Regularisierungen von Differentiell - Algebraischen Systemen – theoretische und numerische Aspekte

Von der Mathematisch-Naturwissenschaftlichen Fakultät
– Fachbereich 1 –

der Rheinisch–Westfälischen Technischen Hochschule Aachen

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften

genehmigte Dissertation

Vorgelegt von

Diplom–Mathematiker
Michael Knorrenschild

aus Duisburg

Referent : Universitätsprofessor Dr. R. Jeltsch
Korreferent : Universitätsprofessor Dr. K. Taubert (Hamburg)

Tag der mündlichen Prüfung: 4. Juli 1988

Ich danke Herrn Prof. Dr. R. Jeltsch für die Anregung zu dieser Arbeit und die ständige Gesprächsbereitschaft. Herrn Prof. Dr. K. Taubert gilt mein Dank für die Übernahme des Korreferats und für viele anregende Diskussionen.
Ferner danke ich meinem Freund und Kollegen Dr. Wolfgang Mackens. In vielen langen Gesprächen mit ihm habe ich über meine mathematische Bildung hinaus profitiert. Frau Gabi Goblet gilt mein Dank für die erstklassige Ausführung der Zeichnungen und der Plots.

Von wesentlicher Bedeutung war die tatkräftige Unterstützung von Frau Gabi Silzer, nicht nur bei Programmierarbeiten und der Erstellung der Plots und Tabellen. Ohne ihren regularisierenden Einfluß auf mich in den vergangenen Jahren hätte diese Arbeit nicht entstehen können.

Inhalt

Einleitung

Differentiell - Algebraische Systeme - im Englischen kurz DAEs (differential – algebraic equations) -, was ist das und wo treten sie überhaupt auf?
Nun, DAEs sind implizite Differentialgleichungen der Form $F(t, x(t), x'(t)) = 0$ und im engeren Sinn solche mit singulärem $F_{x'}$. Darunter fallen auch sogenannte semi - explizite Systeme der Form

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

DAEs sind seit einigen Jahren Gegenstand ausgiebiger Untersuchungen (siehe z. B. [GriMä]). Anwendungen sind zahlreich vorhanden; diese werden in verschiedenen Kapiteln angerissen. Warum haben nun aber DAEs erst in jüngster Zeit die Aufmerksamkeit der Mathematiker auf sich gelenkt?
Der Grund ist darin zu sehen, daß die Anwender, die auf DAEs stießen, diese mit allerlei Transformationen auf Systeme von gewöhnlichen Differentialgleichungen zurückführten, was i. a. prinzipiell möglich ist. So traten nach außen hin gar keine DAEs auf.
Transformationen haben jedoch allseits bekannte Nachteile; sie verändern die Struktur des Systems, verwandeln dünn besetzte Matrizen in voll besetzte und schleppen gelegentlich irgendwelche Eigenschaften ein, die physikalisch keinen Sinn geben (z. B. Instabilitäten). Und so gibt es ein berechtigtes Interesse, DAEs dem Rechner in ihrer ursprünglich gegebenen Form vorzulegen.
Zur numerischen Lösung wurden in letzter Zeit diverse Verfahren vorgeschlagen, die fast ausschließlich Varianten von Verfahren für gewöhnliche Differentialgleichungen sind. Dabei zeigt sich, daß gerade die Verfahren für steife Differentialgleichungen besonders geeignet sind. Dies erscheint nicht weiter verwunderlich, da DAEs als Grenzfall von steifen oder singulär gestörten Differentialgleichungen angesehen werden können:

$$(2) \qquad \begin{aligned} y'(t) &= f(t, y(t), z(t), \varepsilon) \qquad & f : \mathbb{R}^{m+k+2} \to \mathbb{R}^m \\ \varepsilon z'(t) &= g(t, y(t), z(t), \varepsilon) \qquad & g : \mathbb{R}^{m+k+2} \to \mathbb{R}^k, \end{aligned}$$

und nun $\varepsilon \to 0$. Rein formal kann ein solches System natürlich schnell hingeschrieben werden. Sollen die Lösungen des gestörten Systems (2) aber auch gegen die des sogenannten reduzierten Problems (1) konvergieren (in welchem Sinne auch immer), so müssen gewisse Stabilitätseigenschaften gegeben sein. Wenn (1) aus einem System (2) entstanden ist (z. B. durch Vernachlässigung gewisser geringer Einflüsse), so darf angenommen werden, daß Stabilität gegeben ist. Andernfalls wäre (1) nicht physikalisch sinnvoll. Die Entscheidung, ob und welche Größen in einem Modell der Form (2) vernachlässigt werden dürfen, bleibt aber dem Anwender überlassen, der das mathematische Modell für sein Problem bereitstellt.

Der Mathematiker findet daher gelegentlich DAEs vor, die "nicht korrekt gestellt" sind; z. B. hängt die Lösung nicht stetig vom Anfangswert ab. Er erwartet daher Probleme bei der numerischen Berechnung (und wird häufig in dieser Erwartung nicht enttäuscht).

Abhilfe schafft hier eine genauere Betrachtung der Modellentstehung. Es zeigt sich nämlich oft, daß man in solchen Fällen bei der Modellierung über das Ziel hinausgeschossen ist und zu viele kleine Einflüsse vernachlässigt hat, und daß nicht immer "gering" gleichbedeutend mit "unwichtig" ist. Dies trifft auch in Situationen zu, bei denen man sich der Herkunft als reduziertes Problem einer singulär gestörten Differentialgleichung nicht mehr bewußt ist, z. B. bei der Analyse elektrischer Netzwerke. Fügt man nun diese "parasitären" Größen wieder ein, paßt das Modell also besser der Wirklichkeit an, so kann man ein "korrekt gestelltes" Problem erhalten. Dieses ist zumindest in der Theorie leichter numerisch zu lösen. Sicher handelt man sich neue Unannehmlichkeiten ein wie sie beim Rechnen mit kleinen Größen auf dem Computer üblicherweise auftreten. Aber wir hoffen, daß die neuen Schwierigkeiten geringer als die alten sind, und außerdem ist der Stand der Dinge bei der numerischen Lösung von DAEs so, daß man keine Möglichkeit außer acht lassen sollte.

Diese Arbeit stellt Methoden vor, die eventuell zu Unrecht vernachlässigten kleinen Parameter wieder einzufügen, und analysiert theoretische und praktische Auswirkungen. Die prinzipiellen Ideen werden dabei in Kapitel 1 entwickelt. Es zeigt sich dabei, daß dazu keinerlei Kenntnis der Herkunft der DAE nötig ist. Man kann diese Einfügung ganz formal durchführen und erhält unter gewissen Voraussetzungen stabile Systeme. Die vorgestellten Methoden sind durch die Modellierung elektrischer Netzwerke motiviert, wie ausführlich dargelegt wird.
In den Kapiteln 2 und 3 werden diese Techniken auf kompliziertere DAEs übertragen; nämlich solche, die üblicherweise nicht "korrekt gestellt" sind.
In Kapitel 4 wird dann das Verhalten von Runge - Kutta - Verfahren auf diesen gestörten DAEs analysiert. Als Resultat erhalten wir für große Klassen von Verfahren gewisse gleichmäßige Abschätzungen (für die "einfacheren" DAEs) für den lokalen und den globalen Fehler der mit den Verfahren berechneten Näherungen.
Diese Ergebnisse werden in Kapitel 5 mit numerischen Beispielen konfrontiert. Diese demonstrieren die Ergebnisse von Kapitel 4 in der Praxis. Außerdem zeigt die Anwendung der Techniken aus Kapitel 2 bei Programmen mit Schrittweitensteuerung vielversprechende Resultate.

Was wichtig ist, sieht man nicht ... .
– A. de Saint - Exupéry, Der kleine Prinz (1946)

0. Grundlegendes

Wir stellen hier die grundlegenden Begriffe bereit und beschreiben verschiedene Eigenschaften der zu untersuchenden Gleichungen.

0.1 Definition (*Index-0 - Differentialgleichung, konsistente Anfangswerte*)
Sei $F : \mathbb{R}^{2n+1} \to \mathbb{R}^n$.
Ein Anfangswert $x_a \in \mathbb{R}^n$ heißt *konsistent für* $F(t, x(t), x'(t)) = 0 :\iff$
es gibt ein $b > a$ und ein $x \in C^1([a, b])$ mit $x(a) = x_a$ und $F(t, x(t), x'(t)) = 0$ für alle $t \in [a, b]$.
Die Differentialgleichung $F(t, x(t), x'(t)) = 0$ hat den *Index 0* $:\iff$
die Matrix $F_{x'}(y)$ ist für alle $y \in \mathbb{R}^{2n+1}$ regulär.

Bemerkung:
Genauso gut kann man auch alles lokal definieren: "lokal Index-0", falls $F_{x'}$ lokal regulär. Im Index-0 - Fall kann lokal nach $x'(t)$ aufgelöst werden (Satz über implizite Funktionen), so daß im wesentlichen eine ungewöhnlich geschriebene gewöhnliche Differentialgleichung vorliegt. Beachte, daß insbesondere gewöhnliche Differentialgleichungen Index-0 haben.

Ansonsten werden wir uns aber im wesentlichen auf die Betrachtung von impliziten Differentialgleichungen in sogenannter semi - expliziter und auch kanonischer Form (s. u.) beschränken.
Außerdem werden wir der Einfachheit halber o. B. d. A. nur den Fall $t \in [0, 1]$ betrachten und uns i. a. auf autonome Systeme beschränken. Nur in wenigen Fällen ist eine besondere Unterscheidung zwischen den Argumenten (t, x) und x sinnvoll.

0.2 Definition (*DAEs in semi - expliziter und kanonischer Form*)

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

mit $y(t) \in \mathbb{R}^m,\ z(t) \in \mathbb{R}^k,\ t \in [0, 1]$ heißt *DAE in semi - expliziter Form* .
Hinweis: DAE = Differential - Algebraic Equation; Differentialgleichung mit angekoppelter algebraischer Nebenbedingung. Im weiteren werden wir auch die Differentialgleichung $F(t, x, x') = 0$ aus 0.1 als DAE bezeichnen.

Obige DAE hat *kanonische Form* : $\Longleftrightarrow$
es gibt $k_1 \in \{0, \ldots, k\}$ und $g_1 : \mathbb{R}^{m+k_1+1} \to \mathbb{R}^{k_1}$ und $g_2 : \mathbb{R}^{m+1} \to \mathbb{R}^{k-k_1}$, so daß

$$g(t, y, z) = \begin{pmatrix} g_1(t, y, z_1) \\ g_2(t, y) \end{pmatrix}, \qquad \text{wobei } z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},\ z_1 \in \mathbb{R}^{k_1}.$$

Für die Übertragung der für semi - explizite Systeme erzielten Resultate auf allgemeinere DAEs ist es nötig zu wissen, unter welchen Bedingungen und wie diese Systeme ineinander überführt werden können.

0.3 Lemma: (*Transformation auf semi - explizite Form*)

Vor.:
$B : \mathbb{R}^{n+1} \to \mathbb{R}^{n\times n}$, $F : \mathbb{R}^{n+1} \to \mathbb{R}^n$; $C : \mathbb{R} \to \mathbb{R}^{n\times n}$, $A : \mathbb{R}^{n+1} \to \mathbb{R}^{n\times n}$ *reguläre Matrizen; C sei differenzierbar;* $u(t) := C(t)^{-1}x(t),\ t \in \mathbb{R}$
für ein $r \in \mathbb{N}$ und alle $x \in \mathbb{R}$ sei

$$A(t, x)B(t, x)C(t) = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix},$$

wobei I_r die $r \times r$ Einheitsmatrix ist.

Beh.:
Das System $B(t, x)x' = F(t, x)$ läßt sich äquivalent als semi - explizite DAE in $u(t)$ formulieren.

Beweis:
Wir haben

$$\begin{aligned} \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} u'(t) &= A(t, x(t))B(t, x(t))C(t)u'(t) \\ &= A(t, x(t))B(t, x(t))((C(t)u(t))' - C'(t)u(t)) \\ &= A(t, x(t))B(t, x(t))(x'(t) - C'(t)u(t)) \\ &= A(t, x(t))F(t, x(t)) - A(t, x(t))B(t, x(t))C'(t)u(t) \\ &= A(t, C(t)u(t))F(t, C(t)u(t)) - \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} C(t)^{-1}C'(t)u(t) \end{aligned}$$

und das ist offensichtlich ein System in semi - expliziter Form. ■

<u>Bemerkung:</u>
Wichtig ist die Beobachtung, daß eine DAE nur auf semi - explizite Form transformiert werden kann, wenn $rg\, B(x) = const$ gilt ($rg\, B$ ist der Rang von B), denn sonst kann es $A(t,x)$, $C(t)$ mit den verlangten Eigenschaften gar nicht geben.
Grundsätzlich kann man auch allgemeinere Matrizen $C(t,x)$ zulassen. Nur sind dann keine Kriterien für die Existenz derartiger Transformationen bekannt.
Hat man jedoch lineare Abbildungen $A(t,x), C(t)$ mit

$$A(t,x)(R(B(t,x))) = \mathbb{R}^r \times \{\mathcal{O}_{n-r}\} \text{ und } A(t,x)(R(B(t,x))^\perp) = \{\mathcal{O}_r\} \times \mathbb{R}^{n-r}$$

und

$$C(t)(\{\mathcal{O}_r\} \times \mathbb{R}^{n-r}) = N(B(t,x)) \text{ und } C(t)(\mathbb{R}^r \times \{\mathcal{O}_{n-r}\}) = N(B(t,x))^\perp,$$

so erfüllen diese die Anforderungen aus 0.3, falls $rg\, B(t,x) = r$ ist. Hierbei sind $\mathcal{O}_r$ und $\mathcal{O}_{n-r}$ die Nullvektoren der angegebenen Dimension und $R(.)$ und $N(.)$ bezeichnen Bildraum bzw. Kern einer Matrix.
Solche A, C existieren offensichtlich z. B., falls $R(B(t,x)) = const.$ und $N(B(t,x)) = const.$ sind.
Falls $\frac{\partial}{\partial x}B(t,x) = 0$ (d. h. $B(t,x) = B(t)$), so reicht die Voraussetzung $rg\, B(t,x) = const. = r$ aus um die Existenz der Matrizen $A(t)$ und $C(t)$ mit den verlangten Eigenschaften zu garantieren. Diese sind dann genauso glatt wie $B(t,x)$ [Do].

Zu einem besseren Verständnis von DAEs ist es nützlich Zusammenhänge mit gewöhnlichen Differentialgleichungen herzustellen. Man könnte z. B. versuchen, durch Differentiation der Nebenbedingung $g(t,y,z) = 0$ die vermissten Ableitungen z' wieder in das System einzubringen. Dazu reicht aber unter Umständen einmaliges Differenzieren nicht aus. Die Klasse der DAEs in kanonischer Form, für die dieser Weg aber zu einer gewöhnlichen Differentialgleichung führt, wird im folgenden definiert.

0.4 <u>Definition</u> (*DAEs mit Index-k*)
Gegeben sei eine DAE in kanonischer Form gemäß Def. 0.2:

$$y' = f(y,z) \tag{0.4.1a}$$
$$0 = g_1(y, z_1) \tag{0.4.1b}$$
$$0 = g_2(y). \tag{0.4.1c}$$

Dann hat (0.4.1) den Index-k : $\Longleftrightarrow$
$g_{1z_1}(y, z_1)$ ist regulär auf $\mathbb{R}^{m+k_1}$, die DAE

$$
\begin{aligned}
y' &= f(y,z) \\
z_1' &= -(g_{1z_1}(y,z_1))^{-1} g_{1y}(y,z_1) f(y,z) \\
0 &= g_2'(y) f(y,z)
\end{aligned}
\tag{0.4.2}
$$

hat kanonische Form und Index- $(k - -1)$.

0.5 Bemerkungen:
1. Wieder steht einer Definition der Eigenschaft "lokaler Index-k" nichts im Wege. Ebenso macht eine Erweiterung auf nicht autonome Probleme keine Schwierigkeiten.
2. Beachte, daß (0.4.2) durch Differentiation von (0.4.1b) und (0.4.1c) aus (0.4.1) hervorgeht. Damit ist sofort klar, daß - wenn man einen Anfangswert $(y(0), z(0))^T$ vorgibt - gilt:

$$
\begin{aligned}
&(y(t), z(t))^T \text{ erfüllt } (0.4.1) \iff \\
&\qquad 0 = g_1(y(0), z_1(0)) = g_2(y(0)) \text{ und } (y(t), z(t))^T \text{ erfüllt } (0.4.2).
\end{aligned}
$$

Eine DAE vom Index-0 (siehe 0.1) kann i. a. recht leicht durch scharfes Hinsehen als solche erkannt werden; bei Index-k ($k \geq 1$) DAEs wird dies nur in Ausnahmefällen gelingen. Wir können aber 0.4 benutzen, um einen Algorithmus für die Bestimmung des Index anzugeben.

0.6 Algorithmus (*Index - Reduktionsalgorithmus, IRA*) [GePe]
Gegeben sei eine Index-k - DAE ($k \in \mathbb{N}$) gemäß 0.3.

(A) Falls es sich um eine gewöhnliche Differentialgleichung handelt, liegt Index-0 vor; STOP. Andernfalls gehe zu (B).
(B) Reduziere den Index um 1 gemäß Definition 0.4 und gehe mit der neuen DAE zu (A).

0.7 Folgerung:
Für eine DAE in kanonischer Form gilt:
Die DAE hat Index-k $\iff$ der IRA angewandt auf diese DAE stoppt nach k Schritten.

Bemerkung:
In [GePe] wurde Algorithmus 0.6 für lineare DAEs $A(t)x' + B(t)x = f(t)$ formuliert und ebenso die Anwendbarkeit für nichtlineare DAEs erwähnt.
In jedem Reduktionsschritt mit dem IRA wird also der y - Vektor um einige (oder auch gar keine) Komponenten des z - Vektors ergänzt:

$$y_{neu} := \begin{pmatrix} y_{alt} \\ z_1 \end{pmatrix}.$$

Kommen bei der Behandlung einer Index-k - DAE mittels IRA im i-ten Schritt (also $i \leq k$) die Variablen $z_{i_1}, \ldots, z_{i_j}$ zum y - Vektor hinzu, so bezeichen wir diese als

Index-i - Variablen

der ursprünglichen Index-k-DAE. Insbesondere sind also die Komponenten des y - Vektors der Ausgangs- DAE stets Index-0-Variablen.

Kann man bei einer semi - expliziten DAE die Index-i - Variablen nicht als solche identifizieren, da keine kanonische Form vorliegt, so kann man immer noch versuchen durch Linksmultiplikation mit regulären Matrizen und/oder Koordinatentransformationen (vgl. Lemma 0.3) das System auf kanonische Form zu bringen. (Dies wird z. B. später auch in Satz 2.8 geschehen.) In vielen Fällen sind jedoch die Index-i - Variablen leicht zu erkennen - auch ohne Koordinatentransformation. Dazu folgende Beispiele:

0.8 Beispiel: (*Netzwerk - DAE*)
Gegeben ist das folgende elektrische Netzwerk mit einer Stromquelle $I(t)$, einer Kapazität C, zwei Induktivitäten L_1 und L_2 und einem Widerstand R:

Das ganze ist dem Buch von Adby [Ad] entnommen (Example 7.2). (Weitere Beobachtungen bei DAEs aus Netzwerken folgen in Kap. 1.)
Bezeichnet man wie angegeben die Spannungen und Ströme in den einzelnen Zweigen mit v_j bzw. i_j, so ergeben sich zur Beschreibung des Netzwerks folgende Gleichungen (durch Verwendung der Kirchhoffschen Regeln):

$$\begin{array}{lcc} L_1 i_1' - L_2 i_2' - R i_5 & = 0 & \text{obere Masche} \\ L_1 i_1' + v_4 - v_3 & = 0 & \text{untere Masche} \\ C v_3' + i_1 + i_2 & = 0 & \text{Knoten 2} \\ i_1 + i_5 - I & = 0 & \text{Knoten 1} \\ i_2 - i_5 & = 0 & \text{Knoten 3.} \end{array}$$

Bezeichnet man $y := (i_1, i_2, v_3)^T$, $z := (i_5, v_4)^T$, so erhält man eine semi - explizite DAE, nämlich

$$\begin{aligned} \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} &= \begin{pmatrix} L_1 & -L_2 \\ L_1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} R z_1 \\ y_3 - z_2 \end{pmatrix} \\ y_3' &= C^{-1}(-y_1 - y_2) \\ 0 &= y_1 + z_1 - I \\ 0 &= y_2 - z_1. \end{aligned} \tag{0.8.1}$$

Diese ist aber nicht in kanonischer Form, denn

$$\begin{pmatrix} y_1 + z_1 - I \\ y_2 - z_1 \end{pmatrix}_z = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}$$

und das ist nicht regulär. Die letzte Gleichung in (0.8.1) kann aber auch durch $y_2 - I + y_1 = 0$ ersetzt werden (unter Benutzung der vorletzten Gleichung in (0.8.1)). Damit erhalten wir

$$\begin{aligned} \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} &= \begin{pmatrix} L_1 & -L_2 \\ L_1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} Rz_1 \\ y_3 - z_2 \end{pmatrix} \\ y_3' &= C^{-1}(-y_1 - y_2) \\ 0 &= y_1 + z_1 - I \\ 0 &= y_2 + y_1 - I \end{aligned} \tag{0.8.2}$$

und diese ist in kanonischer Form ($g_1(y, z_1) := y_1 + z_1 - I,\ g_2(y) := y_1 + y_2 - I$). Man sieht sofort, daß z_1 Index-1 - Variable ist. Weiter rechnet man leicht nach, daß z_2 Index-2 - Variable ist und damit (0.8.2) als Index-2 - DAE erkannt ist.

0.9 Beispiel: (*Pendel - DAE*)
Wir betrachten die folgende DAE, die die Bewegung eines Pendels beschreibt (siehe auch Kap. 3 und [Rh], [LöPe]):

$$\begin{aligned} y' &= f(y,z) := \begin{pmatrix} y_3 \\ y_4 \\ -y_1 z \\ -y_2 z + 1 \end{pmatrix} \\ 0 &= y_1^2 + y_2^2 - 1 \end{aligned} \tag{0.9.1}$$

Hierbei ist $y = (y_1, y_2, y_3, y_4)^T$.
Diese DAE hat Index-3, wie die Anwendung des IRA zeigt:
der erste Schritt mit dem IRA liefert:

$$\begin{array}{rcl} y' &=& f(y,z) \\ 0 &=& 2y_1y_1' + 2y_2y_2' \end{array} \iff \begin{array}{rcl} y' &=& f(y,z) \\ 0 &=& y_1y_3 + y_2y_4 \end{array} \tag{0.9.2}$$

Offensichtlich ist noch mindestens ein weiterer Schritt mit dem IRA erforderlich. Dieser liefert:

$$\begin{array}{rcl} y' &=& f(y,z) \\ 0 &=& y_1'y_3 + y_1y_3' + y_2'y_4 + y_2y_4' \end{array} \iff \begin{array}{rcl} y' &=& f(y,z) \\ 0 &=& y_3^2 - y_1^2 z + y_4^2 - y_2^2 z + y_2 \end{array} \tag{0.9.3}$$

Ein erneuter Schritt bringt

$$\begin{array}{rcl} y' &=& f(y,z) \\ 0 &=& 2y_3y_3' - 2y_1y_1'z - y_1^2 z' + 2y_4y_4' - 2y_2y_2'z - y_2^2 z' + y_2' \end{array}$$

$$\iff$$

$$\begin{array}{rcl} y' &=& f(y,z) \\ 0 &=& -2y_3y_1z - 2y_1y_3z - y_1^2 z' - 2y_2y_4z + 2y_4 - 2y_2y_4z - y_2^2 z' + y_4 \end{array}$$

Da aber auf der exakten Lösung von (0.9.1) $y_1^2 + y_2^2 = 1$ gilt, also in einer Umgebung dieser Lösung $y_1^2 + y_2^2 \neq 0$ gilt, ist die letzte DAE dort äquivalent zu

$$
\begin{aligned}
y' &= f(y,z) \\
z' &= (y_1^2 + y_2^2)^{-1}(-4y_3y_1z - 4y_2y_4z + 3y_4)
\end{aligned} \tag{0.9.4}
$$

Dies ist eine gewöhnliche Differentialgleichung, also eine Index-0 - DAE. Da drei Schritte mit dem IRA erforderlich waren, handelt es sich bei der Ausgangsgleichung um eine Index-3 - DAE. Außerdem ist z Index-3 - Variable. Wir haben dann insgesamt wegen Folgerung 0.7:

Das System (0.9.i) hat Index-$(4-i)$ für $i = 1, \ldots, 4$.

