Errors in the numerical solution of nonlinear differential-algebraic systems of index 2

M. Arnold* K. Strehmel† R. Weiner†

Abstract

The solution of higher index differential-algebraic systems does not depend continuously on small perturbations of the system that arise e. g. in the numerical solution because of round-off errors and errors in the iterative solution of systems of nonlinear equations. Estimates for the influence of these errors on the solution can be used to implement numerical methods in a way that the error-growth during integration does not exceed the user-supplied tolerances. We give sharp error bounds for differential and algebraic components of a system of index 2 that are substantially smaller than the bounds from the definition of perturbation index. The results are illustrated with various examples and are applied to compare two different discretization methods.

Zusammenfassung

Die Lösung von differential-algebraischen Systemen höheren Indexes hängt nicht stetig von kleinen Störungen des Systems ab (bei der numerischen Lösung werden derartige Störungen beispielsweise durch Rundungsfehler und durch Abbruchfehler bei der iterativen Lösung nichtlinearer Gleichungssysteme hervorgerufen). Unter Verwendung von oberen Schranken für diese Fehlerterme können Diskretisierungsverfahren so implementiert werden, daß der Fehler in der numerischen Lösung im Rahmen der geforderten Genauigkeit verbleibt. Für Index-2-Systeme in Hessenbergform werden für die differentiellen und die algebraischen Komponenten scharfe Fehlerschranken bewiesen, die wesentlich kleiner als die entsprechenden Fehlerschranken aus der Definition des Störungsindex sind. Die theoretischen Ergebnisse werden durch mehrere Beispiele veranschaulicht und dienen als Grundlage für den Vergleich zweier Diskretisierungsverfahren.

Mathematics Subject Classification (1991): 65L05

Keywords: differential-algebraic systems, perturbation index, higher index DAEs

1 Introduction

It is non-trivial to solve initial value problems for higher index differential-algebraic systems numerically. Standard software often fails or there are even for small tolerances large errors in the solution (see e. g. [Arn95b, Example 1]). One of the well known reasons of this phenomenon is the fact that – in contrast to ODE theory – even the analytical solution of an initial value problem does not depend continuously on perturbations in the differential-algebraic system. The discrete analogue is an *additional error term* in the numerical solution that is caused by the amplification of small errors (e. g. round-off errors or errors in the iterative solution of nonlinear equations). This additional error term tends to infinity if the stepsize is decreased.

*University of Rostock, Department of Mathematics, Postfach, 18051 Rostock, Germany

†Martin–Luther–University Halle, Department of Mathematics and Computer Science, Postfach, 06099 Halle/S., Germany

Usually the influence of perturbations $\delta(x)$ on the analytical solution is described by the perturbation index m ([HLR89]) that bounds this influence in terms that include derivatives $\delta^{(j)}(x)$ up to order $m-1$. Often similar bounds can be obtained for the influence of perturbations δ on the numerical solution (e. g. [Mär89], [HLR89]). For systems of perturbation index m these bounds contain the term $\frac{1}{h^{m-1}}\delta$ instead of $\|\delta^{(m-1)}(x)\|$ in the bound for the analytical solution.

For systems of index 2 and 3 that are of Hessenberg form the convergence results for BDF and implicit Runge–Kutta methods (e. g. [BCP89], [HLR89], [HW91]) suggest, however, that these bounds are very pessimistic for the differential components of the solution. In this paper we consider an initial value problem for a differential-algebraic system of index 2:

$$\begin{array}{rcll} y'(x) & = & f(y(x),z(x)) & , \ x \in [x_0,x_e] \ , \ y(x_0)=y_0 \ , \ z(x_0)=z_0 \ , \\ 0 & = & g(y(x)) & , \ g(y_0)=0 \ , \ [g_y f](y_0,z_0)=0 \ . \end{array} \tag{1}$$

(For simplicity we restrict ourselves to autonomous systems but the results remain valid if f and g depend explicitly on x.)

We show that the differential components of system (1) are less affected by small perturbations than the algebraic components. The same effect is observed for the numerical solution if an appropriate discretization method is used. Roughly summarized there is an additional error term of size $\mathcal{O}(\frac{1}{h}\theta^2+\mu\frac{1}{h^2}\theta^2)$ in the differential components and of size $\mathcal{O}(\frac{1}{h}\theta)$ in the algebraic components of the numerical solution if there are (small) perturbations of size $\mathcal{O}(\theta)$ in the algebraic part of (1). Here h denotes the stepsize and μ an upper bound for $\|f_{zz}\|$ (see (3)). The constant μ reflects whether the system is nonlinear in the algebraic variables ($\mu=\mathcal{O}(1)$) or approximately linear ($\mu \ll 1$). For systems that are linear in the algebraic variables the upper bound μ vanishes. The results of this perturbation analysis can be used to implement the numerical method in such a way that errors caused by the iterative solution of nonlinear equations remain within the user-supplied tolerances. If there are several analytically equivalent formulations of a differential-algebraic problem then the perturbation analysis allows to select an appropriate formulation already *before* discretization.

In Section 2 we improve the bounds for the influence of perturbations on the differential components of the analytical solution. For backward Euler method we prove in Section 3 that in general the influence of perturbations on the numerical solution can be bounded in terms that are the discrete analogue of the bounds for the analytical solution. The main results of the perturbation analysis for analytical and numerical solution are given in Theorems 1 and 3, respectively. In contrast to the bounds from the definition of perturbation index our results reflect the influence of small perturbations in the right way for *all* solution components. We illustrate this fact with various examples (Section 4) and with some numerical tests (Section 6). A summary of the theoretical results is given in Section 5.

The investigations for systems that are linear in the algebraic components give more insight in pros and cons of the "stabilized index-1 formulation" of system (1) ([Füh88]). This formulation leads to discretization methods that make use of both the original algebraic constraint $g(y)=0$ and its time derivative $\frac{d}{dx}g(y)=0$ (see Remark 1b and Section 5). Coupling the time derivative $\frac{d}{dx}g(y)=0$ to system (1) the additional error term in the differential components is reduced from $\mathcal{O}(\frac{1}{h}\theta^2+\mu\frac{1}{h^2}\theta^2)$ to $\mathcal{O}(\frac{1}{h}\theta^2)$ but it tends nevertheless to infinity if $h\to 0$.

2 Perturbations in differential-algebraic systems of index 2

Systems of the form (1) occur in various applications, e. g. in optimal control theory and in trajectory prescribed path control problems ([BCP89]). Systems (1) that depend only linearly on the algebraic components (i. e. $f(\eta,\zeta)=f_0(\eta)+f_z(\eta)\zeta$) found special interest in the literature since the numerical solution may be easier than in the general case. Some practical problems can be transformed in this special form: systems of ODE's with invariants ([Gea86]), systems of the form $B(y)y' = f(y)$ that arise in electrical circuit analysis and in chemical reaction kinetics ([Lub89]) and the *Gear–Gupta–Leimkuhler* (GGL) *formulation* of the model equations for constrained mechanical systems with potential forces ([GGL85], [HW91, Section VI.5]).

In the perturbation analysis we suppose, that (1) has a solution $(y, z) : [x_0, x_e] \to \mathbb{R}^{n_y} \times \mathbb{R}^{n_z}$ and that functions f and g are well-defined and sufficiently differentiable in a neighbourhood

$$\mathcal{U} := \{ (\eta, \zeta) : \|\eta - y(x)\| \le \Delta_y , \|\zeta - z(x)\| \le \Delta_z \}$$

of $(y(x), z(x))^T$, $(\Delta_y, \Delta_z > 0)$. In general we will consider vectors (η, ζ) in a small neighbourhood of $(y(x), z(x))$, but if (1) is linear in the algebraic components then $f(\eta, \zeta)$, $g(\eta)$ and its derivatives are well defined whenever η is close to $y(x)$. Then there is no need to restrict the perturbation analysis to vectors (η, ζ) that are close to $(y(x), z(x))$, the constant Δ_z in the definition of $\mathcal{U}$ can be chosen arbitrary large. As a consequence $f(\eta, \zeta)$, $f_y(\eta, \zeta)$, ... are not uniformly bounded in $\mathcal{U}$. Instead we assume that the derivatives of f and g and these functions itself are bounded by a constant $M < \infty$ of moderate size if they are evaluated at $(\eta, z(x))$ with $\|\eta - y(x)\| \le \Delta_y$. Furthermore the *index-2 condition*

$$[g_y f_z](\eta, z(x)) \qquad \text{is non-singular} \ , \ \|([g_y f_z](\eta, z(x)))^{-1}\| \le M < \infty \tag{2}$$

is supposed to be satisfied for all η with $\|\eta - y(x)\| \le \Delta_y$. Finally, we assume that the derivatives of f_z w. r. t. y and z are bounded by

$$\|f_{yz}(\eta, \zeta)\| \le M \quad , \quad \|f_{zz}(\eta, \zeta)\| \le \mu \quad , \quad ((\eta, \zeta) \in \mathcal{U}) \tag{3}$$

with a constant $\mu \le M$.

Remark 1

a) A typical example for systems (1) that are nonlinear in z but differ only slightly from a system that is linear in z (i. e. $0 < \mu \ll 1$) is the GGL formulation of model equations for constrained mechanical systems with small friction forces: if e. g. the friction forces are given by $\mu_0 \|F_N\|_2^2 \cdot \psi(q, q')$ with a function ψ that depends on position and velocity coordinates (μ_0 ... coefficient of friction, $q = q(x)$ ... position coordinates) then the GGL formulation has form (1) with $f(\eta, \zeta) = f_0(\eta) + \mu_0 \zeta^T f_1(\eta) \zeta \cdot f_2(\eta) + f_3(\eta) \zeta$ since the constraint forces F_N are linear in ζ (see also [Arn95b, Remark 5d and Example 4]). We get in (3) $\mu = \mathcal{O}(1)\mu_0$ and the constant Δ_z can be chosen as $\Delta_z = \mathcal{O}(\frac{1}{\mu})$. For $\mu \to 0$ the error bounds of Theorems 1 and 3 tend to the error bounds for systems that are linear in z (i. e. $\mu = 0$).

b) The analytical solution of (1) satisfies $g(y(x)) = 0$ and thus $\frac{d}{dx} g(y(x)) = g_y(y(x)) y'(x) = [g_y f](y, z) = 0$. If this derivative of the algebraic constraint is added to system (1) then an overdetermined system is obtained, ([Füh88]). To integrate this overdetermined system with a standard code like DASSL ([BCP89]) artificial variables u are introduced in the differential part

$$\begin{aligned} y' &= f(y, z) - g_y^T(y) u \\ 0 &= g(y) \\ 0 &= [g_y f](y, z) \end{aligned} \tag{4}$$

(similar to the procedure for index-3 systems that is discussed in detail in [GGL85] and [FL91]). Because of

$$0 = \frac{d}{dx} g(y(x)) = g_y(y) \cdot y' = [g_y f](y, z) - [g_y g_y^T](y) \cdot u = -[g_y g_y^T](y) \cdot u$$

this system is analytically equivalent to (1) with $u(x) \equiv 0$. Because not only the original constraint $g(y) = 0$ but also the “index-1 constraint” $\frac{d}{dx} g(y) = 0$ are considered in the *stabilized index-1 formulation* (4) of system (1) there is some hope that typical difficulties that arise in the numerical solution of higher index systems can be avoided if discretization methods are applied to (4) and *not directly* to the analytically equivalent system (1). This strategy is discussed in detail in Section 5. Because of the index-2 condition (2) the equation $[g_y f](y, z) = 0$ can be solved w. r. t. z: $z = G(y)$. Then system (4) is seen to be an index-2 system (1) that is linear w. r. t. the new algebraic variables u ($f_0(\eta) := f(\eta, G(\eta))$, $f_z(\eta) := -g_y^T(\eta)$). Note, that in contrast to other re-formulations of (1) the stabilized index-1 formulation (4) contains the original algebraic constraint $g(y) = 0$, i. e. the numerical solution will not drift off the manifold $\{ \eta : g(\eta) = 0 \}$.

In this section we consider systems (1) of Hessenberg form and functions $(\hat{y}(x), \hat{z}(x))$ with

$$\begin{aligned} \hat{y}'(x) &= f(\hat{y}(x), \hat{z}(x)) + \delta(x) \quad , \quad x \in [x_0, x_e] \ , \\ \theta(x) &= g(\hat{y}(x)) \qquad\qquad\quad , \quad \hat{y}(x_0) = \hat{y}_0 \ , \quad \hat{z}(x_0) = \hat{z}_0 \end{aligned} \tag{5}$$

and residuals that are continuous ($\delta(x)$) and continuously differentiable ($\theta(x)$), respectively. If the expression at the right hand side of (6) is sufficiently small then standard perturbation index theory gives the estimate

$$\|\hat{y}(x) - y(x)\| + \|\hat{z}(x) - z(x)\| \le C_0(\|\hat{y}_0 - y_0\| + \max_{\xi \in [x_0, x]} \|\delta(\xi)\| + \max_{\xi \in [x_0, x]} \|\theta'(\xi)\|) \tag{6}$$

with a constant C_0 that is independent of the perturbations ([HLR89, p. 4], [HW91, p. 480f]).

