x(t))\big) &= \theta(t)\,, && (20)\\ -\,QQ_1(t)(P\hat{x}'(t)-Px'(t)) + QG_2^{-1}(t)\big(g(t,\hat{x}(t))-g(t,x(t))\big) &= Q[G_2^{-1}\delta](t)\,. && (21)\end{aligned}$$

With the structural condition equation (20) is equivalent to

$$s(t,P\hat{x}(t)) - s(t,Px(t)) = \theta(t) \quad (22)$$

and (applying $\Psi(1)-\Psi(0)=\int_0^1\Psi'(\vartheta)\,d\vartheta$ to $\Psi(\vartheta):=s(t,Px_\vartheta(t))$ with $x_\vartheta(t) := x(t)+\vartheta(\hat{x}(t)-x(t))$) to

$$Q_1(t)P(\hat{x}(t)-x(t)) + [Q_1G_2^{-1}](t)\int_0^1\big(g_x(t,Px_\vartheta(t))-g_x(t,Px(t))\big)\,d\vartheta\cdot P(\hat{x}(t)-x(t)) = \theta(t)\;, \quad (23)$$

(see (13)). To transform (21) equation (22) is differentiated:

$$s_t(t,P\hat{x}(t)) - s_t(t,Px(t)) + s_x(t,P\hat{x}(t))P\hat{x}'(t) - s_x(t,Px(t))Px'(t) = \theta'(t)\;\;.$$

Because of $s_x(t,Px(t)) = [Q_1G_2^{-1}g_x](t,Px(t))$ we get with (13)

$$\begin{aligned}&\Big(Q_1(t) + \big(s_x(t,P\hat{x}(t)) - s_x(t,Px(t))\big)\Big)(P\hat{x}'(t)-Px'(t)) + \\ &\quad + \big(s_x(t,P\hat{x}(t)) - s_x(t,Px(t))\big)Px'(t) + s_t(t,P\hat{x}(t)) - s_t(t,Px(t)) = \theta'(t)\;\;.\end{aligned} \quad (24)$$

Similar to (23) the term $G_2^{-1}(t)(g(t,\hat{x})-g(t,x))$ in (19) and (21) can be expressed as

$$\begin{aligned}G_2^{-1}(t)\big(g(t,\hat{x}(t))-g(t,x(t))\big) &= G_2^{-1}(t)\cdot\int_0^1 g_x(t,x_\vartheta(t))\,d\vartheta\cdot P(\hat{x}(t)-x(t)) + \\ &+ Q(\hat{x}(t)-x(t)) + G_2^{-1}(t)\cdot\int_0^1\big(g_x(t,x_\vartheta(t))-g_x(t,x(t))\big)\,d\vartheta\cdot Q(\hat{x}(t)-x(t))\end{aligned} \quad (25)$$

since $g_x(t,x(t)) = B(t)$, (see (11)).

Let us suppose for the moment that $\hat{x}(t)$ is close to $x(t)$:

$$\|P(\hat{x}(t)-x(t))\| + \|Q(\hat{x}(t)-x(t))\| \leq \gamma \quad (26)$$

with a sufficiently small positive constant γ. At the end of the proof we show that the assumptions of the theorem imply (26). If γ in (26) is small then (23) results in

$$\|Q_1(t)(\hat{x}(t)-x(t))\| - \frac{1}{2}(\|PP_1(t)(\hat{x}(t)-x(t))\| + \|Q_1(t)(\hat{x}(t)-x(t))\|) \le \|\theta(t)\|$$

since $s_x(t,.) = [Q_1G_2^{-1}g_x](t,.)$ is lipschitzian (apply triangle inequality, $Q_1(t)P = Q_1(t)$ and $P(\hat{x}-x) = PP_1(t)(\hat{x}-x) + PQ_1(t)(\hat{x}-x)$). So we have

$$\|Q_1(t)(\hat{x}(t)-x(t))\| \le \|PP_1(t)(\hat{x}(t)-x(t))\| + 2\|\theta(t)\|$$

and

$$\|P(\hat{x}(t)-x(t))\| \le (1+\|P\|)\|PP_1(t)(\hat{x}(t)-x(t))\| + 2\|P\|\,\|\theta(t)\| \quad , \tag{27}$$

i. e. the error in the differential components Px is bounded in terms of the residual $\theta(t)$ and the error in PP_1x. In the same way

$$\|PP_1(t)(P\hat{x}'(t)-Px'(t))\| \le \mathcal{O}(1)(\|P(\hat{x}(t)-x(t))\| + \|\delta(t)\|) + \mathcal{O}(\gamma)\|Q(\hat{x}(t)-x(t))\| \tag{28}$$

follows from (19) and (25), (note that $PP_1(t)Q = PQ = 0$ since $Q_1(t)Q \equiv 0$). With this estimate equality (24) can be transformed to

$$\begin{aligned} \|Q_1(t)(P\hat{x}'(t)-Px'(t))\| &\le \mathcal{O}(1)(\|PP_1(t)(P\hat{x}'(t)-Px'(t))\| + \|P(\hat{x}(t)-x(t))\| + \|\theta'(t)\|) \\ &\le \mathcal{O}(1)(\|P(\hat{x}(t)-x(t))\| + D(t)) + \mathcal{O}(\gamma)\|Q(\hat{x}(t)-x(t))\| \end{aligned} \tag{29}$$

since $\|s_x(t,P\hat{x}) - s_x(t,Px)\| = \mathcal{O}(\gamma)$ and $P\hat{x}' - Px' = PP_1(P\hat{x}'-Px') + PQ_1(P\hat{x}'-Px')$. Estimate (29) is used to bound the error in the nullspace components Qx in terms of $D(t)$ and the error in the differential components Px : with (25), (29) and $QQ = Q$ we get from (21)

$$\|Q(\hat{x}(t)-x(t))\| \le \mathcal{O}(1)\big(\|P(\hat{x}(t)-x(t))\| + D(t)\big) \tag{30}$$

and

$$\|\hat{x}(t)-x(t)\| \cdot \|Q(\hat{x}(t)-x(t))\| \le \mathcal{O}(1)\big(\|P(\hat{x}(t)-x(t))\| + D(t)^2\big) \tag{31}$$

since $\hat{x} - x = P(\hat{x}-x) + Q(\hat{x}-x)$ and $D(t) \ll 1$. ($D(t)^2$ is part of the right hand side of (17) and thus supposed to be small.)

(ii) In part (i) we proved error bounds for components PQ_1x and Qx exploiting the structure of DAE (3). To complete the proof we now study the error propagation in components PP_1x following standard lines of ODE theory. We set $u(t) := PP_1(t)(\hat{x}(t)-x(t))$ and apply Gronwall's Lemma to

$$u(t) = u(t_0) + \int_{t_0}^{t} \frac{d}{d\tau}u(\tau)\,d\tau \quad .$$

With (19), (25), (27) and (31) the estimate

$$\begin{aligned} \frac{d}{d\tau}u(\tau) &= PP_1'(\tau)P(\hat{x}(\tau)-x(\tau)) + PP_1(\tau)(P\hat{x}'(\tau)-Px'(\tau)) \\ &= P[P_1G_2^{-1}\delta](\tau) + \mathcal{O}(1)(\|P(\hat{x}(\tau)-x(\tau))\| + \|\hat{x}(\tau)-x(\tau)\| \cdot \|Q(\hat{x}(\tau)-x(\tau))\|) \\ &= P[P_1G_2^{-1}\delta](\tau) + \mathcal{O}(1)\|u(\tau)\| + \mathcal{O}(1)(\|\theta(\tau)\| + D(\tau)) \end{aligned} \tag{32}$$

is obtained and proves (17) with a constant C_0 that depends on γ of (26) but is independent of t, $\hat{x}(t)$ and $\delta(t)$. Estimate (18) follows from (30).