Index-0 - Differentialgleichungen bereiten numerisch keine besonderen Probleme; auch für Index-1 - Systeme gibt es inzwischen spezielle Programmpakete (DASSL [Pe2], LIMEX [DeuHaZu]). Probleme treten jedoch bei DAEs mit höherem Index auf. Siehe z. B. [Pe1]. Dies ist auch kein Wunder, denn wir wissen ja, daß die Berechnung der Lösung einer Index-k - DAE eine Berechnung von (k-1)-ten Ableitungen erfordert, was bekanntlich fehleranfällig ist. Hinzu kommen weitere Unannehmlichkeiten wie die unvollständige Kenntnis von Anfangswerten, keine stetige Abhängigkeit von diesen und vieles mehr. Am besten wäre es offensichtlich, den Index einer DAE zu reduzieren (optimal wäre Index-0). Eine praktische Verwendung des IRA ist wegen der auftretenden analytischen Ableitungen und der möglichen Zerstörung erhaltungswürdiger Strukturen (wie z. B. das Auftreten dünn besetzter Matrizen) aber nicht sinnvoll und daher kein Weg, diese Probleme zu umgehen. Zudem sehen wir aus der Bemerkung 0.5.2, daß die Lösungsmenge einer DAE zu einem Anfangswert bei den einzelnen Transformationen des IRA nur dann nicht geändert wird, wenn dieser Anfangswert gewisse Bedingungen erfüllt. Diese Bedingungen stellen sich aber erst im Laufe der Durchführung des IRA heraus, so daß sich als zusätzliches Problem das Finden eines geeigneten, d. h. konsistenten Anfangswertes stellt. Später werden wir dann sehen, welche Bedingungen ein Anfangswert erfüllen muß, um konsistent zu sein.
Wir suchen daher andere Möglichkeiten um den Index zu reduzieren, insbesondere ableitungsfreie Methoden.
Dies geschieht durch geschicktes Einfügen kleiner Parameter, die in praktischen Beispielen oftmals eine physikalische Bedeutung haben. Die dadurch bewirkte Reduktion des Index hat auch noch den Vorteil, daß sich die Anforderungen an einen konsistenten Anfangswert abschwächen.
Von einer "vernünftigen" Störung der DAE erwarten wir nun das wie folgt beschriebene Verhalten.

0.10 Definition: (ε - *stabil*)
Sei $F : \mathbb{R}^{2n+2} \to \mathbb{R}^n$,

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

und $x_0(t),\ t \in [a, b]$ sei Lösung von $(*)$ für $\varepsilon = 0$; sei $x_a := x_0(a)$, U Unterraum von $\mathbb{R}^n$.

Dann heißt $(*)$ ε - *stabil* auf $[a, b]$ $:\Longleftrightarrow$

Es gibt $\varepsilon_0 > 0$ mit

$(*)$ besitzt für $0 < \varepsilon \leq \varepsilon_0$ zu jedem konsistenten Anfangswert $\tilde{x}_{a\varepsilon} \in \mathbb{R}^n$ mit $\|\tilde{x}_{a\varepsilon} - x_a\| < \varepsilon_0$ eine Lösung $x_\varepsilon(t)$ mit $x_\varepsilon(a) = \tilde{x}_{a\varepsilon}$ und

$$\lim x_\varepsilon(t) = x_0(t)$$

gleichmäßig auf $[a, b]$ für

$$\varepsilon + \|\tilde{x}_{a\varepsilon} - x_a\| \to 0.$$

$(*)$ heißt ε - *stabil* auf $[a, b]$ bez. U $:\Longleftrightarrow$

$(*)$ ist ε - stabil und

es gibt $\delta > 0, \varepsilon_0 > 0$ mit

$(*)$ besitzt für $0 < \varepsilon \leq \varepsilon_0$ zu jedem $u_a \in U$ mit $\|u_a\| < \delta$, so daß $x_a + u_a$ konsistent für $(*)$ ist, eine Lösung $x_\varepsilon(t)$ mit $x_\varepsilon(a) = x_a + u_a$ und

$$\lim_{\varepsilon \to 0+} x_\varepsilon(t) = x_0(t), \quad \text{gleichmäßig auf } [c, d] \text{ mit } a < c < d \leq b.$$

U ist also der Unterraum, in dessen Richtung ein Anfangswert geringfügig gestört werden darf, ohne daß die gleichmäßige Konvergenz auf kompakten Teilintervallen von $]a, b]$ verloren geht.

Im Prinzip werden wir als hinreichendes Kriterium für diese Eigenschaft stets folgenden Satz benutzen.

0.11 Satz : [LeLe]

Vor.:

die Gleichung

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

habe für $\varepsilon = 0$ *zum Anfangswert* y_a, z_a *eine Lösung* $(y_0(t), z_0(t))$ *auf* $[a, b]$ *mit* $y_0(a) = y_a, z_0(a) = z_a$;
hierbei ist für ein $\delta_0 > 0$ *und*

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

f, f_y, f_z, g, g_y, g_z *stetig auf* R , *für alle* $t \in [a, b]$ *gelte:*

$$\lambda \text{ Eigenwert von } g_z(t, y_0(t), z_0(t), 0) \Rightarrow Re(\lambda) < 0$$

Beh.: $(*)$ *ist* ε *- stabil auf* $[a,b]$ *bez.* $U := \{\mathcal{O}_m\} \times \mathbb{R}^{n-m}$.

Bemerkung dazu:
Hier darf also der Anfangswert z_a geringfügig gestört werden, ohne daß die gleichmäßige Konvergenz auf kompakten Teilintervallen von $]a,b]$ verloren geht.

1. Reduktion von Index-1 auf Index-0

Als erstes werden wir nun versuchen, die gewissermaßen einfachsten DAEs - die vom Index - 1 - so zu stören, daß der Index auf 0 sinkt und das neue System ε - stabil ist. Zunächst werden wir dieses Vorhaben für semi - explizite DAEs durchführen und danach auch allgemeinere betrachten. Dabei wird sich zeigen, daß Regularisierungen nach der hier vorgestellten Methode im Prinzip "immer" möglich sind. Außerdem stellt sich heraus, daß die vorgestellten Ansätze durchaus einen physikalischen Hintergrund haben.

Vorweg eine elementare Folgerung aus der Definition des Index (0.4).

1.1 Lemma: (*Charakterisierung von Index-1 - DAEs*)

Beh.:
Die DAE

$$
\begin{array}{llll}
(1.1.1a) & y'(t) = f(t, y(t), z(t)) & & f : \mathbb{R}^{m+k+1} \to \mathbb{R}^m \\
(1.1.1b) & 0 = g(t, y(t), z(t)) & & g : \mathbb{R}^{m+k+1} \to \mathbb{R}^k
\end{array}
$$

mit f stetig und g differenzierbar hat Index-1 :
$\Longleftrightarrow$ *die Matrix $g_z(t, y, z)$ ist regulär für alle $(t, y, z) \in \mathbb{R}^{m+k+1}$* .

Bemerkung:
Natürlich kann man auch entsprechende lokale Versionen formulieren für den Fall, daß f, g nicht auf ganz $\mathbb{R}^{m+k+1}$ definiert sind.
Lemma 1.1 charakterisiert offensichtlich alle Index-1 - DAEs. Zu berücksichtigen sind später noch solche DAEs, die auf semi - explizite Index-1 - DAEs transformiert werden können (abwarten bis Satz 1.8).
Im Index-1 - Fall kann man schon Aussagen über Lösbarkeit machen; denn lokal kann man in einer Umgebung eines Punktes (t_0, y_0, z_0) mit $g(t_0, y_0, z_0) = 0$ nach z auflösen (Satz über implizite Funktionen). Man hat dort etwa

$$g(t, y, z) = 0 \iff z = h(t, y).$$

Das hat

$$(1.1.1a),\ (1.1.1.b) \iff y'(t) = f(t, y(t), h(t, y(t))),\ z(t) = h(t, y(t))$$

zur Folge und natürlich hat letztere Gleichung lokal eine Lösung, falls nur f und h stetig sind.
Fazit: semi - explizite Index-1 - DAEs haben zu Anfangswerten, die (1.1.1b) erfüllen, lokal eine Lösung. Und damit ist auch sofort klar, daß ein Anfangswert für (1.1.1) genau dann konsistent ist, wenn er (1.1.1b) erfüllt.

Zuerst präsentieren wir eine ganz einfache Regularisierung, die dann noch verallgemeinert wird.

1.2 Satz:

Vor.:

$$\begin{array}{rcll} y'(t) & = & f(y(t),z(t)) & \qquad f:\mathbb{R}^{m+k}\to\mathbb{R}^m \\ 0 & = & g(y(t),z(t)) & \qquad g:\mathbb{R}^{m+k}\to\mathbb{R}^k \end{array}$$

f stetig, g differenzierbar, $D_1 \subseteq \mathbb{R}^m,\ D_2 \subseteq \mathbb{R}^k,\ D_1,\ D_2$ offen und $\bar{D}_1,\ \bar{D}_2$ kompakt, obige DAE habe Index-1 auf $D_1 \times D_2,\ h: D_1 \to D_2$ bijektiv mit

$$g(y,z)=0,\ y\in D_1,\ z\in D_2 \iff z=h(y),$$

(h ist eine Auflösungsfunktion, die wegen der Index-1 Voraussetzung zumindest lokal existiert) und für ein $L_h > 0$

$$\|h(y)-h(\tilde{y})\| \le L_h\|y-\tilde{y}\| \quad \textit{für alle } y,\tilde{y}\in D_1,$$

$y_0 \in D_1,\ z_0 := h(y_0),\ b>0,\ \bar{y}:[0,b]\to D_1,\ \bar{y}(0)=y_0,\ \bar{z}(t):=h(\bar{y}(t))$ für $t\in[0,b]$ mit

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

Beh.:

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

ist für alle $\varepsilon > 0$ Index-0 - DAE und es gibt ein $C > 0$ so, daß :
für alle $(\tilde{y}_0,\tilde{z}_0)\in D_1\times\mathbb{R}^k$ gibt es ein $\tilde{b}\in]0,b]$ und eine Lösung $(y_\varepsilon(t),z_\varepsilon(t))$ von $(*)$ auf $[0,\tilde{b}]$ mit

$$\|y_\varepsilon(t)-\bar{y}(t)\| \le C\|\tilde{y}_0-y_0\| \qquad \textit{für alle } t\in[0,\tilde{b}].$$

Weiter gibt es zu jedem $a\in]0,\tilde{b}]$ ein $\omega_a:\mathbb{R}\times\mathbb{R}^3_{\ge 0}\to\mathbb{R}_{\ge 0}$ mit

$$\begin{array}{rcllll} \lim\omega_a(t,\varepsilon,d1,d2) & = & 0 & \textit{für } \varepsilon+d1+d2\to 0 & \textit{gleichmäßig für} & t\in[0,\tilde{b}] \\ \lim\omega_a(t,\varepsilon,d1,0) & = & 0 & \textit{für } \varepsilon\to 0 & \textit{gleichmäßig für} & t\in[a,\tilde{b}] \end{array}$$

und

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

für alle $t \in]0, \tilde{b}]$.
Insbesondere ist $(*)$ ε *- stabil auf* $[0, \tilde{b}]$ *bez.* $\{\mathcal{O}_m\} \times \mathbb{R}^{n-m}$.

<u>Beweis:</u>
Die Index-0 - Eigenschaft folgt sofort aus Definition 0.1.
Sei nun $(\tilde{y}_0, \tilde{z}_0) \in D_1 \times \mathbb{R}^k$. Bekanntlich gibt es nun $\tilde{b} \in]0, b]$ und $\tilde{y}(t)$ mit $\tilde{y}(0) = \tilde{y}_0$ und

$$\tilde{y}'(t) = f(\tilde{y}(t), h(\tilde{y}(t))) \qquad \text{für alle } t \in [0, \tilde{b}].$$

Wir definieren nun $y_\varepsilon(t) := \tilde{y}(t)$ auf $[0, \tilde{b}]$ und z_ε als Lösung des Anfangswertproblems

$$z_\varepsilon(t) + \varepsilon z_\varepsilon'(t) = h(\tilde{y}(t)),\ z_\varepsilon(0) = \tilde{z}_0,\ t \in [0, \tilde{b}]. \tag{1.2.1}$$

Dann erfüllen y_ε und z_ε offensichtlich $(*)$. Wir haben nun noch die Fehlerabschätzung zu zeigen.
Die Lösung der Differentialgleichung $y' = f(y, h(y))$ hängt bekanntlich lipschitzstetig vom Anfangswert ab. Es gibt daher ein $C > 0$ mit

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

womit der erste Teil schon bewiesen ist. Daraus folgt sofort

$$\|h(\tilde{y}(t)) - h(\bar{y}(t))\| \leq L_h C \|\tilde{y}_0 - y_0\| \text{ gleichmäßig auf } [0, \tilde{b}]. \tag{1.2.2}$$

Wendet man Satz 0.11 nun auf (1.2.1) an (das geht, da die zu prüfenden Eigenwerte gerade -1 sind), so erhält man, falls $a \in]0, \tilde{b}]$,

$$\|z_\varepsilon(t) - h(\tilde{y}(t))\| = \omega_a(t, \varepsilon, \|\tilde{z}_0 - h(\tilde{y}_0)\|, \|\tilde{y}_0 - y_0\|) \tag{1.2.3}$$

mit einem ω_a, das die in der Behauptung genannten Eigenschaften besitzt. Aus (1.2.2) und (1.2.3) folgt nun sofort die Behauptung. ■

Wir können sofort noch eine allgemeinere Version angeben, die auch für Anwendungen interessant ist (siehe Beispiel 1.7).

1.3 Satz:

Vor.:
Wie in Satz 1.2;
$F \in C^1(\mathbb{R}^k, \mathbb{R}^k)$, *F bijektiv*
für alle $z \in \mathbb{R}^k$ *gelte:* λ *Eigenwert von* $F'(z) \Rightarrow Re(\lambda) > 0$

Beh.:

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

ist für alle $\varepsilon > 0$ *Index-0 - DAE und hierfür gilt die gleiche Behauptung wie in Satz 1.2.*

Beweis:
Daß $(*)$ Index-0 hat, sieht man leicht wie folgt:
Sei $\tilde{g}(y, z, z', \varepsilon) := g(y, F^{-1}(F(z) + \varepsilon z'))$. Dann reicht es wegen Definition 0.1 die Regularität von $\tilde{g}_{z'}$ nachzuweisen. Es ist aber hier

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

und mit $(F^{-1})'(F(z) + \varepsilon z') = F'(F^{-1}(F(z) + \varepsilon z'))$ und der Regularität von F' (nach Vor.) ist alles klar.
Der Rest des Beweises läuft wie in 1.2; nur wird (1.2.1) ersetzt durch

$$\begin{array}{rclrcll} F^{-1}(F(z_\varepsilon(t)) + \varepsilon z_\varepsilon'(t)) &=& h(\tilde{y}(t)), & z_\varepsilon(0) &=& \tilde{z}_0, & t \in [0, \tilde{b}] \\ \Longleftrightarrow \quad F(z_\varepsilon(t)) + \varepsilon z_\varepsilon'(t) &=& F(h(\tilde{y}(t))), & z_\varepsilon(0) &=& \tilde{z}_0, & t \in [0, \tilde{b}] \end{array}$$

Wegen der Voraussetzung an die Eigenwerte von $F'(z)$ bleibt auch hier Satz 0.11 anwendbar. ∎

Nun wollen wir den gleichen Sachverhalt bei Multiparameterstörungen untersuchen. Wir betrachten zunächst den Fall, daß die Störparameter untereinander konstantes Verhältnis haben; sind aber an Kriterien interessiert, die nicht die Kenntnis dieser Verhältnisse erfordern. Dazu:

1.4 Definition: (*strikt diagonaldominant*)
Eine $n \times n$-Matrix (a_{ij}) heißt *strikt diagonaldominant* genau dann, wenn

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

1.5 Satz: (Multiparameterversion I von Satz 1.3)

Vor.:

Wie in Satz 1.3;
für alle $z \in \mathbb{R}^k$ sei die Matrix $F'(z)$ strikt diagonaldominant;
für $i = 1, \ldots, k$ sei $\varepsilon_i := d_i \cdot \varepsilon$, wobei $d_i > 0$; $D := diag(d_1, \ldots, d_k)$;

Beh.:

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

ist für alle $\varepsilon > 0$ Index-0 - DAE und hierfür gilt die gleiche Behauptung wie in Satz 1.2.

Beweis:
Wir setzen

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

und verifizieren die Voraussetzungen von Satz 1.3.
Aus der strikten Diagonaldominanz in Verbindung mit dem Satz von Gerschgorin folgt
für alle $z \in \mathbb{R}^k$ gilt: λ Eigenwert von $F'(z) \Rightarrow Re(\lambda) > 0$.
Das gleiche gilt für die Matrix $D^{-1} \cdot F'(z) = \tilde{F}'(z)$, denn diese ist auch wieder strikt diagonaldominant. Das war's schon; nun kann Satz 1.3 auf $\tilde{F}$ angewandt werden und wir erhalten das gewünschte Resultat. ∎

Und nun ein Satz, der Konvergenz für $\varepsilon_1, \ldots, \varepsilon_k \rightarrow 0$ garantiert, ohne Annahme über konstantes Verhältnis der ε_i.

1.6 Satz: (Multiparameterversion II von Satz 1.3)

Vor.:

Wie in Satz 1.5;
für alle $z \in \mathbb{R}^k$ sei $F(z) = (F_1(z_1), \ldots, F_k(z_k))^T$ und $F_i'(z_i) > 0$ für $i = 1, \ldots, k$ (letzteres folgt auch aus der strikten Diagonaldominanz)

Beh.:

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

mit $\mathcal{E} = diag(\varepsilon_1, \ldots, \varepsilon_k)$, $\quad \varepsilon_i > 0$
ist stets Index-0 - DAE und hierfür gilt die gleiche Behauptung wie in Satz 1.2, wenn man ε durch $\mathcal{E}$ und

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

durch

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

ersetzt.

<u>Beweis:</u>
Wie in 1.2; nur wird $z_{\mathcal{E}}$ nicht durch (1.2.1) definiert, sondern durch

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

definiert.
Nun sind die Gleichungen aber gegenseitig entkoppelt. Wir können daher auf die einzelnen Gleichungen Satz 1.3 anwenden und kommen auf die gleiche Weise zum Ziel wie im Beweis zu Satz 1.3. ∎

<u>Bemerkung:</u>
Die Sätze 1.2 bis 1.6 können mit offensichtlichen geringfügigen Modifikationen auch für nichtautonome DAEs formuliert werden.

Der zunächst vielleicht künstlich erscheinenden Einfügung von Störparametern kommt in vielen Fällen eine physikalische Bedeutung zu. Dies soll im folgenden Beispiel verdeutlicht werden.

1.7 <u>Beispiel</u> zum Bezug zur physikalischen Realität:

DAEs bei der Analyse elektrischer Netzwerke

Gegeben sei ein Netzwerk $\mathcal{N}$, das aus Kondensatoren, (spannungskontrollierten) Widerständen und Stromquellen besteht.
$\mathcal{N}$ sei vollständig in Zweige zerlegt, die jeweils genau eines dieser Bauelemente enthalten. Wir interessieren uns dabei für die Spannungen in den einzelnen Zweigen. Der Vektor $(y(t), z(t))^T$ soll die Spannungen in den Zweigen zum Zeitpunkt t enthalten, und zwar sollen in $y(t)$ die aus Kondensatorzweigen und in $z(t)$ die aus Zweigen mit einem Widerstand stehen. Bei m Kondensatorzweigen und k Zweigen mit einem Widerstand ist dann $(y(t), z(t))^T \in \mathbb{R}^{m+k}$.
Die Anwendung der Knotenregel führt dann zunächst auf eine DAE der Form
$B(t)x'(t) = G(x(t))$ mit $x(t) = (y(t), z(t))^T$. $B(t)$ ist dabei eine $(m+k) \times (m+k)$-Matrix, deren Elemente Kapazitäten sind (die wir als zeitabhängig zulassen wollen). Für Details zur Aufstellung dieser Gleichungen siehe z. B. [DesKuh], [Ad]. Wir wollen diese Gleichung wie in Lemma 0.3 beschrieben auf semi - explizite Form bringen. Wir haben in diesem Fall den Vorteil, daß wir genau wissen, welche Komponenten von $x(t)$ differenziert in

der DAE vorkommen, nämlich aufgrund der physikalischen Gesetze (die besagen, daß in Zweigen mit Kondensatoren Differentialgleichungen den Spannung - Strom - Zusammenhang beschreiben, während dies in Zweigen nur mit einem Widerstand rein algebraische Gleichungen sind) sind dies genau die Komponenten von $y(t)$. Dies erspart die Anwendung einer Koordinatentransformation mittels der Matrix $C(t)$ (vergl. Lemma 0.3). Hier reicht also die Linksmultiplikation mit einer Matrix $A(t)$ aus, um auf die gewünschte semi - explizite Form zu kommen. Wenn wir zusätzlich annehmen, daß $rg\,B(t) = m$ ist, gelangen wir schließlich zu (beachte die Bemerkung zu Lemma 0.3)

$$\begin{array}{lrcll} (1.7.1) \qquad & y'(t) & = & f(y(t), F(z(t))) & \qquad y(t) \in \mathbb{R}^m \\ & 0 & = & g(y(t), F(z(t))) & \qquad z(t) \in \mathbb{R}^k. \end{array}$$

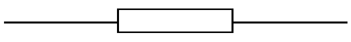
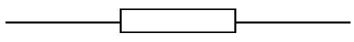
Hierbei ist

$$F(z(t)) := (l_1(z_1(t)), ..., l_k(z_k(t)))^T,$$

wobei die Kennlinien der Widerstände durch die Funktionen l_i gegeben sind (z_i sind die Komponenten von z).
Nun ist es so, daß die Kirchhoffschen Gesetze einen **Idealzustand** beschreiben, der in der Realität nie so gegeben ist. Vielmehr werden dabei z. B. gewisse physikalische Eigenschaften der im Netzwerk verwendeten Materialien vernachlässigt.
Wir konzentrieren unsere Aufmerksamkeit hier auf folgendes Phänomen:

Jede Potentialdifferenz (z. B. zwischen benachbarten Leitern) erzeugt ein schwaches elektrisches Feld, verhält sich also, als ob eine Kapazität $C = \varepsilon$ (Größenordnung ca. 10^{-12} Farad laut [DesKuh], S.60) parallel geschaltet ist.