Example 1 In general estimate (6) is optimal for the algebraic components z: If system (1) is *linear* with $f(\eta, \zeta) = A\eta + B\zeta + a$, $g(\eta) = C\eta - b$ we get

$$\theta'(x) = \frac{d}{dx} g(\hat{y}(x)) = C\hat{y}'(x) = CA\hat{y} + CB\hat{z} + C\delta(x)$$

and because of the index-2 condition (2)

$$\hat{z} = -(CB)^{-1}CA\hat{y} - (CB)^{-1}Ca + (CB)^{-1}(\theta'(x) - C\delta(x)) \quad , \quad z = -(CB)^{-1}CAy - (CB)^{-1}Ca \ .$$

This is inserted in the differential part and results in the system of ODE’s

$$\hat{y}' = PA\hat{y} + Pa + B(CB)^{-1}\theta'(x) + P\delta(x) \quad , \quad y' = PAy + Pa$$

with the notation $P := I - B(CB)^{-1}C$. The application of Gronwall’s Lemma to

$$\hat{y}(x) - y(x) = \hat{y}_0 - y_0 + \int_{x_0}^{x} PA(\hat{y}(\xi) - y(\xi))d\xi + B(CB)^{-1}(\theta(x) - \theta(x_0)) + \int_{x_0}^{x} P\delta(\xi)d\xi$$

gives the estimate

$$\|\hat{y}(x) - y(x)\| = \mathcal{O}(1)(\|\hat{y}_0 - y_0\| + \max_{\xi \in [x_0, x]} \|\theta(\xi)\| + \max_{\xi \in [x_0, x]} \| \int_{x_0}^{\xi} P\delta(s)\, ds\|)$$

and thus

$$\|\hat{z}(x) - z(x)\| = \mathcal{O}(1)(\|\hat{y}_0 - y_0\| + \max_{\xi \in [x_0, x]} \|\theta(\xi)\| + \max_{\xi \in [x_0, x]} \|\theta'(\xi)\| + \max_{\xi \in [x_0, x]} \|\delta(\xi)\|) \ .$$

I. e., the algebraic components do not depend continuously on small perturbations $\theta(x)$ of the algebraic part of (1). Because of estimate (6) system (1) is said to be of *perturbation index 2*.

The derivative θ' of the perturbation enters the algebraic components directly but the differential components only *via the algebraic components*. So the influence of perturbations on the differential solution components y is much smaller and depends essentially on the coupling of the algebraic components z to the system i. e. on the structure of function f (e. g. $\|\hat{y}-y\| = \mathcal{O}(\|\delta\|)+\mathcal{O}(\|\theta\|)$ in Example 1). This is of special interest since errors are propagated over the interval $[x_0, x_e]$ only in the differential solution components y. The algebraic components z are determined by the equations $[g_y f](y,z) = 0$.

In the following theorem a refined analysis gives sharp estimates for the influence of perturbations on the differential solution components. In these bounds the residuum δ in the differential part is splitted into $\delta(x) = P(x)\delta(x) + (I - P(x))\delta(x)$ with the function

$$P(x) := I - [f_z(g_y f_z)^{-1} g_y](y(x), z(x)) \tag{7}$$

that represents a projection into the tangent space of the surface $\{\eta : g(\eta) = 0\}$, ([HLR89, p. 35]). So $P(x)\delta(x)$ is the projection of residuum δ in this tangent space (see Example 1). The analysis is restricted to functions $(\hat{y}(x), \hat{z}(x))$ that remain in $\mathcal{U}$ and have a non-singular Jacobian $[g_y f_z](\hat{y}(x), \hat{z}(x))$ for all $x \in [x_0, x_e]$. This can be guaranteed if $\overline{\mu}D(x)$ is sufficiently small. Here we used the notations

$$D(x) := \max_{\xi\in[x_0,x]} \|\delta(\xi)\| + \max_{\xi\in[x_0,x]} \|\theta'(\xi)\| \tag{8}$$

and $\overline{\mu} := \mu + \frac{1}{\Delta_z}$ (set $\overline{\mu} := 0$ if $\mu = 0$, i. e. if the system is linear in the algebraic variables). Note, that often $\mu\Delta_z = \mathcal{O}(1)$ such that $\overline{\mu} = \mathcal{O}(\mu)$, (see Remark 1a).

Theorem 1
a) There is a constant C_0 so that for all functions $(\hat{y}(x), \hat{z}(x))$ with (5), $\delta \in \mathcal{C}([x_0, x_e])$, $\theta \in \mathcal{C}^1([x_0, x_e])$ and $\|\delta(x)\| \le C_\delta$, $x \in [x_0, x_e]$ the estimate

$$\|\hat{y}(x) - y(x)\| \le C_0\Big(\|\hat{y}_0 - y_0\| + \int P\delta + \max_{\xi\in[x_0,x]} \|\theta(\xi)\| + \max_{\xi\in[x_0,x]} \|\theta(\xi)\| \cdot D(x) + \mu D(x)^2\Big) \tag{9}$$

with

$$\int P\delta := \max_{\xi\in[x_0,x]} \| \int_{x_0}^{\xi} P(s)\delta(s)ds\|$$

is satisfied for all $x \in [x_0, x_e]$ provided that $\overline{\mu}D(x)$ and the expression at the right hand side of (9) are sufficiently small (the constant C_0 is independent of $\hat{y}$, $\hat{z}$, μ and the perturbations δ, θ but it may depend on M and C_δ).

b) If $f(\eta,\zeta) = f_0(\eta) + f_z \cdot \zeta$ and $f_z \equiv const$ then estimate (9) holds with $D(x) := 0$.

Proof Part a) of the theorem is proven applying Gronwall's Lemma to function

$$v(x) := \Phi(\hat{y}(x)) - \Phi(y(x)) \quad \text{with} \quad \Phi(\eta) := \eta - [f_z(g_y f_z)^{-1} g](\eta, z(x)) \ . \tag{10}$$

At the beginning of the proof we suppose for all $\xi \in [x_0, x] \subseteq [x_0, x_e]$

$$\|\hat{y}(\xi) - y(\xi)\| + \mu\|\hat{z}(\xi) - z(\xi)\| \le \gamma \quad , \quad \|\hat{z}(\xi) - z(\xi)\| \le \Delta_z \tag{11}$$

with a sufficiently small positive constant γ. At the end of part a) of the proof we will show that this additional assumption is always satisfied if $\overline{\mu}D(x)$ and the expression at the right hand side of (9) are sufficiently small.

We have

$$v(x) = v(x_0) + \int_{x_0}^{x} \frac{d}{d\xi} v(\xi) d\xi \tag{12}$$

and

$$\begin{aligned} \frac{d}{d\xi} v(\xi) &= \Phi_\eta(\hat{y}(\xi))\hat{y}'(\xi) - \frac{\partial}{\partial z}[(f_z(g_y f_z)^{-1})(\hat{y}, z)](g(\hat{y}), z'(\xi)) - \Phi_\eta(y(\xi))y'(\xi) \\ &= (I - [f_z(g_y f_z)^{-1} g_y](\hat{y}, z))\hat{y}'(\xi) - \frac{\partial}{\partial y}[(f_z(g_y f_z)^{-1})(\hat{y}, z)](g(\hat{y}), \hat{y}'(\xi)) + \\ &\quad + \mathcal{O}(\|\theta(\xi)\|) - P(\xi)y'(\xi) \end{aligned} \tag{13}$$

(note, that $g(y(\xi)) = 0$ and $g(\hat{y}(\xi)) = \theta(\xi)$). Here we use the tensor notation

$$\frac{\partial}{\partial y}[(f_z(g_y f_z)^{-1})(y, z)](\zeta, \eta) := \left(\sum_{j=1}^{n_y} \sum_{k=1}^{n_z} \frac{\partial \phi_{ik}}{\partial y_j} \zeta_k \eta_j \right)_{i=1}^{n_y} \tag{14}$$

for vectors $\zeta = (\zeta_k)_k \in \mathbb{R}^{n_z}$ and $\eta = (\eta_j)_j \in \mathbb{R}^{n_y}$. The functions ϕ_{ik} denote the elements of matrix $[f_z(g_y f_z)^{-1}](y, z)$.

We do not suppose that $\hat{y}'$ is bounded by a constant of moderate size, so this term requires a careful analysis:

$$\begin{aligned} \hat{y}'(\xi) &= f(\hat{y}, \hat{z}) + \delta(\xi) = f(\hat{y}, z) + f_z(\hat{y}, z)(\hat{z} - z) + f(\hat{y}, \hat{z}) - f(\hat{y}, z) - f_z(\hat{y}, z)(\hat{z} - z) + \delta(\xi) \\ &= f(\hat{y}, z) + f_z(\hat{y}, z)(\hat{z} - z) + \mathcal{O}(\mu)\|\hat{z} - z\|^2 + \delta(\xi) \end{aligned} \tag{15}$$

(apply $\|\Psi(1) - \Psi(0) - \Psi'(0)\| = \| \int_0^1 (\Psi'(\vartheta) - \Psi'(0))\, d\vartheta\| \le \frac{1}{2} \max_{\vartheta \in (0,1)} \|\Psi''(\vartheta)\|$ to $\Psi(\vartheta) := f(\hat{y}, z + \vartheta(\hat{z} - z))$). From (15) we get

$$\hat{y}'(\xi) = \mathcal{O}(1) + \mathcal{O}(1)\|\hat{z}(\xi) - z(\xi)\| + \mathcal{O}(\|\delta(\xi)\|) \ , \tag{16}$$

$$\begin{aligned} &(I - [f_z(g_y f_z)^{-1} g_y](\hat{y}, z))\hat{y}'(\xi) = \\ &\qquad = (I - [f_z(g_y f_z)^{-1} g_y](\hat{y}, z))(f(\hat{y}, z) + \delta(\xi)) + \mathcal{O}(\mu)\|\hat{z} - z\|^2 \\ &\qquad = P(\xi)f(y, z) + P(\xi)\delta(\xi) + \mathcal{O}(1 + \|\delta(\xi)\|)\|\hat{y} - y\| + \mathcal{O}(\mu)\|\hat{z} - z\|^2 \end{aligned} \tag{17}$$

and

$$\theta'(\xi) = \frac{d}{d\xi} g(\hat{y}(\xi)) = g_y(\hat{y})\hat{y}'(\xi) = [g_y f](\hat{y}, z) + [g_y f_z](\hat{y}, z)(\hat{z} - z) + \mathcal{O}(\mu)\|\hat{z} - z\|^2 + g_y(\hat{y})\delta(\xi) \ .$$

Because of (11) we have $\mu\|\hat{z} - z\|^2 \ll \|\hat{z} - z\|$ and these equations can be solved w. r. t. $\hat{z} - z$:

$$\hat{z}(\xi) - z(\xi) = \mathcal{O}(1)(\|[g_y f](\hat{y}, z)\| + \|\delta(\xi)\| + \|\theta'(\xi)\|) = \mathcal{O}(1)(\|\hat{y}(\xi) - y(\xi)\| + D(\xi)) \tag{18}$$

since $[g_y f](\hat{y}, z) = [g_y f](y, z) + \mathcal{O}(1)\|\hat{y} - y\|$ and $[g_y f](y, z) = g_y(y)y' = \frac{d}{d\xi} g(y(\xi)) = 0$. Thus the term $\mu\|\hat{z} - z\|^2$ in (17) is bounded by

$$\mu\|\hat{z} - z\|^2 = \mathcal{O}(\mu)\|\hat{y} - y\|^2 + \mathcal{O}(\mu D(\xi))\|\hat{y} - y\| + \mathcal{O}(\mu D(\xi)^2) = \mathcal{O}(1)\|\hat{y} - y\| + \mathcal{O}(\mu D(\xi)^2) \tag{19}$$

(note, that $\|\hat{y} - y\| \le \gamma < 1$ and $\overline{\mu} D(\xi)$ was supposed to be small). Inserting (18) and (19) in (16) and (17), respectively, we get in (13)

$$\frac{d}{d\xi} v(\xi) = P(\xi)\delta(\xi) + \mathcal{O}(1)(\|\hat{y} - y\| + \|\theta(\xi)\| + \|\theta(\xi)\| \cdot D(\xi) + \mu D(\xi)^2) \ .$$

