10/4-11-90 JS ①

UCID 21964
ISCR 1989-01

Differential-Algebraic Equations as Stiff Ordinary Differential Equations

Michael Knorrenschild

May 1989

Institute for
Scientific Computing Research

Lawrence Livermore National Laboratory

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (615) 576-8401, FTS 626-8401.

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161

Price Code	Page Range
A01	Microfiche
Papercopy Prices	
A02	1- 10
A03	11- 50
A04	51- 75
A05	76-100
A06	101-125
A07	126-150
A08	151-175
A09	176-200
A10	201-225
A11	226-250
A12	251-275
A13	276-400
A14	301-325
A15	401-425
A16	351-375
A17	376-400
A18	401-425
A19	426-450
A20	451-475
A21	476-500
A22	501-525
A23	526-550
A24	551-575
A25	576-600
A99	601 & UP

UCID--21964

DE90 007922

Differential-Algebraic Equations as Stiff Ordinary Differential Equations[1]

Michael Knorrenschild[2]

Abstract: In this paper we show that differential-algebraic systems of index-1 can always be viewed as reduced problems from singular perturbed ODEs. Applying implicit Runge-Kutta methods to the singular perturbed system, we gain new insight into the relationship of order-reduction phenomena observed for stiff ODEs to that for differential-algebraic equations. We show that the order of convergence achieved for index-1-differential/algebraic equations is at least the order of B-convergence.

Subject classification: AMS(MOS): 65L05, 34A99

Keywords: Differential algebraic equations, singular perturbations, implicit Runge-Kutta methods, stiff ordinary differential equations, B-convergence.

[1] This report is almost identical with Report #56, June 1989, of the Institut für Geometrie und Praktische Mathematik, Rheinisch Westfälisch Technische Hochschule, 5100 Aachen, Federal Republic of Germany. This work is part of the author's Ph.D. dissertation written under supervision of Prof. R. Jeltsch.

[2] Portions of this work were performed at the Institute for Scientific Computing Research under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48. Michael Knorrenschild's present address (January 1990): Department of Mathematics and Statistics, Simon Fraser University, Burnaby, B.C. V5A 1S6, Canada.

2b

1. Introduction

In the past years considerable progress has been made in the development of numerical methods for differential-algebraic equations (DAEs) of the form

$$\begin{aligned} &(1.1a) \qquad & y'(t) &= f(t,y(t),z(t)) \qquad & f &: \mathbb{R}^{m+k+1} \to \mathbb{R}^m \\ &(1.1b) \qquad & 0 &= g(t,y(t),z(t)) \qquad & g &: \mathbb{R}^{m+k+1} \to \mathbb{R}^k \end{aligned}$$

One area of interest where DAEs arise is the study of singular perturbed ordinary differential equations

$$(1.2) \qquad \begin{aligned} y'(t) &= f(t,y(t),z(t),\varepsilon) \\ \varepsilon z'(t) &= g(t,y(t),z(t),\varepsilon) \end{aligned}$$

where $\varepsilon > 0$ is a small parameter. Then the so-called "reduced" system is obtained by setting $\varepsilon = 0$. Under certain conditions the solution of the reduced system is a good approximation to the solution of (1.2). However in practice not all DAEs derive from singular perturbations. In many applications ($1.1b$) arises as a condition that describes geometrical constraints or conservation of energy.

However (1.1) can obviously be formally embedded in a singular perturbed problem like (1.2) defining $f(.,.,.,0) = f(.,.,.)$, $g(.,.,.,0) = g(.,.,.)$. Ignoring the question whether this makes sense in view of the mathematical background, this gives a widely used trick to derive numerical methods for DAEs from numerical methods for ordinary differential equations:

apply a method for ordinary differential equations to (1.2), put $\varepsilon = 0$ in the resulting recursion formula (if this is feasible) and obtain a numerical method for (1.1). E. g. in the case of the implicit Euler method that would yield

$$\begin{aligned} y_{n+1} &= y_n + h \cdot f(t_{n+1}, y_{n+1}, z_{n+1}) \\ 0 &= g(t_{n+1}, y_{n+1}, z_{n+1}) \end{aligned}$$

where $h > 0$ is the stepsize and $t_{n+1} = t_n + h$, $y_n \approx y(t_n)$, $z_n \approx z(t_n)$ ($y(t), z(t)$ is the exact solution of (1.1) at time t). In a similar way a large class of implicit Runge-Kutta methods can be applied as well as multistep-methods like the BDF-methods. It is easy to see that this approach does not work with explicit methods.

As these methods are all derived from numerical methods for stiff ordinary differential equations it is natural to compare the gained results with the corresponding results for stiff ODEs. Recently some work in this direction has been done ([Pe1], [BuPe], [HaLuRo], [Ro]).

However, the convergence theory for these methods for DAEs was based on an analysis performed on the DAE-system; the derivation-process via stiff ODEs has not been taken into account. On the other hand, it is well known that the order of convergence that these method achieve on DAEs is roughly the same as the order for certain classes of stiff ODEs. In addition some aspects in the proofs of the results for stiff ODEs indicate that this is

not a coincidence (see [BuPe]). However the missing link so far has been the connection between some singular perturbed system and the "reduced system" (1.1) since it turned out that for the purpose of analysis (1.2) is a rather clumsy form.
The outline of this paper is as follows:
In section 2. we give for a large class of DAEs a singular perturbed system that reduces to the DAE for $\varepsilon = 0$ and show that the solution of the perturbed system converges to that of the DAE in a certain sense.
In section 3. we derive discrete analogues to the theorems presented in the previous section. This allows a unified approach to the convergence of implicit Runge-Kutta methods for stiff ODEs on the one hand and DAEs on the other. This way we clarify the relation between the order of B-convergence (see [DV]) and the order of convergence for DAEs.

2. The Regularization-Process

We restrict ourselves to the study of the autonomous case and to the socalled *index-1-case*

2.1 **Definition:**

The DAE

$$\begin{array}{lrcll} (2.1.1a) & y'(t) & = & f(y(t),z(t)) & f \in C^{(}\mathbb{R}^{m+k},\mathbb{R}^{m}) \\ (2.1.1b) & 0 & = & g(y(t),z(t)) & g \in C^{1}(\mathbb{R}^{m+k},\mathbb{R}^{k}) \end{array}$$

is said to have *index*-1 iff the jacobian $g_z(y,z)$ is regular on $\mathbb{R}^{m+k}$.

In essence the index-1-property guarantees local solvability of the DAE: if an initial value (y_0, z_0) with $g(y_0, z_0) = 0$ is given, then there is a mapping G defined in a neighbourhood of y_0 such that

$$g(y,z) = 0 \iff z = G(y)$$

in a neighbourhood of (y_0, z_0) and $z_0 = G(y_0)$.
Using this we can rewrite the initial value problem for the DAE as

$$\begin{array}{rclrcl} y'(t) & = & f(y(t),G(y(t))) & y(0) & = & y_0 \\ z(t) & = & G(y(t)) & z(0) & = & z_0 = G(y_0) \end{array}$$

and then we have local solvability. More precise we will use the following property of an index-1-DAE:

(A1) there is $S_1 \subseteq \mathbb{R}^m$, $S_2 \subseteq \mathbb{R}^k$, S_1, S_2 open and $\bar{S}_1$, $\bar{S}_2$ compact, $G \in C^1(S_1, S_2)$ bijective with

$$g(y,z) = 0 \iff z = G(y) \qquad \text{for all } y \in S_1,\, z \in S_2$$

While looking for a singular perturbed system that reduces for "$\varepsilon = 0$" to a given DAE we should take care that the solutions of the singular perturbed problem converge in some sense towards the solution of the DAE as $\varepsilon \to 0$. The convergence property we can expect is described by the following definition.

2.2 **Definition:**

Let $F : \mathbb{R}^{2n+2} \to \mathbb{R}^n$,

$$(*) \qquad F(t, x(t), x'(t), \varepsilon) = 0.$$

(i) Let ε be fixed. $x_0 \in \mathbb{R}^n$ is a *consistent initial value* for $(*)$ iff $(*)$ has a solution $x \in C^1([a,b], \mathbb{R}^n)$ for some $b > a$ with $x(a) = x_0$

Now let $\bar{x}(t)$, $t \in [a,b]$ be a solution of $(*)$ for $\varepsilon = 0; x_a := \bar{x}(a)$, U a subspace of $\mathbb{R}^n$.