Finally, we fix a sufficiently small value of γ in (26) and restrict the perturbation analysis to functions $\hat{x}(t)$ with residuals $\delta(t)$ such that the right hand sides of (17) and (18) are both $\leq \gamma/2$ for all $t \in [t_0, t_e]$ ("...provided that the expression at the right hand side of (17) is sufficiently small"). Then (17) and (18) prove (26) to be valid for all $t \in [t_0, t_e]$. ■

In Theorem 1 we have $D(t) \ll 1$, i. e. $D(t)^2 \ll D(t)$ and the bound for components Px is substantially smaller than for components Qx. The critical error term $D(t)$ that does not depend continuously on the residual $\delta(t)$ appears originally in the nullspace components Qx, only. It is introduced into the differential components Px via the nonlinear coupling with function $g(t,x)$. If this function is *linear* w. r. t. Qx with $g(t,x) = g_0(t,Px) + G(t)Qx$ (e. g. linear DAEs with time varying coefficients) then bound (17) can be further decreased and the differential components Px depend *continuously* on the residuals $\delta(t)$:

Corollary 1 *If* $g(t,x) = g_0(t,Px) + G(t)Qx$ *then the statement of Theorem 1 remains valid with* $D(t) := 0$ *.*

Proof With this special structure of $g(t,x)$ equation (25) simplifies to

$$G_2^{-1}(t)\big(g(t,\hat{x}(t)) - g(t,x(t))\big) = Q(\hat{x}(t) - x(t)) + G_2^{-1}(t) \cdot \big(g_0(t,P\hat{x}(t)) - g_0(t,Px(t))\big)$$

such that $\frac{d}{d\tau}u(\tau)$ in part (ii) of the proof of Theorem 1 does not depend on $Q(\hat{x}(t) - x(t))$. (Instead of (26) we use here $\|P(\hat{x}(t) - x(t))\| \leq \gamma$.) ■

Remark 2

a) Similar results were proved in [2] on the basis of [14, Theorem 3.4]. There we splitted the error bounds for components Px again w. r. t. PP_1x and PQ_1x. To prove these estimates we used instead of $G_2(t)$ the matrix $A_2(t) := A_1(t) + (B(t) - A_1(t)PP_1'(t))PQ_1(t)$ ([14, Section 2]; note, however, that in [14, Section 3] the definition $A_2(t) := G_2(t)$ in Lemma 3.1 is not compatible with transformation (3.9)). The slight improvement of the error bounds for components PP_1x might be interesting for linear DAEs but it makes the notation much more complicated. In the present paper we skip these details and give error bounds like (17) since our interest is focussed on the critical error term $\theta'(t)$ and on the equivalence of error bounds for analytical and numerical solution.

b) Theorem 1 and Corollary 1 generalize the corresponding results for systems (1) in Hessenberg form ([4, Theorem 1a,c] that is based on the proof of convergence for Runge–Kutta methods in [9]). These error bounds can in general not be improved (see [1, Example 2.2]).

c) For DAEs (1) in Hessenberg form with $\varphi(y,z) = \varphi_0(y) + \varphi_z(y)z$ (i. e. the system is linear in the algebraic components z but the coefficient φ_z depends on the differential components y) the error bounds for y (i. e. for Px) given in [4, Theorem 1b] are substantially smaller than that of Theorem 1. At the present moment we are not aware of a generalization of this result to quasilinear DAEs.

5 The global error of BDF applied to quasilinear DAEs of index 2

In this section variable stepsize variable order BDF

$$A(t_n)\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{n,i}x_{n-i} + g(t_n,x_n) = 0 \quad , \quad (n \geq k) \quad , \tag{33}$$

$x_n := x(t_n)$, $(n = 0, 1, \ldots, k-1)$ are considered. We prove convergence with the classical order from ODE theory and study the influence of small perturbations on the numerical solution. The key result of this section is Theorem 3 that gives error bounds for the numerical solution. These bounds can be interpreted as discrete analogues of Theorem 1 and explain the test results of Section 2. The results of this section are closely related to [17, Section 3]. Essential differences are improved error bounds for the differential components Px and less restrictive assumptions on the size of perturbations in Theorem 3.

The stepsize sequence h_n in (33) is defined by $h_n := t_n - t_{n-1}$ with $t_0 < t_1 < \ldots < t_N := t_e$. We suppose that h_n is bounded by $0 < h_{min} \le h_n \le h_{max}$, $(k \le n \le N)$ and that this stepsize sequence yields to a stable variable stepsize BDF for ODEs (see e. g. [10, Section III.5]). Then the coefficients $\alpha_{n,i}$ are bounded, the global error of the BDF applied to ODEs can be bounded in terms of the local error.

The numerical solution is given implicitly by (33), in each time step a system of nonlinear equations has to be solved w. r. t. x_n. The discretization method is well-defined since (33) is (locally) uniquely solvable for sufficiently small stepsizes h provided that the vectors x_{n-i}, $(i = 1, \ldots, k)$ are in a small neighbourhood of the analytical solution; more precisely

$$\gamma_P := \sum_{i=1}^{k} \|P(x_{n-i} - x(t_{n-i}))\| \quad \text{and} \quad \Theta := \sum_{i=1}^{k} \|[Q_1 G_2^{-1} g](t_{n-i}, x_{n-i})\| \tag{34}$$

are supposed to be small (it holds $[Q_1 G_2^{-1} g](t_{n-i}, x(t_{n-i})) = 0$, see (10)).

Theorem 2 *The system (33) of nonlinear equations is (locally) uniquely solvable for stepsizes $h_n \le h_0$ and vectors x_{n-i}, $(i = 1, \ldots, k)$ with*

$$\gamma_P \le \gamma_0 \sqrt{h_n} \quad , \quad \Theta \le \delta_0 h_n \tag{35}$$

and sufficiently small positive constants h_0, γ_0 and δ_0. The solution x_n satisfies

$$\|P(x_n - x(t_n))\| \le C_\gamma (h_n \|\tau_n\| + \gamma_P) \;\; , \;\; \|Q(x_n - x(t_n))\| \le C_\gamma (\|\tau_n\| + \gamma_P + \frac{1}{h_n}\gamma_P^2 + \frac{1}{h_n}\Theta) \tag{36}$$

with a constant C_γ that is independent of h_n, t_n, γ_P and Θ and the local discretization error

$$\tau_n := A(t_n) \frac{1}{h_n} \sum_{i=0}^{k} \alpha_{n,i} x(t_{n-i}) + g(t_n, x(t_n)) \quad .$$

Proof x_n can be computed by a simplified Newton method with matrix $\frac{\alpha_0}{h_n} A(t_n) + g_x(t_n, \xi^{(0)})$ and initial guess $\xi^{(0)} := x(t_n)$. The details are given in Section 6. ■

Theorem 2 proves that the system (33) of nonlinear equations is solvable under suitable assumptions. If this system is solved numerically then the actually computed approximation $\hat{x}_n$ of the solution will not satisfy (33) but

$$A(t_n) \frac{1}{h_n} \sum_{i=0}^{k} \alpha_{n,i} \hat{x}_{n-i} + g(t_n, \hat{x}_n) = \delta_n \tag{37}$$

with a (small) residuum δ_n that represents e. g. round-off errors and errors caused by stopping the iterative solution of (33). Similar to the definition of $\theta(t)$ in (16) we set

$$\theta_m := [Q_1 G_2^{-1}](t_m) \delta_m, \; (m \ge k) \;\; , \;\; \theta_m := Q_1(t_m)(\hat{x}_m - x(t_m)), \; (m = 0, 1, \ldots, k-1) \quad . \tag{38}$$

