Karlsruher Institut für Technologie

INSTITUT FÜR ANGEWANDTE UND NUMERISCHE MATHEMATIK 1

DIPLOMARBEIT

Numerische Verfahren für mechanische Mehrkörpersysteme

Betrachtung einer speziellen
Differential-Algebraischen-Gleichung vom Index 2

Autor:
Björn Braun

Betreuerin:
Prof. Dr. Marlis Hochbruck

Abgabe: 4. November 2013

Abstract

In dieser Diplomarbeit werden verschiedene Ansätze zur Lösung einer aus der Mechanik stammenden Differential-Algebraischen-Gleichung vom Index 2 untersucht. Es wird zum einen das direkte Lösen der Differential-Algebraischen-Gleichung durch Runge-Kutta-Verfahren behandelt, hierfür werden Ordnungsbedingungen für die auf Differential-Algebraische-Gleichungen angewendeten Verfahren hergeleitet. Besonderes Augenmerk fällt auf zu Kollokationsverfahren äquivalente Runge-Kutta-Verfahren. Zum anderen wird eine in [Dep13] vorgestellte Regularisierungsmethode aufgegriffen, welche letztlich das mechanische System in Abhängigkeit von einem Parameter ϵ neu modelliert. Es wird gezeigt, dass der Grenzfall für $\epsilon \to 0$ zur Lösung der Differential-Algebraischen-Gleichung führt, sowie welchen Einfluss Störungen in den Anfangswerten besitzen. Dies führt zum Begriff der *Boundary Layers*.

Ich erkläre, dass ich diese Arbeit selbstständig angefertigt und nur die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder Sinn nach anderen Werken, gegebenenfalls auch elektronischen Medien, entnommen sind, sind von mir durch Angabe der Quelle als Entlehnung kenntlich gemacht. Entlehnungen aus dem Internet sind durch Angabe der Quelle und des Zugriffsdatums sowie dem Ausdruck der ersten Seite belegt; sie liegen zudem für den Zeitraum von 2 Jahren entweder auf einem elektronischen Speichermedium im PDF-Format oder in gedruckter Form vor.

Inhaltsverzeichnis

Vorwort

Diese vorliegende Diplomarbeit entstand aus einer Zusammenarbeit des Instituts für angewandte und numerische Mathematik und dem Institut für technische Mechanik des Karlsruher Institutes für Technologie (KIT).
Sie befasst sich mit einem in [Dep13] betrachteten, aus der theoretischen Mechanik stammenden Typ Differential-Algebraischer-Gleichungen vom Index 2, der bei der Beschreibung von Mehrkörpersystemen mit Roll- und Reibungsvorgängen im Fall der Haftreibung vorliegt.
Untersucht werden sowohl die theoretischen Grundlagen des numerischen Lösens dieser Differential-Algebraischen-Gleichung durch Runge-Kutta-Methoden, als auch eine in [Dep13] vorgeschlagene Regularisierung. Die anschauliche Entsprechung der vorgeschlagenen Regularisierung besteht darin, den Kontaktpunkt nicht mehr als Kontakt solider Körper zu betrachten, sondern die Oberfläche als System von Federn und viskoser Dämpfung zu modellieren, deren Parameter zur Regularisierung genutzt werden. Aus mathematischer Sicht nähert diese Regularisierung die Index 2 Differential-Algebraische-Gleichung durch eine (im Allgemeinen steife) explizite Differentialgleichung an. Aus physikalischer Sicht beschreibt sie einen anderen Modellansatz für die betrachteten Reibungsvorgänge als durch die strikte Differential-Algebraische-Gleichung. Von diesem anderen Modellansatz wird erhofft, die reellen Vorgänge bei einem derartigen Vorgang besser zu beschreiben, da die Modellbildung mit den strengen Nebenbedingungen einer Differential-Algebraischen-Gleichung an ihre Grenzen stößt. Es werden zwei Sonderfälle dieser Regularisierung betrachtet, namentlich der Fall *Strong Damping*, welcher letztlich der Standardform eines singulär gestörten Problems entspricht, sowie der Fall *Principal Damping*, eine in [Dep13] neu vorgeschlagene, verallgemeinerte singuläre Störung.
Die vorgestellte Regularisierung nach [Dep13] ist Teil eines Projektes, allgemeine Roll- und Reibungsvorgänge in einem ganzheitlichen System zu modellieren. Im Zuge dieser Diplomarbeit wird sich auf die Betrachtung des Haftreibungsfalles beschränkt.

Es werden im Folgenden die mathematischen Grundlagen zu Differential-Algebraischen-Gleichungen vorgestellt. Anschließend wird gezeigt, wie Runge-Kutta-Verfahren direkt auf Differential-Algebraische-Gleichungen angewendet werden können, sowie welchen Einfluss die Regularisierung nach [Dep13] auf die Lösung hat. Es werden Konvergenzuntersuchungen der verwendeten numerischen Methoden und Regularisierungen angestellt, zunächst theoretisch, im Anschluss als numerisches Experiment mit einem Modellbeispiel. Als numerisches Referenzverfahren wird insbesondere eine Modifikation eines 3-stufigen Runge-Kutta-Verfahrens der Ordnung 5 (Radau5) näher betrachtet, welche sowohl auf die Differential-Algebraischen-Gleichungen als auch auf die regularisierte Differentialgleichung angewandt werden kann. Im Detail ist dies eine in Matlab-Code übertragene Fortranversion aus [Hai10][†], die im Zuge dieser Diplomarbeit genutzt und zum Teil modifiziert wurde.

[†] Der ursprüngliche FORTRAN-Code, vorgeschlagen etwa in [HLR89], wurde von Ch. Engstler in Matlab-Code übertragen, siehe auch [Hai].

Kapitel 1

Grundlagen

Explizite, autonome Differentialgleichungen (DGL) erster Ordnung, also Probleme der Form

$$\dot{y} = F(y) \tag{1.1}$$

mit einer gesuchten Funktion $y : \mathbb{R} \to \mathbb{R}^n$ und einer gegebenen Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$, die je nach Problemstellung verschiedene Anforderungen an Stetigkeit oder Differenzierbarkeit erfüllt, sind die wohl am besten untersuchten gewöhnlichen Differentialgleichungen. Insbesondere lassen sich (explizite) Systeme höherer Ordnung oder nichtautonome Probleme durch Einführung zusätzlicher Variablen auf diese Form bringen, entsprechend beschränken sich die Konvergenzuntersuchungen vieler numerischer Lösungsverfahren auf diese Form.
Allgemeine gewöhnliche Differentialgleichungen, aus oben genannten Gründen werden hier nur autonome Probleme erster Ordnung betrachtet, haben die Form

$$F(\dot{y}, y) = 0 \tag{1.2}$$

mit $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. Diese müssen nicht explizit nach $\dot{y}$ auflösbar sein, etwa wenn $\frac{dF}{d\dot{y}}$ singulär ist. Dieser Fall liegt bei einer allgemeineren Klasse gewöhnlicher Differentialgleichungen vor, den Differential-Algebraischen-Gleichungen.
In diesem ersten Kapitel werden grundlegende Möglichkeiten vorgestellt, Differential-Algebraischen-Gleichung zu klassifizieren. Dies führt zu den Begriffen *Differentiations-* und *Störungsindex*. Weiterhin wird der in dieser Diplomarbeit behandelte Typ Differential-Algebraische-Gleichung aus dem mechanischen Problem hergeleitet. Dieser wird im Folgenden mit *das Modellproblem* bezeichnet. Anschließend werden die Indizes dieses Problems bestimmt.

1.1 Differential-Algebraische-Gleichungen und ihr Index

Sogenannte Differential-Algebraische-Gleichungen sind gewöhnliche, im Allgemeinen implizite Differentialgleichungen der Form (1.2) mit hinreichend oft differenzierbarem F. Typische Ursprünge derartiger Gleichungen sind diverse Anwendungen etwa aus Kontrolltheorie oder Mechanik. Die in dieser Diplomarbeit als Modellproblem dienende Differential-Algebraische-Gleichung entstammt der theoretischen Mechanik. Für den Fall singulärer partieller Ableitung $\frac{dF}{d\dot{y}}$ ist dieses System nicht rein algebraisch in eine explizite Differentialgleichung umformbar. Dies sind die sogenannten echten Differential-Algebraischen-Gleichungen, die im Folgenden behandelt werden.
Eine Möglichkeit mit Differential-Algebraischen-Gleichungen umzugehen, ist die Umformung in eine explizite DGL. Hierfür ist es bei echten Differential-Algebraischen-Gleichungen notwendig, das DGL-System zusätzlich zu den algebraischen Umformungen auch (eventuell mehrfach) abzuleiten. Dies bietet zugleich eine Möglichkeit, Differential-Algebraische-Gleichungen zu klassifizieren:

Definition 1.1 Der *Differentiationsindex* einer Differential-Algebraischen-Gleichung (1.2) ist die minimale Anzahl von Ableitungen von F, die benötigt wird, um (1.2) mit Hilfe dieser Ableitungen und ansonsten rein algebraisch in eine explizite Differentialgleichung umzuformen.

Unechte Differential-Algebraische-Gleichungen benötigen keine Ableitung und haben damit Index 0. Für höhere Indizes betrachten wir nun eine spezielle Form: Sei eine Differential-Algebraische-Gleichung aus der Grundform (1.2) algebraisch umformbar zu

$$\begin{cases} \dot{\bar{y}} &= f(y,z) \\ 0 &= g(y,z), \end{cases} \tag{1.3}$$

wobei (y,z) eine neue Variable ist, die durch Variablentransformation aus der bisherigen Variable y entsteht. Diese spezielle Form ist insbesondere noch später wichtig. Anschaulich entsprechen diese Gleichungen einer expliziten Differentialgleichung bezüglich $\bar{y}$ mit einer zusätzlichen algebraischen Nebenbedingung bzw. die Einschränkung der Lösung auf eine Mannigfaltigkeit, die durch g definiert wird. Der Index der Differential-Algebraischen-Gleichung (1.3) hängt nun von den Eigenschaften von g ab. Einmalige Differentiation der zweiten Zeile von (1.3) liefert

$$0 = g_y(y,z)\dot{y} + g_z(y,z)\dot{z} = g_y(y,z)f(y,z) + g_z(y,z)\dot{z}. \tag{1.4}$$

Hierbei und im Folgenden seien g_y bzw. g_z die partiellen Ableitungen von $g(y,z)$ nach y respektive z. Ist g_z in einer Umgebung der Lösung nichtsingulär, kann diese Gleichung nach $\dot{z}$ aufgelöst werden und es liegt ein Index 1 Problem vor. Falls nicht, lässt sich (1.4) analog zur ursprünglichen Differential-Algebraischen-Gleichung algebraisch umformen zu

$$\begin{cases} \dot{z}_1 &= g_1(y,z_1,z_2) \\ 0 &= g_2(y,z_1,z_2) \end{cases},$$

wobei z wiederum in z_1 und z_2 zerlegt, bzw. transformiert wurde. Falls $\frac{\partial g_2}{\partial z_2}$ in einer Umgebung der Lösung nichtsingulär ist, dann hat das Gesamtsystem den Index 2. Falls nicht, ergeben sich analog die höheren Indizes.

Bemerkung 1 *Über den Differentiationsindex*

- *Der Differentiationsindex ist invariant unter Äquivalenz erhaltenden, algebraischen Transformationen. Dies ermöglicht auch die obige Methode zur Bestimmung des Index von (1.2): Die minimale Anzahl der Ableitungen von f, g, g_1, g_2 und algebraischen Abwandlungen davon zu zählen ist äquivalent dazu, die minimal benötigten Ableitungen von F zu bestimmen.*
- *Das obige Verfahren, das nach Ableitung der algebraischen Bedingung den Anteil der Variablen, bezüglich dem diese Gleichung invertierbar ist, vom Rest separiert, kann auch als eine Art analytisches Gauß-Verfahren aufgefasst werden. Man sortiert damit nach Variablen mit aufsteigendem Index. Dies wird insbesondere als Vorarbeit für die Übergabe an diverse numerische Verfahren vorausgesetzt; etwa der später verwendete Radau5-Code benötigt zum Lösen von Differential-Algebraischen-Gleichungen vom Index größer 1 die Dimension der (sortierten) Index 1, 2 und 3-Variablen.*

Eine weitere Möglichkeit der Klassifizierung ergibt sich durch das Verhalten der Differential-Algebraischen-Gleichung bei Störung, der *Störungsindex*:

Definition 1.2 Gegeben sei eine Differential-Algebraische-Gleichung (1.2) mit Lösung $y(t)$ auf einem Intervall $[0, \bar{t}]$. Für beliebiges $\tilde{y}(t)$ bezeichne die Funktion $\delta(t)$ den Defekt, der sich bei Einsetzen von $\tilde{y}$ in die Differential-Algebraische-Gleichung ergibt, d.h.

$$F(\dot{\tilde{y}}(t), \tilde{y}(t)) = \delta(t).$$

Der *Störungsindex* ist die kleinste natürliche Zahl $m \in \mathbb{N}$, sodass es auf $[0, \bar{t}]$ eine Fehlerabschätzung für $\tilde{y}$ der Form

$$\|y(t) - \tilde{y}(t)\| \leq C \left(\|y(0) - \tilde{y}(0)\| + \sum_{i=1}^{m} \max_{0 \leq \xi \leq \bar{t}} \left\| \delta^{(i-1)}(\xi) \right\| \right)$$

gibt, soweit die Störung $\delta(t)$ hinreichend klein ist.

Diese Definition eines Index, in [HLR89] eingeführt, wurde im Hinblick auf das numerische Lösungsverhalten der Differential-Algebraischen-Gleichungen gewählt und spielt insbesondere bei der Anpassung von Runge-Kutta-Verfahren für Differential-Algebraische-Gleichungen noch eine Rolle. Der Sonderfall $m = 0$ kann mit einer -1-ten Ableitung von δ als Integralgleichung aufgefasst werden, wird hier jedoch nicht weiter benötigt. Der Störungsindex unterscheidet sich potentiell (aber nicht notwendig) vom Differentiationsindex, tatsächlich stimmen beide Indizes für das im Folgenden betrachtete Problem überein, dies wird später gezeigt. Allerdings gibt es auch Beispiele (siehe [Hai10], S. 461) von Differential-Algebraischen-Gleichungen, bei denen Differentiations- und Störungsindex beliebig weit auseinander liegen können.
Heuristisch sind Differential-Algebraische-Gleichungen besser zu lösen, je kleiner ihr Index ist. Deutlich wird dies besonders bei für Differential-Algebraische-Gleichungen angepasste Runge-Kutta-Verfahren: Ab (Störungs-)Index 2 gibt es, wie sich später zeigen wird, bereits Probleme, die Konvergenz des im Verfahren benutzten vereinfachten Newton-Verfahrens zu zeigen. Weiterhin haben numerische Verfahren häufig nur noch eine verminderte Konvergenzordnung bezüglich der Komponenten von höherem Index.

1.2 Das betrachtete Modellproblem

Das in [Dep13] untersuchte grundlegende Modellproblem ist ein mechanisches Mehrkörpersystem, wobei die Kontaktpunkte zwischen sich berührenden Körpern im Zustand der Haftreibung sind. Anschaulich bedeutet dies, dass kein „Schleifen“ stattfindet (Gleitreibung), sondern ein echtes „Rollen“. Es geht also keine mechanische Energie in Reibung verloren, sondern es ergeben sich zusätzliche Zwangsbedingungen an Lage oder Relativgeschwindigkeiten.
Mathematisch beschreiben kann man dieses System etwa durch die Lagrangegleichungen; gegeben sei hierzu die Lagrangefunktion (noch ohne Reibungsterm)

$$L(q, \dot{q}) = E_{\text{Kin}}(q, \dot{q}) - E_{\text{Pot}}(q) \tag{1.5}$$

als Differenz von kinetischer Energie E_{Kin} und potentieller Energie E_{Pot} des Systems, abhängig von generalisierten Koordinaten q und deren Ableitung. Die Bewegungsgleichungen ergeben sich durch die Lagrangegleichungen

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = \frac{\partial L}{\partial q_i}. \tag{1.6}$$

Allgemein haben diese die Form

$$M(q)\ddot{q} = F(q, \dot{q}, t)$$

mit einer verallgemeinerten Massematrix $M(q)$ und dem verallgemeinerten Kräftevektor $F(q, \dot{q}, t)$. Wieder kann man dieses System durch die zusätzliche Variable t mit $\dot{t} = 1$ in ein autonomes System überführen. Daher betrachten wir im Folgenden o.B.d.A. nur den autonomen Fall ohne t. Zusätzliche Nebenbedingungen der Form $\bar{g}(q) = 0$ können in der Lagrangefunktion durch Lagrangemultiplikatoren λ ergänzt werden, da der Lagrangeformalismus letztlich auf dem Prinzip der *Wirkungsminimierung* der Funktion L besteht (Vergleiche mit Kapitel 14 in [Fli09]). (1.5) wird damit zu

$$L(q, \dot{q}) = E_{\text{Kin}}(q, \dot{q}) - E_{\text{Pot}}(q) - \bar{g}(q)^T \lambda.$$

Einsetzen in die Lagrangegleichungen, wobei λ als weitere Variable aufgefasst wird, führt damit zu

$$M(q)\ddot{q} = F(q, \dot{q}) - G(q)^T \lambda \tag{1.7}$$
$$0 = \bar{g}(q), \tag{1.8}$$

wobei $G(q) = \frac{\partial \bar{g}(q)}{\partial q}$. Die zweite Zeile ergibt sich, da $\frac{\partial L}{\partial \lambda} = 0$.
Im Allgemeinen können die Zwangsbedingungen für Haftreibung jedoch nicht in einer nur von der Koordinate q abhängigen Form*geschrieben werden, sondern hängen auch von $\dot{q}$ ab. Sie sind dann in der Form

$$0 = G(q)\dot{q}. \tag{1.9}$$

In diesem Fall kann man die Herleitung der Bewegungsgleichung mit einem formalen $\bar{g}(q)$ analog durchführen, muss aber am Schluss (1.8) durch (1.9) ersetzen.
Löst man (Invertierbarkeit der Matrix $M(q)$ vorausgesetzt) (1.7) nach $\ddot{q}$ auf und führt das System (1.7), (1.9) durch Einführen einer neuen Variable für $\dot{q}$ in ein System erster Ordnung über, ergibt sich schließlich eine Differential-Algebraische-Gleichung der Form

$$\begin{cases} \dot{y} &= f(y, \lambda) \\ 0 &= g(y), \end{cases} \tag{1.10}$$

wobei $y = (q, \dot{q}), g(y) = G(q)\dot{q}, f(y, \lambda) = (\dot{q}, M(q)^{-1}(F(q, \dot{q}) - G(q)^T\lambda))^T$.

Grundsätzlich gehen wir von folgenden Modellannahmen auf einem beschränkten Zeitintervall $[0, \bar{t}]$ aus:

Modellannahmen

1. $M(q) \in \mathbb{R}^{n \times n}$ ist symmetrisch und positiv definit in einer hinreichenden Lösungsumgebung.
2. $G(q) \in \mathbb{R}^{m \times n}, m \leq n$, hat vollen Rang m.
3. $g_y(y) f_\lambda(y, \lambda) \in \mathbb{R}^{m \times m}$ ist in einer Umgebung der Lösung invertierbar.
4. Die Lösungen $y = y(t)$ und $\lambda = \lambda(t)$, sowie f und g sind hinreichend oft differenzierbar.
5. Insbesondere ist f Lipschitz-stetig und g_y sowie f_y und f_λ sind in einer Umgebung um die Lösung beschränkt.

Mit diesen Annahmen lassen sich Differentiations- und Störungsindex des Modellproblems bestimmen.

*Das wären sogenannte holonome Zwangsbedingungen. Im Gegensatz dazu bedeuten nichtholonome Zwangsbedingungen, dass es kein $\bar{g}(q)$ gibt, sodass $\frac{d\bar{g}(q)}{dt} = G(q)\dot{q}$, um dem Titel von [Dep13] Tribut zu zollen.

1.2.1 Differentiationsindex

Betrachten wir das System (1.10). Während die erste Zeile bereits eine explizite Differentialgleichung für y darstellt, muss dies für λ noch aus der zweiten Zeile hergeleitet werden. Einmaliges Differenzieren nach t liefert

$$0 = g_y(y)\dot{y} = g_y(y)f(y,\lambda).$$

Diese Gleichung enthält zwar noch kein $\dot{\lambda}$, jedoch eine Abhängigkeit von λ, die sich durch erneutes Differenzieren ausnutzen lässt (zugunsten der Lesbarkeit nun ohne Argumente):

$$0 = g_{yy}(f,f) + g_y f_y f + g_y f_\lambda \dot{\lambda} \tag{1.11}$$

Da nach Modellannahme der Term $g_y f_\lambda$ invertierbar ist, lässt sich (1.11) nach $\dot{\lambda}$ auflösen und wir erhalten, nach nun insgesamt zweimaligem Differenzieren (von Teilen) der Differential-Algebraischen-Gleichung, die explizite Differentialgleichung

$$\begin{aligned}\dot{y} &= f\\ \dot{\lambda} &= -(g_y f_\lambda)^{-1}(g_{yy}(f,f) + g_y f_y f),\end{aligned}$$

somit liegt Differentiationsindex 2 vor.

1.2.2 Störungsindex

Sei (y,λ) eine Lösung der Differential-Algebraischen-Gleichung (1.10). $(\tilde{y},\tilde{\lambda})$ sei eine gestörte Lösung, die die Differential-Algebraische-Gleichung nur mit einer Abweichung $(\delta_1,\delta_2) = \delta = \delta(t)$ erfüllt, d.h.

$$\dot{\tilde{y}} = f(\tilde{y},\tilde{\lambda}) + \delta_1 \tag{1.12}$$
$$0 = g(\tilde{y}) + \delta_2. \tag{1.13}$$

Eine Abschätzung für die Differenz $\| (y,\lambda) - (\tilde{y},\tilde{\lambda}) \|$ zwischen der echten und der gestörten Lösung lässt sich mit Hilfe des Satzes von der Impliziten Funktion und dem Lemma von Gronwall gewinnen. Differentiation von (1.13) ergibt

$$\begin{aligned}0 &= g_y(\tilde{y})\dot{\tilde{y}} + \dot{\delta}_2\\ &= g_y(\tilde{y})f(\tilde{y},\tilde{\lambda}) + g_y(\tilde{y})\delta_1 + \dot{\delta}_2\\ &=: F(\tilde{y},\delta_1,\dot{\delta}_2,\tilde{\lambda}).\end{aligned}$$

Dies kann als implizite Gleichung für $\tilde{\lambda}$ aufgefasst werden. Da nach Voraussetzung $\frac{\partial F}{\partial \tilde{\lambda}} = g_{\tilde{y}} f_{\tilde{\lambda}}(\tilde{y},\tilde{\lambda})$ auf einer Umgebung um (y,λ) invertierbar ist, ist der Satz von der impliziten Funktion (etwa Theorem 8.2 aus [AE08]) für die Variable $\tilde{\lambda}$ anwendbar. Dieser liefert, außer Aussagen über die lokale Existenz und Eindeutigkeit einer Lösung $\tilde{\lambda}$ als Funktion von $\tilde{y},\delta_1,\delta_2$, also $\tilde{\lambda} := h(\tilde{y},\delta_1,\dot{\delta}_2)$, auch noch Aussagen über die Ableitung von $\tilde{\lambda}$: Sei $z \in \left\{y,\delta_1,\dot{\delta}_2\right\}$, dann gilt

$$\begin{aligned}\frac{\partial h}{\partial z}(\tilde{y},\delta_1,\dot{\delta}_2) &= -\left(\frac{\partial F}{\partial \lambda}(\tilde{y},\delta_1,\dot{\delta}_2,\tilde{\lambda})\right)^{-1}\frac{\partial F}{\partial z}(\tilde{y},\delta_1,\dot{\delta}_2,\tilde{\lambda})\\ &= -\left(g_y f_\lambda(\tilde{y},\tilde{\lambda})\right)^{-1}\frac{\partial F}{\partial z}(\tilde{y},\delta_1,\dot{\delta}_2,\tilde{\lambda}).\end{aligned}$$

Aufgrund der Modellannahmen ergibt sich hieraus für $\tilde{\lambda}$ eine Lipschitz-Stetigkeit mit einer Konstante L.
Analoge Umformung des ungestörten Problems (1.10) liefert

$$\begin{aligned} 0 &= F(y, 0, 0, \lambda) \\ \lambda &= h(y, 0, 0). \end{aligned}$$

Zusammen ergibt sich als Fehlerabschätzung für λ:

$$\begin{aligned} \left\| \tilde{\lambda} - \lambda \right\| &= \left\| h(\tilde{y}, \delta_1, \dot{\delta}_2) - h(y, 0, 0) \right\| \\ &\leq L \left\| (\tilde{y} - y, \delta_1 - 0, \dot{\delta}_2 - 0) \right\| \\ &\leq L \left(\|\tilde{y} - y\| + \|\delta_1\| + \|\dot{\delta_2}\| \right). \end{aligned}$$

Setzt man nun die Zeitabhängigkeit ein und ersetzt die Normen auf der rechten Seite durch Maximumsnorm bezüglich des Zeitintervalls $[0, \bar{t}]$, folgt schließlich auf diesem Zeitintervall

$$\left\| \tilde{\lambda}(t) - \lambda(t) \right\| \leq L \left(\|\tilde{y}(t) - y(t)\| + \max_{0 \leq \xi \leq \bar{t}} \|\delta_1(\xi)\| + \max_{0 \leq \xi \leq \bar{t}} \| \dot{\delta}_2(\xi) \| \right). \tag{1.14}$$

Dies ist beinahe die Form der Definition des Störungsindex, es stört nur noch die Abhängigkeit von $e(t) := \tilde{y}(t) - y(t)$. Um eine Abschätzung für die y-Komponente zu erhalten, betrachte nun diese Fehlerfunktion $e(t)$. Aus der ersten Zeile von (1.10) und (1.12) ergibt sich:

$$\dot{e} = \dot{\tilde{y}} - \dot{y} = f(\tilde{y}, \tilde{\lambda}) + \delta_1 - f(y, \lambda).$$

Formelle Integration über t und Ausnutzen der Lipschitz-Bedingung an f mit einer Konstante C, (im Folgenden seien allgemein $C_i, i \in \mathbb{N}$ positive Konstanten) sowie Abschätzung (1.14) für $\| \tilde{\lambda} - \lambda \|$ liefert

$$e(t) = e(0) + \int_0^t f(\tilde{y}, \tilde{\lambda}) + \delta_1(s) - f(y, \lambda) ds$$

und damit

$$\begin{aligned} \|e(t)\| &\leq \|e(0)\| + C \left(\int_0^t \|\tilde{y} - y\| + \| \tilde{\lambda} - \lambda \| \, ds \right) + \left\| \int_0^t \delta_1(s) ds \right\| \\ &\overset{\text{s.o.}}{\leq} \|e(0)\| + C \left(\int_0^t \|\tilde{y} - y\| + L \left(\|\tilde{y} - y\| + \|\delta_1(s)\| + \| \dot{\delta}_2(s) \| \right) ds \right) + \left\| \int_0^t \delta_1(s) ds \right\| \\ &\leq \|e(0)\| + C_1 \int_0^t \|e(s)\| \, ds + C_2 \int_0^t \|\delta_1(s)\| + \| \dot{\delta}_2(s) \| \, ds \\ &=: \alpha(t) + C_1 \int_0^t \|e(s)\| \, ds. \end{aligned}$$

Dies ist eine Abschätzung für $\|e(t)\|$ mit einer nach Definition monoton wachsenden Funktion $\alpha(t)$. Damit ist der Spezialfall der Gronwallschen Ungleichung mit monoton steigender Funktion $\alpha(t)$ anwendbar (vgl. [Tes12], Lemma 2.7). Hieraus folgt auf dem beschränkten Zeitintervall $[0, \bar{t}]$, dass

$$\begin{aligned} \|e(t)\| &\leq C_2 \underbrace{e^t}_{\leq e^{\bar{t}}} \left(\|e(0)\| + \int_0^t \|\delta_1(s)\| + \| \dot{\delta}_2(s) \| \, ds \right) \\ &\leq C_3 \left(\|e(0)\| + \int_0^t \|\delta_1(s)\| + \| \dot{\delta}_2(s) \| \, ds \right) \\ &\leq C_3 \left(\|e(0)\| + \int_0^{\bar{t}} \max_{\xi \in [0,\bar{t}]} \|\delta_1(\xi)\| + \max_{\xi \in [0,\bar{t}]} \| \dot{\delta}_2(\xi) \| \, ds \right) \\ &\leq C_4 \left(\|e(0)\| + \max_{\xi \in [0,\bar{t}]} \|\delta_1(\xi)\| + \max_{\xi \in [0,\bar{t}]} \| \dot{\delta}_2(\xi) \| \right). \end{aligned}$$

Wegen $e(t) = \tilde{y}(t) - y(t)$ entspricht dies der Definition für Störungsindex 2 für die y-Komponente der Differential-Algebraischen-Gleichung, gleichzeitig ergibt sich durch Einsetzen dieser Abschätzung in (1.14)

$$\left\| \tilde{\lambda}(t) - \lambda(t) \right\| \leq C_5 \left(\|\tilde{y}(0) - y(0)\| + \max_{0 \leq \xi \leq t} \|\delta_1(\xi)\| + \max_{0 \leq \xi \leq t} \| \dot{\delta}_2(\xi) \| \right),$$

also ebenfalls Störungsindex 2 der λ-Komponente und damit insgesamt Störungsindex 2. Somit stimmen Differentiation- und Störungsindex des Modellproblems überein und im Folgenden wird nur noch der Begriff *Index* benutzt.

Kapitel 2

Runge-Kutta-Verfahren für Differential-Algebraische-Gleichungen

Um Differential-Algebraische-Gleichungen zu lösen, gibt es verschiedenste Ansätze, die auch vom Index des zu lösenden Problems abhängen. Die in Anbetracht des Differentiationsindex naheliegendste Variante ist es, das Problem gemäß der Definition des Index durch Differentiation auf explizite Form zu bringen. Gehen wir wieder von einem System der Form (1.3) aus, geht durch die Differentiation der zweiten Zeile allerdings die Bedingung, dass die Lösung auf der durch g definierten Mannigfaltigkeit liegen muss, verloren bzw. wird durch abgeleitete Formen ersetzt. Auch wenn dies bei exakter, analytischer Lösung keinen Unterschied macht, kann die numerische Lösung die Mannigfaltigkeit aufgrund von Verfahrens- oder Rundungs-Fehlern verlassen (siehe Abb. 4.5, im Vergleich zur Lösung der Differential-Algebraischen-Gleichung in Abb. 4.4). Um dies zu verhindern, müssen die Nebenbedingungen in den Lösungsverfahren berücksichtigt werden. Runge-Kutta-Verfahren lassen sich für implizite Differentialgleichungen und damit insbesondere Differential-Algebraische-Gleichungen modifizieren. Diese Möglichkeit wird in diesem Kapitel untersucht.