Besonders bei geringen Abmessungen eines Netzwerks (integrierte Schaltkreise!!) wird dieser Effekt eine Rolle spielen.
Wir sehen also, daß ein genaueres Modell für einen Widerstand ein Widerstand parallel geschaltet mit einer sehr kleinen Kapazität wäre. Das legt folgendes Vorgehen nahe:
Das durch (1.7.1) beschriebene Netzwerk $\mathcal{N}$ wird durch Einfügen einiger parasitärer Kapazitäten zu einem erweiterten Netzwerk $\tilde{\mathcal{N}}$, das dann (hoffentlich!) durch eine Index-0 - DAE beschrieben wird. Diese Erweiterung geschieht wie folgt:

Ersetze  durch 

Dabei werden **alle** Widerstände ersetzt. Die Netzwerkgleichung (1.7.1) muß aber nun modifiziert werden, und zwar so:
Sei der Widerstand kontrolliert durch die Spannung z_i, $i \in \{1, \ldots, k\}$ und die Kennlinie gegeben durch den Graphen der Funktion l_i. Dann muß in (1.7.1) $l_i(z_i(t))$ durch $l_i(z_i(t)) + \varepsilon_i z_i'(t)$ ersetzt werden. Insgesamt führt das auf die Ersetzung von $F(z)$ in (1.7.1) durch $F(z) + \mathcal{E}z'$ mit $\mathcal{E} = diag(\varepsilon_1, \ldots, \varepsilon_k)$. Man erhält also:

$$\begin{array}{lrl} (1.7.2) \qquad & y'(t) & = f(y(t), F(z(t)) + \mathcal{E}z'(t)) \\ & 0 & = g(y(t), F(z(t)) + \mathcal{E}z'(t)) \end{array}$$

Das ist aber gerade nichts anderes als das in 1.6 untersuchte regularisierte System, sofern man streng monoton steigende Kennlinien voraussetzt.

Bemerkungen:
1. Mit der hier an einer Klasse von Netzwerken vorgestellten Regularisierung lassen sich in analoger Weise auch Netzwerke, die nur Induktivitäten, Spannungsquellen, und (stromkontrollierte) Widerstände enthalten, behandeln.
2. Die Methode der Einfügung parasitärer Elemente wird bereits in [ChRo] verwendet, um Netzwerkgleichungen in Normalform (d. h. als gewöhnliche Differentialgleichung) zu formulieren (ohne Konvergenzanalyse). Eine mathematische Analyse der Netzwerkgleichungen nach Einfügen von parasitären Elementen (Konvergenz!) findet man für den Index-1 - Fall und unter der Annahme $\varepsilon_i/\varepsilon_j = const.$ schon in [ChAl].

Bis hierhin haben wir nur semi - explizite DAEs untersucht. Naheliegend ist aber nun die Frage, wie sich die Regularisierungsansätze auf allgemeinere sogenannte voll - implizite DAEs übertragen lassen. Diese allgemeineren DAEs werden natürlich solche DAEs sein, die auf semi - explizite Form gebracht werden könnten. Die Herleitung entsprechender Resultate geschieht derart, daß zuerst auf semi - explizite Form transformiert, dann wie gehabt regularisiert und dann wieder zurücktransformiert wird. Diese naheliegende Methode wird auch schon bei Ascher [As] benutzt, um Resultate für Randwertprobleme von gewöhnlichen Differentialgleichungen auf DAEs zu übertragen.

1.8 Satz

Vor.:
A, B, C, F wie in Lemma 0.3; $r := rg\, B(t,x)$, $t \in [a,b]$,
$B(t,x)x' = F(t,x)$ sei gemäß Lemma 0.3 mittels $A(t)$ und $C(t)$ auf eine (semi - explizite) Index-1 - DAE transformierbar
sei zur Abkürzung

$$h(t,x(t)) := x(t) + \varepsilon C(t) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} (C(t)^{-1}x(t))', \qquad N(t) := kern\, B(t,x)$$

Beh.:

$$(*) \qquad B(t, h(t,x(t)))\ (h(t,x(t)))' = F(t, h(t,x(t)))$$

hat für alle $\varepsilon > 0$ Index-0 und ist ε-stabil auf $[0,1]$ bez. $U := N(a)$.

Beweis:
Nach dem in Lemma 0.3 gesagten ist klar, daß das Ausgangsproblem $B(t,x)x' = F(t,x)$ als semi - explizites System in der Variablen $u(t) := C(t)^{-1}x(t)$ geschrieben werden kann.

Die für semi - explizite DAEs schon bekannte Störungstechnik sieht nun die Ersetzung von $u(t)$ durch

$$u(t) + \varepsilon \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} u'(t)$$

vor. Das bedeutet, daß $x(t) = C(t)u(t)$ ersetzt wird durch

$$\begin{aligned} C(t)u(t) + \varepsilon C(t) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} u'(t) &= \\ = x(t) + \varepsilon C(t) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} (C(t)^{-1}x(t))' &= h(t, x(t)). \end{aligned}$$

Dies ist also der Term, der anstelle von $x(t)$ in die Ausgangsgleichung eingesetzt wird. Tun wir das hier, so erhalten wir $(*)$.
Beim Anfangswert u_a sind kleine Störungen aus dem Raum $\{\mathcal{O}_r\} \times \mathbb{R}^{n-r}$ zulässig (vgl. Satz 0.10), d. h. für den gestörten Anfangswert $\tilde{u}_a$ gilt

$$\tilde{u}_a - u_a = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} v \qquad \text{für ein } v \in \mathbb{R}^n.$$

Für den Anfangswert $x_a = C(a)u_a$ und seine Störung $\tilde{x}_a := C(a)\tilde{u}_a$ gilt dann:

$$\tilde{x}_a - x_a = C(a) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} v \qquad \text{für ein } v \in \mathbb{R}^n.$$

Setzt man nun

$$P(t) := C(t) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} C(t)^{-1},$$

so erkennt man leicht, daß $P(t) : \mathbb{R}^n \to N(t)$ und $P(t)x = x$ für alle $x \in N(t)$. Hieraus folgt nun $R(P(t)) = N(t)$ und wir erkennen, daß genau die Störungen $\tilde{x}_a - x_a \in N(a)$ zulässig sind. ■

1.9 Bemerkungen:
1. Man beachte, daß in $(*)$ keine zweite Ableitung von $x(t)$ auftritt (sonst wäre $(*)$ ja auch keine DAE im Sinne der Definitionen in Kapitel 0). Das liegt daran, daß

$$\begin{aligned} & B(t, h(t, x(t)))\ (h(t, x(t)))' = \\ & = B(t, h(t, x(t)))\ (x(t) + \varepsilon P(t)x''(t) + \varepsilon C(t) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} (C(t)^{-1})'x(t))' \\ & = B(t, h(t, x(t)))\ (x(t) + \varepsilon C(t) \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} (C(t)^{-1})'x(t))' \end{aligned}$$

wobei $P(t)$ die im Beweis benutzte Projektion auf $N(t)$ ist.

2.Weiter sehen wir, daß im Falle $C(t) = const$ $(*)$ ganz einfach wie folgt lautet

$$\begin{aligned} & B(t, x(t) + \varepsilon P(t)x'(t))\,(x(t) + \varepsilon P(t)x'(t))' &&= F(t, x(t) + \varepsilon P(t)x'(t)) \\ \Longleftrightarrow\quad & B(t, x(t) + \varepsilon P(t)x'(t))x'(t) &&= F(t, x(t) + \varepsilon P(t)x'(t)). \end{aligned}$$

3. Für lineare DAEs mit konstanten Koeffizienten

$$Ax'(t) + Bx(t) = q(t) \tag{1.9.1}$$

wurde schon die "Büschel - Regularisierung" vorgeschlagen [Ca]:

$$(A + \varepsilon B)x'(t) + Bx(t) = q(t). \tag{1.9.2}$$

Diese erhält ihre Berechtigung daraus, daß Systeme (1.9.1) zu konsistenten Anfangswerten eindeutig lösbar sind, falls das Matrizenbüschel (A, B) regulär ist, d. h. $det(A + \lambda B) \not\equiv 0$ (siehe z. B. [GriMä]).
Dies ist jedoch schon für variable Koeffizienten i. a. nicht mehr richtig ([GriMä]), so daß eine universelle Verwendbarkeit von (1.9.2) fraglich erscheint.

2. Reduktion von Index-2 auf Index-1

Wir wollen nun versuchen die Ideen aus Kap. 1 auf Index-2 - DAEs zu übertragen. Dabei verwenden wir Ansätze, die die semi - explizite Form nicht zerstören. Im Unterschied zum Index-1 - Fall müssen wir hier jedoch zugestehen, daß die Regularisierungen nicht "immer" funktionieren. Die notwendigen Voraussetzungen lassen aber trotzdem noch eine Fülle von Beispielen zu. Auch läßt sich in manchen Fällen dem eingefügten Störparameter ein physikalischer Sinn geben. Danach demonstrieren wir wie die Resultate für ein semi - explizites System auf allgemeinere DAEs übertragen werden können.

Zunächst sehen wir uns also einmal semi - explizite Index-2 - DAEs an.

2.1 Lemma: (*Index-2 - DAEs in kanonischer Form*)

Vor.:
$f \in C^1(\mathbb{R}^n, \mathbb{R}^m),\ g \in C^1(\mathbb{R}^m, \mathbb{R}^k),\ m + k = n,$
$g_1 \in C^1(\mathbb{R}^{m+k_1}, \mathbb{R}^{k_1}),\ g_2 \in C^1(\mathbb{R}^m, \mathbb{R}^{k_2}),\ k_1 + k_2 = k;$

$$(I) \quad \begin{array}{rcl} y'(t) & = & f(y(t), z(t)) \\ 0 & = & g(y(t)) \end{array} \qquad \textit{mit } g_y \cdot f_z \textit{ regulär}$$

$$(II) \quad \begin{array}{rcl} y'(t) & = & f(y(t), z_1(t), z_2(t)) \\ 0 & = & g_1(y(t), z_1(t)) \\ 0 & = & g_2(y(t)) \end{array} \qquad \textit{mit } (g_1)_{z_1} \textit{ und}(g_2)_y \cdot f_{z_2} \textit{ regulär}$$

Beh.:
(I) und (II) haben Index-2 und ein Anfangswert $(y(0), z(0))^T$ ist genau dann konsistent für (I) bzw. (II), wenn

$$g(y(0)) = g'(y(0))f(y(0), z(0)) = 0$$

bzw.

$$g_1(y(0), z_1(0)) = g_2(y(0)) = g_2'(y(0))f(y(0), z_1(0), z_2(0)) = 0.$$

Beweis: Mittels IRA (Algorithmus 0.6). ■

Bemerkung:
(I) kann als Spezialfall von (II) mit $k_1 = 0$ und $z = z_2$ angesehen werden. Im Fall (II) ist z_1 *Index-1 - Variable* und z_2 *Index-2 - Variable.*

Für die Untersuchung von singulär gestörten Index-2 - DAEs bedienen wir uns des folgenden Lemmas, das eine Version von Satz 0.11 für Index-1 - DAEs darstellt.

2.2 Lemma:

Vor.:
Sei $F \in C^1(\mathbb{R}^{n+1}, \mathbb{R}^m),\ G \in C^2(\mathbb{R}^{n+1}, \mathbb{R}^k),\ k := n - m;\ \varepsilon_0 > 0$

$$(*) \qquad \begin{array}{rcl} y'(t) & = & F(y(t), z(t), \varepsilon) \\ 0 & = & G(y(t), z(t), \varepsilon) \end{array}$$

habe für $\varepsilon = 0$ *eine Lösung* $(\bar{y}, \bar{z})$ *auf* $[0, 1]$; $(*)$ *sei für* $\varepsilon = 0$ *Index-2 - DAE (also insb. in kanonischer Form);*
$l \in \{0, \ldots, k-1\}$; $\delta > 0$, $S := B((\bar{y}([0,1]), \bar{z}([0,1]))^T, \delta)$, *("Schlauch um die exakte Lösung")*

$$\tilde{G}(y, z, \varepsilon) := G_z(y, z, \varepsilon) \begin{pmatrix} I_l & 0 \\ 0 & \varepsilon^{-1} I_{k-l} \end{pmatrix}$$

sei auf $S \times [0, \varepsilon_0]$ *regulär; sei* $M(y, z) \in \mathbb{R}^{(k-l)\times(k-l)}$ *definiert durch*

$$\begin{pmatrix} 0 \\ M(y, z) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & I_{k-l} \end{pmatrix} \left[\tilde{G}(y, z, 0)^{-1} G_y(y, z, 0) F(y, z, 0) \right]_{z_2}$$

wobei $z = (z_1, z_2)^T$ *mit* $z_1 \in \mathbb{R}^l, z_2 \in \mathbb{R}^{k-l}$
und die Eigenwerte von $M(y, z)$ *haben für alle* $(y, z)^T \in \mathbb{R}^n$ *positiven Realteil*
(der Fall $l = 0$ *ist so zu verstehen, daß die entsprechenden Teilmatrizen fehlen)*

Beh.:
$(*)$ *hat für alle* $\varepsilon \in]0, \varepsilon_0]$ Index-1 *und ist* ε*- stabil auf* $[0, 1]$ *bez.* $\{\mathcal{O}_{m+l}\} \times \mathbb{R}^{k-l}$.

Beweis:
Idee: Wir transformieren $(*)$ mittels IRA auf Index-1 und wenden Satz 0.11 an.
Also los:
Ein Schritt mit dem IRA ausgehend von $(*)$ mit $\varepsilon \in]0, \varepsilon_0]$ liefert

$$\begin{array}{rcl} y' & = & F(y, z, \varepsilon) \\ 0 & = & G_y(y, z, \varepsilon) F(y, z, \varepsilon) + G_z(y, z, \varepsilon) z' \end{array}$$

$$\Longleftrightarrow \qquad \begin{array}{rcl} y' & = & F(y, z, \varepsilon) \\ \begin{pmatrix} I_l & 0 \\ 0 & \varepsilon I_{k-l} \end{pmatrix} z' & = & -\tilde{G}(y, z, \varepsilon)^{-1} G_y(y, z, \varepsilon) F(y, z, \varepsilon) \end{array}$$

Nun sehen wir zunächst ein, daß $(*)$ offensichtlich Index-1 hat, denn nach einem Schritt mit dem IRA gelangten wir auf ein Index-0 - System.
Ein Blick auf Satz 0.11 sagt uns nun, daß wir fertig sind, denn die Anforderungen an die Eigenwerte sind nach Voraussetzung gegeben. ■

2.3 Satz:

Vor.:
(II) (wie in Lemma 2.1) habe eine Lösung $(\bar{y}, \bar{z})$ *auf* $[0,1]$;
$f \in C^2$, $g_2 \in C^2$, $F \in C^1(\mathbb{R}^m, \mathbb{R}^m)$ *bijektiv;*
für alle $(y, z) \in \mathbb{R}^n$ *sei* $g_2'(y)F'(y)^{-1}f_{z_2}(y,z)$ *regulär und*

$$\lambda \textit{ Eigenwert von } \left[[g_2'(y)F'(y)^{-1}f_{z_2}(y,z)]^{-1} \cdot g_2'(y)f(y,z)\right]_{z_2} \quad \Rightarrow \quad Re\lambda > 0$$

hierbei ist $z := (z_1, z_2)^T$

Beh.:
Es gibt $\varepsilon_0 > 0$ *so, daß*

$$(II') \qquad \begin{array}{rcl} y' & = & f(y,z) \\ 0 & = & g_1(F^{-1}(F(y) + \varepsilon f(y,z)), z_1) \\ 0 & = & g_2(F^{-1}(F(y) + \varepsilon f(y,z))) \end{array}$$

für alle $\varepsilon \in]0, \varepsilon_0]$ *Index-1 hat und* ε *- stabil auf* $[0,1]$ *ist.*

Beweis:
Natürlich wollen wir Lemma 2.2 anwenden. Vorher sei zur Abkürzung

$$h(y,z,\varepsilon) := F^{-1}(F(y) + \varepsilon f(y,z))$$

und merken uns $h(y,z,0) = y$ und für $i \in \{1,2\}$

$$h_{z_i}(y,z,\varepsilon) = (F^{-1})'(F(y) + \varepsilon f(y,z))\varepsilon f_{z_i}(y,z).$$

Mit den Bezeichnungen aus 2.2 sehen wir aus

$$G_z(y,z,\varepsilon) = \begin{pmatrix} g_{1y}(h(y,z,\varepsilon), z_1)h_{z_1}(y,z,\varepsilon) + g_{1z_1}(h(y,z,\varepsilon), z_1) & g_{1y}(h(y,z,\varepsilon), z_1)h_{z_2}(y,z,\varepsilon) \\ g_2'(h(y,z,\varepsilon))h_{z_1}(y,z,\varepsilon) & g_2'(h(y,z,\varepsilon))h_{z_2}(y,z,\varepsilon) \end{pmatrix}$$

daß sich im Hinblick auf die Anforderungen an $\tilde{G}$ in 2.2 die Wahl $l := k_1$ aufdrängt und es ist dann wegen $(F^{-1})'(F(y)) = F'(y)^{-1}$

$$\tilde{G}(y,z,0) = \left[G_z(y,z,\varepsilon)\begin{pmatrix} I_l & 0 \\ 0 & \varepsilon^{-1}I_{k-l} \end{pmatrix}\right]_{\varepsilon=0} = \begin{pmatrix} g_{1z_1}(y,z_1) & g_{1y}(y,z_1)F'(y)^{-1}f_{z_2}(y,z) \\ 0 & g_2'(y)F'(y)^{-1}f_{z_2}(y,z) \end{pmatrix}$$

Diese Matrix ist nach Voraussetzung regulär und aus Stetigkeitsgründen ist dies auch für $\tilde{G}(y,z,\varepsilon)$ der Fall, sofern $\varepsilon \in [0,\varepsilon_0]$ und für hinreichend kleine $\varepsilon_0 > 0$, $\delta > 0$. Die Voraussetzungen an $\tilde{G}$ in 2.2 sind damit erfüllt und wir haben

$$\begin{pmatrix} 0 & 0 \\ 0 & I_{k_2} \end{pmatrix} \tilde{G}(y,z,0)^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & [g_2'(y)F'(y)^{-1}f_{z_2}(y,z)]^{-1} \end{pmatrix}.$$

Definiert man $M(y,z) \in \mathbb{R}^{(k-l)\times(k-l)}$ durch

$$\begin{aligned} \begin{pmatrix} 0 \\ M(y,z) \end{pmatrix} &= \begin{pmatrix} 0 & 0 \\ 0 & I_{k_2} \end{pmatrix} \left[\tilde{G}(y,z,0)^{-1} G_y(y,z,0) F(y,z,0)\right]_{z_2} \\ &= \left[\begin{pmatrix} 0 & 0 \\ 0 & [g_2'(y)F'(y)^{-1}f_{z_2}(y,z)]^{-1} \end{pmatrix} \begin{pmatrix} g_{1y}(y,z_1) \\ g_2'(y) \end{pmatrix} f(y,z)\right]_{z_2}, \end{aligned}$$

so erhält man

$$M(y,z) = \left[[g_2'(y)F'(y)^{-1}f_{z_2}(y,z)]^{-1} \cdot g_2'(y)f(y,z)\right]_{z_2}$$

und die Eigenwerte dieser Matrix müssen wir nun auf positiven Realteil untersuchen, was unter den gegebenen Voraussetzungen aber leicht fällt. ■

<u>Bemerkungen:</u>
1. Die Aussage von Satz 2.3 bleibt erhalten, wenn man in (II') $f(y,z)$ durch $f(y,z,\varepsilon)$ mit f stetig in ε und $f(y,z,0) = f(y,z)$ ersetzt. Dies ist klar, denn für die Anwendung von Lemma 2.2 ist nur $F(y,z,0)$ interessant und das wird ja nicht geändert.
2. Man beachte, daß der Begriff ε - Stabilität konsistente Anfangswerte erfordert. Das heißt in diesem Fall, daß die Anfangswerte von ε abhängig sein müssen. Sind diese jedoch nahe beim Anfangwert des ungestörten Problems, so besagt 2.3, daß die Lösung des regularisierten Problems nahe bei der des reduzierten Problems ist.
Seien z. B. im Fall $F = Id$ $\bar{y}(0) = y_0, \bar{z}(0) = z_0$ die Anfangswerte für $\varepsilon = 0$, so könnte man einen geeigneten Anfangswert für das regularisierte Problem bestimmen gemäß

$$y_{0\varepsilon} + \varepsilon f(y_{0\varepsilon}, z_{0\varepsilon}) = y_0, \qquad z_{0\varepsilon} = z_0.$$

Daraus sollten sich ohne Schwierigkeiten $y_{0\varepsilon}$ und $z_{0\varepsilon}$ bestimmen lassen, die noch dazu von ε stetig abhängen (Satz über implizite Funktionen). Auch eine nur ungefähre Kenntnis von y_0, z_0 würde das nicht ändern.
3. Lötstedt ([Lö2]) analysiert Regularisierungen von Index-2 - DAEs vom Typ (I) in 2.1 in folgender Form

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

unter der Voraussetzung $g_y \cdot f_z$ negativ definit. Diese Voraussetzung erscheint uns jedoch etwas unnatürlich, wenn man die Herkunft aus einer DAE berücksichtigt. Offensichtlich ist es für diese Stabilitätsbedingung nicht egal, ob die Nebenbedingung der DAE $g(t,y)=0$ oder $-g(t,y)=0$ lautete. Hier gehen also Eigenschaften der Formulierung der DAE ein, die absolut willkürlich sind. Unsere Stabilitätsbedingung (die aus Satz 2.3) hat durchaus physikalische Hintergründe (wie man gleich sehen wird) und ist ganz offensichtlich **nicht** abhängig vom Vorzeichen von g. Sie erscheint auch nur auf den ersten Blick schwieriger nachzuprüfen. Die Bedingung ist nämlich für den Fall $F = Id$ offensichtlich erfüllt, falls $[f_{z_2}(y,z)]_{z_2} \equiv 0$ ist (denn dann ist $[(g_2' f_{z_2})^{-1} g_2' f]_{z_2} \equiv I$). Das ist in vielen Fällen gegeben:

2.4 <u>Beispiel:</u> (*Linienmethode für partielle Differentialgleichungen*)
Die Diskretisierung der Navier-Stokes Gleichungen mittels der Linienmethode führt auf

$$\begin{aligned} Mu' + (K + N(u))u + Cp &= F(u,p) \\ C^T u &= 0 \end{aligned}$$

mit einer positiv definiten Matrix M (für Details siehe [GreLeSa]). In unserer Notation ist $y = u,\ z = p$.
Eine übliche Zusatzvoraussetzung (laut [LöPe]) ist:
$F_p \equiv 0$ (d.h. $F(u,p) = F(u)$) und C hat linear unabhängige Spalten.
Wir setzen der Übersicht halber

$$f(y,z) := M^{-1}(F(y) - Cz - (K + N(y))y), \qquad g(y) := C^T y.$$

Dann haben wir ein Index-2 - System vor uns (denn $C^T M^{-1} C$ ist regulär). Weiter ist $[f_z(y,z)]_z \equiv 0$ und mit der vorangehenden Bemerkung ist klar, daß unsere Regularisierung mit $F = Id$ anwendbar ist.