The small errors δ_n are propagated and amplified during integration. Theorem 3 gives bounds for the error $\hat{x}_n - x_n$ at a fixed time t_n in terms of the residuals δ_m, $(m \le n)$ and the errors $\hat{x}_m - x_m$, $(m = 0, 1, \ldots, k-1)$ in the initial values. Such error bounds can only be given if both sequences (x_n) and $(\hat{x}_n)$ remain in the neighbourhood of one and the same analytical solution of (3). This can be guaranteed if

$$\begin{aligned} \|P(\hat{x}_n - x(t_n))\| &\le C^*_\gamma (h_n + \gamma^*_P + h_n(\delta + \frac{1}{h}\theta)) \\ \|Q(\hat{x}_n - x(t_n))\| &\le C^*_\gamma (h_n + \gamma^*_P + \frac{1}{h_n}(\gamma^*_P)^2 + \delta + \frac{1}{h}\theta) \end{aligned} \tag{39}$$

is satisfied with

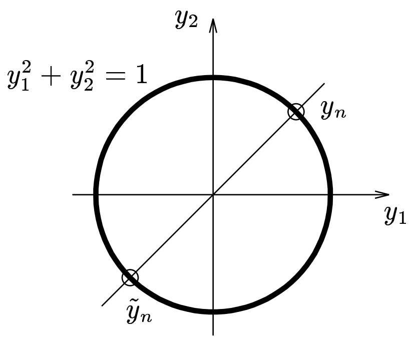
$$\delta := \max_m \|\delta_m\| \ , \ \theta := \max_m \|\theta_m\| \ , \ \frac{1}{h}\theta := \max_m \frac{1}{h_m}\|\theta_m\| \ , \ \gamma^*_P := \sum_{i=1}^{k} \|P(\hat{x}_{n-i} - x(t_{n-i}))\|$$

and a constant C^*_γ that is independent of h_n, t_n, δ, $\frac{1}{h}\theta$ and γ^*_P (see also (36)).

Example 1 The initial value problem $y_1(t_0) = y_1^{(0)}$ for DAE

$$y_1'(t) = -2y_1 z \ , \ y_2'(t) = -2y_2 z \ , \ y_1^2 + y_2^2 = 1$$

has the unique solution $y(t) \equiv y(t_0)$, $z(t) \equiv 0$. Backward Euler method solves this DAE exactly but (33) is only locally but not globally uniquely solvable. For a given vector $y_{n-1} = (\, y_{n-1,1} \, , \, y_{n-1,2} \,)^T$ solutions are $y_n = y_{n-1}$, $z_n = 0$ and $\tilde{y}_n = -y_{n-1}$, $\tilde{z}_n = \frac{1}{h_n}$ (see figure). Error bounds for $\hat{y}_n - y_n$ can be given, obviously, for vectors $\hat{y}_n$ in a neighbourhood of y_n but not for those in a neighbourhood of $\tilde{y}_n$. Condition (39) is an equivalent to the condition $\hat{x} \in \mathcal{C}^1_N[t_0, t_e]$ in Theorem 1.



Following the ideas of Theorem 2 it is proved that there is always (at least) one vector $\hat{x}_n$ that satisfies (39) if $h_n \le h_0$,

$$\gamma^*_P \le \gamma_0 \sqrt{h_n} \ , \ \delta + \frac{1}{h}\theta \le \delta_0$$

and the (positive) constants h_0, γ_0 and δ_0 are sufficiently small. Therefore the perturbation analysis is restricted to sequences $(\hat{x}_n)$ with (39).

Theorem 3 *There is a constant C_0 such that for all sequences $(\hat{x}_n)$ with (37), (39) and*

$$\begin{aligned} \Delta^h_0 := \sum_{m=0}^{k-1} \|PP_1(t_m)(\hat{x}_m - x(t_m))\| &\le \gamma_0 h_{min}^{2/3} \ , \\ \|P[P_1 G_2^{-1}](t_m)\delta_m\| &\le \gamma_0 h_{min}^{2/3} \ , \ (m \ge k) \ , \\ D_h := \delta + \frac{1}{h}\theta &\le \gamma_0 h_{min}^{1/3} \end{aligned} \tag{40}$$

the estimates

$$\|P(\hat{x}_n - x_n)\| \le C_0(\Delta^h_0 + \sum_{m=k}^{n} h_m \|P[P_1 G_2^{-1}](t_m)\delta_m\| + h_{max}\delta + \theta + D_h^2) =: \Delta_P \tag{41}$$

$$\|Q(\hat{x}_n - x_n)\| \le C_0(\Delta^h_0 + \delta + \frac{1}{h}\theta + \frac{1}{h}\Delta_P^2) \tag{42}$$

are satisfied for all n with $k \leq n \leq N$ provided that h_{max} and the constant γ_0 in (40) are sufficiently small.

Proof Theorem 3 is the discrete counterpart to Theorem 1 and we will follow the proof of Theorem 1 as far as possible. Again we give in part (i) estimates for the error in components Q_1x and Qx and study in part (ii) the error propagation in components PP_1x with standard techniques from ODE theory. This includes the proof of convergence for BDF applied to (3).

(i) Multiplying (33) and (37) by $P[P_1G_2^{-1}](t_n)$, $[Q_1G_2^{-1}](t_n)$ and $QG_2^{-1}(t_n)$, respectively, the discrete versions of (19), (20), (21) are obtained. At the moment we restrict ourselves to sequences (x_n), $(\hat{x}_n)$ with

$$\|P(x_m - x(t_m))\| \leq C_\gamma h_{min}^{2/3} \ , \ \|P(\hat{x}_m - x(t_m))\| \leq C_\gamma h_{min}^{2/3} \ , \ (k \leq m < n) \ , \tag{43}$$

(see (26)). Then the nonlinear system of equations (33) is uniquely solvable and both x_n and $\hat{x}_n$ are close to $x(t_n)$ if h_{max} and γ_0 are sufficiently small (Theorem 2 and condition (39)). Following step by step the proof for the analytical solution we get the estimates

$$[Q_1G_2^{-1}](t_n)(g(t_n, \hat{x}_n) - g(t_n, x_n)) = \theta_n \tag{44}$$

(see (20)),

$$\|P(\hat{x}_n - x_n)\| \leq (1 + \|P\|)\|PP_1(t_n)(\hat{x}_n - x_n)\| + 2\|P\|\,\|\theta_n\| \tag{45}$$

(see (27)) and

$$\|Q(\hat{x}_n - x_n)\| \leq \mathcal{O}(1)\|QQ_1(t_n)\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{n,i}P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(1)(\|P(\hat{x}_n - x_n)\| + \|\delta_n\|) \tag{46}$$

(see (21) and (25)). A bound for $\frac{1}{h_n}Q_1(t_n)P(\hat{x}_{n-i} - x_{n-i})$ is obtained from $Q_1(t_n) = Q_1(t_{n-i}) + \mathcal{O}(h_n)$, $Q_1(t_m)(\hat{x}_m - x_m) = \theta_m$, $(m < k)$ and

$$\begin{aligned} Q_1(t_m)P(\hat{x}_m - x_m) &= Q_1(t_m)P(\hat{x}_m - x_m) - \big([Q_1G_2^{-1}g](t_m, P\hat{x}_m) - [Q_1G_2^{-1}g](t_m, Px_m)\big) + \theta_m \\ &= [Q_1G_2^{-1}](t_m) \cdot \int_0^1 \big(g_x(t_m, Px(t_m)) - g_x(t_m, Px_m^\vartheta)\big)\, d\vartheta \cdot P(\hat{x}_m - x_m) + \theta_m \end{aligned}$$

with $x_m^\vartheta := x_m + \vartheta(\hat{x}_m - x_m)$, ($m \geq k$, see (44)), i. e.