Es werden Ordnungsbedingungen für auf Differential-Algebraische-Gleichungen angewendete Runge-Kutta-Verfahren hergeleitet. Damit wird gezeigt, dass es bei gewissen s-stufigen Runge-Kutta-Verfahren möglich ist, bei Problemen vom Index 2 eine hohe $(2s-1)$ Konvergenzordnung zumindest bezüglich der Variablen vom Index 0 zu erhalten. Bei den Variablenanteilen von Index 2 wird immerhin Konvergenzordnung von s erreicht. Insbesondere bei zu Kollokationsverfahren äquivalenten Runge-Kutta-Verfahren wie etwa vom Typ RadauIIA ist dies relativ einfach nachzuweisen, dies wird in Korollar 2 getan.

2.1 Formelle Anwendung von Runge-Kutta-Verfahren auf Differential-Algebraische-Gleichungen

Der prinzipielle Ansatz, durch den man Runge-Kutta-Verfahren auf Differential-Algebraische-Gleichungen anwenden kann, ist eine Modifikation, um statt expliziter Differentialgleichungen der Form (1.1) auch implizite Differentialgleichungen der Form

$$M\dot{y} = f(y)$$

lösen zu können, wobei M eine konstante, quadratische Matrix ist. Für unser Modellproblem (1.10) ergibt sich somit eine singuläre Matrix

$$M = \mathrm{diag}(\underbrace{1,\ldots,1}_{\dim y},\underbrace{0,\ldots,0}_{\dim \lambda}).$$

Grundsätzlich wird bei einem impliziten Runge-Kutta-Verfahren in jedem Schritt ein nichtlineares Gleichungssystem gelöst. Durch das Einführen einer (hier singulären) Matrix wird dieses nur in ein anderes nichtlineares Gleichungssystem überführt. Dessen Lösung sollte dann Lösung unserer Differential-Algebraischen-Gleichung sein. Tatsächlich ist dies mit gewissen Einschränkungen an das Verfahren der Fall; die für Probleme von Index 2 relevanten Aussagen wurden aus [HLR89] und [HNWXX] entnommen.

Betrachten wir ein s-stufiges Runge-Kutta-Verfahren der allgemeinen Form

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array},$$

wobei $A \in \mathbb{R}^{s\times s}$, $b \in \mathbb{R}^s$ und $c \in \mathbb{R}^s$ die Koeffizientenmatrix und -vektoren sind. Für autonome, explizite Differentialgleichungen der Form $\dot{y} = f(y)$ errechnet sich ein Schritt des Verfahrens mit Schrittweite h wie folgt:

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i Y'_{ni} \tag{2.1}$$

$$Y'_{ni} = f(Y_{ni}), \tag{2.2}$$

wobei die inneren Stufen Y_{ni} durch

$$Y_{ni} = y_n + h\sum_{j=1}^{s} a_{ij} Y'_{nj}\ ,\ i = 1,\dots,s \tag{2.3}$$

gegeben sind. Hiermit ergibt sich insgesamt für die Y'_{ni} das i.A. nichtlineare Gleichungssystem

$$Y'_{ni} = f(y_n + h\sum_{j=1}^{s} a_{ij} Y'_{nj})\ ,\ i = 1,\dots,s$$

oder äquivalent

$$Y_{ni} = y_n + h\sum_{j=1}^{s} a_{ij} f(Y_{nj})\ ,\ i = 1,\dots,s.$$

Für unser Modellproblem

$$M\begin{pmatrix}\dot{y}\\ \dot{\lambda}\end{pmatrix} = \begin{pmatrix}\dot{y}\\ 0\end{pmatrix} = \begin{pmatrix}f(y,\lambda)\\ g(y)\end{pmatrix} \tag{2.4}$$

ergibt sich dann analog

$$\begin{pmatrix}y_{n+1}\\ \lambda_{n+1}\end{pmatrix} = \begin{pmatrix}y_n\\ \lambda_n\end{pmatrix} + h\sum_{i=1}^{s} b_i \begin{pmatrix}Y'_{ni}\\ \Lambda'_{ni}\end{pmatrix}, \tag{2.5}$$

$$M\begin{pmatrix}Y'_{ni}\\ \Lambda'_{ni}\end{pmatrix} = \begin{pmatrix}f(Y_{ni},\Lambda_{ni})\\ g(Y_{ni})\end{pmatrix},, \tag{2.6}$$

mit den inneren Stufen

$$\begin{pmatrix}Y_{ni}\\ \Lambda_{ni}\end{pmatrix} = \begin{pmatrix}y_n\\ \lambda_n\end{pmatrix} + h\sum_{j=1}^{s} a_{ij} \begin{pmatrix}Y'_{nj}\\ \Lambda'_{ni}\end{pmatrix},, \tag{2.7}$$

und damit das Gleichungssystem

$$M \begin{pmatrix} Y'_{ni} \\ \Lambda'_{ni} \end{pmatrix} = \begin{pmatrix} Y'_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} f\left(y_n + h\sum_{j=1}^{s} a_{ij} Y'_{nj}, \lambda_n + h\sum_{j=1}^{s} a_{ij}\Lambda'_{nj}\right) \\ g\left(y_n + h\sum_{j=1}^{s} a_{ij} Y'_{nj}\right) \end{pmatrix}, \; i = 1, \ldots, s$$

oder äquivalent als Gleichungssystem für die inneren Stufen

$$\begin{pmatrix} Y_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} y_n + h\sum_{j=1}^{s} a_{ij} f\left(Y_{nj}, \Lambda_{nj}\right) \\ g\left(Y_{ni}\right) \end{pmatrix}, \; i = 1, \ldots, s\,, \tag{2.8}$$

wobei sich hier in der zweiten Zeile der Unterschied zwischen der Behandlung von gewöhnlichen expliziten Differentialgleichungen und Differential-Algebraischen-Gleichungen zeigt; statt Gleichungen, die die gesuchten inneren Stufen Λ_{ni} explizit enthalten, ist hier nur die implizite Bedingung $0 = g(Y_{ni})$ möglich.

Damit ist prinzipiell gezeigt, wie ein Runge-Kutta-Verfahren auf Differential-Algebraische-Gleichungen angewendet werden kann. Allerdings sind die Resultate bezüglich Existenz und Eindeutigkeit einer Lösung für jeden Verfahrensschritt und generellem Konvergenzverhalten nicht direkt übertragbar. Diese werden im folgenden Abschnitt betrachtet.

2.2 Konvergenzanalyse

Die bei Index-2-Problemen relevanten Fragen nach Existenz, Eindeutigkeit und Konvergenz der Runge-Kutta-Lösungen werden in [HLR89] ausführlich behandelt. Im Folgenden werden die wichtigsten Lemmata gezeigt. Zunächst ergibt sich für die Frage nach Existenz und Eindeutigkeit:

Theorem 1 *Theorem 4.1 aus [HLR89]*
Die Koeffizientenmatrix A des Runge-Kutta-Verfahrens sei invertierbar und das Modellproblem (1.10) erfülle insbesondere Modellannahme 3. Seien weiterhin (y_n, λ_n) gegeben, die

$$g(y_n) = \mathcal{O}(h^2), \quad g_y(y_n) f(y_n, \lambda_n) = \mathcal{O}(h) \tag{2.9}$$

erfüllen. Dann besitzt das nichtlineare Gleichungssystem (NLGS) (2.8) für genügend kleines h eine lokal eindeutige Lösung (Y_n, Λ_n), für die gilt

$$Y_{ni} - y_n = \mathcal{O}(h), \quad \Lambda_{ni} - \lambda_n = \mathcal{O}(h). \tag{2.10}$$

Beweis: Der Beweis ist zweigeteilt, zunächst zeigt man Existenz einer Lösung für hinreichend kleines h, sowie das Erfüllen von (2.10), anschließend die Eindeutigkeit. Allgemein wird nur ein einziger Verfahrensschritt betrachtet, dies erlaubt, an einigen Stellen auf den Index n zu verzichten.
Betrachten wir zunächst die Homotopie zwischen einem Schritt mit Werten (y_n, λ_n) und dem nächsten Schritt.

$$\begin{pmatrix} Y_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} y_n + h\sum_{j=1}^{s} a_{ij} f\left(Y_{nj}, \Lambda_{nj}\right) \\ g\left(Y_{ni}\right) \end{pmatrix} + (\tau - 1) \begin{pmatrix} h\sum_{j=1}^{s} a_{ij} f\left(y_n, \lambda_n\right) \\ g\left(y_n\right) \end{pmatrix} \tag{2.11}$$

Für $\tau = 1$ liegt gerade wieder das NLGS (2.8) vor, für $\tau = 0$ ist $(Y_{ni}, \Lambda_{ni}) = (y_n, \lambda_n)$ eine Lösung, da nach Umformung der Homotopie zu

$$\begin{pmatrix} Y_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} y_n + h\sum_{j=1}^{s} a_{ij} \left(f\left(Y_{nj}, \Lambda_{nj}\right) - f\left(y_n, \lambda_n\right)\right) \\ g\left(Y_{ni}\right) - g\left(y_n\right) \end{pmatrix}$$

die zweite Zeile damit offensichtlich erfüllt ist, sowie die Summen in der ersten Zeile Nullsummen sind und nur noch $Y_{ni} = y_n$ übrig bleibt. Fasst man Y_{ni} und Λ_{ni} als Lösung von (2.11) auf, d.h. als Funktionen vom Parameter τ, kann man (2.11) nach τ ableiten und erhält

$$\begin{pmatrix} \dot{Y}_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} h \sum_{j=1}^{s} a_{ij} \left(f_y \left(Y_{nj}, \Lambda_{nj} \right) \dot{Y}_{nj} + f_\lambda \left(Y_{nj}, \Lambda_{nj} \right) \dot{\Lambda}_{nj} \right) \\ g_y \left(Y_{ni} \right) \dot{Y}_{ni} \end{pmatrix} + \begin{pmatrix} h \sum_{j=1}^{s} a_{ij} f \left(y_n, \lambda_n \right) \\ g \left(y_n \right) \end{pmatrix}. \tag{2.12}$$

Hierbei und im Kontext dieses Beweises stehe $\dot{Y}$ für die Ableitung $\frac{d}{d\tau} Y$. Nach Zusammenfassung, Division der zweiten Zeile durch h und Ersetzen von $\dot{Y}_{ni}$ in der zweiten Zeile durch die erste, ergibt sich

$$\begin{pmatrix} \dot{Y}_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} h \sum_{j=1}^{s} a_{ij} \left(f_y \left(Y_{nj}, \Lambda_{nj} \right) \dot{Y}_{nj} + f_\lambda \left(Y_{nj}, \Lambda_{nj} \right) \dot{\Lambda}_{nj} + f \left(y_n, \lambda_n \right) \right) \\ g_y \left(Y_{ni} \right) \sum_{j=1}^{s} a_{ij} \left(f_y \left(Y_{nj}, \Lambda_{nj} \right) \dot{Y}_{nj} + f_\lambda \left(Y_{nj}, \Lambda_{nj} \right) \dot{\Lambda}_{nj} + f \left(y_n, \lambda_n \right) \right) + \frac{g(y_n)}{h} \end{pmatrix}$$

oder, in Matrixschreibeweise als Zusammenfassung aller $i = 1, \ldots, s$ Gleichungen:

$$\begin{pmatrix} \mathbb{1} - h(A \otimes \mathbb{1}) F_y & -h(A \otimes \mathbb{1}) F_\lambda \\ -G_y (A \otimes \mathbb{1}) F_y & -G_y (A \otimes \mathbb{1}) F_\lambda \end{pmatrix} \begin{pmatrix} \dot{Y} \\ \dot{\Lambda} \end{pmatrix} = \begin{pmatrix} h(A\mathbb{1} \otimes f(y_n, \lambda_n)) \\ G_y (A\mathbb{1} \otimes f(y_n, \lambda_n)) + \frac{1}{h} \mathbb{1} \otimes g(y_n) \end{pmatrix}. \tag{2.13}$$

Hierbei wurde die Notation $Y = Y_n = (Y_{n1}, \ldots, Y_{ns})^T$, $\Lambda = \Lambda_n = (\Lambda_{n1}, \ldots, \Lambda_{ns})^T$, $G_y = \text{blockdiag}(g_y(Y_{n1}), \ldots, g_y(Y_{ns}))$ und analog dazu F_y und F_λ verwendet, weiterhin bezeichne $\mathbb{1}$ die n-dimensionale Einheitsmatrix und $\otimes$ das Kroneckerprodukt.
Da mit Modellannahme 5 f, g und deren Ableitungen beschränkt sind, bzw. Lipschitz-Stetigkeit gilt, folgt

$$g_y(Y_{ni}) a_{ij} f_\lambda(Y_{nj}, \Lambda_{nj}) = a_{ij} g_y(Y_{nj}) f_\lambda(Y_{nj}, \Lambda_{nj}) + \mathcal{O}(\|Y_{ni} - Y_{nj}\|).$$

Mit einer Konstanten d, sodass $\|Y_{ni} - Y_{nj}\| \leq d$ gilt, lässt sich der rechte, untere Block der Matrix in (2.13) schreiben als

$$-G_y (A \otimes \mathbb{1}) F_\lambda = -(A \otimes \mathbb{1}) G_y F_\lambda + \mathcal{O}(d),$$

was nach Voraussetzung 3 für genügend kleines d, aber unabhängig von h, invertierbar mit beschränkter Inversen ist, denn die Modellannahmen gelten alle „in einer Umgebung um die Lösung“, diese werde mit U bezeichnet. Für hinreichend kleines h und d und vorausgesetzt, dass alle Y_{ni} und Λ_{ni} in U liegen, ist die Matrix in (2.13) also näherungsweise eine untere Block-Dreiecksmatrix der Form

$$\begin{pmatrix} \mathbb{1} + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}.$$

Als Folge der Modellannahmen für f und g ist diese stetig. Insgesamt ist diese Matrix also näherungsweise eine untere Block-Dreiecksmatrix mit oberem rechten Block $\mathcal{O}(h)$ und invertierbaren Diagonalelementen (der erste Diagonalblock hat die Gestalt $\mathbb{1} + \mathcal{O}(h)$ und ist damit für hinreichend kleines h auch invertierbar mit beschränkter Inversen $\mathbb{1} + \mathcal{O}(h)$). Damit hat die Matrix eine Inverse der Form

$$\begin{pmatrix} \mathbb{1} + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}. \tag{2.14}$$

Da nach Modellannahmen die Funktionen f, g und die beteiligten Ableitungen stetig sind, ist diese Inverse (für hinreichend kleines h) ebenfalls stetig nach Korollar 4.50 in [Pla10] (Stetigkeit der Matrixinversion). Setzt man dies in (2.13) ein, ergibt sich

$$\begin{aligned} \begin{pmatrix} \dot{Y} \\ \dot{\Lambda} \end{pmatrix} &= \begin{pmatrix} \mathbb{1} + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix} \begin{pmatrix} h(A\mathbb{1} \otimes f(y_n, \lambda_n)) \\ G_y (A\mathbb{1} \otimes f(y_n, \lambda_n)) + \frac{1}{h} \mathbb{1} \otimes g(y_n) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{O}(h) \\ \mathcal{O}\left(G_y (A\mathbb{1} \otimes f(y_n, \lambda_n)) \right) + \mathcal{O}(h) \end{pmatrix}, \end{aligned} \tag{2.15}$$

wobei genutzt wurde, dass nach Voraussetzung $g(y_n) = \mathcal{O}(h^2)$, was den Faktor $\frac{1}{h}$ eliminiert. Die Differentialgleichung (2.13) mit den Anfangswerten $Y(0) = (y_n, \dots, y_n)$ und $\Lambda(0) = (\lambda_n, \dots, \lambda_n)$ kann also auf eine explizite Differentialgleichung umgeformt werden, die aufgrund der Stetigkeit nach dem Existenzsatz von Peano eine Lösung in U auf einem Intervall $\tau \in [0, \tau^*)$ mit $\tau^* > 0$ besitzt, die so lange fortgesetzt werden kann, wie Y und Λ in U bleiben. Außerdem gilt wegen

$$\begin{pmatrix} Y(\tau) \\ \Lambda(\tau) \end{pmatrix} = \begin{pmatrix} \mathbb{1} y_n \\ \mathbb{1} \lambda_n \end{pmatrix} + \int_0^\tau \begin{pmatrix} \dot{Y}(t) \\ \dot{\Lambda}(t) \end{pmatrix} dt,$$

dass $Y_{ni}(\tau) = y_n + \mathcal{O}(\tau h)$. Für h hinreichend klein ist die Lösung für Y_{ni} also auch für $\tau = 1$ in U, damit ist die Lösung bis $\tau = 1$ erweiterbar. Hiermit ist die Existenz einer Lösung der Homotopie (2.11) für $\tau = 1$, also gerade im Fall des ursprünglichen NLGS (2.8), zunächst für die Y_{ni} bewiesen. Gleichzeitig ist für die Y_{ni} (2.10) gezeigt. Damit folgt

$$\begin{aligned} G_y(A\mathbb{1} \otimes f(y_n, \lambda_n)) &= \mathbb{1} \otimes g_y(y_n)(A\mathbb{1} \otimes f(y_n, \lambda_n)) + \mathcal{O}\left(\max_i \underbrace{\|y_n - Y_{ni}\|}_{=\mathcal{O}(h) \text{ s.o.}} \right) \\ &= \mathcal{O}\underbrace{(g_y(y_n) f(y_n, \lambda_n))}_{=\mathcal{O}(h) \text{ nach (2.9)}} + \mathcal{O}(h) \\ &= \mathcal{O}(h). \end{aligned}$$

Einsetzen in (2.15) ergibt, dass $\dot{\Lambda} = \mathcal{O}(h)$. Hieraus folgt die Behauptung des Satzes ganz analog zu oben auch für die Λ_{ni}.
Um die Eindeutigkeit zu zeigen, nehmen wir zunächst an, es gebe zusätzlich zu den Lösungen (Y_{ni}, Λ_{ni}) von (2.8) noch die Lösungen $(\hat{Y}_{ni}, \hat{\Lambda}_{ni})$. Seien die Differenzen dieser Lösungen bezeichnet durch $\Delta Y_i := \hat{Y}_{ni} - Y_{ni}$ und $\Delta\Lambda_i := \hat{\Lambda}_{ni} - \Lambda_{ni}$. Einsetzen der ersten Zeile von (2.8) in ΔY_i ergibt (y_n hebt sich weg)

$$\Delta Y_i = h \sum_{j=1}^{s} a_{ij} \left(f\left(Y_{nj}, \Lambda_{nj}\right) - f\left(\hat{Y}_{nj}, \hat{\Lambda}_{nj}\right) \right)$$

und mit Abschätzen der Differenzen von f mit der in den Modellannahmen geforderten Lipschitz-Bedingung, sowie von A durch etwa Maximumsnorm, ergibt sich hieraus mit einer Konstante $C < \infty$:

$$\|\Delta Y\| \leq hC \left(\|\Delta\Lambda\| + \|\Delta Y\| \right), \tag{2.16}$$

wobei hier $\|\Delta Y\| = \max_i \|\Delta Y_i\|$, analog für $\Delta\Lambda$. Durch Taylorentwicklung um y_n erhält man

$$\begin{aligned} 0 &= g(\hat{Y}_{ni}) - g(Y_{ni}) \\ &= g(y_n) + g_y(y_n)\left(\hat{Y}_{ni} - y_n\right) - \left(g(y_n) + g_y(y_n)\left(Y_{ni} - y_n\right)\right) + \underbrace{\mathcal{O}(g_{yy}(\xi))}_{\text{beschr. in } U} \underbrace{\|\Delta Y_i\|^2}_{\mathcal{O}(h^2)} \\ &= g_y(y_n)\left(\hat{Y}_{ni} - Y_{ni}\right) + \mathcal{O}(h^2) \\ &\overset{(2.8)}{=} g_y(y_n) h \sum_{j=1}^{s} a_{ij} \underbrace{\left(f\left(\hat{Y}_{nj}, \hat{\Lambda}_{nj}\right) - f\left(Y_{nj}, \Lambda_{nj}\right)\right)}_{\text{Taylor: } f_y(y_n,\lambda_n)\Delta Y_j + f_\lambda \Delta\Lambda_j + \mathcal{O}(\underbrace{\Delta\Lambda_i^2}_{\mathcal{O}(h^2)} + \Delta Y_i^2)} + \mathcal{O}(h^2) \\ &= h \sum_{j=1}^{s} a_{ij} g_y(y_n) \left(f_\lambda(y_n, \lambda_n)\Delta\Lambda_j + f_y(y_n, \lambda_n)\Delta Y_j\right) + \mathcal{O}(h^2) \\ &= h(\mathcal{O}(\Delta\Lambda) + \mathcal{O}(\Delta Y)) + \mathcal{O}(h^2). \end{aligned}$$

Damit haben $\Delta\Lambda$ und ΔY die gleichen Größenordnungen, also $\|\Delta\Lambda\| \leq D\,\|\Delta Y\|$ mit einer von h unabhängigen Konstanten $D > 0$. Einsetzen in (2.16) liefert

$$\|\Delta Y\| \leq hC\left(\|\Delta\Lambda\| + \|\Delta Y\|\right) \leq hC(1+D)\,\|\Delta Y\|\,.$$

Dies ist für hinreichend kleines h ein Widerspruch, falls $\Delta Y \neq 0$, damit gilt wegen $\|\Delta\Lambda\| \leq D\,\|\Delta Y\|$ auch $\Delta\Lambda = 0$, somit stimmen die beiden Lösungen (Y, Λ) und $(\hat{Y}, \hat{\Lambda})$ überein. □

Dass hierbei nicht gefordert wird, dass (y_n, λ_n) die algebraische Nebenbedingung, also die zweite Zeile von (1.10) und deren Ableitung, exakt erfüllt, ist dem in der Praxis numerischen Lösen von (2.8) geschuldet. Für gewöhnlich wird diese Gleichung hierbei nicht exakt, sondern mittels Newtonverfahren nur näherungsweise gelöst, weshalb eine gewisse Toleranz bezüglich dem Erfüllen der algebraischen Bedingung von (1.10) für Existenz und Eindeutigkeit der Lösung existenziell für das numerische Durchführen des Verfahrens ist. Nichtsdestotrotz ergeben sich beim Durchführen des Newtonverfahrens bei Problemen vom Index 2 oder höher noch weitere Probleme, die im folgenden Abschnitt behandelt werden.
Insbesondere zu bemerken ist, dass dieses Lemma nicht auf explizite Runge-Kutta-Verfahren anwendbar ist, da deren Koeffizientenmatrix eine strikt untere Dreiecksmatrix und somit nicht invertierbar ist.

Ein weiteres Lemma gibt eine Abschätzung bei Störung des Systems an, also wenn (2.8) nicht exakt, sondern (etwa durch Rundungs- oder Verfahrensfehler) vergleichbar zur Definition des Störungsindex eine Abweichung besitzt. Dieses Lemma wird im Beweis für das darauf folgende Theorem, Ordnungsbedingungen für den lokalen Fehler, benötigt.

Lemma 1 *Theorem 4.2 aus [HLR89]*
Seien die Y_{ni} und Λ_{ni} aus (2.8) gegeben, außerdem gelten alle Annahmen aus Theorem 1. Zusätzlich gebe es gestörte Werte $\hat{Y}_{ni}$ und $\hat{\Lambda}_{ni}$, die (2.8) nur mit Abweichung erfüllen, d.h.

$$\begin{pmatrix} \hat{Y}_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{y}_n + h\sum_{j=1}^{s} a_{ij} f\left(\hat{Y}_{nj}, \hat{\Lambda}_{nj}\right) \\ g\left(\hat{Y}_{ni}\right) \end{pmatrix} + \begin{pmatrix} h\delta_i \\ \theta_i \end{pmatrix},$$

wobei für den gestörten Anfangswert $\hat{y}_n$ gelte, dass $\hat{y}_n - y_n = \mathcal{O}(h^2)$, außerdem seien die Störungen gering, d.h. $\|\delta\| := \max_i \|\delta_i\| = \mathcal{O}(h)$ und $\|\theta\| := \max_i \|\theta_i\| = \mathcal{O}(h^2)$.
Dann gelten für hinreichend kleines h die Abschätzungen

$$\left\|\hat{Y}_{ni} - Y_{ni}\right\| \leq C\left(\|\hat{y}_n - y_n\| + h\,\|\delta\| + \|\theta\|\right), \tag{2.17}$$

$$\left\|\hat{\Lambda}_{ni} - \Lambda_{ni}\right\| \leq \frac{C}{h}\left(g_y(y_n)(\hat{y}_n - y_n) + h\,\|\hat{y}_n - y_n\| + h\,\|\delta\| + \|\theta\|\right). \tag{2.18}$$

Beweis: Betrachten wir wieder eine Homotopie, diesmal zwischen exakter Lösung und gestörter Lösung, also

$$\begin{pmatrix} Y_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} y_n + h\sum_{j=1}^{s} a_{ij} f\left(Y_{nj}, \Lambda_{nj}\right) \\ g\left(Y_{ni}\right) \end{pmatrix} + (1-\tau)\begin{pmatrix} \hat{y}_n - y_n + h\delta_i \\ \theta_i \end{pmatrix}.$$

Hier liegt für $\tau = 1$ das ungestörte System (2.8) und für $\tau = 0$ das gestörte System vor. Fasst man wiederum die Lösungen (Y_{ni}, Λ_{ni}) aus der Homotopie als Funktionen in Abhängigkeit von τ auf und leitet danach ab, ergibt sich

$$\begin{pmatrix} \dot{Y}_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} h\sum_{j=1}^{s} a_{ij}\left(f_y\left(Y_{nj}, \Lambda_{nj}\right)\dot{Y}_{nj} + f_\lambda\left(Y_{nj}, \Lambda_{nj}\right)\dot{\Lambda}_{nj}\right) \\ g_y\left(Y_{ni}\right)\dot{Y}_{ni} \end{pmatrix} - \begin{pmatrix} \hat{y}_n - y_n + h\delta_i \\ \theta_i \end{pmatrix}.$$

Wiederum sei im Kontext dieses Beweises $\dot{Y}$ definiert durch $\frac{dY}{d\tau}$. Ersetzen von $\dot{Y}_{ni}$ in der zweiten Zeile durch die erste und Teilen durch h in der zweiten Zeile ergibt

$$\begin{pmatrix} \dot{Y}_{ni} \\ 0 \end{pmatrix} = \begin{pmatrix} h \sum_{j=1}^{s} a_{ij} \left(f_y \left(Y_{nj}, \Lambda_{nj} \right) \dot{Y}_{nj} + f_\lambda \left(Y_{nj}, \Lambda_{nj} \right) \dot{\Lambda}_{nj} \right) - \hat{y}_n + y_n - h\delta_i \\ g_y \left(Y_{ni} \right) \left(\sum_{j=1}^{s} a_{ij} (f_y \left(Y_{nj}, \Lambda_{nj} \right) \dot{Y}_{nj} + f_\lambda \left(Y_{nj}, \Lambda_{nj} \right) \dot{\Lambda}_{nj}) - \frac{\hat{y}_n - y_n}{h} - \delta_i \right) - \frac{\theta_i}{h} \end{pmatrix}.$$

Wiederum Darstellung als Matrix-Vektor Produkt, in der Notation analog zu Theorem 1 ergibt:

$$\begin{pmatrix} \mathbb{1} - h(A \otimes \mathbb{1}) F_y & -h(A \otimes \mathbb{1}) F_\lambda \\ -G_y(A \otimes \mathbb{1}) F_y & -G_y(A \otimes \mathbb{1}) F_\lambda \end{pmatrix} \begin{pmatrix} \dot{Y} \\ \dot{\Lambda} \end{pmatrix} = - \begin{pmatrix} \mathbb{1} \otimes (\hat{y}_n - y_n) + h\delta \\ \frac{1}{h} \left(G_y(\mathbb{1} \otimes (\hat{y}_n - y_n) + h\delta) + \theta \right) \end{pmatrix}.$$

Die Matrix auf der linken Seite ist exakt wieder die Matrix von (2.13) aus dem Beweis von Theorem 1, also invertierbar mit beschränkter Inverser der Form (2.14). Damit kann man die aus der Homotopie entstandene Differentialgleichung zu einer expliziten Differentialgleichung aufstellen:

$$\begin{aligned} \begin{pmatrix} \dot{Y} \\ \dot{\Lambda} \end{pmatrix} &= - \begin{pmatrix} \mathbb{1} + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix} \begin{pmatrix} \mathbb{1} \otimes (\hat{y}_n - y_n) + h\delta \\ \frac{1}{h} \left(G_y(\mathbb{1} \otimes (\hat{y}_n - y_n) + h\delta) + \theta \right) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{O}(h\delta) + \mathcal{O}(\hat{y}_n - y_n) + \mathcal{O}(\theta) \\ \mathcal{O}(\hat{y}_n - y_n) + \mathcal{O}(h\delta) + \mathcal{O}\left(\frac{1}{h} G_y \left(\mathbb{1} \otimes (\hat{y}_n - y_n) \right) \right) + \mathcal{O}\left(\frac{\theta}{h} \right) \end{pmatrix} \end{aligned}$$

Wiederum folgt aus dem Satz von Peano die Existenz einer Lösung und über

$$\begin{pmatrix} Y \\ \Lambda \end{pmatrix} = \begin{pmatrix} \hat{Y} \\ \hat{\Lambda} \end{pmatrix} + \int_{\tau=0}^{1} \begin{pmatrix} \dot{Y} \\ \dot{\Lambda} \end{pmatrix} d\tau$$

und die soeben hergeleiteten Größenordnungen der Ableitungen von Y und Λ ergeben sich direkt die geforderten Abschätzungen (2.17) und (2.18). □

Der nächste Schritt ist das Abschätzen des lokalen Verfahrensfehlers, hierfür ergeben sich Ordnungsbedingungen an die Koeffizienten des Runge-Kutta-Verfahrens, teilweise analog zu jenen für explizite Differentialgleichungen, jedoch nicht äquivalent.