If this expression and

$$\hat{y}(\xi) - y(\xi) = v(\xi) + \mathcal{O}(1)\|g(\hat{y}(\xi))\| = v(\xi) + \mathcal{O}(1)\|\theta(\xi)\|$$

are substituted in (12) then estimate (9) follows from Gronwall's Lemma immediately with a constant C_0 that does not depend on constant γ from (11). Together with (18) we have for all $x \in [x_0, x_e]$

$$\|\hat{y}(x) - y(x)\| + \mu\|\hat{z}(x) - z(x)\| \leq C_0^*(err_y + \mu D(x)) \;\; , \;\; \|\hat{z}(x) - z(x)\| \leq \Delta_z \cdot C_0^*(\frac{err_y}{\Delta_z} + \frac{1}{\Delta_z} D(x))$$

where err_y denotes the right hand side of (9), i. e. the error bound for components y. The additional assumption (11) is satisfied for all $\xi \in [x_0, x_e]$ if $\overline{\mu} D(x)$ and the expression err_y at the right hand side of (9) are small enough to guarantee $C_0^*(err_y + \mu D(x_e)) \leq \gamma/2$ and $C_0^*(\frac{err_y}{\Delta_z} + \frac{1}{\Delta_z} D(x_e)) \leq 1$. So the proof of part a) is completed.

b) To prove part b) of the theorem the propagation of perturbations is studied in the projection $\Phi(x, \eta) := P(x) \cdot \eta$ that satisfies in a neighbourhood of the exact solution

$$\begin{aligned} \Phi(x,\eta) - \Phi(x, y(x)) &= P(x)(\eta - y(x)) = \\ &= \eta - y(x) - [f_z(g_y f_z)^{-1}](y(x)) \cdot \left(\int_0^1 (g_y(y) - g_y(y + \vartheta(\eta - y))) d\vartheta \cdot (\eta - y(x)) + g(\eta) \right) \end{aligned}$$

since $\int_0^1 g_y(y + \vartheta(\eta - y))(\eta - y)\, d\vartheta = g(y + 1 \cdot (\eta - y)) - g(y + 0 \cdot (\eta - y)) = g(\eta)$. So we have

$$\|\Phi(x,\eta) - \Phi(x, y(x))\| = \|\eta - y(x)\| + \mathcal{O}(\|\eta - y(x)\|^2) + \mathcal{O}(\|g(\eta)\|) \; . \tag{20}$$

Because of $P(x) f_z = (I - [f_z(g_y f_z)^{-1} g_y](y(x))) f_z \equiv 0$ it holds

$$[\Phi_\eta f_z](x, \eta) \equiv 0 \tag{21}$$

so that the total derivatives of $\Phi(x, \hat{y}(x))$ and $\Phi(x, y(x))$ w. r. t. x do not depend on the algebraic components z. The statement of the theorem is now proven similarly to part a) using the identity

$$\begin{aligned} \frac{d}{d\xi}(\Phi(\xi, \hat{y}(\xi)) - \Phi(\xi, y(\xi))) &= \Phi_x(\xi, \hat{y}(\xi)) - \Phi_x(\xi, y(\xi)) + \Phi_\eta(\xi, \hat{y}(\xi))\hat{y}'(\xi) - \Phi_\eta(\xi, y(\xi))y'(\xi) \\ &= \Phi_x(\xi, \hat{y}) - \Phi_x(\xi, y) + [\Phi_\eta f_0](\hat{y}) - [\Phi_\eta f_0](y) + \Phi_\eta(\xi, \hat{y})\delta(\xi) \end{aligned}$$

and $\Phi_\eta(\xi, \hat{y})\delta(\xi) = P(\xi)\delta(\xi)$. ■

Remark 2

a) The basic idea in the proof of Theorem 1 is the separate study of error propagation not only for differential and algebraic solution components but additionally within the differential components the splitting in components *along* the surface $\{ \eta : g(\eta) = 0 \}$ (i. e. the projection $P(x)(\hat{y}(x) - y(x))$) and *orthogonal* to the surface (this component is defined by $g(\hat{y}(x)) = \theta(x)$).

b) A similar approach is used in [HLR89] to prove convergence of implicit Runge–Kutta methods and in [Arn95a] to prove error bounds for BDF applied to quasilinear DAEs of index 2. In both references $f(\hat{y}, \hat{z})$ is linearized along (y, z) . The linearization of $f(\hat{y}, \hat{z})$ along $(\hat{y}, z)$ in the proof of Theorem 1a requires different techniques and results in error bounds that reflect the error in the differential components correctly both for systems (1) with $\mu = \mathcal{O}(1)$ and with $0 \leq \mu \ll 1$.

c) Similar to ODE theory it can be proven that for given $(\hat{y}_0, \hat{z}_0)$, $\delta(x)$ and $\theta(x)$ there are uniquely defined functions $(\hat{y}(x), \hat{z}(x))$ with (5) provided that $\overline{\mu}D(x)$ and the expressions on the right hand side of (9) are sufficiently small and $g|_{\hat{y}_0} = \theta(x_0)$, $\frac{d}{dx}g|_{(\hat{y}_0,\hat{z}_0)} = \theta'(x_0)$.

Theorem 1 shows that the differential components are substantially less affected by perturbations than the algebraic ones. But it will be illustrated in Examples 2 and 4 below that in general *both* algebraic and differential components are influenced by the first derivative of perturbations in the algebraic equations. Only for systems with a very special structure the differential components y are index-1 variables in the sense of the perturbation index, i. e. these components depend continuously on perturbations of system (1). This is proven in part b) of Theorem 1 for systems with $f(\eta,\zeta) = f_0(\eta) + f_z \cdot \zeta$. Corollary 1 shows the same result for systems that are linear in the algebraic variables if the algebraic constraints g define a hyper-surface $\{\, \eta \,:\, g(\eta) = 0 \,\} \subset \mathbb{R}^{n_y}$, i. e. if $n_z = 1$. Note, that in contrast to Theorem 1 this Corollary has *no* fellow for the numerical solution (see also Example 3).

Corollary 1 *The differential components y depend continuously on small perturbations of system (1) if the differential part is linear w. r. t. the algebraic components z ($f(\eta,\zeta) = f_0(\eta) + f_z(\eta)\cdot\zeta$) and the algebraic part consists of one scalar algebraic constraint g ($n_z = 1$).*

Proof We follow the lines of the proof that was given for part b) of Theorem 1. The corollary is proven if there is a smooth function Φ so that the equations (20) and (21) are satisfied. This function Φ is defined as solution of the initial value problem

$$\frac{d}{d\tau}\tilde{\eta}(\tau;\eta) = -[f_z(g_y f_z)^{-1}](\tilde{\eta})\cdot g(\eta) \quad , \quad (\tau \in [0,1]) \quad , \quad \tilde{\eta}(0;\eta) = \eta \tag{22}$$

at $\tau_e = 1$: $\Phi(\eta) := \tilde{\eta}(1;\eta)$, (Φ does not depend on x). The index-2 condition (2) guarantees that (22) is uniquely solvable for all η in a sufficiently small neighbourhood of the exact solution. Because of $\Phi(\eta) - \eta = \mathcal{O}(\|g(\eta)\|)$ and $\Phi(\eta) = \eta$ if $g(\eta) = 0$ equation (20) is satisfied.

The trajectory $\tilde{\eta}(\tau;\eta)$ is orthogonal to all manifolds $\{\, y \,:\, g(y) = const \,\}$ and $\Phi(\eta)$ is the intersection of this trajectory with the manifold $\{\, y \,:\, g(y) = 0 \,\}$. (This function Φ was introduced in another context by J. Wensch [Wen92].)

Along the trajectory $\tilde{\eta}(\tau;\eta)$ function Φ is constant since

$$g(\tilde{\eta}(\tau_0;\eta)) = g(\tilde{\eta}(0;\eta)) + \int_0^{\tau_0} g_y(\tilde{\eta})\frac{d\tilde{\eta}}{d\tau}\,d\tau = g(\eta) - \tau_0 g(\eta) = (1-\tau_0)g(\eta)$$

and

$$\tilde{\eta}(\frac{\tau-\tau_0}{1-\tau_0};\tilde{\eta}(\tau_0;\eta)) = \tilde{\eta}(\tau;\eta) \quad ,$$

i. e.

$$\Phi(\tilde{\eta}(\tau_0;\eta)) = \tilde{\eta}(1;\tilde{\eta}(\tau_0;\eta)) = \tilde{\eta}(1;\eta) = \Phi(\eta) \quad . \tag{23}$$

To complete the proof we finally show $[\Phi_\eta f_z](\eta) \equiv 0$, i. e. (21) is satisfied. The directional derivative $\Phi_\eta f_z$ of function Φ vanishes identically because the trajectory $\tilde{\eta}(\tau;\eta + hf_z(\eta))$ starting from $\eta + hf_z(\eta)$ crosses a $\mathcal{O}(h^2)$–neighbourhood of η if $|h|$ is sufficiently small (see below). Then we have $\Phi(\eta + hf_z(\eta)) - \Phi(\eta) = \mathcal{O}(h^2)$ and

$$[\Phi_\eta f_z](\eta) = \lim_{h\to 0}\frac{\Phi(\eta + hf_z(\eta)) - \Phi(\eta)}{h} = 0$$

because Φ is Lipschitzian and (23) is satisfied along $\tilde{\eta}(\tau;\eta + hf_z(\eta))$.

It holds $\tilde{\eta}(\tau_0;\eta+hf_z(\eta))-\eta=\mathcal{O}(h^2)$ for $\tau_0:=h\dfrac{[g_yf_z](\eta)}{g(\eta+hf_z(\eta))}$:

$$\begin{aligned}\tilde{\eta}(\tau_0;\eta+hf_z(\eta)) &= \tilde{\eta}(0;\eta+hf_z(\eta))+\int_0^{\tau_0}\frac{d\tilde{\eta}}{d\tau}\,d\tau \\ &= \eta+hf_z(\eta)-\int_0^{\tau_0}[f_z(g_yf_z)^{-1}](\eta)\cdot g(\eta+hf_z(\eta))\,d\tau+\mathcal{O}(h^2) \\ &= \eta+hf_z(\eta)-\tau_0\frac{g(\eta+hf_z(\eta))}{[g_yf_z](\eta)}f_z(\eta)+\mathcal{O}(h^2)=\eta+\mathcal{O}(h^2)\ .\end{aligned}$$

τ_0 is well-defined if $|h|$ is sufficiently small since $g(\eta+hf_z(\eta))\neq 0$ if $g(\eta)\neq 0$ (if $g(\eta)=0$ then $g(\eta+hf_z(\eta))=g(\eta)+h[g_yf_z](\eta)+\mathcal{O}(h^2)\neq 0$ if $|h|\leq h_0$ and $h\neq 0$).
Note, that the assumption $n_z=1$ (i. e. g is scalar) is implicitly used in the definition of τ_0. ■

3 Perturbations in the numerical solution of differential-algebraic systems

Discretization and the implementation of discretization methods on a computer introduce errors in the solution of initial value problems that are propagated and amplified during integration. The solution does not depend continuously on these errors (see Section 2). The idea of perturbation index theory is that the amplification of these perturbations in the numerical solution can be bounded similarly to Theorem 1 if appropriate discretization methods are used. In this section we prove these error bounds for backward Euler method. This gives also insight in the structure of error estimates for higher order methods (implicit Runge–Kutta methods, BDF).

For system (1) backward Euler method is defined by ([Gea71])

$$\begin{aligned}\frac{y_{n+1}-y_n}{h} &= f(y_{n+1},z_{n+1}) \\ 0 &= g(y_{n+1})\end{aligned}\tag{24}$$

and requires in each step of integration the solution of a system of nonlinear equations. Implicit Runge–Kutta methods for (1) are studied in great detail in [HLR89]. Applied to backward Euler these results prove that the method converges with order 1. [HLR89, Theorem 4.1] shows that the system (24) of nonlinear equations is locally uniquely solvable w. r. t. (y_{n+1},z_{n+1}) if $\|g(y_n)\|+h\|[g_yf](y_n,z_n)\|=\mathcal{O}(h^2)$. Within the framework of the present paper a similar result can be proven with less restrictive assumptions on (y_n,z_n) :

Theorem 2 *Let* (y_n,z_n) *be given with*

$$\|y_n-y(x_n)\|+\mu\|z_n-z(x_n)\|\leq\gamma\quad,\quad\|z_n-z(x_n)\|\leq\frac{1}{2}\Delta_z\quad,\quad\Delta+\overline{\mu}\frac{\Delta}{h}\leq\Delta_0$$

and $\Delta:=\|g(y_n)\|+h\|[g_yf](y_n,z_n)\|$ *. Then the nonlinear system of equations (24) is for stepsizes* $h\leq h_0$ *(locally) uniquely solvable w. r. t.* (y_{n+1},z_{n+1}) *if the positive constants* h_0, γ *and* Δ_0 *are sufficiently small. It holds*

$$\|y_{n+1}-y_n\|\leq C_{h\Delta}(h+\Delta)\quad,\quad\|z_{n+1}-z_n\|\leq C_{h\Delta}(h+\frac{\Delta}{h})\tag{25}$$

with a constant $C_{h\Delta}$ *that is independent of* h, Δ, μ *and* (y_n,z_n).