2.5 <u>Beispiel:</u> (*Destillationskolonnen*)
Zur Beschreibung von Destillationskolonnen dient folgendes System [Mr1]:

$$\begin{array}{rcll} y_1' &=& F(y_1,z_1,z_2) & \text{"Molbilanzen"} \\ A_1(y_1,y_2)y_1' + A_2(y_1,y_2)y_2' &=& G(y_1,y_2,z_1,z_2) & \text{"Wärmebilanzen"} \\ 0 &=& h(y_1,y_2) & \text{"Druckgleichungen"} \\ 0 &=& k(y_1,y_2,z_1) & \text{übrige Beziehungen} \end{array}$$

wobei $y_1 \in \mathbb{R}^{m_1}$, $y_2 \in \mathbb{R}^{m_2}$, $m = m_1 + m_2$, $A_1 : \mathbb{R}^m \rightarrow \mathbb{R}^{m_2 \times m_1}$, $A_2 : \mathbb{R}^m \rightarrow \mathbb{R}^{m_2 \times m_2}$.
A_2 ist regulär und wir können daher setzen

$$\begin{pmatrix} f_1(y_1,y_2,z_1,z_2) \\ f_2(y_1,y_2,z_1,z_2) \end{pmatrix} := \begin{pmatrix} I & 0 \\ A_1(y_1,y_2) & A_2(y_1,y_2) \end{pmatrix}^{-1} \begin{pmatrix} F(y_1,z_1,z_2) \\ G(y_1,y_2,z_1,z_2) \end{pmatrix}$$

$$g_1 := k, \quad g2 := h$$

und die Gegebenheiten sind nun gerade so, daß dieses System die Gestalt (II) von Lemma 2.1 hat und damit vom Index-2 ist. z_1 ist Index-1 - Variable und z_2 Index-2 - Variable. Bei manchen Modellen erfüllt das oben definierte $f = (f_1, f_2)^T \qquad (f_{z_2})_{z_2} \equiv 0$ [Mr2], so daß das System ε - stabil für die Regularisierung mit $F = Id$ ist.

Auch für diesen in 2.3 vorgestellten Regularisierungsansatz läßt sich dem künstlich eingeführten Parameter ε eine physikalische Bedeutung zuordnen. Dies wird in einem weiteren Beispiel verdeutlicht.

2.6 <u>Beispiel:</u>
Physikalische Interpretation der vorgestellten Regularisierungen für Index-2 bei elektrischen Netzwerken
Betrachte die folgende einfache Schaltung mit einer Induktivität L und einer Stromquelle $I_0(t)$

Die zugehörigen Netzwerkgleichungen sind dann

$$\begin{aligned} Li_1' + v_0 &= 0 \qquad \text{Maschenregel} \\ i_1 &= I_0 \qquad \text{Knotenregel} \end{aligned}$$

Mit den Bezeichnungen $y := i_1,\ z := v_0,\ f(y,z) := -\frac{1}{L}z,\ g(y) := y - I_0$ haben wir das System in der uns wohlbekannten semi - expliziten Form (I) bzw. – mit $z_2 := z$ – (II) vorliegen. Dieses hat dann wegen $g'f_z = 1 \cdot -\frac{1}{L}$ (was natürlich regulär ist) Index-2.
Ein physikalisch besseres Modell (vgl. 1.7) wäre nun das folgende erweiterte Netzwerk, was durch die Gleichungen

$$\begin{aligned} Li_1' + v_0 &= 0 \qquad \text{Maschenregel} \\ Li_1' &= v_R \\ i_R &= \varepsilon \cdot v_R \end{aligned}$$

beschrieben wird. Nun ist weiter

$$i_1 + i_R = I_0 \qquad (\text{ Knotenregel }) \qquad \begin{aligned} &\Rightarrow i_1 + \varepsilon \cdot v_R = I_0 \\ &\Rightarrow i_1 + \varepsilon \cdot L i_1' = I_0 \\ &\Rightarrow i_1 + \varepsilon \cdot (-v_0) = I_0 \end{aligned}$$

und damit können wir das gestörte System in den Variablen i_1, v_0 formulieren als

$$\begin{aligned} L i_1' &= -v_0 \\ 0 &= i_1 + \varepsilon \cdot (-v_0) - I_0. \end{aligned}$$

In unserer Notation ist das aber nichts anderes als

$$\begin{aligned} y' &= f(y,z) \\ 0 &= g(y + \varepsilon L f(y,z)) \end{aligned}$$

Das ist aber gerade (II') aus 2.3 mit $F(y) = L^{-1}y$. Die ε - Stabilität dieses Systems ist klar, denn

$$\left[(g'(F')^{-1} f_z)^{-1} g' f\right]_z = \left[(1 \cdot L(-L^{-1}))^{-1} 1(-L^{-1}z)\right]_z = L^{-1} > 0.$$

Übrigens ist bei Netzwerken mit linearen Widerständen (d.h. $U = IR$) stets die Bedingung aus obiger Bemerkung erfüllt (klar !), so daß die ε - Stabilität der Regularisierung für $F = Id$ allgemein gesichert ist (siehe die Bemerkung nach Satz 2.3).
Ein ähnliches Beispiel wird auch von Lötstedt [Lö2] angegeben. Für weitere weniger triviale Beispiele von Netzwerken mit Index > 1 siehe z. B. [Ad] (Stichwort "excess element circuits").

2.7 Beispiel: *DAEs in der Mechanik*
Einige kurze Bemerkungen zu DAEs in der Mechanik findet man in Kap. 3. Üblicherweise sind selbige zunächst vom Index-3, aber wir wissen ja, daß man den Index durch einen Schritt mit dem IRA um eins reduzieren kann. Man erhält dann [Lö2]:

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= F(t,y_1,y_2) + G(y_1)z \\ 0 &= G^T(y_1)y_2 . \end{aligned} \tag{2.7.1}$$

Das ist eine Index-2 - DAE in semi - expliziter Form. Die eben vorgestellte Regularisierung würde auf

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= F(t,y_1,y_2) + G(y_1)z \\ 0 &= G^T(y_1+\varepsilon y_2)(y_2 + \varepsilon[F(t,y_1,y_2) + G(y_1)z]) \end{aligned}$$

und damit

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= F(t,y_1,y_2) + G(y_1)z \\ \varepsilon G^T(y_1+\varepsilon y_2)G(y_1)z &= -G^T(y_1+\varepsilon y_2)(y_2 + \varepsilon F(t,y_1,y_2)) \end{aligned}$$

führen. Die letzte Gleichung könnte, falls G^TG vollen Rang hat, nach z aufgelöst werden. Unter der gleichen Voraussetzung (und noch etwas mehr) ist auch die Analyse in [Lö2] anwendbar. Dort wird die Regularisierung wie folgt vorgenommen:

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= F(t,y_1,y_2) + G(y_1)z \\ \varepsilon z &= -\left[G^T(y_1)y_2 + \kappa\Phi(y_1)\right] . \end{aligned}$$

Hierbei ist $G^T = \Phi'$ und $\kappa > 0$ eine Konstante. Für diesen Fall läßt sich auch eine physikalische Interpretation angeben [Lö2]. Wir wollen noch darauf aufmerksam machen, daß (2.7.1) wieder die in obiger Bemerkung angegebene Voraussetzung erfüllt, so daß auch hier unsere Regularisierung ε - stabil ist (mit $F = Id$).

Wir wollen nun auch Satz 2.3 auf voll - implizite Systeme übertragen. Dies geschieht - wie in Kapitel 1 -, indem wir zunächst das System auf semi - explizite Form transformieren, dann Satz 2.3 benutzen, und anschließend wieder zurücktransformieren. Hier sind aber zwei verschiedene Transformationen nötig; dies macht die Sache weniger übersichtlich als im Index-1 - Fall (siehe Kap. 1). Wir werden daher aus Gründen der Überschaubarkeit auf die allgemeinste Version verzichten, beschränken uns auf den Fall $F = Id$ und erzielen dann

2.8 <u>Satz:</u>

<u>Vor.:</u>

$B : \mathbb{R}^{n+1} \to \mathbb{R}^{n\times n},\ F : \mathbb{R}^{n+1} \to \mathbb{R}^n$;

$$(*) \qquad B(t,x)x' = F(t,x)$$

habe eine Lösung $\bar{x}(t)$ auf $[0,1]$;
(Hinweis: die folgenden Voraussetzungen beschreiben lediglich, daß $(*)$ mit einigen Transformationen (wie in Lemma 0.3) in ein Index-2 - System überführt werden kann)
$A, C : \mathbb{R} \to \mathbb{R}^{n\times n}$ reguläre Matrizen; C sei differenzierbar; $u(t) := C(t)^{-1}x(t)$ für $t \in \mathbb{R}$, für ein $m \in \mathbb{N}$ und alle $x \in \mathbb{R}$ sei

$$A(t)B(t,x)C(t) = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix};$$

$k := n - m,\ \ A_1 \in C(\mathbb{R}^{n+1}, \mathbb{R}^{m\times n}),\ \ A_2 \in C^1(\mathbb{R}^{n+1}, \mathbb{R}^{k\times n})$ mit

$$A(t)F(t, C(t)u) = \begin{pmatrix} A_1(t,u) \\ A_2(t,u) \end{pmatrix}$$

$C_1 \in C^1(\mathbb{R}, \mathbb{R}^{m\times n}),\ \ C_2 \in C^1(\mathbb{R}, \mathbb{R}^{k\times n})$ mit

$$C^{-1}(t)C'(t) = \begin{pmatrix} C_1(t) \\ C_2(t) \end{pmatrix}$$

$\tilde{A}, \tilde{C} \in C^1(\mathbb{R}, \mathbb{R}^{k\times k}),\ \ k_1 \in \{0, \ldots, k\}$ mit

$$\tilde{A}(t)A_{2u_2}(t,u)\tilde{C}(t) = \begin{pmatrix} I_{k_1} & 0 \\ 0 & 0 \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \ \text{ mit } \ u_1 \in \mathbb{R}^m,\ u_2 \in \mathbb{R}^k$$

$H \in C^1(\mathbb{R}^{n+1}, \mathbb{R}^k),\quad H_1 \in C^1(\mathbb{R}^{m+k_1+1}, \mathbb{R}^{k_1}),\quad H_2 \in C^1(\mathbb{R}^{m+1}, \mathbb{R}^{k_2})$ seien definiert durch

$$H(t, v_1, v_2) := \begin{pmatrix} H_1(t, v_1, v_{21}) \\ H_2(t, v_1) \end{pmatrix} := \tilde{A}(t)A_2(t, \begin{pmatrix} v_1 \\ \tilde{C}(t)v_2 \end{pmatrix})$$

wobei $v = (v_1, v_2)^T = (v_1, v_{21}, v_{22})^T$ mit $v_1 \in \mathbb{R}^m,\ v_{21} \in \mathbb{R}^{k_1},\ v_{22} \in \mathbb{R}^{k_2}$,
(daß das so definiert werden kann, wird im Beweis klar)
und seien $H_{1v_{21}}(t, v_1, v_{21})$ und

$$H_{2v_1}(t, v_1)\tilde{f}_{v_{22}}(t, v_1, v_2) \ \text{ wobei } \ \tilde{f}(t, v_1, v_2) := A_1(t, \begin{pmatrix} v_1 \\ \tilde{C}(t)v_2 \end{pmatrix}) - C_1(t)\begin{pmatrix} v_1 \\ \tilde{C}(t)v_2 \end{pmatrix}$$

regulär;

nun gelte aber noch folgende Stabilitätsbedingung
die Eigenwerte von

$$\left[(H_{2v_1}(t,v_1)\tilde{f}_{v_{22}}(t,v_1,v_2))^{-1}H_{2v_1}(t,v_1)\tilde{f}(t,v_1,v_2)\right]_{v_{22}}$$

haben positiven Realteil;
zu guter letzt sei zur Abkürzung

$$P(t) := A(t)^{-1}\begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} A(t)$$

Beh.:
$(*)$ *läßt sich gemäß Lemma 0.3 auf ein Index-2 - System transformieren und es gibt ein* $\varepsilon_0 > 0$ *so, daß das System*

$$(**) \quad \begin{aligned} &B(t,x)x' = P(t)F(t,x)+ \\ &(I-P(t))F(t,x+\varepsilon C(t)\begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}\left[A(t)F(t,x) - C(t)^{-1}C'(t)C(t)^{-1}x\right]) \end{aligned}$$

für alle $\varepsilon \in]0,\varepsilon_0]$ *sich in eine Index-1 - DAE transformieren läßt und* ε *- stabil auf* $[0,1]$ *ist.*

Beweis:
Zuerst transformieren wir $(*)$ auf die kanonische Index-2 - Form:
mit der Transformation $x(t) = C(t)u(t)$ sehen wir mit Lemma 0.3, daß $(*)$ äquivalent ist zu

$$(2.8.1) \qquad \begin{pmatrix} u_1' \\ 0 \end{pmatrix} = A(t)F(t,C(t)u) - \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} C(t)^{-1}C'(t)u = \begin{pmatrix} A_1(t,u) - C_1(t)u \\ A_2(t,u) \end{pmatrix}$$

wobei $u = (u_1,u_2)^T$ mit $u_1 \in \mathbb{R}^m,\ u_2 \in \mathbb{R}^k$.
Setzen wir nun

$$v := \begin{pmatrix} I_m & 0 \\ 0 & \tilde{C}(t)^{-1} \end{pmatrix} u,$$

so ist

$$(2.8.1) \iff \begin{matrix} u_1' & = & A_1(t,u) - C_1(t)u \\ 0 & = & H(t,v_1,v_2) \end{matrix} \iff \qquad (2.8.2) \quad \begin{matrix} v_1' & = & \tilde{f}(t,v_1,v_2) \\ 0 & = & H(t,v_1,v_2). \end{matrix}$$

Nebenbei bemerkt ist

$$H_{v_2}(t,v_1,v_2) = \tilde{A}(t)A_{2u_2}(t,\begin{pmatrix} v_1 \\ \tilde{C}(t)v_2 \end{pmatrix})\tilde{C}(t) = \begin{pmatrix} I_{k_1} & 0 \\ 0 & 0 \end{pmatrix}$$

und dies zeigt uns, daß H in (H_1,H_2) wie in der Voraussetzung angegeben aufgespalten werden kann. (2.8.2) ist daher von der Form (II) aus Lemma 2.1 (mit $g_1 = H_1,\ g_2 = H_2$);

daraus sieht man auch, daß $(*)$ Index-2 hat (denn (2.8.2) hat Index-2). Hier ist v_{21} Index - 1 - Variable und v_{22} Index - 2 - Variable. Wir können nun gemäß 2.3 regularisieren, der Einfachheit halber mit $F = Id$. Dies liefert

$$
\begin{aligned}
v_1' &= \tilde{f}(t, v_1, v_2) \\
0 &= H(t, v_1 + \varepsilon \tilde{f}(t, v_1, v_2), v_2) \\
\Longleftrightarrow & \\
u_1' &= A_1(t,u) - C_1(t)u \\
0 &= H(t, u_1 + \varepsilon \left[A_1(t,u) - C_1(t)u\right], \tilde{C}(t)^{-1} u_2) \\
&= \tilde{A}(t) A_2(t, \begin{pmatrix} u_1 + \varepsilon[A_1(t,u) - C_1(t)u] \\ u_2 \end{pmatrix}) \\
\Longleftrightarrow & \\
u_1' &= A_1(t,u) - C_1(t)u \\
0 &= A_2(t, u + \varepsilon \begin{pmatrix} A_1(t,u) - C_1(t)u \\ 0 \end{pmatrix}) \\
\Longleftrightarrow & \\
\begin{pmatrix} u_1' \\ 0 \end{pmatrix} &= \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} \left[A(t)F(t, C(t)u) - C(t)^{-1}C'(t)u\right] \\
&\quad + \begin{pmatrix} 0 & 0 \\ 0 & I_k \end{pmatrix} A(t)F(t, C(t)u + \varepsilon C(t) \begin{pmatrix} A_1(t,u) - C_1(t)u \\ 0 \end{pmatrix}) \\
\Longleftrightarrow & \\
\begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} u' &= \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} \left[A(t)F(t,x) - C(t)^{-1}C'(t)u\right] \\
&\quad + \begin{pmatrix} 0 & 0 \\ 0 & I_k \end{pmatrix} A(t)F(t, x + \varepsilon C(t)\tilde{I}\left[A(t)F(t,x) - C(t)^{-1}C'(t)u\right]) \\
\Longleftrightarrow & \\
B(t,x)C(t)u' &= B(t,x)C(t)A(t)F(t,x) - B(t,x)C'(t)u + \\
&\quad (I - P(t))F(t, x + \varepsilon C(t)\tilde{I}\left[A(t)F(t,x) - C(t)^{-1}C'(t)u\right]) \\
\Longleftrightarrow & \\
B(t,x)x' &= P(t)F(t,x) + \\
&\quad (I - P(t))F(t, x + \varepsilon C(t)\tilde{I}\left[A(t)F(t,x) - C(t)^{-1}C'(t)C(t)^{-1}x\right]),
\end{aligned}
$$

aber das ist ja gerade $(**)$.

Daß die in der Voraussetzung angegebene Stabilitätsbedingung gerade die entsprechende aus Satz 2.3 ist, rechnet man leicht mit (2.8.2) nach. ■

2.9 <u>Bemerkung:</u>

Eine alternative Regularisierung für lineare DAEs (1.9.1) ist

(2.9.1) $$(A(t) + \varepsilon B(t)Q(t))x'(t) + (B(t) + \varepsilon B(t)Q(t)Q'(t))x(t) = q(t)$$

wie von März vorgeschlagen ([Mä1],[Mä2]), wobei $Q(t) = I - P(t)$ mit $P(t)$ wie in 1.8. Hier kann gezeigt werden, daß (2.9.1) Index-1 hat, falls (1.9.1) Index-2 hat. Ist (1.9.1) vom

Index-1, so ist allerdings auch (2.9.1) vom Index-1 ([Mä1]). Diese Regularisierung läßt sich übrigens nicht als Variante der hier vorgestellten Ansätze auffassen. Das sieht man sofort, wenn man ein semi - explizites System betrachtet.
(2.9.1) hat gegenüber ($**$) aus 2.8 den Vorteil, daß keine impulsartigen Übergänge in der Nähe des Anfangswertes auftreten, sofern der Anfangswert gewisse Bedingungen erfüllt ([Mä2]).
Die hier vorgestellten Regularisierungen basieren auf singulär gestörten Differentialgleichungen und daher können impulsartige Übergänge in der Nähe des Anfangswerts nicht vermieden werden.

3. Reduktion von Index-3 auf Index-2

Wir fahren nun fort mit der Untersuchung von Index-3 Systemen. Der technische Aufwand in Kap. 2 läßt hier unangenehme Vorahnungen aufkommen, so daß wir nicht in größt möglicher Allgemeinheit schwelgen wollen. Insbesondere werden wir uns ganz auf semi - explizite DAEs beschränken und auch davon nur eine Teilklasse untersuchen. Für diese wird eine Regularisierung vorgestellt, die in vielen Fällen anwendbar ist. Die Erfahrung mit den Anwendungen von DAEs zeigt, daß Index-3 - DAEs im wesentlichen nur in semi - expliziter Form auftreten und zwar in der Beschreibung von mechanischen Systemen. Diese Anwendungsklasse wird gerade die oben erwähnte Teilklasse sein.

Doch zunächst beschreiben wir Index-3 - DAEs, die keine Index-1 - und Index-2 - Variablen enthalten (vgl. die Bemerkung nach 0.7).

3.1 Lemma:

Vor.:
$f_1 \in C^1(\mathbb{R}^m, \mathbb{R}^{m_1})$; $f_2 \in C^1(\mathbb{R}^n, \mathbb{R}^{m_2})$; $g \in C^2(\mathbb{R}^{m_1}, \mathbb{R}^k)$;
$m + k = n$; $m_1 + m_2 = m$;

$$\begin{aligned} y_1' &= f_1(y_1, y_2) \\ y_2' &= f_2(y_1, y_2, z) \\ 0 &= g(y_1) \end{aligned} \tag{3.1.1}$$

mit $g'(y_1)\, f_{1y_2}(y_1, y_2)\, f_{2z}(y_1, y_2, z)$ regulär für alle $(y_1, y_2, z) \in \mathbb{R}^n$

Beh.:
(3.1.1) *hat Index-3, z ist Index-3 - Variable, und ein Anfangswert $(y_1(0), y_2(0), z(0))^T$ ist genau dann konsistent für* (3.1.1), *wenn mit* $y(t) := (y_1(t), y_2(t))^T$

$$\begin{aligned} 0 &= g(y_1(0)) = g'(y_1(0))\, f_1(y(0)) \quad \textit{und} \\ 0 &= [g'(y_1(0))\, f_1(y(0))]_{y_1}\, f_1(y(0)) + g'(y_1(0))\, f_{1y_2}(y(0))\, f_2(y(0), z(0)). \end{aligned}$$

erfüllt ist.

Beweis:
Mittels IRA (Algorithmus 0.6). Zweimaliges Differenzieren von $g(y_1(t)) = 0$ bringt das erwünschte. ■

Bemerkung:
Wie auch in den vorangegangenen Kapiteln formulieren wir alle Aussagen für auf ganz $\mathbb{R}^m$ (u. ä.) definierte Abbildungen. Aus den jeweiligen Beweisen geht aber hervor, daß so gut wie alle Aussagen auch für nur lokal definierte Abbildungen gelten (lokal meint dabei in einer kompakten Umgebung der exakten Lösung). Wo das nicht zutrifft, werden wir es ausdrücklich vermerken.
Außer Acht lassen wollen wir in diesem Kapitel die Index-3 - DAEs, die zusätzlich noch Index-1 - und Index-2 - Variablen enthalten und solche DAEs, die auf diese transformiert werden könnten. Wie schon gesagt haben wir aber die für Anwendungen relevanten DAEs damit dabei.

Mit der Methode in der DAE y durch $y + \varepsilon f(y,z)$ in der Nebenbedingung zu ersetzen haben wir in Kap. 2 schon Erfolge verzeichnet. Es liegt daher nahe es auch hier wieder damit zu versuchen. Für den in Lemma 3.1 vorgestellten Fall funktioniert das auch. Es gilt nämlich

3.2 Satz: (*Regularisierung von gewissen Index-3 - DAEs*)

Vor.:
(3.1.1) wie in 3.1 mit einer Lösung $(\bar{y}_1(t), \bar{y}_2(t), \bar{z}(t))^T$ auf $[0,1]$; jedoch sei hier $f_1,\ f_2 \in C^2,\ g \in C^3$
die Eigenwerte von

$$\left[[g'(y_1)\, f_{1y_2}(y) f_{2z}(y,z)]^{-1} \, [g'(y_1)\, f_1(y)]_y \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix} \right]_z$$

haben für alle $(y,z)^T \in \{(\bar{y}_1(t), \bar{y}_2(t), \bar{z}(t))^T \,|\, t \in [0,1]\}$ (hierbei ist $y := (y_1, y_2)^T$) positiven Realteil

Beh.:
Es gibt $\varepsilon_0 > 0$ so, daß

$$\begin{aligned} y_1' &= f_1(y_1, y_2) \\ y_2' &= f_2(y_1, y_2, z) \\ 0 &= g(y_1 + \varepsilon f_1(y_1, y_2)) \end{aligned} \tag{3.2.1}$$

für alle $\varepsilon \in]0, \varepsilon_0]$ Index-2 hat. z ist Index-2 - Variable und (3.2.1) ist ε - stabil auf $[0,1]$.