Theorem 2 *Theorem 4.3 aus [HLR89]*
Das Runge-Kutta-Verfahren erfülle die Voraussetzungen des letzten Theorems und zusätzlich mit $p, q \in \mathbb{N}, p > q \geq 1$ *die Bedingungen*

$$B(p) := \left(\forall\, k = 1, \ldots, p \quad \textit{gilt} \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k} \right)$$

$$C(q) := \left(\forall\, k = 1, \ldots, q,\ \forall\, i = 1, \ldots, s \quad \textit{gilt} \quad \sum_{j=i}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \right).$$

Dann gilt für die von der Schrittweite h *abhängigen lokalen Fehler* $\delta y_h, \delta \lambda_h$ *:*

$$\begin{aligned} \delta y_h(t) &= \mathcal{O}(h^{q+1}), \quad P(t) \delta y_h(t) = \mathcal{O}(h^{q+2}) \\ \delta \lambda_h(t) &= \mathcal{O}(h^q), \end{aligned}$$

wobei $P(t) = \mathbb{1} - \left(f_\lambda (g_y f_\lambda)^{-1} g_y \right) (y(t), \lambda(t))$.

Die Bedingungen B und C sind nichts weiter, als Ordnungsbedingungen für die Quadraturformeln für die inneren und äußeren Schritte, denn für Monome $g = g(x)$ vom Grad $k - 1$ muss gelten:

$$\int_0^1 g(x) dx = \int_0^1 x^{k-1} dx = \frac{1}{k} \overset{!}{=} \sum_{i=1}^{s} b_i g(c_i) = \sum_{i=1}^{s} b_i c_i^{k-1},$$

womit $B(k)$ erfüllt ist, soweit das Runge-Kutta-Verfahren entsprechende Ordnung besitzt. Damit ist B die Ordnungsbedingung für die Quadraturformel des äußeren Schrittes mit Knoten c_i und Gewichten b_i über das Intervall $[0, 1]$. Außerdem muss gelten

$$\int_0^{c_i} g(x)dx = \int_0^{c_i} x^{k-1}dx = \frac{{c_i}^k}{k} \stackrel{!}{=} \sum_{j=1}^{s} a_{ij} g(c_i) = \sum_{j=1}^{s} a_{ij} c_i^{k-1}.$$

Damit ist analog C die Ordnungsbedingung für die Quadraturformel der inneren Schritte mit Knoten c_i und Gewichten a_{ij} über das Intervall $[0, c_i]$. Dies kann man nutzen, um Theorem 2 zu beweisen:

Beweis: Gegeben seien die Werte $\hat{Y}_{ni} := y(t_n + c_i h)$ und $\hat{\Lambda}_{ni} := \lambda(t_n + c_i h)$ bei einem Schritt des Verfahrens mit exaktem Anfangswert $y_n = y(t_n)$, $\lambda_n = \lambda(t_n)$. Durch Taylorentwicklung der Lösungen $y(t)$ und $\lambda(t)$ (und den Modellannahmen, d.h. insbesondere hinreichende Differenzierbarkeit der beteiligten Funktionen) gilt außerdem

$$y(t_n + t) = y_n + \sum_{i=1}^{q+1} \frac{y^{(i)}(t_n)}{i!} ({}^i + \mathcal{O}(t^{q+1}),$$

$$\lambda(t_n + t) = \lambda_n + \sum_{i=1}^{q+1} \frac{\lambda^{(i)}(t_n)}{i!} t^i + \mathcal{O}(t^{q+1}).$$

Da hier exakte Werte von y verwendet werden, kann man außerdem die Integraldarstellung nutzen:

$$\begin{aligned}
\hat{Y}_{ni} &= y_n + \int_0^{c_i h} y'(t_n + x)dx \\
&= y_n + h \int_0^{c_i} y'(t_n + xh)dx \\
&\stackrel{\text{Taylor}}{=} y_n + h \int_0^{c_i} \left(\sum_{l=1}^{q+1} \frac{y^{(l)}(t_n)}{l!} l(xh)^{l-1} \right) + \mathcal{O}(h^{q+1})dx \\
&\stackrel{\text{Quadratur}}{=} y_n + h \left(\sum_{j=1}^{s} a_{ij} \left(\sum_{l=1}^{q} \frac{y^{(l)}(t_n)}{l!} l(c_j h)^{l-1} \right) + \int_0^{c_i} \left(\frac{y^{(q+1)}(t_n)}{q!} (xh)^q \right) dx \right) + \mathcal{O}(h^{q+2}) \\
&= y_n + h \left(\sum_{j=1}^{s} a_{ij} \underbrace{\sum_{l=1}^{q+1} \frac{y^{(l)}(t_n)}{l!} l(c_j h)^{l-1}}_{\text{Taylorentw. von } y'=f(y,\lambda)} - \sum_{j=1}^{s} a_{ij} \left(\frac{y^{(q+1)}(t_n)}{q!} (c_j h)^q \right) \right) \\
&\quad + h \int_0^{c_i} \left(\frac{y^{(q+1)}(t_n)}{q!} (xh)^q \right) dx + \mathcal{O}(h^{q+2}) \\
&= y_n + h \left(\sum_{j=1}^{s} a_{ij} f(\hat{Y}_{nj}, \hat{\Lambda}_{nj}) + \mathcal{O}(h^{q+1}) \right) - h \sum_{j=1}^{s} a_{ij} \left(\frac{y^{(q+1)}(t_n)}{(q+1)!} (q+1)(c_j h)^q \right) \\
&\quad + h^{q+1} \frac{y^{(q+1)}(t_n)}{(q+1)!} c_i^{q+1} + \mathcal{O}(h^{q+2}) \\
&= y_n + h \left(\sum_{j=1}^{s} a_{ij} f(\hat{Y}_{nj}, \hat{\Lambda}_{nj}) \right) + y^{(q+1)}(t_n) \frac{h^{q+1}}{q!} \left(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^{s} a_{ij} c_j^q \right) + \mathcal{O}(h^{q+2}) \\
&=: y_n + h \left(\sum_{j=1}^{s} a_{ij} f(\hat{Y}_{nj}, \hat{\Lambda}_{nj}) \right) + h\delta_i.
\end{aligned}$$

Damit ist δ_i aus Lemma 1 gefunden und ist von der Größenordnung $\mathcal{O}(h^q)$. Da die Werte der exakten Lösung insbesondere die Nebenbedingungen erfüllen, also $g(\hat{Y}_{ni}) = 0$ gilt, ist θ aus Lemma 1 gleich 0. Da die weiteren Werte gleich gewählt wurden, also $\hat{y}_n = y_n$, ist Lemma 1 anwendbar und liefert

$$Y_{ni} - \hat{Y}_{ni} = \mathcal{O}(h^{q+1}), \qquad \Lambda_{ni} - \hat{\Lambda}_{ni} = \mathcal{O}(h^q). \tag{2.19}$$

Wiederum mit Taylorentwicklung und der Ordnungsbedingung für die äußere Quadraturformel (Ordnung größer gleich $p+1$) ergibt sich

$$\begin{aligned} y(t_n + h) &= y_n + h \int_0^1 f(y(t_n + x), \lambda(t_n + x))dx \\ &= y_n + h \sum_{i=1}^{s} b_i f\left(y(t_n + c_i h), \lambda(t_n + c_i h)\right) + \mathcal{O}(h^{p+1}), \end{aligned}$$

womit folgt

$$y(t_n + h) - y_n - h \sum_{i=1}^{s} b_i f\left(y(t_n + c_i h), \lambda(t_n + c_i h)\right) = \mathcal{O}(h^{p+1}).$$

Zieht man dies nun von

$$y_{n+1} - y_n = h \sum_{i=1}^{s} b_i f(Y_{ni}, \Lambda_{ni})$$

ab, folgt als Darstellung für den lokalen Fehler

$$y_{n+1} - y(t_n + h) = h \sum_{i=1}^{s} b_i \left(f(Y_{ni}, \Lambda_{ni}) - f\left(y(t_n + c_i h), \lambda(t_n + c_i h)\right)\right) + \mathcal{O}(h^{p+1}).$$

Wiederum in der verkürzten Schreibweise mit $\hat{Y}$ und $\hat{\Lambda}$ ergibt sich durch Taylorentwicklung für $\delta y_h(t_n) = y_{n+1} - y(t_n + h)$:

$$\begin{aligned} \delta y_h(t_n) &= h \sum_{i=1}^{s} b_i \left(f(Y_{ni}, \Lambda_{ni}) - f(\hat{Y}_{ni}, \hat{\Lambda}_{ni})\right) + \mathcal{O}(h^{p+1}) \\ &\overset{\text{Taylor}}{=} h \sum_{i=1}^{s} \Big(b_i \Big(f(\hat{Y}_{ni}, \hat{\Lambda}_{ni}) + f_y(\hat{Y}_{ni}, \hat{\Lambda}_{ni})(Y_{ni} - \hat{Y}_{ni}) + f_\lambda(\hat{Y}_{ni}, \hat{\Lambda}_{ni})(\Lambda_{ni} - \hat{\Lambda}_{ni})\Big) \\ &\quad - f(\hat{Y}_{ni}, \hat{\Lambda}_{ni}) + \mathcal{O}(h^{p+1}) + \mathcal{O}(Y_{ni} - \hat{Y}_{ni})^2 + \mathcal{O}(\Lambda_{ni} - \hat{\Lambda}_{ni})^2\Big) \\ &\overset{(2.19)}{=} h \sum_{i=1}^{s} b_i f_\lambda(\hat{Y}_{ni}, \hat{\Lambda}_{ni})(\Lambda_{ni} - \hat{\Lambda}_{ni}) + \mathcal{O}(h^{q+2}) + \mathcal{O}(h^{p+1}) \\ &\overset{\text{Taylor}}{=} h \sum_{i=1}^{s} b_i \left(f_\lambda(y_n, \lambda_n) + \mathcal{O}(h)\right) \underbrace{(\Lambda_{ni} - \hat{\Lambda}_{ni})}_{\mathcal{O}(h^q)} + \mathcal{O}(h^{q+2}) \\ &= h f_\lambda(y_n, \lambda_n) \sum_{i=1}^{s} b_i (\Lambda_{ni} - \hat{\Lambda}_{ni}) + \mathcal{O}(h^{q+2}). \end{aligned} \tag{2.20}$$

Da nach Modellannahme f_λ beschränkt ist, folgt hieraus mit (2.19) die Abschätzung des lokalen Fehlers für y durch $\mathcal{O}(h^{p+1})$ und somit die erste Aussage des Satzes. Für die zweite Abschätzung benötigen wir noch

$$P(t_n) f_\lambda(y_n, \lambda_n) = f_\lambda(y_n, \lambda_n) - (f_\lambda(g_y f_\lambda)^{-1} g_y f_\lambda)(y_n, \lambda_n) = f_\lambda(y_n, \lambda_n) - f_\lambda(y_n, \lambda_n) = 0.$$

Hieraus folgt, in (2.20) eingesetzt, dass $P(t_n)\delta y_h(t_n) = \mathcal{O}(h^{q+2})$, also die zweite Abschätzung des Theorems.
Für die Abschätzung der λ-Komponente ist zunächst noch etwas Vorarbeit nötig, da für diese die Ableitung nicht explizit gegeben ist. Die Gleichungen für einen Schritt des Runge-Kutta-Verfahrens bezüglich der λ-Komponente lauten nach (2.6) und (2.7)

$$\lambda_{n+1} = \lambda_n + h\sum_{i=1}^{s} b_i \Lambda'_{ni} \tag{2.21}$$

$$\Lambda_{ni} = \lambda_n + h\sum_{j=1}^{s} a_{ij}\Lambda'_{nj}, \tag{2.22}$$

oder in Matrix-Vektor-Schreibweise für (2.22) mit Notation $\Lambda_n = (\Lambda_{n1}, \dots, \Lambda_{ns})^T$:

$$\Lambda_n = \lambda_n \mathbb{1} + hA\Lambda'_n.$$

Da A invertierbar ist, kann dies nach Λ'_n aufgelöst und in (2.21) eingesetzt werden:

$$\lambda_{n+1} = \lambda_n + b^T A^{-1}(\Lambda_n - \mathbb{1}\lambda_n). \tag{2.23}$$

Des Weiteren kann man λ mit der noch unbekannten Ableitung formal integrieren, d.h. $\lambda(t_n + x) = \lambda_n + \int_0^x \lambda'(t_n + x)dx$. Approximiert durch die inneren und äußeren Quadraturformeln des Runge-Kutta-Verfahrens (also mit Ordnung p bzw. q) ergibt sich so

$$\lambda(t_n + h) = \lambda_n + h\sum_{i=1}^{s} b_i \lambda'(t_n + c_i h) + \mathcal{O}(h^{p+1}),$$
$$\lambda(t_n + c_i h) = \lambda_n + h\sum_{j=1}^{s} a_{ij}\lambda'(t_n + c_j h) + \mathcal{O}(h^{q+1}).$$

Wiederum kann man die zweite Zeile in Matrix-Vektor Darstellung umformen zu

$$\hat{\Lambda}_n = \mathbb{1}\lambda_n + hA\hat{\Lambda}'_n + \mathcal{O}(h^{q+1}),$$

was aufgrund der Invertierbarkeit von A und für hinreichend kleines h nach $\hat{\Lambda}'_n$ auflösbar ist. Einsetzen in die erste Zeile liefert

$$\lambda(t_n + h) = \lambda_n + b^T A^{-1}(\hat{\Lambda}_n - \mathbb{1}\lambda_n) + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}). \tag{2.24}$$

Der lokale Fehler von λ ergibt sich damit direkt aus der Differenz von (2.23) und (2.24) als

$$\begin{aligned}\lambda_{n+1} - \lambda(t_n + h) &= \lambda_n - \lambda_n + b^T A^{-1}(\Lambda_n - \mathbb{1}\lambda_n - \hat{\Lambda}_n + \mathbb{1}\lambda_n) + \mathcal{O}(h^{q+1}) + \mathcal{O}(h^{p+1}) \\ &= b^T A^{-1} \underbrace{(\Lambda_n - \hat{\Lambda}_n)}_{\mathcal{O}(h^q)\text{nach(2.19)}} + \mathcal{O}(h^{q+1}) \\ &= \mathcal{O}(h^q),\end{aligned}$$

die letzte Aussage des Satzes. □

Mit diesen Aussagen über die lokalen Fehler lässt sich der globale Verfahrensfehler für die y- und λ-Komponente abschätzen.

Zur Formulierung des entsprechenden Theorems wird noch die Stabilitätsfunktion R eines Runge-Kutta-Verfahrens benötigt, diese ist definiert durch $R(z) = 1 + zb^T(\mathbb{1} - zA)^{-1}(1, \dots, 1)^T$.

Bei regulärem A gilt damit

$$\begin{aligned} 1 + zb^T(\mathbb{1} - zA)^{-1}(1,\dots,1)^T &= 1 - b^T A^{-1}\left(\mathbb{1} - z^{-1}A^{-1}\right)^{-1}, \\ &= 1 - b^T A^{-1} + \mathcal{O}(z^{-1}) \end{aligned}$$

und somit $R(\infty) = 1 - b^T A^{-1}(1,\dots,1)^T$.

Theorem 3 *Theorem 4.4 aus [HLR89]*
Bei einer Differential-Algebraischen-Gleichung vom Index 2 der Form (1.10) gelte die Modellannahme 3 und es seien konsistente Anfangswerte gegeben (d.h. die gegebenen Anfangswerte erfüllen die zweite Zeile von (1.10)). Weiter sei die Koeffizientenmatrix A des Runge-Kutta-Verfahrens regulär und es gelte $|R(\infty)| < 1$. Ist der lokale Verfahrensfehler für die y-Komponente abschätzbar durch

$$\delta y_h(t) = \mathcal{O}(h^r), \quad P(t)\delta y_h(t) = \mathcal{O}(h^{r+1}),$$

dann ist das Runge-Kutta-Verfahren konvergent der Ordnung r bezüglich der y-Komponente, d.h.

$$y_n - y(t_n) = \mathcal{O}(h^r)$$

mit $t_n = nh \leq$ Konst. Ist zusätzlich $\delta y_h(t) = \mathcal{O}(h^{r+1})$, dann gilt $g(y_n) = \mathcal{O}(h^{r+1})$.

Der allgemeine Beweis ist etwas umfangreicher und besteht aus fünf Teilen und einem Hilfslemma. Daher folgt der hier gezeigte Beweis nicht wie bisher [HLR89], sondern [Hai10] (Theorem VII.4.5). Dort wird ein modifizierter Beweis für den Sonderfall steif-genauer Runge-Kutta-Verfahren ($a_{si} = b_i$), also ausreichend für die später verwendeten Runge-Kutta-Verfahren, vorgestellt. Dieser ist erheblich kompakter. Der Beweis basiert in beiden Fällen letztlich auf der Beweistechnik *Lady Windermeres Fächer*, wie sie zum Beispiel in [Hoc12], Beweis von Satz 8.5 beschrieben ist.
Die Idee der Beweismethode von Lady Windermeres Fächer ist in Kürze, rekursiv vom lokalen Fehler eine Abschätzung für den globalen Fehler zu finden. Man vergleicht bei einer numerischen Integration, etwa mit einem Runge-Kutta-Verfahren mit N Schritten, den Fehler als Differenz zwischen der echten Lösung und der Lösung aus dem Runge-Kutta-Verfahren, wenn nur der letzte Zeitschritt numerisch integriert wurde (also der lokale Fehler in einem Schritt). Mit diesem Ergebnis vergleicht man das Ergebnis des Runge-Kutta-Verfahrens, wenn die letzten beiden Zeitschritte numerisch integriert wurden und so weiter, bis zum Unterschied der Lösungen, wenn alle bis auf einen, bzw. alle Zeitschritte numerisch statt exakt integriert wurden. All diese Differenzen aufaddiert ergeben eine Fehlerabschätzung zwischen der exakten Lösung und der rein numerischen Lösung, also dem globalen Verfahrensfehler. Die Notation für die jeweiligen Lösungen sei y_N^k, wobei N die Anzahl der gerechneten Zeitschritte und k die Anzahl der hiervon exakt integrierten Zeitschritte ist, d.h. y_N^N ist der exakte Wert nach N Zeitschritten und y_N^0 die vollständig numerisch berechnete Lösung.
Sei nun also ein steif-genaues Runge-Kutta-Verfahren gegeben, d.h. $a_{si} = b_i,\ i = 1,\dots,s$. Betrachtet man das dem Runge-Kutta-Verfahren zugrunde liegende Gleichungssystem (2.5) und (2.7), ergibt sich hieraus, dass $y_{n+1} = Y_{ns}$. Da die Y_{ni} insbesondere $g(Y_{ni}) = 0$ erfüllen müssen, folgt hieraus, dass $g(y_{n+1}) = 0$ gilt, damit ist die Nebenbedingung in jedem Schritt erfüllt. Dies erlaubt eine Taylorentwicklung der 0 der Form

$$0 = g(y_{n+1}) - g(y(t_n + h)) \tag{2.25}$$

$$\overset{\text{Taylor}}{=} g(y_{n+1}) - \left(g(y_{n+1}) + g_y(y_{n+1})(y_{n+1} - y(t_n + h)) + \mathcal{O}\left(y(t_n + h) - y_{n+1}\right)^2\right) \tag{2.26}$$

$$= -g_y(y_{n+1})\delta y_h(t_n) + \mathcal{O}(\delta y_h(t_n))^2 \tag{2.27}$$

$$= -(g_y(y_n) + \underbrace{g_{yy}(y_n)}_{\text{beschr.}}\underbrace{(y_{n+1} - y_n)}_{\mathcal{O}(h)})\delta y_h(t_n) + \mathcal{O}\underbrace{\left(\delta y_h(t_n)^2\right)}_{\mathcal{O}(h^r)} \tag{2.28}$$

$$= -g_y(y_n)\delta y_h(t_n) + \mathcal{O}(h\delta y_h(t_n)), \tag{2.29}$$

damit folgt wegen der Voraussetzung, dass $\delta y_h(t_n) = \mathcal{O}(h^r)$, dass $g_y(y_n)\delta y_h(t_n) = \mathcal{O}(h^{r+1})$. Zusätzlich zur Voraussetzung $P(t)\delta y_h(t) = \mathcal{O}(h^{r+1})$ ergibt sich somit

$$\begin{aligned} P(t_n)\delta y_h(t_n) &= \delta y_h(t_n) - \underbrace{f_\lambda}_{\text{beschr.}} \underbrace{(g_y f_y)^{-1}}_{\text{beschr. n. V.}} \underbrace{g_y(y_n)\delta y_h(t_n)}_{\mathcal{O}(h^{r+1}),\ \text{s.o.}} \\ &= \delta y_h(t_n) + \mathcal{O}(h^{r+1}) \\ &\overset{!}{=} \mathcal{O}(h^{r+1}). \end{aligned}$$

Daraus folgt, dass $\delta y_h(t_n)$ sogar $\mathcal{O}(h^{r+1})$. Mit diesen für den Sonderfall steif-genauer Runge-Kutta-Verfahren geltenden Eigenschaften nun zum Beweis des Theorems.

Beweis: Im ersten Schritt werden die Differenzen $y_N^{k-1} - y_N^k$ durch die lokalen Fehler $\delta y_h(t_k) = y_{k+1}^{k-1} - y_{k+1}^k$ abgeschätzt. Hierfür seien $\tilde{y}_n$ und $\hat{y}_n$ zwei Runge-Kutta-Lösungen mit (hinreichend geringfügig) unterschiedlichen Anfangswerten (Genauer gesagt in unserem Fall zwei verschiedene Lösungen aus Lady Windermeres Fächer, für die ab verschiedenen Schritten numerisch gerechnet wird. Die Differenz, die sie dadurch in jenem Schritt haben, ab dem für beide Lösungen numerisch gerechnet wird, kann mal als Differenz im „Anfangswert“ bei eben diesem Schritt betrachten.). Ihre Differenz sei durch $\Delta y_n := \tilde{y}_n - \hat{y}_n$ gegeben. Es gelte – zunächst als Annahme, die Rechtfertigung folgt später –

$$\|\hat{y}_n - y(t_n)\| \le C_0 h, \quad \|\Delta y_n\| \le C_1 h^2, \tag{2.30}$$

somit auch $\|\hat{y}_n - y(t_n)\| = \mathcal{O}(h)$. Hierbei und im Folgenden seien C_i positive Konstanten. Damit sind die Voraussetzungen für Theorem 1 mit $\theta_i = \delta_i = 0$ erfüllt und man erhält Abschätzungen für die Zwischenstellen eines Schrittes:

$$\left\|\tilde{Y}_{ni} - \hat{Y}_{ni}\right\| \le C_2 \|\Delta y_n\|, \quad \left\|\tilde{\Lambda}_{ni} - \hat{\Lambda}_{ni}\right\| \le C_2 \|\Delta y_n\|. \tag{2.31}$$

Für Δy_{n+1} gilt nach Verfahrensvorschrift:

$$\tilde{y}_{n+1} - \hat{y}_{n+1} = \tilde{y}_n - \hat{y}_n + h \sum_{i=1}^{s} b_i \left(f(\tilde{Y}_{ni}, \tilde{\Lambda}_{ni}) - f(\hat{Y}_{ni}, \hat{\Lambda}_{ni}) \right).$$

Nutzt man zusätzlich eine die Lipschitz-Eigenschaft von f mit der Konstanten L aus, folgt hieraus die Abschätzung

$$\begin{aligned} \|\Delta y_{n+1}\| &\le \|\Delta y_n\| + h \sum_{i=1}^{s} |b_i| \left\| f(\tilde{Y}_{ni}, \tilde{\Lambda}_{ni}) - f(\hat{Y}_{ni}, \hat{\Lambda}_{ni}) \right\| \\ &\le \|\Delta y_n\| + h \sum_{i=1}^{s} |b_i| L \left(\left\|\tilde{Y}_{ni} - \hat{Y}_{ni}\right\| + \left\|\tilde{\Lambda}_{ni} - \hat{\Lambda}_{ni}\right\| \right) \\ &\overset{(2.31)}{\le} \|\Delta y_n\| + h \sum_{i=1}^{s} |b_i| 2C_2 \|\Delta y_n\| \\ &\le (1 + hC_3) \|\Delta y_n\|, \end{aligned}$$

iterativ folgt daraus $\|\Delta y_n\| \le C_4 \|\Delta y_0\|$. Dies können wir für die Methode Lady Windermeres Fächer benutzen. Hierfür betrachten wir die Differenz $y_n^l - y_n^{l+1}$, $n > l$, also einmal der numerischen Lösung ab dem l-ten und einmal ab dem $(l+1)$-ten exakten Schritt. Im $(l+1)$-ten Schritt haben beide Lösungen als Unterschied den lokalen Fehler, der bei y_n^l vom l-ten auf den $(l+1)$-ten Schritt durch das Runge-Kutta-Verfahren entstanden ist, d.h. die Gesamtdifferenz entspricht einer Iteration mit $n-l-1$ Schritten und einem Anfangsfehler $\Delta y_l = \delta y_h(t_l) = \mathcal{O}(h^{r+1})$. Mit dem eben Gezeigten folgt daraus

$$\left\|y_n^l - y_n^{l+1}\right\| \le C_4 \|\Delta y_l\| \le C_5 h^{r+1}.$$

Setzt man nun die Abschätzungskette nach Lady Windermeres Fächer ein, ergibt sich für hinreichend kleines h (insbesondere $nh \leq$ Konst.)

$$\|y_n - y(t_n)\| \leq \sum_{l=0}^{n-1} \left\| y_n^l - y_n^{l+1} \right\| \leq nC_4 h^{r+1} = \mathcal{O}(h^r),$$

also die gesuchte Konvergenzaussage bezüglich der y-Variable.

Bleibt noch die Rechtfertigung der am Anfang getroffenen Annahmen (2.30). Dies lässt sich induktiv über n zeigen.
Induktionsanfang: für $n = 1$ ist der globale gerade der lokale Fehler, dieser liegt nach Voraussetzung und für steif-genaue Verfahren bei $\mathcal{O}(h^{r+1})$.
Induktionsvoraussetzung: Bis zu einem festen n gelten die Abschätzungen (2.30).
Induktionsschluss: Betrachte eine Lösung mit $n + 1$ Schritten. Wie für den Hauptbeweis nachgerechnet, gilt mit (2.30) die Abschätzung $\|y_n - y(t_n)\| = \mathcal{O}(h^r)$, damit können sich die verschiedenen Lösungen zum Zeitpunkt t_n nur um $\mathcal{O}(h^{r+1})$ unterscheiden. Berechnet man ab diesem Zeitpunkt mit allen Lösungen einen weiteren Schritt, ergibt sich durch Theorem 1 mit $\hat{y}_n - y_n = \mathcal{O}(h^{r+1})$ und $\delta_i = \theta_i = 0$ wiederum eine Abschätzung für die Differenz zwischen den verschiedenen Y_{ni}. Diese liegt in der Größenordnung $\mathcal{O}(h^{r+1})$. Da bei einem steif-genauen Verfahren die Gleichheit $y_{n+1} = Y_{ns}$ gilt, ist damit der zweite Teil von (2.30) gezeigt.
Der erste Teil ergibt sich mit einer groben Abschätzung. Da f beschränkt ist, gilt für die exakte Lösung

$$\begin{aligned} \|y(t_n + h)\| &= \left\| y_n + h \int_0^1 f\left(y(t_n + x), \lambda(t_n + x)\right) dx \right\| \\ &\leq \|y_n\| + h \|f\|_\infty = \|y_n\| + \mathcal{O}(h). \end{aligned}$$

Hierbei bezeichnet $\|.\|_\infty$ die Maximumsnorm bezüglich einer hinreichenden Lösungsumgebung. Gleichzeitig gilt für alle numerischen Lösungen mit dem n-ten Schritt $\hat{y}_n$

$$\begin{aligned} \hat{y}_{n+1} &= \underbrace{\hat{y}_n}_{= y_n + \mathcal{O}(h^{r+1}) \text{ s.o.}} + h \sum_{i=1}^{s} b_i \underbrace{f(Y_{ni}, \Lambda_{ni})}_{\text{beschr. in Lösungsumgebung}} \\ &= y_n + \mathcal{O}(h). \end{aligned}$$

Daraus folgt $\|\hat{y}_{n+1} - y(t_{n+1})\| = C_0 h$.
Mit hinreichend groß gewählten Konstanten C_0 und C_1 ist somit (2.30) stets gerechtfertigt und der Rest des Beweises gilt, da die hier vorkommenden Konstanten C_2, C_3 und C_4 unabhängig von C_0 und C_1 sind, welche nur als Voraussetzung für Theorem 1 benötigt wurden. □

Mit Kenntnis über den globalen Fehler der y-Komponente kann man den globalen Fehler der λ-Komponente abschätzen:

Theorem 4 *Theorem 4.6 aus [HLR89]*
Es gelten die gleichen Voraussetzungen wie beim vorigen Theorem 3, insbesondere sei der globale Fehler der y-Komponente $\mathcal{O}(h^k)$ und die Abweichung der algebraischen Nebenbedingung $g(y_n) = \mathcal{O}(h^{k+1})$. Ist zusätzlich der lokale Fehler der λ-Komponente $\mathcal{O}(h^k)$, dann gilt für den globalen Fehler von λ:

$$\lambda_n - \lambda(t_n) = \mathcal{O}(h^k),$$

mit $t_n = nh \leq$ Konst.