$$\begin{aligned} &\|\frac{1}{h_n}Q_1(t_n)P(\hat{x}_{n-i} - x_{n-i})\| \leq \\ &\qquad \leq \mathcal{O}(1)\Big(1 + \frac{1}{h_n}(\|P(\hat{x}_{n-i} - x_{n-i})\| + \gamma_P^{n-i})\Big)\|P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(\frac{1}{h}\theta) \end{aligned} \tag{47}$$

with $\gamma_P^{n-i} := \|P(\hat{x}_{n-i} - x(t_{n-i}))\|$. Note, that this estimate substitutes (28), (29) since there is for $k > 1$ no bound like (28) for $PP_1(t_n)\frac{1}{h_n}\sum_i \alpha_{n,i}(\hat{x}_{n-i} - x_{n-i})$. We get with (43), (46) and (47)

$$\begin{aligned} \|Q(\hat{x}_n - x_n)\| &\leq \mathcal{O}(1)\sum_{i=0}^{k}(1 + \frac{1}{h_n}(\|P(\hat{x}_{n-i} - x_{n-i})\| + \gamma_P^{n-i}))\|P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(D_h) \\ &\leq \mathcal{O}(1)(\frac{1}{h_n^{1/3}}\sum_{i=0}^{k}\|P(\hat{x}_{n-i} - x_{n-i})\| + D_h) \ \leq \ \mathcal{O}(h_n^{1/3}) \end{aligned} \tag{48}$$

and

$$
\begin{aligned}
\|\hat{x}_n - x_n\|\,\|Q(\hat{x}_n - x_n)\| &\leq (\|P(\hat{x}_n - x_n)\| + \|Q(\hat{x}_n - x_n)\|)\|Q(\hat{x}_n - x_n)\| \\
&\leq \mathcal{O}(1)(\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| + D_h^2) \qquad (49)
\end{aligned}
$$

as analogues of (30) and (31), respectively, since both $\frac{1}{h_n^{1/3}} \sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\|$ and D_h are of size $\mathcal{O}(h_n^{1/3})$. Estimates (45) and (48) show, that – as for the analytical solution – the errors in components Q_1x and Qx are bounded in terms of δ, θ, $\frac{1}{h}\theta$ and the errors in components PP_1x.
(ii) In this part we prove the estimate

$$
\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{n,i} u_{n-i} \leq \mathcal{O}(1) \sum_{i=0}^{k} u_{n-i} + \|P[P_1 G_2^{-1}](t_n)\delta_n\| + \mathcal{O}(1)(h_{max}\delta + \theta + D_h^2) \qquad (50)
$$

for the sequence $u_n := PP_1(t_n)(\hat{x}_n - x_n)$. By standard results from ODE theory (e. g. [10, Theorem 5.8]) this implies

$$
u_n \leq \tilde{C}_0(\Delta_0^h + \sum_{m=k}^{n} h_m \|P[P_1 G_2^{-1}](t_m)\delta_m\| + h_{max}\delta + \theta + D_h^2) \qquad (51)
$$

and thus with (45) error bound (41) in the theorem. At the end we verify that the additional condition (43) is always satisfied if h_{max} and γ_0 in the theorem are sufficiently small.
The discrete counterpart of $\frac{d}{d\tau}u(\tau)$ is given by

$$
\begin{aligned}
&\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{n,i} u_{n-i} = \\
&= PP_1(t_n)\frac{1}{h_n}\sum_{i=0}^{k} \alpha_{n,i}(\hat{x}_{n-i} - x_{n-i}) + \frac{1}{h_n}\sum_{i=1}^{k} \alpha_{n,i} P(P_1(t_{n-i}) - P(t_n))P(\hat{x}_{n-i} - x_{n-i}) \\
&= -P[P_1 G_2^{-1}](t_n)(g(t_n, \hat{x}_n) - g(t_n, x_n)) + P[P_1 G_2^{-1}](t_n)\delta_n + \mathcal{O}(1)\sum_{i=1}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| \\
&\leq \mathcal{O}(1)\|P(\hat{x}_n - x_n)\| + \mathcal{O}(1)(\|\hat{x}_n - x_n\| + \|\hat{x}_n - x(t_n)\|)\|Q(\hat{x}_n - x_n)\| + \qquad (52) \\
&\quad + \|P[P_1 G_2^{-1}](t_n)\delta_n\| + \mathcal{O}(1)\sum_{i=1}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\|
\end{aligned}
$$

(see (19), (25) and (32)).
Consider now the special sequence $\hat{x}_n := x(t_n)$. If the analytical solution $x(t)$ is inserted in (37) then the residual δ_n is given by the local discretization error $\delta_n = \tau_n = \mathcal{O}(h_{min})$,
$\theta_n = [Q_1 G_2^{-1}](t_n)\tau_n = [Q_1 G_2^{-1} g](t_n, x(t_n)) = 0$ and with (45), (49) estimate (52) simplifies to

$$
\begin{aligned}
\frac{1}{h_n}\sum_{i=0}^{k} \alpha_{n,i} u_{n-i} &\leq \mathcal{O}(1)\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(1)\|\hat{x}_n - x_n\|\,\|Q(\hat{x}_n - x_n)\| + \mathcal{O}(\|\tau_n\|) \\
&\leq \mathcal{O}(1)\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(\|\tau_n\|) \leq \mathcal{O}(1)\sum_{i=0}^{k} u_{n-i} + \mathcal{O}(\|\tau_n\|) \ .
\end{aligned}
$$

With (51) we get $u_n \leq \tilde{C}_0 \max_m \|\tau_m\|$ and thus (see (45) and (48))

$$
\|x_n - x(t_n)\| = \|x_n - \hat{x}_n\| \leq C_0 \max_m \|\tau_m\| \qquad (53)
$$

for $k \le n \le N$ provided that (43) is satisfied for all $m < n$. Finally we prove by induction that condition (43) is valid for all $m \le N$ if h_{max} is chosen sufficiently small to get $C_0 \max_m \|\tau_m\| \le C_\gamma h_{min}^{2/3}$ in (53). So we have – as a by-product of the perturbation analysis – the proof of convergence for BDF applied to DAEs (3) of index 2.
Let us return to arbitrary sequences $(\hat{x}_n)$ with (37), (39). Because of $\hat{x}_m - x(t_m) = (\hat{x}_m - x_m) + (x_m - x(t_m)) = \hat{x}_m - x_m + \mathcal{O}(h_{min})$ inequalities (48) and (52) can be transformed to

$$\|Q(\hat{x}_n - x_n)\| \le \mathcal{O}(1)\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(\frac{1}{h_n})\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\|^2 + \mathcal{O}(D_h) \qquad (54)$$

and

$$\frac{1}{h_n}\sum_{i=0}^{k} \alpha_{n,i} u_{n-i} \le \mathcal{O}(1)\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| + \mathcal{O}(1)(h_n + \|\hat{x}_n - x_n\|)\|Q(\hat{x}_n - x_n)\| + \\ + \|P[P_1G_2^{-1}](t_n)\delta_n\| \ , \qquad (55)$$

respectively. The combination of (55) with (49) and (54) results in inequality (50) and – since the BDF was supposed to be stable – in error bound (41) if the additional condition (43) is satisfied for all $m < n$. Then there is a constant C_1 with

$$\|P(\hat{x}_n - x(t_n))\| \le \|P(\hat{x}_n - x_n)\| + \|P(x_n - x(t_n))\| \le C_1(\gamma_0 h_{min}^{2/3} + h_{min})$$