<u>Beweis:</u>
Wir verwenden wieder die Notation $y = (y_1, y_2)^T$.
Um (3.2.1) auf Index-2 zu prüfen, können wir die Kriterien aus Kap. 2 verwenden und diese sagen uns, daß es reicht mit $\tilde{g}(y) := g(y_1 + \varepsilon f_1(y))$

$$\tilde{g}'(y) \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix}_z$$

als regulär nachzuweisen.
Es gilt nun

$$\begin{aligned}\tilde{g}'(y) \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix}_z &= \tilde{g}'(y) \begin{pmatrix} 0 \\ f_{2z}(y,z) \end{pmatrix}_z = \tilde{g}_{y_2}(y)\, f_{2z}(y,z) \\ &= \varepsilon\, g'(y_1 + \varepsilon f_1(y))\, f_{1y_2}(y)\, f_{2z}(y,z)\end{aligned}$$

und diese Matrix ist regulär (dies sieht man unter Benutzung der Voraussetzung (der aus 3.1) und eines Stetigkeitsargumentes).
Nun weisen wir die ε - Stabilität von (3.2.1) nach.
Da wir für $\varepsilon > 0$ ein Index-2 - System vor uns haben, erzielen wir durch Anwendung des IRA (zwei Schritte) ein Index-0 - System, das eine singulär gestörte gewöhnliche Differentialgleichung ist. Auf diese können wir wieder den üblichen Satz loslassen.
Ein Schritt des IRA für (3.2.1) liefert

$$\begin{aligned} y_1' &= f_1(y) \\ y_2' &= f_2(y,z) \\ 0 &= g'(y_1 + \varepsilon f_1(y))\, [(I + \varepsilon f_{1y_1}(y)) f_1(y) + \varepsilon f_{1y_2}(y) f_2(y,z)] =: G(y,z,\varepsilon) \end{aligned}$$

und ein weiterer Schritt bringt

$$\begin{aligned} y_1' &= f_1(y) \\ y_2' &= f_2(y,z) \\ 0 &= G_y(y,z,\varepsilon) \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix} + G_z(y,z,\varepsilon)\, z'. \end{aligned} \tag{3.2.2}$$

Nun ist

$$G_z(y,z,\varepsilon) = \varepsilon\, g'(y_1 + \varepsilon f_1(y))\, f_{1y_2}(y) f_{2z}(y,z)$$

und das hatten wir oben schon als regulär erkannt. Wir sehen damit

$$(3.2.2) \quad \Longleftrightarrow$$

$$\begin{aligned} y_1' &= f_1(y) \\ y_2' &= f_2(y,z) \\ \varepsilon z' &= -\left[g'(y_1 + \varepsilon f_1(y)) f_{1y_2}(y) f_{2z}(y,z)\right]^{-1} G_y(y,z,\varepsilon) \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix} \end{aligned} \tag{3.2.3}$$

und (3.2.3) ist nach Satz 0.10 ε - stabil, falls die Eigenwerte von

$$M(y,z) := \left[[g'(y_1)\, f_{1y_2}(y) f_{2z}(y,z)]^{-1}\, G_y(y,z,0) \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix} \right]_z$$

für alle (y, z) auf der Kurve $(\bar{y}(t), \bar{z}(t))^T$ positiven Realteil haben. Es ist nun aber

$$M(y,z) = \left[[g'(y_1)\, f_{1y_2}(y) f_{2z}(y,z)]^{-1}\, [g'(y_1)\, f_1(y)]_y \begin{pmatrix} f_1(y) \\ f_2(y,z) \end{pmatrix} \right]_z$$

und das zu überprüfende steht gerade in der Voraussetzung. ■

<u>Bemerkung:</u>
Bezüglich konsistenter Anfangswerte für das regularisierte System gilt das in Kap. 2 gesagte entsprechend.

Im Vergleich zu den schon behandelten Index-1 - und Index-2 - Systemen mag das hier untersuchte Index-3 - System recht speziell aussehen. Jedoch sind **alle** uns bekannten Index-3 - DAEs genau von der Form (3.1.1). Einen breiten Raum nehmen dabei die aus der Mechanik herrührenden DAEs ein. Auf diese wollen wir hier kurz eingehen.

3.3 <u>Beispiel:</u> (*DAEs in der Mechanik*)
Die Bewegung von mechanischen Systemen kann durch folgende DAE beschrieben werden ("holonome Nebenbedingungen"):

$$\begin{aligned} y_1' &= y_2 \\ M(y_1) y_2' &= F(t, y_1, z) \\ G(y_1) &= 0. \end{aligned} \tag{3.3.1}$$

Hierbei ist y_1 der Vektor, der die Position des oder der beteiligten Körper enthält, y_2 ist einfach die Ableitung von y_1 (denn (3.3.1) entsteht durch Umschreiben einer Differentialgleichung zweiter Ordnung auf erste Ordnung), z entspricht den Zwangskräften. $M(y_1)$ ist eine (reguläre) Massenmatrix. Die Nebenbedingungsgleichung gibt geometrische Einschränkungen für die Positionen des oder der Körper wieder. Für Details siehe z. B. [Fü]. Man vergleiche mit der Pendel - Gleichung (0.8.2). Dort besagt die Nebenbedingung gerade, daß der Körper sich auf einer Kreisbahn mit Radius 1 bewegt.
Ganz offensichtlich ist (3.3.1) gerade ein Spezialfall von (3.1.1) ($f_1(y_1, y_2) := y_2$, $f_2(t, y_1, y_2, z) := M(y_1)^{-1} F(t, y_1, z)$, $g := G$ zeigt das; die Frage, ob das System in der autonomen oder nicht-autonomen Form formuliert wird, spielt für die hier vorgenommenen Index - und Stabilitätsuntersuchungen keine Rolle). In vielen Fällen (z. B. eben auch bei der Pendel - Gleichung) ist F von der Form $F(t, y_1, z) = F_1(t, y_1) z + F_2(t, y_1)$ (siehe die Beispiele in [Lö1], [Lö2], [LöPe], [Fü]) und damit führt die oben vorgestellte Regularisierung auf ein ε - stabiles System.

4. Runge - Kutta - Verfahren für regularisierte Index-1 - DAEs

Wir wollen nun untersuchen, wie sich numerische Verfahren verhalten, wenn man sie auf regularisierte DAEs anwendet, wobei wir uns auf semi - explizite Systeme beschränken wollen. Wir konzentrieren uns hier auf Einschrittverfahren und dabei auf Runge - Kutta - Verfahren. Ähnliche Resultate wie die hier vorgestellten wurden schon von Lötstedt [Lö1], [Lö2] für BDF - Formeln erhalten. Wir zeigen, daß für die linearen Regularisierungen für Index-1 - DAEs eine große Klasse von Runge - Kutta - Verfahren bez. ε gleichmäßige Abschätzungen des globalen Fehlers zulassen. Für die nichtlinearen Störungen läßt sich immerhin noch die numerische Lösung nach einem Schritt gleichmäßig beschränken. Außerdem deuten wir mit einem Beispiel an, daß ähnliche Aussagen auch für Regularisierungen von Index-2 - DAEs nicht auszuschließen sind.

Zunächst aber einige Definitionen.

4.1 Definition/Lemma: (*Direktes Produkt von Matrizen*) [La], [DekVe]
Sei $A \in \mathbb{R}^{s\times s},\ B \in \mathbb{R}^{m\times m}$.
(i)
Wir definieren

$$A \otimes B := \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1s}B \\ a_{21}B & a_{22}B & \cdots & a_{2s}B \\ \vdots & \vdots & & \vdots \\ a_{s1}B & a_{s2}B & \cdots & a_{ss}B \end{pmatrix}$$

(ii) *Rechenregeln* [La]
Seien weiter $C \in \mathbb{R}^{s\times s},\ D \in \mathbb{R}^{m\times m}$. Dann gilt:

$$\begin{array}{rcll} (A\otimes B)(C\otimes D) &=& AC\otimes BD & \\ A\otimes B &=& (A\otimes I_m)(I_s\otimes B) & \\ (A\otimes B)^{-1} &=& A^{-1}\otimes B^{-1} & \text{falls } A,B \text{ regulär sind.} \end{array}$$

4.2 <u>Definition/Lemma:</u> (*Runge - Kutta - Verfahren*)

(i)

(s, A, b) mit $s \in \mathbb{N},\ A \in \mathbb{R}^{s\times s},\ b \in \mathbb{R}^s$ heißt *Runge - Kutta - Verfahren* .

Setze noch für $i = 1,\dots,s$: $c_i := \sum\limits_{j=1}^{s} a_{ij}$.

(ii)

Runge - Kutta - Verfahren für implizite Differentialgleichungen $F(t, x, x') = 0$ werden wie folgt definiert:

Löse

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

und setze

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

Man hofft dabei, daß

$$x_{n+1} \approx x(t_{n+1}), \quad \text{falls} \quad x_n \approx x(t_n).$$

Hierbei ist $x(t)$ die exakte Lösung an der Stelle t und $t_{n+1} = t_n + h$.

(iii) (äquivalente) Alternative zu (ii)

Sei A regulär mit $D = (d_{ij}) = A^{-1}$.

Löse

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

und setze

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

Hoffnung wie in (ii).

(iv)

Ein Runge - Kutta - Verfahren (s, A, b) habe für $p \in \mathbb{N}$ die Eigenschaft

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

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

Seien die Bezeichnungen wie in (ii) bzw. (iii) und sei x_{n+1} der ausgehend von $x_n := x(t_n)$ berechnete Wert.

Das Runge - Kutta - Verfahren hat die Eigenschaft $A(p)$ $:\Longleftrightarrow$ es gibt ein $h_0 > 0$ und ein $C > 0$ mit

$$\|x_{n+1} - x(t_{n+1})\| \le Ch^{p+1} \text{ für alle } h \in]0, h_0].$$

Wir sagen auch:
das Runge - Kutta - Verfahren hat *lokale Fehlerordnung* p.

Es gilt für alle $\bar{x} \in C^{p+1}([t_0, t_0 + h])$*, falls* $C(p)$ *und* $B(p)$ *gegeben ist (siehe z. B.* [DekVe] *):*

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

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

Beweis:

$$\begin{aligned}
\bar{x}(t_0 + c_i h) - \bar{x}(t_0) - h \sum_{j=1}^{s} a_{ij} \bar{x}'(t_0 + c_j h) = \\
&= \sum_{l=1}^{p} \frac{1}{l\,!} \bar{x}^{(l)}(t_0) c_j^l h^l - h \sum_{j=1}^{s} a_{ij} \sum_{l=0}^{p-1} \frac{1}{l\,!} \bar{x}^{(l+1)}(t_0) c_j^l h^l + O(h^{p+1}) \\
&= \sum_{l=1}^{p} \frac{1}{l\,!} \bar{x}^{(l)}(t_0) c_j^l h^l - \sum_{l=1}^{p} \frac{1}{(l-1)\,!} \bar{x}^{(l)}(t_0) h^l \sum_{j=1}^{s} a_{ij} c_j^{l-1} + O(h^{p+1}) \\
&= \sum_{l=1}^{p} \frac{1}{(l-1)\,!} \bar{x}^{(l)}(t_0) h^l \left[\frac{1}{l} c_j^l - \sum_{j=1}^{s} a_{ij} c_j^{l-1} \right] + O(h^{p+1}) \\
&= O(h^{p+1}) \qquad \text{wegen } \ C(p)
\end{aligned}$$

und

$$\begin{aligned}
\bar{x}(t_0 + h) - \bar{x}(t_0) - h \sum_{i=1}^{s} b_i \bar{x}'(t_0 + c_i h) = \\
&= \sum_{l=1}^{p} \frac{1}{l\,!} h^l \bar{x}^{(l)}(t_0) - h \sum_{i=1}^{s} b_i \sum_{l=0}^{p-1} \frac{1}{l\,!} h^l \bar{x}^{(l+1)}(t_0) c_i^l + O(h^{p+1}) \\
&= \sum_{l=1}^{p} \frac{1}{(l-1)\,!} h^l \bar{x}^{(l)}(t_0) \left[\frac{1}{l} - \sum_{i=1}^{s} b_i c_i^{l-1} \right] + O(h^{p+1}) \\
&= O(h^{p+1}) \qquad \text{wegen } \ B(p).
\end{aligned}$$

<u>Bemerkung:</u>
Daß (ii) und (iii) beide das gleiche x_{n+1} liefern (falls die angegebenen Gleichungssysteme lösbar sind und A regulär ist), sieht man wie folgt:

$$X_i = x_n + h \sum_{j=1}^{s} a_{ij} X_j' \qquad \text{für} \qquad i = 1, \quad \ldots \quad , s$$

$$\begin{aligned} \Longleftrightarrow \qquad \vec{X} \quad &= \quad e \otimes x_n + h(A \otimes I_n)\vec{X}' \\ \Longleftrightarrow \qquad \vec{X}' \quad &= \quad \tfrac{1}{h}(A^{-1} \otimes I_n)(\vec{X} - e \otimes x_n) \\ \Longleftrightarrow \qquad X_i' \quad &= \quad \tfrac{1}{h} \sum_{j=1}^{s} d_{ij}(X_j - x_n) \text{ für } i = 1, \ldots, s. \end{aligned}$$

Hierbei ist

$$e := (1, \ldots, 1)^T \in \mathbb{R}^s, \quad \vec{X} := (X_1^T, \ldots, X_s^T)^T \in \mathbb{R}^{ns}, \quad \vec{X}' := (X_1'^T, \ldots, X_s'^T)^T \in \mathbb{R}^{ns}.$$

4.3 <u>Definition/Lemma:</u> [BuHuVe] (*A -stabil, AS - stabil, ASI -stabil*)
Sei (s, A, b) ein Runge - Kutta - Verfahren ; $\mathbb{C}^- := \{z \in \mathbb{C} \,|\, Re(z) \leq 0\}$.
(s, A, b) heißt A - *stabil*, falls mit

$$R(z) := 1 + zb^T(I_s - zA)^{-1}e \qquad \text{für } z \in \mathbb{C}$$

für alle $z \in \mathbb{C}^-$ gilt: $\qquad |R(z)| \leq 1.$
(s, A, b) heißt *ASI - stabil*, falls für alle $z \in \mathbb{C}^-$ die Matrix $I - zA$ regulär ist und $(I - zA)^{-1}$ gleichmäßig beschränkt ist auf $\mathbb{C}^-$.
Gelegentlich sagen wir auch kürzer: A ist ASI - stabil.
(s, A, b) heißt *AS - stabil*, falls für alle $z \in \mathbb{C}^-$ die Matrix $I - zA$ regulär ist und $b^T z(I - zA)^{-1}$ gleichmäßig beschränkt ist auf $\mathbb{C}^-$.

Dann gilt:
Falls die Eigenwerte von A alle positiven Realteil haben, ist (s, A, b) AS - stabil.
Falls A bis auf einen eventuellen einfachen Eigenwert 0 nur Eigenwerte mit positivem Realteil hat, ist (s, A, b) ASI - stabil.
Wenn es ein $d \in \mathbb{R}^s$ gibt mit $b^T = d^T A$, dann

$$\text{ASI - Stabilität} \quad \Longrightarrow \quad \text{AS - Stabilität.}$$

Wir haben vor Runge - Kutta - Verfahren auf die gemäß Kap. 1 gestörten Index-1 - DAEs anzuwenden. Die numerische Lösung $y_{\nu\varepsilon} \approx y_\varepsilon(t_\nu)$, $z_{\nu\varepsilon} \approx z_\varepsilon(t_\nu)$ (Bezeichnungen wie in

Satz 1.2) wird dann zwei kleine Parameter enthalten, nämlich den Störparameter $\varepsilon > 0$ und die Schrittweite $h > 0$. Betrachtet man ε als fest, so erhält man auf dem üblichen Wege (siehe z. B. [DekVe]) Abschätzungen der Form

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

Daraus kann man - da man ja eigentlich die exakte Lösung $(\bar{y}(t), \bar{z}(t))^T$ der DAE approximieren will - noch

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

herleiten. Die Konstante C enthält aber immer höhere Ableitungen der exakten Lösung $(y_\varepsilon(t), z_\varepsilon(t))^T$ und diese sind nicht beschränkt in ε (denn $z'_\varepsilon(t) = \varepsilon^{-1}(\bar{z}(t) - z_\varepsilon(t))$).
Will man den Approximationsfehler Ch^p kleinhalten, muß man dazu h sehr klein im Vergleich zu ε wählen. Das ist unschön, denn kleine Schrittweiten h bedeuten viele Schritte und damit - gerade bei Runge - Kutta - Verfahren - auch hohen Rechenaufwand. Da wir aber wissen, daß $(y_\varepsilon(t), z_\varepsilon(t))^T$ nahe bei $(\bar{y}(t), \bar{z}(t))^T$ liegt, sollte man meinen, daß mit geeigneten Verfahren auch die numerische Lösung $(y_{\nu\varepsilon}, z_{\nu\varepsilon})^T$ nahe bei $(\bar{y}(t), \bar{z}(t))^T$ liegt. Wir hoffen dabei auf Abschätzungen der Form

$$\| \begin{pmatrix} y_{\nu\varepsilon} \\ z_{\nu\varepsilon} \end{pmatrix} - \begin{pmatrix} \bar{y}(t_\nu) \\ \bar{z}(t_\nu) \end{pmatrix} \| \leq O(\varepsilon) + O(h^p),$$

woraus wir umgekehrt auch wieder

$$\| \begin{pmatrix} y_{\nu\varepsilon} \\ z_{\nu\varepsilon} \end{pmatrix} - \begin{pmatrix} y_\varepsilon(t_\nu) \\ z_\varepsilon(t_\nu) \end{pmatrix} \| \leq O(\varepsilon) + O(h^p) + \underbrace{\| \begin{pmatrix} y_\varepsilon(t_\nu) \\ z_\varepsilon(t_\nu) \end{pmatrix} - \begin{pmatrix} \bar{y}(t_\nu) \\ \bar{z}(t_\nu) \end{pmatrix} \|}_{\to 0 \text{ für } \varepsilon \to 0}$$

erhalten könnten. Wir wollen also versuchen, bei der Herleitung der Fehlerabschätzungen keine Ableitungen von $(y_\varepsilon(t), z_\varepsilon(t))^T$ ins Spiel zu bringen. Der Lohn wäre dann eine Fehlerabschätzung gleichmäßig in $\varepsilon \in [0, \varepsilon_0]$ und $h \in [0, h_0]$ und wir hätten bei der numerischen Rechnung keine Einschränkungen der Form $h < \varepsilon$ (oder ähnliches) zu beachten.
Zuerst werden wir untersuchen, was bei der Durchführung eines einzelnen Schrittes passiert. Das mindeste, was wir fordern müssen, ist, daß das Resultat nach einem Schritt gleichmäßig in ε und h beschränkt ist. Dies führt auf die folgende Begriffsbildung.

4.4 Definition (*D-stabile Runge - Kutta - Verfahren*)
Sei (s, A, b) ein Runge - Kutta - Verfahren, $\mathcal{F} \subseteq C(\mathbb{R}^{2n+1}, \mathbb{R}^n)$, $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$.
$x_1(h, F)$ sei der mit dem Runge - Kutta - Verfahren aus (t_0, x_0) mit einem Schritt der Schrittweite h ($h > 0$) zur Gleichung $F(t, x(t), x'(t)) = 0$ berechnete Wert.
Das Runge - Kutta - Verfahren heißt *D-stabil auf* $\mathcal{F}$: $\iff$ es gibt $M > 0$, $h_0 > 0$ mit

$$\|x_1(h, F)\| \leq M \quad \forall F \in \mathcal{F},\ h \in [0, h_0].$$

Bemerkungen:

1. Beachte, daß das in Definition 4.4 erwähnte M durchaus von h_0, t_0 und x_0 abhängen darf.
2. Bisher wurde D-Stabilität nur für gewöhnliche, explizite Differentialgleichungen definiert (siehe die Arbeiten von van Veldhuizen [vV]). Man ([vV], [DekVe], [Ha]) betrachtete dabei

$$y'(t) = E(t)\begin{pmatrix} \varepsilon^{-1}d_1(t) & 0 \\ 0 & d_2(t) \end{pmatrix} E(t)^{-1}y(t), \qquad y(0) = y_0,$$

wobei $t \in [0,T]$, $y(t) \in \mathbb{R}^2$, $E(t) \in \mathbb{R}^{2\times 2}$, $Re\, d_1(t) \leq \tilde{d}_1 < 0$, und d_1, d_2 und E sind genügend oft stetig differenzierbar und dürfen manchmal auch noch von ε abhängen und zwar auch differenzierbar auf $]0, \varepsilon_0]$.
Das obige System wird dann als Testproblem benutzt und D - Stabilität für dieses Problem als notwendiges, aber nicht unbedingt hinreichendes Kriterium für die Brauchbarkeit eines Verfahrens für allgemeinere Systeme mit kleinen Parametern angesehen.

4.5 Satz:

Vor.:
(s, A, b) sei A -, AS -, und ASI - stabiles Runge - Kutta - Verfahren mit A regulär, $C(q)$ und $B(p)$; $d > t_0$

$$y' = f(y,z), \quad 0 = g(y,z), \quad y(t_0) = y_0,\ z(t_0) = z_0$$

mit $f \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, $g \in C^1(\mathbb{R}^n, \mathbb{R}^k)$, $m + k = n$ sei Index-1 - DAE mit exakter Lösung
$(\bar{y}(t), \bar{z}(t))^T \in C^{max\{p+1,q+1\}}([t_0, d], \mathbb{R}^n)$; (insb. $g(y_0, z_0) = 0$),
$M \in \mathbb{R}^{k\times k}$ mit $(Mx, x) \geq 0$ für alle $x \in \mathbb{R}^k$ für ein Skalarprodukt $(.,.)$ auf $\mathbb{R}^k$, $\varepsilon_0 > 0$,
$\delta > 0$, $\|.\|$ die zu $(.,.)$ gehörige Norm,
$S_1 := \bar{B}(\bar{y}([t_0, d]), \delta)$, $S_2 := \bar{B}(\bar{z}([t_0, d]), \delta)$, $G \in C^1(S_1, S_2)$ bijektiv mit

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

Beh.:
(s, A, b) ist D-stabil auf

$$\mathcal{F} = \left\{ \begin{pmatrix} y' - f(y, M^{-1}(Mz + \varepsilon z')) \\ g(y, M^{-1}(Mz + \varepsilon z')) \end{pmatrix} \,\middle|\, \varepsilon \in [0, \varepsilon_0] \right\}$$

Beweis:
Wir wollen zuerst einen Schritt mit dem Verfahren angewandt auf die DAE und ausgehend von $(y_0, \tilde{z}_0) \approx (y_0, z_0)$ rechnen. Dabei müssen wir die Stufenwerte Y_i, Z_i, $i = 1, \ldots, s$ gemäß

$$(4.5.1) \qquad \begin{aligned} Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_j, Z_j) \\ 0 &= g(Y_i, Z_i) \end{aligned} \qquad i = 1, \ldots, s$$

berechnen. Sofern wir uns nur für $Y_i \in S_1$, $Z_i \in S_2$ $(i = 1, \ldots, s)$ interessieren, ist dies gleichbedeutend mit

$$(4.5.2) \qquad \begin{aligned} Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_j, G(Y_j)) \\ Z_i &= G(Y_i). \end{aligned} \qquad i = 1, \ldots, s$$

Auf dieses System wollen wir den Satz über implizite Funktionen (siehe z. B. [Dei],[OrRh]) anwenden, um von h stetig abhängige Lösungen $Y_i(h)$ und damit auch $Z_i(h)$ zu bekommen.
Zu lösen ist also

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

wobei

$$\begin{aligned} F(h, \vec{Y}) &:= (F_1(h, \vec{Y}), \ldots, F_s(h, \vec{Y}))^T \\ F_i(h, \vec{Y}) &:= Y_i - y_0 - h\sum_{j=1}^{s} a_{ij} f(Y_j, G(Y_j)) \qquad \text{für } i = 1, \ldots, s. \end{aligned}$$

Nun ist

$$F(0, \begin{pmatrix} y_0 \\ \vdots \\ y_0 \end{pmatrix}) = 0, \qquad F_{\vec{Y}}(0, \vec{Y}) = I_{ms} \quad \text{für alle } \vec{Y} \in \mathbb{R}^{ms},$$

so daß wir wie gewünscht auflösen können, d. h. es gibt $h_0 > 0$ und $\vec{Y} \in C([0, h_0], \mathbb{R}^{ms})$ mit

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

Nun ist wegen

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i f(Y_i, G(Y_i))$$

klar, daß es $C_y > 0$ gibt mit

$$\|y_1\| \leq C_y \qquad \text{für alle } h \in [0, h_0].$$

Beachte, daß y_1 nicht von ε abhängt.