Beweis: Seien die exakte Lösung nach $n+1$ Zeitschritten, die vollständig numerische Lösung sowie eine numerische Lösung vom exakten n-ten Schritt ausgehend gegeben, hierfür wird die Notation aus dem Beweis von Theorem 3 (Lady Windermeres Fächer) genutzt. Setzt man (2.23) für λ_{n+1}^n und λ_{n+1}^0 ein, ergibt sich als Differenz

$$\begin{aligned}\lambda_{n+1}^0 - \lambda_{n+1}^n &= \lambda_n^0 - \lambda_n^n - b^T A^{-1}(\Lambda_n^0 - \mathbb{1}\lambda_n^0 - \Lambda_n^n + \mathbb{1}\lambda_n^n) \\ &= R(\infty)(\lambda_n^0 - \lambda_n^n) + b^T A^{-1}(\Lambda_n^0 - \Lambda_n^n).\end{aligned}$$

Nutzt man, dass der Unterschied zwischen λ_{n+1}^n und $\lambda_{n+1}^{n+1} = \lambda(t_{n+1})$ nur aus dem im letzten Schritt aufgetretenen, also lokalen, Fehler besteht, der nach Voraussetzung $\mathcal{O}(h^k)$ ist, folgt hieraus

$$\lambda_{n+1}^0 - \lambda_{n+1}^{n+1} = R(\infty)(\lambda_n^0 - \lambda_n^n) + b^T A^{-1}(\Lambda_n^0 - \Lambda_n^n) + \mathcal{O}(h^k). \tag{2.32}$$

Weiterhin ergibt eine Taylorentwicklung von $0 = g(y_n^n)$ mit $g(y_n^0) = \mathcal{O}(h^k)$, dass

$$0 = g(y_n^n) = \underbrace{g(y_n^0)}_{\mathcal{O}(h^{k+1})} + g_y(y_n^0)(y_n^n - y_n^0) + \mathcal{O}\left(\underbrace{(y_n^n - y_n^0)}_{\mathcal{O}(h^k)}{}^2\right),$$

also muss gelten $g_y(y_n^0)(y_n^n - y_n^0) = \mathcal{O}(h^{k+1})$. Damit kann man eine Abschätzung für $\Lambda_{ni}^n - \Lambda_{ni}^0$ aus Theorem 1 mit $\theta_i = \delta_i = 0$ und $y_n^n - y_n^0 = \mathcal{O}(h^k)$ gewinnen und zwar

$$\left\|\Lambda_{ni}^0 - \Lambda_{ni}^n\right\| \leq \frac{C}{h}\left(\mathcal{O}(h^{k+1}) + h\mathcal{O}(h^k)\right) = \mathcal{O}(h^k).$$

Einsetzen in (2.32) liefert damit

$$\lambda_{n+1}^0 - \lambda_{n+1}^{n+1} = R(\infty)(\lambda_n^0 - \lambda_n^n) + \mathcal{O}(h^k).$$

Für steif-genaue Verfahren, also für $b_i = A_{ni}$, ist damit die Aussage bereits gezeigt, da hier

$$\begin{aligned}R(\infty) &= 1 - b^T A^{-1}(1,\ldots,1)^T \\ &\overset{\text{s.o.}}{=} 1 - (A_{\text{n. Zeile}})^T A^{-1}(1,\ldots,1)^T \\ &= 1 - (0,\ldots,0,1)(1,\ldots,1)^T \\ &= 0,\end{aligned}$$

gilt. Ansonsten gilt iterativ und wegen $|R(\infty)| < 1$

$$\begin{aligned}\lambda_2^0 - \lambda_2^2 &= R(\infty)\underbrace{(\lambda_1^0 - \lambda_1^1)}_{\mathcal{O}(h^k),\text{ lok. Fehler}} + \mathcal{O}(h^k) = \mathcal{O}(h^k) \\ \lambda_3^0 - \lambda_3^3 &= R(\infty)\underbrace{(\lambda_2^0 - \lambda_2^2)}_{\mathcal{O}(h^k),\text{ s.o.}} + \mathcal{O}(h^k) = \mathcal{O}(h^k) \\ &\vdots \\ \lambda_n^0 - \lambda_n^n &= \mathcal{O}(h^k),\end{aligned}$$

d.h. der globale Fehler bzgl. der λ-Komponente ist $\mathcal{O}(h^k)$.

□

Hiermit wurden nun insgesamt Bedingungen an Runge-Kutta-Verfahren gegeben, bei deren Erfüllung eine gewisse Konvergenzordnung des Verfahrens für Differential-Algebraische-Gleichungen garantiert ist, allerdings stimmt diese im Allgemeinen nicht mit der allgemeinen Ordnung des Runge-Kutta-Verfahrens für explizite Differentialgleichungen überein. Dennoch ist es nicht notwendig, speziell für Differential-Algebraische-Gleichungen nach neuen Runge-Kutta-Verfahren zu suchen, da es praktischerweise eine Klasse von Runge-Kutta-Verfahren gibt, welche die Bedingungen nach Konstruktion erfüllt. Dies sind die sogenannten Kollokationsverfahren, genauer gesagt zu Kollokationsverfahren äquivalente Runge-Kutta-Verfahren.

2.3 Kollokationsverfahren für Differential-Algebraische-Gleichungen

Grundlage für die folgenden Untersuchungen sind die etwa in [Hoc12], Kapitel 10.5 beschriebenen Kollokationsverfahren. Die Idee eines s-stufigen Kollokationsverfahrens ist es, eine gesuchte Lösung einer Differentialgleichung der Form $\dot{y} = f(y)$ durch eine Art Interpolationspolynom u zu approximieren. Dieses soll in jedem Schritt als Anfangswert den Endwert des letzten Schrittes übernehmen, also $u(t_n) = y_n$, sowie bei den s Zwischenstufen zumindest die richtige Steigung besitzen, also $u'(t_n + c_i h) = f\left(u\left(t_n + c_i h\right)\right)$. Hierbei steht analog zu den Runge-Kutta-Verfahren h für die Schrittweite und die $c_i,\ i = 1, \ldots, s$ für die Zwischenstellen. y_{n+1} wird dann als $u(t_{n+1})$ gesetzt und erneut iteriert. Insbesondere von Bedeutung ist in diesem Zusammenhang Satz 10.12 aus [Hoc12], also dass ein solches Kollokationsverfahren äquivalent zum s-stufigen Runge-Kutta-Verfahren

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

mit

$$a_{ij} = \int_0^{c_i} L_j(x)dx$$

und

$$b_j = \int_0^1 L_j(x)dx, \quad i, j\ = 1, \ldots, s$$

ist. Hierbei sind

$$L_j(x) = \prod_{k=1, k\neq j}^{s} \frac{x - c_k}{c_j - c_k}$$

die Lagrange-Interpolations-Polynome in den Stützstellen c_i. Wie man leicht nachrechnen kann, gilt

$$L_i(c_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \tag{2.33}$$

Insbesondere (siehe [Hoc12], Formel (10.12)) kann man mit dieser Darstellung der Koeffizienten a_{ij} und b_j nachrechnen, dass $B(s)$ und $C(s)$ aus Theorem 2 erfüllt sind:

$$\begin{aligned}\sum_{j=1}^{s} a_{ij} c_j^{k-1} &= \sum_{j=1}^{s} \int_0^{c_i} L_j(x) dx c_j^{k-1} \\ &= \int_0^{c_i} \underbrace{\sum_{j=1}^{s} L_j(x) c_j^{k-1}}_{=:p(x)} dx\end{aligned}$$

Hierbei ist $p(x)$ nach Definition als Summe von Polynomen vom Grad $s-1$ ein Polynom vom Grad $s-1$, das an den s verschiedenen Stützstellen c_j die Werte c_j^{k-1} annimmt, somit ist $p(x)$ eindeutig festgelegt als $p(x) = x^{k-1}$, also

$$\begin{aligned}\sum_{j=1}^{s} a_{ij} c_j^{k-1} &= \int_0^{c_i} x^{k-1} dx \\ &= \frac{1}{k} c_i^k.\end{aligned}$$

Analog gilt

$$\begin{aligned}\sum_{j=1}^{s} b_j c_j^{k-1} &= \sum_{j=1}^{s} \int_0^{1} L_j(x) dx c_j^{k-1} \\ &= \int_0^{1} \underbrace{\sum_{j=1}^{s} L_j(x) c_j^{k-1}}_{=x^{k-1} \text{ vgl. oben}} dx \\ &= \frac{1}{k}.\end{aligned}$$

Hieraus ergibt sich direkt eine erste Folgerung für die Konvergenzordnung dieses Verfahrens, angewendet auf unser Modellproblem:

Korollar 1 *Konvergenzordnung für Kollokationsverfahren*
Gegeben sei ein zu einem s-stufigen Kollokationsverfahren äquivalentes Runge-Kutta-Verfahren der Ordnung $p > s$ und Schrittweite h. Angewendet auf das Modellproblem (1.10) mit allen Modellannahmen gilt für die globalen Fehler

$$\begin{aligned}y_n - y(t_n) &= \mathcal{O}(h^s), \\ \lambda_n - \lambda(t_n) &= \mathcal{O}(h^s).\end{aligned}$$

Außerdem ist der Fehler der Zwangsbedingung, also der zweiten Zeile von (1.10), gegeben durch $g(y_n) = \mathcal{O}(h^{s+1})$.

Beweis: Wie bereits gezeigt wurde, sind $B(s)$ und $C(s)$ aus Theorem 2 erfüllt. Für das Einbringen der Ordnung $p > s$ des Runge-Kutta-Verfahrens kann man etwa [Hoc12] folgen. Hier wird in Satz 8.12 eine notwendige und hinreichende Ordnungsbedingung für Runge-Kutta-Verfahren aufgestellt, die auf elementaren Differentialen und den dadurch repräsentierten *Bäumen* basiert. Für die genauen Details hierzu sei auf besagtes Skript verwiesen. Hier relevant ist, dass die Koeffizienten des Runge-Kutta-Verfahrens der Ordnung $p > s$ die Bedingungen aus Satz 8.12 für einen *Busch* der Ordnung $s+1$

(den für das elementare Differential $f^{(s)}(f,\dots,f)$ stehenden Baum) erfüllen, dass also

$$\sum_{i,j_1,\dots,j_s=1}^{s} b_i a_{ij_1}\dots x a_{ij_s} = \frac{1}{s+1}$$

gilt. Wegen $\sum_{j=1}^{s} a_{ij} = c_i$ ist dies äquivalent zu

$$\sum_{i=1}^{s} b_i c_i \dots c_i = \sum_{i=1}^{s} b_i {c_i}^s = C(s+1).$$

Es gilt also $C(s+1) = \frac{1}{s+1}$. Damit sind die Bedingungen von Satz 2 erfüllt und die lokalen Fehler bzgl. y und λ bzw. $P(t)\delta y_h(t)$ sind von der Ordnung $\mathcal{O}(h^{s+1})$, $\mathcal{O}(h^s)$ bzw. $\mathcal{O}(h^{s+2})$. Die Behauptung folgt dann direkt aus Theorem 3 und Theorem 4. □

Dieses Korollar ist eine direkte Folgerung aus den obigen Theoremen. Es werden lediglich die aus der Quadratur folgenden Eigenschaften der Koeffizienten von zu Kollokationsverfahren äquivalenten Runge-Kutta-Verfahren ausgenutzt. Für den Sonderfall steif-genauer Kollokationsverfahren kann allerdings eine erheblich bessere Konvergenzaussage bezüglich der y-Komponente aufgestellt werden. Hierfür ist es allerdings notwendig, Kollokationsverfahren etwas detaillierter zu betrachten.

Ein s-stufiges Kollokationsverfahren, angewendet auf unser Modellproblem, ergibt folgende Verfahrensvorschrift (vgl. [Hai10], *Definition 4.7*):

1. Finde (Kollokations-) Polynome u_n und v_n vom Grad s, welche folgende Bedingungen erfüllen:

$$\begin{aligned} u_n(t_n) = y_n, \quad v_n(t_n) &= \lambda_n \\ u_n(t_n + c_i h)' &= f\left(u_n(t_n + c_i h), v_n\left(t_n + c_i h\right)\right) \\ 0 &= g\left(u_n\left(t_n + c_i h\right)\right) \qquad\qquad \forall i = 1,\dots,s \end{aligned}$$

2. Setze $y(t_{n+1}) = u_n(t_{n+1})$ und $\lambda(t_{n+1}) = v_n(t_{n+1})$.

Es gilt also insbesondere ein nichtlineares Gleichungssystem für die Werte $u_n(t_n + c_i h)$, $u_n'(t_n + c_i h)$ und $v_n(t_n + c_i h)$ zu lösen. Dies ist das selbe nichtlineare Gleichungssystem wie bei der Anwendung des zu diesem Kollokationsverfahren äquivalenten Runge-Kutta-Verfahren, entsprechend gilt $Y_{ni} = u_n(t_n + c_i h)$ und $\Lambda_{ni} = v_n(t_n + c_i h)$. Damit lässt sich für Verfahren mit $c_i \neq 0,\ i = 1,\dots s$ ein erstes Resultat bezüglich der Approximationsgüte von u_n und v_n an y und λ zeigen:

Lemma 2.1 *Theorem 4.8 aus [Hai10]*
Die s-stufige Kollokationsmethode von oben erfülle $c_i \neq 0,\ i = 1,\dots,s$. Dann gelten für alle $k = 0,1,\dots,s$ auf einem Zeitintervall $[t_n, t_{n+1}]$ die Abschätzungen

$$\left\| u_n^{(k)} - y^{(k)} \right\| \leq C h^{s+1-k}$$
$$\left\| v_n^{(k)} - \lambda^{(k)} \right\| \leq C h^{s-k}.$$

Beweis: Aus den Identitäten für die Y_{ni} und Λ_{ni} folgt eine mögliche Darstellung der Kollokationspolynome durch die Punkte y_n und die Y_{ni} bzw. λ_n und Λ_{ni} als

$$u_n(t_n + th) = L_0(t)y_n + \sum_{i=1}^{s} Y_{ni}L_i(t) \tag{2.34}$$

$$v_n(t_n + th) = L_0(t)\lambda_n + \sum_{i=1}^{s} \Lambda_{ni}L_i(t), \tag{2.35}$$

wobei hier, in der Notation abweichend von den obigen Lagrangepolynomen, der Wert $j = 0$ mit $c_0 = 0$ hinzugenommen wird, also $L_j = \prod_{k=0,k\neq j}^{s} \frac{x-c_k}{c_j-c_k}$. Da $u_n(t)$ und $v_n(t)$ Polynome vom Grad s sind und Einsetzen der Werte c_i für t durch Ausnutzung der Eigenschaften der Lagrangefunktionen (2.33) gerade die Zwischenstellen Y_{ni} und Λ_{ni} ergibt, ist diese Darstellung der Kollokationspolynome gerechtfertigt.
Als weitere Gleichung benötigt man eine Interpolation der exakten Lösung. Unter der Annahme, dass die Lösung y hinreichend oft stetig differenzierbar ist, also insbesondere beschränkte Ableitungen auf dem Intervall $[t_n, t_{n+1}]$ hat, gilt durch Interpolation

$$y(t_n + th) = y_nL_0(t) + \sum_{i=1}^{s} y(t_n + c_ih)L_i(t) + \mathcal{O}(h^{s+1}).$$

Dass hierbei der Fehler der Interpolation in der Größenordnung $\mathcal{O}(h^{s+1})$ ist, folgt etwa nach [Hoc12], Satz 2.5 (hierbei ist $|t - t_i| \leq h$). Insbesondere ist die Interpolationsnäherung aber auch an den $s + 1$ Stützstellen für $t = c_i$ exakt, d.h. die Funktion

$$y(t_n + th) - y_nL_0(t) - \sum_{i=1}^{s} y(t_n + c_ih)L_i(t) \tag{2.36}$$

hat $s + 1$ Nullstellen. Da nach dem Satz von Rolle (siehe Theorem IV.2.3 aus [AE06]) bei hinreichend glatten Funktionen zwischen zwei Nullstellen immer eine Nullstelle der Ableitung existiert, hat die Ableitung von (2.36) noch s Nullstellen, die zweite $s - 1$ Nullstellen und schließlich die k-te Ableitung $s + 1 - k$ Nullstellen, alle im Intervall $[t_n, t_{n+1}]$. Das heißt, die Gleichung

$$h^k y^{(k)}(t_n + th) = y_nL_0^{(k)} + \sum_{i=1}^{s} y(t_n + c_ih)L_i^{(k)}(t)$$

gilt an $s + 1 - k$ Stellen. Da die rechte Seite nach Konstruktion ein Polynom vom Grad $s - k$ ist (ein k-mal abgeleitetes Polynom vom Grad s), kann man sie als Interpolationspolynom vom Grad $s - k$ von der k-ten Ableitung von $h^k y$ auffassen. Sie hat damit einen Interpolationsfehler in der Größenordnung

$$h^{s+1-k} \max_{t^* \in [t_n, t_{n+1}]} \left(\frac{d}{dt} h^k y(t) \right) (t^*) = \mathcal{O}(h^{s+1}),$$

da y nach Voraussetzung hinreichend glatt ist. Somit folgt als Interpolation

$$h^k y^{(k)}(t_n + th) = y_nL_0^{(k)} + \sum_{i=1}^{s} y(t_n + c_ih)L_i^{(k)}(t) + \mathcal{O}(h^{s+1}). \tag{2.37}$$

Zieht man dies von der k-ten Ableitung von (2.34) ab, ergibt sich

$$\begin{aligned} h^k \left(u_n^{(k)}(t_n + th) - y^{(k)}(t_n + th) \right) &= \sum_{i=1}^{s} \underbrace{(Y_{ni} - y(t_n + c_ih))}_{=\mathcal{O}(h^{s+1}) \text{ nach (2.19) mit } q=s} L_i^{(k)}(t) + \mathcal{O}(h^{s+1}) \\ &= \mathcal{O}(h^{s+1}). \end{aligned}$$

Damit folgt $\left(u_n^{(k)}(t_n+th)-y^{(k)}(t_n+th)\right)=\mathcal{O}(h^{s+1-k})$, die erste Behauptung.
Die zweite Behauptung folgt ganz analog, wiederum nach Annahme, dass die Lösung $\lambda(t)$ hinreichend glatt ist, bzw. die Ableitungen lokal beschränkt sind. Der einzige Unterschied ergibt sich in (2.19), welches für $\lambda(t_n+c_ih)-\Lambda_{ni}$ mit $q=s$ die schwächere Größenordnungsabschätzung $\mathcal{O}(h^s)$ liefert, damit erhalten wir eine h-Potenz weniger und letztlich die Fehlergrößenordnung $\mathcal{O}(h^{s-k})$.

□

Theorem 5 *Theorem 4.9 aus [Hai10]*
Es sei ein zu einem Kollokationsverfahren äquivalentes, steif-genaues Runge-Kutta-Verfahren mit $c_i\neq 0,\ i=1,\dots,s$ und $c_s=1$ und regulärer Matrix A gegeben, wobei die Quadraturformel mit Knoten c_i und Gewichten b_i die Ordnung p besitze. Dann gilt für den lokalen Fehler der y-Komponente unseres Modellproblems

$$\delta y_h(t_n)=\mathcal{O}(h^{p+1}).$$

Beweis: Analog zu Lemma 1 kann man die Defekte bei Einsetzen der Approximation in die Differential-Algebraische-Gleichung betrachten, diesmal allerdings mit kontinuierlichen Approximationen u_n und v_n an die Lösungen y und λ, damit auch mit kontinuierlichen Funktionen δ_n und θ_n für die jeweiligen Abweichungen. Da hier nur der lokale Fehler, also der Fehler innerhalb eines Verfahrensschrittes bei exaktem Anfangswert betrachtet wird, wird im Folgenden der Schrittindex n weggelassen und es werde o.B.d.A. nur der erste Schritt betrachtet. In dieser Notation sieht die Darstellung des Defekts folgendermaßen aus:

$$\dot{u}(t)=f\left(u(t),v(t)\right)+\delta(t) \tag{2.38}$$
$$0=g\left(u(t)\right)+\theta(t) \tag{2.39}$$

Da die Kollokationspolynome u und v so gewählt wurden, dass sie die Differential-Algebraische-Gleichung zumindest an den Stützstellen t_0+c_ih mit $i=0,\dots,s$ und $c_0=0$ erfüllen, gilt entsprechend auch $\delta(t_0+c_ih)=\theta(t_0+c_ih)=0$. Weiterhin gilt, da u und v als Polynome glatt und f und g nach Voraussetzung in Umgebung um die exakte Lösung hinreichend glatt sind, dass auch die Defekte mehrfach stetig differenzierbar sein müssen (sie liegen nach Theorem 2.1 für genügend kleines h in einer beliebig kleinen Lösungsumgebung). Nun liefert einmaliges Ableiten von (2.39) nach t:

$$\begin{aligned}0&=g_y(u(t))\dot{u}(t)+\dot{\theta}(t)\\&=g_y(u(t))\left(f(u(t),v(t))+\delta(t)\right)+\dot{\theta}(t).\end{aligned}$$

Dies kann man als Gleichung für u und v (in einer Umgebung um die exakte Lösung) auffassen, also

$$0=g_y(u)\left(f(u,v)+\delta(t)\right)+\dot{\theta}(t)=:F(u,v,\delta,\dot{\theta}).$$

Da F als Kompositionen hinreichend glatter Funktionen ebenfalls glatt ist und $F_v=g_y(u)f_\lambda(u,v)$ nach Modellannahme 3 invertierbar ist, ist F nach dem Satz von der impliziten Funktion lokal nach v auflösbar, also darstellbar in der Form $v(t):=G(u(t),\delta(t),\dot{\theta}(t))$. Hiermit lässt sich auch die exakte Lösung λ darstellen, die die Differential-Algebraische-Gleichung ohne Defekt erfüllt, d.h. $\lambda=G(y,0,0)$. Setzt man beides in (2.38) ein, ergibt sich

$$\begin{aligned}\dot{u}(t)&=f\left(u(t),G\left(u(t),\delta(t),\dot{\theta}(t)\right)\right)+\delta(t)\\\dot{y}(t)&=f\left(y(t),G\left(y(t),0,0\right)\right).\end{aligned}$$

Für den gesuchten lokalen Fehler $y_1-y_0=u(t_1)-y(t_1)$ benötigt man eine Darstellung der Differenz $u-y$. Eine Möglichkeit, dies aus den oben aufgestellten Gleichungen abzuleiten, ist die nichtlineare

Variation-der-Konstanten Formel (siehe [HNWXX], Theorem 14.5). Dessen Aussage lautet in Kürze: Für die Lösungen a und b von

$$\begin{aligned}\dot{a}(t) &= \bar{f}(t,a)\\ \dot{b}(t) &= \bar{f}(t,b) + h(t,b)\end{aligned}$$

mit Anfangswerten $a(t_0) = b(t_0) =: a_0$ und hinreichend glatten Funktionen $\bar{f}, h$, dann gilt

$$b(t) = a(t) + \int_{t_0}^{t} \frac{\partial a}{\partial a_0}(t,s,b(s))h(s,b(s))ds.$$

Hierbei ist die Ableitung $\frac{\partial a}{\partial a_0}$ der Lösung a nach dem Anfangswert a_0 im Sinne und der Notation von Theorem 14.3 aus [HNWXX].
Im vorliegenden Problem entspricht dies

$$\begin{aligned}\dot{y}(t) &= f(y(t), G(y(t),0,0))\\ \dot{u}(t) &= f(u(t), G(u(t),0,0) + h(t,u(t)),\end{aligned}$$

wobei für h gilt:

$$\begin{aligned}&h(t,u(t))\\ &= \delta(t) + f\left(u(t), G(u(t),\delta(t),\dot{\theta}(t))\right) - f\left(u(t), G(u(t),0,0)\right)\\ &= \delta(t) + \int_0^1 \frac{d}{d\tau} f\left(u(t), G(u(t),\tau\delta(t),\tau\dot{\theta}(t))\right) d\tau\\ &= \int_0^1 f_\lambda\left(u(t), G\left(u(t),\tau\delta(t),\tau\dot{\theta}(t)\right)\right)\left(G_2\left(u(t),\tau\delta(t),\tau\dot{\theta}(t)\right)\delta(t) + G_3(u(t),\tau\delta(t),\tau\dot{\theta}(t))\dot{\theta}(t)\right) d\tau\\ &\quad + \delta(t)\\ &=: Q_1(t)\delta(t) + Q_2(t)\dot{\theta}(t),\end{aligned}$$

Hierbei bezeichnen G_2, bzw. G_3 die Ableitungen von G nach der $2.$ bzw. $3.$ Komponente. Damit folgt nach Theorem 14.5 aus [HNWXX] für die Differenz $u - y$, dass

$$\begin{aligned}u(t) - y(t) &= \int_{t_0}^{t} \frac{\partial y}{\partial y_0}(t,s,u(s))h(s,u(s))ds\\ &= \int_{t_0}^{t} \frac{\partial y}{\partial y_0}(t,s,u(s))\left(Q_1(s)\delta(s) + Q_2(s)\dot{\theta}(s)\right) ds\\ &=: \int_{t_0}^{t} S_1(t,s)\delta(s) + S_2(t,s)\dot{\theta}(s)ds.\end{aligned}$$

Auswertung an der Stelle $t_0 + h$, also für $u(t) = y_1$ und partielle Integration angewendet auf $S_2(s)\dot{\theta}(s)$ liefert

$$\begin{aligned}y_1 - y(t_0+h) &= \int_{t_0}^{t_1} S_1(t_1,s)\delta(s) + S_2(t_1,s)\dot{\theta}(s)ds\\ &= \int_{t_0}^{t_1} \underbrace{S_1(t_1,s)\delta(s) - \frac{\partial S_2}{\partial t}(t_1,s)\theta(s)}_{=:\sigma(s)} ds + \left[S_2(t_1,s)\theta(s)\right]_{s=0}^{t_1}.\end{aligned}$$

Nach Konstruktion gilt $\theta(t_0) = 0$ und nach Voraussetzung $c_s = 1$, also ebenfalls $\theta(t_1) = \theta(t_0 + c_s h) = 0$. Damit fällt der zweite Teil weg und es bleibt das Integral $\int_{t_0}^{t_1} \sigma(s)ds$. Dieses Integral kann man etwa durch die Quadraturformel b_i, c_i integrieren, dies ergibt

$$y_1 - y(t_0+h) = \sum_{i=1}^{s} b_i\sigma(t_0 + c_i h) + \mathrm{err}(\sigma).$$

Nach Konstruktion von $\sigma(t)$, welches nur Anteile mit Faktor $\delta(t)$ oder $\theta(t)$ enthält, gilt, dass $\sigma(t_0 + c_i h) = 0$ für alle $i = 1, \ldots, s$. Bleibt also noch der Quadraturfehler err(σ); dieser ist nach Voraussetzung an die Quadraturordnung p von der Größenordnung $h^{p+1} \max_{t\in[t_0,t_1]} \left\|\sigma^{(p)}(t)\right\|$. Die p-te Ableitung von σ behinhaltet maximal p. Ableitungen von f, g, δ und θ. f und g sind nach Modellannahme in einer Umgebung um die exakte Lösung hinreichend glatt, und nach Theorem 5 liegen u und v in dieser Umgebung, damit sind, wie oben bereits argumentiert, ebenfalls δ und θ hinreichend glatt also auf dem gesamten Lösungsintervall gleichmäßig beschränkt. Somit gilt err$(\sigma) = \mathcal{O}(h^{p+1})$. Daraus folgt insgesamt $y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1})$, die gesuchte Aussage über den lokalen Fehler.

□

Diese neue Abschätzung für den lokalen Fehler kann man nun wiederum nutzen, um eine bessere Aussage für den globalen Fehler, zumindest für die y-Komponente zu erhalten:

Korollar 2 *Globaler Fehler der y-Komponente steif-genauer Kollokationsverfahren*
Unter gleichen Voraussetzungen wie für Theorem 5 mit $p > 1$, also insbesondere mit $\delta y_h(t_n) = \mathcal{O}(h^{p+1})$, gilt für unser Modellproblem, dass der globale Fehler der y-Komponente gegeben ist, durch

$$y(t_n) - y_n = \mathcal{O}(h^p)$$

für $nh <$ Konst.

Beweis: Ziel des Beweises ist es, die Voraussetzungen aus Theorem 3 zu zeigen. Wie im Beweis von Theorem 3 nachgerechnet wurde, gilt auch hier (2.29) und damit $g_y(y_n)\delta y_h(t_n) = \mathcal{O}(h^{p+2})$, nach Voraussetzung für den lokalen Fehler. Hieraus kann man auf $P(t_n)\delta y_h(t_n)$ schließen, denn es gilt

$$\begin{aligned} P(t_n)\delta y_h(t_n) &= \delta y_h(t_n) - \underbrace{f_\lambda(y(t_n), \lambda(t_n))}_{\text{beschr.}} \underbrace{(g_y f_y\,(y(t_n), \lambda(t_n)))^{-1}}_{\text{beschr.}} \underbrace{g_y(y_n)\delta y_h(t_n)}_{\mathcal{O}(h^{p+2}),\ \text{s.o.}} \\ &= \underbrace{\delta y_h(t_n)}_{\mathcal{O}(h^{p+1})} + \mathcal{O}(h^{p+2}) \\ &= \mathcal{O}(h^{p+1}). \end{aligned}$$

Damit sind die Voraussetzungen für Theorem 3 gegeben und es folgt der globale Fehler als $\mathcal{O}(h^p)$. □

Bemerkung 2 *Damit ist insgesamt gezeigt, dass ein zu einem s-stufigen Kollokationsverfahren äquivalentes, steif-genaues Runge-Kutta-Verfahren der Ordnung $p > s$ mit invertierbarer Koeffizientenmatrix gilt, dass*

$$\begin{aligned} y_n - y(t_n) &= \mathcal{O}(h^p) \\ \lambda_n - \lambda(t_n) &= \mathcal{O}(h^s), \end{aligned}$$

für $nh <$ Konst. Dies ist insbesondere bei unserem Modellproblem interessant, da wir letztlich nur an den verallgemeinerten Koordinaten des mechanischen Systems, hier den y-Koordinaten interessiert sind. Die $\lambda-$Koordinaten sind nur Hilfsvariablen für die Zwangsbedingungen. Insofern ist die geringere Konvergenzordnung für den λ-Anteil verschmerzbar, im Gegenzug die hohe Konvergenzordnung des y-Anteils erfreulich. Für diesen hat etwa das später betrachtete RadauIIA-5-Verfahren die selbe Konvergenzordnung wie bei expliziten Differentialgleichungen.

Wie bereits erwähnt, gibt es aber bei der numerischen Durchführung noch Probleme, was das Lösen des nichtlinearen Gleichungssystems (2.8) betrifft. Zwar zeigt Theorem 1 lokale Existenz und Eindeutigkeit der Lösung, jedoch ist dies noch keine Aussage über die Lösbarkeit von (2.8) mit numerischen Verfahren.

2.4 Implementierung und numerische Lösbarkeit

Um auf die genaue Problematik (und die Abhilfe nach [HLR89]) einzugehen, wird zunächst zusammengefasst, wie insbesondere implizite Runge-Kutta-Verfahren allgemein implementiert werden. Hierbei gibt es große Unterschiede zu expliziten Verfahren, da bei expliziten Verfahren aufgrund der strikten unteren Dreiecksgestalt der Koeffizientenmatrix kein nichtlineares Gleichungssystem zu lösen ist, sondern die inneren Stufen iterativ auszurechnen sind.

