Die BDF für nichtlineare Algebro–Differentialgleichungen vom Index 2

Diplomarbeit

angefertigt von

Caren Tischendorf

am
Institut für Angewandte Mathematik
Humboldt-Universität zu Berlin

Betreuer: Prof. Dr. sc. nat. R. März

Berlin, den 27. Oktober 1992

An dieser Stelle möchte ich Frau Prof. R. März für die interessante Aufgabenstellung, die anregenden Diskussionen zu dieser Problematik und ihre stetige, freundliche Unterstützung danken.

Mein Dank gilt ebenfalls Herrn Dr. R. Lamour für die Überlassung seines Programms *NLSOLV* zur Lösung nichtlinearer Gleichungssysteme.

Inhaltsverzeichnis

Einleitung

In der gegenwärtigen Wissenschaft und Technik tauchen immer häufiger Probleme auf, die sich in natürlicher Weise durch Algebro–Differentialgleichungen (ADG) beschreiben lassen, d.h. mit Hilfe impliziter gewöhnlicher Differentialgleichungen der Form

$$f(x'(t), x(t), t) = 0,$$

deren partielle Ableitung $f'_y(y, x, t)$ singulär ist, mathematisch modelliert werden können. Stellvertretend hierfür seien an dieser Stelle die Modellierung mechanischer Mehrkörpersysteme, die Analyse elektrischer Netzwerke, die Simulation chemischer Prozesse und Probleme der optimalen Steuerung genannt. Bereits in den 70er Jahren wurden die ersten Versuche unternommen, solche Algebro–Differentialgleichungen numerisch zu behandeln. In vielen Fällen bewährt hat sich davon der Vorschlag von Gear [71], die ADGen mit Hilfe der BDF (backward differentiation formulas, dt.: rückwärts genommene Differenzenformeln) zu lösen. Jedoch lieferte die Praxis auch Beispiele, bei denen die BDF und andere bis dahin bekannte numerische Verfahren versagten. Dies war u.a. ein Grund dafür, daß man sich seit etwa den 80er Jahren eingehender mit der Theorie der Algebro–Differentialgleichungen beschäftigte. Das unterschiedliche numerische Verhalten von Lösungen von ADGen erforderte eine Klassifizierung dieser, welche durch den Indexbegriff gegeben ist. Mit einer ausführlichen Darstellung dessen und der damit verbundenen Probleme befaßt sich der erste Abschnitt dieser Arbeit. Dabei wird u.a. dargelegt, daß zur Lösung von ADGen vom Index 2 neben der eigentlichen Integration auch Differentiationen notwendig werden, wodurch die Aufgabe zu einem "schlecht gestellten" Problem im Sinne von Hadamard (Louis [89]) wird. Damit numerische Verfahren wie die BDF erfolgreich angewandt werden können, ist nun von entscheidender Bedeutung, wie diese Differentiationen in das Problem eingehen.

Ziel der Arbeit war, die Klasse von Index–2–Problemen herauszuarbeiten, für die die inzwischen in der Praxis häufig verwendeten BDF tatsächlich zuverlässige Ergebnisse liefern. Ausgangspunkt für meine Untersuchungen waren Resultate von Gear, Gupta & Leimkuhler [85], Lötstedt & Petzold [86], Brenan & Engquist [89], März [92] u.a. Auf diese Arbeiten wird am Ende des ersten Abschnitts etwas näher eingegangen. In ihren Abhandlungen wird deutlich, daß Verfahren höherer Ordnung (≥ 2) eine einmalige Differentiation, wie sie bei den Index–2–Problemen auftritt, relativ gut bewältigen. Problematisch wird es beim impliziten Euler–Verfahren, insbesondere für nichtlineare Systeme. Dennoch schien es, daß zumindest Hessenberg–Systeme vom Index 2 der Form

$$\begin{aligned} u'(t) + g(u(t), v(t)) &= 0 \\ h(u(t)) &= 0 \end{aligned}$$

mit dem impliziten Euler–Verfahren erfolgreich numerisch gelöst werden können. Die theoretische Analyse, mit der sich der zweite Abschnitt und weitergehend

der dritte Abschnitt beschäftigen, ließ nun deutlich werden, daß selbst eine Einschränkung auf diese Klasse nicht immer die gewüschten Resultate liefert. Ein entsprechendes Beispiel dafür ist im dritten Abschnitt angegeben. Vielmehr stellte sich bei den Untersuchungen heraus, daß eine gewisse Linearitätsbedingung für bestimmte Komponenten erfüllt sein muß. Mit den im dritten Abschnitt gezeigten Konvergenz– und Stabilitätsresultaten wird neben den eben angesprochenen Fragen erstmalig bewiesen, daß die Einzugsbereiche des Newton–Verfahrens für die hier auftretenden nichtlinearen Gleichungen unabhängig von der Schrittweite sind. Erst damit ist gesichert, daß das implizite Euler–Verfahren tatsächlich durchführbar ist.
Im letzten Abschnitt wird eine Implementierung der BDF mit variabler Schrittweite und variabler Ordnung vorgestellt, die die theoretischen Ergebnisse dahingehend berücksichtigt, daß für die Schrittweiten– und Ordnungssteuerung lediglich die differentiellen Komponenten herangezogen werden. Auf diese Weise können einerseits Rechenzeiten minimiert und andererseits einige Probleme, bei denen Verfahren mit herkömmlicher Steuerung versagten, gelöst werden.

1 Einführende Betrachtungen

1.1 Lösungs- und Index-Begriff für ADGen

Wir betrachten die allgemeine nichtlineare Algebro–Differentialgleichung

$$f(x'(t), x(t), t) = 0, \quad f : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m, \tag{1.1}$$

wobei $x' := \frac{dx}{dt}$ bezeichnet, und die partielle Ableitung $f'_y(y, x, t)$ konstanten Rang auf dem Definitionsgebiet von f hat. Es würde genügen (vgl. Rabier & Rheinboldt [91]), daß diese Rangbedingung auf einer offenen Umgebung von $f^{-1}(0)$ in $\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}$ erfüllt ist. Der Einfachheit und Übersichtlichkeit halber gelte die Rangbedingung in den hier betrachteten Fällen auf dem Definitionsgebiet von f,

$$D_f := \mathbb{R}^m \times D \times \mathcal{I},$$

hier bezeichnen $D \in \mathbb{R}^m$ ein offenes Gebiet und $\mathcal{I} \in \mathbb{R}$ ein offenes Zeitintervall. Eine "klassische" Lösung einer solchen Algebro–Differentialgleichung wäre eine auf $\mathcal{I}$ stetig differenzierbare Funktion $x(\cdot)$, die der Gleichung (1.1) genügt. Die Untersuchungen von ADGen haben nun gezeigt, daß es für die Lösbarkeit solcher impliziten Gleichungen einerseits notwendig sein kann, daß Teile der Lösung höhere Glattheitsbedingungen erfüllen müssen, und andererseits bestimmte Teile der Lösung nur stetig zu sein brauchen. Ein weiteres Problem gegenüber den (regulären) gewöhnlichen Differentialgleichungen besteht darin, daß zu beliebigen Anfangswerten für ADGen nicht unbedingt eine Lösung existiert. Nicht zuletzt gibt es bei manchen ADGen erhebliche numerische Probleme, diese zu lösen. Es entstand die Frage, ob sich die Algebro–Differentialgleichungen in bestimmter Art und Weise charakteresieren lassen, so daß die Kriterien Aufschluß darüber geben, wie Lösungsraum und Anfangswerte zu wählen sind, um überhaupt Lösbarkeit der ADGen erreichen zu können. Eine solche Charakteresierung ist der Index einer ADG. Bei den verschiedenen Untersuchungen zu den Algebro–Differentialgleichungen ist es nun zu unterschiedlichen Indexbegriffen und demzufolge auch unterschiedlichen Lösungsbegriffen gekommen. Ausgangsbasis sämtlicher Definitionen sind jedoch die Ergebnisse des Studiums linearer ADGen mit konstanten Koeffizienten

$$Ax'(t) + Bx(t) = q(t), \tag{1.2}$$

hier sind $A, B \in L(\mathbb{R}^m)$, A singulär. Die Lösung eines solchen Gleichungssystems steht in unmittelbarem Zusammenhang zu den Eigenschaften des Matrixbüschels $\{A, B\}$. So ist für die Eindeutigkeit von Anfangswertproblemen, die der Gleichung (1.2) genügen, notwendig, daß das Matrixbüschel $\{A, B\}$ regulär ist, d.h. das Polynom $p(\lambda) := \det(\lambda A + B)$ nicht identisch verschwindet (vgl. Griepentrog

& März [86]). Nun lassen sich reguläre Matrixbüschel $\{A, B\}$ mittels regulärer Matrizen $E, F \in L(\mathbb{R}^m)$ zu einem System

$$EAF = \mathrm{diag}(I, J),\ EBF = \mathrm{diag}(W, I)$$

transformieren (vgl. Gantmacher [54]), wobei $W \in L(\mathbb{R}^k)$, und $J \in L(\mathbb{R}^{m-k})$ ist eine nilpotente Blockmatrix mit Jordanblöcken der Form

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

Dementsprechend erhält man durch geeignete Koordinatentransformation ein zu (1.2) äquivalentes System

$$u'(t) + Wu(t) = p(t) \tag{1.3a}$$
$$Jv'(t) + v(t) = r(t). \tag{1.3b}$$

Die erste Gleichung (1.3a) stellt offenbar eine reguläre gewöhnliche Differentialgleichung dar, während die Lösung der zweiten Gleichung (1.3b) die Form

$$v(t) = \sum_{k=0}^{\mu-1} (-1)^k (J^k r(t))^{(k)} \tag{1.4}$$

hat, wenn $r(\cdot)$ entsprechend oft differenzierbar und μ die Nilpotenz der Jordan–Block–Matrix J ist. Dieses μ ist unabhängig von der Wahl der Transformation und wird der *Index des Matrix-Büschels* $\{A, B\}$ genannt. Entsprechend definiert man den *Index der ADG* (1.2) ebenfalls als diese natürliche Zahl μ. Das System (1.3a), (1.3b) und dessen Lösung (vgl. (1.4)) machen nun folgendes deutlich:

(i) Algebro–Differentialgleichungen verkörpern sowohl Integrationsprobleme als auch Differentiationsprobleme. Teile der rechten Seite müssen hinreichend oft differenzierbar sein. Der Lösungsraum ist so zu wählen, daß bestimmte Komponenten hinreichend glatt sind.

(ii) Einige Komponenten der Lösung sind algebraisch bestimmt. Das bedeutet für die Lösbarkeit von Anfangswertproblemen, daß die Anfangswerte nicht frei wählbar sind, sie müssen "konsistent" mit der ADG sein.

Auf diesen Fakten bauen nun die verschiedenen Indexkonzepte für allgemeinere ADGen auf, so entstanden

- der Differentiations–Index (Gear & Petzold [83], Gear, Gupta & Leimkuhler [85], Brenan, Campbell & Petzold [89], Gear [90]),

- der Traktabilitäts–Index (Griepentrog & März [86], März[90], März[92]),
- der Störungs–Index (Hairer & Lubich & Roche [89]),
- der geometrische Index (Rheinboldt [84], Rabier &Rheinboldt [91], Reich [90], Griepentrog [91]).

Die Aufzählung beansprucht keine Vollständigkeit, beinhaltet aber wohl die wesentlichen bisherigen Konzepte zu diesem Thema. Sie unterscheiden sich durch die Herangehensweise an die ADGen. So orientieren sich die drei erstgenannten an analytischen Gesichtspunkten, die sowohl zu Existenzaussagen als auch zu praktischen numerischen Ergebnissen führen. Beim geometrischen Index werden die ADGen als gewöhnliche Differentialgleichungen auf Mannigfaltigkeiten betrachtet und damit die natürliche Herkunft der ADGen erklärt. Die praktische Anwendung der geometrischen Resultate scheint momentan noch eng an die Entwicklung der Computeralgebra gebunden zu sein.
Bei all diesen Arbeiten sind die Indexbegriffe im Falle allgemeiner nichtlinearer ADGen (1.1) an bestimmte Eigenschaften der Funktion f gekoppelt. So wird z.B. häufig vorausgesetzt, daß die gesamte Funktion f hinreichend oft differenzierbar ist. In dieser Arbeit verwende ich den Traktabilitäts–Index, der sich dadurch auszeichnet, daß er mit minimalen Glattheitsforderungen an die Funktion f, als auch an die Lösung auskommt. Dafür wird jedoch die Eigenschaft von f gebraucht, daß der Nullraum von $f'_y(y,x,t)$ von (y,x) unabhängig ist und glatt von t abhängt. Es sei hier bemerkt, daß dies in der überwiegeden Zahl der bisherigen praktischen Anwendungen der Fall ist. Mit dieser Voraussetzung existiert ein Projektor $Q \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ auf den Nullraum $ker(f'_y(y,x,t)$ und es gilt folgende Identität:

$$f(y,x,t) - f(P(t)y,x,t) = \int_0^1 f'_y(sy + (1-s)P(t)y, x, t)Q(t)\,ds \;=\; 0, \quad (y,x,t) \in D_f,$$

wobei $P := I - Q$ bezeichnet. So kann die Gleichung (1.1) auch geschrieben werden als

$$f((Px)'(t) - P'(t)x(t), x(t), t) = 0, \tag{1.5}$$

und der Funktionenraum, zu dem die Lösungen gehören sollten, ist der folgende:

$$C^1_N(\mathcal{I}_0, \mathbb{R}^m) := \{x \in C(\mathcal{I}_0, \mathbb{R}^m) : Px \in C^1(\mathcal{I}_0, \mathbb{R}^m)\}, \tag{1.6}$$

hier sei $\mathcal{I}_0 \subseteq \mathcal{I}$ ein gewisses Intervall.

1.2 Numerische Verfahren für Anfangswertprobleme von ADGen

Für die numerische Behandlung von Algebro–Differentialgleichungen ist es naheliegend, die für reguläre gewöhnliche Differentialgleichungen bewährten Verfah-

ren zu untersuchen. Die erste Frage, die sich sofort ergibt, ist die Frage nach der Durchführbarkeit dieser Verfahren. Die implizite Gestalt der Algebro–Differentialgleichungen erfordert zunächst die Untersuchung, unter welchen Bedingungen ein Verfahren eine Lösung liefert. Daran anschließend erhebt sich die Frage, in welcher Beziehung die numerische Lösung zur exakten Lösung der ADG steht, wenn die ADG überhaupt eine Lösung besitzt. Im Gegensatz zu den regulären gewöhnlichen Differentialgleichungen bereitet bereits die Entscheidung der ersten Frage erhebliche Probleme, was jedoch nicht anders zu erwarten ist, wenn man berücksichtigt, daß schon die Frage nach der Lösbarkeit der ADG nicht in jedem Fall beantwortet werden kann. Wendet man sich der zweiten Frage zu, so wird man mit einem neuen Problem konfrontiert. Man stellt fest, daß bei Verkleinerung der Schrittweite eines numerischen Verfahrens die Approximation der exakten Lösung nicht unbedingt besser werden muß, wie man es zunächst (ohne eingehendere Analyse der Struktur der ADGen) erwarten könnte. Der Grund dafür ist das oben genannte Eingehen von Differentiationsproblemen in eine ADG. Gerade dieser Fakt bereitet bei in der Praxis auftretenden Problemen große Schwierigkeiten und erfordert umfangreiche theoretische und numerische Untersuchungen zu den ADGen. Betrachtet man noch einmal das System (1.3a), (1.3b) und Gleichung (1.4), so ist eine Differentiation bestimmter Komponenten erst bei einem Index ≥ 2 notwendig. Deshalb treten die entscheidenden numerischen Schwierigkeiten erst bei solchen Problemen auf, und diese Probleme werden in der Literatur *Systeme höheren Index* genannt. Schon bei allgemeinen Index-2-Systemen wird eine besondere Steuerung der numerischen Verfahren notwendig, was in den nächsten Abschnitten meiner Arbeit auch zum Ausdruck kommen wird.

1.3 Bisherige Resultate zu den BDF

In einer Reihe von Arbeiten wurden die BDF schon für Index-2-Probleme untersucht. Gegenstand der ersten Untersuchungen waren u.a. die linearen Index–2–ADGen mit variablen Koeffizienten

$$A(t)x'(t) + B(t)x(t) = q(t). \tag{1.7}$$

Dabei zeigte sich, daß sich schon diese relativ einfache Klasse von Aufgaben mit Hilfe der BDF nicht mehr vollständig bewältigen läßt. Zur Illustration dessen möchte ich hier ein Beispiel anführen, welches von Gear & Petzold [84] konstruiert wurde. Wir betrachten die Index–2–ADG

$$\begin{pmatrix} 0 & 0 \\ 1 & \eta t \end{pmatrix} x' + \begin{pmatrix} 1 & \eta t \\ 0 & 1+\eta \end{pmatrix} x = q(t).$$

Hierbei ist η ein beliebiger Parameter und $q(\cdot)$ eine beliebige stetige Funktion, dessen erste Komponente $q_1(\cdot)$ stetig differenzierbar ist. Die Lösung dieses Systems

ist dann eindeutig durch

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

gegeben. Wendet man nun das implizite Euler–Verfahren auf die ADG an, so ergibt sich für den Fall $\eta \neq -1$ folgende Lösung:

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

Das Verfahren ist dann auf einem abgeschlossenen Intervall $[t_0, t_N]$ schwach instabil und konvergent, falls $\eta > -0.5$, und exponentiell instabil, falls $\eta < -0.5$. Für $\eta = -1$ versagt das Verfahren. Leicht zu sehen sind diese Aussagen für die speziellen rechten Seiten

$$q_1(t) = q_2(t) = ct,$$

betrachtet auf dem abgeschlossenen Intervall $[0, 1]$. In diesem Fall ergibt sich bei exaktem Anfangswert $x_{2,0} = x_2(0) = -c$ und $t_N = 1$:

$$x_{2,N} - x_2(t_N) = ch\eta \left[\left(\frac{\eta}{1+\eta} \right)^N - 1 \right].$$

Diese Tatsache beweist, daß die BDF nur für eine eingeschränkte Problemklasse von Index–2–ADGen geeignet sind. Die wohl schwächste Voraussetzung, die eine solche Aufgabe erfüllen muß, um mit den BDF erfolgreich behandelt werden zu können, ist die von R. März angegebene Bedingung, daß der Nullraum von $A(\cdot)$ kostant ist. An dieser Stelle sollen die in der Arbeit von März [92] bewiesenen Konvergenz– und Stabilitätsresultate für solche linearen Index–2–ADGen angeführt werden, um deutlich zu machen, welchen Einfluß die beim Index 2 auftretenden Differentiationen auf die numerische Lösung haben.

Es seien die BDF mit der Ordnung s für die regulären gewöhnlichen Differentialgleichungen auf der Klasse von Zerlegungen des abgeschlossenen Intervalls $[t_0, t_N]$ *stabil. Weiter besitze die ADG (1.7) den Traktabilitäts–Index 2 und der Nullraum* $ker(A(t))$ *sei von t unabhängig. Dann liefern die BDF eine Lösung* x_j *(j=s,...,N), für die folgende scharfe Fehlerabschätzung gilt:*

$$\begin{aligned} &\max_{j \geq s} |P(x_*(t_j) - x_j)| \leq \\ &\quad \leq S_P \{ \max_{j \leq s-1} |P(x_*(t_j) - x_j)| + \max_{j \geq s} |\tau_j - \delta_j| \} \end{aligned}$$

und

$$|Q(x_*(t_j) - x_j)| \leq$$

$$\begin{aligned} \leq \quad & S_Q\{\max_{j\leq s-1}|P(x_*(t_j)-x_j)|+\max_{j\geq s}|\tau_j-\delta_j|\} \\ & +\frac{1}{h_j}|QQ_1(t_j)\sum_{i=0}^{s}\alpha_{ji}Q_1(t_{j-i})G_2(t_{j-i})^{-1}\tilde{\delta}_{j-i}|, \end{aligned}$$

wobei $x_*(\cdot)$ die exakte Lösung der Gleichung (1.7) ist, S_P und S_Q gewisse Konstanten sind, $\tilde{\delta}_j := G_2(t_j)(x_*(t_j)-x_j)$ für $j=0,\ldots,s-1$ die Fehler in den Startwerten darstellen, $\tilde{\delta}_j := \delta_j$, δ_j die Rundungsfehler und τ_j für $j \geq s$ die lokalen Fehler sind.

Die hierbei auftretenden Projektoren P, Q und Q_1 sind wie folgt definiert:

- Q ist ein beliebiger konstanter Projektor auf $ker(A(t))$, $P := I - Q$.
- $Q_1(t)$ ist der kanonische Projektor auf $ker(A_1(t))$ längs $S_1(t)$, wobei $A_1(t) := A(t) + B(t)Q$ und $S_1(t) := \{z \in I\!R^m : B(t)Pz \in im(A(t))\}$.

Schließlich ist hierbei $G_2(t) := A_1(t) + B(t)PQ_1(t)$.

Da diese Darstellung auf den ersten Blick etwas kompliziert erscheint, möchte ich hierzu einige Erläuterungen angeben. Die Projektion Q bewirkt eine Projektion auf die Komponenten, die lediglich algebraisch, d.h. nicht differenziert, in das System eingehen. Die Projektion $PP_1(t)$ liefert eine Projektion auf die Komponenten, für die eine reguläre gewöhnliche Differentialgleichung implizit im System gegeben ist (vgl. hierzu (1.3a)). Die Matrix $G_2(t)$ besitzt die Eigenschaft, daß sie genau dann regulär ist, wenn das System den Index 2 besitzt.
Aus den angegebenen Abschätzungen läßt sich nun ablesen:

- Sind die Startwerte exakt und treten keine Rundungsfehler auf, so gilt:

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

  d.h. die BDF konvergieren formal mit der erwarteten Ordnung.

- Stabilität erreicht man nur für die P–Komponente; in der algebraischen Komponente ist eine schwache Instabilität zu sehen, die darauf zurückzuführen ist, daß bei der Berechnung dieser Komponenten implizit Differentiationen (hier in diskretisierter Form) auftreten. An dieser Stelle sei daran erinnert, daß Differentiationsprobleme schlecht gestellte Probleme sind.

Für die linearen Index–2–ADGen ist das Verhalten der BDF also bereits vollständig geklärt. Wie sieht es nun mit den nichtlinearen Problemen dieser Form aus? Wesentliche Resultate hierzu sind in den Arbeiten von

- Gear, Leimkuhler & Gupta [85],

- Lötstedt & Petzold [86],
- Brenan & Engquist [88],
- März [92]

zu finden.
In der Arbeit von Gear, Leimkuhler & Gupta [85] wurde für Probleme der Form

$$x' + g_1(x, y, t) = 0 \tag{1.8a}$$
$$g_2(x, t) = 0 \tag{1.8b}$$

die Konvergenz der BDF der Ordnung k nach $k+1$ Schritten unter folgenden Voraussetzungen gezeigt:

- Die Funktionen g_1 und g_2 besitzen in einer Umgebung der Lösung des Systems (1.8a), (1.8b) stetige partielle Ableitungen bis zur benötigten Ordnung.
- Die Schrittweiten h_n sind von der Form, daß : $\frac{max\, h_n}{min\, h_n} \leq const.$
- Die Schrittweiten seien so gewählt, daß die für reguläre gewöhnliche Differentialgleichungen bekannte Stabilitätsbedingung erfüllt ist.
- Falls g_1 nichtlinear in y oder g_2 nichtlinear in x ist, so gelte weiter:
  1. η_1, τ und e_0 sind $O(h)$–genau.
  2. η_2 ist $O(h^2)$–genau.
  3. Die Matrix $[{g_2}'_x(x,t){g_1}'_y(x_*(t), y_*(t), t)]^{-1}$ existiert und ist beschränkt für x in einer Umgebung der Lösung $x_*(t)$.