If backward Euler method is implemented on a computer then the actually computed solution $(\hat{y}_n,\hat{z}_n)$ does in general not satisfy equations (24) exactly but with (small) residuals δ_{n+1}, θ_{n+1}

that are caused e. g. by round-off errors and by stopping the iterative solution of the system of nonlinear equations:

$$\begin{aligned} \frac{\hat{y}_{n+1} - \hat{y}_n}{h} &= f(\hat{y}_{n+1}, \hat{z}_{n+1}) + \delta_{n+1} \\ \theta_{n+1} &= g(\hat{y}_{n+1}) \end{aligned} \tag{26}$$

Remark 3 Estimates for $\|\hat{y}_n - y_n\|$ can be obtained only if the sequences (y_n, z_n) and $(\hat{y}_n, \hat{z}_n)$ remain in the neighbourhood of one and the same analytical solution of (1). This can be guaranteed if the increment in one step of integration is bounded by

$$\begin{aligned} \|\hat{y}_{n+1} - \hat{y}_n\| &\le C^*_{h\Delta}(h + \|g(\hat{y}_n)\| + h\|[g_y f](\hat{y}_n, \hat{z}_n)\| + h\delta + \theta) \ , \\ \|\hat{z}_{n+1} - \hat{z}_n\| &\le C^*_{h\Delta}(h + \tfrac{1}{h}\|g(\hat{y}_n)\| + \|[g_y f](\hat{y}_n, \hat{z}_n)\| + \delta + \tfrac{1}{h}\theta) \ , \end{aligned} \tag{27}$$

$$\delta := \max_n \|\delta_n\| + \|[g_y f](\hat{y}_0, \hat{z}_0)\| \quad , \quad \theta := \max_n \|\theta_n\| + \|g(\hat{y}_0)\|$$

with a constant $C^*_{h\Delta}$ that is independent of h, δ, θ and μ. Extending the ideas of Theorem 2 it can be proven that there is at least one $(\hat{y}_{n+1}, \hat{z}_{n+1})$ with (26) and (27) if $h \le h_0$, $\|\hat{y}_n - y(x_n)\| + \mu\|\hat{z}_n - z(x_n)\| \le \gamma$, $\|\hat{z}_n - z(x_n)\| \le \frac{1}{2}\Delta_z$ and $\Delta^* + \overline{\mu}\frac{\Delta^*}{h} \le \Delta_0$ and the constants γ, h_0 and Δ_0 are sufficiently small, $(\Delta^* := \|g(\hat{y}_n)\| + h\|[g_y f](\hat{y}_n, \hat{z}_n)\| + \|\theta_{n+1}\| + h\|\delta_{n+1}\|)$. In the following we suppose that the sequence $(\hat{y}_n, \hat{z}_n)$ satisfies (27) – this is the discrete analogue to the assumption $\hat{y} \in \mathcal{C}^1([x_0, x_e])$, $\hat{z} \in \mathcal{C}([x_0, x_e])$ in Section 2.

Inequality (27) gives some estimate for the influence of perturbations in one single step of integration. In the following theorem the *propagation* of the perturbations during integration is studied:

Theorem 3
a) There is a constant C_0 so that for all sequences $(\hat{y}_n, \hat{z}_n)$ with (26), (27) and $\delta \le C_\delta$ the estimate

$$\|\hat{y}_n - y_n\| \le C_0(\|\hat{y}_0 - y_0\| + \int_h P\delta + \theta + \theta D_h + \mu D_h^2) \tag{28}$$

is satisfied for all $n \ge 0$, $x_0 + nh \le x_e$ with

$$\int_h P\delta := h \sum_{m=0}^{n-1} \|P_m \delta_{m+1}\| + h\delta \quad , \quad D_h := \delta + \frac{1}{h}\theta$$

provided that the stepsize h, the expression at the right hand side of (28) and $\overline{\mu} D_h$ are sufficiently small. P_m denotes the projector $P(x_m)$ that was defined in (7). The constant C_0 may again depend on M and C_δ and is independent of h, δ, θ and μ.

b) If $f(\eta, \zeta) = f_0(\eta) + f_z \cdot \zeta$ and $f_z \equiv const$ then estimate (28) holds with $D_h := 0$.

Furthermore in both cases the estimate $\|\hat{z}_n - z_n\| \le C_0(\|\hat{y}_0 - y_0\| + \delta + \frac{1}{h}\theta)$ for the algebraic components is satisfied.

Proof As for the analytical solution the error propagation is not studied in $\hat{y}_m - y_m$ directly but in some projection

$$v_m := \Phi_h(\hat{y}_m) - \Phi_h(y_m) \quad \text{with} \quad \Phi_h(\eta) := P(y_m, z_m)(\eta - [f_z(g_y f_z)^{-1} g](\eta, z_m)) \ . \tag{29}$$

Here the projector $P(y_m, z_m) := I - [f_z(g_y f_z)^{-1} g_y](y_m, z_m)$ is evaluated along the numerical solution (y_m, z_m) .

a) Similar to condition (11) in the proof of Theorem 1 we restrict ourselves to sequences (y_m, z_m) and $(\hat{y}_m, \hat{z}_m)$ that remain for all $m \le n$ in a sufficiently small neighbourhood of the exact solution:

$$\begin{gathered} \|y_m - y(x_m)\| + \|z_m - z(x_m)\| \le \gamma \ , \ \ \|[g_y f](y_m, z_m)\| \le C_h \cdot h \ , \\ \|\hat{y}_m - y(x_m)\| + \mu\|\hat{z}_m - z(x_m)\| \le \gamma, \ \|[g_y f](\hat{y}_m, \hat{z}_m)\| \le C_h(\gamma + D_h), \ \|\hat{z}_m - z(x_m)\| \le \tfrac{1}{2}\Delta_z \end{gathered} \tag{30}$$

This additional assumption will be justified at the end of the proof of part a). In (30) the constants γ and C_h are independent of h, D_h and μ, constant γ is supposed to be sufficiently small, C_h is of size $\mathcal{O}(1)$ and is fixed below.

With these assumptions equations (24) are locally uniquelly solvable, it holds $\|y_{m+1} - y_m\| = \mathcal{O}(h + C_h h^2)$, $\|z_{m+1} - z_m\| = \mathcal{O}(h + C_h h)$, (see Theorem 2). Here the constants in the $\mathcal{O}(.)$-terms are independent of C_h. In the same way we get from condition (27)

$$\begin{aligned} \|\hat{y}_{m+1} - \hat{y}_m\| &= \mathcal{O}(h) + \mathcal{O}(hD_h) + \mathcal{O}(h)\|\|[g_y f](\hat{y}_m, \hat{z}_m)\| \\ &= \mathcal{O}(h + C_h h\gamma + (1 + C_h)hD_h) = \mathcal{O}((1 + C_h)\sqrt{h}) \\ \|\hat{z}_{m+1} - \hat{z}_m\| &= \mathcal{O}(h) + \mathcal{O}(C_h\gamma) + \mathcal{O}((1 + C_h)D_h) \\ \theta\|\hat{y}_{m+1} - \hat{y}_m\| &= \mathcal{O}(h)(1 + C_h)(\theta + \theta D_h) \end{aligned} \tag{31}$$

since $hD_h = h\delta + \theta \le C_\delta h + \sqrt{h} \cdot \sqrt{\theta \cdot D_h} = \mathcal{O}(\sqrt{h})$.

Studying the error propagation in the analytical solution we used the relation $g_y(y)y' = 0$. The discrete counterpart is given by

$$g_y(y_{m+1})(y_{m+1} - y_m) = \int_0^1 (g_y(y_{m+1}) - g_y(y_m + \vartheta(y_{m+1} - y_m)))\, d\vartheta \cdot (y_{m+1} - y_m) + g(y_{m+1}) - g(y_m) \tag{32}$$

(apply $\Psi(1) - \Psi(0) = \int_0^1 \Psi'(\vartheta)\, d\vartheta$ to $\Psi(\vartheta) := g(y_m + \vartheta(y_{m+1} - y_m))$). With (32) we get

$$[g_y f](y_{m+1}, z_{m+1}) = \frac{1}{h} g_y(y_{m+1})(y_{m+1} - y_m) = \mathcal{O}(\frac{1}{h})\|y_{m+1} - y_m\|^2 = \mathcal{O}(h + C_h h^2) \le C_h h \tag{33}$$

and

$$\begin{aligned} [g_y f](\hat{y}_{m+1}, \hat{z}_{m+1}) &= \mathcal{O}(\frac{1}{h})\|\hat{y}_{m+1} - \hat{y}_m\|^2 + \mathcal{O}(D_h) \\ &= \mathcal{O}(\,(1 + C_h)^2(1 + C_\delta + \frac{1}{h}\theta)(h + hD_h)\,) + \mathcal{O}(D_h) \le C_h(\gamma + D_h) \end{aligned} \tag{34}$$

if the constant C_h in (30) is chosen sufficiently large and $h \le h_0$, $\theta + \theta D_h \le \Delta_0$ with (small) positive constants h_0, Δ_0 (that may depend on γ and C_h).
Substituting $y_m \to \hat{y}_m$, $y_{m+1} \to y_m$ we obtain from (32) and $g(\hat{y}_m) = \theta_m = \mathcal{O}(\theta)$

$$\begin{aligned} \|(I - P(y_m, z_m))(\hat{y}_m - y_m)\| &\le \mathcal{O}(1)\|\hat{y}_m - y_m\|^2 + \mathcal{O}(\theta) \le \frac{1}{2}\|\hat{y}_m - y_m\| + \mathcal{O}(\theta) \\ &\le \frac{1}{2}\|P(y_m, z_m)(\hat{y}_m - y_m)\| + \frac{1}{2}\|(I - P(y_m, z_m))(\hat{y}_m - y_m)\| + \mathcal{O}(\theta) \end{aligned}$$

and thus $\|(I - P(y_m, z_m))(\hat{y}_m - y_m)\| \le \|v_m\| + \mathcal{O}(\theta)$ and

$$\|\hat{y}_m - y_m\| \le 2\|v_m\| + \mathcal{O}(\theta) \tag{35}$$

(see the definition (29) of v_m). I. e. the error in the differential components is bounded in terms of θ and $\|v_m\|$.
The equivalent of (32) for the perturbed sequence is

$$g_y(\hat{y}_{m+1})(\hat{y}_{m+1} - \hat{y}_m) = \mathcal{O}(1)\|\hat{y}_{m+1} - \hat{y}_m\|^2 + g(\hat{y}_{m+1}) - g(\hat{y}_m) = g(\hat{y}_{m+1}) - g(\hat{y}_m) + \mathcal{O}(h) \ , \tag{36}$$

(see (31)). Finally we consider the difference of (32) and the corresponding equation with $y_m \to \hat{y}_m$, $y_{m+1} \to \hat{y}_{m+1}$ and get

$$\begin{aligned} g_y(\hat{y}_{m+1})(\hat{y}_{m+1} - y_{m+1}) - (g(\hat{y}_{m+1}) - g(\hat{y}_m)) - g_y(y_{m+1})(y_{m+1} - y_m) = & \\ = \quad & \mathcal{O}(1)(\|\hat{y}_{m+1} - \hat{y}_m\| + \|y_{m+1} - y_m\|)(\|\hat{y}_{m+1} - y_{m+1}\| + \|\hat{y}_m - y_m\|) \\ = \quad & \mathcal{O}(h + hD_h)(\|\hat{y}_{m+1} - y_{m+1}\| + \|\hat{y}_m - y_m\|) \ . \end{aligned} \tag{37}$$