(ii) $(*)$ is said to be ε - *stable on* $[a,b]$ $:\iff$
There is an $\varepsilon_0 > 0$ such that

$(*)$ has for each $0 < \varepsilon \leq \varepsilon_0$ and for each consistent initial value $x_{a\varepsilon} \in \mathbb{R}^n$ with $\|x_{a\varepsilon} - x_a\| < \varepsilon_0$ a solution $x_\varepsilon(t)$ such that $x_\varepsilon(a) = x_{a\varepsilon}$ and

$$\lim x_\varepsilon(t) = \bar{x}(t)$$

uniformly on $[a,b]$ for

$$\varepsilon + \|x_{a\varepsilon} - x_a\| \to 0.$$

(iii) $(*)$ is said to be ε - *stable on* $[a,b]$ *for* U $:\iff$
$(*)$ is ε - stable and
there is $\delta > 0, \varepsilon_0 > 0$ such that

$(*)$ has for each $0 < \varepsilon \leq \varepsilon_0$ and for each $u_a \in U$, $\|u_a\| < \delta$, $x_a + u_a$ consistent for $(*)$, a solution $x_\varepsilon(t)$ with $x_\varepsilon(a) = x_a + u_a$ and

$$\lim_{\varepsilon \to 0+} x_\varepsilon(t) = \bar{x}(t), \qquad \text{uniformly on } [c,d] \text{ where } a < c < d \leq b.$$

In other words small perturbations out of the subspace U added to the initial value do not destroy the uniform convergence on compact subintervals of $]a,b] := \{x \in \mathbb{R} \,|\, a < x \leq b\}$.

In general we will use the following criterion to guarantee ε - stability. We denote for a set M $B(M,\delta) := \{x \mid \|x - y\| < \delta \text{ for some } y \in M\}$.

2.3 **Theorem:** [LeLe]

Let the singular perturbed ordinary differential equation

$$(*) \qquad \begin{aligned} y'(t) &= f(t,y(t),z(t),\varepsilon) \qquad & y(t) \in \mathbb{R}^m \\ \varepsilon z'(t) &= g(t,y(t),z(t),\varepsilon) \qquad & z(t) \in \mathbb{R}^{n-m} \end{aligned}$$

have for $\varepsilon = 0$ and the initial value (y_a, z_a) a solution $(y_0(t), z_0(t))$ on $[a,b]$ with $y_0(a) = y_a, z_0(a) = z_a$;
where for a $\delta_0 > 0$ and

$$R := [a,b] \times B(y_0([a,b]),\delta_0) \times B(z_0([a,b]),\delta_0) \times [0,\delta_0[$$

f, f_y, f_z, g, g_y, g_z are continous on R; for all $t \in [a,b]$ we assume

$$\lambda \textit{ eigenvalue of } g_z(t,y_0(t),z_0(t),0) \Rightarrow Re(\lambda) < 0$$

Then $(*)$ is ε - stable on $[a,b]$ for $U := \{\mathcal{O}_m\} \times \mathbb{R}^{n-m}$.

Our goal is to give for any index-1-DAE a singular perturbed problem $F(t,x,x',\varepsilon) = 0$ that is ε - stable, i. e. we want to regularize our system. Regularizations for linear DAEs have already been studied in [Ca], but the analysis uses some requirement on the spectrum of the involved matrices and seems in general not a practical way. Here we will approach the regularization problem in a natural way and we are able to do so under very general assumptions.
Since we want to use Theorem 2.3 which is formulated for explicit systems we have to take care that $F(t,x,x',\varepsilon) = 0$ is solvable for the derivative x'. Systems which allow this will be called *index-0- systems*. At first we present the basic regularization which will be generalized later.

2.4 **Theorem:**

Let the DAE (2.1.1) *have index-1, $f, g \in C^1$, and (A1) be satisfied. We assume that for a $L_G > 0$*

$$||G(y) - G(\tilde{y})|| \le L_G ||y - \tilde{y}|| \quad \textit{for all } y, \tilde{y} \in D_1,$$

Let $y_0 \in D_1$, $z_0 := G(y_0)$, $b > 0$, $\bar{y} : [0,b] \to D_1$, $\bar{y}(0) = y_0$, $\bar{z}(t) := G(\bar{y}(t))$ for $t \in [0,b]$ with

$$\bar{y}'(t) = f(\bar{y}(t), \bar{z}(t)) \quad \textit{for all } t \in [0,b]$$

let $F \in C^1(\mathbb{R}^k, \mathbb{R}^k)$, F bijective and
for all $z \in \mathbb{R}^k$ $\qquad \lambda$ eigenvalue of $F'(z) \Rightarrow Re(\lambda) > 0$
Then

$$(2.4.1) \qquad \begin{aligned} y'(t) &= f(y(t), F^{-1}(F(z(t)) + \varepsilon z'(t))) \\ 0 &= g(y(t), F^{-1}(F(z(t)) + \varepsilon z'(t))) \end{aligned}$$

has for all $\varepsilon > 0$ index-0 and there is a $C > 0$ such that:
for all $(\tilde{y}_0, \tilde{z}_0) \in D_1 \times \mathbb{R}^k$ there is a $\tilde{b} \in]0, b]$ and a solution $(y_\varepsilon(t), z_\varepsilon(t))$ of (2.4.1) on $[0, \tilde{b}]$ with

$$\|y_\varepsilon(t) - \bar{y}(t)\| \leq C\|\tilde{y}_0 - y_0\| \qquad \text{for all } t \in [0, \tilde{b}].$$

Furthermore for each $a \in]0, \tilde{b}]$ there is a $\omega_a : \mathbb{R} \times \mathbb{R}^3_{\geq 0} \to \mathbb{R}_{\geq 0}$ with

$$\begin{array}{rclllr} \lim \omega_a(t, \varepsilon, d1, d2) & = & 0 & \text{for } \varepsilon + d1 + d2 \to 0 & \text{uniformly on} & t \in [0, \tilde{b}] \\ \lim \omega_a(t, \varepsilon, d1, 0) & = & 0 & \text{for } \varepsilon \to 0 & \text{uniformly on} & t \in [a, \tilde{b}] \end{array}$$

and

$$\|z_\varepsilon(t) - \bar{z}(t)\| \leq \omega_a(t, \varepsilon, \|\tilde{z}_0 - G(\tilde{y}_0)\|, \|\tilde{y}_0 - y_0\|) + L_G C\|\tilde{y}_0 - y_0\|$$

for all $t \in]0, \tilde{b}]$.
In particular (2.4.1) is ε - stable on $[0, \tilde{b}]$ for $\{\mathcal{O}_m\} \times \mathbb{R}^{n-m}$.

Proof:
The index-0-property is easily seen: Put $\tilde{g}(y, z, z', \varepsilon) := g(y, F^{-1}(F(z) + \varepsilon z'))$. It is sufficient to prove the regularity of $\tilde{g}_{z'}$. Here we have

$$\tilde{g}_{z'}(y, z, z', \varepsilon) = \varepsilon g_z(y, F^{-1}(F(z) + \varepsilon z'))(F^{-1})'(F(z) + \varepsilon z')$$

and observing $(F^{-1})'(F(z) + \varepsilon z') = F'(F^{-1}(F(z) + \varepsilon z'))$ and the regularity of F' (which is guaranteed by assumption) we are done.
Let $(\tilde{y}_0, \tilde{z}_0) \in S_1 \times \mathbb{R}^k$. From the classical theory we know that there is a $\tilde{b} \in]0, b]$ and $\tilde{y}(t)$ with $\tilde{y}(0) = \tilde{y}_0$ and

$$\tilde{y}'(t) = f(\tilde{y}(t), G(\tilde{y}(t))) \qquad \text{for all } t \in [0, \tilde{b}].$$