Es gelten dann darüberhinaus die Abschätzungen:

$$\|e_n^x\| = O(e_0 + \tau + \eta_1 + \eta_2),$$
$$\|e_n^y\| = O(e_0 + \tau + \eta_1 + \frac{\eta_2}{h}).$$

Hierbei bezeichnen e_n^x bzw. e_n^y die Fehler in den x– bzw. y–Komponenten, e_0 die Fehler in den Startwerten, τ den maximalen lokalen Fehler, η_1 bzw. η_2 die maximalen Fehler, die in jedem Integrations–Schritt durch Rundung und approximative Lösung der nichtlinearen Gleichungen in der 1. bzw. 2. Gleichung entstehen, und h die maximale Schrittweite.
Für eine etwas allgemeinere Problemklasse erzielten Lötstedt & Petzold [86] folgende Resultate:
Betrachtet wurden Systeme der Form

$$F_1(x, x', y, t) = 0 \tag{1.9a}$$
$$F_2(x, y, t) = 0 \tag{1.9b}$$

Unter den Voraussetzungen, daß für alle t

- die Schrittweite h konstant ist,
- die Ordnung $k \leq 6$ ist,
- die partiellen Ableitungen von F_1 und F_2 in Richtung x, x' und y existieren und beschränkt sind,
- die partielle Ableitung ${F_1}'_{x'}$ eine quadratische Matrix ist und deren Inverse existiert,
- die Inverse des Schur–Komplements (der zu den BDF gehörenden Iterationsmatrix) ${F_2}'_y - h{F_2}'_x(\alpha_0 {F_1}'_{x'} + h{F_1}'_x)^{-1}{F_1}'_y$ existiert,
- die von Null verschiedenen Zeilen der Matrix ${F_2}'_y$ linear unabhängig sind,
- die Fehler in den Startwerten einer Genauigkeit von $O(h^k)$ genügen,
- die Fehler, die durch Rundung und Newton–Abbruch bei Lösung der nichtlinearen Gleichung aus (1.9a) bzw. (1.9b) entstehen, die Genauigkeit $O(h^k)$ bzw. $O(h^{k+1})$ besitzen,

ist die diskrete Lösung nach $k+1$ Schritten genau mit der Größenordnung $O(h^k)$.

In der Arbeit von Brenan & Engquist [88] wurden diese Ergebnisse für semiexplizite Probleme der Form

$$y' = E(t, y, u), \tag{1.10a}$$
$$0 = H(t, y), \tag{1.10b}$$

dahingehend verbessert, daß die angegebene Konvergenz bereits vom ersten Schritt an gilt, vorausgesetzt, daß die Anfangswerte numerisch konsistent sind, d.h.: Die Startwerte in der y–Komponente besitzen eine Genauigkeit von $O(h^{k+1})$.

Offen blieb bei diesen Arbeiten, wie sich die Einzugsbereiche für die Konvergenz des Newton–Verfahrens zur Lösung der nichtlinearen Gleichungen in Bezug auf die Schrittweite verhalten und damit auch die Frage, ob die BDF tatsächlich durchführbar sind. Diese Frage wird teilweise in der Arbeit von März [92] beantwortet, in der auch Stabilitätsabschätzungen für nichtlineare Index–2–Systeme bewiesen wurden.
Für Probleme der allgemeinen Form

$$f(x', x, t) = 0 \tag{1.11}$$

wird dort unter den Voraussetzungen, daß

- die Schrittweiten h_j von der Form sind, daß: $\frac{max\, h_j}{min\, h_j} \leq const$,
  und die BDF für reguläre gewöhnliche DGLen stabil sind,

- die Ordnung $s > 1$ ist,
- die partiellen Ableitungen f'_y und f'_x in einer Umgebung der Trajektorie der Lösung x_* existieren und Lipschitzstetig sind,
- der Nullraum $ker(f'_y(y, x, t))$ konstant ist und
- der Bildraum $im(f'_y(y, x, t))$ unabhängig von y ist,

gezeigt, daß die BDF durchführbar und schwach instabil sind. Falls die Startwerte in der Q_1–Komponente eine Genauigkeit von $O(h^{s+1})$ und in den anderen Komponenten die Genauigkeit $O(h^s)$ besitzen, so konvergiert das Verfahren mit der Genauigkeit $O(h^s)$.
Die hier auftauchenden Projektoren P, Q und Q_1 sind entsprechend den oben angegebenen für den Fall linearer ADGen für die in der Lösung x_* linearisierte ADG von (1.11) gewählt.
Mit der dort gewählten Beweisstrategie ließen sich die Resultate nur für den Fall der Ordnung $s \geq 2$ zeigen. An dieser Stelle sei erwähnt, daß auch in den Beweisen der vorangegangenen zitierten Arbeiten der Fall der Ordnung $s = 1$ gesondert behandelt werden mußte. Es blieb also offen, ob das implizite Euler–Verfahren für solche Problemklassen auch durchführbar ist, und wie sein Stabilitätsverhalten aussieht. Die nächsten beiden Abschnitte geben Antworten auf diese Fragen.

2 Vorbereitende Analyse der Index-2-ADGen

Die Systeme von Algebro–Differentialgleichungen, die hier untersucht werden sollen, sind die quasilinearen ADGen der Form:

$$A(t)\,x'(t) + g(x(t),t) = 0\,. \tag{2.1}$$

Hierbei seien $A : \mathcal{I} \to L(\mathbb{R}^m)$ eine stetige Matrixfunktion, $\mathcal{I} \subseteq \mathbb{R}$ ein offenes Intervall. Zusätzlich sei aus den in Abschnitt 1.3 genannten Gründen $ker(A(t))$ von t unabhängig und bezeichne $Q : \mathcal{I} \to L(\mathbb{R}^m)$ einen konstanten Projektor auf $ker(A(t))$, $P := I - Q$.
Weiter seien $g : D \times \mathcal{I} \to \mathbb{R}^m$ eine beliebige stetige Funktion, $D \subseteq \mathbb{R}^m$ ein offenes Gebiet, und $g(\cdot, t)$ stetig differenzierbar für alle $t \subseteq \mathcal{I}$.

Sei $x_* \in C^1_N$ eine Lösung des Systems (2.1). Nun besitzt die ADG (2.1) den Traktabilitäts–Index 2 lokal um x_*, wenn für alle (x,t) in einer Umgebung U von $(x_*(t),t)$ in $\mathbb{R}^m \times \mathcal{I}$ die Matrix $A_1(x,t) := A(t) + g'_x(x,t)Q$ singulär ist, $\mathrm{rang}(A_1(x,t))$ konstant ist und folgende Beziehung gilt:

$$ker(A_1(x,t)) \cap S_1(x,t) = \{0\} \quad \forall (x,t) \in U,$$

wobei $S_1(x,t) := \{z \in \mathbb{R}^m : g'_x(x,t)Pz \in im(A_1(x,t))\}$.

Um nun ADGen vom Index 2 näher untersuchen zu können, sollen zunächst einige hilfreiche algebraische Überlegungen angestellt werden.

2.1 Algebraische Hilfsbetrachtungen

Einen grundlegenden Zusammenhang zwischen den beim Traktabilitäts-Index auftretenden Räumen und der Auswahl entsprechender Projektoren liefert folgendes Lemma, das sich direkt aus Griepentrog & März [86], Theorem A.13. und Lemma A.14., ableiten läßt.

Lemma 2.1 *Seien $\bar{A}, \bar{B}, \bar{Q} \in L(\mathbb{R}^m)$ gegeben, $\bar{Q}^2 = \bar{Q}$, $\mathrm{im}(\bar{Q}) = ker(\bar{A})$, d.h. $\bar{Q}$ sei ein Projektor auf $ker(\bar{A})$. Bezeichne $\bar{S} := \{z \in \mathbb{R}^m : \bar{B}z \in im(\bar{A})\}$. Dann sind folgende Bedingungen äquivalent:*

(i) *Die Matrix $\bar{G} := \bar{A} + \bar{B}\bar{Q}$ ist regulär.*

(ii) $\mathbb{R}^m = \bar{S} \oplus ker(\bar{A})$.

(iii) $\bar{S} \cap ker(\bar{A}) = \{0\}$.

Wenn $\bar{G}$ regulär ist, so gilt für den kanonischen Projektor $\bar{Q}_s$ (d.h. $\bar{Q}_s$ projeziert $\mathbb{R}^m$ auf $ker(\bar{A})$ längs $\bar{S}$) die Beziehung:

$$\bar{Q}_s = \bar{Q}\bar{G}^{-1}\bar{B}.$$

Beweis:

$(i)\rightarrow(ii)$ Zunächst läßt sich $\mathbb{R}^m$ darstellen als $\bar{S} + ker(\bar{A})$, denn für beliebige $z \in \mathbb{R}^m$ gilt:

$$z = (I - \bar{Q}\bar{G}^{-1}\bar{B})z + \bar{Q}\bar{G}^{-1}\bar{B}z =: z_1 + z_2. \quad (*)$$

Nun liegt z_2 offenbar in $ker(\bar{A})$, da $\bar{Q}$ ein Projektor auf $ker(\bar{A})$ ist. Für z_1 erhält man, daß

$$\bar{B}z_1 = (I - \bar{B}\bar{Q}\bar{G}^{-1})\bar{B}z = \bar{A}\bar{G}^{-1}\bar{B}z \in im(\bar{A}),$$

d.h. $z_1 \in \bar{S}$.
Es bleibt zu zeigen, daß $\bar{S} \cap ker(\bar{A}) = \{0\}$. Sei dazu $x \in \bar{S} \cap ker(\bar{A})$ beliebig. Dann gilt $x = \bar{Q}x$ und es existiert ein $z \in \mathbb{R}^m$, so daß

$$\bar{A}z = \bar{B}x = \bar{B}\bar{Q}x \text{ und somit } \bar{G}^{-1}\bar{A}z = \bar{G}^{-1}\bar{B}\bar{Q}x,$$

d.h. $(I - \bar{Q})z = \bar{Q}x$, also $0 = \bar{Q}x = x$.

$(ii)\rightarrow(iii)$ Dies gilt trivial nach Definition.

$(iii)\rightarrow(i)$ Sei $x \in \mathbb{R}^m$ so gewählt, daß $\bar{G}x = 0$, d.h. $\bar{B}\bar{Q}x = -\bar{A}x$ und somit $\bar{Q}x \in \bar{S}$. Da nun andererseits $\bar{Q}x$ in $\ker(\bar{A})$ liegt, so gilt nach Voraussetzung $x \in ker(\bar{Q})$. Dies bedeutet wiederum $\bar{A}x = 0$, also $x \in im(\bar{Q})$. Damit muß $x = 0$ gelten, und $\bar{G}$ ist regulär.

Aufgrund der Eindeutigkeit der Zerlegung $(*)$, folgt die letzte Behauptung sofort. □

Lemma 2.2 *Sei $x_* \in C_N^1$ eine Lösung des Systems (2.1) und besitze die ADG (2.1) den Traktabilitäts-Index 2 lokal um x_*. Sei außerdem $t \in \mathcal{I}$ beliebig, aber fest gewählt, $P = I - Q$ und $Q_1(t)$ der kanonische Projektor auf* $\ker(A_1(t))$, *$A_1(t) := A_1(x_*(t), t)$, d.h. $Q_1(t)$ projeziert längs $S_1(t) := S_1(x_*(t), t)$.*
Dann gilt:

(i) Die Matrix $A_2(t) := A_1(t) + g_x'(x_*(t), t)PQ_1(t)$ ist regulär.

(ii) $\mathbb{R}^m = S_1(t) \oplus ker(A_1(t))$

(iii) $Q_1(t) = Q_1(t)A_2(t)^{-1}g_x'(x_*(t), t)P, \qquad Q_1(t)Q \equiv 0.$

Beweis: Die Behauptung folgt aus Lemma 2.1, wenn man

$$\bar{A} := A_1(t),\ \bar{B} := g'_x(x_*(t),t)P \text{ und } \bar{Q}_s := Q_1(t)$$

setzt.

□

Bemerkung: Dieses Lemma impliziert: Wenn die ADG (2.1) den Traktabilitäts-Index 2 lokal um x_* besitzt, so besitzt auch die in x_* linearisierte ADG den Traktabilitäts-Index 2.

2.2 Implizites Euler-Verfahren

Sei $x_* \in C^1_N$ eine Lösung des Systems (2.1) und besitze die ADG (2.1) den Traktabilitäts-Index 2 lokal um x_*. Sei weiter π eine Zerlegung des abgeschlossenen Intervalls $[t_0, T] \subset \mathcal{I}$ mit folgenden Eigenschaften:

$$\begin{aligned} \pi : t_0 < t_1 < \cdots < t_N = T, \qquad (2.2)\\ h_{min} \le t_\ell - t_{\ell-1} \le h_{max},\ h_{min} > 0,\ \ell = 1,2,...,N. \end{aligned}$$

Das implizite Eulerverfahren läßt sich dann für ADGen der Form (2.1) in folgender Weise formulieren:

$$x_0 - x_*(t_0) = \delta_0 \qquad (2.3a)$$

$$A(t_\ell)\frac{x_\ell - x_{\ell-1}}{h_\ell} + g(x_\ell, t_\ell) = \delta_\ell \quad ;\ \ell = 1,2,...,N. \qquad (2.3b)$$

Hier bezeichnen δ_0 die Störung im Anfangswert und δ_ℓ $(\ell = 1,2,...,N)$ die Störungen, die durch die numerische Rechnung im ℓ-ten Schritt auftreten. Mit h_ℓ wird wie üblich die Schrittweite im ℓ-ten Schritt bezeichnet, d.h. $h_\ell = t_\ell - t_{\ell-1}$. Zur Vereinfachung der Untersuchungen werden für $\ell = 1,2,...,N$ folgende Bezeichnungen eingeführt:

$$\begin{aligned} \tilde{x}_\ell &:= x_\ell - x_*(t_\ell)\\ \tau_\ell &:= A(t_\ell)\frac{x_*(t_\ell) - x_*(t_{\ell-1})}{h_\ell} + g(x_*(t_\ell), t_\ell)\\ \tilde{g}_\ell(y) &:= g(y + x_*(t_\ell), t_\ell)\,. \end{aligned}$$

Letztere liefert eine stetige Funktion $\tilde{g}_\ell$, die auf $\tilde{D}_\ell$ definiert ist, wobei $\tilde{D}_\ell := \{y \in \mathbb{R}^m : y + x_*(t_\ell) \in D\}$ für jedes $\ell = 1,2,...,N$ eine Nullumgebung ist. Es sei bemerkt, daß das so definerte τ_ℓ den lokalen Fehler des impliziten Eulerverfahrens im ℓ-ten Schritt darstellt und damit bekannter Weise von der Größenordnung $O(h_\ell)$ ist, was durch anschließende Taylorreihenentwicklung leicht zu sehen ist:

$$Px_*(t_{\ell-1}) = Px_*(t_\ell) - h_\ell\,(Px_*)'(t_\ell) + O(h_\ell^2)\,.$$

Da $x_*(t_\ell)$ das System (2.1) löst, so ergibt sich sofort

$$\begin{aligned}\tau_\ell &= A(t_\ell)\frac{x_*(t_\ell)-x_*(t_{\ell-1})}{h_\ell}+g(x_*(t_\ell),t_\ell)\\ &= A(t_\ell)\left(\frac{Px_*(t_\ell)-Px_*(t_{\ell-1})}{h_\ell}-(Px_*)'(t_\ell)\right)\\ &= O(h_\ell)\,.\end{aligned}$$

Es sei daraufhingewiesen, daß für den lokalen Fehler gilt: $\tau_\ell \in im(A(t_\ell))$. Dieser Fakt erlangt an späterer Stelle Bedeutung.

Jetzt hat das System (2.3a),(2.3b) folgende Form,

$$\begin{aligned}x_0-x_*(t_0)-\delta_0 &= 0\\ A(t_\ell)\frac{\tilde{x}_\ell-\tilde{x}_{\ell-1}}{h_\ell}+\tilde{g}_\ell(\tilde{x}_\ell)-\tilde{g}_\ell(0)+\tau_\ell-\delta_\ell &= 0 \quad ;\ \ell=1,2,...,N\,,\end{aligned}$$

welche äquivalent ist zu:

$$x_0-x_*(t_0)-\delta_0 = 0 \tag{2.4a}$$

$$A_\ell\frac{\tilde{x}_\ell-\tilde{x}_{\ell-1}}{h_\ell}+B_\ell\tilde{x}_\ell+\Phi_\ell(\tilde{x}_\ell)+\tau_\ell-\delta_\ell = 0 \quad ;\ \ell=1,2,...,N\ , \tag{2.4b}$$

wobei

$$\begin{aligned}A_\ell &:= A(t_\ell),\\ B_\ell &:= \tilde{g}_\ell'(0),\\ \Phi_\ell(y) &:= \tilde{g}_\ell(y)-\tilde{g}_\ell(0)-\tilde{g}_\ell{}'(0)y \quad ;\ \ell=1,2,...,N\ .\end{aligned} \tag{2.5}$$

Die zuletzt eingeführten Größen haben folgende gewünschten Eigenschaften:

(1) Das Matrix-Büschel $\{A_\ell, B_\ell\}$ besitzt den Traktabilitäts-Index 2 für $\ell = 1,2,...,N$.

(2) Für $\ell = 1,2,...,N$ gilt: $\Phi_\ell(0) = 0$.

(3) Ist $g(\cdot,t)$ als Funktion des Ortes auf D von der Klasse C^1, d.h. einmal stetig differenzierbar, so ist Φ_ℓ stetig differenzierbar und es gilt: $\Phi_\ell'(0) = 0,\ \ell = 1,2,...,N$.

Ziel ist es nun, die Systeme in (2.4b) für jeden Schritt so aufzuspalten, daß sich zunächst die differentiellen Komponenten der Lösung berechnen lassen und dann in Abhängigkeit von diesen die algebraischen Komponenten ermittelt werden können. Dabei seien die differentiellen Komponenten der Lösung diejenigen, die in abgeleiteter Form in die ADG eingehen. Entsprechend werden die anderen

algebraisch genannt. Mit anderen Worten: Px_* stellen die differentiellen und Qx_* die algebraischen Komponenten der Lösung x_* dar. Nun sind Systeme höheren Index' (also auch Index-2-Systeme) gerade dadurch charakterisiert, daß es keine algebraische Transformation gibt, die den differentiellen und den algebraischen Teil vollständig voneinander trennt. Man erreicht jedoch unter Umständen eine Aufspaltung in 3 Teile,

(a) eigentliche Differentialgleichung (diskretisiert)

(b) Zusammenhang von algebraischen und differentiellen Komponenten

(c) rein algebraische Gleichung.

Dies soll im folgenden gezeigt werden. Seien

$$\begin{aligned} A_{1,\ell} &:= A_\ell + B_\ell Q, \\ P_{1,\ell} &:= I - Q_{1,\ell} \\ \text{und } A_{2,\ell} &:= A_{1,\ell} + B_\ell P Q_{1,\ell}, \end{aligned}$$

wobei $Q_{1,\ell}$ der kanonische Projektor auf $ker(A_{1,\ell})$ längs

$$S_{1,\ell} := \{z \in I\!R^m : B_\ell P z \in im(A_{1,\ell})\} \qquad \text{ist.}$$

Mit Lemma 2.2 sind folgende Beziehungen leicht nachzuvollziehen:

$$\begin{aligned} A_{2,\ell}^{-1} A_\ell &= P_{1,\ell} P \\ A_{2,\ell}^{-1} B_\ell &= A_{2,\ell}^{-1} B_\ell P P_{1,\ell} + A_{2,\ell}^{-1} B_\ell P Q_{1,\ell} + A_{2,\ell}^{-1} B_\ell Q \\ &= A_{2,\ell}^{-1} B_\ell P P_{1,\ell} + Q_{1,\ell} + Q. \end{aligned}$$

Multipliziert man die ℓ-te Gleichung von (2.4b) mit der regulären Matrix $A_{2,\ell}^{-1}$, so erhält man die dazu äquivalente Gleichung

$$P_{1,\ell} P \frac{\tilde{x}_\ell - \tilde{x}_{\ell-1}}{h_\ell} + A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{x}_\ell + Q_{1,\ell} \tilde{x}_\ell + Q \tilde{x}_\ell + A_{2,\ell}^{-1} \Phi_\ell(\tilde{x}_\ell) + A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0. \tag{2.6}$$

Unter Beachtung der Tatsache, daß $PP_{1,\ell}$, $QP_{1,\ell}$ und $Q_{1,\ell}$ Projektoren sind, läßt sich die Gleichung (2.6) durch Multiplikation mit diesen Projektoren in folgendes System äquivalent umformen:

$$\begin{aligned} PP_{1,\ell} \frac{\tilde{x}_\ell - \tilde{x}_{\ell-1}}{h_\ell} + PP_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{x}_\ell & \\ + PP_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{x}_\ell) + PP_{1,\ell} A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) &= 0 \end{aligned} \tag{2.7a}$$

$$\begin{aligned} -QQ_{1,\ell} \frac{\tilde{x}_\ell - \tilde{x}_{\ell-1}}{h_\ell} + Q\tilde{x}_\ell + QP_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{x}_\ell & \\ + QP_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{x}_\ell) + QP_{1,\ell} A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) &= 0 \end{aligned} \tag{2.7b}$$

$$Q_{1,\ell} \tilde{x}_\ell + Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{x}_\ell) - Q_{1,\ell} A_{2,\ell}^{-1} \delta_\ell = 0 \tag{2.7c}$$

In der letzten Gleichung (2.7c) verschwindet glücklicher Weise der Einfluß des lokalen Fehlers, da $\tau_\ell \in im(A_\ell)$ (wie bereits weiter oben bemerkt wurde) und $Q_{1,\ell} A_{2,\ell}^{-1} A_\ell \equiv 0$.
Unser vorläufiges Ziel ist fast erreicht. Bisher wurde lediglich davon Gebrauch gemacht, daß die in der Lösung x_* linearisierte ADG den Traktabilitäts–Index 2 besitzt. Andererseits gibt es Beispiele (siehe R. März [91]), die zeigen, daß für die eindeutige Lösbarkeit einer solchen ADG der Index 2 in einer offenen Umgebung gegeben sein sollte. Das folgende Lemma (vgl. Lemma 2.2 in März [91]) gibt hierfür eine hinreichende Bedingung an, die wir auch im weiteren benutzen werden.