This estimate is used to get an error bound for the algebraic components: as in (15) we have

$$f(\hat{y}_{m+1}, \hat{z}_{m+1}) = f(\hat{y}_{m+1}, z_{m+1}) + f_z(\hat{y}_{m+1}, z_{m+1})(\hat{z}_{m+1} - z_{m+1}) + \mathcal{O}(\mu)\|\hat{z}_{m+1} - z_{m+1}\|^2 \tag{38}$$

and thus

$$[g_y f](\hat{y}_{m+1}, \hat{z}_{m+1}) = [g_y f](\hat{y}_{m+1}, z_{m+1}) + [g_y f_z](\hat{y}_{m+1}, z_{m+1})(\hat{z}_{m+1} - z_{m+1}) + \mathcal{O}(\mu)\|\hat{z}_{m+1} - z_{m+1}\|^2 \ .$$

These equations can be solved w. r. t. $\hat{z}_{m+1} - z_{m+1}$ since $[g_y f_z](\hat{y}_{m+1}, z_{m+1})$ is non-singular ($\|z_{m+1} - z(x_{m+1})\| = \mathcal{O}(\gamma) + \mathcal{O}(h)$) and $\mu\|\hat{z}_{m+1} - z_{m+1}\| = \mathcal{O}(\gamma) + \mathcal{O}(h) + \mathcal{O}(\mu D_h) \ll 1$:

$$\|\hat{z}_{m+1} - z_{m+1}\| = \mathcal{O}(1)\|[g_y f](\hat{y}_{m+1}, \hat{z}_{m+1}) - [g_y f](\hat{y}_{m+1}, z_{m+1})\| \ .$$

Because of $f(\hat{y}_{m+1}, \hat{z}_{m+1}) = \frac{1}{h}(\hat{y}_{m+1} - \hat{y}_m) - \delta_{m+1}$, $f(y_{m+1}, z_{m+1}) = \frac{1}{h}(y_{m+1} - y_m)$ and $[g_y f](\hat{y}_{m+1}, z_{m+1}) = [g_y f](y_{m+1}, z_{m+1}) + \mathcal{O}(1)\|\hat{y}_{m+1} - y_{m+1}\|$ this estimate and (37) yield to

$$\|\hat{z}_{m+1} - z_{m+1}\| = \mathcal{O}(1 + D_h)(\|\hat{y}_{m+1} - y_{m+1}\| + \|\hat{y}_m - y_m\|) + \mathcal{O}(D_h) \tag{39}$$

and

$$\mu\|\hat{z}_{m+1} - z_{m+1}\|^2 = \mathcal{O}(1)(\|\hat{y}_{m+1} - y_{m+1}\| + \|\hat{y}_m - y_m\|) + \mathcal{O}(\mu D_h^2) \ , \tag{40}$$

(note, that μD_h and μD_h^2 are supposed to be small and $\|\hat{y}_m - y_m\| = \mathcal{O}(\gamma)$, $\|\hat{y}_{m+1} - y_{m+1}\| = \mathcal{O}(\gamma) + \mathcal{O}(\sqrt{h})$).
With estimates (32), (35)–(40) a bound for $v_{m+1} - v_m$ is obtained:

$$\begin{aligned} & v_{m+1} - v_m = \\ & \quad = P(y_{m+1}, z_{m+1})(\hat{y}_{m+1} - \hat{y}_m - [f_z(g_y f_z)^{-1}](\hat{y}_{m+1}, z_{m+1})(g(\hat{y}_{m+1}) - g(\hat{y}_m)) - (y_{m+1} - y_m)) + \\ & \qquad + \mathcal{O}(h)\|\hat{y}_m - y_m\| + \mathcal{O}(h\theta) + \\ & \qquad + P(y_{m+1}, z_{m+1})([f_z(g_y f_z)^{-1}](\hat{y}_{m+1}, z_{m+1}) - [f_z(g_y f_z)^{-1}](\hat{y}_m, z_m))g(\hat{y}_m) \\ & \quad = P(y_{m+1}, z_{m+1})\Big((I - [f_z(g_y f_z)^{-1} g_y](\hat{y}_{m+1}, z_{m+1}))(\hat{y}_{m+1} - \hat{y}_m) - P(y_{m+1}, z_{m+1})(y_{m+1} - y_m)\Big) + \\ & \qquad + P(y_{m+1}, z_{m+1})[f_z(g_y f_z)^{-1}](\hat{y}_{m+1}, z_{m+1}) \cdot (g_y(\hat{y}_{m+1})(\hat{y}_{m+1} - \hat{y}_m) - (g(\hat{y}_{m+1}) - g(\hat{y}_m))) + \\ & \qquad + \mathcal{O}(h)\|\hat{y}_m - y_m\| + (\mathcal{O}(1)\|\hat{y}_{m+1} - \hat{y}_m\| + \mathcal{O}(h))\theta \\ & \quad = h(\Psi(\hat{y}_{m+1}) - \Psi(y_{m+1})) + hP(y_{m+1}, z_{m+1})\delta_{m+1} + (\mathcal{O}(h\|\delta_{m+1}\|) + \mathcal{O}(h))\|\hat{y}_{m+1} - y_{m+1}\| + \\ & \qquad + \mathcal{O}(h)\|\hat{y}_m - y_m\| + \mathcal{O}(h\mu)\|\hat{z}_{m+1} - z_{m+1}\|^2 + \mathcal{O}(h\theta) + \mathcal{O}(h\theta D_h) \end{aligned}$$

since $P(y_{m+1}, z_{m+1})f_z(\hat{y}_{m+1}, z_{m+1}) = P(y_{m+1}, z_{m+1})(f_z(\hat{y}_{m+1}, z_{m+1}) - f_z(y_{m+1}, z_{m+1}))$. Here $\Psi(\eta)$ stands for $P(y_{m+1}, z_{m+1})(I - [f_z(g_y f_z)^{-1} g_y](\eta, z_{m+1}))f(\eta, z_{m+1})$. Now we insert into this expression the bounds for $\|\hat{y}_{m+1} - y_{m+1}\|$, $\|\hat{z}_{m+1} - z_{m+1}\|$ and get

$$\begin{aligned} \|v_{m+1}\| \quad &= \quad \|v_m\| + \mathcal{O}(h)\|\hat{y}_m - y_m\| + \mathcal{O}(h)\|\hat{y}_{m+1} - y_{m+1}\| + \mathcal{O}(h\mu)\|\hat{z}_{m+1} - z_{m+1}\|^2 + \\ & \qquad + h\|P_m \delta_{m+1}\| + \mathcal{O}(h^2\delta) + \mathcal{O}(h\theta) + \mathcal{O}(h\theta \cdot D_h) + \mathcal{O}(h)d_h\|\delta_{m+1}\| \\ \|v_{m+1}\| \quad &= \quad (1 + \mathcal{O}(h))\|v_m\| + h\|P_m \delta_{m+1}\| + \mathcal{O}(h)(d_h\|\delta_{m+1}\| + h\delta + \theta + \theta D_h + \mu D_h^2) \end{aligned} \tag{41}$$

with $d_h := \|y_{m+1} - y(x_{m+1})\| + \|z_{m+1} - z(x_{m+1})\|$.
This estimate is used to prove convergence of backward Euler method: consider for the moment the special sequence $(\hat{y}_n, \hat{z}_n) := (\, y(x_n), z(x_n)\,)$ then δ represents the local discretization error, $\delta = \mathcal{O}(h)$, and $\theta = \max_n \|g(y(x_n))\| = 0$, i. e. $D_h = \mathcal{O}(h)$. Estimate (41) simplifies to $\|v_{m+1}\| = (1 + \mathcal{O}(h))\|v_m\| + \mathcal{O}(h\delta)$ and we get $\|v_n\| = \mathcal{O}(1)\|v_0\| + \mathcal{O}(\delta) = \mathcal{O}(h)$ by standard techniques. With (35) and (39) this proves $y_n = y(x_n) + \mathcal{O}(h)$, $z_n = z(x_n) + \mathcal{O}(h)$, i. e. convergence with order 1. Thus the numerical solution (y_n, z_n) remains for all $h \le h_0$ in a γ–neighbourhood of the analytical solution if h_0 is sufficiently small, i. e. the first and second condition in (30) are always satisfied ($\|[g_y f](y_n, z_n)\| \le C_h h$ follows from (33)).

We now return to arbitrary sequences $(\hat{y}_n, \hat{z}_n)$ with (26) and (27) and get $d_h = \mathcal{O}(h)$ in (41). Recursive application of this estimate yields to

$$\|v_n\| = \mathcal{O}(1)(\|v_0\| + \int_h P\delta + \theta + \theta \cdot D_h + \mu D_h^2)$$

and thus also to (28) and the corresponding error bound for the algebraic components. Again these estimates and (34) can be used to prove by induction that the additional assumption (30) is satisfied for all m with $x_0 + mh \le x_e$ whenever $\overline{\mu} D_h$ and the expressions at the right hand side of (28) are sufficiently small: $\hat{y}_m - y(x_m) = \hat{y}_m - y_m + \mathcal{O}(h)$, $\hat{z}_m - z(x_m) = \hat{z}_m - z_m + \mathcal{O}(h)$. This completes the proof of part a) of the theorem.

b) If $f_z = const$ we have $\mu = 0$ and $P(y_{m+1}, z_{m+1}) f_z = 0$, i. e. $v_m = P(y_m, z_m)(\hat{y}_m - y_m)$ and v_m is not affected by $\hat{z}_m$. Thus the error bound for the differential components depends continuously on the perturbations. ■

Remark 4

a) For simplicity of notation we assumed the stepsize h to be fixed, but all results of Theorem 3 remain valid for variable stepsizes. The equivalence of the error estimates in Theorems 1 and 3 suggests that similar error bounds can be proven also for more complex discretization methods. Note, however, that the analysis for backward Euler method is simpler than for higher order methods (implicit Runge–Kutta methods, BDF) since (24) is a one-step one-stage method and the numerical solution satisfies $g(y_n) = 0$.

b) In [HLR89] the influence of perturbations on the Runge–Kutta solution is studied as part of the convergence analysis. They consider the pure nonlinear case $\mu = \mathcal{O}(1)$ and perturbations of size $\delta = \mathcal{O}(h)$, $\theta = \mathcal{O}(h^2)$, i. e. $D_h = \mathcal{O}(h)$ (then the right hand side of (28) simplifies to $C_0(\|\hat{y}_0 - y_0\| + \int_h P\delta + \theta)$). In contrast to these results we do not linearize the system along the analytical solution but both along (y_n, z_n) and $(\hat{y}_n, z_n)$ and get realistic error bounds both for $\mu = \mathcal{O}(1)$ and for $0 \le \mu \ll 1$. Furthermore Theorem 3 gives error bounds not only for large but also for (very) small stepsizes (e. g. $h < 10^{-4}$ in Example 5 with $\mu_0 = 1$) since the assumptions on the size of perturbations are substantially less restrictive ($\theta \cdot D_h = o(1)$, $\mu D_h^2 = o(1)$, $\overline{\mu} D_h = o(1)$).

4 Examples

The bounds that are known from the definition of the perturbation index reflect the influence of small perturbations on the algebraic components correctly (Example 1). In this section we illustrate that the bounds of Theorems 1 and 3 are optimal in this sense for *both* differential and algebraic components. Using a similar example of Lubich ([Lub89]) we start with a system (1) with *linear* algebraic constraints g (Example 2). The proof that in general the differential components do not depend continuously on small perturbations of the system is much more complicated if f has the special structure $f(\eta,\zeta) = f_0(\eta) - g_y^T(\eta)\cdot\zeta$ (cf. Remark 1b): In Example 3 the amplification of perturbations in the numerical solution is studied for a system with one nonlinear algebraic constraint. A system with two nonlinear algebraic constraints illustrates that also the *analytical* solution y of such systems depends on derivatives of the perturbation (Example 4).