We now define $y_\varepsilon(t) := \tilde{y}(t)$ on $[0, \tilde{b}]$ and z_ε as the solution of the initial value problem

$$(2.4.2) \qquad \begin{array}{rclcrclc} F^{-1}(F(z_\varepsilon(t)) + \varepsilon z'_\varepsilon(t)) & = & G(\tilde{y}(t)), & & z_\varepsilon(0) & = & \tilde{z}_0, & t \in [0, \tilde{b}] \\ \iff F(z_\varepsilon(t)) + \varepsilon z'_\varepsilon(t) & = & F(G(\tilde{y}(t))), & & z_\varepsilon(0) & = & \tilde{z}_0, & t \in [0, \tilde{b}] \end{array}$$

Obviously y_ε and z_ε satisfy (2.4.1). It remains to prove the error estimate.
It is well known that the solution of $y' = f(y, G(y))$ depends Lipschitz-boundedly on the initial value. Therefore there is a $c > 0$ with

$$\|\tilde{y}(t) - \bar{y}(t)\| \leq C\|\tilde{y}_0 - y_0\|, \ t \in [0, \tilde{b}],$$

which proves the first part already. This implies immediately

$$(2.4.3) \qquad \|G(\tilde{y}(t)) - G(\bar{y}(t))\| \leq L_G C\|\tilde{y}_0 - y_0\| \text{ uniformly on } [0, \tilde{b}].$$

If we invoke Theorem 2.3 on (2.4.2), (the assumptions of which are satisfied since the eigenvalues of $-F'(z)$ have negative real part), we arrive for $a \in]0, \tilde{b}]$ at

$$\|z_\varepsilon(t) - G(\tilde{y}(t))\| = \omega_a(t, \varepsilon, \|\tilde{z}_0 - G(\tilde{y}_0)\|, \|\tilde{y}_0 - y_0\|) \tag{2.4.4}$$

with a certain ω_a that has the desired properties. From (2.4.3) and (2.4.4) the claim follows immediately. ■

Now we want to have a brief look at multiparameter-perturbations. The simplest situation is when all perturbation parameters involved have a constant ratio. But we are interested in criteria that do not involve these ratios (since from a practical point of view knowledge about the ratios may not be available). The essential assumption here is the diagonal dominance.

2.5 **Definition:**

We call a $n \times n$-Matrix (a_{ij}) *strictly diagonally dominant* iff

$$a_{ii} > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \qquad \text{for } i = 1, \dots, n.$$

2.6 **Theorem:** (Multiparameterversion I of Theorem 2.4)

Let the assumptions of Theorem 2.4 be satisfied;
assume that for all $z \in \mathbb{R}^k$ the matrix $F'(z)$ is strictly diagonally dominant;
for $i = 1, \dots, k$ put $\varepsilon_i := d_i \cdot \varepsilon$, where $d_i > 0$; $D := diag(d_1, \dots, d_k)$;
Then

$$(*) \qquad \begin{array}{rcl} y'(t) & = & f(y(t), F^{-1}(F(z(t)) + \varepsilon D z'(t))) \\ 0 & = & g(y(t), F^{-1}(F(z(t)) + \varepsilon D z'(t))) \end{array}$$

has for all $\varepsilon > 0$ index-0 and the same claim as in 2.4 holds.

Proof:
Put

$$\tilde{F}(z) := D^{-1} F(z) \qquad \text{for } z \in \mathbb{R}^n$$

and verify the assumptions of 2.4.
From the strict diagonal dominance and the Gershgorin-theorem we have for all $z \in \mathbb{R}^k$: λ eigenvalue of $F'(z) \Rightarrow Re(\lambda) > 0$.
The same holds for the matrix $D^{-1} \cdot F'(z) = \tilde{F}'(z)$ since it is strictly diagonally dominant too. Now we are done, since we can apply 2.4 with $\tilde{F}$ and we get the desired result. ■

Under more restrictive assumptions we have the following result which guarantees convergence for $\varepsilon_1, \ldots, \varepsilon_k \to 0$ without requiring the ratios of the ε_i being constant.

2.7 Theorem: (Multiparameterversion II of Theorem 2.4)

Let the assumptions of Theorem 2.6 be satisfied;
in addition assume $F(z) = (F_1(z_1), \ldots, F_k(z_k))^T$ for all $z \in \mathbb{R}^k$ and $F_i'(z_i) > 0$ for $i = 1, \ldots, k$ (the latter follows also from the strict diagonally dominance)
Then

$$(*) \qquad \begin{aligned} y'(t) &= f(y(t), F^{-1}(F(z(t)) + \mathcal{E} z'(t))) \\ 0 &= g(y(t), F^{-1}(F(z(t)) + \mathcal{E} z'(t))) \end{aligned}$$

where $\mathcal{E} = diag(\varepsilon_1, \ldots, \varepsilon_k)$, $\quad \varepsilon_i > 0$
has always index-0 and the claim in 2.6 holds, if one replaces ε by $\mathcal{E}$ and

$$\omega_a(t, \varepsilon, \|\tilde{z}_0 - G(\tilde{y}_0)\|, \|\tilde{y}_0 - y_0\|)$$

by

$$\omega_a(t, \|\mathcal{E}\|, \|\tilde{z}_0 - G(\tilde{y}_0)\|, \|\tilde{y}_0 - y_0\|).$$

Proof:
The proof is analogous to 2.4; but now $z_{\mathcal{E}}$ is not defined by (2.4.1), but by

$$\begin{aligned} & F^{-1}(F(z_{\mathcal{E}}(t)) + \mathcal{E} z_{\mathcal{E}}'(t)) &= \quad & G(\tilde{y}(t)), & z_{\mathcal{E}}(0) &= \tilde{z}_0, \quad t \in [0, b] \\ \iff \quad & F(z_{\mathcal{E}}(t)) + \mathcal{E} z_{\mathcal{E}}'(t) &= \quad & F(G(\tilde{y}(t))), & z_{\mathcal{E}}(0) &= \tilde{z}_0, \quad t \in [0, b] \\ \iff \quad & F_i(z_{\mathcal{E}i}(t)) + \mathcal{E} z_{\mathcal{E}i}'(t) &= \quad & F_i(G(\tilde{y}(t))), & z_{\mathcal{E}i}(0) &= \tilde{z}_{0i}, \quad t \in [0, b] \\ & & & & & \text{for } i = 1, \ldots, k \end{aligned}$$

But these equations are decoupled. Therefore we can proceed as in the proof to 2.6, by dealing with the scalar equations rather than the complete system. ∎

Remark:
Theorems 2.4-2.7 may be formulated with obvious modifications also for nonautonomous DAEs.

The presented introduction of perturbation parameters is by no means some kind of artificial technique as it may seem at a first glance. Actually, these small parameters can be given a physical meaning:
DAEs arise for example in electrical network analysis. If we consider a large class of networks, namely that composed solely of capacitors, resistors and current sources, then the

vector of unknowns is given by the voltages in the capacitor-branches (the y-component) and the voltages in the resistor-branches (the z-component). Using Kirchhoff's laws we arrive at a DAE. But this model is somewhat idealized. It does not take into account so-called parasitic effects (like lead inductance, stray capacitance ...). To refine the model one can introduce parasitic elements (by replacing each resistor by a resistor in parallel with a small capacity "ε").
This procedure leads exactly to our regularized system: 2.4 describes the case when all resistors have nonlinear, strictly monotone increasing characteristics and all parasitic capacitances have the same order. However to be more exact one should allow parasitic elements of different magnitudes (i.e. introducing ε_i, $i = 1, \ldots$) and this is exactly the situation considered in 2.6, 2.7 (and the situation in the networks described leads always to a function F that satisfies the requirements in Theorem 2.7). A detailed discussion on regularization aspects for DAEs arising in electrical network analysis can be found in [Kn] and will be continued in a forthcoming paper.