Lemma 2.3 *Es sei $x_* \in C_N^1$ eine Lösung des Systems (2.1) und besitze die in x_* linearisierte ADG den Traktabilitäts-Index 2. Dann sind mit den zuvor eingeführten Projektoren folgende Bedingungen äquivalent:*

(i) $Q_1(t)A_2^{-1}(t)(g(y,t) - g(Py,t)) = 0$
für y in einer Umgebung $U_0(t) \subset D$ von $x_*(t)$, $t \in \mathcal{I}$.

(ii) $S(t)(g(y,t) - g(Py,t)) \in im(S(t)B(t)Q)$
für y in einer Umgebung $U_0(t) \subset D$ von $x_*(t)$ und $(I - S(t))$ ein beliebiger Projektor auf $im(A(t))$, $t \in \mathcal{I}$.

Ferner gilt unter Voraussetzung von einer der beiden Bedingungen:
Für jedes abgeschlossene Intervall $\mathcal{I}_0 \subset \mathcal{I}$ existiert ein Radius $\delta > 0$, so daß die ADG (2.1) in der Umgebung $U := \bigcup_{t \in \mathcal{I}_0} B((x_*(t),t),\delta)$ den Traktabilitäts–Index 2 lokal um x_* besitzt.

Beweis: Es wird zunächst die Äquivalenz der angegebenen Bedingungen gezeigt. Es sei $t \in \mathcal{I}$ fest gewählt, $y \in U_0(t)$ und $z := g(y,t) - g(Py,t)$.
$(i) \rightarrow (ii)$

$$\begin{aligned} S(t)z &= S(t)A_2(t)A_2^{-1}(t)z = S(t)B(t)QA_2^{-1}(t)z + S(t)B(t)PQ_1(t)A_2^{-1}(t)z \\ &= S(t)B(t)QA_2^{-1}(t)z \in \mathrm{im}(S(t)B(t)Q). \end{aligned}$$

$(ii) \rightarrow (i)$ Da $\mathrm{im}(A(t)) \subset \ker(Q_1(t)A_2^{-1}(t))$ und $(I - S(t))$ ein Projektor auf $\mathrm{im}(A(t))$ ist, so gilt die Identität

$$Q_1(t)A_2^{-1}(t) \equiv Q_1(t)A_2^{-1}(t)S(t).$$

Damit gilt:

$$\begin{aligned} Q_1(t)A_2^{-1}(t)z &= Q_1(t)A_2^{-1}(t)S(t)z = Q_1(t)A_2^{-1}(t)S(t)B(t)Qy, \text{ gew. } y \in \mathbb{R}^m \\ &= Q_1(t)A_2^{-1}(t)S(t)A_1(t)y = Q_1(t)A_2^{-1}(t)A_1(t)y = 0. \end{aligned}$$

Analog zum Beweis von Lemma 2.2 in März [91] läßt sich auch die letzte Behauptung zeigen. Sei wieder $t \in \mathcal{I}$ fest gewählt. So gilt aufgrund der Voraussetzungen für $y \in U_0(t)$:

$$Q_1(t)A_2^{-1}(t)g_y'(y,t)Q = Q_1(t)A_2^{-1}(t)g_y'(Py,t)PQ = 0.$$

Dies bedeutet, daß für $A_1(y,t) := A(t) + g_y'(y,t)Q$ die Relation

$$Q_1(t)A_2^{-1}(t)A_1(y,t) = 0$$

erfüllt ist. Da $Q_1(t)$ der kanonische Projektor auf $\ker(A_1(t))$ längs $S_1(t) = \{z \in I\!R^m : B(t)Pz \in \mathrm{im}(A_1(t))\}$ ist, so ist nun

$$\mathrm{im}(A_2^{-1}(t)A_1(y,t)) \subset S_1(t).$$

Demzufolge gilt

$$rang(A_2^{-1}(t)A_1(y,t)) \leq dim(S_1(t)) = rang(P_1(t)),$$

und mit der Kenntnis, daß

$$A_2^{-1}(t)A_1(x_*(t),t) = A_2^{-1}(t)A_1(t) = P_1(t)$$

stets erfüllt ist, finden wir jetzt eine Umgebung $U_1(t) \subset U_0(t)$ von $x_*(t)$, so daß der Rang der Matrix $A_2^{-1}(t)A_1(y,t)$ für $y \in U_1(t)$ konstant ist. Also ist auch die Dimension des Raumes $\ker(A_1(y,t))$ für $y \in U_1(t)$ konstant. Sei jetzt $Q_1(y,t)$ der orthogonale Projektor auf $\ker(A_1(y,t)$, $y \in U_1(t)$. Dieser hängt dann stetig von y ab, da $A_1(y,t)$ stetig von y abhängt. Damit ist auch

$$A_2(y,t) := A_1(y,t) + g_y'(y,t)PQ_1(y,t)$$

stetig bezüglich y.
Da die in x_* linearisierte ADG den Traktabilitäts-Index 2 besitzt, so ist außerdem

$$A_2(x_*(t),t) := A_1(x_*(t),t) + g_x'(x_*(t),t)PQ_1(t)$$

regulär und stetig bezüglich t. Somit hängt

$$A_2(y,t) := A_1(y,t) + g_y'(y,t)PQ_1(y,t)$$

stetig von (y,t) ab. Ferner gilt die gleichmäßige Stetigkeit auf $\{(y,t) \mid \|y - x_*(t)\| \leq \delta_0,\ t \in \mathcal{I}_0\}$, falls $\mathcal{I}_0 \subset \mathcal{I}$ ein abgeschlossenes Intervall ist und $\delta_0 > 0$ so gewählt ist, daß $\{y \mid \|y - x_*(t)\| < \delta_0,\ t \in \mathcal{I}_0\} \subset D$. Da die Determinantenfunktion ebenfalls gleichmäßig stetig auf kompakten Mengen ist, so findet man nun einen Radius $\delta > 0$, so daß für alle $t \in \mathcal{I}_0$ die Matrix $A_2(y,t)$ mit $y \in B(x_*(t),\delta)$ regulär ist. □

Bemerkung: Die im obigen Lemma angegebenen Bedingungen sind z.B. für Index–2–Hessenberg–Systeme trivial erfüllt. Eine ausführlichere Diskussion dieser Bedingungen ist in März[91] nachzulesen.

Für die weiteren Betrachtungen seien die Voraussetzungen von Lemma 2.3 erfüllt. Sei im folgenden für $\ell = 1, 2, ..., N$:

$$\tilde{u}_\ell := PP_{1,\ell}\tilde{x}_\ell, \quad \tilde{v}_\ell := Q_{1,\ell}\tilde{x}_\ell, \quad \tilde{w}_\ell := Q\tilde{x}_\ell,$$

Dann hat das System (2.7a)–(2.7c) die Gestalt:

$$\begin{aligned} \frac{\tilde{u}_\ell - \tilde{u}_{\ell-1}}{h_\ell} + \frac{1}{h_\ell}P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ + PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell + \tilde{w}_\ell) + PP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \quad (2.8a) \end{aligned}$$

$$\begin{aligned} -Q\frac{\tilde{v}_\ell - \tilde{v}_{\ell-1}}{h_\ell} - \frac{1}{h_\ell}Q(Q_{1,\ell} - Q_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + QP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ + \tilde{w}_\ell + QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell + \tilde{w}_\ell) + QP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \quad (2.8b) \end{aligned}$$

$$\tilde{v}_\ell + Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell) - \tilde{\delta}_\ell = 0, \quad (2.8c)$$

Das Gleichungssystem (2.8a)-(2.8c) hat jetzt die erwarteten Eigenschaften. Die 1. Gleichung widerspiegelt die eigentlich zugrundeliegende diskretisierte Differentialgleichung in der u–Komponente. Die 3. Gleichung ist rein algebraisch und ermöglicht die Bestimmung der v–Komponente, die lediglich algebraisch von den anderen Komponenten abhängt. Die 2. Gleichung schließlich repräsentiert den algebraischen Zusammenhang zwischen der algebraischen (w–) Komponente und den differentiellen Komponenten, wobei letztere zum Teil (Qv) differenziert (in diskretisierter Form) eingehen.

3 Stabilitäts- und Konvergenzresultate

Bevor die erzielten Stabilitäts- umd Konvergenzresultate angegeben werden, sollen hier zwei Lemmata bewiesen werden, die an späterer Stelle gebraucht werden.

Lemma 3.1 *Seien $E_1, E_2, \cdots, E_n, F$ $(n \in \mathcal{N}, n \geq 2)$ Banachräume und f eine C^1-Abbildung einer offenen Teilmenge A von $E_1 \times E_2 \times \cdots \times E_n$ in F. Im Punkt $x^0 := (x_1^0, x_2^0, \cdots, x_n^0)$ von A sei $f(x^0) = 0$, und die partielle Ableitung $D_1 f(x^0)$ sei ein linearer Homöomorphismus von E_1 auf F. Dann gibt es eine zusammenhängende offene Umgebung U von $(x_2^0, \cdots, x_n^0)$ in $E_2 \times \cdots \times E_n$ und eine eindeutig bestimmte C^1–Abbildung u von U in E_1 derart, daß die Gleichung $u(x_2^0, \cdots, x_n^0) = x_1^0$ erfüllt ist und für alle $(x_2, \cdots, x_n) \in U$ die Beziehungen $(u(x_2, \cdots, x_n), x_2, \cdots, x_n) \in A$ und $f(u(x_2, \cdots, x_n), x_2, \cdots, x_n) = 0$ gelten.*
Weiter gilt für jedes $(x_2, \cdots, x_n) \in U$ folgende Ungleichung:

$$\begin{aligned} \|u(x_2, \cdots, x_n) - u(x_2^0, \cdots, x_n^0)\| \;\leq\; & (2\|D_{x_2} u(x_2^0, \cdots, x_n^0)\| + 1)\|x_2 - x_2^0\| \\ & + \quad \cdots \\ & + (2\|D_{x_n} u(x_2^0, \cdots, x_n^0)\| + 1)\|x_n - x_n^0\| \end{aligned}$$

Beweis: Die Existenz einer solchen Umgebung U und einer solchen (in U eindeutig bestimmten) Abbildung ergibt sich aus dem Satz über implizite Funktionen. Es soll nun gezeigt werden, daß die angegebene Ungleichung gilt. Dazu sei $(s_2, \cdots, s_n) \in E_2 \times \cdots \times E_n$ so gewählt, daß der Punkt $(x_2^0 + s_2, \cdots, x_n^0 + s_n)$ in der Umgebung U liegt. Sei weiter $s_1 := u(x_2^0 + s_2, \cdots, x_n^0 + s_n) - u(x_2^0, \cdots, x_n^0)$. Nach Voraussetzung gilt dann die Beziehung

$$f(u(x_2^0, \cdots, x_n^0) + s_1, x_2^0 + s_2, \cdots, x_n^0 + s_n) = 0 \quad .$$

Außerdem erhält man $s_1 \to 0$, falls $(s_2, \cdots, s_n) \to (0, \cdots, 0)$. Da nun f in $(x_2^0, \cdots, x_n^0)$ stetig differenzierbar ist, so existiert zu jedem $\delta > 0$ ein $r > 0$ derart, daß für alle $(s_2, \cdots, s_n)$ mit $\|s_i\| \leq r$, $i = 2, \cdots, n$, die Ungleichung

$$\|f(x^0 + s) - f(x^0) - \sum_{i=1}^{n} D_i f(x^0) s_i\| \leq \delta \sum_{i=1}^{n} \|s_i\|$$

erfüllt ist, wenn $s := (s_1, s_2, \cdots, s_n)$. Nach Definition ist diese Ungleichung mit folgender äquivalent:

$$\|\sum_{i=1}^{n} D_i f(x^0) s_i\| \leq \delta \sum_{i=1}^{n} \|s_i\| \quad .$$

Da $D_1 f(x^0)$ ein linearer Homöomorphismus von E_1 auf F ist, so läßt sich aus der vorigen Beziehung schließen, daß

$$\|s_1 + \sum_{i=2}^{n} (D_1 f(x^0))^{-1} D_i f(x^0) s_i\| \leq \delta \|(D_1 f(x^0))^{-1}\| \sum_{i=1}^{n} \|s_i\| \quad .$$

Sei nun δ so gewählt worden, daß $\delta \|(D_1 f(x^0))^{-1}\| \leq \frac{1}{2}$. Dann folgt mit Hilfe der Dreiecksungleichung:

$$\|s_1\| - \sum_{i=2}^{n} \|(D_1 f(x^0))^{-1} D_i f(x^0)\| \, \|s_i\| \leq \frac{1}{2} \sum_{i=1}^{n} \|s_i\| \quad ,$$

also

$$\|s_1\| \leq \sum_{i=2}^{n} (2\|(D_1 f(x^0))^{-1} D_i f(x^0)\| + 1) \|s_i\| \quad .$$

Sei o.B.d.A. U bereits so klein gewählt worden, daß $U \subseteq B(x^0, r)$, dann widerspiegelt die letzte Gleichung nach Definition von $s_1, s_2, \cdots, s_n$ gerade die Behauptung.

□

Lemma 3.2 *Sei $x_* \in C_N^1$ eine Lösung des Systems (2.1). Dann gilt für die in (2.5) definierten Funktionen Φ_ℓ ($\ell = 1, 2, ..., N$):*
Für jedes $\epsilon > 0$ existiert ein (von der Zerlegung unabhängiger) Radius $\delta(\epsilon)$, so daß für alle $z \in \mathbb{R}^m$ mit $\|z\| \leq \delta(\epsilon)$ gilt:

(i) $\|\Phi_\ell(z)\| \leq \epsilon \, \|z\|, \quad \|\Phi'_\ell(z)\| \leq \epsilon$

(ii) Wenn $Q_1(t) A_2^{-1}(t) g(\cdot, t)$ stetig differenzierbar ist, so gilt auch:
$\|Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(z)\| \leq \epsilon \, \|z\|, \quad \|Q_{1,\ell} A_{2,\ell}^{-1} \Phi'_\ell(z)\| \leq \epsilon$

Beweis: Zunächst ist g'_x nach Voraussetzung stetig. Ferner ist $x_*(\cdot)$ stetig auf $[t_0, T]$ und damit die Menge

$$\{ \, x_*(t) \mid t \in [t_0, T] \, \}$$

kompakt in $\mathbb{R}^m$. Dann existiert ein Radius $r > 0$, so daß die Menge

$$M := \{ \, (z + x_*(t)) \mid t \in [t_0, T], \ \|z\| \leq r, \ z \in \mathbb{R}^m \, \}$$

eine kompakte Teilmenge von D ist. Dann ist g'_x gleichmäßig stetig auf M und es gilt:
$\forall \, \epsilon > 0 \ \exists \, \delta(\epsilon) > 0, \ \delta(\epsilon) < r \ \forall \, t \in [t_0, T]$:

$$\|z\| \leq \delta(\epsilon) \ \Rightarrow \ \|g(z + x_*(t), t) - g(x_*(t), t) - g'_x(x_*(t), t) z\| \leq \epsilon \, \|z\| \text{ und}$$

$$\|z\| \leq \delta(\epsilon) \ \Rightarrow \ \|g'_x(z + x_*(t), t) - g'_x(x_*(t), t)\| \leq \epsilon.$$

Nach Definition von Φ_ℓ folgt hieraus die Behauptung *(i)*.
Analog läßt sich auch *(ii)* beweisen.

□

3.1 Allgemeiner Konvergenzsatz für das implizite Euler-Verfahren

Der folgende Satz stellt das Hauptresultat dieser Arbeit dar. Es wird gezeigt, daß bei hinreichend genauen Startwerten und hinreichend kleinen Schrittweiten das implizite Euler–Verfahren für ADGen der Form (2.1) eine Lösung liefert, die gegen die exakte Lösung konvergiert, wenn nur die Änderung der numerischen Störungen im Verhältnis zur Schrittweite klein bleibt. Dies war nach den bisherigen Ergebnissen der Untersuchungen zum impliziten Euler–Verfahren zu erwarten gewesen. Überraschend ist jedoch, daß die Störungen i.a. auch wesentlichen Einfluß auf die differentiellen Komponenten der Lösung des Verfahrens haben können. Daß dieser Fakt weder an der Beweistechnik noch an der allgemeineren Form der ADG liegt, macht das im Anschluß an den Beweis des folgenden Satzes angeführte Beispiel einer ADG in Hessenberg–Form deutlich.

Satz 3.3 *Sei $x_* \in C_N^1$ eine Lösung des Systems (2.1) und besitze die ADG (2.1) den Traktabilitäts-Index 2. Mit den Bezeichnungen aus Lemma 2.2 sei $Q_1(\cdot)$ stetig differenzierbar auf $\mathcal{I}$ und $Q_1(t)A_2^{-1}(t)g(\cdot,t)$ zweimal stetig differenzierbar. Sei außerdem für alle $t \in \mathcal{I}$ die Bedingung*

$$S(t)(g(y,t) - g(Py,t)) \in im(S(t)B(t)Q),$$

$y \in U_0(t)$, $U_0(t) \subset D$ Umgebung von $x_*(t)$, $(I - S(t))$ ein beliebiger Projektor auf $im(A(t))$, aus Lemma 2.3 erfüllt.

Dann existieren Konstanten $\epsilon > 0$, $C > 0$ und $H_{max} > 0$, so daß für jede Zerlegung (2.2) mit $h_{max} \le H_{max}$ folgende Implikation wahr ist:
Wenn die Relationen

$$\begin{aligned} \|\delta_\ell\| &\le \epsilon, \quad \ell = 0,1,...,N \\ \frac{1}{h_1}\|Q_{1,0}(A_{2,0}^{-1}(g(x_0,t_0) - g(x_*(t_0),t_0)) - \delta_0)\| &\le \epsilon, \\ \frac{1}{h_\ell}\|Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - Q_{1,\ell-1}A_{2,\ell-1}^{-1}\delta_{\ell-1}\| &\le \epsilon, \quad \ell = 2,...,N \end{aligned}$$

erfüllt sind, so liefert das implizite Eulerverfahren eine Lösung x_ℓ zum Zeitpunkt t_ℓ, $\ell = 1,2,...,N$ und es gilt:

(i)

$$\begin{aligned} \max_{\ell=1,2,...,N}\|P(x_*(t_\ell) - x_\ell)\| \le C \Big[&\|x_*(t_0) - x_0\| + \max_{\ell=1,2,...,N}\|\delta_\ell - \tau_\ell\| \\ &+ \|Q_{1,0}(A_{2,0}^{-1}(g(x_0,t_0) - g(x_*(t_0),t_0)) - \delta_0)\| \\ &+ \max_{\ell=2,...,N}\frac{1}{h_\ell}\|Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - Q_{1,\ell-1}A_{2,\ell-1}^{-1}\delta_{\ell-1}\| \Big] \end{aligned}$$

(ii)

$$\begin{aligned}\max_{\ell=1,2,\ldots,N}\|Q(x_*(t_\ell)-x_\ell)\| \le C\ \Big[\ &\|x_*(t_0)-x_0\| + \max_{\ell=1,2,\ldots,N}\|\delta_\ell-\tau_\ell\| \\ &+\frac{1}{h_1}\|Q_{1,0}(A_{2,0}^{-1}(g(x_0,t_0)-g(x_*(t_0),t_0))-\delta_0)\| \\ &+\max_{\ell=2,\ldots,N}\frac{1}{h_\ell}\|Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - Q_{1,\ell-1}A_{2,\ell-1}^{-1}\delta_{\ell-1}\|\ \Big]\ .\end{aligned}$$

Bemerkung: Die in der P–Komponente auftretende Instabilität hat ihre Ursache in der Aufsummierung der Defekte in den ableitungsfreien Gleichungen

$$\|Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - Q_{1,\ell-1}A_{2,\ell-1}^{-1}\delta_{\ell-1}\|.$$

Die Instabilität in der Q–Komponente rührt zusätzlich vom Eingehen von durch die Differentiation bedingten Termen der Form

$$\frac{Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - Q_{1,\ell-1}A_{2,\ell-1}^{-1}\delta_{\ell-1}}{h_\ell}$$

her.

Beweis: Sei zunächst $\ell \in \{2,...,N\}$ fest gewählt und der ℓ-te Schritt des impliziten Euler-Verfahrens entsprechend der Herleitung in Kapitel 3 äquivalent umgeformt in das System:

$$\begin{aligned}&\frac{\tilde u_\ell-\tilde u_{\ell-1}}{h_\ell}+\frac{1}{h_\ell}P(P_{1,\ell}-P_{1,\ell-1})(\tilde u_{\ell-1}+P\tilde v_{\ell-1})+PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde u_\ell \\ &\qquad +PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde u_\ell+P\tilde v_\ell+\tilde w_\ell)+PP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell-\delta_\ell)=0 \qquad (3.1a)\end{aligned}$$

$$\begin{aligned}&-Q\frac{\tilde v_\ell-\tilde v_{\ell-1}}{h_\ell}-\frac{1}{h_\ell}Q(Q_{1,\ell}-Q_{1,\ell-1})(\tilde u_{\ell-1}+P\tilde v_{\ell-1})+QP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde u_\ell \\ &\qquad +\tilde w_\ell+QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde u_\ell+P\tilde v_\ell+\tilde w_\ell)+QP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell-\delta_\ell)=0 \qquad (3.1b)\end{aligned}$$

$$\tilde v_\ell+Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde u_\ell+P\tilde v_\ell)-\tilde\delta_\ell=0, \qquad (3.1c)$$

wobei $\tilde\delta_\ell := Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell$. In Anlehnung an die Gleichung (3.1c) definieren wir die Funktion:

$$F_\ell(v,u,\delta) := v-\delta+Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u+Pv). \qquad (3.2)$$

Dann ist F_ℓ zweimal stetig differenzierbar,

$$F_\ell(0)=0,\ F'_{\ell_v}(0)=I,$$

und wir können folgende Behauptung zeigen:

Behauptung 1.
Es existieren ein von der Zerlegung unabhängiger Radius α und eine eindeutig bestimmte C^2-Funktion

$$f_\ell(u,\delta) \;:\; B(0,\alpha) \;\to\; B(0,\rho)$$

mit den Eigenschaften:

(i) $F_\ell(f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell),\tilde{u}_\ell,\tilde{\delta}_\ell) = 0$

(ii) $f_\ell(0) = 0,\;\; f_\ell'(0) = (0, I)$

(iii) $f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell) \equiv Q_{1,\ell} f(\tilde{u}_\ell,\tilde{\delta}_\ell)$

(iv) $\|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)\| \le \|\tilde{u}_\ell\| + 3\|\tilde{\delta}_\ell\|$.