Im Grunde führt der folgende Abschnitt durch einige Anpassungen den Abschnitt IV.8 aus [Hai10] (Implementierung von Runge-Kutta-Verfahren für implizite Probleme) und Kapitel 7 aus [HLR89] (Besonderheiten der Lösung des NLGS bei Index 2 Problemen) zusammen.
Betrachten wir hierfür nochmal die allgemeine Aufstellung eines Runge-Kutta-Verfahren der Form (2.1) bis (2.3). Um den Einfluss von Rundungsfehlern zu verringern, werden statt der inneren Stufen Y_i die Differenzen zwischen den Stufen und den Schritten $z_i = Y_i - y_n$ iteriert. Zur Bestimmung dieser Differenzen ergibt sich durch Einsetzen in (2.3) das NLGS

$$z_j = h \sum_{i=1}^{s} a_{ij} f(y_n + z_i).$$

Sind diese Differenzen berechnet, lassen sich aus ihnen direkt die inneren Stufen Y_i bestimmen und die Gleichung (2.1) kann unverändert angewendet werden. Allerdings kann man die hierfür notwendigen s weiteren Funktionsauswertungen vermeiden, falls die Koeffizientenmatrix A invertierbar ist. Stellt man die Gleichung für die z_i in Matrix-Vektor Schreibweise dar, d.h.

$$Z := \begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix} = h(A \otimes \mathbb{1}) \begin{pmatrix} f(y_n + z_1) \\ \vdots \\ f(y_n + z_s) \end{pmatrix} = h(A \otimes \mathbb{1}) \begin{pmatrix} f(Y_1) \\ \vdots \\ f(Y_s) \end{pmatrix} =: h(A \otimes \mathbb{1}) F(Z), \tag{2.40}$$

so kann man (2.1) umschreiben zu

$$y_{n+1} = y_n + h(b_1, \ldots, b_s) \begin{pmatrix} f(Y_1) \\ \vdots \\ f(Y_s) \end{pmatrix} = y_n + (b_1, \ldots, b_s) A^{-1} \begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix},$$

womit sich das Bestimmen von y_{n+1} nur durch Matrix-Vektor-Multiplikationen erledigen lässt.
Betrachten wir nun das NLGS (2.40). Durch Subtraktion von Z lässt sich dies auf eine iterationsfähige Form für das Newtonverfahren bringen:

$$0 = -Z + h(A \otimes \mathbb{1}) F(Z) =: \tilde{F}(Z)$$

Unter gewissen Voraussetzungen an F und für h klein genug, sind die Voraussetzungen an den Satz von Kantorovich (etwa Theorem 5.4.2 [HA09]) erfüllt und es ergibt sich die Iterationsvorschrift

$$Z_{n+1} = Z_n - (\tilde{F}_Z(Z_n))^{-1} \tilde{F}(Z_n),$$

oder wieder als Differenz und mit F dargestellt:

$$\begin{gathered} \tilde{F}_Z(Z_n)(Z_{n+1} - Z_n) = -\tilde{F}(Z_n) \\ \Leftrightarrow \\ (\mathbb{1} - h(A \otimes \mathbb{1}) F_Z(Z_n)) \Delta Z_n = Z_n - h(A \otimes \mathbb{1}) F(Z_n) \\ \Delta Z_n = Z_{n+1} - Z_n, \end{gathered}$$

hierbei ist

$$F_Z(Z) = \begin{pmatrix} f_y(y_n + z_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & f_y(y_n + z_s) \end{pmatrix}.$$

Eine bei hinreichend kleiner Schrittweite gerechtfertigte Näherung ist, jede der hier verwendeten Ableitungen $f_y(y_n + z_i)$ durch eine Einzige anzunähern (vgl. mit dem vereinfachten Newtonverfahren), etwa durch $f_y(y_n) =: J$. Hierdurch vereinfacht sich die Rekursion zu

$$(\mathbb{1} - hA \otimes J)\Delta Z_n = Z_n - h(A \otimes \mathbb{1})F(Z_n).$$

Will man ein implizites Problem mit einer Matrix M lösen, ergeben sich die Gleichungen analog mit $M^{-1}f$ statt f. Im Falle einer singulären Matrix, also dem für Differential-Algebraische-Gleichungen interessanten Fall, ist dies jedoch nur formal möglich, da ein M^{-1} nicht existiert. Daher muss es durch anschließendes Multiplizieren der Gleichungen mit M eliminiert werden, was schließlich zur Rekursion

$$(\mathbb{1} \otimes M - hA \otimes J)\Delta Z_n = (\mathbb{1} \otimes M)Z_n - h(A \otimes \mathbb{1})F(Z_n)$$

führt.
Betrachten wir nun unser Modellproblem mit f und g. O.B.d.A. gelte bei einem Schritt $(y_n, \lambda_n) = 0$, ansonsten kann man das Verfahren wieder auf die Differenzen anwenden. Hiermit ergibt sich aus der Rekursion (2.8) durch Ersetzen von $0 = g(Y_{ni})$ durch den bei regulärem A äquivalenten Term $0 = h\sum_{j=1}^{s} a_{ij} g(Y_{ni})$ und Ausnutzen von $g_\lambda = 0$ die Matrix

$$(\mathbb{1} \otimes M - hA \otimes J) = \begin{pmatrix} \mathbb{1} - hA \otimes f_y & -hA \otimes f_\lambda \\ -hA \otimes g_y & 0 \end{pmatrix}, \tag{2.41}$$

wobei f_y, f_λ, g_y hier für die genäherten Ableitungen an der Stelle y_n, λ_n stehen. Das Iterieren des vereinfachten Newtonverfahrens erfordert eine Lösung des Gleichungssystems, also indirekt ein Invertieren von (2.41). Diese Inverse hat (vgl. [HLR89], Seite 93) die Form

$$\begin{pmatrix} \left(\mathbb{1} \otimes (\mathbb{1} - f_\lambda(g_y f_\lambda)^{-1} g_y) + \mathcal{O}(h)\right) & \left(-h^{-1}A^{-1} \otimes f_\lambda(g_y f_\lambda)^{-1} + \mathcal{O}(1)\right) \\ \left(-h^{-1}A^{-1} \otimes (g_y f_\lambda)^{-1} g_y + \mathcal{O}(1)\right) & \left(-h^{-2}A^{-2} \otimes (g_y f_\lambda)^{-1} + \mathcal{O}(h^{-1})\right) \end{pmatrix}.$$

Diese Matrix ist insbesondere unbeschränkt für $h \to 0$, womit eine Konvergenz des Newtonverfahrens nach dem Satz von Kantorovich (Theorem 5.4.2 [HA09]) nicht gewährleistet werden kann.
Tatsächlich konvergiert das Newtonverfahren allerdings doch. Um dies zu beweisen, fassen wir das verwendete vereinfachte Newtonverfahren als Fixpunktiteration auf. Betrachten wir im Detail die Verfahrensfunktion für einen Schritt des zu einer Fixpunktiteration umgestellten Newtonverfahrens. Diese ist gegeben durch

$$\begin{aligned} &\Phi(Y, \Lambda) \\ &= \begin{pmatrix} Y \\ \Lambda \end{pmatrix} - (\mathbb{1} - hA \otimes J)^{-1} \begin{pmatrix} Y - h(A \otimes \mathbb{1})f(Y, \Lambda) \\ -h(A \otimes \mathbb{1})g(Y) \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{1} \otimes f_\lambda(g_y f_\lambda)^{-1}(g_y Y - g(Y)) - hA \otimes f_\lambda(g_y f_\lambda)^{-1} g_y f(Y, \Lambda) + \mathcal{O}(h)(Y) + \mathcal{O}(h^2)(\Lambda) \\ \Lambda + h^{-1}A^{-1} \otimes (g_y f_\lambda)^{-1}(g_y Y - g(Y)) - \mathbb{1} \otimes (g_y f_\lambda)^{-1} g_y f(Y, \Lambda) + \mathcal{O}(1)(Y) + \mathcal{O}(h)(\Lambda) \end{pmatrix}. \end{aligned}$$

Die Ableitung dieser Verfahrensfunktion nach Y und Λ ist damit (hierbei benutzt : $\frac{\partial \Lambda}{\partial \Lambda} = \mathbb{1} = (g_y f_\lambda)^{-1} g_y f_\lambda$ und die Größenordnungen aus Theorem 1, sowie $'$ als Notation für die Ableitung nach

Y und Λ)

$$\begin{aligned}
&\Phi'(Y,\Lambda)\\
&= \begin{pmatrix} \mathbb{1}\otimes f_\lambda(g_y f_\lambda)^{-1}(g_y - g_y(Y)) + \mathcal{O}(h) & -hA\otimes f_\lambda(g_y f_\lambda)^{-1} g_y f_\lambda(Y,\Lambda) + \mathcal{O}(h^2)\\ h^{-1}A^{-1}\otimes (g_y f_\lambda)^{-1}(g_y - g_y(Y)) + \mathcal{O}(1) & \mathbb{1}\otimes (g_y f_\lambda)^{-1} g_y (f_\lambda - f_\lambda(Y,\Lambda)) + \mathcal{O}(h)\end{pmatrix}\\
&\overset{\text{Thm 1}}{=} \begin{pmatrix}\mathcal{O}(h) & \mathcal{O}(h^2)\\ \mathcal{O}(1) & \mathcal{O}(h)\end{pmatrix}.
\end{aligned}$$

Diese Matrix hat allerdings auch bei beliebig kleinem h die Norm $\mathcal{O}(1)$. Für Konvergenz nach dem Banachschen Fixpunktsatz (etwa Theorem 4.3 in [Hoc12]) benötigt man aber eine Kontraktion, also, dass die Norm der Ableitungsmatrix kleiner 1 ist und idealerweise durch $\mathcal{O}(h)$ abgeschätzt werden kann. Abhilfe schafft (vgl. [HLR89], S. 94) statt der Standardnorm des $\mathbb{R}^n$ die äquivalente Norm

$$\left\|\begin{pmatrix} Y\\ \Lambda\end{pmatrix}\right\|_D =: \left\| D\begin{pmatrix} Y\\ \Lambda\end{pmatrix}\right\|$$

mit der regulären Skalierungsmatrix $D = \operatorname{diag}(\mathbb{1}, h\mathbb{1})$ zu benutzen. Die hiervon induzierte Matrixnorm ergibt sich durch

$$\begin{aligned}
\|A\|_D &= \max_{\|x\|_D=1}\frac{\|Ax\|_D}{\|x\|_D} = \max_{\|Dx\|=1}\frac{\|DAx\|}{\|Dx\|} = \max_{\|Dx\|=1}\frac{\left\|DAD^{-1}Dx\right\|}{\|Dx\|} = \max_{\|y\|=1}\frac{\left\|DAD^{-1}y\right\|}{\|y\|}\\
&= \left\|DAD^{-1}\right\|.
\end{aligned}$$

Analoge Rechnung für $D\Phi' D^{-1}$ liefert

$$D\Phi'(Y,\Lambda)D^{-1} = \begin{pmatrix}\mathcal{O}(h) & \mathcal{O}(h)\\ \mathcal{O}(h) & \mathcal{O}(h)\end{pmatrix}.$$

Die Matrix Φ' hat also in der $\|.\|_D$-Norm die Größenordnung $\mathcal{O}(h)$ und die Fixpunktiteration konvergiert nach dem Banachschen Fixpunktsatz, falls h hinreichend klein ist. Bei der Implementierung wird dies dadurch berücksichtigt, dass die Norm der Index-2-Variablen zum Test des Abbruchkriteriums des vereinfachten Newtonverfahrens mit h multipliziert wird, also die Konvergenz nicht durch die Standardnorm, sondern durch diese modifizierte (aber äquivalente) Norm überprüft wird.
Weitere numerische Verbesserungen erhält das Verfahren durch Transformationen, durch die letztlich A^{-1} eine blockdiagonale Form erhält (vgl. [Hai10], Seite 121ff). Da A invertierbar und vor allem für das jeweilige Verfahren fest gewählt ist, sind diese Transformationen bereits bei der Implementierung berechenbar und ins Programm übertragbar. Diese Änderungen haben jedoch keine so gravierenden Auswirkungen auf Konvergenz oder Lösungsverhalten wie die oben gezeigte Skalierung des Fehlers für das vereinfachte Newtonverfahren, sondern dienen meist „nur“ der Recheneffizienz. Daher wird eine detailliertere Betrachtung an dieser Stelle unterlassen. Nachzulesen sind diese Modifikationen in [Hai10] (S. 118 ff). Bei den später verwendeten Matlab-Codes wurden diese Transformationen allerdings genutzt.

Kapitel 3

Regularisierung nach J. Deppler und A. Fidlin

Eine andere Herangehensweise an Differential-Algebraische-Gleichungen als das direkte numerische Lösen (etwa wie gezeigt durch Runge-Kutta-Verfahren) ist, sie bereits im Vorfeld durch eine explizite Differentialgleichung zu approximieren. Diese Idee spiegelt sich etwa bei sogenannten *singulär gestörten Problemen* (vgl. etwa Kap. 6, [Hai10]) wieder. Eine verbreitete Form hiervon ist

$$\begin{cases} \dot{y} &= f(y,\lambda) \\ \epsilon\dot{\lambda} &= g(y,\lambda) \end{cases}. \tag{3.1}$$

Für $\epsilon \neq 0$ ist dies eine explizite Differentialgleichung, deren Lösung analytisch oder numerisch bestimmt werden kann. Für den Grenzfall $\epsilon = 0$ ändert diese Gleichung ihren Charakter zu einer Differential-Algebraischen-Gleichung (daher *singuläre* Störung). Tatsächlich gibt es, unter gewissen Voraussetzungen an die beteiligten Funktionen, Aussagen über den Zusammenhang zwischen der Lösung für ein $\epsilon \neq 0$ und der Lösung der zugrundeliegenden Differential-Algebraischen-Gleichung, die hier zum Teil auch zitiert werden, dies führt schließlich zu Theorem 6.
Die in [Dep13] vorgeschlagene Variante für unser Modellproblem geht von einer Annäherung des physikalischen Modells aus und führt hieraus letztlich zu einer etwas allgemeineren Form singulär gestörter Probleme als der oben Gezeigten. Diese kann einerseits für den Grenzfall $\epsilon \to 0$ wieder als Näherung an die Differential-Algebraische-Gleichung aufgefasst werden, beschreibt andererseits für eine geeignete (kleine) Wahl von ϵ ein *realistischeres* mechanisches Modell als die sehr idealisierte Darstellung durch die Differential-Algebraische-Gleichung. Diese Untersuchung eines realistischeren mechanischen Modells ist die eigentliche Motivation für [Dep13]. Die exakte Erfüllung der Nebenbedingung $0 = g(y)$ ist spätestens auf mikroskopischer Ebene nicht mehr gegeben und muss daher durch ein anderes mechanisches Modell ersetzt werden. Außerdem ermöglicht es die Auffassung als singulär gestörte Differential-Algebraische-Gleichung, unter gewissen Bedingungen an die beteiligten Funktionen, das mechanische Modell durch die Differential-Algebraische-Gleichung zu approximieren.

3.1 Grundlegender Ansatz

Betrachten wir wiederum das Modellproblem in der ursprünglichen Darstellung (1.7), (1.9), also

$$\begin{aligned} \ddot{q} &= M^{-1}(q)(F(q,\dot{q}) - G(q)^T\lambda) \\ 0 &= G(q)\dot{q}. \end{aligned}$$

Die physikalische Bedeutung der algebraischen Bedingung (1.9) ist der Zwang zur Rollreibung, also, dass der Körper im Kontaktpunkt auf der Oberfläche nicht „rutscht“. Diese strikte Einschränkung wird

dadurch ersetzt, dass man eine gewisse Bewegung $\dot{z}$ im Kontaktpunkt zulässt, d.h. (1.9) wird modifiziert zu

$$\dot{z} = G(q)\dot{q}. \tag{3.2}$$

Dieses Verletzen der Zwangsbedingung wird allerdings eingeschränkt; aus dem physikalischen Modell heraus wird der Kontakt zwischen Körper und Oberfläche durch ein sogenanntes *Kelvin-Voigt-Material* (siehe [Ber08], S.20) modelliert, eine Parallelschaltung aus Feder und viskoser Dämpfungseinheit. Dies führt zu folgender Bewegungsgleichung für z:

$$\lambda = \frac{c}{\epsilon} z + \frac{d}{\epsilon^\kappa} \dot{z}$$

und nun insgesamt zum System

$$\ddot{q} = M^{-1}(q)\left(F(q,\dot{q}) - G(q)^T\lambda\right) \tag{3.3}$$
$$\dot{z} = G(q)\dot{q} \tag{3.4}$$
$$\lambda = \frac{c}{\epsilon} z + \frac{d}{\epsilon^\kappa} \dot{z}. \tag{3.5}$$

Hierbei sind c, d, ϵ und κ positive reellwertige Parameter, mit denen Feder- und Dämpfungskonstanten des Modells erfasst werden. Die Idee ist, dass für genügend kleines ϵ die Federhärte und die Viskosität so hoch sind, dass Relativbewegung $\dot{z}$ gegen Null gezwungen wird und damit die Zwangsbedingung (1.9) möglichst gering verletzt wird. Gleichzeitig entspricht dieses neue Problem einem Problem vom geringeren Differentiationsindex 1, denn einmaliges Ableiten von (3.4) und (3.5) führt zu (der Übersichtlichkeit halber nun ohne Argumente)

$$\begin{aligned}
\ddot{z} &= \dot{G}\dot{q} + G\ddot{q} \\
&\overset{(3.3)}{=} \dot{G}\dot{q} + GM^{-1}(F - G^T\lambda) \\
\dot{\lambda} &= \frac{c}{\epsilon}\dot{z} + \frac{d}{\epsilon^\kappa}\ddot{z} \\
&\overset{(3.4)}{=} \frac{c}{\epsilon} G\dot{q} + \frac{d}{\epsilon^\kappa}\left(\dot{G}\dot{q} + GM^{-1}(F - G^T\lambda)\right).
\end{aligned} \tag{3.6}$$

Überführt man das System (3.3), (3.4) und (3.6) durch Einführen der Variable $y = (y_1, y_2)^T := (q, \dot{q})^T$ in ein System erster Ordnung, ergibt sich

$$\begin{aligned}
\dot{y}_1 &= y_2 && =: f_1(y,\lambda) \\
\dot{y}_2 &= M^{-1}(y_1)(F(y) - G(y_1)^T\lambda) && =: f_2(y,\lambda) \\
\dot{\lambda} &= \frac{c}{\epsilon} G(y_1)y_2 + \frac{d}{\epsilon^\kappa}\left(\dot{G}(y_1)y_2 + GM^{-1}(F(y) - G(y_1)^T\lambda)\right) && =: g_{\kappa,\epsilon}(y,\lambda) \\
\dot{z} &= G(y_1)y_2 && =: h(y).
\end{aligned}$$

Setzen wir nun $f := (f_1, f_2)^T$, erhalten wir die explizite Differentialgleichung

$$\dot{y} = f(y,\lambda) \tag{3.7}$$
$$\dot{\lambda} = g_{\kappa,\epsilon}(y,\lambda) \tag{3.8}$$
$$\dot{z} = h(y). \tag{3.9}$$

Dieses bezeichnen wir im Folgenden als *regularisiertes Problem* der zu Grunde liegenden ursprünglichen Differential-Algebraischen-Gleichung (erhalten durch analoge Umformung auf ein System erster Ordnung)

$$\begin{aligned}
\dot{y} &= f(y,\lambda) \\
0 &= h(y).
\end{aligned}$$

Hierbei fällt auf, dass bei der expliziten Differentialgleichung die Bewegungsgleichungen für y und λ unabhängig von z sind, d.h. wenn nur nach Lösungen für y und λ gesucht wird (den eigentlich für das Modellproblem relevanten verallgemeinerten Koordinaten), genügt es, sich auf das System (3.7), (3.8) zu beschränken.
Streng genommen besitzt das System (3.3), (3.4) und (3.5) also Differentiationsindex 1, sofern man direkt der Definition des Differentiationsindex folgt und das System in eine explizite Differentialgleichung bezüglich aller Variablen y, z und λ umformen will. Eliminiert man allerdings λ durch (3.5) und somit $\dot{z}$ durch $h(y)$, also $\lambda = \frac{c}{\epsilon}z + \frac{d}{\epsilon^\kappa}h(y)$, erhält man das System

$$\dot{y}_1 = y_2 \tag{3.10}$$

$$\dot{y}_2 = M^{-1}(y_1)\left(F(y) - G(y_1)^T\left(\frac{c}{\epsilon}z + \frac{d}{\epsilon^\kappa}h(y)\right)\right) \tag{3.11}$$

$$\dot{z} = h(y). \tag{3.12}$$

Damit ist das regularisierte System ohne Differentiation äquivalent zu einer expliziten Differentialgleichung, kann also als System vom Index 0 aufgefasst werden.
Insgesamt hat man also bei der Regularisierung nach [Dep13] die Wahl zwischen der Lösung eines Index 1 Problems und der Lösung einer expliziten Differentialgleichung bezüglich der Variablen y und λ oder y und z. Da bei entsprechender Wahl der Anfangswerte für die verwendeten Variablen letztlich alle Varianten analytisch äquivalent sind, kann man sich für den jeweiligen Verwendungszweck die geeignete Variante wählen. So bietet sich für numerisches Lösen etwa (3.10), (3.11), (3.12) an (geringerer Index), für die Untersuchung des Grenzwertes $\epsilon \to 0$ sind je nach Ansatz evtl. die anderen Varianten vorteilhafter.
Es verbleibt nun allerdings die Frage, in welchem Zusammenhang die Lösung des regularisierten Problems mit der Lösung des ursprünglichen Problems steht. Wünschenswert wäre, dass die regularisierte Lösung in einer ϵ-Umgebung zur nichtregularisierten Lösung steht, dies jedoch zu beweisen, erweist sich als unerwartet problematisch.
Zur Untersuchung der Konvergenz der regularisierten Lösung gegen die nichtregularisierte Lösung für $\epsilon \to 0$ wird sich hier – wie auch in [Dep13] – auf die beiden Spezialfälle $\kappa \in \left\{1, \frac{1}{2}\right\}$ beschränkt. Dies hat teils historische, teils praktische Gründe.
So ist der Fall $\kappa = 1$ bereits relativ gut untersucht und wurde bisher zur Modellierung von derartigen Reibungsproblemen herangezogen (vgl. [Lö85]); insbesondere vereinfacht sich in diesem Fall das Differentialgleichungssystem erheblich, da nur noch eine einzige $\epsilon-$Potenz vorkommt und letztlich die Standardform* eines singulär gestörten Problems vorliegt.
Der Fall $\kappa = \frac{1}{2}$ andererseits ist insofern interessant, als dass er aus physikalischer Sicht einer Art aperiodischem Grenzfall der Schwingung/Dämpfung im Auflagepunkt entspricht, damit also die Verletzung der Zwangsbedingung in irgendeiner Weise am „schnellsten“ oder „besten“ unterdrücken sollte. Allerdings hat man es in diesem Fall dann mit verschiedenen ϵ-Potenzen innerhalb der Differentialgleichung zu tun, was einen Konvergenzbeweis mit herkömmlichen Mitteln erschwert.
Betrachten wir die beiden Fälle nun getrennt und beginnen mit dem Fall $\kappa = 1$.

3.2 Sonderfall $\kappa = 1$, *Strong Damping*

Betrachten wir das System (3.7), (3.8). Für den Fall $\kappa = 1$ kann man (3.8) mit ϵ multiplizieren und erhält

$$\dot{y} = f(y, \lambda) \tag{3.13}$$

$$\epsilon\dot{\lambda} = cG(y_1)y_2 + d\left(\dot{G}(y_1)y_2 + G(y_1)M^{-1}(y)(F(y) - G(y_1)^T\lambda)\right) \qquad =: g(y, \lambda). \tag{3.14}$$

*Die in der Literatur behandelten Formen singulär gestörter Differentialgleichungen haben fast ausnahmslos die Form $\dot{y} = f(y, z)$, $\epsilon\dot{z} = g(y, z)$, bestenfalls noch mit einer zusätzlichen glatten Abhängigkeit von ϵ, vergleiche [Hop68].

Dies entspricht einer verbreiteten Grundform von singulär gestörten Problemen zum *zugrundeliegenden ungestörten Problem* (der Differential-Algebraischen-Gleichung), also dem Fall $\epsilon = 0$

$$\dot{y} = f(y, \lambda) \tag{3.15}$$
$$0 = g(y, \lambda). \tag{3.16}$$

Es liegt also nahe, entsprechende Resultate für singulär gestörte Probleme zu benutzen. Insbesondere bietet sich hier Theorem 6 an.
Die obige Differential-Algebraische-Gleichung besitzt (bei geeigneten Anfangswerten) eine Lösung, denn es gilt

$$g_\lambda(y, \lambda) = -dG(y)M^{-1}(y)G(y)^T.$$

Aus den Modellannahmen folgt, dass G vollen Rang hat und M positiv definit ist. Damit ist g_λ invertierbar. Diese Tatsache ist auch später noch wichtig. Damit ist insbesondere λ als Funktion von y festgelegt, denn $0 = g(y, \lambda)$ lässt sich nach λ auflösen:

$$\begin{aligned} & 0 = cG(y_1)y_2 + d\left(\dot{G}(y_1)y_2 + G(y_1)M^{-1}(y)\left(F(y) - G(y_1)^T\lambda\right)\right) \\ & \quad \Leftrightarrow \\ & \lambda = \left(dG(y_1)M^{-1}(y)G(y_1)^T\right)^{-1} c\left(G(y_1)y_2 + d\left(\dot{G}(y_1)y_2 + G(y_1)M^{-1}(y)F(y)\right)\right) \end{aligned} \tag{3.17}$$

Dies kann man in die Differentialgleichung für y einsetzen und erhält eine lokal Lipschitz-stetige Differentialgleichung, vorausgesetzt g, f etc. sind glatt genug. Also besitzt die Differential-Algebraische-Gleichung bei geeigneten Anfangswerten eine auf einem beschränkten Intervall eindeutige Lösung.

Nun kann man als ersten Ansatz eine Reihenentwicklung in ϵ machen, d.h. man sucht zunächst Lösungen der Form

$$\begin{pmatrix} y \\ \lambda \end{pmatrix} = \sum_{i=0}^{\infty} \epsilon^i \begin{pmatrix} y_i \\ \lambda_i \end{pmatrix},$$

mit der Absicht, eine Lösung zu konstruieren, deren Anteile zur nullten ϵ-Potenz die Lösung der Differential-Algebraischen-Gleichung sind und der Rest für $\epsilon \to 0$ verschwindet. Einsetzen in (3.13), (3.14) und Taylorentwicklung um die Punkte y_0 bzw. λ_0 liefert

$$\begin{aligned} \sum_{i=0}^{\infty} \epsilon^i \dot{y}_i &= f(y_0, \lambda_0) + f_y(y_0, \lambda_0) \sum_{i=1}^{\infty} \epsilon^i y_i + f_\lambda(y_0, \lambda_0) \sum_{i=1}^{\infty} \epsilon^i \lambda_i + \mathcal{O}(\epsilon^2)(f_{yy} \dots) \\ \sum_{i=0}^{\infty} \epsilon^{i+1} \dot{\lambda}_i &= g(y_0, \lambda_0) + g_y(y_0, \lambda_0) \sum_{i=1}^{\infty} \epsilon^i y_i + g_\lambda(y_0, \lambda_0) \sum_{i=1}^{\infty} \epsilon^i \lambda_i + \mathcal{O}(\epsilon^2)(g_{yy} \dots). \end{aligned}$$

Koeffizientenvergleich bezüglich der ϵ-Potenzen liefert für die nullte Potenz:

$$\dot{y}_0 = f(y_0, \lambda_0) \tag{3.18}$$
$$0 = g(y_0, \lambda_0) \tag{3.19}$$

y_0 und λ_0 entsprechen damit tatsächlich Lösungen der unregularisierten Differential-Algebraischen-Gleichung (3.15), (3.16), die nach oben eindeutig existieren. Insbesondere legt der Anfangswert für y_0 den Anfangswert für λ_0 nach (3.17) fest.
Die erste Potenz liefert

$$\begin{aligned} \dot{y}_1 &= f_y(y_0, \lambda_0)y_1 + f_\lambda(y_0, \lambda_0)\lambda_1 \\ \dot{\lambda}_0 &= g_y(y_0, \lambda_0)y_1 + g_\lambda(y_0, \lambda_0)\lambda_1. \end{aligned}$$

Da g_λ wie bereits begründet invertierbar ist, lässt sich hier die zweite Zeile nach λ_1 auflösen und liefert bei Kenntnis von y_0 und λ_0 insgesamt eine explizite lineare Differentialgleichung für y_1, die nach Modellannahmen für f und g eine eindeutige Lösung besitzt. Dies lässt sich iterativ fortsetzen, wobei jetzt die höheren Ableitungen der Taylorentwicklung ebenfalls einen Beitrag liefern. Da hier jedoch die Argumente und damit insbesondere die ϵ-Potenzen potenziert werden, kommen in den Gleichungen nur y- und λ-Terme von geringerem Index dazu, welche in der Iteration nach Potenz bereits davor schon bestimmt wurden, d.h. für die zweite Potenz kommen nur Terme mit y_1 und λ_1 dazu, etc. Allgemein liefert der Koeffizientenvergleich

$$\begin{aligned}\dot{y}_i &= f_y(y_0,\lambda_0)y_i + f_\lambda(y_0,\lambda_0)\lambda_i + \phi_i(y_1,\dots,y_{i-1},\lambda_1,\dots,\lambda_{i-1})\\ \dot{\lambda}_{i-1} &= g_y(y_0,\lambda_0)y_i + g_\lambda(y_0,\lambda_0)\lambda_i + \psi_i(y_1,\dots,y_{i-1},\lambda_1,\dots,\lambda_{i-1}),\end{aligned}$$

was durch die Invertierbarkeit von g_λ wieder nach λ_i aufgelöst werden kann und eine explizite lineare Differentialgleichung für y_i ergibt.
Dieser Ansatz ist allerdings noch nicht hinreichend, wenn man Freiheiten bei der Wahl der Anfangswerte für λ fordert. Bereits bei der nullten Potenz ist aufgrund der im Allgemeinen nichttrivialen Funktion g erkennbar, dass nicht jeder beliebige Anfangswert y_0, λ_0 eingesetzt werden kann, da $g(y_0,\lambda_0) = 0$ gelten muss. Auch bei den weiteren Koeffizientenvergleichen wird λ_i nicht durch eine Differentialgleichung bestimmt, sondern ergibt sich durch algebraische Rechnung explizit aus y_i und Funktionen von niedrigerem Index. Insgesamt sind so zwar die Anfangswerte für die y_i frei wählbar, jedoch sind damit die Anfangswerte für die λ_i fest vorgegeben, damit liegt kein allgemeiner Ansatz für einen beliebigen Anfangswert (in Umgebung der exakten Lösung der Differential-Algebraischen-Gleichung) vor. Abhilfe hierfür schafft eine Erweiterung des Ansatzes durch Terme, die einerseits den obigen Reihenansatz in jeder ϵ-Potenz ergänzen, andererseits zeitlich exponentiell fallen, also auf das gesamte Lösungsverhalten nach einer kurzen *Einschwingzeit* keinen Einfluss mehr haben. In der Literatur wird dies auch als *Boundary Layer* bezeichnet. (vgl. [Hai10], S 389ff). Hierfür ist allerdings noch etwas Vorarbeit erforderlich.
Für die folgenden Schritte und Folgerungen werden einige Sätze und Definitionen aus [HNWXX] benötigt, die insbesondere für den Beweis von Theorem 6 notwendig sind. Grundsätzlich sind dies Werkzeuge, um Abschätzungen für *differentialgleichungsähnliche* Strukturen zu erhalten. Diese werden benötigt, um aus sich ergebenden Gleichungen mit Normen, also nicht notwendigerweise differenzierbaren Funktionen, Abschätzungen gegenüber gewöhnlichen Differentialgleichungen mit differenzierbarer Lösung zu erhalten.