(see assumption (40)). If h_{max} and γ_0 are sufficiently small ($C_1(\gamma_0 h_{min}^{2/3} + h_{min}) \le C_\gamma h_{min}^{2/3}$) then $\|P(\hat{x}_n - x(t_n))\| \le C_\gamma h_{min}^{2/3}$ can again be proved by induction. Eventually the proof is completed inserting error bound (41) for the differential components Px into estimate (54) to get error bound (42) for the nullspace components. ■

As for the analytical solution the error bound for the differential components Px is – especially in the critical case $h \ll 1$ – substantially smaller than that for the nullspace components Qx, $((\frac{1}{h}\theta)^2 \ll \frac{1}{h}\theta)$. For special types of DAEs the differential components may even depend continuously on the perturbations:

Corollary 2 *If $g(t,x) = g_0(t,Px) + G(t)Qx$ then the statements of Theorems 2 and 3 remain valid if the definition of D_h is changed to $D_h := 0$ and the conditions (35) and (40) are substituted by the less restrictive conditions "$\gamma_P \le \gamma_0$" and*

$$\Delta_0^h \le \gamma_0 \ , \ \|P[P_1G_2^{-1}](t_m)\delta_m\| \le \gamma_0 \ , \ (m \ge k) \ , \ \theta \le \gamma_0 \ ,$$

respectively.

Proof
a) Because of the special structure of $g(t,x)$ the simplified Newton method solves the system (33) of nonlinear equations in one single step w. r. t. the nullspace components (independently of the initial guess): $\Phi(\xi) = \Phi(P\xi)$ and $J(\Phi(\xi_2) - \Phi(\xi_1)) = \mathcal{O}(1)\|P(\xi_2 - \xi_1)\|$ instead of (59) in the proof of Theorem 2.
b) Similar to the analytical solution $\frac{1}{h_n}\sum_i \alpha_{n,i}u_{n-i} = \frac{1}{h_n}\sum_i \alpha_{n,i}PP_1(t_{n-i})(\hat{x}_{n-i} - x_{n-i})$ is independent of $Q(\hat{x}_n - x_n)$ since $P[P_1G_2^{-1}](t) \cdot G(t)Q = P[P_1G_2^{-1}B](t)Q = PQ = 0$. We get instead of (55)

$$\frac{1}{h_n}\sum_{i=0}^{k} \alpha_{n,i} u_{n-i} \le \mathcal{O}(1)\sum_{i=0}^{k} \|P(\hat{x}_{n-i} - x_{n-i})\| + \|P[P_1G_2^{-1}](t_n)\delta_n\|$$

and the corollary is proved following the lines of the proof of Theorem 3. ■

Remark 3

a) Theorem 3 and Corollary 2 generalize corresponding results for backward Euler method applied to DAEs (1) in Hessenberg form ([4, Theorem 3a,c]). Both analytically given examples ([1, Example 3.2]) and the numerical tests of Section 2 illustrate that the error bounds reflect the influence of perturbations on the numerical solution correctly for *all* solution components. In general these error bounds are optimal, i. e. they can not be improved. As for the analytical solution smaller error bounds can be proved for systems (1) in Hessenberg form with $\varphi(y,z) = \varphi_0(y) + \varphi_z(y)z$ (see Remark 2c and [4, Theorem 3b]).

b) The sensitivity of BDF discretizations w. r. t. small errors was studied in [14], [17] for quasilinear DAEs (3) and – in the context of convergence analysis – by several authors for systems (1) in Hessenberg form (see [11, Section VI.6] for a list of references). These results differ from Theorem 3 by the more restrictive assumption "$\delta + \frac{1}{h}\theta = \mathcal{O}(h_{min})$" instead of (40). This additional assumption simplifies the proofs of Theorems 2 and 3 substantially and results in a bound Δ_P for the differential components Px that depends *continuously* on the perturbations since $D_h^2 = (\delta + \frac{1}{h}\theta)^2 = \mathcal{O}(h_{min}(\delta + \frac{1}{h}\theta)) = \mathcal{O}(h_{min}\delta + \theta)$ (see (41)). For systems in Hessenberg form this estimate could be obtained generalizing [11, Theorem 6.5], for quasilinear DAEs the error bound $\Delta_P = \mathcal{O}(\ldots + D_h)$ of [17, Theorem 3.1] is improved by the factor h.

c) The less restrictive assumption "$\delta + \frac{1}{h}\theta \leq \gamma_0 h_{min}^{1/3}$" in our perturbation analysis was motivated by numerical tests with very small stepsizes (region C in Figure 1) that explain why variable stepsize codes often fail if the stepsize becomes too small (even if the algebraic components are excluded from the stepsize control). The results for perturbations of size $D_h = \mathcal{O}(h_{min})$ seem to be misleading since they suggest that the differential components depend continuously on perturbations and the additional error term appears in the algebraic components, only. Thus it makes sense to analyse error propagation for all perturbations with $D_h \leq \gamma_0 h_{min}^{\alpha}$, and $\alpha \geq 0$ as small as possible. A lower limit is given by the conditions that ensure solvability of the BDF equations ($D_h^2 \leq \gamma_0 h_{min}^{1/2}$, i. e. $D_h \leq \gamma_0 h_{min}^{1/4}$, see Example 2).

6 Technical details

In this section we summarize basic properties of the various projectors that are used in the theoretical investigations and give the proof of Theorem 2.

Lemma 1 (cf. [14, Appendix])

a) Matrix $G_2(t) = A_1(t) + B(t)PQ_1(t) = A(t) + B(t)Q + B(t)PQ_1(t)$ *is non-singular.*

b) It holds $Q_1(t) = [Q_1 G_2^{-1} BP](t) = [G_2^{-1} BPQ_1](t)$.

c) The structural condition (12) implies $Q_1(t) = [Q_1 G_2^{-1}](t) g_x(t, Px(t))P$.

Proof

a) If $G_2(t)\xi = 0$ then $A_1(t)\xi + [BPQ_1](t)\xi = 0$ and $B(t)P \cdot Q_1(t)\xi = -A_1(t)\xi \in \operatorname{im} A_1(t)$, i. e. $Q_1(t)\xi \in S_1(t)$ (see Section 3). $Q_1(t)$ projects along $S_1(t)$ onto $N_1(t)$, i. e. $Q_1(t) \cdot Q_1(t)\xi = 0$, $Q_1(t)\xi = 0$ and because of $G_2(t)\xi = 0$ also $A_1(t)\xi = 0$.