Die Aussage *(iii)* ist eine einfache Folgerung von *(i)*. Die Richtigkeit der Aussagen *(i),(ii)* bzw. *(iv)* wäre offensichtlich mit Hilfe des Satzes über implizite Funktionen bzw. Lemma 3.1, wenn der Radius α abhängig von ℓ gewählt werden kann. Um nun zu beweisen, daß diese Aussagen auch für einen von der Zerlegung unabhängigen Radius α gelten, soll folgende "Anwendung" des Banach'schen Fixpunktsatzes (siehe Dieudonné [85], S.249) verwendet werden:

Lemma 3.4 *Es seien E, F zwei Banachräume, U bzw. V eine offene Kugel in E bzw. F mit dem Mittelpunkt 0 und dem Radius ρ bzw. α. Ferner sei v eine stetige Abbildung von $U \times V$ in F derart, daß $\|v(x,y_1) - v(x,y_2)\| \le k\,\|y_1 - y_2\|$ für $x \in U$, $y_1 \in V$, $y_2 \in V$ erfüllt ist; dabei sei k eine der Bedingung $0 \le k < 1$ genügende Konstante. Ist dann $\|v(x,0)\| \le \rho(1-k)$ für jedes $x \in U$, so existiert eine eindeutig bestimmte Abbildung f von U in V derart, daß $f(x) = v(x,f(x))$ für jedes $x \in U$ erfüllt ist. Die Abbildung f ist auf U stetig.*

Für die Abbildung $p_\ell(v,u,\delta) := v - F_\ell(v,u,\delta)$ gilt zunächst:

$$\begin{aligned}
&\|p_\ell(v_1,u,\delta) - p_\ell(v_2,u,\delta)\| \\
&\quad= \|Q_{1,\ell}A_{2,\ell}^{-1}(\Phi_\ell(u+Pv_1) - \Phi_\ell(u+Pv_2))\| \\
&\quad= \|\int_0^1 Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell'(u+Pv_2+s(Pv_1-Pv_2))ds(Pv_1-Pv_2)\| \\
&\quad\le \int_0^1 \|Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell'(u+Pv_2+s(Pv_1-Pv_2))\|ds\|P\|\|v_1-v_2\|\,. \qquad (3.3)
\end{aligned}$$

Nach Lemma 3.2 existieren nun Radien α_1 und ρ (unabhängig von der Zerlegung), so daß für alle $u \in B(0,\alpha_1)$ und alle $(Pv_2 + s(Pv_1 - Pv_2)) \in B(0,\rho)$

$$\|Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell'(u+Pv_2+s(Pv_1-Pv_2))\| \le \frac{1}{2\|P\|}$$

gilt. Da die Kugel $B(0,\rho)$ in $\mathbb{R}^m$ konvex ist, so gilt die letzte Ungleichung offenbar für alle $s \in [0,1]$, wenn $v_1 \in B(0,\rho)$ und $v_2 \in B(0,\rho)$. Die Ungleichung (3.3) liefert dann für $u \in B(0,\alpha_1)$, $v_1 \in B(0,\rho)$ und $v_2 \in B(0,\rho)$:

$$\|p_\ell(v_1,u,\delta) - p_\ell(v_2,u,\delta)\| \leq \frac{1}{2}\|v_1 - v_2\| \,. \tag{3.4}$$

Andererseits gilt:

$$\begin{aligned} \|p_\ell(0,u,\delta)\| &= \|\delta - Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u)\| \\ &\leq \|\delta\| + \|Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u)\| \end{aligned}$$

Nun existiert wieder nach Lemma 3.2 ein von der Zerlegung unabhängiger Radius α_2, so daß

$$\|Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u)\| \leq \|u\|$$

für jedes $u \in B(0,\alpha_2)$ gilt. Man erhält:

$$\|p_\ell(0,u,\delta)\| \leq \|\delta\| + \|u\| \tag{3.5}$$

Wählt man nun $\alpha := min\{\frac{1}{4}\rho, \alpha_1, \alpha_2\}$, so liefern die Ungleichungen (3.4) und (3.5) die gewünschten Abschätzungen:

$$\begin{aligned} \|p_\ell(0,u,\delta)\| &< \frac{1}{2}\rho \\ \|p_\ell(v_1,u,\delta) - p_\ell(v_2,u,\delta)\| &\leq \frac{1}{2}\|v_1 - v_2\| \end{aligned}$$

für $(u,\delta) \in B((0,0),\alpha)$ und $v_1, v_2 \in B(0,\rho)$.
Wenden wir jetzt Lemma 3.4 an, so erhalten wir:
Es existiert eine eindeutig bestimmte Abbildung f_ℓ, die die Kugel $B(0,\alpha)$ in die Kugel $B(0,\rho)$ derart abbildet, daß

$$f_\ell(u,\delta) = p_\ell(f_\ell(u,\delta),u,\delta)$$

für alle $(u,\delta) \in B(0,\alpha)$ erfüllt ist, und diese Abbildung f_ℓ ist stetig. Wir erhalten weiter, daß

$$F_\ell(f_\ell(u,\delta),u,\delta) = 0 \quad .$$

Es bleibt zu zeigen, daß f_ℓ eine C^2-Funktion ist. Dies läßt sich aber in üblicher Art und Weise wie beim Beweis des Satzes über implizite Funktionen unabhängig von ℓ zeigen.
Verfolgt man den Beweis von Lemma 3.1, so genügt es für die Abschätzung *(iv)* zu zeigen, daß es einen von der Zerlegung unabhängigen Radius r gibt, so daß für $(\tilde{u}_\ell, \tilde{\delta}_\ell) \in \tilde{D}_\ell \times \mathbb{R}^m$ mit $\|\tilde{u}_\ell\| \leq r$, $\|\tilde{\delta}_\ell\| \leq r$ die Ungleichung

$$\|F_\ell(f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell),\tilde{u}_\ell,\tilde{\delta}_\ell) - f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell) + \tilde{\delta}_\ell\| \leq \frac{1}{2}(\|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)\| + \|\tilde{u}_\ell\| + \|\tilde{\delta}_\ell)\|)$$

erfüllt ist. Nach Definition von F_ℓ ist diese Ungleichung äquivalent zu

$$\|Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell))\| \leq \frac{1}{2}(\|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)\| + \|\tilde{u}_\ell\| + \|\tilde{\delta}_\ell)\|). \tag{3.6}$$

Es existiert aber wieder nach Lemma 3.2 ein (von der Zerlegung unabhängiger) Radius r, so daß für $\|\tilde{u}_\ell\| \leq r$, $\|\tilde{\delta}_\ell\| \leq r$ mit $\|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)\| \leq r$ gilt:

$$\|Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell))\| \leq \frac{1}{2(1+\|P\|)}(\|\tilde{u}_\ell\| + \|P\|\,\|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)\|).$$

Sei nun o.B.d.A. $\rho < r$, dann folgt unmittelbar die Ungleichung (3.6) und damit die Behauptung 1.

Die soeben bewiesene Aussage gilt natürlich auch für den vorangegangen Euler-Schritt, d.h. für die C^2–Funktion

$$F_{\ell-1}(v,u,\delta) := v - \delta + Q_{1,\ell-1}A_{2,\ell-1}^{-1}\Phi_{\ell-1}(u+Pv). \tag{3.7}$$

gilt:

$$F_{\ell-1}(0) = 0,\ F'_{\ell-1_v}(0) = I,$$

und es existiert wieder eine eindeutig bestimmte C^2-Funktion

$$f_{\ell-1}(u,\delta)\ :\ B(0,\alpha)\ \rightarrow\ B(0,\rho)$$

mit folgenden Eigenschaften:

$$\begin{aligned} F_{\ell-1}(f_{\ell-1}(\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1}),\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1}) &= 0 \\ f_{\ell-1}(0) = 0,\ f'_{\ell-1}(0) &= (0,I) \\ f_{\ell-1}(\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1}) &\equiv Q_{1,\ell-1}f(\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1})\ . \end{aligned}$$

Berücksichtigt man diese Resultate, so bleibt anstelle des Systems (3.1a)-(3.1c) folgendes System zu lösen:

$$\begin{aligned} \frac{\tilde{u}_\ell - \tilde{u}_{\ell-1}}{h_\ell} + PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ + \frac{1}{h_\ell}P(P_{1,\ell}-P_{1,\ell-1})(\tilde{u}_{\ell-1}+P\tilde{v}_{\ell-1}) \\ + PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)+\tilde{w}_\ell) + PP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) &= 0 \end{aligned} \tag{3.8a}$$

$$\begin{aligned} -Q\frac{f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell) - f_{\ell-1}(\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1})}{h_\ell} + QP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ - \frac{1}{h_\ell}Q(Q_{1,\ell}-Q_{1,\ell-1})(\tilde{u}_{\ell-1}+P\tilde{v}_{\ell-1}) + \tilde{w}_\ell \\ + QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)+\tilde{w}_\ell) + QP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) &= 0 \end{aligned} \tag{3.8b}$$

$$\tilde{v}_\ell - f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell) = 0. \tag{3.8c}$$

Um nun die Gleichung (3.8b) nach $\tilde{w}_\ell$ auflösen zu können, ist es offenbar notwendig, die Differenz von $f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) - f_{\ell-1}(\tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1})$ näher auf ihr Verhältnis zur Schrittweite h_ℓ hin zu untersuchen. Dazu schreiben wir die Gleichung

$$F_{\ell-1}(\tilde{v}_{\ell-1}, \tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1}) = 0$$

in folgender Form:

$$\tilde{v}_{\ell-1} - \tilde{\delta}_{\ell-1} + Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) - \zeta_\ell = 0 \ ,$$

wobei

$$\zeta_\ell := Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) - Q_{1,\ell-1} A_{2,\ell-1}^{-1} \Phi_{\ell-1}(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}).$$

Dann gilt entsprechend den Eigenschaften von F_ℓ für die Funktion

$$\Psi_\ell(v, u, \delta, \zeta) := v - \delta + Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(u + Pv) - \zeta \ ,$$

daß Ψ_ℓ von der Klasse C^2 ist und

$$\Psi_\ell(0) = 0, \ \Psi'_{\ell_v}(0) = I.$$

Es existieren ein von der Zerlegung unabhängiger Radius γ und eine eindeutig bestimmte C^2-Funktion

$$\psi_\ell(u, \delta, \zeta) \ : \ B(0, \gamma) \ \rightarrow \ B(0, \gamma')$$

mit den Eigenschaften:

$$\begin{aligned} \Psi_\ell(\psi_\ell(u, \delta, \zeta), u, \delta, \zeta) &= 0 \\ \psi_\ell(0) = 0, \ \psi'_\ell(0) &= (0, I, I) \\ \|\psi_\ell(u, \delta, \zeta)\| &\leq \|u\| + 3\|\delta\| + 3\|\zeta\|. \end{aligned} \tag{3.9}$$

Außerdem haben wir nach Definition von Ψ_ℓ:

$$F_\ell(\tilde{v}_\ell, \tilde{u}_\ell, \tilde{\delta}_\ell) = \Psi_\ell(\tilde{v}_\ell, \tilde{u}_\ell, \tilde{\delta}_\ell, 0) \ .$$

Aufgrund der Eindeutigkeit der Funktionen f_ℓ und ψ_ℓ gilt:

$$f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) = \psi_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell, 0)$$

für

$$\|\tilde{u}_\ell\| \ \leq \ \min\{\alpha, \gamma\} \quad \text{und} \quad \|\delta_\ell\| \ \leq \ \min\{\alpha, \gamma\} \ .$$

Andererseits war ζ_ℓ gerade so definiert, daß :

$$F_{\ell-1}(\tilde{v}_{\ell-1}, \tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1}) = \Psi_\ell(\tilde{v}_{\ell-1}, \tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1}, \zeta_\ell) \ .$$

Die Eindeutigkeit der Funktionen $f_{\ell-1}$ und ψ_ℓ liefert:

$$f_{\ell-1}(\tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1}) = \psi_\ell(\tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1}, \zeta_\ell)$$

für

$$\|\tilde{u}_{\ell-1}\| \;\leq\; \min\{\beta, \gamma\}\ , \quad \|\delta_{\ell-1}\| \;\leq\; \min\{\beta, \gamma\} \quad \text{und} \quad \|\zeta_\ell\| \;\leq\; \gamma\ .$$

Es bleibt die Frage, unter welchen Bedingungen, die Relation $\|\zeta_\ell\| \;\leq\; \gamma$ gilt. Deshalb wird jetzt ζ_ℓ genauer untersucht:

$$\zeta_\ell = Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) - Q_{1,\ell-1} A_{2,\ell-1}^{-1} \Phi_{\ell-1}(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1})$$

Nun gilt für beliebige $y \in \{\tilde{D}_\ell \cap \tilde{D}_{\ell-1}\}$:

$$\begin{aligned} Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(y) - Q_{1,\ell-1} A_{2,\ell-1}^{-1} \Phi_{\ell-1}(y) \;&=\; \hat{g}(x_*(t_\ell) + y, t_\ell) - \hat{g}(x_*(t_{\ell-1}) + y, t_{\ell-1}) \\ &\quad - (\hat{g}(x_*(t_\ell), t_\ell) - \hat{g}(x_*(t_{\ell-1}), t_{\ell-1})) \\ &\quad - (\hat{g}'_x(x_*(t_\ell), t_\ell) - \hat{g}'_x(x_*(t_{\ell-1}), t_{\ell-1}))y \end{aligned}$$

$$= \int_0^1 (\hat{g}'(x_*(z(s)) + y, z(s)) - \hat{g}'(x_*(z(s)), z(s)) - (\hat{g}')'_x(x_*(z(s)), z(s))y)\, ds\, h_\ell\ ,$$

wobei $\hat{g}(x,t) := Q_1(t) A_2^{-1}(t) g(x,t)$ und $z(s) := st_\ell + (1-s)t_{\ell-1}$.
Da $\hat{g}$ zweimal stetig differenzierbar bezüglich x ist, so gilt weiter:
$\forall\, c \geq 0\ \exists\, \varrho \geq 0\ \forall\, 0 \leq s \leq 1$:

$$\|\hat{g}'(x_*(z(s)) + y, z(s)) - \hat{g}'(x_*(z(s)), z(s)) - (\hat{g}')'_x(x_*(z(s)), z(s))y\| \;\leq c\ \|y\|\ ,$$

wenn $\|y\| \leq \varrho$.
Wir erhalten für ζ_ℓ:

$$\|\zeta_\ell\| \;\leq c\, h_\ell(\|\tilde{u}_{\ell-1}\| + \|P\tilde{v}_{\ell-1}\|)\ , \tag{3.10}$$

wenn $\|\tilde{u}_{\ell-1}\| \leq \varrho$ und $\|P\tilde{v}_{\ell-1}\| \leq \varrho$. Dies bedeutet, daß $\|\zeta_\ell\| \;\leq \gamma$ gilt, falls die Schrittweite h_ℓ hinreichend klein ist.

Setzen wir nun die entsprechenden Funktionswerte der Funktion ψ_ℓ anstelle von $f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) - f_{\ell-1}(\tilde{u}_{\ell-1}, \tilde{\delta}_{\ell-1})$ in die Gleichung (3.8b) ein, so ergibt sich:

$$\begin{aligned} &-Q\int_0^1 \psi'_\ell(s\tilde{u}_\ell + (1-s)\tilde{u}_{\ell-1}, s\tilde{\delta}_\ell + (1-s)\tilde{\delta}_{\ell-1}, (1-s)\zeta_\ell)\, ds\, (\frac{\tilde{u}_\ell - \tilde{u}_{\ell-1}}{h_\ell}, \frac{\tilde{\delta}_\ell - \tilde{\delta}_{\ell-1}}{h_\ell}, \frac{-\zeta_\ell}{h_\ell}) \\ &-\frac{1}{h_\ell} Q(Q_{1,\ell} - Q_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + \tilde{w}_\ell + QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) + \tilde{w}_\ell) \\ &\qquad + QP_{1,\ell}A_{2,\ell}^{-1}B_\ell\tilde{u}_\ell + QP_{1,\ell}A_{2,\ell}^{-1}\tau_\ell - QP_{1,\ell}A_{2,\ell}^{-1}\delta_\ell \;=\; 0\ . \end{aligned}$$

Nun läßt sich mit Hilfe von (3.8a) $\frac{\tilde{u}_\ell - \tilde{u}_{\ell-1}}{h_\ell}$ ersetzen durch

$$\frac{1}{h_\ell} P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) - PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell$$
$$- PP_{1,\ell}A_{2,\ell}^{-1}\tau_\ell + PP_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) + \tilde{w}_\ell)$$

Zur kürzeren Schreibweise seien:

$$\mu_\ell := \frac{\tilde{\delta}_\ell - \tilde{\delta}_{\ell-1}}{h_\ell}, \qquad \nu_\ell := \frac{-\zeta_\ell}{h_\ell}. \tag{3.11}$$

Jetzt können wir (3.8b) nach $\tilde{w}_\ell$ auflösen. Dazu definieren wir folgende Funktion:

$$\begin{aligned}
&K_\ell(w, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) := \\
&\quad - Q \int_0^1 \psi_\ell'(su + (1-s)\tilde{u}_{\ell-1}, \tilde{\delta}_\ell + (s-1)h_\ell\mu_\ell, (s-1)h_\ell\nu_\ell)\, ds * \\
&\quad\quad * \Big(\frac{1}{h_\ell} P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) - PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}u \\
&\quad\quad\quad + PP_{1,\ell}A_{2,\ell}^{-1}(\delta - \tau) - PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w), \mu_\ell, \nu_\ell \Big) \\
&\quad - \frac{1}{h_\ell} Q(Q_{1,\ell} - Q_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + w \\
&\quad + QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w) + QP_{1,\ell}A_{2,\ell}^{-1}B_\ell u - QP_{1,\ell}A_{2,\ell}^{-1}(\delta - \tau),
\end{aligned}$$

für die gilt: K_ℓ ist zweimal stetig differenzierbar,

$$K_\ell(0) = 0,\ K_{\ell_w}'(0) = I.$$

Behauptung 2.
Es existieren ein von der Zerlegung unabhängiger Radius σ und eine eindeutig bestimmte C^1-Funktion

$$k_\ell(u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \ :\ B(0, \sigma) \ \to\ B(0, \eta)$$

mit den Eigenschaften:

(i) $K_\ell(k_\ell(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell), \tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) = 0,$
$k_\ell(0) = 0$

(ii) $k_\ell'(0) = (-QP_{1,\ell}A_{2,\ell}^{-1}B_\ell,\ QP_{1,\ell}A_{2,\ell}^{-1},\ 0,\ \frac{1}{h_\ell}Q(Q_{1,\ell} - Q_{1,\ell-1}),\ \frac{1}{h_\ell}Q(Q_{1,\ell} - Q_{1,\ell-1}),$
$Q,\ Q)$

(iii) $k_\ell(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \equiv Qk_\ell(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)$

(iv) $\|k_\ell(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)\|$

$\leq (1 + 2\|QP_{1,\ell}A_{2,\ell}^{-1}B_\ell\|)\ \|\tilde{u}_\ell\| + (1 + 2\|QP_{1,\ell}A_{2,\ell}^{-1}\|)\ \|\delta_\ell - \tau_\ell\| + \|\tilde{\delta}_\ell\|$

$+ (1 + 2\frac{1}{h_\ell}\|Q(Q_{1,\ell} - Q_{1,\ell-1})\|)\ (\|\tilde{u}_{\ell-1}\| + \|\tilde{v}_{\ell-1}\|) + (1 + 2\|Q\|)(\|\mu_\ell\| + \|\nu_\ell\|)\|$

Für die Richtigkeit dieser Behauptung bleibt wieder zu zeigen, daß der Radius σ unabhängig von der Zerlegung gewählt werden kann. Die dazu anzustellenden Überlegungen sind im wesentlichen analog zu den Betrachtungen zu Behauptung 1. für F_ℓ. Jedoch sind noch einige zusätzliche Fakten nachzuweisen. Deshalb sollen sie im folgenden angeführt werden.
Wir wollen wieder Lemma 3.4 benutzen und definieren:

$$q_\ell(w, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) := w - K_\ell(w, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell).$$

(1) Dann gilt:

$$\begin{aligned}
&\|q_\ell(w_1, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) - q_\ell(w_2, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)\| \\
&\quad \leq \|Q \int_0^1 \psi_\ell'(su + (1-s)\tilde{u}_{\ell-1}, \tilde{\delta}_\ell + (s-1)h_\ell\mu_\ell, (s-1)h_\ell\nu_\ell)\, ds * \\
&\quad *[-PP_{1,\ell}A_{2,\ell}^{-1}(\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w_1) - \Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w_2)], 0, 0)\| \\
&\quad + \|QP_{1,\ell}A_{2,\ell}^{-1}(\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w_1) - \Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w_2))\|. \quad (3.12)
\end{aligned}$$

(a) Zunächst soll die Differenz

$$\|\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w_1) - \Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + w_2)\|$$

abgeschätzt werden. Sei dazu $z := u + Pf_\ell(u, \tilde{\delta}_\ell)$.

$$\begin{aligned}
&\|\Phi_\ell(z + w_1) - \Phi_\ell(z + w_2)\| \\
&= \|\int_0^1 \Phi_\ell'(z + \tau w_1 + (1-\tau)w_2)\, d\tau\, (w_1 - w_2)\|
\end{aligned}$$

Sei $\kappa > 0$ bliebig, aber fest. Dann existieren nach Lemma 3.2 Radien $\sigma_1(\kappa)$ und $\eta_1(\kappa)$ (unabhängig von der Zerlegung), so daß für alle $z \in B(0, \sigma_1(\kappa))$ und $w_1, w_2 \in B(0, \eta_1(\kappa))$ gilt:

$$\|\Phi_\ell'(z + \tau w_1 + (1-\tau)w_2)\| \leq \kappa.$$

Von z wissen wir:

$$\begin{aligned}
\|z\| &= \|u + Pf_\ell(u, \tilde{\delta}_\ell)\| \\
&\leq \|u\| + \|P\|(\|u\| + 3\|\tilde{\delta}_\ell\|),
\end{aligned}$$

wenn $(u, \tilde{\delta}_\ell) \in B(0, \alpha)$.
Also existiert ein $\sigma_2(\kappa) > 0$ (von der Zerlegung unabhängig), so daß

für $(u,\tilde{\delta}_\ell) \in B(0,\sigma_2(\kappa))$ die Elementbeziehung $z \in B(0,\sigma_1(\kappa))$ gilt. Damit erhalten wir folgende Ungleichung für $(u,\tilde{\delta}_\ell) \in B(0,\sigma_2(\kappa))$ und $w_1, w_2 \in B(0,\eta_1(\kappa))$:

$$\|\Phi_\ell(u+Pf_\ell(u,\tilde{\delta}_\ell)+w_1)-\Phi_\ell(u+Pf_\ell(u,\tilde{\delta}_\ell)+w_2)\| \le \kappa\,\|w_1-w_2\|. \quad (3.13)$$