Eine erste Definition stellt eine Verallgemeinerung des Ableitungsbegriffes für nicht notwendig differenzierbare Funktionen dar, die sogenannte *Dini-Ableitung*. Man verwendet hierfür den Limes Superior respektive Inferior des Differenzenquotienten.

Definition 3.1 *Dini-Ableitung*
Sei $y(t)$ eine stetige Funktion, dann bezeichnen

$$\begin{aligned}D_+y(t) &:= \liminf_{h\to 0^+} \frac{y(t+h)-y(t)}{h}\\ D^+y(t) &:= \limsup_{h\to 0^+} \frac{y(t+h)-y(t)}{h}\end{aligned}$$

die Dini-Ableitungen von $y(t)$.

Da bei differenzierbaren Funktionen der Limes des Differenzenquotienten existiert und eindeutig ist, gilt dann insbesondere $D_+y(t) = D^+y(t) = \dot{y}(t)$.
Weiterhin wird im Folgenden die *logarithmische Norm* benötigt.

Definition 3.2 *Definition 10.4 aus [HNWXX]*
Sei A eine quadratische Matrix und $\|\cdot\|$ die euklidische Norm, dann bezeichnet

$$\begin{aligned}\mu(A) :&= \lim_{h\to 0^+} \frac{\|\mathbb{1} + hA\| - 1}{h} \\ &= \max\left\{\lambda \;\middle|\; \lambda \text{ Eigenwert von } \frac{1}{2}(A + A^T)\right\}\end{aligned}$$

die logarithmische Norm von A.

Die in der Definition benutzten Gleichheiten folgen aus Theorem 10.5 aus [HNWXX].

Beweis: Da in der euklidischen Norm gilt

$$\|A\|^2 = \text{ maximaler Eigenwert von } A^T A,$$

folgt

$$\begin{aligned}\lim_{h\to 0^+} \frac{\|\mathbb{1} + hA\| - 1}{h} &= \lim_{h\to 0^+} \frac{1}{h}\left(\sqrt{\lambda_{\max}\left((\mathbb{1} + hA)^T(\mathbb{1} + hA)\right)} - 1\right) \\ &= \lim_{h\to 0^+} \frac{1}{h}\left(\sqrt{\lambda_{\max}\left(\mathbb{1} + h\left(A + A^T\right) + h^2 A^T A\right)} - 1\right) \\ &= \lim_{h\to 0^+} \frac{1}{h}\left(\sqrt{1 + h\lambda_{\max}\left((A + A^T) + hA^T A\right)} - 1\right) \\ &\overset{\text{Taylor}}{=} \lim_{h\to 0^+} \frac{1}{h}\left(1 + h\frac{1}{2}\lambda_{\max}\left(A + A^T\right) + \mathcal{O}(h^2) - 1\right) \\ &= \lim_{h\to 0^+} \frac{1}{2}\lambda_{\max}\left(A + A^T\right) + \mathcal{O}(h) \\ &= \frac{1}{2}\lambda_{\max}\left(A + A^T\right)\end{aligned}$$

□

Weiterhin sind noch zwei Lemmata später von Bedeutung, das erste ist eine Verallgemeinerung des Lemmas 10.1 aus [HNWXX]

Lemma 3.3 *Übung 10.7 aus [HNWXX]*
Es seien zwei Funktionen u und m aus $C(\mathbb{R}, \mathbb{R}^n)$, sowie eine Funktion $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ gegeben. Weiterhin seien auf dem nach links abgeschlossenen Intervall $[0, T)$, wobei T nicht notwendig endlich sein muss, die Bedingungen

1. $D_+ m_i(t) \leq g_i(t, m(t))$
2. $D_+ u_i(t) > g_i(t, u(t))$
3. $u(0) = m(0)$
4. *Aus* $s_j \leq \tilde{s}_j$ *folgt* $g_i(t, s_1, \ldots, s_n) \leq g_i(t, s_1, \ldots, \tilde{s}_j, \ldots, s_n),\ j \neq i$

erfüllt, jeweils für $i = 1, \ldots, n$, wobei u_i und m_i die i-ten Komponenten von u und m sind. Dann gilt für alle $i = 1, \ldots, n$ und für alle $t \in [0, T)$

$$m_i(t) \leq u_i(t). \tag{3.20}$$

Beweis: Angenommen, (3.20) werde für mindestens eine Komponente verletzt. Da u und m nach Voraussetzung stetige Funktionen sind, gilt diese Verletzung auf einem nach links offenen Intervall $(\tilde{t}, \tilde{t}_2) \subset [0, T)$, nach links begrenzt durch ein $\tilde{t}$, an dem für diese Komponente Gleichheit in (3.20) gilt. Für den Beweis wird nun der kleinste dieser Werte betrachtet, d.h.

$$t_0 := \min_{i=1,\dots,n} \max \left\{ \tilde{t} \in [0,T) \,\middle|\, \forall\, 0 \le t < \tilde{t} \text{ gilt } m_i(t) \le u_i(t) \right\}.$$

Insbesondere gilt also auf dem Intervall $[0, t_0]$

$$m_i(t) \le u_i(t) \,\forall\, 0 \;\le\; t \;\le\; t_0. \tag{3.21}$$

Die Existenz von t_0 folgt, da $\left\{ \tilde{t} \in [0,T) \,\middle|\, \forall\, 0 \le t \le \tilde{t} \text{ gilt } m_i(t) \le u_i(t) \right\}$ nach Bedingung 3 mindestens aus $t = 0$ besteht, außerdem besteht diese Menge dank der Stetigkeitsbedingung an m und u aus einem abgeschlossenen Intervall, also existiert auch ein Maximum. O.B.d.A. werde das Minimum bei $i = 1$ angenommen. Es gilt also $m_1(t_0) = u_1(t_0)$ und $m_1(t_0 + h) > u_1(t_0 + h)$ für genügend kleines $h > 0$. Damit folgt die Äquivalenz

$$\begin{gathered} m_1(t_0+h) - m_1(t_0) > u_1(t_0+h) - u_1(t_0) \\ \Leftrightarrow \\ \frac{m_1(t_0+h) - m_1(t_0)}{h} > \frac{u_1(t_0+h) - u_1(t_0)}{h} \end{gathered}$$

und somit

$$D_+ m_1(t_0) \ge D_+ u_1(t_0).$$

Gleichzeitig muss aber gelten

$$\begin{aligned} D_+ m_1(t_0) &\overset{\text{Bed. 1}}{\le} g_1(t_0, m_1(t_0), \dots, m_n(t_0)) = g_1(t_0, u_1(t_0), \underbrace{m_2(t_0) \dots, m_n(t_0)}_{\text{je } \le u_i(t_0) \text{ nach} 3.21}) \\ &\overset{\text{Bed. 4}}{\le} g_1(t_0, u(t_0)) \overset{\text{Bed. 2}}{<} D_+ u_1(t_0), \end{aligned}$$

ein Widerspruch. □

Ein weiteres Lemma wird direkt im Anschluss dazu benötigt, um die Erweiterung des Reihenansatzes für das Modellproblem zu rechtfertigen.

Lemma 3.4 *Theorem 10.6 aus [HNWXX]*
Seien $f(t,y) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, $y(t)$ und $v(t) : \mathbb{R} \to \mathbb{R}^n$ hinreichend glatte Funktionen, wobei $\dot{y}(t) = f(t, y(t))$, sowie $l(t)$ und $\delta(t)$ in $C(\mathbb{R})$ stetige Funktionen auf $0 \le t$ sind. Es gelten weiterhin die Bedingungen:

$$\begin{aligned} \mu\left(f_y(t,\eta)\right) &\le l(t) \qquad \forall \eta \in [y(t), v(t)] \\ \|\dot{v}(t) - f(t, v(t))\| &\le \delta(t) \\ \|y(0) - v(0)\| &\le C \end{aligned}$$

mit einer positiven Konstanten C. Dann gilt mit $L(t) := \int_0^t l(x)dx$ und für $t \ge 0$:

$$\|y(t) - v(t)\| \le e^{L(t)} \left(C + \int_0^t e^{-L(x)} \delta(x) dx \right).$$

Beweis: Da y und v nach Voraussetzung stetige Funktionen sind, ist für jedes fest gewählte $t > 0$ die Menge $\{(y(s), v(s)) \mid s \in [0,t]\}$ kompakt, damit folgt aus der Glattheit von f, dass $\frac{\partial f_i}{\partial y_j}(t,\eta)$ für alle i,j und $\eta \in [y(t), v(t)]$ beschränkt ist. Damit ist die Matrixdarstellung von $\frac{\partial f}{\partial y}$ in der Definition 3.2 wohldefiniert. Weiterhin gilt für $m(t) := \|y(t) - v(t)\| \geq 0$ mit $h > 0$

$$\begin{aligned}
m(t+h) &= \|y(t+h) - v(t+h)\| \\
&\overset{\text{Taylor}}{=} \|y(t) - v(t) + h\,(\dot{y}(t) - \dot{v}(t))\| + \mathcal{O}(h^2) \\
&\leq \|y(t) - v(t) + h\,(f(t,y(t)) - f(t,v(t)))\| + h\delta(t) + \mathcal{O}(h^2) \\
&\overset{\text{zws}}{=} \left\| y(t) - v(t) + h\,(y(t) - v(t))\,\frac{\partial f}{\partial y}(t,\eta) \right\| + h\delta(t) + \mathcal{O}(h^2) \quad \text{für ein } \eta \in [y(t), v(t)] \\
&\leq \max_{\eta \in [y(t),v(t)]} \left\| \mathbb{1} - h\frac{\partial f}{\partial y}(t,\eta) \right\| m(t) + h\delta(t) + \mathcal{O}(h^2).
\end{aligned}$$

Nach Subtraktion von $m(t)$ und Division durch h, liefert der Limes:

$$\begin{aligned}
D_+ m(t) &= \lim_{h \to 0^+} \frac{m(t+h) - m(t)}{h} \\
&\leq \lim_{h \to 0^+} \max_{\eta \in [y(t),v(t)]} \underbrace{\frac{\left\| \mathbb{1} - h\frac{\partial f}{\partial y}(t,\eta) \right\| - 1}{h}}_{\mu\left(\frac{\partial f}{\partial y}(t,\eta)\right) + \mathcal{O}(h) \text{ nach Def. und Bew. 3.2}} m(t) + \delta(t) + \mathcal{O}(h) \\
&\leq l(t) m(t) + \delta(t) \\
&=: g(t, m(t))
\end{aligned}$$

Als Analogon zu dieser Dini-Differentialgleichung betrachte die gewöhnliche Differentialgleichung

$$\dot{u}(t) = g(t, u(t)) + \epsilon$$

mit $\epsilon > 0$ und Anfangswert $u(0) = m(0) =: C$. Die Lösung dieser inhomogenen linearen Differentialgleichung ergibt sich durch die Variation-der-Konstanten Formel durch

$$u(t) = e^{L(t)} \left(C + \int_0^t e^{-L(x)} (\delta(x) + \epsilon) dx \right),$$

wobei $L(t)$ wie oben definiert ist. Für $\epsilon \to 0$ ist dies die zu zeigende Ungleichung, allerdings für u statt für m. Somit bleibt zu zeigen, dass $m(t) \leq u(t)\ \forall t > 0$. Hierfür bietet sich die eindimensionale Version von Lemma 3.3 an (letztlich Theorem 10.1 aus [HNWXX]). Zu zeigen sind wegen der Eindimensionalität nur die Bedingungen 1,2 und 3:

1. $D_+ m(t) \leq g(t, m(t))$ gilt nach obiger Herleitung der Dini-Differentialgleichung.
2. Da $u(t)$ als Lösung einer gewöhnlichen Differentialgleichung differenzierbar ist, stimmen Dini- und gewöhnliche Ableitung überein, damit $D_+ u(t) = \dot{u}(t) = g(t, u(t)) + \epsilon > g(t, u(t))$
3. Der Anfangswert wurde entsprechend gewählt.

Damit ist Lemma 3.3 anwendbar und die Behauptung gezeigt. □

Mit diesen Hilfsmitteln ist es nun möglich, den obigen Potenzreihenansatz für die Lösung von (3.13), (3.14) für allgemeine Anfangswerte zu erweitern, d.h. mit einem Ansatz der Form

$$\begin{pmatrix} y(t) \\ \lambda(t) \end{pmatrix} = \sum_{i=0}^{\infty} \epsilon^i \begin{pmatrix} y_i(t) \\ \lambda_i(t) \end{pmatrix} + \sum_{i=0}^{\infty} \epsilon^i \begin{pmatrix} \epsilon \eta_j(\frac{t}{\epsilon}) \\ \zeta_j(\frac{t}{\epsilon}) \end{pmatrix},$$

wobei die y_i und λ_i dieselben wie im ersten Potenzreihenansatz sind und wie gewünscht $\|\eta_i(t)\| \leq C_i e^{-\kappa_i t}$ und $\|\zeta_i(t)\| \leq C_i e^{-\kappa_i t}$ mit $\kappa_i, C_i > 0$ erfüllen sollen. Einsetzen in die Differentialgleichung (3.13),(3.14) liefert

$$\underbrace{\sum_{i=0}^{\infty} \epsilon^i \dot{y}_i(t)}_{f\left(\sum_{i=0}^{\infty}\epsilon^i y_i(t), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(t)\right)} + \sum_{i=0}^{\infty} \epsilon^i \dot{\eta}_j\left(\frac{t}{\epsilon}\right) = f\left(\sum_{i=0}^{\infty} \epsilon^i y_i(t) + \epsilon^{i+1}\eta_i\left(\frac{t}{\epsilon}\right), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(t) + \epsilon^i \zeta_i\left(\frac{t}{\epsilon}\right)\right) \tag{3.22}$$

$$\underbrace{\epsilon \sum_{i=0}^{\infty} \epsilon^i \dot{\lambda}_i(t)}_{g\left(\sum_{i=0}^{\infty}\epsilon^i y_i(t), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(t)\right)} + \sum_{i=0}^{\infty} \epsilon^i \dot{\zeta}_j\left(\frac{t}{\epsilon}\right) = g\left(\sum_{i=0}^{\infty} \epsilon^i y_i(t) + \epsilon^{i+1}\eta_i\left(\frac{t}{\epsilon}\right), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(t) + \epsilon^i \zeta_i\left(\frac{t}{\epsilon}\right)\right). \tag{3.23}$$

Das Argument $\frac{t}{\epsilon}$ der neu hinzugekommenen Ansatzfunktionen ist zwar hilfreich dabei, diese Ansatzfunktionen schnell gegen 0 konvergieren zu lassen, ist aber hinderlich für einen Koeffizientenvergleich wie oben. Daher bietet sich die Substitution $\xi = \frac{t}{\epsilon}$ an. Setzt man diese in (3.22), (3.23) ein, folgt

$$\sum_{i=0}^{\infty} \epsilon^i \dot{\eta}_j(\xi) = f\left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi) + \epsilon^{i+1}\eta_i(\xi), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(\epsilon\xi) + \epsilon^i \zeta_i(\xi)\right) - f\left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(\epsilon\xi)\right)$$
$$\sum_{i=0}^{\infty} \epsilon^i \dot{\zeta}_j(\xi) = g\left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi) + \epsilon^{i+1}\eta_i(\xi), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(\epsilon\xi) + \epsilon^i \zeta_i(\xi)\right) - g\left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi), \sum_{i=0}^{\infty}\epsilon^i \lambda_i(\epsilon\xi)\right).$$

Taylorentwicklung der Funktionen auf der rechten Seite um die Punkte $(y_0(0), \lambda_0(0)+\zeta_0(\xi))$ respektive $(y_0(0), \lambda_0(0))$ ergibt

$$\begin{aligned}
&\sum_{i=0}^{\infty} \epsilon^i \dot{\eta}_j(\xi) \\
&= f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) + f_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi) - y_0(0) + \sum_{i=0}^{\infty} \epsilon^{i+1}\eta_i(\xi)\right) \\
&\quad + f_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \epsilon^i \lambda_i(\epsilon\xi) - \lambda_0(0) + \sum_{i=1}^{\infty} \epsilon^i \zeta_i(\xi)\right) - f(y_0(0), \lambda_0(0)) \\
&\quad - f_y(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi) - y_0(0)\right) - f_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \epsilon^i \lambda_i(\epsilon\xi) - \lambda_0(0)\right) \\
&\quad + \mathcal{O}\left(\epsilon^2 (f_{yy} \dots)\right) \\
&= f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - f(y_0(0), \lambda_0(0)) \\
&\quad + f_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(y_i(0) + \epsilon\xi \dot{y}_i(0)\right) + \dot{y}_0(0)\epsilon\xi + \sum_{i=0}^{\infty} \epsilon^{i+1}\eta_i(\xi)\right) \\
&\quad + f_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(\lambda_i(0) + \epsilon\xi \dot{\lambda}_i(0)\right) + \epsilon\xi \dot{\lambda}_0(0) + \sum_{i=1}^{\infty} \epsilon^i \zeta_i(\xi)\right) \\
&\quad - f_y(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(y_i(0) + \epsilon\xi \dot{y}_i(0)\right) + \dot{y}_0(0)\epsilon\xi\right) \\
&\quad - f_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \epsilon^i \left(\lambda_i(0) + \epsilon\xi \dot{\lambda}_i(0)\right) + \epsilon\xi \dot{\lambda}_0(0)\right) + \mathcal{O}\left(\epsilon^2 (f_{yy} \dots)\right)
\end{aligned} \tag{3.24}$$

$$
\begin{aligned}
&\sum_{i=0}^{\infty} \epsilon^i \dot{\zeta}_j(\xi) \\
&= g(y_0(0), \lambda_0(0) + \zeta_0(\xi)) + g_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi) - y_0(0) + \sum_{i=0}^{\infty} \epsilon^{i+1} \eta_i(\xi) \right) \\
&\quad + g_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \epsilon^i \lambda_i(\epsilon\xi) - \lambda_0(0) + \sum_{i=1}^{\infty} \epsilon^i \zeta_i(\xi) \right) - g(y_0(0), \lambda_0(0)) \\
&\quad - g_y(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \epsilon^i y_i(\epsilon\xi) - y_0(0) \right) - g_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \epsilon^i \lambda_i(\epsilon\xi) - \lambda_0(0) \right) \qquad (3.25) \\
&\quad + \mathcal{O}\left(\epsilon^2 (g_{yy} \dots)\right) \\
&= g(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - g(y_0(0), \lambda_0(0)) \\
&\quad + g_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(y_i(0) + \epsilon\xi \dot{y}_i(0) \right) + \dot{y}_0(0)\epsilon\xi + \sum_{i=0}^{\infty} \epsilon^{i+1} \eta_i(\xi) \right) \\
&\quad + g_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(\lambda_i(0) + \epsilon\xi \dot{\lambda}_i(0) \right) + \epsilon\xi \dot{\lambda}_0(0) + \sum_{i=1}^{\infty} \epsilon^i \zeta_i(\xi) \right) \\
&\quad - g_y(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(y_i(0) + \epsilon\xi \dot{y}_i(0) \right) + \dot{y}_0(0)\epsilon\xi \right) \\
&\quad - g_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \epsilon^i \left(\lambda_i(0) + \epsilon\xi \dot{\lambda}_i(0) \right) - \dot{\lambda}_0(0)\epsilon\xi \right) + \mathcal{O}\left(\epsilon^2 (g_{yy} \dots)\right). \qquad (3.26)
\end{aligned}
$$

Hier kann man wieder ϵ-Koeffizienten vergleichen und erhält für die nullte Potenz:

$$\dot{\eta}_0(\xi) = f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - f(y_0(0), \lambda_0(0)) =: \tilde{f}(\xi, \zeta_0) \qquad (3.27)$$

$$\dot{\zeta}_0(\xi) = g(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - g(y_0(0), \lambda_0(0)). \qquad (3.28)$$

Setzt man nun die Modellannahme $\mu(g_\lambda(y, \lambda)) \leq -1$ in einer Umgebung um die Lösung voraus (mit entsprechender Skalierung genügt jede negative Schranke, wie später noch gezeigt wird), kann man Theorem 3.4 auf (3.28) für die λ-Komponente der Differentialgleichung anwenden. Mit der Annahme $\mu(g_\lambda(y, \lambda)) \leq -1$, sowie da nach der Herleitung der y_i und λ_i-Funktionen gilt $\dot{\zeta}_0(\xi) - g(y_0(0), \lambda_0(0) + \zeta_0(\xi)) = -g(y_0(0), \lambda_0(0)) = 0$, ist Theorem 3.4 anwendbar. Unter Verwendung der Nullfunktion als Vergleichsfunktion (Im Theorem mit y bezeichnet, damit $C = \|\zeta_0(0)\|$) liefert dieses

$$\|\zeta_0(\xi) - 0\| = \|\zeta_0(\xi)\| \leq \|\zeta_0(0)\| \, e^{-\xi}.$$

Damit erfüllt ζ_0 die geforderte Eigenschaft der schnellen Konvergenz gegen 0 im Sinne von $\|\zeta_0(\xi)\| \leq \|\zeta_0(0)\| \, e^{-\xi}$. Da die rechte Seite von (3.27) nicht von η_0 abhängt, also $\frac{\partial \tilde{f}(\xi, \zeta_0)}{\partial \eta_0} = 0$, erfüllt $\tilde{f}$ bezüglich der η_0 Komponente natürlich jede Lipschitz-Bedingung und nach dem Satz von Picard und Lindelöff hat somit (3.27) eine eindeutige Lösung $\eta_0(\xi)$ in Abhängigkeit eines Anfangswertes. Weiterhin ist $\| \tilde{f} \|$ exponentiell fallend, da

$$
\begin{aligned}
\| \tilde{f}(\xi) \| &= \|f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - f(y_0(0), \lambda_0(0))\| \\
&\overset{\text{lip}}{\leq} C \underbrace{\|\zeta_0(\xi)\|}_{\leq \|\zeta_0(0)\| e^{-\xi}} \\
&\leq C \, \|\zeta_0(0)\| \, e^{-\xi}
\end{aligned}
$$

mit einer Lipschitz-Konstante $C > 0$ und ist damit auf $\mathbb{R}$ integrierbar. Somit ist $\eta_0(\xi)$ zunächst einmal beschränkt mit eindeutigem Grenzwert für $\xi \to \infty$. Fordert man nun wie oben, dass $\eta_0(\xi)$ exponentiell

gegen 0 fällt, folgt zumindest, dass $\eta_0(\infty) = 0$ ist. Insgesamt hat man

$$\begin{aligned}\eta_0(\xi) &= \underbrace{\eta_0(\infty)}_{=0} + \int_{\infty}^{\xi} \tilde{f}(x)dx \\ &= -\int_{\xi}^{\infty} \tilde{f}(x)dx. \end{aligned} \tag{3.29}$$

Hierdurch ist $\eta_0(\xi)$ und insbesondere $\eta_0(0)$ eindeutig bestimmt und man kann die Norm von $\eta_0(\xi)$ abschätzen durch

$$\begin{aligned}\|\eta_0(\xi)\| &\leq \int_{\xi}^{\infty} \left\|\tilde{f}(x)\right\| dx \\ &\overset{\text{s.o.}}{\leq} C \|\zeta_0(0)\| \int_{\xi}^{\infty} e^{-x} dx \\ &= C \|\zeta_0(0)\| e^{-\xi}. \end{aligned}$$

Die Funktion η_0 ist damit wie vorausgesetzt abschätzbar durch $\|\eta_0(\xi)\| \leq C_0 e^{-\kappa_0 \xi}$ mit positiven Konstanten κ_0 und C_0, die so gewählt wurden, dass sie die Abschätzung sowohl für η_0 und ζ_0 erfüllen, insbesondere $C_0 \geq \min\{C\zeta_0(0), \zeta_0(0)\}$.
Die erste Potenz des Koeffizientenvergleichs in ϵ von (3.24) und (3.26) liefert

$$\begin{aligned}\dot{\eta}_1(\xi) &= f_y\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right)\eta_0(\xi) + f_\lambda\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right)\zeta_1(\xi) \\ &\quad + \left(y_1(0) + \xi \dot{y}_0(0)\right) \underbrace{\left(f_y\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right) - f_y(y_0(0), \lambda_0(0))\right)}_{\mathcal{O}(\zeta_0(\xi))} \\ &\quad + \left(\lambda_1(0) + \xi \dot{\lambda}_0(0)\right) \underbrace{\left(f_\lambda\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right) - f_y(y_0(0), \lambda_0(0))\right)}_{\mathcal{O}(\zeta_0(\xi))} \end{aligned} \tag{3.30}$$

$$\begin{aligned}\dot{\zeta}_1(\xi) &= g_y\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right)\eta_0(\xi) + g_\lambda\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right)\zeta_1(\xi) \\ &\quad + \left(y_1(0) + \xi \dot{y}_0(0)\right) \underbrace{\left(g_y\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right) - g_y(y_0(0), \lambda_0(0))\right)}_{\mathcal{O}(\zeta_0(\xi))} \\ &\quad + \left(\lambda_1(0) + \xi \dot{\lambda}_0(0)\right) \underbrace{\left(g_\lambda\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right) - g_\lambda(y_0(0), \lambda_0(0))\right)}_{\mathcal{O}(\zeta_0(\xi))}. \end{aligned} \tag{3.31}$$

Da die Funktionen η_0 und ζ_0 bereits bestimmt wurden, ist damit (3.31) eine lineare Differentialgleichung für ζ_1. Bezeichnet man (3.31) als $\dot{\zeta}_1(\xi) = \psi(\xi, \zeta_1)$ und vergleicht dies in Theorem 3.4 mit der Funktion $v = 0$, ergibt sich

$$\mu\left(\frac{\partial \psi(\xi, \nu)}{\partial \nu}\right) = \mu\left(g_\lambda\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right)\right) \overset{\text{n.V.}}{\leq} -1,$$

sogar unabhängig vom Argument ν. Die Abschätzung gilt damit insbesondere in Lösungsumgebung. Außerdem gilt mit den oben gezeigten Abschätzungen für $\eta_0(\xi)$ und $\zeta_0(\xi)$ (beide kleiner gleich $C_0 e^{-\xi}$)

$$\|\dot{v}(0) - \psi(\xi, 0)\| = \|g_y\left(y_0(0), \lambda_0(0) + \zeta_0(\xi)\right)\eta_0(\xi) + \mathcal{O}(\zeta_0(\xi))\| \leq C e^{-\kappa_0 \xi}.$$

und schließlich $\|v(0) - \zeta(0)\| \leq \|\zeta(0)\|$, damit ergibt Theorem 3.4 für den Fall $\kappa_0 \neq 1$

$$\begin{aligned}\|\zeta(\xi)\| &\leq e^{-\xi}\left(\|\zeta(0)\| + C\int_0^{\xi} e^{x} e^{-\kappa_0 x} dx\right) \\ &\leq e^{-\xi}\left(\|\zeta(0)\| + \frac{C}{1-\kappa_0}\left(e^{(1-\kappa_0)\xi} - 1\right)\right) \\ &\leq \hat{C} e^{-\kappa_1 \xi} \end{aligned}$$

bzw. für den Fall $\kappa_0 = 1$

$$\begin{aligned}\|\zeta(\xi)\| &\leq e^{-\xi}\left(\|\zeta(0)\| + C\int_0^\xi e^x e^{-x} dx\right)\\ &\leq e^{-\xi}\left(\|\zeta(0)\| + C\xi\right)\\ &\leq \hat{C}e^{-\kappa_1\xi}\end{aligned}$$

mit positiven Konstanten C und $\hat{C}$. Da damit auch die rechte Seite von (3.30) und deren Ableitung mit $Ce^{-\kappa_0\xi}$ abschätzbar ist, folgt analog zu η_0 die Existenz und Eindeutigkeit einer mit $Ce^{-\kappa_1}$ abschätzbaren Lösung, diese wird wieder durch $\zeta_1(\xi)$ festgelegt und man kann beide abschätzen durch $C_1 e^{\kappa_1\xi}$ mit geeignet gewählten Konstanten C_1 und κ_1. Iterativ folgen die höheren Potenzen aus (3.30), (3.31). Insgesamt kann man nun Anfangswerte für die y_i und ζ_i frei bzw. frei in Umgebung um die Lösung wählen. Damit sind dann die Anfangswerte für λ_i und η_i bestimmt, d.h. wählt man Anfangswerte der Form

$$y(0) = \sum_{i=0}^{\infty} \epsilon^i y_i^0 \tag{3.32}$$

$$\lambda(0) = \sum_{i=0}^{\infty} \epsilon^i \lambda_i^0, \tag{3.33}$$

gilt nach dem oben hergeleiteten, dass

$$\begin{aligned}y_0^0 &= y_0(0)\\ y_j^0 &= y_j(0) + \eta_{j-1}(0)\\ \lambda_j^0 &= \lambda_j(0) + \zeta_j(0).\end{aligned}$$

Ein gegebenes y_0^0 legt $\lambda_0(0)$ nach (3.19) fest, damit wird mit gegebenem λ_0^0 der Wert $\zeta_0(0)$ und hieraus aus (3.29) $\eta_0(0)$ der Wert für $\eta_0(0)$ festgelegt. Damit sind y_0^0 und λ_0^0 in einer Umgebung um die Lösung frei wählbar. Sie bestimmen die einzelnen Anfangswerte und damit die kompletten Ansatzfunktionen vom Index Null eindeutig. Durch analoge Iteration folgen die Anfangswerte der Ansatzfunktionen von höherem Index. Insbesondere ist es mit zur Differential-Algebraischen-Gleichung konsistenten, nicht von ϵ unabhängigen Werten möglich, $\eta_i(0) = \zeta_i(0) = 0$ für alle i zu erhalten.
Damit ist theoretisch eine Lösung des Ausgangsproblems gefunden. Diese ist jedoch in der Praxis schwer zu bestimmen, da für jede ϵ-Potenz des Ansatzes eine eigene Differentialgleichung gelöst werden müsste. Insbesondere sind zwar die Ansatzfunktionen durch $C_i e^{-\kappa_i\xi}$ abschätzbar, jedoch können die κ_i beliebig klein, bzw. die C_i beliebig groß werden. Von mehr praktischem Nutzen ist daher eine Abschätzung für eine Ansatzfunktion in der Form einer endlichen Summe, also nur die ersten N Indizes der soeben bestimmten Lösung. Aussagen hierüber macht das folgende Theorem.