3. Implicit Runge-Kutta Methods on Regularized DAEs

In this section we study the behaviour of implicit Runge-Kutta schemes when applied to our perturbed DAEs. Similar work for BDF-methods has been done by Lötstedt [Lö1], [Lö2]. We derive for a large class of methods applied to (2.4.1) with linear F bounds for the global discretization error that are uniform in h and ε. The results extend easily to the situation considered in Theorems 2.6 and 2.7 (i.e. the multiparameter-case).
In the following we denote the Kronecker-product of square matrices by $A \otimes B$ (see e. g. [DV]). Runge-Kutta methods are given by (s, A, b) where s is the number of stages, A is the $s \times s$ - coefficient matrix and b is the vector of weights. We define $c_i := \sum_{j=1}^{s} a_{ij}$.
Application of that method to an implicit differential equation $F(t, x, x') = 0$ then gives:

Solve

$$F\left(t_n + c_i h, x_n + h \sum_{j=1}^{s} a_{ij} X'_j, X'_i\right) = 0 \quad i = 1, \ldots, s$$

and compute

$$x_{n+1} = x_n + h \sum_{i=1}^{s} b_i X'_i.$$

This gives approximations x_n to the true solution $\bar{x}(t_n)$ where $t_n = t_0 + nh$. Equivalently if A is regular we can proceed as follows ($D := (d_{ij}) = A^{-1}$):

Solve

$$F\left(t_n + c_i h, X_i, \frac{1}{h}\sum_{j=1}^{s} d_{ij}(X_j - x_n)\right) = 0, \quad i = 1, \ldots, s$$

and compute

$$x_{n+1} = x_n + \sum_{i=1}^{s} b_i \sum_{j=1}^{s} d_{ij}(X_j - x_n).$$

Furthermore we will use the "simplifying conditions": For $p \in \mathbb{N}$ a Runge-Kutta method (s, A, b) satisfies

$$B(p) \quad :\Longleftrightarrow \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k} \quad \text{for } k = 1, \ldots, p$$

$$C(p) \quad :\Longleftrightarrow \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k \quad \text{for } i = 1, \ldots, s;\ k = 1, \ldots, p$$

Later we will need the coefficients

$$d_i := \frac{c_i^{q+1}}{(q+1)!} - \frac{\sum_{j=1}^{s} a_{ij} c_j^q}{q!}, \qquad d := (d_1, \ldots, d_s)^T.$$

We denote the property that the local error is of order p by $A(p)$, i. e. let x_{n+1} be the approximation that the method computes within one step starting with $x_n := \bar{x}(t_n)$. Then $A(p)$ means

$$\|x_{n+1} - \bar{x}(t_{n+1})\| \leq C h^{p+1} \quad \text{for all } h \in [0, h_0]$$

for a certain $h_0 > 0,\ C > 0$.

We know (see e. g. [DV]) that for all $\bar{x} \in C^{p+1}([t_0, t_0 + h])$, property $C(p)$ implies

$$\bar{x}(t_0 + c_i h) = \bar{x}(t_0) + h \sum_{j=1}^{s} a_{ij} \bar{x}'(t_0 + c_j h) + O(h^{p+1}) \qquad i = 1, \ldots, s$$

and $B(p)$ implies

$$\bar{x}(t_0 + h) = \bar{x}(t_0) + h \sum_{i=1}^{s} b_i \bar{x}'(t_0 + c_i h) + O(h^{p+1}).$$

While applying our method to the perturbed system we expect the numerical solution $y_{\nu\varepsilon} \approx y_\varepsilon(t_\nu)$, $z_{\nu\varepsilon} \approx z_\varepsilon(t_\nu)$ (notation as in 2.4) to be dependent on two small parameters namely the perturbation-parameter $\varepsilon > 0$ and the stepsize $h > 0$. If ε is considered to be fixed we can apply the classical theory and get bounds like

$$\left\| \begin{pmatrix} y_{\nu\varepsilon} \\ z_{\nu\varepsilon} \end{pmatrix} - \begin{pmatrix} y_\varepsilon(t_\nu) \\ z_\varepsilon(t_\nu) \end{pmatrix} \right\| \le Ch^p.$$

Since we want to have approximations for the true solution of the DAE $(\bar{y}(t), \bar{z}(t))^T$ we can derive from this (by simply using the triangle-inequality)

$$\left\| \begin{pmatrix} y_{\nu\varepsilon} \\ z_{\nu\varepsilon} \end{pmatrix} - \begin{pmatrix} \bar{y}(t_\nu) \\ \bar{z}(t_\nu) \end{pmatrix} \right\| \le Ch^p + \underbrace{\left\| \begin{pmatrix} y_\varepsilon(t_\nu) \\ z_\varepsilon(t_\nu) \end{pmatrix} - \begin{pmatrix} \bar{y}(t_\nu) \\ \bar{z}(t_\nu) \end{pmatrix} \right\|}_{\to 0 \text{ for } \varepsilon \to 0}$$

But the constant C always involves higher derivatives of the true solution of the perturbed system $(y_\varepsilon(t), z_\varepsilon(t))^T$ and these are in general unbounded in ε (since essentially z'_ε is of the form $z'_\varepsilon(t) = \varepsilon^{-1}(\bar{z}(t) - z_\varepsilon(t))$).
If under these circumstances we want to keep the discretization error Ch^p small we have to choose h very small compared to ε. Obviously this is not desirable in real computation. But since we know that $(y_\varepsilon(t), z_\varepsilon(t))^T$ is close to $(\bar{y}(t), \bar{z}(t))^T$ we expect that appropiate methods will yield numerical solutions $(y_{\nu\varepsilon}, z_{\nu\varepsilon})^T$ quite close to $(\bar{y}(t), \bar{z}(t))^T$ even if ε is small compared to h. We hope to arrive at estimates like

$$\left\| \begin{pmatrix} y_{\nu\varepsilon} \\ z_{\nu\varepsilon} \end{pmatrix} - \begin{pmatrix} \bar{y}(t_\nu) \\ \bar{z}(t_\nu) \end{pmatrix} \right\| \le C_1\varepsilon + C_2 h^p,$$

where C_1 does not depend on h and C_2 does not depend on ε. This way we would get

$$\left\| \begin{pmatrix} y_{\nu\varepsilon} \\ z_{\nu\varepsilon} \end{pmatrix} - \begin{pmatrix} y_\varepsilon(t_\nu) \\ z_\varepsilon(t_\nu) \end{pmatrix} \right\| \le C_1\varepsilon + C_2 h^p + \underbrace{\left\| \begin{pmatrix} y_\varepsilon(t_\nu) \\ z_\varepsilon(t_\nu) \end{pmatrix} - \begin{pmatrix} \bar{y}(t_\nu) \\ \bar{z}(t_\nu) \end{pmatrix} \right\|}_{\to 0 \text{ for } \varepsilon \to 0}$$

That means we have to avoid derivatives of $(y_\varepsilon(t), z_\varepsilon(t))^T$ in our error bound. The reward would be an error bound uniform in $\varepsilon \in [0, \varepsilon_0]$ and $h \in [0, h_0]$ and we would not have to worry about restrictions like $h < \varepsilon$ (or similar). The appropiate idea in this context is the *B-convergence* theory. There one deals with systems of ordinary differential equations $y' = f(y)$ where f satisfies a one-sided Lipschitz condition $(f(y) - f(z), y - z) \le \mu\|y - z\|^2$, where $(.,.)$ is a scalar product. The crucial point here is that μ may be negative. The goal of the B-convergence approach is to derive global error bounds Ch^p where C is not allowed to depend on the usual Lipschitz constant but only on the one-sided μ. That way one gets error bounds that are uniform on *classes* of right hand sides f. This is important for stiff systems where it is characteristic that the usual Lipschitz constant is very large whereas the one-sided may be of moderate size. In our perturbed problem we deal basically with systems $\varepsilon z' = -z + \ldots$ and that means that the usual Lipschitz constant is essentially $1/\varepsilon$, whereas the one-sided Lipschitz constant could be taken as $-1/\varepsilon$ or also $\mu = 0$. Setting $\mu = 0$ represents a uniform one-sided Lipschitz constant suitable for every $\varepsilon > 0$ and promises to lead us in the direction of uniform error bounds.

Since the requirement not to make use of the classical Lipschitz constant imposes restrictions on the kind of error bounds obtained it is not surprising that the order of B-convergence of a method may be lower than its classical order (it may happen that a method is not B-convergent at all, but convergent in the classical sense). In many cases the order of B-convergence reflects the practical behaviour of a method applied to a stiff system better than the classical order. For a detailed discussion on these aspects see the excellent book by Dekker and Verwer ([DV]).