Es bleibt nun die analoge Aussage für z_1 zu zeigen.
Wir setzen vorher $M_\varepsilon := \frac{1}{\varepsilon}M$.
Ein Schritt mit dem Verfahren für das gestörte System erfordert nun Stufenwerte $Y_{i\varepsilon}, Z_{i\varepsilon}$ mit

$$\left.\begin{aligned} Y_{i\varepsilon} &= y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_{j\varepsilon}, Z_{j\varepsilon} + M_\varepsilon^{-1} Z'_{j\varepsilon}) \\ 0 &= g(Y_{i\varepsilon}, Z_{i\varepsilon} + M_\varepsilon^{-1} Z'_{i\varepsilon}) \end{aligned}\right. \qquad i = 1, \ldots, s$$

wobei wie üblich $Z'_{i\varepsilon}$ implizit definiert ist durch

$$Z_{i\varepsilon} = \tilde{z}_0 + h\sum_{j=1}^{s} a_{ij} Z'_{j\varepsilon} \qquad i = 1, \ldots, s \tag{4.5.3}$$

(vergleiche die Bemerkung nach Definition 4.2).
Wir können also $Y_{i\varepsilon},\ Z_{i\varepsilon}$ aus $Y_i,\ Z_i$ wie folgt berechnen:

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

und definieren $Z_{i\varepsilon}$ als Lösung von

$$\begin{aligned} Z_{i\varepsilon} + M_\varepsilon^{-1} Z'_{i\varepsilon} &= Z_i, & i &= 1, \ldots, s \\ \Longleftrightarrow \qquad Z'_{i\varepsilon} &= M_\varepsilon(Z_i - Z_{i\varepsilon}), & i &= 1, \ldots, s. \end{aligned} \tag{4.5.4}$$

Setzt man nun

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

so führen (4.5.3) und (4.5.4) auf

$$\vec{Z}_\varepsilon = E\tilde{z}_0 + h(A \otimes I_k)(I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon). \tag{4.5.5}$$

Setzt man nun

$$\bar{Z}_0 := \begin{pmatrix} \bar{z}(t_0 + c_1 h) \\ \vdots \\ \bar{z}(t_0 + c_s h) \end{pmatrix}, \quad \bar{Z}'_0 := \begin{pmatrix} \bar{z}'(t_0 + c_1 h) \\ \vdots \\ \bar{z}'(t_0 + c_s h) \end{pmatrix}, b^T \otimes I_k = (\, b_1 \cdot I_k \ldots b_s \cdot I_k \,) \in \mathbb{R}^{k\times ks}$$

und analog $c^T \otimes I_k$, so erhalten wir daraus mit 4.2,(iv)

$$\begin{aligned} \bar{Z}_0 &= E\bar{z}(t_0) + h(A \otimes I_k)\bar{Z}'_0 + O(h^{q+1}) \\ \bar{z}(t_0 + h) &= \bar{z}(t_0) + h(b^T \otimes I_k)\bar{Z}'_0 + O(h^{p+1}). \end{aligned} \tag{4.5.6}$$

Außerdem berechnet sich ja der Näherungswert $\tilde{z}_1$, der bei Anwendung auf das gestörte Problem ausgehend von $\tilde{z}_0$ entsteht, zu

$$\tilde{z}_1 = \tilde{z}_0 + h(b^T \otimes I_k)(I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon). \tag{4.5.7}$$

Für die Fehler $e_1 := \tilde{z}_1 - \bar{z}(t_0 + h)$ und $e_0 := \tilde{z}_0 - z_0 = \tilde{z}_0 - \bar{z}(t_0)$ erhalten wir dann aus (4.5.6) und (4.5.7):

$$e_1 = e_0 + h(b^T \otimes I_k)((I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}_0') + O(h^{p+1}). \tag{4.5.8}$$

Außerdem liefert (4.5.5) mit (4.5.6)

$$\vec{Z}_\varepsilon - \bar{Z}_0 = Ee_0 + h(A \otimes I_k)((I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}_0') + O(h^{q+1}). \tag{4.5.9}$$

Um dieses in (4.5.8) verwerten zu können, fehlt uns noch ein Zusammenhang zwischen $\bar{Z}_0$ und $\vec{Z}_0$. Dazu zeigen wir nun folgende Aussage:

Sei $\alpha := \max\limits_{i=1,\ldots,s} \sum\limits_{j=1}^{s} |a_{ij}| > 0$, L_f und L_G seien Lipschitzkonstanten für $f(.,G(.))$ bzw. G auf S_1. Dann gilt für $i = 1, \ldots, s$

$$h < \frac{1}{L_f \alpha} \Longrightarrow \|Z_i - \bar{z}(t_0 + c_i h)\| \le \frac{L_G}{1 - hL_f\alpha} \left[\|y_0 - \bar{y}(t_0)\| + O(h^{q+1}) \right]. \tag{4.5.10}$$

Beweis dazu:
Sei $l \in \{1, \ldots, s\}$ so gewählt, daß

$$\|Y_l - \bar{y}(t_0 + c_l h)\| = \max_{j=1,\ldots,s} \|Y_j - \bar{y}(t_0 + c_j h)\|$$

Nach 4.2,(iv) hat man zunächst

$$\begin{aligned}\bar{y}(t_0 + c_l h) &= \bar{y}(t_0) + h \sum_{j=1}^{s} a_{lj} \bar{y}'(t_0 + c_j h) + O(h^{q+1}) \\ &= \bar{y}(t_0) + h \sum_{j=1}^{s} a_{lj} f(\bar{y}(t_0 + c_j h), G(\bar{y}(t_0 + c_j h))) + O(h^{q+1}).\end{aligned}$$

Zusammen mit (4.5.2) (mit $i = l$) folgt dann unter Verwendung von Dreiecksungleichung und Lipschitzbedingung für $f(.,G(.))$

$$\|Y_l - \bar{y}(t_0 + c_l h)\| \le \|y_0 - \bar{y}(t_0)\| + hL_f\alpha \|Y_l - \bar{y}(t_0 + c_l h)\| + O(h^{q+1}).$$

Sei nun $i \in \{1, \ldots, s\}$. Es gilt:

$$\begin{aligned}\|Z_i - \bar{z}(t_0 + c_i h)\| &= \|G(Y_i) - G(\bar{y}(t_0 + c_i h))\| \le L_G \|Y_i - \bar{y}(t_0 + c_i h)\| \\ &\le L_G \|Y_l - \bar{y}(t_0 + c_l h)\| \\ &\le \frac{L_G}{1 - hL_f\alpha} \|y_0 - \bar{y}(t_0)\| + O(h^{q+1}).\end{aligned}$$

Damit ist (4.5.10) vollständig bewiesen.

O. B. d. A. sei das oben gewählte $h_0 < \frac{1}{L_f \alpha}$. In unserem Fall ist nach Voraussetzung $y_0 = \bar{y}(t_0)$ (aber dieses wird hier zum ersten Mal benutzt!), so daß wir sofort erkennen, daß

$$\bar{Z}_0 = \vec{Z}_0 + O(h^{q+1})$$

gilt. (Hierbei kann man sich $\mathbb{R}^{ks}$ mit

$$|||\vec{X}||| := \max_{i=1,\ldots,s} \|X_i\|$$

normiert denken. Aber jede andere Norm tut es auch, da es auf Konstanten, die aus der Normäquivalenz herrühren, nicht ankommt.)
Und das setzen wir nun in (4.5.9) ein und erhalten

$$\vec{Z}_\varepsilon - \vec{Z}_0 = Ee_0 + h(A \otimes I_k)((I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}_0') + O(h^{q+1}).$$

Das bedeutet aber nichts anderes als

$$\begin{aligned} & (I_{ks} + h(A \otimes M_\varepsilon))(\vec{Z}_\varepsilon - \vec{Z}_0) = Ee_0 - h(A \otimes I_k)\bar{Z}_0' + O(h^{q+1}) \\ \Longleftrightarrow \quad & (I_s \otimes M_\varepsilon)(\vec{Z}_0 - \vec{Z}_\varepsilon) - \bar{Z}_0' = \\ & \quad (I_s \otimes M_\varepsilon)(I_{ks} + h(A \otimes M_\varepsilon))^{-1} \left[h(A \otimes I_k)\bar{Z}_0' \right. \\ & \quad \left. -Ee_0 + O(h^{q+1}) - (I_{ks} + h(A \otimes M_\varepsilon))(I_s \otimes M_\varepsilon)^{-1}\bar{Z}_0'\right] \\ & = (I_s \otimes M_\varepsilon)(I_{ks} + h(A \otimes M_\varepsilon))^{-1}\left[-Ee_0 - (I_s \otimes M_\varepsilon)^{-1}\bar{Z}_0' + O(h^{q+1})\right]. \end{aligned}$$

Das in (4.5.8) eingesetzt liefert

$$\begin{aligned} e_1 = e_0 + h(b^T \otimes I_k)(I_s \otimes M_\varepsilon)(I_{ks} + h(A \otimes M_\varepsilon))^{-1} \cdot \\ \cdot\left[-Ee_0 - (I_s \otimes M_\varepsilon)^{-1}\bar{Z}_0' + O(h^{q+1})\right] + O(h^{p+1}). \end{aligned} \tag{4.5.11}$$

Setzen wir noch

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

so erhalten wir aus (4.5.11)

$$e_1 = [I_k - N(h,\varepsilon)E]\, e_0 - N(h,\varepsilon)\left[(I_s \otimes M_\varepsilon)^{-1}\bar{Z}_0' + O(h^{q+1})\right] + O(h^{p+1}). \tag{4.5.12}$$

Um den Beweis abzuschließen reicht es die bezüglich h und ε gleichmäßige Beschränktheit von e_1 nachzuweisen.
Wir zitieren nun einfach zwei Lemmata aus [BuHuVe] (3.6, 3.7), die uns dem Ziel entscheidend näher bringen:
A - Stabilität und ASI - Stabilität $\Longrightarrow$

$$\|I_k - N(h,\varepsilon)E\| \leq 1 + \gamma_2 h \tag{4.5.13}$$

AS - Stabilität und ASI - Stabilität $\Longrightarrow$

$$\|N(h,\varepsilon)x\| \leq \gamma_3\|x\| \tag{4.5.14}$$

für alle $x \in \mathbb{R}^k$ und Konstanten γ_2, γ_3 und das alles für h hinreichend klein.

(Anmerkung: Um aus den zitierten Lemmata auf (4.5.13) und (4.5.14) zu gelangen, muß man die dort gegebenen Aussagen auf die Differentialgleichung $U' = -M_\varepsilon U$ anwenden. Wegen der Voraussetzung an M ist auch $(M_\varepsilon x, x) \geq 0$, so daß die Anwendung der zitierten Lemmata möglich ist. Siehe auch Bemerkung 4.8.3.)

Und dies liefert uns nun die Behauptung, denn alle in (4.5.12) auftretenden Terme sind offensichtlich beschränkt in h und in ε. ■

Bemerkungen:

1. Die Voraussetzungen von Satz 4.5 sind z. B. für die Gauss - Legendre, Radau-IA -, Radau-IIA - und Lobatto-IIIC - Formeln erfüllt ([Ha]).
2. Im Fall, daß man das euklidische Skalarprodukt gewählt hat, folgt aus $(Mx, x) \geq 0$ für alle x, daß die Eigenwerte von M alle Realteil ≥ 0 haben. Denn:
Sei $x \in \mathbb{R}^k$, $x \neq 0$, $\lambda \in \mathbb{C}$ mit $Mx = \lambda x$. Mit ($x^H := (\bar{x}_1, \ldots, \bar{x}_s)^T$, komponentenweise konjugiert komplex) ist dann auch $x^H M^T = \bar{\lambda} x^H$. Und damit

$$0 \leq (Mx, x) + (x, M^T x) = x^H \lambda x + \bar{\lambda} x^H x = 2 Re \lambda x^H x.$$

Wir haben im Beweis zu 4.5 praktisch eine Darstellung des lokalen Fehlers hergeleitet. Da drängt sich sofort die Frage auf, ob es ähnliches auch für den globalen Fehler gibt. Die Antwort darauf ist

4.6 Satz:

Vor.:
Wie in Satz 4.5; zusätzlich erfülle das Runge - Kutta - Verfahren noch $A(l)$ *mit* $l \geq q+1$, $t > t_0$, $h_0 > 0$
für $h \leq h_0$ *und* $\nu \in \mathbb{N}$ *mit* $t = t_0 + \nu h$ *seien* $(y_i^{(\nu)}, z_i^{(\nu)})^T$, $i = 1, \ldots, \nu$ *die mit dem Verfahren mit der Schrittweite* $h > 0$, *angewandt auf*

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

und ausgehend von $(y_0, \tilde{z}_0) \in S_1 \times S_2$ *berechneten Näherungen (also Startwert in der* y *- Komponente exakt und in der* z *- Komponente gestört); es sei*

$$(y_i^{(\nu)}, z_i^{(\nu)})^T \in S_1 \times S_2, \qquad i = 1, \dots, \nu$$

(Werte verlassen den Schlauch nicht), $y_\nu := y_\nu^{(\nu)}$, $z_\nu := z_\nu^{(\nu)}$, $e_0 := \tilde{z}_0 - \bar{z}(t_0)$

Beh.:

Es gilt

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

Beweis:

Aus dem Beweis zu 4.5 sehen wir, daß in der y - Komponente nichts anderes geschieht, als daß die gewöhnliche Differentialgleichung $y' = f(y, G(y))$ mit dem Verfahren behandelt wird. Insofern ist sofort klar, daß in dieser Komponente die gleiche Konvergenzordnung wie bei gewöhnlichen Differentialgleichungen erreicht wird. Und das ist nach Voraussetzung l. Wir haben also nur noch den Fehler in der z- Komponente zu untersuchen. Dazu wenden wir im Prinzip Satz 4.5 ν - mal an, für jeden Schritt einmal. Dort haben wir ja einen Schritt von t_0 nach $t_1 = t_0 + h$ gerechnet. Das gleiche erhält man, wenn man von $t_j = t_0 + jh$ nach $t_{j+1} = t_j + h$ rechnet. Dabei läßt sich das meiste genauso übernehmen; bei der Verwendung von (4.5.10) ist jedoch zu beachten, daß diesmal nicht $y_j = \bar{y}(t_j)$ gilt, da wir für $j \geq 1$ schon Diskretisierungsfehler gemacht haben. Jedoch ist der Fehler in der y- Komponente

$$y_j = \bar{y}(t_j) + O(h^l) \qquad \text{wegen } A(l).$$

Und so folgt (vgl. Beweis zu 4.5)

$$\bar{Z}_j = \vec{Z}_j + O(h^{q+1}) + O(h^l) = \vec{Z}_j + O(h^{q+1}) \qquad \text{da n. V. } l \geq q+1.$$

Die Rechnungen im Beweis von 4.5 werden also nicht beeinflußt und man erhält schließlich das Analogon zur Rekursionsformel (4.5.12)

$$\text{(4.6.1)} \qquad e_{j+1} = [I_k - N(h,\varepsilon)E]\, e_j - N(h,\varepsilon) \left[(I_s \otimes M_\varepsilon)^{-1} \bar{Z}_j' + O(h^{q+1}) \right] + O(h^{p+1}).$$

Nun ist noch

$$\bar{Z}_j' = E\bar{z}'(t_j) + O(h)$$

und damit

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

Damit läßt sich (4.6.1) schreiben als

$$e_{j+1} = [I_k - N(h,\varepsilon)E]\, e_j - N(h,\varepsilon) \left[\varepsilon E O(1) + O(\varepsilon h) + O(h^{q+1}) \right] + O(h^{p+1}).$$

Mit vollständiger Induktion erhalten wir

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

Zusammen mit (4.5.13) und (4.5.14) folgt

$$\begin{aligned}
\|e_\nu\| &\leq (1+\gamma_2 h)^\nu \|e_0\| + [(1+\gamma_2 h)^\nu + 1]\, O(\varepsilon) \\
&\qquad + \sum_{j=0}^{\nu-1} (1+\gamma_2 h)^j \left[\gamma_3(O(\varepsilon h) + O(h^{q+1})) + O(h^{p+1})\right] \\
&\leq e^{\gamma_2 h\nu}\|e_0\| + (e^{\gamma_2 h\nu} + 1)O(\varepsilon) + \nu e^{\gamma_2 h\nu}\left[O(\varepsilon h) + O(h^{q+1}) + O(h^{p+1})\right].
\end{aligned}$$

Nun ist aber $\nu h = t - t_0$ und damit unabhängig von h und ν, so daß wir insgesamt schließlich

$$\|e_\nu\| \leq e^{\gamma_2(t-t_0)}\|e_0\| + (e^{\gamma_2(t-t_0)} + 1)O(\varepsilon) + e^{\gamma_2(t-t_0)}\left[O(\varepsilon) + O(h^q) + O(h^p)\right]$$

erhalten und damit ist alles klar. ■

Nachdem diese Vorarbeit erst einmal geleistet ist, fällt uns nun auch leicht eine Multiparameterversion in die Hände:

4.7 Satz:

Vor.:
Wie in Satz 4.6, nur wird diesmal in der DAE ε durch $\mathcal{E}$ ersetzt, wobei $\mathcal{E} := diag(\varepsilon_1, \ldots, \varepsilon_k)$ mit $\varepsilon_i > 0$ für $i = 1, \ldots, k$; M sei Diagonalmatrix; $(.,.)$ das euklidische Skalarprodukt

Beh.:

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

Beweis:
Wir setzen $M_\varepsilon := \mathcal{E}^{-1} \cdot M = M \cdot \mathcal{E}^{-1}$ und stellen fest, daß für alle $x \in \mathbb{R}^k$:

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

denn nach Voraussetzung ist für alle $y \in \mathbb{R}^k$ $0 \leq (My, y) = \sum_{i=1}^{k} m_{ii} y_i^2$, was $m_{ii} \geq 0$ für $i = 1, \ldots, k$ zur Folge hat. Mit diesem M_ε kann der Beweis von 4.6 einfach übernommen werden; zu beachten ist dabei nur

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

Alles andere folgt dann wie gehabt. ■

4.8 <u>Bemerkungen:</u>

1. Wir erhalten also als gleichmäßige Konvergenzordnung für semi - explizite regularisierte Index-1 - DAEs in der z - Komponente:

$$C(q),\ B(p) \Longrightarrow Konvergenzordnung = min\{p, q\}.$$

Die Konvergenzordnung in der y - Komponente ist ja die gleiche wie bei gewöhnlichen Differentialgleichungen, also l, falls $A(l)$ gegeben ist.
Bekanntlich [DekVe] gilt nun bei Stufenzahl s

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

so daß man für diese Verfahren gleichmäßige Konvergenz in ε hat, falls $s \geq 2$. Bezüglich der einstufigen Verfahren läßt sich folgendes beobachten:

die Gauss - Formel mit $s = 1$ ist gleichmäßig konvergent mit Ordnung 1 in h (Satz 4.6),
zur Radau-IA - Formel mit $s = 1$:
diese erfüllt nicht C(1), was der Fall wäre, wenn die c_i gemäß unserer Definition 4.2 gesetzt wären. Diese Definition der c_i geht aber an keiner Stelle des Beweises explizit ein (höchstens als $C(1)$, wenn dies gegeben ist), so daß der Satz gültig bleibt, aber nur auf gleichmäßige "Konvergenz der Ordnung 0" in h schließen läßt, d. h. gleichmäßige Beschränktheit;
zur Radau-IIA - Formel mit $s = 1$:
dies ist das implizite Euler - Verfahren; es erfüllt zwar $C(1)$, aber ansonsten gilt das oben zur einstufigen Radau-IA - Formel gesagte;
zur Lobatto-IIIC - Formel:
diese Formeln gibt es per Definition nur mit $s \geq 2$.

Insgesamt erhalten wir für die gleichmäßige Konvergenzordnung der oben genannten s - stufigen Verfahren:

$$\begin{aligned} \text{für die Gauss - Formeln} &: s \\ \text{für die Radau-IA - Formeln} &: s-1 \\ \text{für die Radau-IIA - Formeln} &: s \quad \text{falls } s \geq 2; \quad 0 \text{ falls } s=1 \\ \text{für die Lobatto-IIIC - Formeln} &: s-1 \quad \text{stets ist } s \geq 2 \end{aligned}$$

Die gleiche Konvergenzordnung gilt für die ungestörten Index-1 - DAEs, denn für $\varepsilon = 0$ bleibt ja alles richtig. Diese Ordnung ist für DAEs aber nicht optimal, wie wir nach der Lektüre von [Pe3] und [BuPe] wissen. Z. B. hat das implizite Euler - Verfahren Konvergenzordnung 1 für Index-1 - DAEs (siehe z. B. [BuPe], [Pe3], [GriMä]). Zum Nachweis einer in ε gleichmäßigen höheren Konvergenzordnung (wobei unklar ist, ob das überhaupt möglich ist), müßte man die Terme $O(h^{q+1})$ und $O(h^l)$ in obigem Beweis genauer untersuchen. Wir wollen an dieser Stelle jedoch darauf verzichten.

2. In [BuHuVe] wird eine gleichmäßige Konvergenzordnung für die Klasse von Differentialgleichungen

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

wobei

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

hergeleitet (gleichmäßig in α und β). Der Beweis von Satz 4.6 folgt der dort verwendeten Idee, und wir erhalten ja auch (genau wie dort) Ordnung $p = q$, falls $B(p)$, $C(p)$ gegeben sind. Unter einer Zusatzvoraussetzung wird in [BuHuVe] mit $B(p+1)$, $C(p)$ sogar auf Ordnung $p+1$ geschlossen; der dortige Beweis läßt sich jedoch nicht ohne weiteres auf unsere Situation übertragen, da die $O(h^{q+1})$ Terme sich anders als unsere zusammensetzen.

3. Vergleich von Satz 4.5 mit einem früher erzielten Resultat von Griepentrog [Gri]:

Dort werden Systeme der Form

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

betrachtet, bei denen die Eigenwerte der konstanten Matrix M alle negativen Realteil haben. Hergeleitet wird die gleichmäßige (in ε) Konvergenz von Einschrittverfahren für dieses System.

Ein Vergleich mit unserer gestörten Gleichung ergibt folgendes:

Die Regularisierung in $(*)$ ist im Gegensatz zu unserer linear in z'. Andererseits steckt hinter unserer letzten Endes auch ein (noch einfacherer) linearer Ansatz, dessen Linearität sich aber erst durch die Index-1 - Annahme bemerkbar macht. Beim Beweis von 4.6 wird daher praktisch eine äußerst einfache lineare, aber *diskrete* Differentialgleichung gelöst. Die stetige Abhängigkeit der ganzen rechten Seite von ε stellt keinen entscheidenden Unterschied dar; der Beweis von 4.6 läßt sich leicht entsprechend modifizieren.