Theorem 6 *Thm. VI.3.2 aus [Hai10]*
Gegeben sei das Problem (3.13), (3.14) mit geeigneten Anfangswerten für y und λ der Form (3.32), (3.33). In einer von ϵ unabhängigen Umgebung um die Lösung $y_0(t), \lambda_0(t)$ des reduzierten Problems (dh. für $\epsilon = 0$) gelte $\mu(g_\lambda) \leq -1$, insbesondere liege y_0^0, λ_0^0 in dieser Umgebung. Dann besitzt (3.13), (3.14) auf einem Zeitintervall $0 \leq t \leq \bar{t}$ für hinreichend kleines ϵ eine eindeutige Lösung der Form

$$y(t) = \sum_{j=0}^{N} \epsilon^j y_j(t) + \epsilon \sum_{j=0}^{N-1} \epsilon^j \eta_j\left(\frac{t}{\epsilon}\right) + \mathcal{O}(\epsilon^{N+1})$$

$$\lambda(t) = \sum_{j=0}^{N} \epsilon^j \lambda_j(t) + \sum_{j=0}^{N} \epsilon^j \zeta_j\left(\frac{t}{\epsilon}\right) + \mathcal{O}(\epsilon^{N+1}).$$

Hierbei sind die Koeffizientenfunktionen $y_j, \lambda_j, \eta_j, \zeta_j$ unabhängig von ϵ, außerdem gilt $\|c(t)\| \leq \hat{C}_j e^{-\kappa_j t}$, $c \in \{\eta_j, \zeta_j\}$ mit $\kappa_j, \hat{C}_j > 0$.

Beweis: Wie oben bereits hergeleitet wurde, gibt es eine durch die Anfangswerte eindeutig bestimmte Lösung von (3.13), (3.14), die durch die Reihenentwicklung

$$y(t) = \sum_{j=0}^{\infty} \epsilon^j y_j(t) + \epsilon \sum_{j=0}^{\infty} \epsilon^j \eta_j \left(\frac{t}{\epsilon}\right),$$
$$\lambda(t) = \sum_{j=0}^{\infty} \epsilon^j \lambda_j(t) + \sum_{j=0}^{\infty} \epsilon^j \zeta_j \left(\frac{t}{\epsilon}\right)$$

gegeben ist, wobei die jeweiligen Funktionen $y_i, \lambda_i, \eta_i, \zeta_i$ iterativ bestimmbar und unter obigen Voraussetzungen eindeutig bestimmt sind. Betrachtet man hiervon nur die abgeschnittene Reihe

$$\hat{y}(t) = \sum_{j=0}^{N} \epsilon^j y_j(t) + \epsilon \sum_{j=0}^{N} \epsilon^j \eta_j \left(\frac{t}{\epsilon}\right)$$
$$\hat{\lambda}(t) = \sum_{j=0}^{N} \epsilon^j \lambda_j(t) + \sum_{j=0}^{N} \epsilon^j \zeta_j \left(\frac{t}{\epsilon}\right)$$

und setzt sie in die DGL (3.13), (3.14) ein, kann man wiederum nach einer Taylorentwicklung und der Substitution mit ξ einen ϵ-Koeffizientenvergleich durchführen. Bei diesem Koeffizientenvergleich unterscheiden sich allerdings die ersten N Koeffizienten nicht von jenen Koeffizienten, die sich bei der Entwicklung von (3.22), (3.23) ergeben, da bei der j-te ϵ-Potenz im Koeffizienten nur Summanden y_i, λ_i, ζ_i von maximal Index j und η_i von maximal Index $j-1$ vorkommen. Sie besitzen bis zu diesem Index also die selben Koeffizienten wie bei der Entwicklung nach Einsetzen der vollen Reihe in die Differentialgleichung

$$f(\hat{y}, \hat{\lambda}) = f\left(\sum_{j=0}^{N} \epsilon^j y_j(t) + \epsilon \sum_{j=0}^{N} \epsilon^j \eta_j \left(\frac{t}{\epsilon}\right), \sum_{j=0}^{N} \epsilon^j \lambda_j(t) + \sum_{j=0}^{N} \epsilon^j \zeta_j \left(\frac{t}{\epsilon}\right)\right)$$
$$\overset{\text{Taylor}}{=:} \epsilon^0 \underbrace{\psi_0(y_0, \lambda_0, \zeta_0)}_{=\dot{y}_0(t)+\dot{\eta}_0(t/\epsilon)\text{nach(3.22)}} + \epsilon^1 \cdots + \epsilon^N \underbrace{\psi_N(y_{j\le N}, \lambda_{j\le N}, \eta_{j<N}, \zeta_{j\le N})}_{=\dot{y}_N(t)+\dot{\eta}_N(t/\epsilon)\text{ nach (3.22)}}$$
$$+ \mathcal{O}\left(\epsilon^{N+1} \psi_{N+1}(y_{j\le N}, \lambda_{j\le N}, \eta_{j\le N}, \zeta_{j\le N})\right).$$

Da hier im Restglied nur endlich viele Ansatzfunktionen (bis Index N) vorkommen, man also insbesondere eine obere Schranke für die $\hat{C}_i$ und eine untere Schranke für die κ_i finden kann, kann man mit den Modellannahmen für g und f das Restglied beschränken. Analoge Rechnung führt zur gleichen Aussage für $\dot{\hat{\lambda}}$ aus (3.23) und man erhält insgesamt:

$$\dot{\hat{y}}(t) = f(\hat{y}(t), \hat{\lambda}(t)) + \mathcal{O}(\epsilon^{N+1}) \tag{3.34}$$
$$\epsilon\dot{\hat{\lambda}}(t) = g(\hat{y}(t), \hat{\lambda}(t)) + \mathcal{O}(\epsilon^{N+1}). \tag{3.35}$$

Zieht man hiervon die ursprüngliche DGL ab, führt dies zur Gleichung

$$\dot{\hat{y}}(t) - \dot{y}(t) = f(\hat{y}(t), \hat{\lambda}(t)) - f(y(t), \lambda(t)) + \mathcal{O}(\epsilon^{N+1}) \tag{3.36}$$
$$\epsilon\left(\dot{\hat{\lambda}}(t) - \dot{\lambda}(t)\right) = g(\hat{y}(t), \hat{\lambda}(t)) - g(y(t), \lambda(t)) + \mathcal{O}(\epsilon^{N+1}). \tag{3.37}$$

Mit der Lipschitz-Bedingung an f (mit Konstanten $L_1, L_2 > 0$) und mit Dreiecksungleichung erhält man aus (3.36) schließlich

$$\left\|\dot{\hat{y}}(t) - \dot{y}(t)\right\| \le L_1 \left\|\hat{y}(t) - y(t)\right\| + L_2 \left\|\hat{\lambda}(t) - \lambda(t)\right\| + C_1\epsilon^{N+1}. \tag{3.38}$$

Um die Ableitung auf der linken Seite dieser Ungleichung aus der Norm zu bekommen, kann man wieder die Dini-Ableitung benutzen, da man mit dieser die Norm einer Ableitung abschätzen kann, etwa mit der umgekehrten Dreiecksungleichung:

$$\frac{\|y(t+h)\| - \|y(t)\|}{h} \leq \left\| \frac{y(t+h) - y(t)}{h} \right\|,$$

somit $D_+ \|y(t)\| \leq \|\dot{y}(t)\|$ bzw. auf (3.38) angewandt

$$D_+ \|\hat{y}(t) - y(t)\| \leq L_1 \|\hat{y}(t) - y(t)\| + L_2 \left\|\hat{\lambda}(t) - \lambda(t)\right\| + C_1 \epsilon^{N+1}.$$

Für eine entsprechende Abschätzung für (3.37) könnte man auf gleiche Weise eine Lipschitz-Abschätzung durchführen, jedoch benötigt man im Folgenden eine schärfere Abschätzung. Analog zu Formeln (10.17) und (10.18) aus [HNWXX] gilt mit $m(t) := \| \hat{\lambda}(t) - \lambda(t) \|$:

$$\begin{aligned}
m(t+h) &= \left\|\hat{\lambda}(t+h) - \lambda(t+h)\right\| \\
&\overset{\text{Taylor}}{=} \left\|\hat{\lambda}(t) - \lambda(t) + h\left(\dot{\hat{\lambda}}(t) - \dot{\lambda}(t)\right)\right\| + \mathcal{O}(h^2) \\
&\overset{(3.37)}{=} \left\|\hat{\lambda}(t) - \lambda(t) + \frac{h}{\epsilon}\left(g(\hat{y}(t), \hat{\lambda}(t)) + \mathcal{O}(\epsilon^{N+1}) - g(y(t), \lambda(t))\right)\right\| + \mathcal{O}(h^2) \\
&= \left\|\hat{\lambda}(t) - \lambda(t) + \frac{h}{\epsilon}\left(g(\hat{y}(t), \hat{\lambda}(t)) - g(y(t), \hat{\lambda}(t)) + g(y(t), \hat{\lambda}(t)) - g(y(t), \lambda(t))\right)\right\| \\
&\quad + \mathcal{O}(h^2) + \frac{h}{\epsilon}\mathcal{O}(\epsilon^{N+1}) \\
&\leq \left\|\hat{\lambda}(t) - \lambda(t) + \frac{h}{\epsilon}\left(g(y(t), \hat{\lambda}(t)) - g(y(t), \lambda(t))\right)\right\| + \mathcal{O}(h^2) \\
&\quad + \frac{h}{\epsilon}\left(\mathcal{O}(\epsilon^{N+1}) + \left\|g(\hat{y}(t), \hat{\lambda}(t)) - g(y(t), \hat{\lambda}(t))\right\|\right) \\
&\overset{\text{ZWS}}{\leq} \max_{\mu \in [\lambda(t), \hat{\lambda}(t)]} \left\|\hat{\lambda}(t) - \lambda(t) + \frac{h}{\epsilon} g_\lambda(y(t), \mu)\left(\hat{\lambda}(t) - \lambda(t)\right)\right\| + \mathcal{O}(h^2) \\
&\quad + \frac{h}{\epsilon}\left(\mathcal{O}(\epsilon^{N+1}) + \left\|g(\hat{y}(t), \hat{\lambda}(t)) - g(y(t), \hat{\lambda}(t))\right\|\right) \\
&\overset{\text{Lip.}}{\leq} \max_{\mu \in [\lambda(t), \hat{\lambda}(t)]} \left\|\mathbb{1} + \frac{h}{\epsilon} g_\lambda(y(t), \mu)\right\| m(t) + \mathcal{O}(h^2) + \frac{h}{\epsilon}\left(\mathcal{O}(\epsilon^{N+1}) + L_3 \|\hat{y}(t) - y(t)\|\right).
\end{aligned}$$

Daraus folgt mit Einsetzen der Definition der Dini-Ableitung für $m(t)$

$$\begin{aligned}
&\epsilon D_+ m(t) \\
&= \epsilon \liminf_{h \to 0^+} \frac{m(t+h) - m(t)}{h} \\
&\leq \liminf_{h \to 0^+} \frac{\max_{\mu \in [\lambda(t), \hat{\lambda}(t)]} \left\|\mathbb{1} + \frac{h}{\epsilon} g_\lambda(y(t), \mu)\right\| m(t) + \frac{h}{\epsilon} L_3 \|\hat{y}(t) - y(t)\| - m(t)}{\frac{h}{\epsilon}} + \mathcal{O}(\epsilon^{N+1}) \\
&\leq \epsilon \max_{\mu \in [\lambda(t), \hat{\lambda}(t)]} \underbrace{\liminf_{h \to 0^+} \frac{\left\|\mathbb{1} + \frac{h}{\epsilon} g_\lambda(y(t), \mu)\right\| - 1}{h}}_{= \mu(\frac{1}{\epsilon} g_\lambda) = \frac{1}{\epsilon}\mu(g_\lambda) \leq -\frac{1}{\epsilon} \text{ n.V.}} m(t) + \mathcal{O}(\epsilon^{N+1}) + L_3 \|\hat{y}(t) - y(t)\| \\
&= L_3 \|\hat{y}(t) - y(t)\| - m(t) + \mathcal{O}(\epsilon^{N+1}).
\end{aligned}$$

Es ergibt sich also insgesamt das Dini-Differentialgleichungs-System

$$D_+ \|\hat{y}(t) - y(t)\| \leq L_1 \|\hat{y}(t) - y(t)\| + L_2 \left\|\hat{\lambda}(t) - \lambda(t)\right\| + C_1 \epsilon^{N+1} \tag{3.39}$$

$$\epsilon D_+ \left\|\hat{\lambda}(t) - \lambda(t)\right\| \leq L_3 \|\hat{y}(t) - y(t)\| - \left\|\hat{\lambda}(t) - \lambda(t)\right\| + C_2 \epsilon^{N+1}. \tag{3.40}$$

Dies führt zum System gewöhnlicher Differentialgleichungen

$$\begin{pmatrix} \dot{u} \\ \epsilon\dot{v} \end{pmatrix} = \begin{pmatrix} L_1 & L_2 \\ L_3 & -1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \epsilon^{N+1} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \tag{3.41}$$

mit Anfangswerten $u_0 = \|\hat{y}(0) - y(0)\| = \mathcal{O}(\epsilon^{N+1})$ und $v_0 = \left\|\hat{\lambda}(0) - \lambda(0)\right\| = \mathcal{O}(\epsilon^{N+1})$.

Um zu zeigen, dass Lemma 3.3 anwendbar ist, benötigt es eine geringfügige Modifikation: Statt $(\dot{u}, \dot{v})^T =: g(u, v)$ aus (3.41) sei für die gesuchten Variablen u und v die Differentialgleichung

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} L_1 & L_2 \\ \frac{1}{\epsilon}L_3 & \frac{-1}{\epsilon} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \tilde{C}_1\epsilon^{N+1} \\ \tilde{C}_2\epsilon^N \end{pmatrix} =: \tilde{g}(u, v) \tag{3.42}$$

mit $\tilde{C}_1$ und $\tilde{C}_2$ geringfügig größer als C_1 bzw. C_2 zu erfüllen. Insbesondere gilt also komponentenweise $g(u, v) < \tilde{g}(u, v)$. Hiermit sind die Bedingungen von Lemma 3.3 erfüllt, denn

1. $D_+ \|\hat{y}(t) - y(t)\| \leq g_1(\|\hat{y}(t) - y(t)\|)$ nach (3.39) und
   $D_+ \left\|\hat{\lambda}(t) - \lambda(t)\right\| \leq g_2(D_+ \left\|\hat{\lambda}(t) - \lambda(t)\right\|)$ nach (3.40)

2. $D_+(u(t), v(t))^T \overset{u,v \text{ diffbar}}{=} (\dot{u}(t), \dot{v}(t))^T = \tilde{g}(u(t), v(t)) \overset{\text{komp.weise}}{>} g(u(t), v(t))$

3. Anfangswerte wurden gleich gewählt

4. $L_1u + L_2v + C_1\epsilon^{N+1} \leq L_1u + L_2(v + h) + C_1\epsilon^{N+1}$ für $h \geq 0$, ebenso
   $L_3u - v + C_2\epsilon^{N+1} \leq L_3(u + h) - v + C_2\epsilon^{N+1}$ für $h \geq 0$, da die Lipschitz-Konstanten L_i positiv sind.

Damit ist Lemma 3.3 anwendbar und es folgt, dass $\|\hat{y}(t) - y(t)\| \leq u(t)$ und $\|\hat{\lambda}(t) - \lambda(t)\| \leq v(t)$, Abschätzungen für u und v können also direkt auf Abschätzungen für die Differenz zwischen den abgeschnittenen Reihen und der exakten Lösung übernommen werden. Die Lösung des linearen Differentialgleichungssystems (3.42) hat die Form *homogene Lösung plus partikuläre Lösung*, wobei die partikuläre Lösung aufgrund der konstanten Inhomogenität ebenfalls konstant gewählt werden kann als $u_{\text{part}} = -\frac{\tilde{C}_1 + L_2\tilde{C}_2}{L_1 + L_2L_3}\epsilon^{N+1}$ und $v_{\text{part}} = \left(\tilde{C}_2 - L_3\frac{\tilde{C}_1 + L_2\tilde{C}_2}{L_1 + L_2L_3}\right)\epsilon^{N+1}$ und damit $\mathcal{O}(\epsilon^{N+1})$ ist. Für die inhomogenen Lösungen benötigt man die Eigenwerte α_i der Matrix

$$\begin{pmatrix} L_1 & L_2 \\ \frac{L_3}{\epsilon} & \frac{-1}{\epsilon} \end{pmatrix}.$$

Ihre Eigenwerte sind

$$\alpha_\pm = \frac{1}{2}\left(L_1 - \frac{1}{\epsilon}\left(1 \pm \sqrt{1 + \epsilon(2L_1 + 4L_2L_3) + L_1\epsilon^2}\right)\right).$$

Dies sind für hinreichend kleines ϵ zwei verschiedene Eigenwerte. Die inhomogenen Lösungen ergeben sich also wie gewohnt in der Form $e^{\alpha_i t}\nu_i$, wobei die ν_i die zu den α_i gehörigen Eigenvektoren der Matrix sind. Offensichtlich ist der zu „+“ gehörende Eigenwert für hinreichend kleines ϵ negativ, damit ist die erste homogene Lösung exponentiell fallend. Der zum „−“ gehörende Eigenwert ist zwar positiv, jedoch kann man zeigen, dass er beschränkt ist. Eine Taylorentwicklung des Wurzelterms ergibt, dass

$$\begin{aligned} \sqrt{1 + \epsilon(2L_1 + 4L_2L_3) + L_1\epsilon^2} &= 1 + \frac{1}{2}\left(\epsilon(2L_1 + 4L_2L_3) + L_1\epsilon^2\right) + \mathcal{O}(\epsilon^2) \\ &= 1 + \epsilon\left(L_1 + 2L_2L_3\right) + \mathcal{O}(\epsilon^2) \end{aligned}$$

und damit

$$\begin{aligned}\alpha_- &= \frac{1}{2}\left(L_1 - \frac{1}{\epsilon}\left(1 - \sqrt{1 + \epsilon(2L_1 + 4L_2L_3) + L_1\epsilon^2}\right)\right)\\ &= \frac{1}{2}\left(L_1 - \frac{1}{\epsilon}\left(-\epsilon\left(L_1 + 2L_2L_3\right) + \mathcal{O}(\epsilon^2)\right)\right)\\ &= L_1 + L_2L_3 + \mathcal{O}(\epsilon).\end{aligned}$$

Damit ist der zweite Eigenwert zwar positiv, aber auch bzw. gerade für beliebig kleines ϵ beschränkt. Insgesamt folgt für die allgemeine Lösung

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = c_1 e^{\alpha_+ t}\nu_1 + c_2 e^{\alpha_- t}\nu_2 + \begin{pmatrix} u_{\text{part}} \\ v_{\text{part}} \end{pmatrix}.$$

Zur Bestimmung der Konstanten c_1 und c_2 setzt man die Anfangswerte ein, damit folgt

$$\begin{pmatrix} u(0) \\ v(0) \end{pmatrix} = c_1\nu_1 + c_2\nu_2 + \underbrace{\begin{pmatrix} u_{\text{part}} \\ v_{\text{part}} \end{pmatrix}}_{\mathcal{O}(\epsilon^{N+1})} \stackrel{!}{=} \mathcal{O}(\epsilon^{N+1}).$$

Da die beiden Eigenvektoren ν_i o.B.d.A. normiert sind und außerdem linear unabhängig sind, müssen also c_1 und c_2 von der Größenordnung $\mathcal{O}(\epsilon^{N+1})$ sein, um diese Gleichung zu erfüllen. Da weiterhin $e^{\alpha_+ t}$ exponentiell fallend und $e^{\alpha_- t}$ wie gezeigt auf dem untersuchten beschränkten Intervall $[0, \bar{t}]$ von ϵ unabhängig beschränkt ist, gilt damit auf diesem Intervall allgemein

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = \mathcal{O}(\epsilon^{N+1}).$$

Da diese Abschätzung nach Lemma 3.3 auf $\|\hat{y}(t) - y(t)\|$ und $\| \hat{\lambda}(t) - \lambda(t) \|$ übertragbar ist, folgt hieraus und mit $\|\eta_N t\| \leq$ Konst. die Behauptung. □

Eine direkte Folgerung dieses Theorems für $N = 0$ ist, dass sich die Lösung des regularisierten Problems für $t > 0$ zumindest in einer ϵ-Umgebung der reduzierten Lösung befindet, für $t = 0$ lässt sich dies nur für den y-Anteil der Lösung sagen, der λ-Anteil hat eine durch $\|\zeta_0\|_\infty + \mathcal{O}(\epsilon)$ beschränkte Abweichung, die allerdings für $t > 0$ exponentiell gegen 0 strebt.

Um Lemma 6 auf unser Modellproblem anzuwenden zu können, muss allerdings noch gezeigt werden, dass $\mu(g_\lambda) \leq -1$ gilt, sowie, dass die Lösung des reduzierten Problems gleichzeitig Lösung des nichtregularisierten Problems (1.10) – unserem Ausgangsproblem – ist. Beides kann einfach nachgerechnet werden:
Betrachten wir g_λ. Nach (3.14) gilt

$$g_\lambda = -dGM^{-1}G^T.$$

Wegen Modellannahme 1 ist M symmetrisch und positiv definit, damit auch M^{-1}. Wegen $(GM^{-1}G^T)^T = GM^{-T}G^T = GM^{-1}G^T$, ist g_λ symmetrisch. Die logarithmische Norm von g_λ ist also gerade der größte Eigenwert der Matrix g_λ. Da nach Modellannahme 2 G vollen Rang hat, gilt außerdem

$$\begin{aligned} & x^T M^{-1} x > 0 \ \forall\ x \in \mathbb{R}^n \backslash \{0\} \\ \Rightarrow\ & x^T M^{-1} x > 0 \ \forall\ x := G^T y,\ y \in \mathbb{R}^n \backslash \{0\} \ \overset{\text{Modellann. 2}}{(\Rightarrow x \neq 0)} \\ \Rightarrow\ & y^T G M^{-1} G^T y > 0 \ \forall\ y \in \mathbb{R}^n \backslash \{0\}. \end{aligned} \tag{3.43}$$

Somit ist $GM^{-1}G^T$ positiv definit. Da $d > 0$ ist, ist damit g_λ negativ definit in einer Umgebung um die Lösung $y(t)$ unseres Problems auf dem abgeschlossenen Zeitintervall $[0, \bar{t}]$. Da wir uns in endlichen Dimensionen und auf einem kompakten Zeitintervall befinden, folgt daraus die Existenz einer Zahl $a \in \mathbb{R},\ a > 0$, sodass für die logarithmische Norm von g_λ gilt $\mu(g_\lambda) = -a$. Für $a \geq 1$ erfüllt dies die gesuchte Voraussetzung des Theorems 6, für $0 < a < 1$ ergibt sich die gesuchte Schranke durch Neuskalierung in (3.14): Multiplikation beider Seiten mit $\frac{1}{a}$ und Verwenden der neuen Konstanten $\epsilon_{\text{neu}} := \frac{\epsilon}{a}$ und $d_{\text{neu}} := \frac{d}{a}$ statt ϵ und d beschreibt ein äquivalentes System wie das ursprüngliche (3.14), erfüllt jedoch nun $\mu(g_\lambda) < -1$ und die Aussagen des Satzes gelten analog für das neu skalierte ϵ, das ja ohnehin beliebig klein werden soll.
Bleibt noch zu zeigen, dass das reduzierte Problem, also

$$\begin{aligned}\dot{y} &= f(y, \lambda)\\ 0 &= cG(y_1)\dot{y}_1 + d(\dot{G}(y_1)\dot{y}_1 + GM^{-1}(y)(F(y) - G(y_1)^T\lambda)),\end{aligned}$$

äquivalent zum unregularisierten Problem, in gleicher Schreibweise dargestellt

$$\begin{aligned}\dot{y} &= f(y, \lambda)\\ 0 &= G(y_1)\dot{y}_1,\end{aligned}$$

ist. Da jeweils die ersten Zeilen identisch sind, müssen also die zweiten Gleichungen äquivalent sein. Nutzt man, dass

$$\begin{aligned}\frac{d}{dt}G(y_1)y_2 &= \dot{G}(y_1)y_2 + G(y_1)\dot{y}_2\\ &= \dot{G}(y_1)y_2 + G(y_1)(M^{-1}(y)(F(y) - G(y_1)^T))\end{aligned}$$

und substituiert die Funktion $u(t) := G(y_1)y_2$, bleibt somit die Äquivalenz

$$0 = u \tag{3.44}$$

$$\Leftrightarrow$$

$$0 = cu + d\dot{u} \tag{3.45}$$

zu zeigen. Da (3.45) eine lineare Differentialgleichung mit konstanten Koeffizienten ist, lässt sich die Lösung direkt ablesen als $u(t) = u(0)e^{-t\frac{c}{d}}$. Setzt man die Definition von u ein, erhält man den Anfangswert $u(0) = G(y_1(0))y_2(0)$. Da zumindest die Anfangswerte für y im Modellproblem die Zwangsbedingung (3.44) erfüllen müssen, ergibt sich $u(0) = 0$ und damit hat (3.45) die eindeutige Lösung $u(t) = 0$, was gerade (3.44) entspricht. Damit ist die Äquivalenz von (3.44) und (3.45) und folglich die Äquivalenz der reduzierten Lösung mit der unregularisierten Lösung des Modellproblems gezeigt. Somit ist Lemma 6 anwendbar. Der Fall $\kappa = 1$ liefert also eine Regularisierung, mit der die Lösung des Problems für $\epsilon \to 0$ gegen die Lösung des unregularisierten Problems konvergiert (zumindest für $t > 0$).
Für $N = 0$ in Theorem 6 und nach dem soeben gezeigten, ist die Aussage aus [Dep13] für den Fall *Strong Damping* bewiesen, namentlich, dass die Lösung der durch *Strong Damping* regularisierten Differential-Algebraischen-Gleichung in einer ϵ-Umgebung um die Lösung der ursprünglichen Differential-Algebraischen-Gleichung liegt.

3.3 Sonderfall $\kappa = 0.5$, *Principal Damping*

Wie bereits erwähnt, erwies sich der Fall $\kappa = 0.5$ als unerwartet schwierig. Zwar lassen diverse Theoreme (insbesondere das oben verwendete Theorem 6) für singulär gestörte Probleme auf der rechten

Seite noch eine Abhängigkeit von ϵ zu, also es sind Probleme der Form

$$\dot{y} = f(y, \lambda, \epsilon)$$
$$\epsilon\dot{\lambda} = g(y, \lambda, \epsilon)$$

erlaubt, allerdings unter der Einschränkung, dass die ϵ-Abhängigkeit, insbesondere im Grenzbereich $\epsilon \to 0$, glatt ist. In unserem Fall ist dies jedoch nicht gegeben, da sich bei $\kappa = 0.5$ auf der rechten Seite der zweiten Zeile der Term $\sqrt{\epsilon}$ ergibt (nicht stetig differenzierbar im Punkt $\epsilon = 0$). Die Problematik einer singulär gestörten Differentialgleichung mit mehr als zwei verschiedenen ϵ-Potenzen innerhalb der selben Gleichung ist, von besagter erlaubter glatter Abhängigkeit abgesehen, in der Literatur nicht auffindbar. An der zusätzlichen ϵ-Potenz, $\sqrt{\epsilon}$, scheitern diverse naheliegende Ansätze, ein zu Theorem 6 äquivalentes Lemma auch für diesen Fall zu beweisen ebenso wie der Versuch, eine Fehlerabschätzung zwischen den Fällen $\kappa = 1$ und $\kappa = 0.5$ zu erhalten (mit der Absicht, hierdurch eine überzählige ϵ-Potenz eliminieren zu können).

Im ersten nun folgenden Abschnitt wird dargelegt, welche Schwierigkeiten sich dabei ergeben, ein zu Theorem 6 analoges Theorem für singulär gestörte Probleme der Form des Principal Damping herzuleiten. Das größte Problem hierbei ist, dass die logarithmische Norm von g_λ nun nicht mehr durch eine von ϵ unabhängige negative Konstante abgeschätzt werden kann, sondern den Faktor $\sqrt{\epsilon}$ erhält. Um dies zu kompensieren, müssen starke Zusatzbedingungen an die Differential-Algebraische-Gleichung und ihre Lösung gestellt werden, die nur schwer zu erfüllen oder nachzuweisen sind.

Im zweiten Abschnitt wird gezeigt, wie sich dieses Problem umgehen lässt. Die spezielle Struktur des Modellproblems wird dazu ausgenutzt, das Problem durch Einführen einer weiteren Variable auf eine Form zu bringen, auf welche sich Theorem 6 anwenden lässt. Dies führt zu einer allgemeinen Aussage über die Konvergenz des als Principal Damping regularisierten Modellproblems, für die sich nur zusätzliche Bedingungen an die Wahl der Parameter c und d ergibt. Die Kernaussage hierfür ist schließlich Korollar 4.