Example 2 Consider the initial value problem

$$\begin{array}{rclcrcl} y_1' & = & z_1 & , & 0 & = & y_1 \quad , \quad x \in [0,1] \ , \\ y_2' & = & z_2 & , & 0 & = & y_2 \quad , \\ y_3' & = & y_2 z_1 - y_1 z_2 + \mu(z_1^2+z_2^2) & , & \multicolumn{3}{l}{y(0) = (0,0,0)^T \ , \ z(0) = (0,0)^T \ ,} \end{array} \tag{42}$$

with the solution: $y(x) = (0,0,0)^T$, $z(x) = (0,0)^T$.
Let (small) positive constants Δ, Θ and ε be given. For functions

$$\hat{y}(x) = \Big(\Theta \sin\tfrac{x}{\varepsilon}\, ,\ \Theta\cos\tfrac{x}{\varepsilon}\, ,\ (\Theta D + \mu D^2)\,x\Big)^T \quad , \quad \hat{z}(x) = \Big(D\cos\tfrac{x}{\varepsilon}\, ,\ -D\sin\tfrac{x}{\varepsilon}\Big)^T$$

with $D := \Delta + \frac{\Theta}{\varepsilon}$ the residuals in (5) are given by $\delta(x) = (-\Delta\cos\frac{x}{\varepsilon}\, , \Delta\sin\frac{x}{\varepsilon}\, , 0)^T$, $\theta(x) = (\Theta\sin\frac{x}{\varepsilon}\, , \Theta\cos\frac{x}{\varepsilon})^T$. So we have $\|\hat{y}_0 - y_0\| = \mathcal{O}(\Theta)$, $\|\delta(x)\| = \mathcal{O}(\Delta)$, $\|\theta(x)\| = \mathcal{O}(\Theta)$, $\|\theta'(x)\| = \mathcal{O}(\frac{\Theta}{\varepsilon})$ and $|\hat{y}_3(x) - y_3(x)| = \mathcal{O}(\theta D + \mu D^2)$, $\|\hat{z}(x) - z(x)\| = \mathcal{O}(D)$.
This system is solved exactly by backward Euler method: $y_n = y(x_n)$, $z_n = z(x_n)$. For the sequence $(\hat{y}_n, \hat{z}_n)$ with

$$\hat{y}_n = \Big(\Theta\sin n\tfrac{\pi}{2}\ ,\ \Theta\cos n\tfrac{\pi}{2}\ ,\ nh(\Theta D_h + 2\mu D_h^2)\Big)^T \ , \ \hat{z}_n = \sqrt{2}D_h\cdot\Big(\cos(2n-1)\tfrac{\pi}{4}\ ,\ -\sin(2n-1)\tfrac{\pi}{4}\Big)^T$$

and $D_h := \Delta + \frac{1}{h}\Theta$ the residuals are of size $\delta = \mathcal{O}(\Delta)$, $\theta = \mathcal{O}(\Theta)$. It holds $|\hat{y}_{3,n} - y_{3,n}| = \mathcal{O}(\theta D_h + \mu D_h^2)$ and $\|\hat{z}_n - z_n\| = \mathcal{O}(D_h)$, i. e. instead of $\|\delta(x)\|$ and $\|\theta'(x)\|$ in the difference of functions $y_3(x)$, $\hat{y}_3(x)$ and $z(x)$, $\hat{z}(x)$ the terms δ and $\frac{1}{h}\theta$ appear in the difference of sequences $(y_{3,n})$, $(\hat{y}_{3,n})$ and (z_n), $(\hat{z}_n)$.

A detailed analysis of the influence of perturbations on the solution is of special interest if the differential-algebraic system is constructed using some stabilizing procedure for the numerical solution of higher-index systems (cf. Remark 1b). For systems (1) with $\mu = 0$ and $f(\eta,\zeta) = f_0(\eta) - g_y^T(\eta)\cdot\zeta$ that arise e. g. in the integration of overdetermined systems of index 2 ([Füh88]) both analytical and numerical solution y depend continuously on perturbations if the algebraic constraints g are linear (Theorems 1b and 3b are applicable here because of $g_y^T \equiv const$). But in general the terms $\|\theta\|\cdot\|\theta'\|$ and $\frac{1}{h}\theta^2$ in part a) of these theorems do really appear in the perturbed solution. To illustrate this phenomenon systems with nonlinear constraints must be considered.

Example 3 The initial value problem

$$\begin{array}{rcl} y' & = & -g_y^T(y)z \\ 0 & = & g(y) \end{array} \quad , \quad g(y) = g(y_1,y_2) = y_1 + y_2^2 - 1 \quad , \quad x \in [0,1] \tag{43}$$

with $y(0) = (\,0\,,\,1\,)^T$ and $z(0) = 0$ has the exact solution $y_1(x) \equiv y_1(0) = 0$, $y_2(x) \equiv y_2(0) = 1$, $z(x) \equiv z(0) = 0$ and is integrated exactly by backward Euler method: $y_n = y(x_n)$, $z_n = z(x_n)$. Lemma 1 (see below) defines a sequence (η_n) that grows with $\mathcal{O}(\frac{1}{h}\Theta^2)$ if $nh = 1$ until an (arbitrary large) number Y is reached. If $(\hat{y}_n, \hat{z}_n)$ is defined as $\hat{y}_{2,n} := \eta_n$, $\hat{y}_{1,n} := 1 + (-1)^n\Theta - \hat{y}_{2,n}^2$,

$\hat{z}_{n+1} = -\frac{1}{h}(\hat{y}_{1,n+1} - \hat{y}_{1,n})$ then we get in (26) for system (43) the residuals $\delta_{n+1} = 0$, $\theta_{n+1} = (-1)^{n+1}\Theta$, i. e. $\delta = 0$, $\theta = \mathcal{O}(\Theta)$. At the end of interval $[0,1]$ it holds

$$\hat{y}_{2,n} \geq \hat{y}_{2,n-2} + \kappa\Theta^2 \geq \ldots \geq \hat{y}_{2,0} + \tfrac{n}{2}\kappa\Theta^2 = y_{2,n} + \mathcal{O}(\tfrac{1}{h}\theta^2)$$

with $nh = 1$ or $\hat{y}_{2,n} \geq Y$.

Result There is no component of the solution vector (y, z) of system (43) that depends continuously on small perturbations of the discretization method. Note, however, that the error in the differential components y of the *analytical* solution is bounded by Corollary 1.

Lemma 1 *Let the function $\varphi(\eta, \eta^*)$ be defined by*

$$\varphi(\eta, \eta^*) := (\eta - \eta^*)(\eta^* + \eta + \frac{1}{2\eta}) \ .$$

For any given $Y > 1$ there are positive constants $\kappa = \kappa(Y)$ and $\Theta_0 = \Theta_0(Y)$ so that for arbitrary $\Theta \in (0, \Theta_0]$ a sequence (η_n) can be defined with

$$\eta_0 = 1 \ , \ \eta_{n+1} = \eta_n + \mathcal{O}(\Theta) \ , \ \varphi(\eta_{n+1}, \eta_n) = 2(-1)^{n+1}\Theta \ , \ (n \geq 0)$$

and

$$\eta_{n+2} \geq \min(Y, \eta_n + \kappa\Theta^2) \ , \ (n \geq 0 \ , \ n \ \textit{even}) \ .$$

Proof The sequence (η_n) is recursively defined as sequence of solutions of the nonlinear equations $\varphi(\eta, \eta_n) = 2(-1)^{n+1}\Theta$.
The lemma is proven by induction. Let $n \geq 0$ be even and $\eta_n \geq 1$. At first we consider the case $\eta_n \leq Y$ and define η_{n+1} as smallest zero of $\varphi(\eta, \eta_n) - 2(-1)^{n+1}\Theta$ in

$$I_{n+1} := [\eta(-1), \eta(1)] \quad \text{with} \quad \eta(\tau) := \eta_n + \beta(\eta_n)(-1)^{n+1}\Theta + (\alpha(\eta_n) + \tau\epsilon)\Theta^2$$

and

$$\beta(\eta) := \tfrac{4\eta}{1+4\eta^2} \ , \ \alpha(\eta) := -\tfrac{1}{2}\beta^3(\eta)(1 - \tfrac{1}{2\eta^2}) \ , \ \epsilon := \tfrac{\beta^3(Y)}{8Y^2} \ .$$

I. e. $\eta_{n+1} = \eta(\tau_0)$ with a $\tau_0 \in [-1, 1]$. In the same way η_{n+2} is defined as smallest zero of $\varphi(\eta, \eta_{n+1}) - 2(-1)^{n+2}\Theta$ in $I_{n+2} := [\eta_n + \frac{\epsilon}{2}\Theta^2, \eta_n + (4 + 2\epsilon)\Theta^2]$. It remains the proof that these η_{n+1} and η_{n+2} are well defined. The functions $\varphi(\eta, \eta_n)$ and $\varphi(\eta, \eta_{n+1})$ are continuous. It is proven by Taylor expansion that $\varphi(\eta, \eta_n) - 2(-1)^{n+1}\Theta$ changes sign in I_{n+1} and $\varphi(\eta, \eta(\tau)) - 2(-1)^{n+2}\Theta$ changes sign in I_{n+2} (for *arbitrary* $\tau \in [-1, 1]$) if $\Theta_0 > 0$ is sufficiently small:

$$\varphi(\eta(\tau), \eta_n) = (\eta(\tau) - \eta_n)\frac{2}{\beta(\eta_n)} + (\eta(\tau) - \eta_n)^2(1 - \frac{1}{2\eta_n\eta(\tau)}) = 2(-1)^{n+1}\Theta + \tau\frac{2\epsilon}{\beta(\eta_n)}\Theta^2 + \mathcal{O}(\Theta^3)$$

and

$$\begin{aligned}
\varphi(\eta, \eta(\tau)) &= (\eta - \eta(\tau))\frac{2}{\beta(\eta_n)} + (\eta - \eta(\tau))((\eta - \eta_n)(1 - \frac{1}{2\eta\eta_n}) + (\eta(\tau) - \eta_n)) \\
&= 2(-1)^{n+2}\Theta + \left(\frac{2}{\beta(\eta_n)}(-\alpha(\eta_n) + \lambda - \tau\epsilon) - \beta^2(\eta_n)\right)\Theta^2 + \mathcal{O}(\Theta^3) \\
&= 2(-1)^{n+2}\Theta + \frac{2}{\beta(\eta_n)}(-\frac{\beta^3(\eta_n)}{4\eta_n^2} + \lambda - \tau\epsilon)\Theta^2 + \mathcal{O}(\Theta^3)
\end{aligned}$$

for $\eta = \eta_n + \lambda\Theta^2$. If $1 \leq \eta_n \leq Y$ then

$$-4 \leq -4(\frac{4\eta_n}{1 + 4\eta_n^2}) \cdot (\frac{1}{1 + 4\eta_n^2})^2 = -\frac{\beta^3(\eta_n)}{4\eta_n^2} \leq -\frac{\beta^3(Y)}{4Y^2} = -2\epsilon$$

i. e.

$$\begin{aligned}
\varphi(\eta_n + \frac{\epsilon}{2}\Theta^2, \eta(\tau)) &\leq 2(-1)^{n+2}\Theta + \frac{2}{\beta(\eta_n)}(-2\epsilon + \frac{\epsilon}{2} + \epsilon)\Theta^2 + \mathcal{O}(\Theta^3) < 2(-1)^{n+2}\Theta \\
\varphi(\eta_n + (4+2\epsilon)\Theta^2, \eta(\tau)) &\geq 2(-1)^{n+2}\Theta + \frac{2}{\beta(\eta_n)}(-4 + 4 + 2\epsilon - \epsilon)\Theta^2 + \mathcal{O}(\Theta^3) > 2(-1)^{n+2}\Theta
\end{aligned}$$

for $\Theta \leq \Theta_0$ and arbitrary $\tau \in [-1, 1]$.