Was bringt nun ein Vergleich der Resultate?
Griepentrog [Gri] weist gleichmäßige Konvergenz für alle L - stabilen Runge - Kutta - Verfahren nach. Das sind solche, die A-stabil sind und für die

$$\lim_{z \to -\infty} R(z) = 0 \qquad \text{mit } R(z) := 1 + z b^T (I_s - zA)^{-1} e$$

gilt. (R bezeichnet man als Stabilitätsfunktion des Runge - Kutta - Verfahrens.)
Es wird aber keine Fehlerdarstellung gegeben.
Für unsere Regularisierung haben wir eine Fehlerdarstellung und Ordnungsbedingungen angegeben. Außerdem sind unsere Voraussetzungen an die Runge - Kutta - Verfahren in gewisser Weise weniger einschränkend, denn die in der Praxis gebräuchlichen Runge - Kutta - Verfahren sind sowohl A - als auch AS - und ASI - stabil, müssen aber nicht zwangsläufig auch L - stabil sein. Dazu folgendes Beispiel aus [Pe3], [HW]:
Betrachte das semi - explizite Runge - Kutta - Verfahren definiert durch

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

Dies ist A - stabil [HW].
Die Eigenwerte von A haben natürlich positiven Realteil, so daß wegen
4.3 das Verfahren AS - und ASI - stabil ist. Jedoch ist es nicht L - stabil, denn

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

Satz 4.6 ist also anwendbar, nicht aber der Satz in [Gri].
Unser Verfahren erfüllt $B(4)$, $C(1)$, $A(3)$ und mit Satz 4.6 ist klar, daß die gleichmäßige Konvergenzordnung in der y - Komponente 3 ist und in der z - Komponente (mindestens) 1.

Nun wenden wir uns den in Kapitel 1 vorgestellten nichtlinearen Regularisierungen zu. Wir wollen unter ähnlichen Voraussetzungen wie in Satz 1.3 D - Stabilität für Runge - Kutta - Verfahren nachweisen. Dazu brauchen wir auch noch einige Eigenschaften der verwendeten Funktionen.

4.9 Definition (*monotone Operatoren*)
Sei H Hilbertraum mit Skalarprodukt $(.,.)$, $\|.\|$ die vom Skalarprodukt induzierte Norm, $D \subseteq H$, $F : D \to H$.
F heißt *stark monoton auf* $D :\iff$ es gibt ein $c > 0$ mit

$$(F(x) - F(y), x - y) \geq c\|x - y\|^2 \qquad \text{für alle } x, y \in D.$$

4.10 Satz:

Vor.:
H Hilbertraum, $F : H \to H$, F stetig und stark monoton

Beh.:
F ist Homöomorphismus

Beweis:
Siehe [OrRh], Theorem 6.4.4. (für den Fall $\mathbb{R}^n$) und [Dei], Theorem 11.2 (für den allgemeinen Fall).

Nun noch einige Notationen für das Rechnen in $\mathbb{R}^{ns}$.

4.11 Definition/Lemma: [DekVe] (*Skalarprodukte in* $\mathbb{R}^{ns}$)
Sei $D = diag(d_1, \ldots, d_n)$ mit $d_i > 0$ für $i = 1, \ldots, n$. $(.,.)$ sei ein Skalarprodukt auf $\mathbb{R}^n$, und

$$X := \begin{pmatrix} X_1 \\ \vdots \\ X_s \end{pmatrix} \in \mathbb{R}^{ns}, \quad X_i := \begin{pmatrix} X_{i1} \\ \vdots \\ X_{in} \end{pmatrix} \in \mathbb{R}^n \text{ für } i = 1, \ldots, s$$

$$X^{(l)} := \begin{pmatrix} X_{1l} \\ \vdots \\ X_{sl} \end{pmatrix} \in \mathbb{R}^s \quad l = 1, \ldots, n \qquad \text{(alle } l \text{ -ten Komponenten zusammengefaßt)}$$

und $Y \in \mathbb{R}^{ns}$ analog. Sei $(.,.)$ das euklidische Skalarprodukt und $\|.\|$ die zugehörige Norm. Dann definieren wir ein Skalarprodukt $[.,.]_D$ und eine Norm $\|.\|_D$ auf $\mathbb{R}^{ns}$ durch

$$[X, Y]_D := \sum_{i=1}^{n} d_i (X_i, Y_i), \qquad \|X\|_D := \sqrt{\sum_{i=1}^{n} d_i \|X_i\|^2}.$$

Sei $A \in \mathbb{R}^{s \times s}$. Dann gilt für alle $X \in \mathbb{R}^{ns}$:

$$(4.11.1) \qquad [(A \otimes I)X, X]_D = \frac{1}{2} \sum_{l=1}^{n} (X^{(l)})^T (DA + A^T D) X^{(l)}$$

und mit $\psi_D(A) := \min\limits_{x \neq 0} \frac{(DAx, x)}{(x, x)}$ für A regulär:

$$(4.11.2) \qquad \begin{aligned} DA + A^T D \text{ positiv semidefinit } &\iff \psi_D(A) \geq 0 \\ \iff \psi_D(A^{-1}) \geq 0 &\iff DA^{-1} + (A^T)^{-1} D \text{ positiv semidefinit} \end{aligned}$$

Beweis:
Für einen Vektor Y sei $(Y)_i$ die i-te Komponente von Y. (4.11.1) erhält man wie folgt:

$$\begin{aligned}
[(A\otimes I)X,X]_D &= \sum_{i=1}^{s} d_i \sum_{j=1}^{s} a_{ij}(X_j,X_i) &&= \sum_{i=1}^{s} d_i \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{n} X_{jl}X_{il} \\
&= \sum_{l=1}^{n}\sum_{i=1}^{s} d_i X_{il} \sum_{j=1}^{s} a_{ij}X_{jl} &&= \sum_{l=1}^{n}\sum_{i=1}^{s} (DX^{(l)})_i (AX^{(l)})_i \\
&= \sum_{l=1}^{n} (DX^{(l)})^T AX^{(l)} &&= \sum_{l=1}^{n} (X^{(l)})^T D^T AX^{(l)} \\
&= \tfrac{1}{2}\sum_{l=1}^{n} (X^{(l)})^T D^T AX^{(l)} + (X^{(l)})^T A^T DX^{(l)} \\
&= \tfrac{1}{2}\sum_{l=1}^{n} (X^{(l)})^T (DA + A^T D)X^{(l)}.
\end{aligned}$$

Zu (4.11.2): Die erste und die letzte Äquivalenz sind klar und die zweite findet man in [DekVe], Corollary 5.1.4..

Nun sind wir bereit für eine D - Stabilitätsaussage für nichtlineare Regularisierungen.

4.12 Satz:

Vor.:
$y' = f(y,z)$, $0 = g(y,z)$, $y(0) = y_0$, $z(0) = z_0$ sei Index-1 - DAE wie in 4.5, $g(y_0,z_0) = 0$,
(s,A,b) Runge - Kutta - Verfahren mit A regulär und
es gibt eine Diagonalmatrix D mit positiven Diagonalelementen und

$$DA + A^T D \text{ ist positiv semidefinit;}$$

$F : \mathbb{R}^k \to \mathbb{R}^k$ stetig, bijektiv und stark monoton mit Monotoniekonstante c_F bezüglich des euklidischen Skalarprodukts $(.,.)$, $\|.\|$ sei die zugehörige Norm

Beh.:
Das Runge - Kutta - Verfahren ist D-stabil auf

$$\mathcal{F} = \left\{ \begin{pmatrix} y' - f(y, F^{-1}(F(z) + \varepsilon z')) \\ g(y, F^{-1}(F(z) + \varepsilon z')) \end{pmatrix} \,\middle|\, \varepsilon > 0 \right\}$$

<u>Beweis:</u> (Vgl. Beweis zu 4.5)
Das Runge - Kutta - Verfahren liefere ausgehend von den Startwerten $y_0,\ z_0$ für die DAE die Stufenwerte $Y_i,\ Z_i,\ i = 1, \ldots, s$.
Anwendung auf ein Element aus $\mathcal{F}$ führt bei Startwerten $y_0,\ \tilde{z}_0$ auf (für alle $i = 1, \ldots, s$)

$$\begin{aligned} Y_{i\varepsilon} &= Y_i \\ F^{-1}(F(Z_{i\varepsilon}) + \varepsilon Z'_{i\varepsilon}) &= Z_i \\ \Longleftrightarrow \qquad F(Z_{i\varepsilon}) + \varepsilon Z'_{i\varepsilon} &= F(Z_i). \end{aligned}$$

Nun ist wie immer

$$\vec{Z}'_\varepsilon := \begin{pmatrix} Z'_{1\varepsilon} \\ \vdots \\ Z'_{s\varepsilon} \end{pmatrix} = \frac{1}{h}(A^{-1} \otimes I_k)(\vec{Z}_\varepsilon - e \otimes \tilde{z}_0).$$

Also ist eine Nullstelle von

$$G(\vec{Z}_\varepsilon) := \underbrace{\begin{pmatrix} F(Z_{1\varepsilon}) \\ \vdots \\ F(Z_{s\varepsilon}) \end{pmatrix}}_{=: \vec{F}(\vec{Z}_\varepsilon)} + \frac{\varepsilon}{h}(A^{-1} \otimes I_k)(\vec{Z}_\varepsilon - e \otimes \tilde{z}_0) - \underbrace{\begin{pmatrix} F(Z_1) \\ \vdots \\ F(Z_s) \end{pmatrix}}_{= \vec{F}(\vec{Z}_0).}$$

zu suchen.
Wir folgen nun dem Beweis von Theorem 5.3.12 in [DekVe].
Es ist für alle $X, Y \in \mathbb{R}^{ks}$

$$\begin{aligned} [G(X) - G(Y), X - Y]_D = [\vec{F}(X) - \vec{F}(Y), X - Y]_D &+ \underbrace{\frac{\varepsilon}{h}}_{\geq 0} \underbrace{[(A^{-1} \otimes I_k)(X - Y), X - Y]_D}_{\geq 0 \ wegen\ 4.11} \\ \geq [\vec{F}(X) - \vec{F}(Y), X - Y]_D &= \sum_{i=1}^{s} d_i (F(X_i) - F(Y_i), X_i - Y_i) \\ \geq c_F \|X - Y\|_D^2. \end{aligned}$$

Nach Satz 4.10 hat G dann genau eine Nullstelle $\vec{Z}_\varepsilon \in \mathbb{R}^{ks}$. Sei $X_0 := e \otimes \tilde{z}_0$. Dann ist X_0 unabhängig von h und ε und es gilt $G(X_0) = \vec{F}(X_0) - \vec{F}(\vec{Z}_0)$ und damit

$$\begin{aligned} c_F \|\vec{Z}_\varepsilon - X_0\|_D^2 &\leq [G(\vec{Z}_\varepsilon) - G(X_0), \vec{Z}_\varepsilon - X_0]_D \\ &= -[\vec{F}(X_0) - \vec{F}(\vec{Z}_0), \vec{Z}_\varepsilon - X_0]_D \\ &\leq \|\vec{F}(X_0) - \vec{F}(\vec{Z}_0)\|_D \cdot \|\vec{Z}_\varepsilon - X_0\|_D \\ \Rightarrow \qquad \|\vec{Z}_\varepsilon - X_0\|_D &\leq \frac{1}{c_F} \|\vec{F}(X_0) - \vec{F}(\vec{Z}_0)\|_D \\ \Rightarrow \qquad \|\vec{Z}_\varepsilon - X_0\|_D &\leq M_1 \text{ für ein gewisses } M_1 \text{ und alle } \varepsilon > 0,\ h \in]0, h_0]. \end{aligned}$$

Wir haben also gezeigt, daß es für jedes $h_0 > 0$ ein $M > 0$ gibt mit

$$\|\vec{Z}_\varepsilon\|_D \leq M \qquad \text{gleichmäßig auf} \quad \mathbb{R}_{>0} \times]0, h_0].$$

Daß die y - Komponente gleichmäßig beschränkt ist, folgt ganz genauso wie im Beweis zu Satz 4.5.
Für die z - Komponente haben wir

$$z_1 = \tilde{z}_0 + (b^T \otimes I_k) h \vec{Z}'_\varepsilon = \tilde{z}_0 + (b^T \otimes I_k)(A^{-1} \otimes I_k)(\vec{Z}_\varepsilon - e \otimes \tilde{z}_0)$$

und damit ist die gleichmäßige Beschränktheit auch der z - Komponente klar. ■

Bemerkungen:

1. Die Voraussetzungen an F in Satz 4.12 implizieren, daß für alle $x, y \in \mathbb{R}^k$ gilt (siehe [Dei], [OrRh]):

$$((F'(y) - c_F I_k)x, x) \geq 0.$$

Genauso wie in Bemerkung 2. nach Satz 4.5 schließt man dann, daß die Eigenwerte von $F'(y)$ alle Realteil $\geq c_F > 0$ haben. Damit ist auch die Voraussetzung an F in Satz 1.3 gegeben.

2. Welche Runge - Kutta - Verfahren erfüllen nun die Voraussetzungen von Satz 4.12? Die gleichen Voraussetzungen sichern ja auch die Lösbarkeit der auftretenden Gleichungssysteme und das ist schon länger bekannt [DekVe]. Dort finden sich dann auch Angaben, auf welche Verfahren diese Voraussetzungen zutreffen. Ergebnis:

Für die Gauss - , die Radau-IA - , und die Radau-IIA - Formeln gilt:

Es gibt ein D wie verlangt mit: $DA + A^T D$ ist positiv definit.

Für die Lobatto-IIIC - Formeln sind die Verhältnisse noch unklar:

Das 2-stufige Verfahren erfüllt dies noch (mit $D = I_s$, siehe [DekVe], Example 5.9.1). Für das 3-stufige Verfahren gilt dies schon nicht mehr. Man kann aber ein D finden, so daß $DA + A^T D$ positiv semidefinit ist ([DekVe], Example 5.9.2). Weiter kann man zeigen, daß für alle Lobatto-IIIC - Formeln mit ungerader Stufenzahl die Voraussetzung nicht mit "positiv definit" erfüllt sein kann ([DekVe], Theorem 5.9.4).

3. Ob auch hier eine gleichmäßige Abschätzung für den globalen Fehler möglich ist (analog zu Satz 4.6) muß zunächst offen bleiben. Im Beweis zu Satz 4.6 haben wir uns am Beweis des Resultates in [BuHuVe] für eine Klasse von Differentialgleichungen orientiert, die eine einseitige Lipschitzbedingung im linearen Anteil erfüllt. Es gibt auch ein entsprechendes Resultat für nichtlineare Differentialgleichungen mit

$$(F(z_1) - F(z_2), z_1 - z_2) \leq \beta \|z_1 - z_2\|^2,$$

das dem linearen Fall analoges aussagt [BuHu]. Dieses könnte als Hinweis dienen, daß auch Satz 4.6 für nichtlineare Regularisierungen Gültigkeit besitzt.

Nachdem wir gesehen haben, daß viele numerische Verfahren das analytische Verhalten der in Kap. 1 vorgestellte Regularisierungen nachmodellieren, stellt sich die Frage, ob entsprechendes auch im Index-2 - Fall erwartet werden kann. Wir müssen dabei natürlich mit erheblichem Aufwand rechnen; schon die analytischen Aspekte in Kap. 2 verlangten einiges. Daher werden wir uns hier darauf beschränken zu zeigen, daß diese Frage zumindest nicht mit einem klaren NEIN beantwortet werden kann. Das folgende kleine Beispiel läßt die Lage jedenfalls nicht ganz hoffnungslos erscheinen:

<u>Beispiel:</u>
Wir betrachten die DAE

$$\begin{aligned} y' &= y + z \\ 0 &= y. \end{aligned}$$

Diese hat Index-2, wie man sofort mit Lemma 2.1 feststellt.
Die gemäß Satz 2.3 vorgenommene Regularisierung liefert

$$\begin{aligned} y' &= y + z \\ 0 &= y + \varepsilon(y + z) \end{aligned}$$

und dieses ist auch eine ε - stabile DAE (dies zeigt die Bemerkung nach Satz 2.3). Wenden wir nun hierauf das implizite Euler - Verfahren an, so erhalten wir, ausgehend vom Startvektor $(y_0, z_0)^T$ eine Näherung $(y_1, z_1)^T$ aus

$$\begin{array}{rclcrcl} y_1 &=& y_0 + h(y_1 + z_1), & \qquad & 0 &=& (1+\varepsilon)y_1 + \varepsilon z_1 \\ y_1 &=& y_0 - \frac{h}{\varepsilon} y_1, & & z_1 &=& -\frac{1+\varepsilon}{\varepsilon} y_1 \\ y_1 &=& \frac{\varepsilon}{\varepsilon + h} y_0, & & z_1 &=& -\frac{1+\varepsilon}{\varepsilon + h} y_0. \end{array}$$

Jeder für die gestörte Gleichung konsistente Startvektor erfüllt nun

$$y_0 = -\frac{\varepsilon}{\varepsilon + 1} z_0,$$

so daß wir erhalten

$$y_1 = \frac{\varepsilon}{\varepsilon + h} y_0, \qquad z_1 = \frac{\varepsilon}{\varepsilon + h} z_0,$$

woraus wir ohne Schwierigkeiten

$$y_\nu = \left[\frac{\varepsilon}{\varepsilon + h}\right]^\nu y_0, \qquad z_\nu = \left[\frac{\varepsilon}{\varepsilon + h}\right]^\nu z_0, \qquad \nu \in \mathbb{N}$$

entnehmen. Das gleiche Ergebnis würden wir erhalten, wenn wir das implizite Euler - Verfahren auf die Differentialgleichung $\varepsilon y' = -y$ bzw. $\varepsilon z' = -z$ mit den Anfangswerten y_0 bzw. z_0 anwenden. Diese wiederum entstehen nun durch Regularisierung aus den Index-1 - DAEs $0 = y$ bzw. $0 = z$ gemäß 4.6 mit $M = I_n$. Da das implizite Euler - Verfahren nun $B(1)$, $C(1)$ und $A(1)$ erfüllt, sagt uns Satz 4.6, daß

$$y_n = O(\|y_0\|) + O(\varepsilon) + O(h)$$

und das Entsprechende für z_n. Auch in diesem Fall erzielt man also eine numerische Lösung für das gestörte Problem, die nahe bei der exakten Lösung des reduzierten Problems liegt (letztere ist ja $\bar{y} = \bar{z} = 0$).

Wir haben ja die obige Regularisierung in Kap. 2 physikalisch motiviert. Hier wollen wir darauf hinweisen, daß dieser Typ von Regularisierung in der Praxis gerne verwandt wird, wenn bei der numerischen Lösung einer DAE Probleme auftreten, jedenfalls bei der Netzwerksimulation. Dort weiß man, daß Simulationsprogramme mit bestimmten Netzwerken Schwierigkeiten haben (dies sind Systeme mit Index > 1!) und gibt daher in den Handbüchern dem Benutzer den Rat in einer bestimmten Weise sehr kleine oder sehr große Widerstände in sein Netzwerk einzufügen und dann erst das Simulationsprogramm zu starten. Beispiele findet man in den Handbüchern für *MICRO-CAP II* [MC] und *SPICE* [SP]. Diese Empfehlung darf als weiterer Hinweis darauf angesehen werden, daß die vorgestellten Regularisierungen auch unter numerischen Aspekten sinnvoll sein können.

5. Numerische Beispiele für regularisierte DAEs

Hier soll das Verhalten von verschiedenen Verfahren für regularisierte und nichtregularisierte DAEs anhand numerischer Beispiele demonstriert werden. Dazu verwenden wir neben verschiedenen Runge - Kutta - Verfahren das Programm DASSL von L. Petzold [Pe2], das mit BDF - Formeln arbeitet. Wir werden dabei sehen, daß die Verfahren sich für ein regularisiertes System zumindest nicht schlechter als für das ungestörte System verhalten. In einigen Fällen kann es auch deutliche Verbesserungen geben. Diese Beobachtung wurde auch schon von Ascher [As] gemacht; er empfiehlt, wenn der physikalische Hintergrund so ist, daß die vorgelegte DAE das reduzierte Problem einer singulär gestörten Differentialgleichung ist, stets die Möglichkeit zu beachten, das gestörte Problem numerisch zu lösen (und nicht die DAE). Und nach unserer Überzeugung ist der physikalische Hintergrund sehr oft so (wenn nicht immer).

Die Runge - Kutta - Verfahren wurden in TURBO - Pascal auf IBM - AT implementiert, während DASSL in FORTRAN auf IBM - AT benutzt wurde.
Unser Testbeispiel ist durchweg die Pendel - DAE $(0.9.i)$ mit Index $4-i$ für $i = 2, 3$ (siehe Kap. 0) und ihre regularisierten Versionen mit $M = 1$ ("1×1 Einheitsmatrix") bzw. $F = Id$ (vgl. Satz 4.5).

Zunächst zum Index-1 - Fall:

Da wir in Kap. 4 nachgewiesen haben, daß die mit vielen Runge - Kutta - Verfahren bei regularisierten Index-1 - DAEs erzeugten Näherungswerte in gewisser Hinsicht stetig bezüglich des Störparameters ε sind, erwarten wir in diesen Fall keinen großen Unterschied zwischen den Werten für "$\varepsilon = 0$" und "$\varepsilon > 0$". Das ist natürlich nur der Fall, wenn ε klein ist, aber das nehmen wir ja stets an. Es ist damit zu rechnen, daß sich bei kleiner werdendem ε die numerische Approximation immer mehr der nähert, die das Verfahren für die DAE - das reduzierte Problem - liefert. Das sollte auch für den Fall gelten, daß man inkonsistente Startwerte wählt, denn gerade hier kommt ja der Vorteil der Regularisierung zum Tragen.

5.1 <u>Beispiel:</u> (*Index-1 - Pendel - DAE (0.9.3)*)
Zuerst verwenden wir das bereits in Kap. 4 erwähnte Runge - Kutta - Verfahren :

$$(5.1.1) \qquad A := \begin{pmatrix} \frac{3+\sqrt{3}}{6} & 0 \\ \frac{-\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \end{pmatrix}, \qquad b := \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

aus [HW], das schon in [Pe3] als Testbeispiel diente. Dieses Verfahren erfüllt - wie schon erwähnt - die Voraussetzungen unserer Sätze und es gilt $B(4)$, $C(1)$, $A(3)$. Nach Satz 4.6 dürfen wir bei Anwendung auf das Index-1 - System (0.9.3) mit einer in ε gleichmäßigen Konvergenzordnung von (mindestens) 1 in der z - Komponente und 3 in der y - Komponente rechnen. Die Ordnungen in h wurden mit der Formel

$$(5.1.2) \qquad p \approx \log \left[\frac{y_h - y_{h/2}}{y_{h/2} - y_{h/4}} \right] / \log(2)$$

geschätzt, wobei y_h der mit der Schrittweite h mit dem Verfahren berechnete Wert ist. Als Startvektor in $t = 0$ wurde $(y_1, y_2, y_3, y_4, z)^T = (1, 0, 0, 2, 10)^T$ benutzt und es wurde mit verschiedenen Schrittweiten h bis $t = 1$ gerechnet. Beachte, daß der Startvektor nicht konsistent ist, denn sonst müßte die Nebenbedingung von (0.9.3) erfüllt sein, die da lautet:

$$y_3^2 - y_1^2 z + y_4^2 - y_2^2 z + y_2 = 0.$$

Die damit berechneten Werte sind in *Tabelle 1* wiedergegeben. Für die Rechnungen mit $\varepsilon > 0$ wurden nur die z - Komponenten notiert; die Ergebnisse für die y - Komponenten sind die gleichen wie im Fall $\varepsilon = 0$, denn diese Werte sind unabhängig von ε (siehe den Beweis zu Satz 4.5, 4.6).