Thus $\xi \in \ker A_1(t)$ and $\xi = Q_1(t)\xi = 0$, i. e. matrix $G_2(t)$ is non-singular.

b) We have $[G_2Q_1](t) = [A_1Q_1](t) + [BPQ_1](t) = [BPQ_1](t)$ and thus $Q_1(t) = [G_2^{-1}BPQ_1](t)$ and $[Q_1G_2^{-1}A_1](t) = [Q_1G_2^{-1}](t)(G_2(t) - [BPQ_1](t)) = Q_1(t)(I - Q_1(t)) = 0$.
To prove $Q_1(t)\xi = [Q_1G_2^{-1}BP](t)\xi$ we split $\xi \in \mathbb{R}^{n_x}$ into $\xi = \xi_S + \xi_N$ with $\xi_S \in S_1(t)$, $\xi_N \in N_1(t)$, (remember that $S_1(t) \oplus N_1(t) = \mathbb{R}^{n_x}$). By definition of $S_1(t)$ there is a $\zeta \in \mathbb{R}^{n_x}$ with $B(t)P\xi_S = A_1(t)\zeta$, i. e. $[Q_1G_2^{-1}BP](t)\xi_S = [Q_1G_2^{-1}A_1](t)\zeta = 0 \cdot \zeta = 0$. On the other hand we get $\xi_N = Q_1(t)\xi_N$, i. e.

$$[Q_1G_2^{-1}BP](t)\xi = [Q_1G_2^{-1}BP](t)\xi_N = [Q_1G_2^{-1}BPQ_1](t)\xi_N = [Q_1Q_1](t)\xi_N = Q_1(t)\xi_N = Q_1(t)\xi .$$

c) With $s(t,\xi) = s(t,P\xi)$ we get $s_x(t,\xi) = s_x(t,P\xi)P$ and

$$Q_1(t) = [Q_1G_2^{-1}B](t)P = s_x(t,x(t))P = s_x(t,Px(t))P^2 = [Q_1G_2^{-1}](t)g_x(t,Px(t))P . \qquad \blacksquare$$

Proof of Theorem 2
The simplified Newton method $\xi^{(0)} := x(t_n)$, $\xi^{(l+1)} := \xi^{(l)} - J^{-1} \cdot F(\xi^{(l)})$, $(l \geq 0)$,

$$F(\xi) := A(t_n)(\frac{\alpha_0}{h}\xi + \frac{1}{h}\sum_{i=1}^{k}\alpha_i x_{n-i}) + g(t_n,\xi)$$

and $J := \frac{\alpha_0}{h}A(t_n) + B(t_n)(PQ_1(t_n) + Q)$ induces a mapping $\Phi : \mathcal{U} \to \mathcal{U}$, $\Phi(\xi) := \xi - J^{-1}F(\xi)$ in a suitable neighbourhood $\mathcal{U}$ of $x(t_n)$, (for simplicity we omit index "n" in h and α_i). In part a) we prove that Φ is contractive in $\mathcal{U}$ w. r. t. the scaled norm $\|\xi\|_h := \|P\xi\| + h\|Q\xi\|$. In part b) we show that the sequence $(\xi^{(l)})$ of Newton iterates remains in $\mathcal{U}$ and converges to a solution of (33) (Fixed point Theorem). Throughout the proof we use again the splitting of vectors w. r. t. the images of $PP_1(t_n)$, $PQ_1(t_n)$ and Q as basic tool (see proof of Theorem 1).
a) If h is sufficiently small then J is non-singular since $J = G_2(t_n)(P_1(t_n) + \frac{h}{\alpha_0}Q_1(t_n))(\frac{\alpha_0}{h}P + Q)$ and

$$J^{-1} = (\frac{h}{\alpha_0}P + Q)(P_1(t_n) + \frac{\alpha_0}{h}Q_1(t_n))G_2^{-1}(t_n) \quad . \tag{56}$$

Newton iteration starts from $\xi^{(0)} := x(t_n)$, i. e.

$$F(\xi^{(0)}) = \tau_n + A(t_n)\frac{1}{h}\sum_{i=1}^{k}\alpha_i(x_{n-i} - x(t_{n-i})) \quad .$$

With (9) and (10) we get $[Q_1G_2^{-1}](t_n)F(\xi^{(0)}) = 0$, $hP[P_1G_2^{-1}](t_n)F(\xi^{(0)}) = \mathcal{O}(h\|\tau_n\|) + \mathcal{O}(\gamma_P)$ and

$$\begin{aligned} QG_2^{-1}(t_n)F(\xi^{(0)}) &= \mathcal{O}(\|\tau_n\|) + \mathcal{O}(\frac{1}{h})\|QQ_1(t_n)\sum_{i=1}^{k}\alpha_i(x_{n-i} - x(t_{n-i}))\| \\ &= \mathcal{O}(\|\tau_n\|) + \mathcal{O}(\gamma_P) + \mathcal{O}(\frac{1}{h}\gamma_P^2) + \mathcal{O}(\frac{1}{h}\Theta) \end{aligned}$$

since $Q_1(t_n) = Q_1(t_{n-i}) + \mathcal{O}(h)$, $s(t_{n-i}, x_{n-i}) = \mathcal{O}(\Theta)$, $s(t_{n-i}, x(t_{n-i})) = 0$, (see the transformation from (22) to (23)). Summarizing these estimates the bounds

$$\begin{aligned} \|\xi^{(1)} - \xi^{(0)}\|_h &= \|P(\xi^{(1)} - \xi^{(0)})\| + h\|Q(\xi^{(1)} - \xi^{(0)})\| \;\leq\; C_1(h\|\tau_n\| + \gamma_P) \\ \|Q(\xi^{(1)} - \xi^{(0)})\| &\leq C_1(\|\tau_n\| + \gamma_P + \frac{1}{h}\gamma_P^2 + \frac{1}{h}\Theta) \end{aligned}$$

with a constant C_1 are obtained from $\xi^{(1)} - \xi^{(0)} = -J^{-1}F(\xi^{(0)})$, (56) and $g(t_{n-i}, x(t_{n-i})) = 0$, i. e. $h \cdot \frac{1}{h}\Theta = \mathcal{O}(\gamma_P)$. Note that $\tau_n = \mathcal{O}(h_n^k)$ if the analytical solution $x(t)$ is sufficiently differentiable.
To prove contractivity of Φ w. r. t. $\|.\|_h$ we consider now vectors

$$\xi \in \mathcal{U} := \{ \xi \,:\, \|P(x_n - x(t_n))\| \le C_\gamma(h\|\tau_n\| + \gamma_P) \,,\ \|Q(x_n - x(t_n))\| \le C_\gamma(\|\tau_n\| + \gamma_P + \frac{1}{h}\gamma_P^2 + \frac{1}{h}\Theta) \}$$

with a constant $C_\gamma \ge 2C_1$ that is specified below. The assumptions of the theorem guarantee that $\xi \in \mathcal{U}$ is close to $x(t)$ if h_0, γ_0 and δ_0 are small. Let $\xi_1, \xi_2 \in \mathcal{U}$ be given. To bound $\Phi(\xi_2) - \Phi(\xi_1)$ we split J^{-1} into $(PP_1(t_n) + PQ_1(t_n) + Q)J^{-1}$ and get

$$PP_1(t_n)J^{-1} = \frac{h}{\alpha_0}P[P_1G_2^{-1}](t_n),\ Q_1(t_n)J^{-1} = [Q_1G_2^{-1}](t_n),\ QJ^{-1} = QG_2^{-1}(t_n) + (\frac{\alpha_0}{h} - 1)QQ_1J^{-1}. \tag{57}$$

The transformation (22) $\to$ (23) results here in

$$\begin{aligned} Q_1(t_n)(\Phi(\xi_2) - \Phi(\xi_1)) &= Q_1(t_n)(\xi_2 - \xi_1) - Q_1(t_n)J^{-1}(F(\xi_2) - F(\xi_1)) \\ &= Q_1(t_n)(\xi_2 - \xi_1) - (s(t_n, \xi_2) - s(t_n, \xi_1)) \\ &\le \mathcal{O}(1)(\|P(\xi_1 - x(t_n))\| + \|P(\xi_2 - x(t_n))\|)\|P(\xi_2 - \xi_1)\| \ . \end{aligned} \tag{58}$$