Unfortunately we cannot apply the results of the B-convergence theory immediately to our perturbed DAEs. Although that way we get rid of the (classical) Lipschitz constant we still have higher derivatives of the analytical solution involved in the error bound. Also one can show that in general our perturbed system does not satisfy a one-sided Lipschitz condition with μ independent of ε. But it turns out that we can use very similar techniques if we interlace the exact solution of the DAE (which of course does not depend on ε).

We will denote in the following

$$e := (1, \ldots, 1)^T \in \mathbb{R}^s, \quad v \otimes I_k := \begin{pmatrix} v_1 \cdot I_k \\ \vdots \\ v_s \cdot I_k \end{pmatrix} \in \mathbb{R}^{ks \times k},$$

$$v^T \otimes I_k := (v_1 \cdot I_k, \ldots, v_s \cdot I_k) \in \mathbb{R}^{k \times ks} \quad \text{for } v \in \mathbb{R}^s, \quad E := e \otimes I_k.$$

Before we present our main theorem we will describe the some technical assumptions that are needed. Assuming (A1) is satisfied we will make use of

(A2) The underlying DAE (2.1.1) has an exact solution $(\bar{y}, \bar{z})^T$ on an interval $[t_0, t]$ for an initial value (y_0, z_0) such that $(\bar{y}(t), \bar{z}(t))^T \in S_1 \times S_2$. Note, that this implies $g(y_0, z_0) = 0$). $\bar{y}$, $\bar{z}$, f, g are sufficiently smooth.

(A3) The numerical approximations computed by the underlying method and all stage values arising in this process remain in $S_1 \times S_2$ for the initial value $(y_0, \tilde{z}_0)$, where $\tilde{z}_0$ is sufficiently close to z_0.

For (A2) we refer to the remark after Definition 2.1; (A3) is always satisfied when h is sufficiently small.

The main result of this paper is

3.1 Theorem

Let (s, A, b) be an A-stable Runge-Kutta method such that A have eigenvalues with positive real part only and $C(q)$, $B(p)$ and $A(p)$, where $l \geq p+1$, are satisfied. Let (A1) and (A2) be satisfied.
$M \in \mathbb{R}^{k \times k}$ with $(Mx, x) \geq 0$ for all $x \in \mathbb{R}^k$ for a inner product $(.,.)$ on $\mathbb{R}^k$, $\varepsilon_0 > 0$, $\|.\|$ the corresponding norm, $\delta > 0$,
Let (A3) be satisfied for the method applied to

$$\begin{aligned} y' &= f(y, M^{-1}(Mz + \varepsilon z')) \\ 0 &= g(y, M^{-1}(Mz + \varepsilon z')), \end{aligned}$$

$e_0 := \tilde{z}_0 - \bar{z}(t_0)$
Then

$$\begin{aligned} e_{y\nu} &:= y_\nu - \bar{y}(t) = O(h^p) \\ e_\nu &:= z_\nu - \bar{z}(t) = O(\|e_0\|) + O(\varepsilon) + O(h^q) + O(h^p). \end{aligned}$$

Furthermore, if $\psi(z) := [b^T(I_s - Az)^{-1}e]^{-1} b^T(I_s - Az)^{-1} d$ is uniformly bounded on $\mathbb{C}^-$ then we have

$$e_\nu = O(\|e_0\|) + O(\varepsilon) + O(h^{q+1}) + O(h^p).$$

Note that the crucial point here is that the starting value $\tilde{z}_0$ is not required to be $\bar{z}_0$. In this aspect our approach differs from that in [HaLuRo], [Ro].

Proof:
Without restriction we assume that the h_0 mentioned in the assumption is small enough to meet all the requirements that will be made in the following proof.
At first we want to compute one step with our method starting at $(y_\mu, z_\mu) \in S_1 \times S_2$ applied to (2.1.1), i. e. the $\varepsilon = 0$ case.
The stage values Y_i, Z_i, $i = 1, \ldots, s$ in that case are given as the solution of

$$\begin{aligned} Y_i &= y_\mu + h \sum_{j=1}^{s} a_{ij} f(Y_j, Z_j) \\ 0 &= g(Y_i, Z_i) \end{aligned} \qquad i = 1, \ldots, s \tag{3.1.1}$$

We are just interested in solutions $Y_i \in S_1$, $Z_i \in S_2$ $(i = 1, \ldots, s)$, so this is equivalent with

$$\begin{aligned} Y_i &= y_\mu + h \sum_{j=1}^{s} a_{ij} f(Y_j, G(Y_j)) \\ Z_i &= G(Y_i). \end{aligned} \qquad i = 1, \ldots, s \tag{3.1.2}$$

We want to apply the implicit function theorem to this system in order to derive the existence of solutions $Y_i(h)$ and therefore also $Z_i(h)$ that are continous in h.
The nonlinear system of equations to solve is

$$F(h,\vec{Y})=0 \qquad \vec{Y}:=\begin{pmatrix} Y_1\\ \vdots \\ Y_s\end{pmatrix},$$

where

$$F(h,\vec{Y}):=(F_1(h,\vec{Y}),\dots,F_s(h,\vec{Y}))^T$$
$$F_i(h,\vec{Y}):=Y_i-y_\mu-h\sum_{j=1}^{s}a_{ij}f(Y_j,G(Y_j)) \qquad \text{for } i=1,\dots,s.$$

We have

$$F(0,(y_0,\dots,y_0)^T)=0, \qquad F_{\vec{Y}}(0,\vec{Y})=I_{ms} \quad \text{for all } \vec{Y}\in \mathbb{R}^{ms},$$

which allows us to solve for $Y(h)$ as desired. In particular if h_0 is small enough there exists $\vec{Y}\in C([0,h_0],\mathbb{R}^{ms})$ with

$$F(h,\vec{Y}(h))=0 \qquad \text{on } [0,h_0].$$

Since

$$y_{\mu+1}=y_\mu+h\sum_{i=1}^{s}b_i f(Y_i(h),G(Y_i(h)))$$

it is clear that for h small enough $y_{\mu+1}$ will be in S_1. Also it is obvious that we have done nothing else than applying our method to the ODE $y'=f(y,G(y))$. Using the classical techniques one can derive the assertion for the y-component.

It remains to prove the assertion for the z-component.
Define $M_\varepsilon:=\frac{1}{\varepsilon}M$.
One step with our method for the regularized system requires stage values $Y_{i\varepsilon}$, $Z_{i\varepsilon}$ satisyfing

$$\begin{aligned} Y_{i\varepsilon}&=y_\mu+h\sum_{j=1}^{s}a_{ij}f(Y_{j\varepsilon},Z_{j\varepsilon}+M_\varepsilon^{-1}Z'_{j\varepsilon}) \\ 0&=g(Y_{i\varepsilon},Z_{i\varepsilon}+M_\varepsilon^{-1}Z'_{i\varepsilon})\end{aligned} \qquad i=1,\dots,s$$

where as usual $Z'_{i\varepsilon}$ is implicitly defined by

$$Z_{i\varepsilon}=z_\mu+h\sum_{j=1}^{s}a_{ij}Z'_{j\varepsilon} \qquad i=1,\dots,s \tag{3.1.3}$$

Therefore it is obvious that

$$Y_{i\varepsilon}:=Y_i \qquad i=1,\dots,s$$

and $Z_{i\varepsilon}$ satisfies

$$(3.1.4) \qquad \begin{array}{rcll} Z_{i\varepsilon} + M_\varepsilon^{-1} Z'_{i\varepsilon} & = & Z_i, & i = 1, \ldots, s \\ \Longleftrightarrow \qquad Z'_{i\varepsilon} & = & M_\varepsilon (Z_i - Z_{i\varepsilon}), & i = 1, \ldots, s. \end{array}$$

If we denote

$$\vec{Z}_\varepsilon := \begin{pmatrix} Z_{1\varepsilon} \\ \vdots \\ Z_{s\varepsilon} \end{pmatrix} \in \mathbb{R}^{ks}, \qquad \vec{Z}_0 := \begin{pmatrix} Z_1 \\ \vdots \\ Z_s \end{pmatrix} \in \mathbb{R}^{ks},$$

(3.1.3) and (3.1.4) lead us to

$$(3.1.5) \qquad \vec{Z}_\varepsilon = E z_\mu + h(A \otimes I_k)(I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon).$$