(b) Nun betrachten wir

$$\int_0^1 \psi'_\ell(su+(1-s)\tilde{u}_{\ell-1},\tilde{\delta}_\ell+(s-1)h_\ell\mu_\ell,(s-1)h_\ell\nu_\ell)\,ds$$

etwas genauer. Es war ψ'_ℓ stetig auf $B(0,\gamma)$. Daher ist für beliebige $0<\gamma_1<\gamma$ die Funktion ψ'_ℓ auch stetig auf der kompakten Menge $\bar{B}(0,\gamma_1)$ und somit ψ'_ℓ beschränkt auf $\bar{B}(0,\gamma_1)$. Dies bedeutet, daß es eine Konstante $C>0$ gibt, so daß für $(u,\tilde{u}_{\ell-1},\tilde{\delta}_\ell,h_\ell\mu_\ell,h_\ell\nu_\ell) \in \bar{B}(0,\gamma_1)$ gilt:

$$\|\int_0^1 \psi'_\ell(su+(1-s)\tilde{u}_{\ell-1},\tilde{\delta}_\ell+(s-1)h_\ell\mu_\ell,(s-1)h_\ell\nu_\ell)\,ds\| \le C. \quad (3.14)$$

Unter Berücksichtigung von (3.10) und (3.11) finden wir für hinreichend kleine Schrittweiten h_ℓ einen (von der Zerlegung unabhängigen) Radius $\sigma_3>0$ so, daß die folgende Relation gilt:

$$(u,\tilde{u}_{\ell-1},\tilde{\delta}_\ell,\tilde{\delta}_\ell-\tilde{\delta}_{\ell-1}) \in B(0,\sigma_3) \Rightarrow (u,\tilde{u}_{\ell-1},\tilde{\delta}_\ell,h_\ell\mu_\ell,h_\ell\nu_\ell) \in \bar{B}(0,\gamma_1).$$

Fassen wir (3.12), (3.13), (3.14) zusammen und beachten, daß es eine von der Zerlegung unabhängige untere Schranke >0 für $\|QP_{1,\ell}A_{2,\ell}^{-1}\|$ gibt, so finden wir feste Radien $\sigma_4>0$ und $\eta>0$, so daß für

$$(u,\tilde{u}_{\ell-1},\tilde{\delta}_\ell,\tilde{\delta}_\ell-\tilde{\delta}_{\ell-1}) \in B(0,\sigma_4) \text{ und } w_1,w_2 \in B(0,\eta)$$

gilt:

$$\|q_\ell(w_1,u,\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1},\mu_\ell,\nu_\ell)-q_\ell(w_2,u,\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1},\mu_\ell,\nu_\ell)\| \le \frac{1}{2}\|w_1-w_2\|. \quad (3.15)$$

(2)

$$\begin{aligned}
&\|q_\ell(0,u,\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1},\mu_\ell,\nu_\ell)\| \\
&= \|-Q\int_0^1 \psi'_\ell(su+(1-s)\tilde{u}_{\ell-1},\tilde{\delta}_\ell+(s-1)h_\ell\mu_\ell,(s-1)h_\ell\nu_\ell)\,ds * \\
&\quad * (\frac{1}{h_\ell}P(P_{1,\ell}-P_{1,\ell-1})(\tilde{u}_{\ell-1}+P\tilde{v}_{\ell-1})-PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}u \\
&\quad\quad + PP_{1,\ell}A_{2,\ell}^{-1}(\delta-\tau)-PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u+Pf_\ell(u,\tilde{\delta}_\ell)),\mu_\ell,\nu_\ell) \\
&\quad - \frac{1}{h_\ell}Q(Q_{1,\ell}-Q_{1,\ell-1})(\tilde{u}_{\ell-1}+P\tilde{v}_{\ell-1})+QP_{1,\ell}A_{2,\ell}^{-1}B_\ell u \\
&\quad + QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u+Pf_\ell(u,\tilde{\delta}_\ell))-QP_{1,\ell}A_{2,\ell}^{-1}(\delta-\tau)\| \qquad (3.16)
\end{aligned}$$

(a) Sei wieder $z := u + Pf_\ell(u, \tilde{\delta}_\ell)$. Nach Lemma 3.2 existiert ein (von der Zerlegung unabhängiger) Radius $\sigma_5 > 0$, so daß für alle $z \in B(0, \sigma_5)$ gilt:

$$\|\Phi_\ell(z)\| \leq \|z\|.$$

Sei jetzt $\sigma_6 > 0$ so gewählt, daß $\sigma_6 < \alpha$ und für alle $(u, \tilde{\delta}_\ell) \in B(0, \sigma_6)$ gilt:

$$\|u\| + \|P\|(\|u\| + 3\|\tilde{\delta}_\ell\|) \leq \sigma_5.$$

Dann haben wir:

$$\begin{aligned} \|\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell))\| &= \|\Phi_\ell(z)\| \leq \|z\| \\ &\leq \|u\| + \|P\|(\|u\| + 3\|\tilde{\delta}_\ell\|) \leq \sigma_5. \end{aligned} \quad (3.17)$$

(b) Wie schon in 1.(b) gezeigt wurde, existiert eine Konstante $c_1 > 0$, so daß für $(u, \tilde{u}_{\ell-1}, \tilde{\delta}_\ell, \tilde{\delta}_\ell - \tilde{\delta}_{\ell-1}) \in B(0, \sigma_3)$ gilt:

$$\| \int_0^1 \psi'_\ell(su + (1-s)\tilde{u}_{\ell-1}, \tilde{\delta}_\ell + (s-1)h_\ell\mu_\ell, (s-1)h_\ell\nu_\ell)\, ds\| \leq c_1. \quad (3.18)$$

(c) Da $Q_1(\cdot)$ und $P_1(\cdot)$ nach Voraussetzung stetig differenzierbar auf $\mathcal{I}$ sind, so existiert eine von der Zerlegung unabhängige Konstante c_2, so daß :

$$\frac{1}{h_\ell}\|P_{1,\ell} - P_{1,\ell-1}\| \leq c_2 \text{ und } \frac{1}{h_\ell}\|Q_{1,\ell} - Q_{1,\ell-1}\| \leq c_2. \quad (3.19)$$

Fassen wir (3.16)-(3.19) zusammen und beachten, daß es eine von der Zerlegung unabhängige untere Schranke > 0 für $\|P_{1,\ell}A_{2,\ell}^{-1}\|$ gibt, so finden wir einen festen Radius $\sigma_7 > 0$ und eine Konstante $C > 0$, so daß für

$$(u, \delta - \tau, \tilde{\delta}_\ell, \tilde{\delta}_\ell - \tilde{\delta}_{\ell-1}, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \in B(0, \sigma_7)$$

gilt:

$$\begin{aligned} &\|q_\ell(0, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)\| \leq \\ &\quad C\ (\ \|u\| + \|\delta - \tau\| + \|\tilde{\delta}_\ell\| + \|\tilde{u}_{\ell-1}\| + \|\tilde{v}_{\ell-1}\| + \|\mu_\ell\| + \|\nu_\ell)\|\ ) \end{aligned}$$

Sei nun $\sigma > 0$ so klein, daß für $(u, \delta - \tau, \tilde{\delta}_\ell, \tilde{\delta}_\ell - \tilde{\delta}_{\ell-1}, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \in B(0, \sigma)$ gilt:

$$C\ (\ \|u\| + \|\delta - \tau\| + \|\tilde{\delta}_\ell\| + \|\tilde{u}_{\ell-1}\| + \|\tilde{v}_{\ell-1}\| + \|\mu_\ell\| + \|\nu_\ell)\|\ ) \leq \frac{1}{2}\eta.$$

Dann gilt:

$$\|q_\ell(0, u, \delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)\| \leq \frac{1}{2}\eta. \quad (3.20)$$

Unter Anwendung von Lemma 3.4 folgen aus (3.15) und (3.20) die Aussagen *(i)–(iii)* von Behauptung 2. Für die Abschätzung *(iv)* genügt es nach Beweis von Lemma 3.1 wieder zu zeigen, daß es einen von der Zerlegung unabhängigen Radius r gibt, so daß für $(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \in B((0), r)$ gilt:

$$\begin{aligned}\|K_\ell(k_\ell(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell), \tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \\ - D_{K_\ell}(0)(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)^T\| \\ \leq \frac{1}{2}(\|\tilde{u}_\ell\| + \|\delta_\ell - \tau_\ell\| + \|\tilde{\delta}_\ell\| + \|\tilde{u}_{\ell-1}\| + \|\tilde{v}_{\ell-1}\| + \|\mu_\ell\| + \|\nu_\ell)\|). \quad (3.21)\end{aligned}$$

Beachtet man nun die Definition von K_ℓ und Lemma 3.2, so ist für die Richtigkeit der Ungleichung (3.21) hinreichend, daß es für jedes $\epsilon > 0$ ein $r(\epsilon)$ gibt, so daß für $(u, \tilde{u}_{\ell-1}, \tilde{\delta}_\ell, h_\ell\mu_\ell, h_\ell\nu_\ell) \in B((0), r(\epsilon))$ und alle $\ell = 1, 2, ..., N$ gilt:

$$\|\psi'_\ell(su + (1-s)\tilde{u}_{\ell-1}, \tilde{\delta}_\ell + (s-1)h_\ell\mu_\ell, (s-1)h_\ell\nu_\ell) - \psi'_\ell(0)\| \leq \epsilon. \quad (3.22)$$

Sei zur kürzeren Schreibweise:

$$z := (su + (1-s)\tilde{u}_{\ell-1}, \tilde{\delta}_\ell + (s-1)h_\ell\mu_\ell, (s-1)h_\ell\nu_\ell).$$

Dann gilt:

$$\begin{aligned}\psi'_\ell(z) &= -(I + Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(su + (1-s)\tilde{u}_{\ell-1} + P\psi_\ell(z))P)^{-1} * \\ &\quad * (Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(su + (1-s)\tilde{u}_{\ell-1} + P\psi_\ell(z)), -I, -I)\end{aligned}$$

und

$$\psi'_\ell(0) = (0, I, I).$$

Wendet man wiederum Lemma 3.2 an, so läßt sich (3.22) nachweisen, womit auch die Ungleichung *(iv)* und schließlich die Behauptung 2. gilt.
Es bleibt, Gleichung (3.8a) nach $\tilde{u}_\ell$ aufzulösen, welche nun folgende Gestalt hat:

$$\begin{aligned}\tilde{u}_\ell &= \tilde{u}_{\ell-1} + P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) \\ &\quad - h_\ell PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell + h_\ell PP_{1,\ell}A_{2,\ell}^{-1}(\delta_\ell - \tau_\ell) \quad (3.23) \\ &\quad - h_\ell PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + Pf_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) + k_\ell(\tilde{u}_\ell, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell))\end{aligned}$$

Wir definieren:

$$\begin{aligned}&R_\ell(u, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) := \\ &\quad u - \tilde{u}_{\ell-1} - P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) \\ &\quad + h_\ell PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}u - h_\ell PP_{1,\ell}A_{2,\ell}^{-1}(\delta_\ell - \tau_\ell) \\ &\quad + h_\ell PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(u + Pf_\ell(u, \tilde{\delta}_\ell) + k_\ell(u, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)).\end{aligned}$$

Es gilt: R_ℓ ist stetig differenzierbar,

$$R_\ell(0) = 0, \; R'_{\ell_u}(0) = I + h_\ell PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}.$$

Beachtet man die Eigenschaften von Φ_ℓ, f_ℓ und k_ℓ, so erhält man mit den gleichen Argumenten wie zuvor:
Es existieren für hinreichend kleine h_ℓ ein Radius χ (unabhängig von der Zerlegung) und eine eindeutig bestimmte C^1-Funktion

$$r_\ell(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \ : \ B((0), \chi) \ \rightarrow \ B(0, \chi')$$

mit den Eigenschaften:

$$R_\ell(r_\ell(\delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell), \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) = 0, \ r_\ell(0) = 0$$
$$r_\ell(\delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \equiv PP_{1,\ell} r_\ell(\delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \ .$$

Nun ist die so ermittelte Funktion

$$\begin{aligned}
&\tilde{x}_\ell(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) := \\
&\quad r_\ell(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \\
&\quad + P f_\ell(r_\ell(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell), \tilde{\delta}_\ell) \\
&\quad + k_\ell(r_\ell(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell), \delta_\ell, \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell)
\end{aligned}$$

eine Lösung des Systems (3.1a)–(3.1c), d.h. das implizite Euler-Verfahren liefert eine Lösung x_ℓ zum Zeitpunkt t_ℓ für $\ell = 2, ..., N$.

Betrachten wir wieder die Gleichung (3.23), so erhalten wir für hinreichend kleine Schrittweiten h_ℓ mit Standardargumenten, daß für

$$(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \in B((0), \chi) :$$

$$\begin{aligned}
&\|\tilde{u}_\ell\| \leq \\
&\quad \Big[\, 1 - h_\ell(\|PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\| + L\|PP_{1,\ell}A_{2,\ell}^{-1}\|(2 + \|P\| + 2\|QP_{1,\ell}A_{2,\ell}^{-1}B_\ell\|)) \, \Big]^{-1} \\
&\quad * \Big[\, [1 + \|P(P_{1,\ell} - P_{1,\ell-1})\| \\
&\qquad + h_\ell L\|PP_{1,\ell}A_{2,\ell}^{-1}\|(1 + 2\frac{1}{h_\ell}\|Q(Q_{1,\ell} - Q_{1,\ell-1})\|)]\|\tilde{u}_{\ell-1}\| \\
&\qquad + \ [\ \|P(P_{1,\ell} - P_{1,\ell-1})\| \\
&\qquad + h_\ell L\|PP_{1,\ell}A_{2,\ell}^{-1}\|(1 + 2\frac{1}{h_\ell}\|Q(Q_{1,\ell} - Q_{1,\ell-1})\|) \ ] \ \|\tilde{v}_{\ell-1}\| \\
&\qquad + \ [\ h_\ell\|PP_{1,\ell}A_{2,\ell}^{-1}\| + h_\ell L\|PP_{1,\ell}A_{2,\ell}^{-1}\|(1 + 2\|QP_{1,\ell}A_{2,\ell}^{-1}\|) \ ] \ \|\delta_\ell - \tau_\ell\| \\
&\qquad + h_\ell L\|PP_{1,\ell}A_{2,\ell}^{-1}\|(1 + 3\|P\|) \ \|\tilde{\delta}_\ell\| \\
&\qquad + h_\ell L\|PP_{1,\ell}A_{2,\ell}^{-1}\|(1 + 2\|Q\|)) \ (\|\mu_\ell\| + \|\nu_\ell\|) \, \Big],
\end{aligned} \tag{3.24}$$

wenn L Lipschitzkonstante von Φ_ℓ ist.
Es sei hier bemerkt, daß es ein $L \in I\!R$ gibt, so daß L Lipschitzkonstante von Φ_ℓ auf $\tilde{D}_\ell$ für alle $\ell = 1, 2, ..., N$ ist. Denn für beliebige $y, z \in \tilde{D}_\ell$ gilt:

$$\begin{aligned} & \|\Phi_\ell(y) - \Phi_\ell(z)\| \\ & \leq \|g(y + x_*(t_\ell), t_\ell) - g(z + x_*(t_\ell), t_\ell)\| + \|g'_x(y + x_*(t_\ell), t_\ell)\| \, \|y - z\| \\ & \leq (L_1 + L_2)\|y - z\|, \end{aligned}$$

wobei L_1 Lipschitzkonstante von g und L_2 Lipschitzkonstante von g'_x sind.
Es existieren also eine von der Zerlegung unabhängige Konstante C_1 und eine maximale Schrittweite H_{max}, so daß für $h_\ell < H_{max}$ und $(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}, \mu_\ell, \nu_\ell) \in B((0), \chi)$ gilt:

$$\begin{aligned} \|\tilde{u}_\ell\| \leq (1 - h_\ell C_1)^{-1} * \Big[& (1 + h_\ell C_1)\|\tilde{u}_{\ell-1}\| + h_\ell C_1 \|\tilde{v}_{\ell-1}\| \\ & h_\ell C_1 \|\delta_\ell - \tau_\ell\| + h_\ell C_1 \|\tilde{\delta}_\ell\| + h_\ell C_1 (\|\mu_\ell\| + \|\nu_\ell\|) \Big] . \end{aligned}$$

Unter Berücksichtigung von (3.10), (3.11), und der Tatsache, daß

$$\|\tilde{\delta}_\ell\| = \|Q_{1,\ell} A_{2,\ell}^{-1} \delta_\ell\| = \|Q_{1,\ell} A_{2,\ell}^{-1} (\delta_\ell - \tau_\ell)\|,$$

finden wir eine Konstante $C_2 > 0$, so daß :

$$\begin{aligned} \|\tilde{u}_\ell\| \leq (1 - h_\ell C_2)^{-1} * \Big[& (1 + h_\ell C_2)\|\tilde{u}_{\ell-1}\| + h_\ell C_2 \|\tilde{v}_{\ell-1}\| \\ & h_\ell C_2 \|\delta_\ell - \tau_\ell\| + C_2 \|\tilde{\delta}_\ell - \tilde{\delta}_{\ell-1}\| \Big] . \end{aligned} \tag{3.25}$$

Anfangs hatten wir vorausgesetzt, daß $\ell \geq 2$. So bleibt also noch der Fall $\ell = 1$ zu untersuchen.
Verfolgt man noch einmal alle bisherigen Darlegungen, so lassen sich alle Schlüsse auch für $\ell = 1$ nachvollziehen. Den einzigen Unterschied, den es zu beachten gilt, ist die Tatsache, daß wir keine Funktion f_0 zur Verfügung haben und dementsprechend ζ_1 etwas anders definiert werden muss, nämlich:

$$\zeta_1 := Q_{1,0}\delta_0 - \tilde{\delta}_0 + Q_{1,1} A_{2,1}^{-1} \Phi_1(\tilde{u}_0 + P\tilde{v}_0).$$

Mit den Voraussetzungen des Satzes hat dann ζ_1 auch die Eigenschaften, die für ζ_ℓ ($\ell \geq 2$) benötigt wurden und es folgt mit Behauptung 1, Behauptung 2 und (3.25) die Behauptung des Satzes.

□

3.2 Beispiel für die Verkopplung der verschiedenen Fehler

Wir betrachten folgendes autonome Index-2-Beispiel in Hessenberg-Form:

$$x'_1 + (x_2 - 1)x_3^2 = 0$$

$$\begin{aligned} x_2' - x_3 &= 0 \\ \frac{x_2 - 1}{x_1} &= 0 \end{aligned}$$

mit der Anfangsbedingung

$$x_1(t_0) = 1 \quad .$$

Offenbar gibt es nur die eine Lösung

$$\begin{aligned} x_1 &\equiv 1 \\ x_2 &\equiv 1 \\ x_3 &\equiv 0 \quad . \end{aligned}$$

Es sei hier bemerkt, daß das Beispiel gerade so konstruiert ist, daß $(x_1, 0, 0) = PP_1x$, $(0, x_2, 0) = PQ_1x$ und $(0, 0, x_3) = Qx$ entsprechend der obigen Notation sind, und das System bereits in aufgespaltener Form vorliegt.
Das implizite Euler-Verfahren ergibt bei gegebenem Anfangswert $x_{1,0}$ und Störungen δ_ℓ im ℓ-ten Schritt folgende Iteration:

$$\frac{x_{1,\ell} - x_{1,\ell-1}}{h_\ell} + (x_{2,\ell} - 1)x_{3,\ell}^2 = \delta_{1,\ell} \tag{3.26a}$$

$$\frac{x_{2,\ell} - x_{2,\ell-1}}{h_\ell} - x_{3,\ell} = \delta_{2,\ell} \tag{3.26b}$$

$$\frac{x_{2,\ell} - 1}{x_{1,\ell}} = \delta_{3,\ell} \quad , \tag{3.26c}$$

für $\ell = 1, 2, ..., N$.
Zunächst liefert die Gleichung (3.26c):

$$\begin{aligned} x_{2,\ell} - 1 &= x_{1,\ell}\delta_{3,\ell} \qquad \text{und} \\ x_{2,\ell-1} - 1 &= x_{1,\ell-1}\delta_{3,\ell-1} \end{aligned}$$

für $\ell = 2, ..., N$.
Setzt man dies in Gleichung (3.26b) und (3.26a) ein, so ergibt sich

$$x_{1,\ell}\delta_{3,\ell} - x_{1,\ell-1}\delta_{3,\ell-1} = h_\ell x_{3,\ell} + h_\ell \delta_{2,\ell}$$

und weiter

$$x_{1,\ell} = x_{1,\ell-1} + h_\ell x_{1,\ell}\delta_{3,\ell}\left(\frac{x_{1,\ell}\delta_{3,\ell} - x_{1,\ell-1}\delta_{3,\ell-1}}{h_\ell} - \delta_{2,\ell}\right)^2 + h_\ell \delta_{1,\ell} \quad .$$

Damit nun wünschenswerter Weise

$$x_{1,\ell} = x_{1,\ell-1} + O(h)$$

gelten kann (sei an dieser Stelle die Schrittweite h konstant), ist offenbar notwendig, daß

$$O(1) = x_{1,\ell}\delta_{3,\ell}(\frac{x_{1,\ell}\delta_{3,\ell} - x_{1,\ell-1}\delta_{3,\ell-1}}{h} - \delta_{2,\ell})^2 + \delta_{1,\ell} \quad ,$$

d.h.

$$\frac{x_{1,\ell}\delta_{3,\ell} - x_{1,\ell-1}\delta_{3,\ell-1}}{h} = O(1)$$

gilt. Dies bedeutet, daß selbst die differentiellen Komponenten von der schwachen Instabilität wesentlich beeinflußt werden.

Mit dem im 4. Abschnitt angegebenen Programm wurde dieses Beispiel auch mit konstanten Schrittweiten ($10^{-2} - 10^{-4}$) und dem Anfangspunkt $t_0 = 0$ getestet. Aufgrund der besonders einfachen Gestalt der algebraischen Nebenbedingung sind die Fehler durch den Abbruch des Newtonverfahrens (bei einer Genauigkeit von 10^{-8}) und die Rundungsfehler bei Schrittweiten, die nicht kleiner als 10^{-8} sind, nicht sichtbar. Um nun aber den Einfluß dieser Fehler auf die differentiellen Komponenten zu zeigen, wurden künstlich Rundungsfehler in der zweiten Komponente der Größenordnung 10^{-6} eingeführt. Die Auswirkungen auf die erste Komponente zeigt die folgende Tabelle, in der der absolute Fehler angegeben ist:

Zeitpunkt	Schrittweite		
	1.0E-2	1.0E-3	1.0E-4
1.0E+2	2.372E-11	2.151E-9	2.174E-7
2.0E+2	4.741E-11	4.302E-9	4.347E-7
3.0E+2	7.108E-11	6.452E-9	6.521E-7
4.0E+2	9.477E-11	8.603E-9	8.694E-7
5.0E+2	1.184E-10	1.075E-8	1.086E-6
6.0E+2	1.421E-10	1.290E-8	1.304E-6
7.0E+2	1.658E-10	1.505E-8	1.522E-6
8.0E+2	1.895E-10	1.721E-8	1.739E-6
9.0E+2	2.108E-10	1.935E-8	1.956E-6
1.0E+3	2.345E-10	2.129E-8	2.151E-6

Die erhaltenen Werte widerspiegeln also die theoretischen Ergebnisse. Im Gegensatz zu den allgemeinen Erwartungen und Hoffnungen kann sich eine Verkleinerung der Schrittweite bei bestimmten nichtlinearen Index–2–ADGen nicht nur negativ auf die algebraische Komponente, sondern auch auf die differentielle Komponente auswirken.