Furthermore matrix J was defined such that

$$\begin{aligned} J(\Phi(\xi_2) - \Phi(\xi_1)) &= J(\xi_2 - \xi_1) - (F(\xi_2) - F(\xi_1)) \\ &= B(t_n)(PQ_1(t_n) + Q)(\xi_2 - \xi_1) - (g(t_n, \xi_2) - g(t_n, \xi_1)) \\ &\le \mathcal{O}(1)\|P(\xi_2 - \xi_1)\| + \mathcal{O}(1)(\|\xi_1 - x(t_n)\| + \|\xi_2 - x(t_n)\|)\|Q(\xi_2 - \xi_1)\| \ , \end{aligned} \tag{59}$$

(see also (25)). Combining (57)–(59) we obtain for $\xi_1, \xi_2 \in \mathcal{U}$

$$\begin{aligned} \|P(\Phi(\xi_2) - \Phi(\xi_1))\| &\le \|PP_1(t_n)J^{-1} \cdot J(\Phi(\xi_2) - \Phi(\xi_1))\| + \|PQ_1(t_n)(\Phi(\xi_2) - \Phi(\xi_1))\| \\ &\le (\mathcal{O}(h) + \mathcal{O}(1) \cdot C_\gamma(h_0 + \gamma_0\sqrt{h_0}))\|P(\xi_2 - \xi_1)\| + \\ &\quad + \mathcal{O}(h) \cdot C_\gamma(h_0 + \gamma_0\sqrt{h_0} + \gamma_0^2 + \delta_0)\|Q(\xi_2 - \xi_1)\| \ , \\ \|Q(\Phi(\xi_2) - \Phi(\xi_1))\| &\le \\ &\le \|QG_2^{-1}(t_n) \cdot J(\Phi(\xi_2) - \Phi(\xi_1))\| + \|(\frac{\alpha_0}{h} - 1)Q\|\,\|Q_1(t_n)(\Phi(\xi_2) - \Phi(\xi_1))\| \\ &\le (\mathcal{O}(1) + \mathcal{O}(\frac{1}{h}) \cdot C_\gamma(h_0 + \gamma_0\sqrt{h_0}))\|P(\xi_2 - \xi_1)\| + \\ &\quad + \mathcal{O}(1) \cdot C_\gamma(h_0 + \gamma_0\sqrt{h_0} + \gamma_0^2 + \delta_0)\|Q(\xi_2 - \xi_1)\| \ , \end{aligned} \tag{60}$$

i. e. $\|\Phi(\xi_2) - \Phi(\xi_1)\|_h \le \frac{1}{2}\|\xi_2 - \xi_1\|_h$ if h_0, γ_0 and δ_0 are sufficiently small.
b) Fixed point Theorem proves convergence of $(\xi^{(l)})_{l\ge 0}$ to a vector $x_n \in \mathcal{U}$ with $F(x_n) = 0$ and

$$\|x_n - x(t_n)\|_h = \|x_n - \xi^{(0)}\|_h \le \sum_{l=0}^{\infty} \|\xi^{(l+1)} - \xi^{(l)}\|_h \le \sum_{l=0}^{\infty} \frac{1}{2^l}\|\xi^{(1)} - \xi^{(0)}\|_h \le 2C_1(h\|\tau_n\| + \gamma_P)$$

provided that $(\xi^{(l)})$ *remains in* $\mathcal{U}$. Thus the proof is completed if $\|Q(\xi^{(l)} - x(t_n))\| \le C_\gamma\Delta_Q$, $(l \ge 0)$ with $\Delta_Q := \|\tau_n\| + \gamma_P + \frac{1}{h}\gamma_P^2 + \frac{1}{h}\Theta$. For $\xi_1 := x(t_n)$, $\xi_2 = \xi \in \mathcal{U}$ we get in (58) and (60)

$$\begin{aligned} Q_1(t_n)(\Phi(\xi) - \Phi(x(t_n))) &= \mathcal{O}(1)\|P(\xi - x(t_n))\|^2 \\ \|Q(\Phi(\xi) - \Phi(x(t_n)))\| &\le \mathcal{O}(1)\|P(\xi - x(t_n))\| + \mathcal{O}(\frac{1}{h})\|P(\xi - x(t_n))\|^2 + \frac{1}{2}\|Q(\xi - x(t_n))\| \end{aligned}$$

and thus (if $\xi^{(l)} \in \mathcal{U}$)

$$\begin{aligned}
\|Q(\xi^{(l+1)} - x(t_n))\| &\le \|Q(\Phi(\xi^{(l)}) - \Phi(x(t_n)))\| + \|Q(\Phi(x(t_n)) - x(t_n))\| \\
&\le \frac{1}{2}\|Q(\xi^{(l)} - x(t_n))\| + \\
&\quad + \mathcal{O}(1)(\|P(\xi^{(l)} - x(t_n))\| + \frac{1}{h}\|P(\xi^{(l)} - x(t_n))\|^2 + \|Q(\xi^{(1)} - \xi^{(0)})\|) \\
&\le \frac{1}{2}\|Q(\xi^{(l)} - x(t_n))\| + C_2\Delta_Q \qquad (61)
\end{aligned}$$

with a constant C_2 that is independent of h, γ_p, Θ. If the constant C_γ in the definition of $\mathcal{U}$ is chosen such that $C_\gamma \ge 2(C_1 + C_2)$ then $\|Q(\xi^{(l+1)} - x(t_n))\| \le C_\gamma \Delta_Q$ follows by induction from (61) and $\|Q(\xi^{(1)} - x(t_n))\| = \|Q(\xi^{(1)} - \xi^{(0)})\| \le C_1\Delta_Q$. Thus $x_n := \lim_{l\to\infty} \xi^{(l)}$ exists, $F(x_n) = 0$ and $x_n \in \mathcal{U}$, i. e. (36) is satisfied. Since Φ is contractive x_n is the only fixed point of Φ in $\mathcal{U}$, i. e. $F(\xi) = 0$ is uniquely solvable in $\mathcal{U}$. ■

Remark 4 The special choice of J simplifies notation in the proof of Theorem 2. Convergence of simplified Newton method with the more natural choice $J := F_x(\xi^{(0)}) = \frac{\alpha_0}{h}A(t_n) + B(t_n)$ can be proved in the same way using $F_x(\xi^{(0)}) = (G_2 + \frac{h}{\alpha_0}BPP_1)(P_1 + \frac{h}{\alpha_0}Q_1)(\frac{\alpha_0}{h}P + Q)$ and

$$(F_x(\xi^{(0)}))^{-1} = (\frac{h}{\alpha_0}P + Q)(P_1 + \frac{\alpha_0}{h}Q_1)(I - \frac{h}{\alpha_0}PP_1G_2^{-1}BPP_1)G_2^{-1} + \mathcal{O}(h^2) \ ,$$

(in P_1, ... the argument t_n has been omitted for simplicity). This can be generalized to show that simplified Newton method with matrix $J := \frac{\alpha_0}{h}A(t_n) + g_x(t_n, \xi^{(0)})$ converges for arbitrary initial guess $\xi^{(0)}$ with $\xi^{(0)} = x(t_n) + \mathcal{O}(h^2)$, i. e. system (33) can be solved without knowing the analytical solution $x(t)$.