Using Lemma 2.4.6 in [Hu] we see that (3.1.5) has a solution $\vec{Z}_\varepsilon$ for every $h \geq 0$, $\varepsilon \geq 0$, since

$$(I_{ks} + h(A \otimes M_\varepsilon))^{-1} = (I + hAz)|_{z=M_\varepsilon}^{-1} = (I + hzA)^{-1}|_{z=M_\varepsilon}.$$

The approximation $z_{\mu+1}$ is given by

$$(3.1.6) \qquad z_{\mu+1} = z_\mu + h(b^T \otimes I_k)(I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon).$$

If we denote in addition

$$\bar{Z}_\mu := \begin{pmatrix} \bar{z}(t_\mu + c_1 h) \\ \vdots \\ \bar{z}(t_\mu + c_s h) \end{pmatrix}, \qquad \bar{Z}'_\mu := \begin{pmatrix} \bar{z}'(t_\mu + c_1 h) \\ \vdots \\ \bar{z}'(t_\mu + c_s h) \end{pmatrix},$$

we have (using Taylor-expansion)

$$(3.1.7) \qquad \begin{array}{rcl} \bar{Z}_\mu & = & E\bar{z}(t_\mu) + h(A \otimes I_k)\bar{Z}'_\mu + r_\mu \\ \bar{z}(t_\mu + h) & = & \bar{z}(t_\mu) + h(b^T \otimes I_k)\bar{Z}'_\mu + O(h^{p+1}), \end{array}$$

where

$$r_\mu = (d \otimes I_k)\bar{z}^{(q+1)}(t_\mu)h^{q+1} + O(h^{q+2})$$

For the errors $e_{\mu+1} := z_{\mu+1} - \bar{z}(t_\mu + h)$ and $e_\mu := z_\mu - \bar{z}(t_\mu)$ we derive using (3.1.6) and (3.1.7):

$$(3.1.8) \qquad e_{\mu+1} = e_\mu + h(b^T \otimes I_k)((I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}'_\mu) + O(h^{p+1}).$$

Further (3.1.5) together with (3.1.7) gives

$$(3.1.9) \qquad \vec{Z}_\varepsilon - \bar{Z}_\mu = Ee_\mu + h(A \otimes I_k)((I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}'_\mu) - r_\mu$$

To make use of that in (3.1.8) we need a relation between $\bar{Z}_\mu$ and $\vec{Z}_0$. In order to get this we return temporarily to the events in the y-component.
Remembering the beginning of the proof and defining $F := f(.,G(.))$ and

$$D_i := h\int_0^1 F'(Y_i+\theta(\bar{y}(t_\mu+c_ih)-Y_i))\,d\theta \quad (i=1,\dots,s), \quad D := diag(D_1,\dots,D_s) \in \mathbb{R}^{ks\times ks}$$

we get (as in [BuHuVe]) ($\bar{Y}_\mu$ is defined analogous to $\bar{Z}_\mu$)

$$\text{(3.1.10)} \qquad \bar{Y}_\mu - \vec{Y} = (I_{ks} - (A\otimes I_k)D)^{-1}(E(y_\mu - \bar{y}(t_\mu)) + R_\mu)$$

where

$$R_\mu = (d\otimes I_k)\bar{y}^{(q+1)}(t_\mu)h^{q+1} + O(h^{q+2}).$$

Observe that $I_{ks} - (A\otimes I_k)D$ is regular for h small enough (since $D = O(h)$). Also $(I_{ks} - (A\otimes I_k)D)^{-1} = I_{ks} + O(h)$ and therefore (since $y_\mu - \bar{y}(t_\mu) = O(h^p)$ with $p \geq q+1$) we have $\bar{Y}_\mu - \vec{Y} = O(h^{q+1})$.
Using the mean value theorem for G and the s components in (3.1.10) we get

$$\begin{aligned}
\bar{Z}_\mu - \vec{Z}_0 &= \begin{pmatrix} G'(\bar{y}(t_\mu+c_1h)) & & \\ & \ddots & \\ & & G'(\bar{y}(t_\mu+c_sh)) \end{pmatrix} (\bar{Y}_\mu - \vec{Y}) + O(h^{2q+2}) \\
&= (I_s\otimes G'(\bar{y}(t_\mu))(\bar{Y}_\mu - \vec{Y}) + O(h^{q+2}) \\
&= (I_s\otimes G'(\bar{y}(t_\mu))(EO(h^p) + (d\otimes I_k)O(h^{q+1})) + O(h^{q+2}). \\
&= EG'(\bar{y}(t_\mu))O(h^p) + (d\otimes I_k)G'(\bar{y}(t_\mu))O(h^{q+1}) + O(h^{q+2}) \\
&= EO(h^p) + (d\otimes I_k)O(h^{q+1}) + O(h^{q+2}) =: \tilde{R}_\mu
\end{aligned}$$

Now insert this into (3.1.9) and we have

$$\vec{Z}_\varepsilon - \vec{Z}_0 = Ee_\mu + h(A\otimes I_k)((I_s\otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}'_\mu) + \tilde{R}_\mu - r_\mu.$$

But this means nothing else than

$$\begin{aligned}
& (I_{ks} + h(A\otimes M_\varepsilon))(\vec{Z}_\varepsilon - \vec{Z}_0) = Ee_\mu - h(A\otimes I_k)\bar{Z}'_\mu + \tilde{R}_\mu - r_\mu \\
\Longleftrightarrow \quad & (I_s\otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}'_\mu = \\
& \qquad (I_s\otimes M_\varepsilon)(I_{ks}+h(A\otimes M_\varepsilon))^{-1}[h(A\otimes I_k)\bar{Z}'_\mu \\
& \qquad -Ee_\mu - \tilde{R}_\mu + r_\mu - (I_{ks} + h(A\otimes M_\varepsilon))(I_s\otimes M_\varepsilon)^{-1}\bar{Z}'_\mu] \\
& = (I_s\otimes M_\varepsilon)(I_{ks} + h(A\otimes M_\varepsilon))^{-1}\left[-Ee_\mu - (I_s\otimes M_\varepsilon)^{-1}\bar{Z}'_\mu - \tilde{R}_\mu + r_\mu\right].
\end{aligned}$$

Inserting into (3.1.8) leads us to

$$
\begin{aligned}
e_{\mu+1} &= e_\mu + h(b^T \otimes I_k)(I_s \otimes M_\varepsilon)(I_{ks} + h(A \otimes M_\varepsilon))^{-1}\cdot \\
&\quad \cdot[-Ee_\mu - (I_s \otimes M_\varepsilon)^{-1}\bar{Z}'_\mu + r_\mu - \tilde{R}_\mu] + O(h^{p+1}).
\end{aligned}
\tag{3.1.11}
$$

If we denote

$$N(h,\varepsilon) := h(b^T \otimes I_k)(I_s \otimes M_\varepsilon)(I_{ks} + h(A \otimes M_\varepsilon))^{-1} \in \mathbb{R}^{k \times ks},$$

we finally arrive at

$$e_{\mu+1} = [I_k - N(h,\varepsilon)E]\, e_\mu - N(h,\varepsilon)\left[(I_s \otimes M_\varepsilon)^{-1}\bar{Z}'_\mu + r_\mu - \tilde{R}_\mu\right] + O(h^{p+1}). \tag{3.1.12}$$

Since we have

$$\bar{Z}'_\mu = E\bar{z}'(t_\mu) + O(h)$$

it is clear that

$$
\begin{aligned}
(I_s \otimes M_\varepsilon)^{-1}\bar{Z}'_\mu &= (I_s \otimes M_\varepsilon^{-1})\,[E\bar{z}'(t_\mu) + O(h)] = E \cdot M_\varepsilon^{-1}\bar{z}'(t_\mu) + \varepsilon(I_s \otimes M^{-1})O(h) \\
&= \varepsilon E \cdot M^{-1}\bar{z}'(t_\mu) + O(\varepsilon h).
\end{aligned}
$$

Now we can write (3.1.12) as

$$e_{\mu+1} = [I_k - N(h,\varepsilon)E]\, e_\mu - N(h,\varepsilon)\left[\varepsilon EO(1) + O(\varepsilon h) + r_\mu - \tilde{R}_\mu\right] + O(h^{p+1}).$$