3.3 Stabilität für spezielle Probleme

Die Aussage des Satzes 3.3 läßt sich für spezielle Probleme in der gewünschten Weise verbessern, daß der Einfluß der Störungen durch die numerische Rechnung

in der differentiellen Komponente Px unabhängig von der Schrittweite beschränkt bleibt, wie dies z.B. für lineare Index-2-Probleme schon gut bekannt ist (vgl. Abschnitt 1.3). Hier soll nun eine etwas größere Klasse angegeben werden, für die dies der Fall ist.

Satz 3.5 *Sei $x_* \in C_N^1$ eine Lösung des Systems (2.1) und besitze die ADG (2.1) den Traktabilitäts-Index 2. Mit den Bezeichnungen aus Lemma 2.2 sei $Q_1(\cdot)$ stetig differenzierbar auf $\mathcal{I}$ und $Q_1(t)A_2^{-1}(t)g(\cdot,t)$ stetig differenzierbar. Sei außerdem in einer Umgebung der Kurve $x_*(\cdot)$ die nichtlineare Funktion g darstellbar als*

$$g(x,t) = \hat{g}(Px,t) + \hat{B}(t)Qx,$$

hänge also nur linear von der algebraischen Komponente Qx ab.
Dann existieren Konstanten $\epsilon > 0$, $C > 0$ und $H_{max} > 0$, so daß für jede Zerlegung (2.2) mit $h_{max} \le H_{max}$ folgende Implikation wahr ist:
Wenn die Relationen

$$\|\delta_\ell\| \le \epsilon, \quad \ell = 0, 1, ..., N$$

erfüllt sind, so liefert das implizite Eulerverfahren eine Lösung x_ℓ zum Zeitpunkt t_ℓ, $\ell = 1, 2, ..., N$, und es gilt:

(i)

$$\max_{\ell=1,2,...,N} \|P(x_*(t_\ell) - x_\ell)\| \le C \left[\|x_*(t_0) - x_0\| + \max_{\ell=1,2,...,N} \|\delta_\ell - \tau_\ell\| \right]$$

(ii)

$$\begin{aligned} \max_{\ell=1,2,...,N} \|Q(x_*(t_\ell) - x_\ell)\| \;\le\; & C \Big[\|x_*(t_0) - x_0\| + \max_{\ell=1,2,...,N} \|\delta_\ell - \tau_\ell\| \\ & + \frac{1}{h_1} \|Q_{1,0}A_{2,0}^{-1}(g(x_0,t_0) - g(x_*(t_0),t_0) - \delta_0)\| \\ & + \max_{\ell=2,...,N} \frac{1}{h_\ell} \|Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell - Q_{1,\ell-1}A_{2,\ell-1}^{-1}\delta_{\ell-1}\| \Big] . \end{aligned}$$

Beweis: Sei $\ell \in \{1, ..., N\}$ fest gewählt und der ℓ-te Schritt des impliziten Euler-Verfahrens wieder entsprechend der Herleitung in Kapitel 2 äquivalent umgeformt in das System:

$$\begin{aligned} & \frac{\tilde{u}_\ell - \tilde{u}_{\ell-1}}{h_\ell} + \frac{1}{h_\ell} P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + PP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ & \qquad + PP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell + \tilde{w}_\ell) + PP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \qquad (3.27a) \end{aligned}$$

$$\begin{aligned} & -Q\frac{\tilde{v}_\ell - \tilde{v}_{\ell-1}}{h_\ell} - \frac{1}{h_\ell} Q(Q_{1,\ell} - Q_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + QP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ & \qquad + \tilde{w}_\ell + QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell + \tilde{w}_\ell) + QP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \qquad (3.27b) \end{aligned}$$

$$\tilde{v}_\ell + Q_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell) - \tilde{\delta}_\ell = 0, \qquad (3.27c)$$

wobei $\tilde{\delta}_\ell := Q_{1,\ell} A_{2,\ell}^{-1} \delta_\ell$. In dem jetzt betrachteten Fall vereinfacht sich dieses System zu

$$\frac{\tilde{u}_\ell - \tilde{u}_{\ell-1}}{h_\ell} + \frac{1}{h_\ell} P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + PP_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{u}_\ell + PP_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell) + PP_{1,\ell} A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \tag{3.28a}$$

$$-Q\frac{\tilde{v}_\ell - \tilde{v}_{\ell-1}}{h_\ell} - \frac{1}{h_\ell} Q(Q_{1,\ell} - Q_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + QP_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{u}_\ell + \tilde{w}_\ell + QP_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell) + QP_{1,\ell} A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \tag{3.28b}$$

$$\tilde{v}_\ell + Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_\ell + P\tilde{v}_\ell) - \tilde{\delta}_\ell = 0, \tag{3.28c}$$

denn aufgrund der speziellen Gestalt von g ergibt sich für Φ_ℓ:

$$\Phi_\ell(y) = \Phi_\ell(Py) \; \forall \, y \in \tilde{D}_\ell.$$

Entscheidendend für die qualitative Verbesserung der Ergebnisse in dem jetzt betrachteten Fall ist die Tatsache, daß die Gleichung (3.28a) nicht mehr von $\tilde{w}_\ell$ abhängt und somit die Diskretisierung der "eigentlichen" Differentialgleichung nicht mehr mit der algebraischen Komponente verkoppelt ist. Damit können wir auch unsere Strategie etwas ändern. Zunächst lösen wir wieder die Gleichung (3.28c) nach $\tilde{v}_\ell$ auf. Dies ermöglicht es uns bereits die Gleichung (3.28a) als nichtlineare Gleichung in $\tilde{u}_\ell$ zu lösen und am Ende ergibt sich die Komponente $\tilde{w}_\ell$ aus der Gleichung (3.28b) als einfache Zuweisung. Wir definieren also wieder

$$F_\ell(v, u, \delta) := v - \delta + Q_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(u + Pv). \tag{3.29}$$

und es gilt wie im Beweis von Satz 3.3:
Es existieren ein von der Zerlegung unabhängiger Radius α und eine eindeutig bestimmte C^1-Funktion

$$f_\ell(u, \delta) \; : \; B((0,0), \alpha) \; \rightarrow \; B(0, \rho)$$

mit den Eigenschaften:

(i) $F_\ell(f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell), \tilde{u}_\ell, \tilde{\delta}_\ell) = 0$

(ii) $f_\ell(0,0) = 0, \;\; f_\ell'(0,0) = (0, I)$

(iii) $f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) \equiv Q_{1,\ell} f(\tilde{u}_\ell, \tilde{\delta}_\ell)$

(iv)
$$\|f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell)\| \leq \|\tilde{u}_\ell\| + 3\|\tilde{\delta}_\ell\|. \tag{3.30}$$

Jetzt bleibt anstelle des Systems (3.28a)-(3.28c) folgendes System zu lösen:

$$\begin{aligned}
&\tilde{u}_\ell - \tilde{u}_{\ell-1} + h_\ell P P_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{u}_\ell \\
&\quad + P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) \\
&+ h_\ell P P_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_\ell + P f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell)) + h_\ell P P_{1,\ell} A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) \;=\; 0 \quad (3.31a)
\end{aligned}$$

$$\begin{aligned}
&- Q\frac{f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) - \tilde{v}_{\ell-1}}{h_\ell} + Q P_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} \tilde{u}_\ell \\
&\quad - \frac{1}{h_\ell} Q(Q_{1,\ell} - Q_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + \tilde{w}_\ell \\
&+ Q P_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(\tilde{u}_\ell + P f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell)) + Q P_{1,\ell} A_{2,\ell}^{-1}(\tau_\ell - \delta_\ell) \;=\; 0 \quad (3.31b)
\end{aligned}$$

$$\tilde{v}_\ell - f_\ell(\tilde{u}_\ell, \tilde{\delta}_\ell) \;=\; 0. \quad (3.31c)$$

Um nun die Gleichung (3.31a) zu lösen, definieren wir die Abbildung

$$\begin{aligned}
&R_\ell(u, \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}) := \\
&\quad u - \tilde{u}_{\ell-1} + P(P_{1,\ell} - P_{1,\ell-1})(\tilde{u}_{\ell-1} + P\tilde{v}_{\ell-1}) + h_\ell P P_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell} u \\
&\quad - h_\ell P P_{1,\ell} A_{2,\ell}^{-1}(\delta_\ell - \tau_\ell) + h_\ell P P_{1,\ell} A_{2,\ell}^{-1} \Phi_\ell(u + P f_\ell(u, \tilde{\delta}_\ell)).
\end{aligned}$$

Es existieren wieder für hinreichend kleine h_ℓ ein Radius χ (unabhängig von der Zerlegung) und eine eindeutig bestimmte C^1-Funktion

$$r_\ell(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}) \;:\; B((0), \chi) \;\rightarrow\; B(0, \chi')$$

mit den Eigenschaften:

$$\begin{aligned}
&R_\ell(r_\ell(\delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}), \delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}) = 0, \; r_\ell(0) = 0 \\
&\quad r_\ell(\delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}) \equiv P P_{1,\ell} r_\ell(\delta_\ell - \tau_\ell, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}) \;,
\end{aligned}$$

und mit Standardargumenten erhalten wir für hinreichend kleine Schrittweiten h_ℓ, daß für

$$(\delta - \tau, \tilde{\delta}_\ell, \tilde{u}_{\ell-1}, \tilde{v}_{\ell-1}) \in B((0), \chi) :$$

$$\begin{aligned}
&\|\tilde{u}_\ell\| \leq \\
&\quad \Big[\, 1 - h_\ell(\|P P_{1,\ell} A_{2,\ell}^{-1} B_\ell P P_{1,\ell}\| + L\|P P_{1,\ell} A_{2,\ell}^{-1}\|(1 + \|P\|)) \,\Big]^{-1} \\
&\quad * \Big[\, [1 + \|P(P_{1,\ell} - P_{1,\ell-1})\|]\|\tilde{u}_{\ell-1}\| + \|P(P_{1,\ell} - P_{1,\ell-1})\| \, \|\tilde{v}_{\ell-1}\| \\
&\qquad + h_\ell \|P P_{1,\ell} A_{2,\ell}^{-1}\| \, \|\delta_\ell - \tau_\ell\| + h_\ell 3L \|P P_{1,\ell} A_{2,\ell}^{-1}\| \|P\| \, \|\tilde{\delta}_\ell\| \,\Big], \qquad (3.32)
\end{aligned}$$

wenn L Lipschitzkonstante von Φ_ℓ ist.
Schließlich gilt nun für $\tilde{w}_\ell$:

$$\begin{aligned}\tilde{w}_\ell &= Q\frac{f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)-\tilde{v}_{\ell-1}}{h_\ell} - QP_{1,\ell}A_{2,\ell}^{-1}B_\ell PP_{1,\ell}\tilde{u}_\ell \\ &\quad + \frac{1}{h_\ell}Q(Q_{1,\ell}-Q_{1,\ell-1})(\tilde{u}_{\ell-1}+P\tilde{v}_{\ell-1}) \\ &\quad - QP_{1,\ell}A_{2,\ell}^{-1}\Phi_\ell(\tilde{u}_\ell+Pf_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)) - QP_{1,\ell}A_{2,\ell}^{-1}(\tau_\ell-\delta_\ell). \qquad (3.33)\end{aligned}$$

Die so ermittelte Funktion

$$\begin{aligned}&\tilde{x}_\ell(\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1}) := \\ &\quad r_\ell(\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1}) \\ &\quad + Pf_\ell(r_\ell(\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1}),\tilde{\delta}_\ell) \\ &\quad + k_\ell(r_\ell(\delta-\tau,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1}),\delta_\ell,\tau_\ell,\tilde{\delta}_\ell,\tilde{u}_{\ell-1},\tilde{v}_{\ell-1})\end{aligned}$$

ist eine Lösung des Systems (3.27a)–(3.27c), d.h. das implizite Euler-Verfahren liefert eine Lösung x_ℓ zum Zeitpunkt t_ℓ für $\ell = 1,2,...,N$.
Die erste im Satz angegebene Abschätzung *(i)* folgt nun unter Berücksichtigung der Beziehung

$$\|\tilde{\delta}_\ell\| = \|Q_{1,\ell}A_{2,\ell}^{-1}\delta_\ell\| = \|Q_{1,\ell}A_{2,\ell}^{-1}(\delta_\ell-\tau_\ell)\|$$

aus den Ungleichungen (3.30) und (3.32).
Um nun die im Satz angegebene zweite Abschätzung *(ii)* zu erhalten, müssen wir in der Gleichung (3.33) nochmals die Differenz

$$f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)-\tilde{v}_{\ell-1}$$

im Verhältnis zur Schrittweite h_ℓ untersuchen. Dazu seien ζ_ℓ und Ψ_ℓ wie im Beweis von Satz 3.3 definiert und man erhält ebenso für $\ell \geq 2$:
Es existieren ein von der Zerlegung unabhängiger Radius γ und eine Konstante $C > 0$, so daß für $(\tilde{u}_\ell-\tilde{u}_{\ell-1},\tilde{\delta}_\ell-\tilde{\delta}_{\ell-1},\zeta_\ell) \in B((0),\gamma)$:

$$\begin{aligned}&\|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)-\tilde{v}_{\ell-1}\| \\ &\quad = \|f_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell)-f_{\ell-1}(\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1})\| \\ &\quad = \|\psi_\ell(\tilde{u}_\ell,\tilde{\delta}_\ell,\zeta_\ell)-\psi_\ell(\tilde{u}_{\ell-1},\tilde{\delta}_{\ell-1},0)\| \\ &\quad = \int_0^1 \psi_\ell'(s\tilde{u}_\ell+(1-s)\tilde{u}_{\ell-1},s\tilde{\delta}_\ell+(1-s)\tilde{\delta}_{\ell-1},s\zeta_\ell)\,ds(\tilde{u}_\ell-\tilde{u}_{\ell-1},\tilde{\delta}_\ell-\tilde{\delta}_{\ell-1},\zeta_\ell)\| \\ &\quad \leq C(\|\tilde{u}_\ell-\tilde{u}_{\ell-1}\|+\|\tilde{\delta}_\ell-\tilde{\delta}_{\ell-1}\|+\|\zeta_\ell)\|). \qquad (3.34)\end{aligned}$$

Für ζ_1 bekommen wir:
Es existiert nun für jedes $\epsilon > 0$ eine Konstante $\theta > 0$, so daß für $\delta_0 \leq \theta$ gilt:

$$\|\zeta_1\| = \|Q_{1,0}\delta_0-\tilde{\delta}_0+Q_{1,1}A_{2,1}^{-1}\Phi_1(\tilde{u}_0+P\tilde{v}_0)\|$$

$$
\begin{aligned}
&\leq \|Q_{1,1}A_{2,1}^{-1}(\Phi_1(\delta_0) - \Phi_0(\delta_0))\| + \|(Q_{1,1}A_{2,1}^{-1} - Q_{1,0}A_{2,0}^{-1})\Phi_0(\delta_0)\| \\
&\quad + \|Q_{1,0}A_{2,0}^{-1}\Phi_0(\delta_0) - \tilde{\delta}_0 + Q_{1,0}\delta_0\| \\
&\leq \epsilon h_1 + \|Q_{1,0}A_{2,0}^{-1}(\Phi_0(\delta_0) - \delta_0) + Q_{1,0}\delta_0\| \\
&\leq \epsilon h_1 + \|Q_{1,0}(A_{2,0}^{-1}(g(x_0,t_0) - g(x_*(t_0),t_0)) - \delta_0)\| \qquad (3.35)
\end{aligned}
$$

Mit den Gleichungen bzw. Ungleichungen (3.32)-(3.35) folgt nun die Behauptung.

4 Praktische Realisierung der BDF

In den vergangenen Jahren wurden bereits einige effektive Codes mittels der BDF für Anfangswertaufgaben von Algebro–Differentialgleichungen entwickelt. An dieser Stelle seien die Programmpakete LSODI von Hindmarsh [80], DASSL von Petzold [83] und SPRINT von Berzins & Furzeland [85] erwähnt. Jedoch sind diese nur für Systeme vom Index ≤ 1 entwickelt worden und versagen i.a. auch für Systeme höheren Index. Für Systeme höheren Index ist mir nur das Programm RADAU5 von Hairer, Lubich & Roche [89] bekannt, welches auf dem impliziten Runge–Kutta–Verfahren (RADAU IIA) der Ordnung 5 basiert. Es soll nun eine Implementierung der BDF (DAE2SOL), vorgestellt werden, die erwarten läßt, daß diese auch in einer Reihe von Index–2–ADGen erfolgreich ist und unnötige oder nachteilige Schrittweitenverkleinerung vermeidet. An dieser Stelle sei gesagt, daß bei der Programmentwicklung lediglich die Schrittweiten– und Ordnungssteuerung im Blickpunkt stand. Probleme wie die Bestimmung konsistenter Anfangswerte, Wahl einer geeigneten Skalierung und Interpolation der Lösungspunkte wurden hier nicht betrachtet. Dies muß Gegenstand zukünftiger Untersuchungen sein. Deshalb wurde hier auch noch auf Vergleiche mit den anderen ADG–Lösern verzichtet.
Das Programm DAE2SOL realisiert die BDF für ADGen der Form

$$f(x'(t), x(t), t) = 0, \tag{4.1}$$

für die der Nullraum $ker(f_y(y, x, t))$ konstant ist. Die Gründe, warum diese Einschränkung notwendig ist, sind bereits im ersten Abschnitt der Arbeit angegeben. Wesentliche Grundlage für DAE2SOL ist die von Denk [88] vorgestellte numerische Integration von ADGen zur Simulation elektrischer Schaltkreise. Die dort entwickelte äußerst effektive Fehlerschätzung und damit verbundene Schrittweiten– und Ordnungssteuerung wurde an das Verhalten der ADGen höheren Index in der Form angepaßt, daß lediglich die differentiellen Komponenten für die Fehlerschätzung herangezogen werden. Die Idee, die algebraischen Komponenten bei der Steuerung auszuschalten, findet man bereits in Arbeiten von Petzold & Lötstedt [86] und Leimkuhler [86]. Zur Lösung der nichtlinearen Gleichungen konnte ich die von R. Lamour (wiss. Mitarbeiter im Bereich Numerik am Institut für Angewandte Mathematik der Humboldt–Universität Berlin) entwickelte Routine NLSOLV benutzen, die er mir freundlicher Weise zur Verfügung stellte.
In den folgenden Abschnitten werden die wesentlichen Aspekte der Realisierung in kurzer Form etwas näher erläutert. Betrachtet werden die BDF k-ter Ordnung für Probleme (4.1) im n-ten Zeitschritt in der Form

$$f(-\frac{1}{h_n}\sum_{i=0}^{k}\alpha_i x_{n+1-i}, x_{n+1}, t_{n+1}) = 0,$$

wobei x_{n+1} durch Lösung dieser nichtlinearen Gleichung ermittelt wird.

4.1 Berechnung der Koeffizienten

Bei der Verwendung der BDF k-ter Ordnung wird als Näherung für $\dot{x}(t_{n+1})$ der Wert des Polynoms $\dot{P}_k(t)$ an der Stelle t_{n+1} benutzt, wobei $P_k(t)$ ein Polynom k-ten Grades ist, welches die Punkte $(t_{n+1-k}, x_{n+1-k}), ..., (t_{n+1}, x_{n+1})$ interpoliert. Damit ergibt sich für die Koeffizienten

$$\alpha_i = \frac{t_{n+1} - t_n}{t_{n+1} - t_{n+1-i}} \prod_{j=1, j\neq i}^{k} \frac{t_{n+1} - t_{n+1-j}}{t_{n+1-i} - t_{n+1-j}}, \quad i = 1, ...k.$$

Der erste Koeffizient ergibt sich einfach aus der Beziehung

$$\alpha_0 = -\sum_{i=1}^{k} \alpha_i, \tag{4.2}$$

da das Verfahren die Konsistenzordnung k besitzt.
Neben den BDF–Koeffizienten wird für die Lösung der impliziten Gleichungen ein geeigneter Prädiktor benötigt. Diesen erhält man, wenn man das Polynom k-ten Grades $P(t)$ an der Stelle t_{n+1} auswertet, welches die Punkte $(t_{n-k}, x_{n-k}), ..., (t_n, x_n)$ interpoliert. Sei nun

$$x^p_{n+1} = \sum_{i=0}^{k} \gamma_i x_{n-i}$$

dieser Prädiktor. Dann ergibt sich für die Koeffizienten

$$\gamma_i = \prod_{j=0, j\neq i}^{k} \frac{t_{n+1} - t_{n+1-j}}{t_{n+1-i} - t_{n+1-j}}, \quad i = 1, ...k.$$

Der erste Koeffizient läßt sich wieder leicht ermitteln:

$$\gamma_0 = 1 - \sum_{i=1}^{k} \gamma_i. \tag{4.3}$$

Es ist hier erwähnenswert (wie bereits in Denk [88] bemerkt wurde), daß die Beziehungen (4.2) und (4.3) nicht nur den Vorteil haben, daß sie relativ einfach sind, sie sichern auch bei mit Rundungsfehlern behafteten Koeffizienten, daß die Konsistenzbedingung für die Exaktheit von Polynomen k-ten Grades erfüllt ist.

4.2 Fehlerschätzung

Zur Fehler–Schätzung wird eine Näherung des Diskretisierungsfehlers für reguläre gewöhnliche Differentialgleichungen

$$E_{n+1} := h_{n+1}(x_{n+1} - x(t_{n+1}))$$

benutzt. Entsprechend den Darlegungen in Denk [88] gilt die Beziehung

$$E_{n+1} = \frac{h_{n+1}}{t_{n+1} - t_{n-k}}(x_{n+1} - x^p_{n+1}) + O(h^{k+2}). \tag{4.4}$$

Der angegebene Prädiktor eignet sich damit sowohl als Startnäherung für das Newton–Verfahren als auch als Hilfe zur Fehlerschätzung.