Example 2 The condition "$\gamma_P \le \gamma_0\sqrt{h_n}$" is more restrictive than the usual condition "$\gamma_P \le \gamma_0$" for corresponding results for one step methods. To prove that it is *necessary* for $k \ge 2$ consider the DAE

$$y_1'(t) = \frac{1}{2}\psi(z) \ , \ y_2'(t) = \frac{1}{4}\psi(z) \ , \ y_1 + y_2^2 = 1$$

with a function ψ that satisfies $\psi(0) = 0$, $\psi'(0) = 1$ and $\psi'(z) > 0$, $\psi(z) \in [-\epsilon, \epsilon]$ with a constant $\epsilon > 0$. The initial value problem $y_2(0) = 1$ has the constant solution $y_1(t) \equiv 0$, $y_2(t) \equiv 1$, $z(t) \equiv 0$. BDF with exact initial values integrate this initial value problem exactly, but with initial values $y_{n-i,2} = 1$, $(i = 1, \ldots, k; i \ne \iota)$, $y_{n-\iota,2} = 1 + \gamma$, $(|\gamma| \ll 1)$, $g(y_{n-i}) = 0$, $(i = 1, \ldots, k)$ the approximation $y_{n,2} \approx y_2(t_n)$ is defined by $y_{n,2}^2 + 2y_{n,2} - 3 + \frac{\alpha_\iota}{\alpha_0}(4\gamma + \gamma^2) = 0$, i. e. $y_{n,2} = 1 - \frac{\alpha_\iota}{\alpha_0}\gamma - \frac{\alpha_\iota}{4\alpha_0}(1 + \frac{\alpha_\iota}{\alpha_0})\gamma^2 + \mathcal{O}(\gamma^3)$. Here the index $\iota \in \{1, \ldots, k\}$ is defined such that $\alpha_\iota(\alpha_o + \alpha_\iota) \ne 0$ (this is always possible for $k \ge 2$). On the one hand we get $\psi(z_n) \in [-\epsilon, \epsilon]$, on the other hand we get $\psi(z_n) = -\frac{\alpha_\iota}{h}(1 + \frac{\alpha_\iota}{\alpha_0})\gamma^2 + \mathcal{O}(\gamma^3/h)$. Thus $\gamma^2/h = o(1)$ is seen to be necessary to have a solution (y_n, z_n) of the BDF equations.

Summary

The quasilinear DAE (3) has perturbation index 2 along the solution $x(t)$. The perturbation analysis gives error bounds of size $\max_\tau \|\delta(\tau)\| + \max_\tau \|\theta'(\tau)\|$ for the analytical solution and of size $\max_m \|\delta_m\| + \max_m \frac{1}{h_m}\|\theta_m\|$ for the BDF discretization. Here $\delta(t)$ and δ_m are the residuals in (15) and (37), respectively, and $\theta(\tau)$, θ_m denote the projections $[Q_1G_2^{-1}\delta](\tau)$, $[Q_1G_2^{-1}](t_m)\delta_m$, i. e. the residuals in the algebraic constraint (10) (for systems (1) in Hessenberg form: the residuals in $\gamma(y) = 0$). The error bounds for the numerical solution are the discrete analogues of the bounds for the analytical solution. In general *all* solution components are affected by the critical error term $\theta'(\tau)$, $\frac{1}{h_m}\theta_m$, but nevertheless the error bounds for the differential components are substantially smaller than the ones for the algebraic components. For a special class of DAEs (that includes linear DAEs with time varying coefficients) the differential components depend even continuously on the perturbations.

Compared with related results in the literature we use less restrictive assumptions on the size of perturbations. The ideas of the perturbation analysis for BDF applied to systems in Hessenberg form ([11, Section VI.6]) are carried over to quasilinear DAEs (3). We improve the error bound [17, Theorem 3.1] for the differential components by the factor h. Convergence of simplified Newton method for the BDF equations (33) is proved.

For practical computations these results imply that BDF can be applied to quasilinear DAEs (3) of index 2 directly (i. e. without analytical transformations such as index reduction). There is no order reduction. In variable stepsize variable order codes the perturbations (especially that in the algebraic constraints) should be kept small using (at least) double precision arithmetic and (very) small stopping bounds for the iterative solution of the BDF equations (33). The parts of the code that are sensitive to small errors (e. g. stepsize and order control) should be based on the differential solution components. Severe trouble has to be expected if the stepsize becomes too small (e. g. at the beginning of the integration and in regions where the solution changes rapidly). Thus the perturbation analysis gives some further theoretical justification for the way in that standard software is applied to systems of index 2 (see [5, Section 5.4.2] for DASSL and [9, Section 10] for RADAU5).

Acknowledgement I am grateful to Prof. R. März for the motivation to extend the perturbation analysis from systems in Hessenberg form to the more general class of quasilinear DAEs and for the encouragement to enter the world of "DAEs and projectors". Furthermore I would like to thank here and Caren Tischendorf (Berlin) for very helpful discussions about the topic of the present paper.

References

[1] M. Arnold. Stability of numerical methods for differential–algebraic equations of higher index. *Applied Numerical Mathematics*, 13:5–14, 1993.

[2] M. Arnold. Perturbation analysis for differential-algebraic equations of index 2. In W.F. Ames, editor, *Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics*, pages 20–23, Atlanta, 1994. Georgia Institute of Technology.

[3] M. Arnold and H. Netter. Ein modifizierter Korrektor für die stabilisierte Integration differential-algebraischer Systeme mit von Hessenbergform abweichender Struktur. Technical Report IB 515–93–03, DLR, D-5000 Köln 90, 1993.

[4] M. Arnold, K. Strehmel, and R. Weiner. Small perturbations in differential-algebraic systems of index 2. Technical Report 93/1, University of Rostock, Department of Mathematics, 1993.

[5] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial–value problems in differential–algebraic equations.* North-Holland Publ. Co., Amsterdam, 1989.

[6] L.O. Chua and A. Deng. Impasse points. Part I: Numerical aspects. *Intern. J. of Circuit Theory and Applications*, 17:213–235, 1989.

[7] C.W. Gear. The simultaneous numerical solution of differential-algebraic equations. *IEEE Trans. Circuit Theory*, CT-18:89–95, 1971.

[8] C.W. Gear, H.H. Hsu, and L.R. Petzold. Differential-algebraic equations revisited. Proc. Oberwolfach Conference on Numerical Methods for Solving Stiff Initial Value Problems, RWTH Aachen, Bericht des Instituts für Geometrie und Praktische Mathematik, No. 9, 1981.

[9] E. Hairer, Ch. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge–Kutta methods.* Lecture Notes in Mathematics, 1409. Springer-Verlag, Berlin Heidelberg New York, 1989.

[10] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations.I. Nonstiff Problems.* Springer-Verlag, Berlin Heidelberg New York, 2nd edition, 1993.

[11] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations.II. Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin Heidelberg New York, 1991.

[12] R. März. Numerical methods for differential-algebraic equations. *Acta Numerica*, pages 141–198, 1992.

[13] R. März. On linear differential-algebraic equations and linearizations. Technical Report 94–14, Humboldt–University Berlin, Department of Mathematics, 1994.

[14] R. März and C. Tischendorf. Solving more general index-2 differential algebraic equations. *Computers Math. Applic.*, 28:77–105, 1994.

[15] L.R. Petzold and P. Lötstedt. Numerical solution of nonlinear differential equations with algebraic constraints II: practical implications. *SIAM J. Sci. Stat. Comput.*, 7:720–733, 1986.

[16] B. Simeon, C. Führer, and P. Rentrop. Differential-algebraic equations in vehicle system dynamics. *Surveys on Mathematics for Industry*, 1:1–37, 1991.

[17] C. Tischendorf. Feasibility and stability behaviour of the BDF applied to index-2 differential-algebraic equations. Technical Report 93–14, Humboldt–University Berlin, Department of Mathematics, 1993. - submitted to: *ZAMM*.