If $\eta_n > Y$, (n even), then the constants in the $\mathcal{O}(\Theta^3)$–terms could become arbitrary large. But it holds $\varphi(\eta_n - \Theta, \eta_n) < -2\Theta = 2(-1)^{n+1}\Theta$ and $\varphi(\eta_n, \eta_n) = 0$ so that η_{n+1} can be defined as above but with $I^*_{n+1} := [\eta_n - \Theta, \eta_n]$ instead of I_{n+1}. With this η_{n+1} we have $\varphi(\eta_{n+1} + \Theta, \eta_{n+1}) > 2\Theta = 2(-1)^{n+2}\Theta$ and

$$\varphi(\eta_n, \eta_{n+1}) = \frac{\eta_n - \eta_{n+1}}{2\eta_n} - \varphi(\eta_{n+1}, \eta_n) + \frac{\eta_{n+1} - \eta_n}{2\eta_{n+1}} = -2(-1)^{n+1}\Theta - \frac{(\eta_{n+1} - \eta_n)^2}{2\eta_n\eta_{n+1}} < 2(-1)^{n+2}\Theta$$

and η_{n+2} is defined as smallest zero of $\varphi(\eta, \eta_{n+1}) - 2(-1)^{n+2}\Theta$ in $I^*_{n+2} := [\eta_n, \eta_{n+1} + \Theta]$. This sequence satisfies $\eta_{n+2} \geq \min(Y, \eta_n + \frac{\epsilon}{2}\Theta^2)$, so we set $\kappa := \frac{\epsilon}{2}$ to complete the proof. ∎

In Example 3 the amplification of perturbations by a factor $\frac{1}{h}$ is the result of the discretization of (1) by backward Euler method. But in general also time-continuous problems (1) with $f(\eta, \zeta) = f_0(\eta) - g_y^T(\eta) \cdot \zeta$ are affected by this phenomenon if the algebraic part consists of more than one constraints:

Example 4 The initial value problem

$$\begin{aligned} y' &= -g_y^T(y) z \\ 0 &= g(y) \end{aligned} \quad , \quad g(y) = g(y_1, y_2, y_3) = \begin{pmatrix} y_1^2 - y_2 \\ y_1 - y_3 \end{pmatrix} \quad , \quad x \in [0, 1] \tag{44}$$

with $y(0) = (\,0\,,\,0\,,\,0\,)^T$ and $z(0) = (\,0\,,\,0\,)^T$ has a constant solution $y(x) \equiv y(0), z(x) \equiv z(0)$. Following Remark 2 there are uniquely defined functions $(\hat{y}(x), \hat{z}(x))$ with

$$\begin{aligned} \hat{y}' &= -g_y^T(\hat{y})\hat{z} \\ \theta(x) &= g(\hat{y}) \end{aligned} \quad , \quad \theta(x) = \begin{pmatrix} \Theta \sin \frac{x}{\varepsilon} \\ \Theta \cos \frac{x}{\varepsilon} \end{pmatrix}$$

and $\hat{y}(0) = (\,0\,,\,0\,,\,-\Theta\,)^T$, $\hat{z}(0) = (\,-\frac{\Theta}{\varepsilon}\,,\,0\,)^T$ provided that the positive parameters Θ and ε satisfy $\varepsilon \leq \varepsilon_0$ and $\Theta \leq \varepsilon_0 \varepsilon^{0.5}$ with a sufficiently small constant $\varepsilon_0 > 0$ ($\|\theta(x)\| = \mathcal{O}(\Theta)$, $\|\theta'(x)\| = \mathcal{O}(\frac{\Theta}{\varepsilon})$).

Because of $g(y(x)) = 0$ and $g(\hat{y}(x)) = \theta(x)$ we have

$$\hat{y}(1) - y(1) = \Phi(\hat{y}(1)) - \Phi(y(1)) + \mathcal{O}(\Theta)$$

with

$$\Phi(\eta) := \eta - [g_y^T(g_y g_y^T)^{-1} g](\eta) + \frac{1}{2}[(\frac{\partial}{\partial y} g_y^T(g_y g_y^T)^{-1})(g, g_y^T(g_y g_y^T)^{-1} g)](\eta)$$

(the tensor notation was introduced in (14) in the proof of Theorem 1). It holds $y'(x) = 0$, $\Phi(y(0)) = y(0) = 0$ and $\Phi(\hat{y}(0)) = \hat{y}(0) + \mathcal{O}(\Theta) = \mathcal{O}(\Theta)$ and thus

$$\begin{aligned}
\hat{y}(1) - y(1) &= \Phi(\hat{y}(0)) - \Phi(y(0)) + \int_0^1 \frac{d}{dx}(\Phi(\hat{y}(x)) - \Phi(y(x)))\, dx + \mathcal{O}(\Theta) \\
&= \int_0^1 \Phi_\eta(\hat{y}(x))\hat{y}'(x) dx + \mathcal{O}(\Theta) \ .
\end{aligned}$$

Differentiating the algebraic constraint we get $\theta'(x) = g_y(\hat{y})\hat{y}' = -[g_y g_y^T](\hat{y})\hat{z}$ and

$\Phi_\eta(\hat{y}(x))\hat{y}'(x) = [\Phi_\eta g_y^T(g_y g_y^T)^{-1}](\hat{y}(x)) \cdot \theta'(x)$. Expansion of $[\Phi_\eta g_y^T(g_y g_y^T)^{-1}](\eta)$ in a Taylor series w. r. t. $g(\eta)$ results in

$$[\Phi_\eta g_y^T(g_y g_y^T)^{-1}](\eta) = \Psi(\eta) + \mathcal{O}(\|g(\eta)\|^2)$$

with the matrix

$$\Psi(\eta) = \frac{1}{2}\Big([(\frac{\partial}{\partial y} g_y^T(g_y g_y^T)^{-1})(I, g_y^T(g_y g_y^T)^{-1} g)](\eta) - [(\frac{\partial}{\partial y} g_y^T(g_y g_y^T)^{-1})(g, g_y^T(g_y g_y^T)^{-1})](\eta)\Big)$$

that has elements

$$\psi_{kl}(\eta) = \frac{1}{2}\sum_{j=1}^{n_z}\sum_{i=1}^{n_y}(\frac{\partial \phi_{kl}}{\partial y_i}\phi_{ij} - \frac{\partial \phi_{kj}}{\partial y_i}\phi_{il}) g_j \quad .$$

(The functions ϕ_{kl} are the elements of matrix $g_y^T(g_y g_y^T)^{-1}$.)
For system (44) matrix $\Psi(\eta)$ is computed as

$$\Psi(\eta) = \lambda(\eta)\ (\,1\,,\,2\eta_1\,,\,1\,)^T \cdot (\,g_2(\eta)\,,\,-g_1(\eta)\,) \quad \text{with} \quad \lambda(\eta) := \frac{1}{4(1+2\eta_1^2)^2} \quad .$$

We get

$$[\Phi_\eta g_y^T(g_y g_y^T)^{-1}](\hat{y}(x)) \cdot \theta'(x) = \lambda(\hat{y}(x)) \begin{pmatrix} 1 \\ 2\hat{y}_1(x) \\ 1 \end{pmatrix} \cdot (\Theta \cos\frac{x}{\varepsilon}\,,\,-\Theta \sin\frac{x}{\varepsilon}) \begin{pmatrix} \frac{\Theta}{\varepsilon}\cos\frac{x}{\varepsilon} \\ -\frac{\Theta}{\varepsilon}\sin\frac{x}{\varepsilon} \end{pmatrix} + \mathcal{O}(\frac{\Theta^3}{\varepsilon})$$

i. e.

$$\hat{y}_1(1) - y_1(1) = \int_0^1 \lambda(\hat{y}(x))dx \cdot \frac{\Theta^2}{\varepsilon} + \mathcal{O}(\frac{\Theta^3}{\varepsilon}) + \mathcal{O}(\Theta) \quad .$$

If $\hat{y}_1(x) \le 1$ for all $x \in [0,1]$ then $\int_0^1 \lambda(\hat{y}_1(x))dx \ge \frac{1}{36}$. Thus there is a (small) constant $\varepsilon_0 > 0$ so that for at least one $x \in [0,1]$ the inequality $\hat{y}_1(x) - y(x) \ge \frac{1}{72}\frac{\Theta^2}{\varepsilon}$ is satisfied for all $\varepsilon \le \varepsilon_0$ and $\Theta \le \varepsilon_0 \varepsilon^{0.5}$. So the differential component y_1 does not depend continuously on small perturbations $\theta(x)$ of the algebraic part. The same result holds for the other components of the solution since $\hat{y}_2 = \hat{y}_1^2 + \mathcal{O}(\Theta)$, $\hat{y}_3 = \hat{y}_1 + \mathcal{O}(\Theta)$ and $\hat{z} = \mathcal{O}(\frac{\Theta}{\varepsilon})$.

5 Summary

While the perturbation index of a differential-algebraic system describes the influence of small perturbations on *all* solution components simultaneously the real influence on the *differential* components of semi-explicit systems like (1) is much smaller and depends furthermore on the special structure of the system. The same qualitative behaviour can be observed for the discretization by backward Euler method. We summarize these results in Table 1, for simplicity we omitted the perturbations in the initial values. We used the simplifying notations $\int \delta = \max \| \int \delta(s)ds \|$, $\int P\delta = \max \| \int P(s)\delta(s)ds \|$, $\|\delta\| = \max \|\delta(\xi)\|$, $D = D(x) = \|\delta\| + \|\theta'\|$ and so on and $\int_h \delta = h\sum_m \|\delta_{m+1}\|$, $\int_h P\delta = h\sum_m \|P_m\delta_{m+1}\| + h\delta$. The summands h^p, h^q and h^r stand for the (global) discretization error of the numerical method, $p = q = r = 1$ for backward Euler method. The error bounds depend on the special structure of system (1), for systems with $f(\eta,\zeta) = f_0(\eta) + f_z(\eta)\cdot\zeta$ they are valid even with $D := 0$, $D_h := 0$. We did not include the "exotic" case of systems with only one algebraic constraint and $f(\eta,\zeta) = f_0(\eta) + f_z(\eta)\zeta$. Here the differential components of the analytical solution are index-1 variables (i. e. $D = 0$, Corollary 1) but perturbations in the numerical solution are nevertheless amplified (additional error term $\theta D_h = \theta(\delta + \frac{1}{h}\theta)$, see Example 3). The examples of Section 4 illustrate that in general the

	analytical solution $\|\hat{y}(x)-y(x)\|,\|\hat{z}(x)-z(x)\|$	numerical solution $\|\hat{y}_n-y(x_n)\|,\|\hat{z}_n-z(x_n)\|$
ODE's	$C_0 \int \delta$	$C_0(h^p + \int_h \delta)$
systems (1)		
differential components y	$C_0(\int P\delta + \|\theta\| + \|\theta\| D + \mu D^2)$	$C_0(h^q + \int_h P\delta + \theta + \theta D_h + \mu D_h^2)$
algebraic components z	$C_0(\|\delta\| + \|\theta\| + \|\theta'\|)$	$C_0(h^r + \delta + \frac{1}{h}\theta)$

Table 1: Bounds for the influence of small perturbations on the solution of ODE's and differential-algebraic systems (1)

bounds of Table 1 can not be improved. Note, that the results that are summarized in Table 1 can be proven only if the perturbed solution remains in a small neighbourhood of the exact solution. So we supposed in Theorems 1 and 3 that the expressions $\|\theta\|D$, μD^2, $\overline{\mu}D$, θD_h, μD_h^2, and $\overline{\mu}D_h$ are sufficiently small.

The bounds for $\|\hat{y}_n - y(x_n)\|$ give some orientation for the implementation of the discretization method. The additional error term should be kept small during integration, e. g. choosing (very) small stopping bounds for the iterative solution of systems of nonlinear equations ([HLR89, Section 9]). The parts of the code that are highly sensitive to small perturbations (e. g. stepsize control) should be based on that solution components that are robust against perturbations. In the Runge–Kutta code RADAU5 this is done e. g. scaling the error in the algebraic components by h ([HLR89, Section 10]). Error amplification can cause difficulties if the stepsize becomes very small.

From this point of view we compare the straightforward discretization (24) of system (1) with the discretization of the stabilized index-1 formulation (4) that is analytically equivalent to (1) (see Remark 1b). The stabilized index-1 formulation is *linear* in the *new* algebraic variables u, i. e. the additional error term in y is of size $\mathcal{O}(\theta D_h)$. If the differential-algebraic system (1) is nonlinear in the algebraic variables z and $\mu = \max\|f_{zz}\| = \mathcal{O}(1)$ then the additional consideration of $\frac{d}{dx}g(y) = 0$ in the stabilized index-1 formulation reduces the additional error term in y by the factor h. Compared with the straightforward discretization (24) the discretization of the stabilized index-1 formulation is more robust against round-off errors and errors in the solution of systems of nonlinear equations if the stepsize h is very small. On the other hand the dimension of the system is increased by n_z equations. In both cases the (original) algebraic variables z can be computed with the same accuracy as the differential components y from $[g_y f](y,z) = 0$.

We want to point out that the solution does neither depend continuously on perturbations in the original system (1) nor on perturbations in the stabilized index-1 formulation (4). For systems with a very special structure even a stable discretization can become unstable if the time derivative $\frac{d}{dx}g(y) = 0$ is coupled to the system (Example 6). But this "worst case" does not play any part in practical computations since the "bad" term $\frac{1}{h}\theta^2$ is very small for errors θ and stepsizes h that do really occur in practice. For the same reason the amplification of errors in the straightforward discretization (24) of system (1) will not cause trouble if the system is linear in the (original) algebraic variables z or if $\mu \ll 1$.

From a theoretical point of view the stabilized index-1 formulation seems to be very interesting if the algebraic constraints g are linear (then $g_y^T = const$ and $D_h = 0$). But in practice systems (1) with linear constraints are of no special interest because these systems can be reduced to a system of ODE's eliminating some of the differential components y ([Gea86]).

The implications of the perturbation analysis for practical computations can be summarized as follows: If the bounds for stopping the iterative solution of nonlinear equations are sufficiently small then the differential components are usually affected by small perturbations in the same way as in ODE theory. Only for very small stepsizes errors can drastically be amplified during integration if system (1) depends non-linearly on the algebraic components. These difficulties can be overcome adding the derivative $\frac{d}{dx}g(y(x)) = [g_y f](y,z) = 0$ of the algebraic constraint to the system before discretization.