Using an induction argument yields

$$
\begin{aligned}
e_\nu &= (I_k - N(h,\varepsilon)E)^\nu e_0 + \\
&\quad \sum_{j=0}^{\nu-1}(I_k - N(h,\varepsilon)E)^j\left[-N(h,\varepsilon)EO(\varepsilon) - N(h,\varepsilon)(O(\varepsilon h) + r_\mu - \tilde{R}_\mu) + O(h^{p+1})\right] \\
&= (I_k - N(h,\varepsilon)E)^\nu e_0 + ((I_k - N(h,\varepsilon)E)^\nu - I_k)O(\varepsilon) \\
&\qquad + \sum_{j=0}^{\nu-1}(I_k - N(h,\varepsilon)E)^j\left[-N(h,\varepsilon)(O(\varepsilon h) + r_\mu - \tilde{R}_\mu) + O(h^{p+1})\right].
\end{aligned}
$$

Furthermore we have

$$
\begin{aligned}
N(h,\varepsilon) &= h(b^T \otimes I_k)(I_s \otimes M_\varepsilon)(I_{ks} + h(A \otimes M_\varepsilon))^{-1} \\
&= \left[hb^T z(I + hzA)^{-1}\right]_{z=M_\varepsilon}.
\end{aligned}
$$

If we denote the *stability function* of our method by

$$R(z) := 1 + zb^T(I_s - zA)^{-1}e,$$

we then have

$$I_k - N(h,\varepsilon)E = R(-hM_\varepsilon).$$

If we further write the vector of polynomials $p_1(z) := -zb^T(I - zA)^{-1}$ we finally arrive at

$$\begin{aligned} e_\nu &= R(-hM_\varepsilon)^\nu e_0 + (R(-hM_\varepsilon)^\nu - 1)O(\varepsilon) \\ &+ \sum_{j=0}^{\nu-1} R(-hM_\varepsilon)^j \left[p_1(-hM_\varepsilon)(O(\varepsilon h) + r_\mu - \tilde{R}_\mu) + O(h^{p+1})\right] . \end{aligned} \tag{3.1.13}$$

Because of the A-stability we have $|R(-hM_\varepsilon)| \le 1$ for all h, $\varepsilon > 0$. This, and the uniform boundedness of p_1 (according to Lemma 4.2 in [BuHuVe], using the assumption on the eigenvalues of A) and $\nu h = const.$ gives:

$$e_\nu = R(-hM_\varepsilon)^\nu e_0 + O(\varepsilon) + \sum_{j=0}^{\nu-1} R(-hM_\varepsilon)^j p_1(-hM_\varepsilon)(r_\mu - \tilde{R}_\mu) + O(h^p).$$

But

$$\begin{aligned} &\sum_{j=0}^{\nu-1} R(-hM_\varepsilon)^j p_1(-hM_\varepsilon)(r_\mu - \tilde{R}_\mu) \\ &\quad = (1 - R(-hM_\varepsilon)^\nu)(1 - R(-hM_\varepsilon))^{-1} p_1(-hM_\varepsilon)EO(h^p) \\ &\qquad + \sum_{j=0}^{\nu-1} R(-hM_\varepsilon)^j p_1(-hM_\varepsilon)(d \otimes I_k)O(h^{q+1})) \\ &\quad = (1 - R(-hM_\varepsilon)^\nu)O(h^p) + \sum_{j=0}^{\nu-1} R(-hM_\varepsilon)^j p_1(-hM_\varepsilon)(d \otimes I_k)O(h^{q+1})) \end{aligned}$$

and from this the first part of the assertion follows immediately. To prove the second part one uses that

$$\sum_{j=0}^{\nu-1} R(-hM_\varepsilon)^j p_1(-hM_\varepsilon)(d \otimes I_k) = (1 - R(-hM_\varepsilon)^\nu)\psi(-hM_\varepsilon). \qquad \blacksquare$$

Having done that work already we can easily derive a multiparameterversion.

3.2 Theorem:

Let the assumptions of 3.1 be satisfied, but replace this time in the DAE ε by $\mathcal{E}$ where
$\mathcal{E} := diag(\varepsilon_1, \ldots, \varepsilon_k)$ with $\varepsilon_i > 0$ for $i = 1, \ldots, k$; M is assumed to be a diagonal matrix and $(.,.)$ denotes the Euclidean inner product

Then

$$\begin{aligned} e_{y\nu} &:= y_\nu - \bar{y}(t) = O(h^p) \\ e_\nu &:= z_\nu - \bar{z}(t) = O(\|e_0\|) + O(\|\mathcal{E}\|) + O(h^q) + O(h^p) \end{aligned}$$

Furthermore, if $\psi(z) := [b^T(I_s - Az)^{-1}e]^{-1}b^T(I_s - Az)^{-1}d$ is uniformly bounded on $\mathbb{C}^-$ then we have

$$e_\nu = O(\|e_0\|) + O(\|\mathcal{E}\|) + O(h^{q+1}) + O(h^p).$$

Proof:
Denote $M_\varepsilon := \mathcal{E}^{-1} \cdot M = M \cdot \mathcal{E}^{-1}$ and observe that for all $x \in \mathbb{R}^k$:

$$(M_\varepsilon x, x) = \sum_{i=1}^{k} \varepsilon_i^{-1} m_{ii} x_i^2 \geq 0,$$

since we have for all $y \in \mathbb{R}^k$ $0 \leq (My, y) = \sum_{i=1}^{k} m_{ii} y_i^2$. Hence $m_{ii} \geq 0$ for $i = 1, \ldots, k$.
Using this M_ε we can simply use the proof for 3.2; note that

$$\begin{aligned} (I_s \otimes (\mathcal{E}^{-1} \cdot M))^{-1} \bar{Z}'_\mu &= E \cdot M^{-1} \cdot \mathcal{E}\bar{z}'(t_\mu) + (I_s \otimes M^{-1})(I_s \otimes \mathcal{E})O(h) \\ &= EO(\|\mathcal{E}\|) + O(\|\mathcal{E}\|h). \end{aligned}$$

The remainder is obvious. ■

3.3 Corollary:

Let the assumptions of 3.1 resp. 3.2 be satisfied.

Then

$$\begin{aligned} e_{y\nu} &:= y_\nu - y_\varepsilon(t) = O(h^p) \\ e_\nu &:= z_\nu - z_\varepsilon(t) = O(\|e_0\|) + o(\varepsilon) + O(h^q) + O(h^p). \end{aligned}$$

Furthermore, if $\psi(z) := [b^T(I_s - Az)^{-1}e]^{-1}b^T(I_s - Az)^{-1}d$ is uniformly bounded on $\mathbb{C}^-$ then we have

$$e_\nu = O(\|e_0\|) + o(\varepsilon) + O(h^{q+1}) + O(h^p)$$

resp. corresponding assertion for $\mathcal{E}$.

Remarks:

1. Discussion of some methods:

The question for which methods ψ is uniform bounded on $\mathbb{C}^-$ has already been discussed in some detail in [BuHuVe]. For example this is the case for the Radau-IA-, Radau-IIA-, and Lobatto-IIIC-methods. Remember that for the s-stage methods we have (see e. g. [DV]):

$$\begin{aligned} \text{for the Gauss-methods} &: A(2s),\ B(2s),\ C(s) \\ \text{for the Radau-IA-methods} &: A(2s-1),\ B(2s-1),\ C(s-1) \\ \text{for the Radau-IIA-methods} &: A(2s-1),\ B(2s-1),\ C(s) \\ \text{for the Lobatto-IIIC-methods} &: A(2s-2),\ B(2s-2),\ C(s-1), \end{aligned}$$

In the y-component the order of convergence is the same as for ODEs, i. e. l, if we have $A(l)$.