4.3 Schrittweiten– und Ordnungssteuerung

Zunächst muß entschieden werden, ob die berechnete Lösung annähernd im gewünschten Toleranzbereich liegt. Falls dies nicht der Fall sein sollte, muß die Schrittweite h verkleinert werden. Nun haben jedoch die Untersuchungen in Abschnitt 1.3 und 3 gezeigt, daß mindestens in den algebraischen Komponenten eine Instabilität der Größenordnung $\frac{1}{h}$ auftritt. Dies bedeutet, daß eine Verkleinerung der Schrittweite in diesen Komponenten nicht umbedingt wünschenswert ist. Ein gewisser Kompromiß hinsichtlich der Schrittweitenbegrenzung nach oben als auch nach unten wird dadurch erreicht, daß man lediglich die differentiellen Komponenten auf ihre Genauigkeit überprüft und nur diese für den nächsten Schrittweitenvorschlag berücksichtigt. Entsprechend den Darlegungen im 3. Abschnitt ist dies für solche Probleme, die in Satz 3.5 angegeben sind, gerechtfertigt. Die Testuntersuchungen ergaben jedoch auch für andere Probleme bei dieser Vorgehensweise positive Ergebnisse.
Der Benutzer kann die relative und absolute Fehlertoleranz (RELTOL und ABSTOL) vorgeben. Dann wird der Integrationsschritt akzeptiert, falls die Relation

$$|EP^{(j)}_{n+1}| \leq A^{(j)} := RELTOL \cdot max(|Px^{(j)}_{n+1}|, |Px^{(j)}_n|) + ABSTOL$$

erfüllt ist, wobei P ein konstanter Projektor auf den Nullraum $ker(f'_y(y, x, t))$ ist und $EP_{n+1} := P * E_{n+1}$. Der Index (j) bezeichnet an dieser Stelle die j-te Komponente des entsprechenden Vektors.
Die größtmögliche Schrittweite h, damit diese Relation noch erfüllt ist, ergibt sich dann mit Hilfe von (4.4) aus der Beziehung

$$|EP^{(j)}_{n+1}| \cdot \left(\eta^{(j)}\right)^{k+1} = A^{(j)},$$

wobei $\min_j \eta^{(j)} =: \eta =: \frac{h}{t_{n+1}-t_n}$. Dies wäre also ein optimaler Schrittweitenvorschlag für den nächsten Integratiosschritt. Um auch hinsichtlich der Ordnung steuern zu können, werden nach dem gleichen Prinzip zwei weitere Schrittweitenvorschläge h^- und h^+ für die Ordnungen $k-1$ und $k+1$ bestimmt. Das Maximum dieser drei Vorschläge und die dazugehörige Ordnung werden dann beim nächsten Integrationsschritt verwendet. Die bei der Berechnung von h^- und h^+ benötigten

Koeffizienten γ_i^- und γ_i^+ für die entsprechenden Prädiktoren zur Fehlerschätzung lassen sich einfach mittels folgender Relationen ermitteln:

$$\begin{aligned}
\gamma_i^- &= \frac{t_{n-i} - t_{n-k}}{t_{n+1} - t_{n-k}}\gamma_i, \qquad i = 0, ..., k-1 \\
\gamma_i^+ &= \frac{t_{n+1} - t_{n-k-1}}{t_{n-i} - t_{n-k-1}}\gamma_i, \qquad i = 0, ..., k \\
\gamma_{k+1}^+ &= 1 - \sum_{i=0}^{k} \gamma_i^+.
\end{aligned}$$

Auf diese Weise ist der Aufwand für die Ordnungsbestimmung relativ gering und der dazugehörige Schrittweitenvorschlag an die Ordnung angepaßt.

4.4 Der erste Zeitschritt

Beim ersten Integrationsschritt steht nur der Startwert $x(t_0)$ zur Verfügung. Daher ist die eben beschriebene Schrittweitensteuerung nach dem ersten Integrationsschritt noch nicht möglich. Eine Möglichkeit, dennoch den ersten Fehler schätzen zu können, ist die in Denk [88] verwendete Idee: Der erste Zeitschritt wird einerseits mit einem Schritt der Schrittweite $h = t_1 - t_0$ und andererseits mit zwei Schritten der Schrittweite $\frac{h}{2}$ durchgeführt. Dann liefert die Differenz der so ermittelten Werte $x_1[h]$ und $x_1[\frac{h}{2}]$ in erster Näherung eine Schätzung des Fehlers nach einem halben Integrationsschritt. Verfolgt man wieder obige Philosophie, so kann der erste Schritt akzeptiert werden, falls die Relation

$$2 \cdot \left| x_1^{(j)}[h] - x_1^{(j)}[\frac{h}{2}] \right| \leq A^{(j)}$$

für jede Komponente j erfüllt ist, wobei die Fehlertoleranz analog der oben gewählten wie folgt gebildet wird:

$$A^{(j)} = RELTOL \cdot max\left(\left| x_1^{(j)}[h] \right|, \left| x_1^{(j)}[\frac{h}{2}] \right| \right) + ABSTOL.$$

Die neue Schrittweite h_{new} ergibt sich entsprechend aus der Beziehung

$$2 \cdot \left| x_1^{(j)}[h] - x_1^{(j)}[\frac{h}{2}] \right| \cdot \left(\eta^{(j)} \right)^2 = A^{(j)},$$

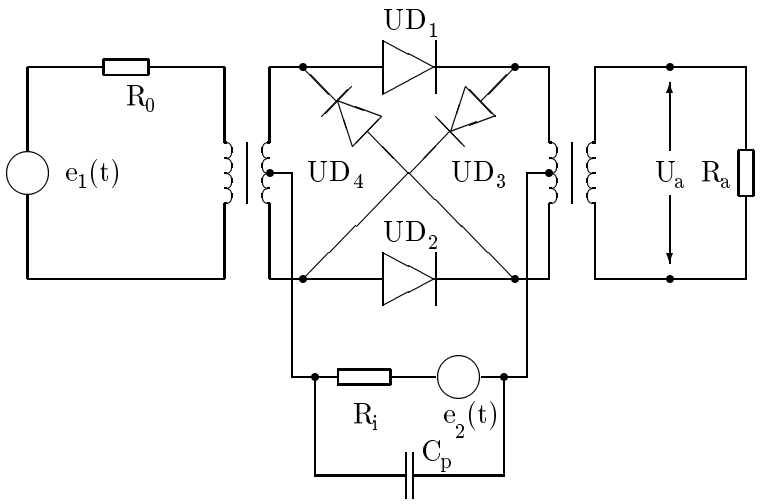
wobei $\min\limits_j \eta^{(j)} =: \eta =: \frac{h_{new}}{h}$.

4.5 Test–Beispiele

An dieser Stelle werden drei Beispiele angeführt, die die ersten Erfahrungen mit dem Programm DAE2SOL widerspiegeln sollen.

Ringmodulator

Der hier betrachtete Ringmodulator stellt einen kleinen Schaltkreis dar, bei dem ein hochfrequentes Signal $e_1(t)$ durch ein niederfrequentes Signal $e_2(t)$ überlagert wird.



Dieser Schaltkreis wurde von Horneber [76] mit Hilfe der folgenden 15 gewöhnlichen Differentialgleichungen simuliert.

$$
\begin{aligned}
C\dot{U}_1 &= I_1 - I_3 \cdot 0.5 + I_4 \cdot 0.5 + I_7 - U_1/R \\
C\dot{U}_2 &= I_2 - I_5 \cdot 0.5 + I_6 \cdot 0.5 + I_8 - U_2/R \\
C_S\dot{U}_3 &= I_3 - G(UD_1) + G(UD_4) \\
C_S\dot{U}_4 &= -I_4 + G(UD_2) - G(UD_3) \\
C_S\dot{U}_5 &= I_5 + G(UD_1) - G(UD_3) \\
C_S\dot{U}_6 &= -I_6 - G(UD_2) + G(UD_4) \\
C_P\dot{U}_7 &= U_7/R_i + G(UD_1) + G(UD_2) - G(UD_3) - G(UD_4) \\
L_h\dot{I}_1 &= -U_1 \\
L_h\dot{I}_2 &= -U_2 \\
L_{S2}\dot{I}_3 &= U_1 \cdot 0.5 - U_3 - R_{g2} \cdot I_3 \\
L_{S3}\dot{I}_4 &= -U_1 \cdot 0.5 + U_4 - R_{g3} \cdot I_4 \\
L_{S2}\dot{I}_5 &= U_2 \cdot 0.5 - U_5 - R_{g2} \cdot I_5 \\
L_{S3}\dot{I}_6 &= -U_2 \cdot 0.5 + U_6 - R_{g3} \cdot I_6
\end{aligned}
$$

$$
\begin{aligned}
L_{S1}\dot{I}_7 &= -U_1 + e_1(t) - (R_0 + R_{g1}) \cdot I_7 \\
L_{S1}\dot{I}_8 &= -U_2 - (R_a + R_{g1}) \cdot I_8,
\end{aligned}
$$

wobei die Charakteristik der Dioden durch

$$G(UD) = 40.67286402 \cdot 10^{-9} \cdot [exp(17.7493332 \cdot UD) - 1]$$

und die Spannungen an den verschiedenen Dioden durch

$$
\begin{aligned}
UD_1 &= U_3 - U_5 - U_7 - e_2(t) \\
UD_2 &= -U_4 + U_6 - U_7 - e_2(t) \\
UD_3 &= U_4 + U_5 + U_7 + e_2(t) \\
UD_4 &= -U_3 - U_6 + U_7 + e_2(t)
\end{aligned}
$$

gegeben sind. Für die technischen Parameter gilt:

$R_{g1} = 36.3\Omega, \quad R_{g2} = R_{g3} = 17.3\Omega, \quad R_0 = R_i = 50\Omega,$
$R_a = 600\Omega, \quad R = 25000\Omega,$
$C = 16 \cdot 10^{-9}F, \quad C_P = 10 \cdot 10^{-9}F,$
$L_h = 4.45H, \quad L_{S1} = 2 \cdot 10^{-3}H, \quad L_{S2} = L_{S3} = 0.5 \cdot 10^{-3}H,$
Trägersignal: $e_2(t) = 2 \cdot \sin(2\pi \cdot 10^4 \cdot t)$
zu modulierendes Signal: $e_1(t) = 0.5 \cdot \sin(2\pi \cdot 10^3 \cdot t)$.

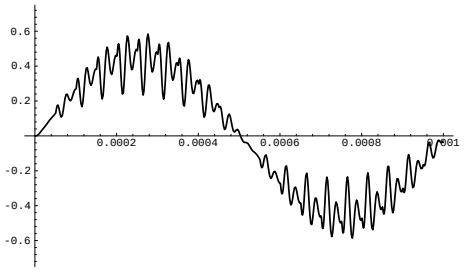
Am Anfangspunkt liegen fogende Werte vor:

$$
\begin{aligned}
U_j(0) &= 0, \qquad j = 1, ..., 7 \\
I_j(0) &= 0, \qquad j = 1, ..., 8.
\end{aligned}
$$

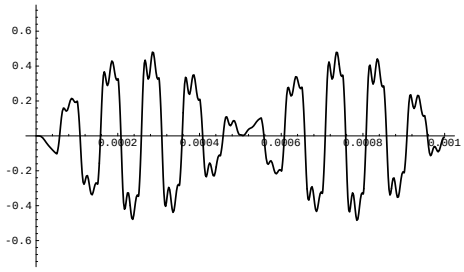
Die hierbei berücksichtigten kapazitiven Widerstände der Dioden C_S liegen bei Werten der Größenordnung $10^{-12}F$, wodurch das System außerordentlich steif wird. Bei den Berechnungen von Horneber war es nur möglich, dieses System für Werte von $C_S \geq 10^{-9}F$ zu lösen. Hinzu kam das Problem der langen Rechenzeiten. Die späteren Untersuchungen von Denk & Rentrop [89] und Hairer, Lubich & Roche [89] haben gezeigt, daß die auftretenden unerwünschten Oszillationen in den Diodenspannungen um so geringer werden, um so kleiner der Paramter C_S ist. Der gemessene Kurvenverlauf wurde am besten durch die Berechnungen für den Fall $C_S = 0$, wodurch das System zu einer Index–2–ADG wird, widergespiegelt. Die Ergebnisse von Denk & Rentrop [89] mit dem Programm DAE34, das Methoden vom Rosenbrock–Wanner–Typ benutzt, dienten mir als Anhaltspunkt für die Resultate, die mit dem Programm DAE2SOL erreicht wurden. Bei einer Toleranz von 10^{-6} wurden bei der Integration bis zum Zeitpunkt 10^{-4} folgende Resultate erreicht:

Programm	Schritte erfolgreich/schlecht	Schrittweitenbereich	Zeit
DAE34	496/84	$1 \cdot 10^{-6} - 4 \cdot 10^{-11}$	$\sim 3\frac{1}{2} min$
DAE2SOL	195/67	$1.8 \cdot 10^{-6} - 1 \cdot 10^{-8}$	$\sim 2 min$

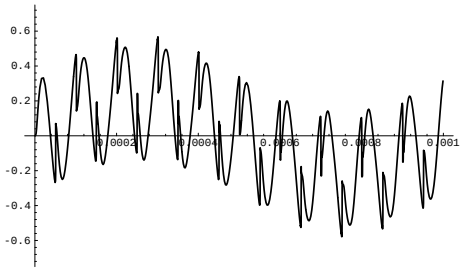
Zu den Zeitangaben ist hinzuzufügen, daß die Berechnungen mit DAE34 auf einem PC des Typs INTEL 8086/8087 CPU und die mit DAE2SOL auf einem PC des Typs INTEL 80386SX/80387SX CPU durchgeführt wurden. Die mit DAE2SOL bei der Integration bis zum Zeitpunkt 10^{-3} mit einer Toleranz (RELTOL und ABSTOL) von 10^{-6} erreichten Ergebnisse sind für einige Komponenten im folgenden dargstellt:



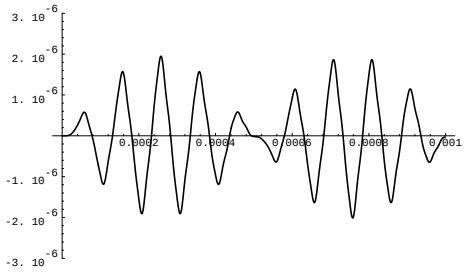
Spannung U_1



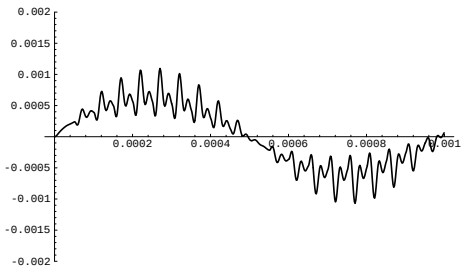
Spannung U_2



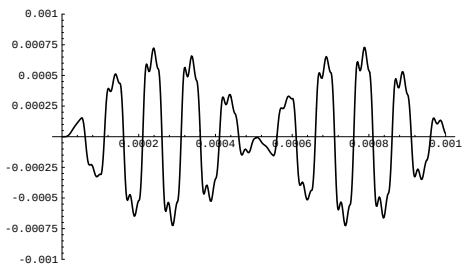
Spannung U_3



Strom I_2



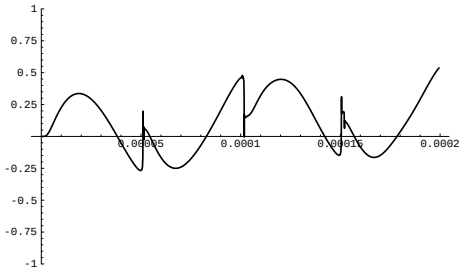
Strom I_7



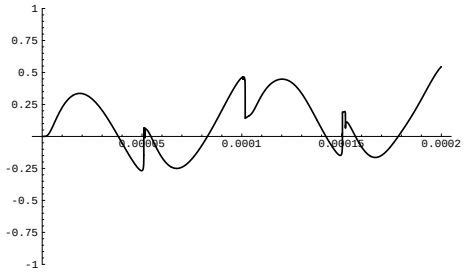
Strom I_8

Da bei der in DAE2SOL verwendeten Steuerung (Steuerung 1) die algebraischen Komponenten ausgeblendet wurden, ist es von Intersesse, wie der Kurvenverlauf

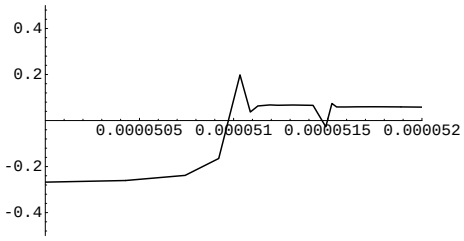
dieser algebraischen Komponenten im Vergleich zur Steuerung aller Komponenten (Steuerung 2) ausfällt. Die Unterschiede werden in der 3. Komponente U_3 sichtbar, wenn man zunächst nur das kurze Intervall $[0, 2 \cdot 10^{-4}]$ betrachtet und davon die Darstellung um den kritischen Punkt um $5 \cdot 10^{-5}$ genauer auflöst.



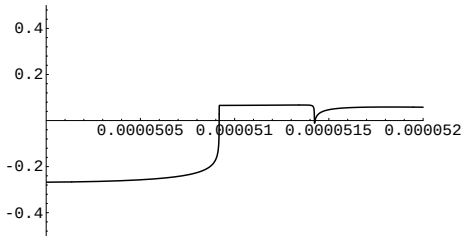
Steuerung 1
U_3 in $[0, 2 \cdot 10^{-4}]$



Steuerung 2
U_3 in $[0, 2 \cdot 10^{-4}]$



Steuerung 1
U_3 in $[5.0 \cdot 10^{-5}, 5.2 \cdot 10^{-5}]$



Steuerung 2
U_3 in $[5.0 \cdot 10^{-5}, 5.2 \cdot 10^{-5}]$

Kurzzeitig treten also Schwankungen in der algebraischen Komponente auf, die jedoch den Kurvenverlauf im gesamten nicht negativ beeinflussen. Der Aufwand, der bei der Steuerung aller Komponenten nötig ist, beträgt ungefähr das 4-fache der anderen Steuerung.

Lineares zeitabhängiges Beispiel

Betrachtet wurde das folgende lineare Index–2–System

$$\begin{aligned} tx_1' - tx_2' - (t+1)x_1 + x_2 &= 0 \\ x_1' - x_2' - x_1 &= t+1 \end{aligned}$$

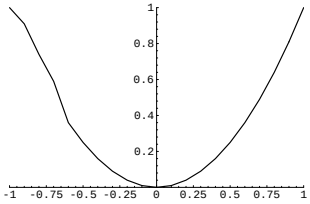
mit dem Anfangspunkt $x_1(-1) = 1$ und $x_2(-1) = 1$. Es liefert die Lösung

$$x_1 = -t, \qquad x_2 = t^2.$$

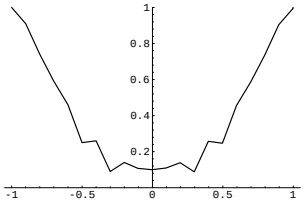
Dieses Beispiel ist ein sehr einfaches System, daß die Bedingungen des Satzes 3.5 erfüllt und mit der Kenntnis der Lösung eine Bestimmung des exakten Fehlers ermöglicht. Damit läßt sich hierfür die Qualität der Fehlerschätzung überprüfen. Bei der Steuerung, die nur auf die differentiellen Komponenten zurückgreift, wurde folgender Projektor P verwendet:

$$P = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}$$

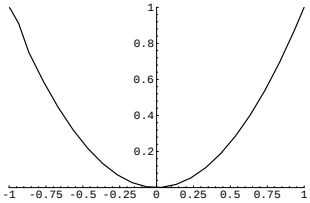
In den folgenden Abbildungen ist die zweite Komponente x_2 dargestellt. Dabei wurden Steuerung 1 (nur differentielle Komponenten) und Steuerung 2 (alle Komponenten) zum Vergleich gegenübergestellt.



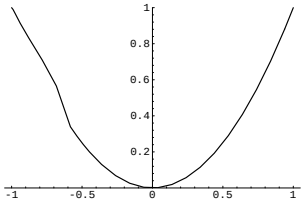
Steuerung 1: x_2 bei
$ABSTOL = RELTOL = 10^{-1}$



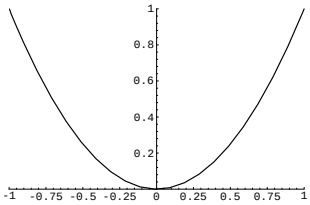
Steuerung 2: x_2 bei
$ABSTOL = RELTOL = 10^{-1}$



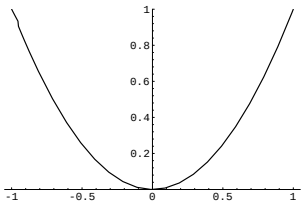
Steuerung 1: x_2 bei
$ABSTOL = RELTOL = 10^{-2}$



Steuerung 2: x_2 bei
$ABSTOL = RELTOL = 10^{-2}$



Steuerung 1: x_2 bei
$ABSTOL = RELTOL = 10^{-3}$



Steuerung 2: x_2 bei
$ABSTOL = RELTOL = 10^{-3}$

Interessant bei diesem Beispiel ist die Beobachtung der Anzahl der Schritte, die

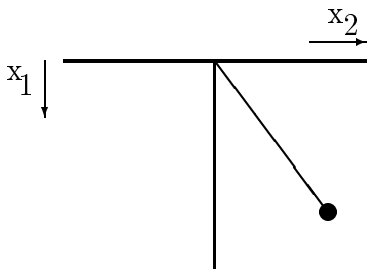
sich für beide Steuerungen bei den unterschiedlichen Fehlertoleranzen ergaben:

	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Steuerung 1	22	24	28	31	35	38
Steuerung 2	22	26	36	40	47	676

Hierbei ist also sowohl hinsichtlich der Qualität der Lösung als auch hinsichtlich des Aufwandes die Steuerung 1 vorzuziehen. Die geradezu explosionsartige Steigerung des Aufwandes bei Steuerung 2 ist auf Probleme bei der Richtungsfindung am Anfang der Lösungskurve zurückzuführen.

Das mathematische Pendel

Mit dem Programm DAE2SOL wurde zunächst auch die Originalversion des mathematischen Pendels getestet, welches ja bekanntlich auf ein ADG–System vom Index 3 führt.

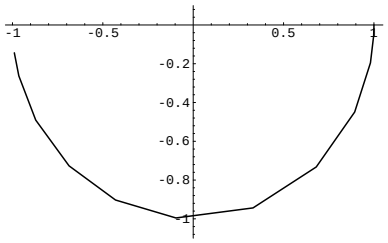


Die folgenden Gleichungen beschreiben das System:

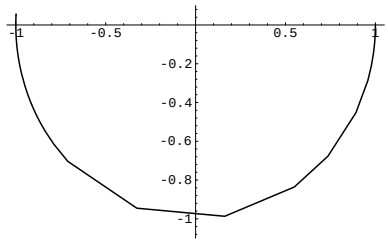
$$x'_1 = v_1 \tag{4.5a}$$
$$x'_2 = v_2 \tag{4.5b}$$
$$v'_1 = -g + 2x_1\lambda \tag{4.5c}$$
$$v'_2 = 2x_2\lambda \tag{4.5d}$$
$$x_1^2 + x_2^2 = 1. \tag{4.5e}$$

Bei den nachfolgenden Rechnungen ist anstelle der Erdbeschleunigung der Wert $g := 13.750371636041$ gewählt worden, wodurch die Periodendauer den Wert 2 besitzt. Dies ermöglicht eine Bestimmung des exakten Fehlers zum Zeitpunkt

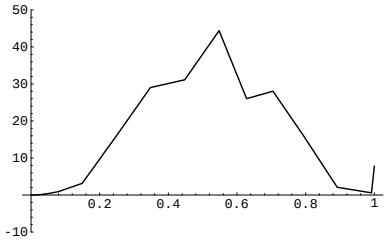
$t = 1$. Die algebraische Komponente (der Lagrange–Parameter λ) geht in diesem Beispiel nur linear in das System ein. Dieser Fakt läßt erwarten, daß die in DAE2SOL gewählte Steuerung geeigneter als eine Steuerung über alle Komponenten ist.
In der Tat bestätigte die Praxis die Erwartungen. Bei einer Steuerung aller Komponenten ist die Integration nur für Toleranzen der Größenordnung 10^{-1} möglich gewesen. Während die Pendelbewegung dabei noch relativ gut simuliert wird, passieren bei der Berechnung des Lagrange–Parameters förmlich Katastrophen. Steuert man nur die differentiellen Komponenten, so wird der Lagrangeparameter wesentlich besser approximiert, wenn auch der Kurvenverlauf noch zu Wüschen übrig läßt.