6 Numerical Tests

In Sections 3 and 4 the influence of small perturbations on the numerical solution was studied in great detail. But we did not discuss the size of the constants in the $\mathcal{O}(.)$–terms. So it remains to illustrate the amplification of perturbations for stepsizes h that are actually used in the integration of systems (1).

We apply implicit Euler method (24) to a system that is perturbed in the algebraic part. For larger stepsizes the error is dominated by the (global) discretization error that is $\mathcal{O}(h)$. But for smaller stepsizes the error grows rapidly if the system is nonlinear in the algebraic variables. The error remains in the size of the perturbations ($\mathcal{O}(\theta)$) if the system is linear in z (i. e. if $\mu = 0$). The (very) small additional error term $\mathcal{O}(\frac{1}{h}\theta^2)$ can not be observed numerically for stepsizes $h \geq 5.0e-7$. But if the perturbations are defined by hand to get a large influence on the numerical solution ($\theta_n = (-1)^n\Theta$) then the $\mathcal{O}(\frac{1}{h}\theta^2)$–term is clearly visible in the error plot (Example 6).

Example 5 We applied implicit Euler method to the initial value problem

$$\begin{aligned} y_1' &= y_1^2 - 0.5y_2 - ((1-\mu_0)y_1 + \mu_0 z)z \\ y_2' &= (2(1-\mu_0)y_1^2 + \mu_0(y_1 z + z^2) + 0.5y_2^2)z \\ 30\theta &= 4y_1^2 + y_2^2 - 4 \quad , \quad x \in [0.5, 0.6] \end{aligned}$$

with the exact solution $y_1(x) = \cos x$, $y_2(x) = 2\sin x$, $z(x) = \cos x$ and an artificial pertur-

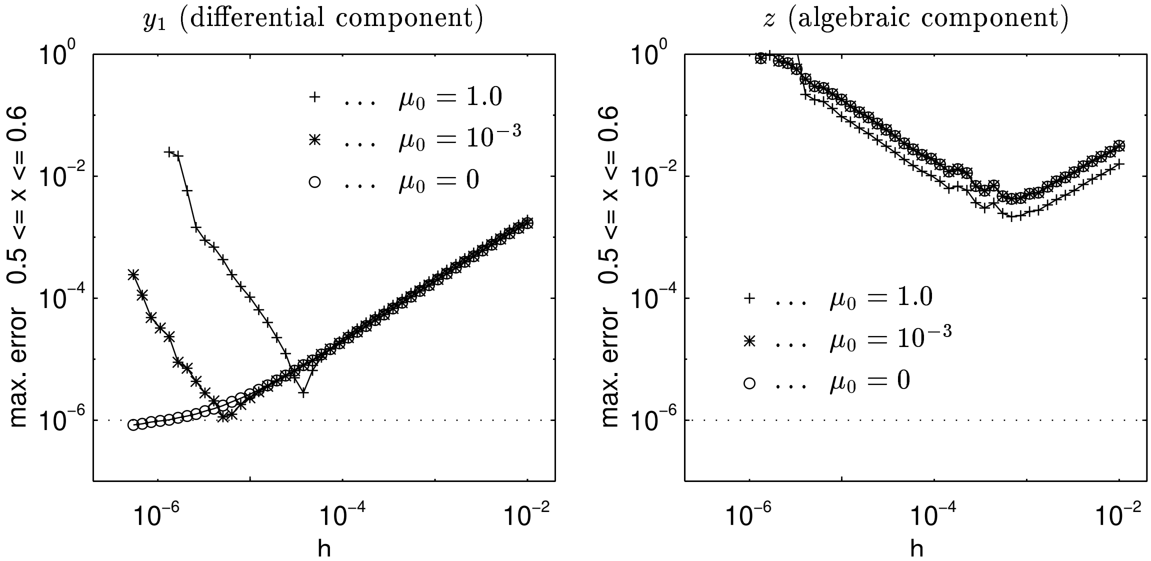

Figure 1: Global error for Example 5

bation $\theta(x)$ that vanishes for the analytical solution and is computed as difference between $\sin x$ in `extended`-arithmetic (10 bytes) and $\sin x$ in `single`-arithmetic (4 bytes), ($30\theta \approx 10^{-6}$, machine eps $\approx 10^{-19}$). The parameter μ_0 controls the structure of the system: if $\mu_0 = 0$ then

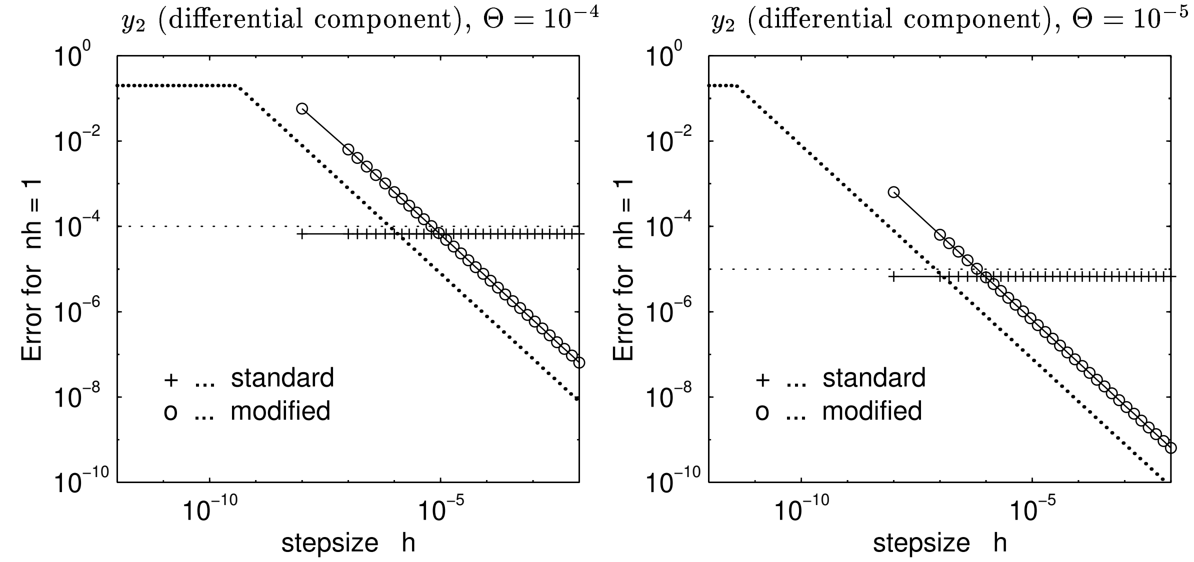

Figure 2: Global error for Example 6

the system is linear in z, otherwise $f(y,z)$ depends nonlinearly on the algebraic variables, the upper bound μ for $\|f_{zz}\|$ is of size $\mathcal{O}(\mu_0)$.
We did not succeed to illustrate the influence of the perturbation on a numerical solution that is computed by a variable stepsize solver since the integration breaks down for smaller tolerances. So we plotted for $\mu_0 = 1$, $\mu_0 = 10^{-3}$ and $\mu_0 = 0$ the errors in one differential and in the algebraic component versus (fixed) stepsize h (Figure 1). For larger stepsizes the error is decreased with $\mathcal{O}(h)$, the error constant is nearly independent of μ_0. For smaller stepsizes the additional error term dominates (it is of size $\mathcal{O}(\frac{1}{h}\theta^2 + \mu\frac{1}{h^2}\theta^2)$ for the differential and of size $\mathcal{O}(\frac{1}{h}\theta)$ for the algebraic components). For $\mu_0 = 0$ the error in y remains in the size of the perturbations.

Example 6 The previous example underlines that the discretization is more robust against perturbations if the system is linear in the algebraic variables. But from a theoretical point of view it is interesting that not only the straightforward implementation but also the discretization of the stabilized index-1 formulation does not depend continuously on perturbations for $h \to 0$. To illustrate the worst possible case that could happen if the time derivative $\frac{d}{dx}g(y) = 0$ is coupled to the system we integrated the initial value problem

$$\begin{aligned} y_1' &= z \\ y_2' &= z \\ 0 &= y_1 + y_2^2 - 1 \quad , \quad x \in [0,1] \end{aligned}$$

with the exact solution $y_1(x) \equiv 0$, $y_2(x) \equiv 1$, $z(x) \equiv 0$. In Figure 2 the straightforward discretization (24) is compared with implicit Euler method applied to the stabilized index-1 formulation (4) ("standard" and "modified" method, respectively). We added an artificial perturbation $\theta_{n+1} = (-1)^{n+1}\Theta$ to the algebraic equation: $g(\hat{y}_{n+1}) = \theta_{n+1}$, $\delta = 0$. The equation $[g_y f](y_{n+1}, z_{n+1}) = 0$ in the modified method is solved exactly, i. e. $z_{n+1} = 0$. The stabilized index-1 formulation (4) of this system is given by (43) in Example 3 (read "$g_y^T(y)u$" instead of "$g_y^T(y)z$").
We plotted for two values of Θ the error in the differential component y_2. Because of the very simple right hand side $f(y,z) = (z,z)^T$ the differential components depend continuously on small perturbations if the straightforward discretization (24) is used ($D_h = 0$), the error is $\mathcal{O}(\Theta)$. As it was proven in Sections 3 and 4 the error is $\mathcal{O}(\frac{1}{h}\Theta^2)$ for the discretization of the stabilized

index-1 formulation (the errors in the algebraic components z are of size $\mathcal{O}(\frac{1}{h}\Theta)$ and $\mathcal{O}(\frac{1}{h}\Theta^2)$, respectively). I. e. instead of decreasing the influence of perturbations on the numerical solution the additional consideration of $\frac{d}{dx}g(y(x)) = 0$ *increases* in this example the error for very small stepsizes.
We included in the plots the lower bound $\min(Y-1, \frac{n}{2}\kappa\Theta^2)$ from Example 3 as dotted lines, ($Y = 1.2$). It is clear to be seen that even for large perturbations (e. g. $\Theta = 10^{-5}$) the error term $\mathcal{O}(\frac{1}{h}\Theta^2)$ is not relevant to practical computations. For stepsizes $h > 10^{-6}$ it is (much) smaller than the other error terms.

Acknowledgement We would like to thank C. Führer who focussed our attention to the comparison of analytically equivalent formulations for a given differential-algebraic problem and St. L. Campbell who suggested the extension of the results for $\mu = 0$ to the case $\mu \ll 1$.

References

[Arn95a] M. Arnold. Applying BDF to quasilinear differential-algebraic equations of index 2: a perturbation analysis. Technical Report, University of Rostock, Department of Mathematics, 1995, in preparation.

[Arn95b] M. Arnold. A perturbation analysis for the dynamical simulation of mechanical multibody systems. *Applied Numerical Mathematics*, to appear, 1995.

[BCP89] 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.

[FL91] C. Führer and B. Leimkuhler. Numerical solution of differential–algebraic equations for constrained mechanical motion. *Numer. Math.*, 59:55–69, 1991.

[Füh88] C. Führer. Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen. Theorie, numerische Ansätze und Anwendungen. Technical report, TU München, Mathematisches Institut und Institut für Informatik, 1988.

[Gea71] C.W. Gear. The simultaneous numerical solution of differential-algebraic equations. *IEEE Trans. Circuit Theory*, CT-18:89–95, 1971.

[Gea86] C.W. Gear. Maintaining solution invariants in the numerical solution of ode's. *SIAM J. Sci. Stat. Comp.*, 7:734–743, 1986.

[GGL85] C.W. Gear, G.K. Gupta, and B. Leimkuhler. Automatic integration of Euler–Lagrange equations with constraints. *J. Comp. Appl. Math.*, 12&13:77–90, 1985.

[HLR89] 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.

[HW91] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations.II. Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin Heidelberg New York, 1991.

[Lub89] Ch. Lubich. Linearly implicit extrapolation methods for differential-algebraic systems. *Numer. Math.*, 55:197–211, 1989.

[Mär89] R. März. Analysis and numerical treatment of differential-algebraic systems. In R.E. Bank, R. Bulirsch, and K. Merten, editors, *Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices. Proc. of a Conference held at the Math. Forschungsinstitut, Oberwolfach, Oct. 30 – Nov. 5, 1988*, pages 27–43, Basel, 1989. Birkhäuser-Verlag.

[Wen92] J. Wensch. Stabilität von Runge–Kutta–Methoden für Algebrodifferentialgleichungen vom Index 2. Martin-Luther-Universität Halle-Wittenberg, FB Mathematik und Informatik, Diplomarbeit, 1992.