Now we are able to state the final results:

The following orders of convergence for s-stage Runge-Kutta methods applied to (2.4.1) with a linear F hold uniform in ε near 0:

$$\begin{aligned} \text{for the Gauss-methods} &: s \quad \text{if } s \geq 2; \quad 2 \ \text{ if } s = 1 \\ \text{for the Radau-IA-methods} &: s \\ \text{for the Radau-IIA-methods} &: s+1 \quad \text{if } s \geq 2; \quad 1 \ \text{ if } s = 1 \\ \text{for the Lobatto-IIIC-methods} &: s \end{aligned}$$

The same orders of convergence are stated in [BuHuVe] as orders of B-convergence for the class of stiff ODEs

$$U'(t) = QU(t) + g(t, U(t)),$$

where

$$(Qu, u) \leq \beta \|u\|^2 \quad \forall u \in \mathbb{R}^n,$$
$$\|g(t,u) - g(t,\tilde{u})\| \leq \alpha \|u - \tilde{u}\| \quad \forall u, \tilde{u} \in \mathbb{R}^n,\ t \in \mathbb{R}.$$

This is of course no surprise since to a great extent our proof followed the lines of the proof in [BuHuVe] and consequently the assumptions in their theorem are exactly identical with our's. Let us point out again that it is not possible to use their theorem directly since our perturbed system does not fit in the above mentioned form of a stiff ODE.

The same order of convergence holds for the unperturbed DAEs since the results remain valid for $\varepsilon = 0$. But these orders are not optimal in all cases for DAEs as we know after reading [Pe3] and [BuPe]. For the Radau-IA-, Radau-IIa-, and Lobatto-IIIC-methods they are indeed optimal, but it is known that the Gauss-methods yield order $s+1$ if s is odd (and order s if s is even). We do not intend to investigate this case here.

2. Comparison of Theorem 3.1 with an earlier result from Griepentrog [Gri]:

He considers

$$(*) \qquad \begin{aligned} y'(t) &= f(t, y(t), z(t), \varepsilon) & y(t) \in \mathbb{R}^m \\ z'(t) &= g(t, y(t), z(t), \varepsilon) + \tfrac{1}{\varepsilon} Mz & z(t) \in \mathbb{R}^k \end{aligned}$$

where the eigenvalues of the constant matrix M have negative real part. His result states uniform (in ε) order of convergence for this system.
In contrast to our regularization the one in $(*)$ is linear in z'. On the other hand the idea behind our approach is an even simpler and also linear regularization ("replace z by $z + \varepsilon M^{-1} z'$") so that in principle we solve a very simple linear but *discrete* differential equation. The continous dependency of the right hand side in $(*)$ presents no essential difference since our results could be easily modified to include this case also.
Griepentrog proves uniform convergence for all L-stable Runge-Kutta methods (L-stable means A-stable and

$$\lim_{z \to -\infty} R(z) = 0$$

holds, where R is the stability function of the method).
But there is no discussion on the rate of convergence. For our regularization we were able to give an error-representation and order conditions. Furthermore our assumptions are less restrictive since there are Runge-Kutta methods that have a coefficient matrix with eigenvalue with positive real part but are not L-stable, e. g. the Gauss-methods. Another example is ([Pe3], [HW]):

Define a semi-explicit method by

$$A := \begin{pmatrix} \frac{3+\sqrt{3}}{6} & 0 \\ \frac{-\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \end{pmatrix}, \qquad b := \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

Obviously the real parts of the eigenvalues of A are positive.
But this method is not L-stable since

$$\lim_{z \to -\infty} R(z) = \frac{1+\sqrt{3}}{2+\sqrt{3}}.$$

So our theorem is applicable whereas Griepentrog's is not. This method satisfies $B(4)$, $C(1)$, $A(3)$ and it can be shown that the corresponding ψ is uniformly bounded on $\mathbb{C}^-$.Therefore Theorem 3.1 shows that the uniform order of convergence in the y-component is 3 and in the z-component 2.

2. Influence of the initial error e_0:
From the error-representation in Theorem 3.1 we see that e_0 is propagated after ν steps with stepsize h as $R(-hM_\varepsilon)^\nu e_0$. Since we assume A-stability this term is always bounded by e_0. For $h > 0$, $\varepsilon > 0$ we have $R(-hM_\varepsilon) < 1$ so that the influence of e_0 decreases with an increasing number of steps. Keeping in mind that we are interested in an approximation for the $\varepsilon = 0$ case this means that the numerical solution of the perturbed system tends to the true solution of the reduced system within $O(\varepsilon)$ even if we use an incorrect initial value in the z-component. (Note that perturbations in the y-component influence the numerical solution in the same way as they do for ordinary differential equations since in that component we simply solve an ordinary differential equation). However in case $\varepsilon = 0$ and we use an L-stable method (i. e. $R(-\infty) = 0$), then after one step the error e_0 no

longer influences the numerical solution.
If the method satisfies $|R(-\infty)| = 1$ (as is the case for the Gauss-methods) then the influence of e_0 on the numerical solution of the reduced system will not decrease; it will remain in the amount of $\|e_0\|$ and therefore the numerical solution will not come close to the true solution but will remain in a tube with diameter $\|e_0\| + O(\varepsilon)$ around it. But if we compute the numerical solution of the perturbed problem the e_0-term will die away (since $R(-hM_\varepsilon) < 1$ even if the method is not L-stable) and therefore the numerical solution will tend to the true solution of the reduced problem within $O(\varepsilon)$. Consequently in this situation it is of significant advantage to solve the perturbed system rather than the unperturbed one. The effects described can actually be observed in numerical experiments (see [Kn]).

Acknowledgement

I would like to thank L. Petzold for a careful reading of a first version of this paper.

References

[BuHu] K. Burrage, W. H. Hundsdorfer:
The Order of B - Convergence
of Algebraically Stable Runge - Kutta Methods
BIT 27 (1987), p. 62 - 71

[BuHuVe] K. Burrage, W. H. Hundsdorfer, J. G. Verwer:
A Study of B - Convergence of Runge - Kutta Methods
Computing 36 (1986), p. 17 - 34

[BuPe] K. Burrage, L. Petzold:
On Order Reduction for Runge - Kutta Methods applied to
Differential/Algebraic Systems and to Stiff Systems of ODEs
Preprint UCRL - 98046, Lawrence Livermore National Laboratory,
Jan. 1988

[Ca] S. L. Campbell:
Regularizations for Linear Time Varying Singular Systems
Automatica 20 (1984), p. 365-370

[DV] K. Dekker, J. G. Verwer:
Stability of Runge-Kutta methods for stiff nonlinear differential equations
North - Holland 1984

[Gri] E. Griepentrog:
Numerische Integration von steifen Differentialgleichungssystemen
mit Einschrittverfahren
Beiträge zur Numerischen Mathematik, 8 (1980), p. 59 -74

[HaBaLu] E. Hairer, G. Bader, Ch. Lubich:
On the stability of semi-implicit methods for ordinary differential equations
BIT 22 (1982), p. 211-232

[HaLuRo] E. Hairer, Ch. Lubich, M. Roche:
Error of Runge - Kutta methods for stiff problems studied
via differential algebraic equations
To appear in BIT

[Hu] W. H. Hundsdorfer:
The numerical solution of nonlinear stiff initial value problems -
an analysis of one-step methods
CWI Tract 12, Amsterdam 1985

[HW] G. Hall, J. M. Watt:
Modern Numerical Methods for Ordinary Differential Equations
Oxford University Press, 1976

[Kn] M. Knorrenschild:
Regularisierungen von Differentiell - Algebraischen Systemen -
theoretische und numerische Aspekte
doctoral thesis, Aachen University of Technology 1988

[LeLe] J. J. Levin, N. Levinson:
Singular Perturbations of Nonlinear Systems of Differential Equations
and an Associated Boundary Layer Equation
J. Ration. Mech. Anal. (now ”Ind. J. Math.) 3 (1954), p. 247-270

[Lö1] P. Lötstedt:
Discretizations of Singular Perturbation Problems by BDF-Methods
Report 99, Uppsala University (1985)

[Lö2] P. Lötstedt:
On the Relation between Singular Perturbation Problems and
Differential-Algebraic Equations
Report 100, Uppsala University (1985)

[Pe1] L. R. Petzold:
Order Results for Implicit Runge-Kutta Methods applied to
Differential/Algebraic Systems
SIAM J. Num. Anal. 23 (1986), p. 837-852

[Ro] M. Roche:
Runge - Kutta and Rosenbrock methods for differential-algebraic
equations and stiff ODEs
doctoral thesis, Univ. de Genève, 1988

☆U.S. GOVERNMENT PRINTING OFFICE: 1990 - 785-103/34010