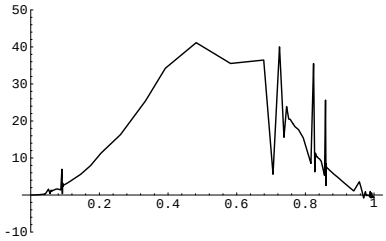
Steuerung 1
Pendelbewegung
$ABSTOL = RELTOL = 10^{-1}$



Steuerung 2
Pendelbewegung
$ABSTOL = RELTOL = 10^{-1}$

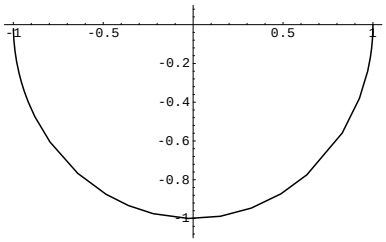


Steuerung 1
Lagrange–Parameter λ
$ABSTOL = RELTOL = 10^{-1}$

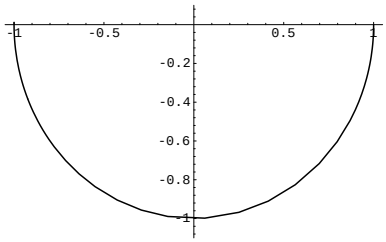


Steuerung 2
Lagrange–Parameter λ
$ABSTOL = RELTOL = 10^{-1}$

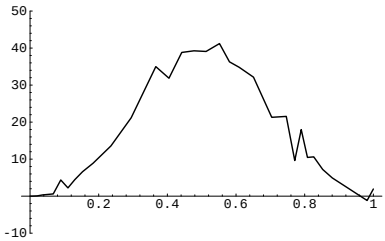
Doch ein wesentlicher Vorteil der Steuerung in DAE2SOL besteht darin, daß die Integration auch für kleinere Toleranzbereiche möglich ist und ab einer Genauigkeit von etwa 10^{-4} auch die Approximation des Lagrange–Parameters akzeptiert werden kann.



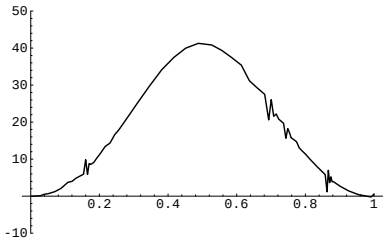
Steuerung 1
Pendelbewegung
$ABSTOL = RELTOL = 10^{-2}$



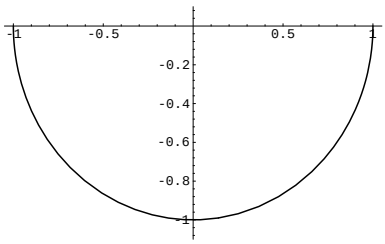
Steuerung 1
Pendelbewegung
$ABSTOL = RELTOL = 10^{-3}$



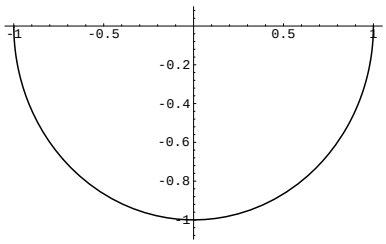
Steuerung 1
Lagrange–Parameter λ
$ABSTOL = RELTOL = 10^{-2}$



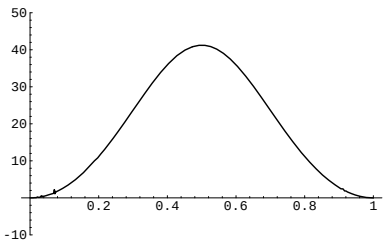
Steuerung 1
Lagrange–Parameter λ
$ABSTOL = RELTOL = 10^{-3}$



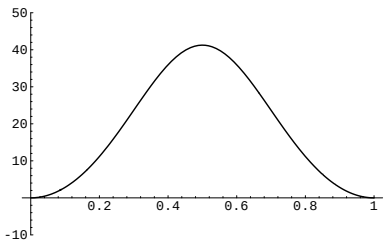
Steuerung 1
Pendelbewegung
$ABSTOL = RELTOL = 10^{-4}$



Steuerung 1
Pendelbewegung
$ABSTOL = RELTOL = 10^{-5}$



Steuerung 1
Lagrange–Parameter λ
$ABSTOL = RELTOL = 10^{-4}$



Steuerung 1
Lagrange–Parameter λ
$ABSTOL = RELTOL = 10^{-5}$

Von Interesse ist an dieser Stelle, welche Ergebnisse man erzielt, wenn man das obige System (4.5a)-(4.5e) auf ein System vom Index 2 zurückführt. Es wurden hier zwei relativ vielversprechende Varianten getestet, die u.a. auch in Brenan, Campell & Petzold [89] diskutiert wurden.
Die erste Variante besteht in der Differentiation der algebraischen Nebenbedingung und anschließenden Substitution der Ableitungen von x_1 und x_2:

$$x_1 v_1 + x_2 v_2 = 0. \tag{4.6}$$

Dann stellen die Gleichungen (4.5a)-(4.5d), (4.6) ein Index–2–System dar.
Die zweite Variante besteht in der zusätzlichen Einführung eines künstlichen Parameters μ, welches eine Stabilisierung des eben erhaltenen Systems zur Folge hat. Anstelle der Gleichungen (4.5a) und (4.5b) werden die Gleichungen

$$x_1' = v_1 + x_1\mu \tag{4.7a}$$
$$x_2' = v_2 + x_2\mu \tag{4.7b}$$

verwendet, und das System (4.7a), (4.7b), (4.5c)- (4.5e) und (4.6) besitzt wieder den Index 2. Es gilt weiter, daß $(x_1, x_2, v_1, v_2, \lambda, \mu)$ eine Lösung dieses Systems genau dann ist, wenn $\mu \equiv 0$ und $(x_1, x_2, v_1, v_2, \lambda)$ eine Lösung des Systems (4.5a)-(4.5d), (4.6) ist.
Die folgenden Tabellen geben Aufschluß über Aufwand und Genauigkeiten der verschiedenen Varianten bei Steuerung 1 (nur differentielle Komponenten) und Steuerung 2 (alle Komponenten):

Anzahl der Schritte (erfolgreich/schlecht):

RELTOL=ABSTOL	10^{-2}	10^{-4}	10^{-6}
Index 2, Var.2, Steuer.1	21/4	56/6	125/4
Index 2, Var.2, Steuer.2	56/54	153/145	761/842
Index 2, Var.1, Steuer.1	21/5	56/6	125/4
Index 2, Var.1, Steuer.2	44/40	60/17	479/517
Index 3, Steuer.1	34/21	81/42	515/467

Absoluter Fehler $|x^{(1)}(1) - x_*^{(1)}(1)|$:

RELTOL=ABSTOL	10^{-2}	10^{-4}	10^{-6}
Index 2, Var.2, Steuer.1	$2.1 \cdot 10^{-4}$	$1.2 \cdot 10^{-9}$	$5.2 \cdot 10^{-12}$
Index 2, Var.2, Steuer.2	$6.6 \cdot 10^{-5}$	$9.8 \cdot 10^{-10}$	$8.2 \cdot 10^{-13}$
Index 2, Var.1, Steuer.1	$1.3 \cdot 10^{-3}$	$4.2 \cdot 10^{-5}$	$2.1 \cdot 10^{-7}$
Index 2, Var.1, Steuer.2	$8.6 \cdot 10^{-3}$	$1.2 \cdot 10^{-6}$	$3.3 \cdot 10^{-6}$
Index 3, Steuer.1	$2.2 \cdot 10^{-4}$	$9.3 \cdot 10^{-8}$	$3.3 \cdot 10^{-11}$

Absoluter Fehler $|x^{(2)}(1) - x_*^{(2)}(1)|$:

RELTOL=ABSTOL	10^{-2}	10^{-4}	10^{-6}
Index 2, Var.2, Steuer.1	$2.0 \cdot 10^{-2}$	$4.9 \cdot 10^{-5}$	$3.2 \cdot 10^{-6}$
Index 2, Var.2, Steuer.2	$1.1 \cdot 10^{-2}$	$4.5 \cdot 10^{-5}$	$1.4 \cdot 10^{-6}$
Index 2, Var.1, Steuer.1	$9.5 \cdot 10^{-3}$	$6.7 \cdot 10^{-5}$	$1.8 \cdot 10^{-6}$
Index 2, Var.1, Steuer.2	$2.5 \cdot 10^{-2}$	$3.4 \cdot 10^{-4}$	$1.5 \cdot 10^{-5}$
Index 3, Steuer.1	$2.1 \cdot 10^{-2}$	$4.3 \cdot 10^{-4}$	$8.1 \cdot 10^{-6}$

Absoluter Fehler $|\lambda(1) - \lambda_*(1)|$:

RELTOL=ABSTOL	10^{-2}	10^{-4}	10^{-6}
Index 2, Var.2, Steuer.1	$2.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-4}$	$4.4 \cdot 10^{-5}$
Index 2, Var.2, Steuer.2	$1.5 \cdot 10^{-1}$	$6.1 \cdot 10^{-4}$	$1.9 \cdot 10^{-5}$
Index 2, Var.1, Steuer.1	$1.2 \cdot 10^{0}$	$9.3 \cdot 10^{-4}$	$5.6 \cdot 10^{-5}$
Index 2, Var.1, Steuer.2	$3.6 \cdot 10^{-1}$	$4.7 \cdot 10^{-3}$	$2.1 \cdot 10^{-4}$
Index 3, Steuer.1	$1.9 \cdot 10^{0}$	$1.4 \cdot 10^{-2}$	$1.4 \cdot 10^{-4}$

Absoluter Fehler $|\mu(1) - \mu_*(1)|$:

RELTOL=ABSTOL	10^{-2}	10^{-4}	10^{-6}
Index 2, Var.2, Steuer.1	$9.4 \cdot 10^{-3}$	$5.8 \cdot 10^{-9}$	$3.5 \cdot 10^{-10}$
Index 2, Var.2, Steuer.2	$1.9 \cdot 10^{-3}$	$5.5 \cdot 10^{-5}$	$9.1 \cdot 10^{-7}$

Absoluter Fehler $\max_j |\mu_j - \mu_*(t_j)|$:

RELTOL=ABSTOL	10^{-2}	10^{-4}	10^{-6}
Index 2, Var.2, Steuer.1	$6.3 \cdot 10^{-2}$	$1.3 \cdot 10^{-3}$	$8.4 \cdot 10^{-5}$
Index 2, Var.2, Steuer.2	$6.5 \cdot 10^{-2}$	$1.3 \cdot 10^{-3}$	$8.9 \cdot 10^{-5}$

Qualitativ liefert also die Index–2–Formulierung des mathematischen Pendels mit dem zusätzlichen Parameter μ die besten Ergebnisse. Hinsichtlich der Art der Steuerung liefert die Steuerung 2 (aller Komponenten) hierbei geringfügig bessere Ergebnisse als die Steuerung 1. Allerdings steigt der Aufwand bei Steuerung 2 mit zunehmender Genauigkeit rapide und ist wesentlich höher als bei Steuerung 1.

4.6 FORTRAN–Code

```
c subroutine dae2sol(m,t0,tend,x0,proj,fyx,fably,fablx,w,*,*,*)
c
c The Subroutine dae2sol solves DAEs of the form
c
c f(x'(t),x(t),t) = 0,
c
c where the partial Jacobian df/dy(y,x,t) is constant,
c with the initial condition x(t0) = x0
c by BDF method with variable stepsize and variable order (max.5).
c
c m dimension of the DAE
c t0 initial point
c tend last time point
c x0 initial solution
c proj(m,P) SUBROUTINE for calculating the project matrix P,
c where P:=I-Q and Q is a constant projector onto ker(df/dy(y,x,t))
c fyx(y,x,t) SUBROUTINE for calculating the DAE-Function f(y,x,t)
c dfy(y,x,t) SUBROUTINE for calculating the Jacobian df/dy(y,x,t)
c dfx(y,x,t) SUBROUTINE for calculating the Jacobian df/dx(y,x,t)
c
c *1 working array to short
c *2 additional opportunity for error messages
c *3 stepsize to small
```

Literatur

M. Berzins & R. M. Furzelands [85], *"A user's manual for SPRINT: Part 1"*, Dept. of Computer Studies Report 199, Leeds University.

K.E. Brenan, S.L. Campbell & L.R. Petzold [89], *"Numerical solution of initial-value problems in differential-algebraic equations"*, North Holland, New York-Amsterdam-London.

K.E. Brenan & B.E. Engquist [88], *"Backward differentiation approximations of nonlinear differential/algebraic systems"*, and Supplement, Math. Comp. 51, 659-676, S7-S16.

S. L. Campbell [85], *"The numerical solution of higher index linear time varying singular systems of differential equations"*, SIAM J. Sci. Stat. Comput. 6, 334-348.

S. L. Campbell [86], *"Index two linear time–varying system of differential equations"*, Circuits Systems Signal Process. 5, 97-107

G. Denk [88], *"Die numerische Integration von Algebro-Differentialgleichungen bei der Simulation elektrischer Schaltkreise mit SPICE2"*, Mathematisches Institut, TU München, Rep. TUM-M8809.

G. Denk & P. Rentrop [89], *"Mathematical Models In Electric Circuit Simulation And Their Numerical Treatment"*, Mathematisches Institut, TU München, Rep. TUM-M8903.

P. Deuflhard, E. Hairer & J. Zugck [87], *"One-step and extrapolation methods for differential-algebraic systems"*, Numer. Math. 51, 501-516.

J. Dieudonné [85], *"Grundzüge der modernen Analysis"*, Volume 1, Deutscher Verlag der Wissenschaften, Berlin.

C. Führer & B. Leimkuhler [89], *"Formulation and numerical solution of the equations of constrained mechanical motion"*, DFVLR-Forschungsbericht 89-08, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt, Oberpfaffenhofen.

F.R. Gantmacher [54], *"Teorija matrits"*, Gosudarstv. Izdat. Techn.-Teor. Lit., Moskva.

C.W. Gear & L.R. Petzold [84], *"ODE methods for the solution of differential/algebraic systems"*, SIAM J. Numer. Anal. 21, 716-728.

C.W. Gear, G.K. Gupta & B. Leimkuhler [85], *"Automatic integration of Euler-Lagrange equations with constraints"*, J. Comp. Appl. Math. 12 & 13, 77-90.

E. Griepentrog [91], *"Index reduction methods for differential-algebraic equations"*, Preprint 91-12, Humboldt-Univ. Berlin, Fachbereich Mathematik.

E. Griepentrog, M. Hanke & R. März [92], *"Towards a better understanding of differential-algebraic equations"*, In E. Griepentrog, M. Hanke & R. März, editors, Berlin Seminar on Differential-Algebraic Equations, pages 2-13, Humboldt-Univ. Berlin, Fachbereich Mathematik

E. Griepentrog & R. März [86], *"Differential-algebraic equations and their numerical treatment"*, Teubner Texte zur Mathematik 88, Leipzig.

E. Hairer, S. P. Nørsett & G. Wanner [87], *"Solving Ordinary Differential Equations I: Nonstiff Problems"*, Springer Series in Computational Mathematics 8.

E. Hairer, Ch. Lubich & M. Roche [89], *"The numerical solution of differential-algebraic systems by Runge-Kutta methods"*, Springer, Lecture Notes in Mathematics 1409.

E. Hairer & G. Wanner [91], *"Solving ordinary differential equations II: Stiff and differential-algebraic problems"*, Springer Series in Computational Mathematics 14.

M. Hanke [90], *"On the asymptotic representation of a regularization approach to nonlinear semiexplicit higher index differential-algebraic equations"*, IMA J. Appl. Math.

B. Hansen [89], *"Comparing different concepts to treat differential-algebraic equations"*, Preprint 220, Humboldt-Univ. Berlin, Sektion Mathematik.

A. C. Hindmarsh [80], *"LSODE and LSODI, two new initial value ordinary differential equation solvers"*, ACM–SIGNUM Newsletters 15, 10-11.

E.-H. Horneber [76], *"Analyse nichtlinearer RLCÜ–Netzwerke mit Hilfe der gemischten Potentialfunktion mit einer systematischen Darstellung der Analyse nichtlinearer dynamischer Netzwerke"*, Dissertation, FB: Elektrotechnik, Univ. Kaiserslautern.

B. J. Leimkuhler [86], *"Error estimates for differential-algebraic equations"*, Technical Report UIUCDCD-R-86-1287, Dept. of Computer Science Univ. of Illinois.

P. Lötstedt & L. Petzold [86], *"Numerical solution of nonlinear differential equations with algebraic constraints I: Convergence results for backward differentiation formulas"*, Math. Comp. 46, 491-516.

A. K. Louis [89] *"Inverse und schlecht gestellte Probleme"*, Teubner Studienbücher: Mathematik.

R. März [89] *"Index-2 differential–algebraic equations"*, Results in Mathematics 15, 148-171.

R. März [90], *"Higher-index differential-algebraic equations: Analysis and numerical treatment"*, Banach Center Publ. 24, 199-222.

R. März [91], *"On quasilinear index 2 differential algebraic equations"*, Preprint 269, Humboldt-Univ. Berlin, Fachbereich Mathematik.

R. März [92], *"Numerical methods for differential-algebraic equations"*, Acta Numerica, 141-198.

L. R. Petzold [83], *"A Description of DASSL: A differential/algebraic system solver"*, Scientific Computing, eds R. S. Stepleman et al., North Holland, Amsterdam, 65-68.

L. R. Petzold & Lötstedt [86], *"Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications"*, SIAM J. Sci. Stat. Comput. 7, 720-733.

P.J. Rabier & W.C. Rheinboldt [91], *"A general existence and uniqueness theory for implicit differential-algebraic equations"*, Differential and Integral Equations 4 (3), 563-582.

S. Reich [90], *"Beitrag zur Theorie der Algebrodifferentialgleichungen"*, Dissertation (A), Techn. Univ. Dresden.

W.C. Rheinboldt [84], *"Differential-algebraic systems as differential equations on manifolds"*, Math. Comp. 43, 473-482.

H. Wriedt [88], *"Über Theorie und Numerik von Algebro–Differenrtial–Gleichungssystemen"*, Diplomarbeit, Universität Hamburg.

Thesen

1. Als Algebro–Differentialgleichung (ADG) bezeichnet man implizite gewöhnliche Differentialgleichungen der Form

$$f(x'(t), x(t), t) = 0, \quad f : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m,$$

deren partielle Ableitung $f'_y(y, x, t)$ singulär ist und konstanten Rang auf dem Definitionsgebiet von f hat. Hierbei sind reguläre gewöhnliche Differentialgleichungen mit algebraischen Gleichungen verknüpft. Algebro–Differentialgleichungen verkörpern sowohl Integrations– als auch Differentiationsprobleme. Der Index einer ADG gibt Aufschluß über die Anzahl der Differentiationen, die zur Lösung der ADG notwendig werden. Der Lösungsbegriff für ADGen ist somit vom Index dieser abhängig.

2. Die numerische Behandlung von Algebro–Differentialgleichungen erfordert Kenntnisse über deren Struktur. Einige Komponenten der Lösung einer ADG sind algebraisch bestimmt. Das bedeutet für die Lösbarkeit von Anfangswertaufgaben, daß die Anfangswerte nicht frei wählbar sind. Sie müssen "konsistent" mit der ADG sein. Ungenauigkeiten bei der Lösung der algebraischen Gleichungen führen zu einem "Abdriften" von der Lösungsmannigfaltigkeit, womit Instabilitäten hervorgerufen werden.

3. Unter den für reguläre gewöhnliche Differentialgleichungen bekannten numerischen Verfahren haben sich zur Lösung von ADGen die BDF in vielen Fällen bewährt. Schwierigkeiten tauchen bei Systemen höheren Index (≥ 2) auf, deren Grad zunimmt, um so größer der Index ist. Bereits bei linearen Index–2–Problemen ist schon die Durchführbarkeit der BDF nicht immer gewährleistet.

4. Die BDF der Ordnung $p \geq 2$ sind für ADGen vom Index 2 der Form

$$f(x'(t), x(t), t) = 0, \qquad ker(f'_y(y, x, t)) \equiv const,$$

durchführbar, konvergent und schwach instabil. Bei linearen Problemen betrifft die Instabilität nur die algebraischen Komponenten, bei denen aufgrund der auftretenden Differentiation auch nichts besseres zu erwarten ist.

5. Das implizite Euler–Verfahren bewältigt als ein Verfahren erster Ordnung die auftretende Differentiation bei solch allgemeinen nichtlinearen Index–2–Systemen nicht ohne Probleme. Selbst bei einfacheren Aufgaben vom Index 2 der Form

$$A(t)x'(t) + g(x(t), t) = 0, \qquad ker(A(t)) \equiv const,$$

ist Durchführbarkeit und Konvergenz nur gewährleistet, wenn noch zusätzliche Bedingungen an die Genauigkeit der Lösung der beim Verfahren entstehenden nichtlinearen Gleichungen in Abhängigkeit von der verwendeten Schrittweite gestellt werden. Die auftretende schwache Instabilität betrifft i.a. nicht mehr nur die algebraischen, sondern alle Komponenten.

6. Die Resultate, die man von den BDF höherer Ordnung kennt, sind für das implizite Euler–Verfahren dann zu erreichen, wenn die algebraischen Komponenten lediglich linear in das System eingehen. In diesem Fall sind die Fehler, die bei der Lösung der nichtlinearen Gleichungen entstehen, nicht mehr mit den Fehlern, die durch das numerische Verfahren hervorgerufen werden, verkoppelt.

7. Bei der praktischen Durchführung der BDF für Index–2–Probleme ist die auftretende Instabilität mindestens in den algebraischen Komponenten zu berücksichtigen. Zu kleine Schrittweiten führen zu schlechteren numerischen Ergebnissen in diesen Komponenten. Eine Möglichkeit dieser Schwierigkeit zu begegnen, ist die Ausblendung der algebraischen Komponenten aus der Fehlerschätzung für die Schrittweiten– und Ordnungssteuerung. Die auf diese Weise erzielten Resultate sind in den getesteten Beispielen zufriedenstellend, jedoch sollten weitere Untersuchungen zu diesem Problem unternommen werden, um für die Praxis zuverlässige Ergebnisse liefern zu können.

Versicherung

Ich versichere, daß ich diese Diplomarbeit selbständig unter Verwendung der angegebenen Literatur verfaßt und keine außer den angegebenen Hilfsmitteln verwendet habe.

Berlin, 27.10.1992

Caren Tischendorf