$\downarrow \varepsilon \setminus h \rightarrow$	0.005	0.01	0.02	Ordnungen
0	−6.83658E−01 7.29802E−01 −1.70524E+00 −1.59742E+00 6.18943E+00	−6.83657E−01 7.29798E−01 −1.70523E+00 −1.59742E+00 6.18949E+00	−6.83646E−01 7.29764E−01 −1.70518E+00 −1.59743E+00 6.18982E+00	3.04 2.99 2.98 3.02 2.40
0.0025	6.20133E+00	6.20139E+00	6.20171E+00	2.49
0.005	6.21311E+00	6.21316E+00	6.21348E+00	2.56
0.01	6.23624E+00	6.23629E+00	6.23659E+00	2.67

Tabelle 1 (s. 5.1)

Die Schätzungen für die Konvergenzordnung in h deuten auf Ordnung 3 in den y - Komponenten hin, während man für die z - Komponente geneigt ist Ordnung 2 anzunehmen. Letztere wäre dann um 1 höher als in Satz 4.6 prognostiziert.
Mit einer zu (5.1.2) analogen Formel läßt sich auch die Konvergenzordnung in ε bei festem h schätzen. Die herkömmliche Theorie würde nur auf eine negative Ordnung schließen lassen (siehe die Erläuterungen in Kap. 4); Satz 4.6 gibt jedoch mindestens Ordnung 1 in ε an (da der Fehler den Term $O(\varepsilon)$ enthält). Natürlich widerspricht das nicht der negativen Ordnung, nur ist diese eben nicht optimal. Ein Blick auf *Tabelle 1* legt aber nahe, daß die Näherungen sich stetig in ε verhalten, zumindest aber beschränkt sind. Eine Ordnungsschätzung gemäß der oben genannten Formel für die mit den obigen Schrittweiten berechneten Werte in den z - Komponenten ergibt jeweils 0.97, also eine recht gute Übereinstimmung mit unserer Prognose. In dieser Hinsicht zumindest scheinen unsere Aussagen (Satz 4.6) nicht verbesserungsfähig zu sein.

Um das Verhalten der mit dem Verfahren (5.1.1) berechneten Werte in der Nähe des (inkonsistenten) Startvektors $(1,0,0,2,10)^T$ zu veranschaulichen, bringen wir noch *Bild 2*. Dort wurde mit verschiedenen Werten für ε wieder (5.1.1) auf (0.9.3) angewandt; dabei wurde von $t_0 = 0$ bis $t = 0.1$ jeweils mit 20 Schritten (und konstanter Schrittweite) gerechnet. Zum Vergleich die "exakte" Lösung zum konsistenten Startvektor $(1,0,0,2,4)^T$. Hier wird die glättende Wirkung der Einfügung von ε sehr deutlich. Die mit $\varepsilon > 0$ berechneten Werte weisen keine Oszillationen mehr auf.

Leider sagen die Sätze in Kap. 4 nichts (d. h. nur gleichmäßige Beschränktheit in ε) für das implizite Euler - Verfahren aus. Da wir aber vermuten, daß dies ein zu behebendes beweistechnisches Problem ist, wollen wir dieses Verfahren auch noch numerisch testen.

5.2 Beispiel: (*Index-1 - Pendel - DAE (0.9.3), impliziter Euler*)

Das implizite Euler - Verfahren ist die Radau-IIA - Formel zur Stufenzahl 1 und lautet in unserer Notation

$$A = (1), \qquad b = (1)$$

oder, etwas klarer anzusehen

$$x_{n+1} = x_n + hF(x_{n+1}) \qquad (\text{ für Differentialgleichungen } x' = F(x) \text{ }).$$

Als Startvektor wurde wieder $(1,0,0,2,10)^T$ benutzt. Es wurden die in *Tabelle 3* aufgeführten Resultate erzielt (wobei für $\varepsilon > 0$ wieder auf die Angabe der y - Komponenten verzichtet wurde; vgl. Beispiel 5.1).

Bild 2 (s. 5.1)

(y_1, z) auf $[0.995, 1] \times [0, 10]$
exakte Lösung zum Startvektor $(1, 0, 0, 2, 4)^T$
numerische Lösung für $\varepsilon = 0$
numerische Lösung für $\varepsilon = 0.01$
numerische Lösung für $\varepsilon = 0.02$

Die Ordnungen wurden wieder mit (5.1.2) geschätzt. Eine Schätzung der Ordnungen in ε ergibt wieder jeweils 0.97 für die verwendeten Schrittweiten. Wir vermuten daher folgende gleichmäßige Ordnungen

Ordnung 1 in h für die y - Komponenten;
Ordnung 1 in h für die z - Komponenten;
Ordnung 1 in ε für die z - Komponenten.

Die Ordnung in h für die y - Komponenten braucht nicht im Stadium der Ungewißheit zu bleiben. Die Beweise der Sätze 4.5 und 4.6 zeigen, daß in der y - Komponente für **jedes** Verfahren die gleiche Ordnung in h wie für gewöhnliche Differentialgleichungen erreicht wird. Was in der z - Komponente genau passiert, sollte Gegenstand späterer Untersuchungen sein.

$\downarrow \varepsilon \setminus h \rightarrow$	0.005	0.01	0.02	Ordnungen
0	−6.76310E−01 7.13240E−01 −1.67167E+00 −1.59028E+00 6.24849E+00	−6.68763E−01 6.96774E−01 −1.63833E+00 −1.58315E+00 6.31181E+00	−6.53084E−01 6.64095E−01 −1.57226E+00 −1.56910E+00 6.45294E+00	1.05 0.99 0.99 0.98 1.16
0.0025	6.26048E+00	6.32386E+00	6.46506E+00	1.16
0.005	6.27233E+00	6.33576E+00	6.47701E+00	1.15
0.01	6.29557E+00	6.35908E+00	6.50038E+00	1.15

Tabelle 3 (s. 5.2)

Damit nicht der Eindruck entsteht, daß die Resultate aus Kap. 4 grundsätzlich **nicht** optimal sind, betrachten wir nun noch

5.3 Beispiel: (*Index-1 - Pendel - DAE (0.9.3), 2 - stufige Gauss - Formel*)

Wir testen unter den gleichen Vorgaben wie oben die 2 - stufige Gauss - Formel, die da lautet (z. B. [DekVe]):

$$A = \begin{pmatrix} 0.25 & 0.25 - 0.5/\sqrt{3} \\ 0.25 + 0.5/\sqrt{3} & 0.25 \end{pmatrix}, \qquad b = (0.5, 0.5)^T.$$

Das liefert die in *Tabelle 4* verzeichneten Werte (wobei für $\varepsilon > 0$ wieder auf die Angabe der y - Komponenten verzichtet wurde; vgl. Beispiel 5.1).
Eine Schätzung der Ordnungen in ε ergibt wieder für alle drei Schrittweiten 0.97. Auffallend ist, daß es dem Verfahren im Fall $\varepsilon = 0$ im Laufe der Rechnung nicht gelingt, den Fehler in der z - Komponente des Startvektors loszuwerden. In den Rechnungen mit $\varepsilon > 0$ gelingt das hingegen ganz hervorragend. Dieses Phänomen läßt sich aber erklären, indem man den Term $O(\|e_0\|)$ im Beweis von Satz 4.6 einmal genauer analysiert:
Im Fall $M = I_k$ (wie hier) ist (vgl. den Beweis zu Satz 4.6)

$$N(h,\varepsilon) = \frac{h}{\varepsilon}(b^T \otimes I_k)(I_{ks} + h(A \otimes \varepsilon^{-1} I_k))^{-1} = \frac{h}{\varepsilon}\left[b^T(I_s + \frac{h}{\varepsilon}A)^{-1}\right] \otimes I_k$$

und damit auch

$$N(h,\varepsilon)E = \frac{h}{\varepsilon}\left[b^T(I_s + \frac{h}{\varepsilon}A)^{-1}\right] e \cdot I_k.$$

$\downarrow \varepsilon \setminus h \rightarrow$	0.005	0.01	0.02	Ord.
0	−6.83657937E−01	−6.83657937E−01	−6.83657945E−01	4.00
	7.29802594E−01	7.29802593E−01	7.29802575E−01	4.00
	−1.70524200E+00	−1.70524200E+00	−1.70524198E+00	4.00
	−1.59742133E+00	−1.59742133E+00	−1.59742137E+00	4.00
	1.21893956E+01	1.21893591E+01	1.21892134E+01	2.00
0.0025	6.20132189E+00	6.20132223E+00	6.20132752E+00	3.97
0.005	6.21310034E+00	6.21310052E+00	6.21310327E+00	3.98
0.01	6.23623565E+00	6.23623574E+00	6.23623720E+00	3.99

Tabelle 4 (s. 5.3)

Erinnert man sich nun an die Stabilitätsfunktion

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

und setzt $z := -\frac{h}{\varepsilon}$ so erhält man

$$e_{i+1} = R(z)e_i - \left[zb^T(I_s - zA)^{-1} \otimes I_k\right]\left[EO(\varepsilon) + O(\varepsilon h) + O(h^{q+1})\right] + O(h^{p+1}).$$

Und daraus sieht man

$$e_\nu = R(z)^\nu e_0 + O(\varepsilon) + O(h^q) + O(h^p).$$

Beachte, daß $R(z)$ von ε abhängt. Es ist nun für alle s - stufigen Gauss - Formeln ([DekVe])

$$\lim_{\varepsilon \to 0} R(z) = 1,$$

so daß man im Fall $\varepsilon = 0$ natürlich

$$e_\nu = e_0 + O(h^q) + O(h^p)$$

hat. Für $\varepsilon > 0$ ist aber wegen der A - Stabilität $R(z) < 1$, so daß dann wieder die Dämpfung des Fehlers e_0 gegeben ist. Man betrachte dazu *Bild 5.* Hierfür wurde von $t_0 = 0$ bis $t = 0.1$ mit 20 Schritten (und konstanter Schrittweite) vom Startvektor $(1, 0, 0, 2, 10)^T$ gerechnet und mit der exakten Lösung verglichen.
Dagegen erfüllen alle Radau-IA - , Radau-IIA - , und Lobatto-IIIC - Formeln $\lim_{\varepsilon \to 0} R(z) = 0$ ([DekVe]), so daß sich der Startfehler in der z - Komponente für ε hinreichend klein (nämlich so klein, daß $|R(z)| < \alpha < 1$) nicht mehr bemerkbar macht. (Dies ist ja bei

Bild 5 (s. 5.3)

(y_1, z) auf $[0.995, 1] \times [2, 12]$
exakte Lösung zum Startvektor $(1, 0, 0, 2, 4)^T$
numerische Lösung für $\varepsilon = 0$
numerische Lösung für $\varepsilon = 0.01$

der analytischen Lösung auch der Fall). Wie schon in Bemerkung 4.8.2 erwähnt, ist für das Verfahren (5.1.1) der obige Grenzwert kleiner als 1, so daß auch dieses die genannte Dämpfungseigenschaft besitzt.

Außerdem ist die Konvergenzordnung in h für $\varepsilon = 0$ deutlich niedriger als für $\varepsilon > 0$. Satz 4.6 läßt auf gleichmäßige Ordnung mindestens 2 in h schließen und unser Testbeispiel gibt einen deutlichen Hinweis, daß diese auch nicht höher sein kann. Hier ist also die nachgewiesene Ordnung offensichtlich optimal.

5.4 Anwendung im Index-2 - Fall:

Wir wollen hier auf die Wiedergabe weiterer Tabellen und Kurven verzichten und uns auf einige Bemerkungen beschränken.
Wir haben zwar keine Aussagen über die Konvergenz verschiedener Verfahren vorliegen, erhoffen aber, daß sich die Verfahren für das regularisierte System (das Index-1 hat) nicht schlechter als für das Original - Index-2 - Problem verhalten (denn z. B. ist die Menge der konsistenten Startwerte nicht so eingeschränkt wie die für das Index-2 - System).
Testrechnungen für das Index-2 - Problem (0.9.2) mit einem inkonsistenten Startvektor lassen keinen relevanten Unterschied zwischen den Rechnungen mit $\varepsilon = 0$ und $\varepsilon > 0$ erkennen. (Hierbei wurde wieder mit $F = Id$ gemäß Satz 2.3 regularisiert).
Man könnte jedoch auch auf die Idee kommen, in das gestörte System (das Index-1 hat) noch ein weiteres ε_1 einzufügen um so Index-0 mit zwei Störparametern zu erreichen. Das würde auf

$$\begin{aligned} y' &= f(y, z + \varepsilon_1 z') \\ 0 &= (y_1 + \varepsilon y_3)(y_3 - \varepsilon y_1(z + \varepsilon_1 z')) + (y_2 + \varepsilon y_4)(y_4 + \varepsilon(1 - y_2(z + \varepsilon_1 z'))) \end{aligned}$$

führen (f wie in (0.9.1) definiert).
Für dieses System beobachten wir die gleichen Effekte wie in den oben genannten Beispielen (was ja auch zu erwarten war). Z. B. liefert das Verfahren (5.1.1) für dieses doppelt gestörte System verglichen mit dem selben Verfahren für das Original - Index-2 - System ein Verhalten ganz ähnlich dem in *Bild 2*.

Wie verhält sich nun ein Programm mit Schrittweitensteuerung? Dazu verwenden wir das von L. Petzold [Pe2] entwickelte Programm DASSL in der DOUBLE - Version. Dort wird mit der Familie der BDF - Formeln gearbeitet und es wird eine Schrittweitensteuerung verwendet, die auf den Index-1 - Fall zugeschnitten ist.

5.5 Beispiel: (*Verhalten von BDF - Formeln*)

Wir haben zwar in dieser Arbeit nicht das Verhalten von BDF - Formeln für regularisierte DAEs untersucht, doch ist das bereits von Lötstedt ([Lö1], [Lö2]) getan worden. Die Ergebnisse sind mit den in Kap. 4 vorgestellten vergleichbar.
Testrechnungen für Index-1 - Systeme und deren regularisierte Varianten zeigen daher nur die "stetige" Abhängigkeit der numerischen Lösung von ε. Es gibt keine erwähnenswerten Unterschiede, weder im Aufwand noch in der Genauigkeit. Wir verzichten daher auf weitere Zahlenkolonnen.
Wie aber ist die Situation bei Index-2 - DAEs?
Bekanntlich ([Pe1]) gibt es bei der Anwendung numerischer Verfahren auf DAEs im Vergleich zu gewöhnlichen Differentialgleichungen nicht nur Unterschiede in der Konvergenzordnung, sondern auch darin, daß die Schrittweitensteuerung modifiziert werden muß.

Die Strategie für die Wahl einer "vernünftigen" Schrittweite hängt aber auch vom Index der zu lösenden DAE ab und ist bisher nur für den Index-1 - Fall befriedigend geklärt. Man muß daher bei der Anwendung von DASSL auf Probleme mit höherem (> 1) Index mit Schwierigkeiten rechnen.
Verwenden wir jedoch das regularisierte Index-2 - System (das ja Index-1 hat), so sollte es aber bei der Schrittweitensteuerung theoretisch keine Probleme geben. In der Praxis ist jedoch das Rechnen mit kleinen Zahlen (wie eben dem Störparameter ε) anfällig gegenüber Rundefehlern und anderen im Laufe der Rechnung zwangsläufig auftretenden Ungenauigkeiten. Lassen wir es auf einige Versuche ankommen.

Wir wendeten DASSL mit verschiedenen Vorgaben von Toleranzen TOL (für den absoluten Fehler) mit verschiedenen Werten von ε auf (0.9.2) an. Dabei wurde stets in $t_0 = 0$ der für die DAE konsistente Startvektor $(1, 0, 0, 2, 4)^T$ vorgegeben und es wurde bis $t = 0.25$ gerechnet. Wir geben nur die Ergebnisse für die Komponenten y_3, y_4, z an. In der ersten Zeile ist die exakte Lösung angegeben und in der letzten Spalte die Anzahl der Schritte, die DASSL aufgrund seiner Schrittweitenstrategie benötigte. Die auffallendsten Ergebnisse sind dabei folgende:

ε	TOL	y_3	y_4	z	n
		-1.13283955	1.93098425	5.51803971	
0	1.E$-$4	*	*	*	*
0.001	1.E$-$4	-1.13262554	1.93033715	5.51719176	478
0	1.E$-$6	**	**	**	**
0.001	1.E$-$6	-1.13284233	1.93098436	5.51804863	76
0.01	1.E$-$6	-1.13310821	1.93094665	5.51887056	70

*: DASSL bricht bei $t = 0.0983$ ab, da mehr als 500 Schritte nötig wären.
**: DASSL bricht bei $t = 0.0182$ ab, da mehr als 500 Schritte nötig wären.

If you cannot solve the proposed problem
try to solve first some related problem.
– G. Polya, How to solve it, Princeton 1945

Bezeichnungen

$\mathbb{N}$, $\mathbb{R}$, $\mathbb{C}$	die natürlichen, reellen, komplexen Zahlen
$\mathbb{R}_{\geq 0}$	$\{x \in \mathbb{R} \mid x \geq 0\}$
$\mathbb{C}^-$	$\{z \in \mathbb{C} \mid Rez \leq 0\}$
DAE	Differential - Algebraic Equation, Def. 0.2, S. 3
IRA	Index - Reduktionsalgorithmus, Alg. 0.6, S. 6
$rg(A)$, A Matrix	Rang von A
$N(A)$, A Matrix	Nullraum von A
$R(A)$, A Matrix	Bildraum von A
$\mathcal{O}_k$	Nullvektor der Länge k
I_k	$k \times k$ Einheitsmatrix
$diag(d_1, \ldots, d_k)$	$k \times k$ Diagonalmatrix mit Diagonalelementen $d_1, \ldots, d_k$
f_z, f'	partielle bzw. totale Ableitung einer Funktion f
$B(x, \delta)$	$\{y \mid \|y - x\| < \delta\}$, "offene Kugel"
$B(X, \delta)$, X Menge	$\bigcup_{x \in X} B(x, \delta)$
$A \otimes B$	direktes Produkt von Matrizen, Def./Lemma 4.1, S. 39
e (Vektor)	Vektor mit $e_i = 1$ für alle i, S. 42
$B(p)$, $C(p)$, $A(p)$	Eigenschaften von Runge - Kutta - Verfahren, S. 40

Literatur

[Ad] P. R. Adby:
Applied Circuit Theory: Matrix and Computer Methods
John Wiley 1980

[As] U. Ascher:
On Numerical Differential Algebraic Problems with
Application to Semiconductor Device Simulation
Dept. of Comp. Sci., Univ. of Brit. Col.,Tech. Report 87 - 27, (1987)

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

[BuHuVe] K. Burrage, W. H. Hundsdorfer, J. A. Verwer:
A Study of B - Convergence of Runge - Kutta Methods
Computing 36 (1986), S. 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), S. 365-370

[ChRo] L. O. Chua, R. A. Rohrer:
On the Dynamic Equations of a Class of Nonlinear RLC Networks
IEEE Trans. on Circuit Theory, Vol. CT-12, No. 4, 1965, S. 475-489

[ChAl] L. O. Chua, G. R. Alexander:
The Effects of Parasitic Reactances on Nonlinear Networks
IEEE Trans. on Circuit Theory, Vol. CT-18, No. 5, 1971, S. 520-532

[Dei] K. Deimling:
Nonlinear Functional Analysis
Springer 1985

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

[DesKuh] C. A. Desoer, E. S. Kuh:
Basic Circuit Theory
McGraw Hill 1969

[DeuHaZu] P. Deuflhard, E. Hairer, J. Zugck:
One-Step and Extrapolation Methods for Differential-Algebraic Systems
Num. Math. 51 (1987), S. 501-516

[Do] V. Dolezal:
The Existence of a Continuous Basis of a Certain Linear Subspace of E_r which depends on a Parameter
Čas. pro Pěst. Mat. 89 (1964), S. 466-469

[Fü] C. Führer:
On the Description of Constrained Mechanical Systems by Differential/Algebraic Equations
In: Proceedings of the Workshop "The Road - Vehicle - System and Related Mathematics" (H. Neunzert ed.) in Lambrecht 1985
B. G. Teubner Stuttgart 1985

[GePe] C. W. Gear, L. R. Petzold:
ODE-Methods for the Solution of Differential/Algebraic Systems
SIAM J. Num. Anal. 21 (1984), S. 716-728

[GreLeSa] P. M. Gresho, R. L. Lee, R. L. Sani:
On the Time - Dependent Solution of the Incompressible Navier - Stokes Equations in Two or and Three Dimensions
in: Recent Advances in Numerical Methods in Fluids, 1980, Pineridge, Swansea

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

[GriMä] E. Griepentrog, R. März:
Differential-Algebraic Equations and their Numerical Treatment
B. G. Teubner, Leipzig 1986

[Ha] E. Hairer:
A Note on D-Stability
BIT 24 (1984), S. 383-386

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

[La] P. Lancaster:
Theory of Matrices
Academic Press 1969

[LeLe] J. J. Levin, N. Levinson:
Singular Perturbations of Nonlinear Systems of Differential Equations
Equations and an Associated Boundary Layer Equation
J. Ration. Mech. Anal. (nun "Ind. J. Math.") 3 (1954), S. 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)

[LöPe] P. Lötstedt, L. R. Petzold:
Numerical Solution of Nonlinear Differential Equations with
Algebraic Constraints
SANDIA Report SAND 83-8877 (1983)

[Mä1] R. März:
On Tractability with Index-2
Humboldt Univ. Berlin, Sektion Math., Preprint 109, 1986

[Mä2] R. März:
Higher - Index Differential - Algebraic Equations:
Analysis and Numerical Treatment
Humboldt Univ. Berlin, Sektion Math., Preprint 159, 1987

[MC] A. W. Thompson:
MICROCAP II - Microcomputer Circuit Analysis Program
Spectrum Software

[Mr1] T. Mrziglod:
Zur Theorie und numerischen Realisierung von Lösungsmethoden bei Differentialgleichungen mit angekoppelten algebraischen Gleichungen
Diplomarbeit, Universität zu Köln 1986/1987

[Mr2] T. Mrziglod:
Private Mitteilung, März 1988

[OrRh] J. M. Ortega, W. C. Rheinboldt:
Iterative Solution of Nonlinear Equations in Several Variables
Academic Press 1970

[Pe1] L. R. Petzold:
Differential/Algebraic Equations are not ODEs
SIAM J. Sci. Stat. Computing 3 (1982), S. 367-384

[Pe2] L. R. Petzold:
A Description of DASSL - a differential-algebraic system solver
SANDIA - Report SAND 82 - 8637 (1982)

[Pe3] L. R. Petzold:
Order Results for Implicit Runge-Kutta Methods applied to Differential/Algebraic Systems
SIAM J. Num. Anal. 23 (1986), S. 837-852

[Rh] W. C. Rheinboldt:
Differential-Algebraic Systems as Differential Equations on Manifolds
Math. Comp. 43 (1984), S. 473-482

[SP] E. E. E. Hoefer, H. Nielinger:
SPICE - Benutzerhandbuch
Springer - Verlag

[vV] M. van Veldhuizen:
D-Stability
SIAM J. Num. Anal. 18 (1981), S. 45-64