3.3.1 Direkte Modifikation von Theorem 6

Die in (3.44), (3.45) gezeigte Äquivalenz ermöglicht es, eine modifizierte Variante von Theorem 6 zu zeigen, allerdings unter starken zusätzlichen Einschränkungen. Insgesamt ist dies ein Phänomen, welches sich durch jegliche zu diesem Thema auffindbare Literatur zieht: Für den Fall einer singulär gestörten Differential-Algebraischen-Gleichung, die nicht exakt die Standardform besitzt, werden starke zusätzliche Annahmen gefordert. Am weitesten verbreitet ist hierbei die Annahme, dass die Lösung der regularisierten Differential-Algebraischen-Gleichung von ϵ unabhängig beschränkt ist. Diese Eigenschaft ist der Differential-Algebraischen-Gleichung im Allgemeinen nur sehr schwer nachzuweisen und erlaubt letztlich nur *a posteriori* Rückschlüsse bei der Lösung: Ist die Lösung unabhängig von ϵ beschränkt, dann ist sie auch konvergent.
Ziel ist es zunächst, wieder eine Potenzreihenentwicklung der Lösung der Differential-Algebraischen-Gleichung zu bestimmen, deren Koeffizientenfunktionen eindeutig bestimmt sind. Der Übersicht halber hier zunächst nochmals die zu lösenden Gleichungen:
Die singulär gestörte Differentialgleichung lautet, in angepasster Notation

$$\begin{aligned}
\dot{y}_1 &= y_2 && =: f_1(y, \lambda) \\
\dot{y}_2 &= M^{-1}(y_1)(F(y) - G(y_1)^T\lambda) && =: f_2(y, \lambda) \\
\dot{\lambda} &= \frac{c}{\epsilon}\underbrace{G(y_1)y_2}_{=:g(y)} + \frac{1}{\epsilon^\kappa}\underbrace{d(\dot{G}(y_1)y_2 + GM^{-1}(F(y) - G(y_1)^T\lambda))}_{=:h(y,\lambda)}.
\end{aligned}$$

O.B.d.A. seien $c = d = 1$, die Sonderfälle $c = 0$ oder $d = 0$ entsprechen ohnehin keinem Kelvin-Voigt-Körper (Federkonstante 0 oder keine viskose Dämpfung). Damit liegt nach Multiplikation mit ϵ in der zweiten Zeile die Differentialgleichung

$$\dot{y} = f(y, \lambda) \tag{3.46}$$

$$\epsilon\dot{\lambda} = g(y) + \sqrt{\epsilon}h(y, \lambda) \qquad \left(= g(y) + \sqrt{\epsilon}\dot{g(y)}\right) \tag{3.47}$$

mit der zugrunde liegenden Differential-Algebraischen-Gleichung

$$\dot{y} = f(y, \lambda) \tag{3.48}$$

$$0 = g(y) \tag{3.49}$$

vor. Diese ist nach (3.44), (3.45) äquivalent zur Differential-Algebraischen-Gleichung (3.15), (3.16). Beginnen wir wie bei *Strong Damping* mit einem Potenzreihenansatz, diesmal aufgrund der anderen Problemstruktur allerdings in der Form

$$y = \sum_{i=0}^{\infty} y_i \sqrt{\epsilon}^i, \tag{3.50}$$

$$\lambda = \sum_{i=0}^{\infty} \lambda_i \sqrt{\epsilon}^i. \tag{3.51}$$

Einsetzen in (3.46), (3.47) und Taylorentwicklung um den Entwicklungspunkt (y_0, λ_0) ergibt

$$\sum_{i=0}^{\infty} \dot{y}_i \sqrt{\epsilon}^i = f(y_0, \lambda_0) + f_y(y_0, \lambda_0) \left(\sum_{i=1}^{\infty} y_i \sqrt{\epsilon}^i \right) + f_\lambda(y_0, \lambda_0) \left(\sum_{i=1}^{\infty} \lambda_i \sqrt{\epsilon}^i \right) + \mathcal{O}\left(\epsilon \left(f_{yy} \dots\right)\right)$$

$$\begin{aligned} \sum_{i=0}^{\infty} \dot{\lambda}_i \sqrt{\epsilon}^{i+2} = {} & g(y_0) + g_y(y_0) \left(\sum_{i=1}^{\infty} y_i \sqrt{\epsilon}^i \right) + \sqrt{\epsilon} h(y_0, \lambda_0) \\ & + h_y(y_0, \lambda_0) \left(\sum_{i=1}^{\infty} y_i \sqrt{\epsilon}^{i+1} \right) + h_\lambda(y_0, \lambda_0) \left(\sum_{i=1}^{\infty} \lambda_i \sqrt{\epsilon}^{i+1} \right) + \mathcal{O}\left(\epsilon \left(g_{yy} \dots\right)\right). \end{aligned}$$

Der Koeffizientenvergleich liefert nun für ϵ^0

$$\begin{aligned} \dot{y}_0 &= f(y_0, \lambda_0), \\ 0 &= g(y_0). \end{aligned}$$

Dies ist unsere ursprüngliche Differential-Algebraische-Gleichung vom Differentiationsindex 2, d.h. nach zweifachem Ableiten der zweiten Zeile kann man dieses System zu einer expliziten Differentialgleichung umformen, die nach Modellannahmen für f, g und insbesondere die nach Modellannahme 3 existierende Inverse von $g_y f_\lambda$ eine lokale Lipschitz-Bedingung erfüllt, somit existiert nach dem Satz von Picard Lindelöff eine lokal eindeutige Lösung, also sind y_0 und λ_0 auf einem beschränkten Zeitintervall $[0, \bar{t}]$ in Abhängigkeit des Anfangswertes eindeutig festgelegt. Insbesondere folgt, aufgrund der in (3.44),(3.45) gezeigten Äquivalenz von $0 = g(y_0)$ zu $0 = g(y_0) + \frac{d}{dt} g(y_0)$ (bei vorausgesetztem Anfangswert $g(y_0(0)) = 0$), dass nach Wahl eines geeigneten Anfangswertes für y_0 der Anfangswert für λ_0 festgelegt ist, da dies der ϵ-Potenz nullter Ordnung des Falles *Strong Damping*, also (3.18), (3.19), entspricht. (Von der Notation: Das $g(y, \lambda)$ im Falle Strong Damping entspricht hier $g(y) + \dot{g(y)} = g(y) + h(y, \lambda)$).

Der Koeffizientenvergleich für $\sqrt{\epsilon}$ ergibt:

$$\dot{y}_1 = f_y(y_0, \lambda_0)y_1 + f_\lambda(y_0, \lambda_0)\lambda_1, \tag{3.52}$$
$$0 = g_y(y_0)y_1 + h(y_0, \lambda_0) \tag{3.53}$$

Dies ist im Grunde auch wieder eine Differential-Algebraische-Gleichung, wobei y_0 und λ_0 oben bereits bestimmt wurden. Insbesondere sind die g und h Terme auf der rechten Seite der zweiten Zeile bereits bestimmt als Funktionen $c_1(t)$ bzw. $c_2(t)$. Ableiten der zweiten Zeile ergibt

$$\begin{aligned} 0 &= \dot{c}_1(t)y_1 + c_1(t)\dot{y}_1 + \dot{c}_2(t) \\ &\overset{3.52}{=} \dot{c}_1(t)y_1 + g_y(y_0)y_1\left(f_y(y_0, \lambda_0)y_1 + f_\lambda(y_0, \lambda_0)\lambda_1\right) \\ &= \tilde{c}(t, y_1, \lambda_1) + \underbrace{g_y(y_0)f_\lambda(y_0, \lambda_0)}_{\text{invertierbar nach Modellannahme 3}}\lambda_1, \end{aligned}$$

hierbei steht $\tilde{c}$ ebenfalls für eine Funktion von bereits bekannten Argumenten. Damit ist die zweite Gleichung nach Ableiten nach λ_1 auflösbar, Einsetzen in die erste Zeile legt schließlich y_1 und λ_1 in Form einer DGL fest.

Der Koeffizientenvergleich für ϵ ergibt

$$\begin{aligned} \dot{y}_2 &= f_y(y_0, \lambda_0)y_2 + f_\lambda(y_0, \lambda_0)\lambda_2 + \phi(y_0, y_1, \lambda_0, \lambda_1), \\ \dot{\lambda}_0 &= g_y(y_0)y_2 + \psi(y_0, y_1, \lambda_0, \lambda_1). \end{aligned}$$

Alle Funktionen bis zum Index 1 sind bereits bestimmt, d.h. auch dies ist wieder eine Differential-Algebraische-Gleichung, die zweite Zeile lässt sich ganz analog zu oben ableiten (wieder ergibt sich der Term $g_{1y}(y_0, \lambda_0)f_\lambda(y_0, \lambda_0)\lambda_2$), damit ist auch hier y_2 und λ_2 eindeutig bestimmt und die ganze Iteration geht so weiter:

$$\begin{aligned} \dot{y}_j &= f_y(y_0, \lambda_0)y_j + f_\lambda(y_0, \lambda_0)\lambda_2 + \phi_j(y_0, \dots, y_{j-1}, \lambda_0, \dots, \lambda_{j-1}) \\ \dot{\lambda}_{j-2} &= g_y(y_0)y_j + \psi_j(y_0, \dots, y_{j-1}, \lambda_0, \dots, \lambda_{j-1}) \end{aligned}$$

Immer ergibt sich eine Differential-Algebraische-Gleichung, deren zweite Zeile nach Ableiten nach dem entsprechenden λ_j auflösbar ist, da dies immer in der Form $g_{1y}(y_0, \lambda_0)f_\lambda(y_0, \lambda_0)\lambda_j$ vorkommt.
Analog zum Fall *Strong Damping* kann dieser Ansatz in der Form

$$\begin{pmatrix} y(t) \\ \lambda(t) \end{pmatrix} = \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \begin{pmatrix} y_i(t) \\ \lambda_i(t) \end{pmatrix} + \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \begin{pmatrix} \sqrt{\epsilon}\eta_j(\frac{t}{\sqrt{\epsilon}}) \\ \zeta_j(\frac{t}{\sqrt{\epsilon}}) \end{pmatrix}$$

erweitert werden, wobei die y_i und λ_i die oben bereits für spezielle (für λ passende) Anfangswerte bestimmten Koeffizientenfunktionen sind. Einsetzen in (3.46),(3.47) ergibt

$$\begin{aligned} &\underbrace{\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{y}_i(t)}_{=f(\dots)} + \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\eta}_j\left(\frac{t}{\sqrt{\epsilon}}\right) \\ &= f\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(t) + \sqrt{\epsilon}\eta_j\left(\frac{t}{\sqrt{\epsilon}}\right)\right), \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(t) + \zeta_j\left(\frac{t}{\sqrt{\epsilon}}\right)\right)\right), \end{aligned}$$

sowie

$$\begin{aligned}
&\underbrace{\epsilon \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\lambda}_i(t)}_{=g(\dots)+\sqrt{\epsilon}h(\dots)} + \sqrt{\epsilon} \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\zeta}_j \left(\frac{t}{\sqrt{\epsilon}}\right) \\
&= g\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(t) + \sqrt{\epsilon}\eta_j\left(\frac{t}{\sqrt{\epsilon}}\right)\right)\right) \\
&\quad + \sqrt{\epsilon} h\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(t) + \sqrt{\epsilon}\eta_j\left(\frac{t}{\sqrt{\epsilon}}\right)\right), \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(t) + \zeta_j\left(\frac{t}{\sqrt{\epsilon}}\right)\right)\right).
\end{aligned}$$

Ersetzt man wieder $\frac{t}{\sqrt{\epsilon}}$ durch ξ, folgt

$$\begin{aligned}
\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\eta}_j(\xi) = f&\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(\sqrt{\epsilon}\xi) + \sqrt{\epsilon}\eta_j(\xi)\right), \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(\sqrt{\epsilon}\xi) + \zeta_j(\xi)\right)\right) \\
&- f\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i y_i(\sqrt{\epsilon}\xi), \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \lambda_i(\sqrt{\epsilon}\xi)\right),
\end{aligned}$$

sowie

$$\begin{aligned}
&\sqrt{\epsilon} \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\zeta}_j(\xi) \\
&= g\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(\sqrt{\epsilon}\xi) + \sqrt{\epsilon}\eta_j(\xi)\right)\right) - \sqrt{\epsilon} h\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i y_i(\sqrt{\epsilon}\xi), \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \lambda_i(\sqrt{\epsilon}\xi)\right) \\
&\quad + \sqrt{\epsilon} h\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(\sqrt{\epsilon}\xi) + \sqrt{\epsilon}\eta_j(\xi)\right), \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(\sqrt{\epsilon}\xi) + \zeta_j(\xi)\right)\right) - g\left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i y_i(\sqrt{\epsilon}\xi)\right).
\end{aligned}$$

Taylorentwicklungen zunächst der Funktionen f, g und h um die Entwicklungspunkte $(y_0(0), \lambda_0(0) + \zeta_0(\xi))$ respektive $(y_0(0), \lambda_0(0))$ und anschließend von y und λ um 0 ergeben:

$$\begin{aligned}
&\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\eta}_j(\xi) \\
&= f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) + f_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(\sqrt{\epsilon}\xi) + \sqrt{\epsilon}\eta_j(\xi)\right) - y_0(0)\right) \\
&\quad + f_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(\sqrt{\epsilon}\xi) + \zeta_j(\xi)\right) - \lambda_0(0)\right) \\
&\quad - f(y_0(0), \lambda_0(0)) - f_y(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i y_i(\sqrt{\epsilon}\xi) - y_0(0)\right) \\
&\quad - f_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \lambda_i(\sqrt{\epsilon}\xi) - \lambda_0(0)\right) + \mathcal{O}\left(\epsilon \left(f_{yy} \dots\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - f(y_0(0), \lambda_0(0)) \\
&\quad + f_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(y_i(0) + \sqrt{\epsilon}\xi \dot{y}_i(0) \right) + \dot{y}_0(0)\sqrt{\epsilon}\xi + \sum_{i=0}^{\infty} \sqrt{\epsilon}^{i+1} \eta_i(\xi) \right) \\
&\quad + f_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(0) + \sqrt{\epsilon}\xi \dot{\lambda}_i(0) \right) + \sqrt{\epsilon}\xi \dot{\lambda}_0(0) + \sum_{i=1}^{\infty} \sqrt{\epsilon}^i \zeta_i(\xi) \right) \\
&\quad - f_y(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(y_i(0) + \sqrt{\epsilon}\xi \dot{y}_i(0) \right) + \dot{y}_0(0)\sqrt{\epsilon}\xi \right) \\
&\quad - f_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(0) + \sqrt{\epsilon}\xi \dot{\lambda}_i(0) \right) + \sqrt{\epsilon}\xi \dot{\lambda}_0(0) \right) + \mathcal{O}\left(\epsilon \left(f_{yy} \dots \right) \right)
\end{aligned}
$$

$$
\begin{aligned}
&\sqrt{\epsilon} \sum_{i=0}^{\infty} \sqrt{\epsilon}^i \dot{\zeta}_j(\xi) \\
&= g(y_0(0)) + g_y(y_0(0)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(\sqrt{\epsilon}\xi) + \sqrt{\epsilon}\eta_j(\xi) \right) - y_0(0) \right) \\
&\quad + \sqrt{\epsilon} \left(h(y_0(0), \lambda_0(0) + \zeta_0(\xi)) + h_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(y_i(\sqrt{\epsilon}\xi) + \sqrt{\epsilon}\eta_j(\xi) \right) - y_0(0) \right) \right) \\
&\quad + \sqrt{\epsilon} h_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(\sqrt{\epsilon}\xi) + \zeta_j(\xi) \right) - \lambda_0(0) - \zeta_0(\xi) \right) \\
&\quad - g(y_0(0)) - g_y(y_0(0)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i y_i(\sqrt{\epsilon}\xi) - y_0(0) \right) - \sqrt{\epsilon} h_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i \lambda_i(\sqrt{\epsilon}\xi) - \lambda_0(0) \right) \\
&\quad - \sqrt{\epsilon} \left(h(y_0(0), \lambda_0(0)) + h_y(y_0(0), \lambda_0(0)) \left(\sum_{i=0}^{\infty} \sqrt{\epsilon}^i y_i(\sqrt{\epsilon}\xi) - y_0(0) \right) \right) + \mathcal{O}\left(\epsilon^{3/2} \left(g_{yy} \dots \right) \right) \\
&= g_y(y_0(0)) \sum_{i=0}^{\infty} \sqrt{\epsilon}^{i+1} \eta_i(\xi) + \sqrt{\epsilon} \left(h(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - h(y_0(0), \lambda_0(0)) \right) \\
&\quad + \sqrt{\epsilon} h_y(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(y_i(0) + \sqrt{\epsilon}\xi \dot{y}_i(0) \right) + \dot{y}_0(0)\sqrt{\epsilon}\xi + \sum_{i=0}^{\infty} \sqrt{\epsilon}^{i+1} \eta_i(\xi) \right) \\
&\quad + \sqrt{\epsilon} h_\lambda(y_0(0), \lambda_0(0) + \zeta_0(\xi)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(0) + \sqrt{\epsilon}\xi \dot{\lambda}_i(0) \right) + \sqrt{\epsilon}\xi \dot{\lambda}_0(0) + \sum_{i=1}^{\infty} \sqrt{\epsilon}^i \zeta_i(\xi) \right) \\
&\quad - \sqrt{\epsilon} h_y(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(y_i(0) + \sqrt{\epsilon}\xi \dot{y}_i(0) \right) + \dot{y}_0(0)\sqrt{\epsilon}\xi \right) \\
&\quad - \sqrt{\epsilon} h_\lambda(y_0(0), \lambda_0(0)) \left(\sum_{i=1}^{\infty} \sqrt{\epsilon}^i \left(\lambda_i(0) + \sqrt{\epsilon}\xi \dot{\lambda}_i(0) \right) + \sqrt{\epsilon}\xi \dot{\lambda}_0(0) \right) + \mathcal{O}\left(\epsilon^{3/2} \left(g_{yy} \dots \right) \right).
\end{aligned}
$$

Hier kann man wieder einen Koeffizientenvergleich durchführen. Da in der zweiten Gleichung keine ϵ-Potenz der Ordnung Null vorkommt, kann man diese komplett durch $\sqrt{\epsilon}$ teilen und erhält danach insgesamt für die nullte Potenz:

$$\dot{\eta}_0(\xi) = f(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - f(y_0(0), \lambda_0(0)) \tag{3.54}$$

$$\dot{\zeta}_0(\xi) = g_y(y_0(0))\eta_0(\xi) + h(y_0(0), \lambda_0(0) + \zeta_0(\xi)) - h(y_0(0), \lambda_0(0)). \tag{3.55}$$

Hierbei ergibt sich allerdings der im Vergleich zum Fall *Strong Damping* gravierende Unterschied, dass in der Differentialgleichung für ζ_0 ebenfalls η_0 vorkommt und umgekehrt, eine Abschätzung wie oben ist daher nicht ohne Weiteres möglich. Immerhin ist die Differentialgleichung (3.54), (3.55) nach Modellannahmen für die diversen Funktionen lokal Lipschitz-stetig und besitzt damit eine eindeutige Lösung. Für die Wahl der Anfangswerte $\eta_0(0) = \zeta_0(0) = 0$ ist diese gerade die Nullfunktion. Damit kann man Theorem (6) mit allerdings sehr eingeschränkter Aussage anwenden:

Korollar 3 *Konvergenzbeweis im Falle Principal Damping bei geeigneter Wahl der Anfangswerte*
Gegeben sei das Problem (3.46), (3.47) *mit geeigneten Anfangswerten für* y *und dann durch* (3.17) *festgelegten Anfangswerten für* λ *und einer Lösung* y, λ *auf einem Zeitintervall* $[0, \bar{t}]$. *In einer von* ϵ *unabhängigen, hinreichend großen Umgebung um die Lösung* y_0, λ_0 *des reduzierten Problems* (3.48),(3.49) *mit den selben Anfangswerten gelte insbesondere Modellannahme 3, außerdem* $\mu(h_\lambda) < -1$ *und* $g(y) = \mathcal{O}(\sqrt{\epsilon}(y - y_0))$. *Dann ist die Lösung von* (3.46), (3.47) *für hinreichend kleines* ϵ *lokal eindeutig und erfüllt*

$$\begin{aligned} y(t) &= y_0(t) + \mathcal{O}(\sqrt{\epsilon}) \\ \lambda(t) &= \lambda_0(t) + \mathcal{O}(\sqrt{\epsilon}). \end{aligned}$$

Beweis: Der Beweis folgt ganz analog zum Beweis von Theorem 6 mit $N = 0$. Aufgrund der anderen Struktur der zugrunde liegenden Differential-Algebraischen-Gleichung, insbesondere weil

$$\mu(g_\lambda + \sqrt{\epsilon} h_\lambda) = \sqrt{\epsilon}\mu(g_\lambda) \leq -\sqrt{\epsilon} \tag{3.56}$$

nicht wie oben durch eine von ϵ unabhängige, negative Konstante abgeschätzt werden kann, wird die zusätzliche Voraussetzung $g(y) = \mathcal{O}(\sqrt{\epsilon}(y - y_0))$ benötigt. Einsetzen des Ansatzes (y_0, λ_0) in die Differentialgleichung ergibt diesmal

$$\begin{aligned} \dot{y}_0(t) &= f(y_0(t), \lambda_0(t)) + \mathcal{O}(\epsilon^1) \\ \epsilon\dot{\lambda}_0(t) &= g(y_0(t)) + \sqrt{\epsilon} h(y_0(t), \lambda_0(t)) + \mathcal{O}(\epsilon^1) \end{aligned}$$

und damit, im Vergleich zu einer Lösung y, λ der regularisierten Differential-Algebraischen-Gleichung,

$$\begin{aligned} \dot{y}_0(t) - y'(t) &= f(y_0(t), \dot{\lambda}_0(t)) - f(y(t), \lambda(t)) + \mathcal{O}(\epsilon^1) \\ \epsilon\left(\dot{\lambda}_0 - \dot{\lambda}(t)\right) &= g(y_0(t)) - g(y(t)) + \sqrt{\epsilon}(h(y_0(t), \lambda_0(t)) - h(y(t), \lambda(t))). \end{aligned}$$

Hiermit kann man wiederum für $\|y_0 - y\|$ und $\|\lambda_0 - \lambda\|$ die Dini-Differentialgleichungen aufstellen. Wegen (3.56) und nach Annahme $g(y) = \mathcal{O}(\sqrt{\epsilon}(y - y_0)$ ergibt sich somit wie oben das Dini-Differentialgleichungssystem

$$\begin{aligned} D_+ \|y_0(t) - y(t)\| &\leq L_1 \|y_0(t) - y(t)\| + L_2 \|\lambda_0(t) - \lambda(t)\| + C_1\epsilon \\ \epsilon D_+ \|\lambda_0(t) - \lambda(t)\| &\leq \sqrt{\epsilon} L_3 \|y_0(t) - y(t)\| - \sqrt{\epsilon} \|\lambda_0(t) - \lambda(t)\| + C_2\epsilon \end{aligned}$$

mit der dazu gehörigen Differentialgleichung

$$\begin{pmatrix} \dot{u} \\ \sqrt{\epsilon}\dot{v} \end{pmatrix} = \begin{pmatrix} L_1 & L_2 \\ L_3 & -1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \epsilon C_1 \\ \sqrt{\epsilon} C_2 \end{pmatrix}.$$

Diese hat partikuläre, konstante Lösungen analog zu oben, aber der Größenordnung $\mathcal{O}(\sqrt{\epsilon})$. Die homogenen Lösungen benötigen die Eigenwerte der Matrix

$$\begin{pmatrix} L_1 & L_2 \\ \frac{L_3}{\sqrt{\epsilon}} & \frac{-1}{\sqrt{\epsilon}} \end{pmatrix}.$$

Dies ist gerade die Matrix aus dem Fall *Strong Damping* mit $\sqrt{\epsilon}$ statt ϵ, entsprechend ergeben sich wieder ein negativer und ein in ϵ beschränkter Eigenwert, damit insgesamt Lösungen für u und v der Größenordnung $\mathcal{O}(\sqrt{\epsilon})$, welche nach Theorem 3.3 die Aussage des Satzes beweisen.

□

Damit ist, allerdings unter genauer Festlegung der Anfangswerte und mit einer in der Praxis schwer vorherzusagende Voraussetzung für $g(y)$, ein zu Theorem 6 analoges Theorem auch für den Fall Principal Damping gezeigt.

3.3.2 Transformation des Problems

Eine letztlich erfolgreichere Herangehensweise an das Problem *Principal Damping*, ist die Überführung der regularisierten Gleichung in eine andere Form, auf die Theorem 6 angewendet werden kann. Dieser Ansatz geht auf eine grundlegende Idee von Herrn Prof. Dr.-Ing. Alexander Fidlin zurück.

Betrachten wir hierfür das als Principal Damping regularisierte Modellproblem. Dieses hat nach Transformation analog zum letzten Abschnitt die Form

$$\dot{y} = f(y, \lambda) \tag{3.57}$$
$$\epsilon\dot{\lambda} = cg(y) + \sqrt{\epsilon}dh(y, \lambda). \tag{3.58}$$

Hierbei entspricht $g(y)$ dem ursprünglichen Term $G(q)\dot{q}$ aus dem Modellproblem, sowie $h(y, \lambda) = g_y f = g_y \dot{y} = \dot{g(y)}$. Insbesondere entspricht das im Falle Strong Damping verwendete g in dieser Notation dem Term $cg(y) + dh(y, \lambda)$.
Zunächst wird dieses Problem durch Neuskalierung von ϵ auf die Form $\epsilon\dot{\lambda} = \tilde{g} + \sqrt{\epsilon}\dot{\tilde{g}}$ gebracht. Hierfür wird Gleichung (3.58) mit d^2/c^2 multipliziert, wir erhalten

$$\frac{d^2}{c^2}\epsilon\dot{\lambda} = \frac{d^2}{c}g(y) + \sqrt{\epsilon}\frac{d^3}{c^2}h(y, \lambda).$$

Einführen des neu skalierten $\epsilon_{\text{neu}} := \frac{d^2}{c^2}\epsilon$ führt nun zur gewünschten Form

$$\epsilon_{\text{neu}}\dot{\lambda} = \frac{d^2}{c}g(y) + \sqrt{\epsilon_{\text{neu}}}\frac{d^2}{c}h(y, \lambda).$$

Damit ist gezeigt, dass für beliebige Parameterwahl $c, d > 0$ die Gleichung (3.58) auf die Form $\epsilon\dot{\lambda} = \tilde{g} + \sqrt{\epsilon}\dot{\tilde{g}}$ überführbar ist. Der Übersichtlichkeit halber setzen wir deshalb o.B.d.A. in (3.58) $c = d = 1$. Später wird lediglich von Bedeutung sein, dass in dieser Notation gilt $h_\lambda = -\frac{d^2}{c}GM^{-1}G^T$ mit den Matrizen G und M aus der Definition des Modellproblems.
Nach wie vor stört die zusätzliche ϵ-Potenz. Um diese zu eliminieren, führen wir eine neue Variable ein:

$$\theta := \sqrt{\epsilon}\dot{\lambda} - h(y, \lambda) + \gamma(\lambda) \tag{3.59}$$

Hierbei ist $\gamma(\lambda)$ eine noch frei zu wählende Funktion, wie sich herausstellen wird, ist $\gamma(\lambda) = -\lambda$ geschickt. Für die neue Unbekannte θ folgt nun:

$$\begin{aligned}\sqrt{\epsilon}\theta &= \epsilon\dot{\lambda} - \sqrt{\epsilon}\left(h(y, \lambda) - \gamma(\lambda)\right)\\ &\overset{(3.58)}{=} g(y) + \sqrt{\epsilon}h(y, \lambda) - \sqrt{\epsilon}\left(h(y, \lambda) - \gamma(\lambda)\right)\\ &= g(y) + \sqrt{\epsilon}\gamma(\lambda).\end{aligned}$$

Einmaliges Differenzieren liefert

$$\begin{aligned}\sqrt{\epsilon}\dot{\theta} &= \underbrace{g_y(y)\dot{y}}_{=h(y,\lambda)} + \sqrt{\epsilon}\gamma_\lambda(\lambda)\dot{\lambda} \\ &\overset{(3.59)}{=} h(y,\lambda) + \gamma_\lambda(\lambda)\left(\theta + h(y,\lambda) - \gamma(\lambda)\right).\end{aligned}$$

Die Wahl $\gamma(\lambda) = -\lambda$ führt zu

$$\begin{aligned}\sqrt{\epsilon}\dot{\theta} &= h(y,\lambda) - (\theta + h(y,\lambda) + \lambda) \\ &= -\theta - \lambda.\end{aligned}$$

Dies, (3.57) und (3.59) ergeben zusammen das singulär gestörte Problem

$$\begin{cases}\dot{y} & = f(y,\lambda) \\ \sqrt{\epsilon}\dot{\lambda} & = \theta + h(y,\lambda) + \lambda \\ \sqrt{\epsilon}\dot{\theta} & = -\theta - \lambda\end{cases}.$$

Fasst man θ und λ zu $\tilde{\lambda}$ zusammen, ist dies ein singulär gestörtes Problem in der bekannten Form

$$\begin{cases}\dot{y} & = f(y,\tilde{\lambda}) \\ \sqrt{\epsilon}\dot{\tilde{\lambda}} & = \tilde{g}(y,\tilde{\lambda})\end{cases}. \tag{3.60}$$

Diese Funktion $\tilde{g}$ übernimmt sämtliche Glattheitsannahmen von h. Die zugrunde liegende Differential-Algebraische-Gleichung hat die Form

$$\begin{aligned}\dot{y} &= f(y,\lambda) \\ 0 &= \theta + h(y,\lambda) + \lambda \\ 0 &= -\theta - \lambda.\end{aligned}$$

Einsetzen der dritten Zeile in die zweite ergibt

$$\begin{aligned}\dot{y} &= f(y,\lambda) \\ 0 &= h(y,\lambda) \quad (= \dot{g(y)}).\end{aligned}$$

Dies ist äquivalent zur ursprünglichen Differential-Algebraischen-Gleichung

$$\begin{aligned}\dot{y} &= f(y,\lambda) \\ 0 &= g(y),\end{aligned}$$

da bei konsistenten Anfangswerten gilt $g(y_0) = 0$, also die Bedingung $0 = h(y,\lambda) = \dot{g(y)}$ zu $g(y) = 0$ äquivalent ist. Dies wurde bereits im Falle Strong Damping bei (3.44), (3.45) bewiesen.

Bleibt zu zeigen, dass $\tilde{g}_{\tilde{\lambda}}(y,\tilde{\lambda})$ in Lösungsumgebung eine negative logarithmische Norm besitzt, um Theorem 6 anwenden zu können. Es gilt

$$\tilde{g}_{\tilde{\lambda}}(y,\tilde{\lambda}) = \begin{pmatrix}\mathbb{1} + h_\lambda(y,\lambda) & \mathbb{1} \\ -\mathbb{1} & -\mathbb{1}\end{pmatrix} =: A$$

Damit ergibt sich

$$A + A^T = \begin{pmatrix}2\cdot\mathbb{1} + h_\lambda(y,\lambda) + h_\lambda(y,\lambda)^T & 0 \\ 0 & -2\cdot\mathbb{1}\end{pmatrix}$$

Es verbleibt also zu zeigen, dass die Eigenwerte von $h_\lambda(y,\lambda) + h_\lambda(y,\lambda)^T$ kleiner als -2 sind. Ist dies gezeigt, lässt sich Theorem VI.3.2 aus [Hai10] direkt anwenden, die einzige Modifikation ist $\sqrt{\epsilon}$ statt ϵ. Betrachten wir hierfür h_λ. In der hier verwendeten Notation gilt

$$h_\lambda(y,\lambda) = -\frac{d^2}{c} G(y) M(y)^{-1} G(y)^T.$$

Im Fall Strong Damping wurde bereits nachgerechnet, dass $-GM^{-1}G^T$ symmetrisch und negativ definit ist. Damit ergibt sich

$$\begin{aligned}
\mu\left(h_{\tilde{\lambda}}\left(y,\tilde{\lambda}\right)\right) &= \mu\left(\begin{pmatrix} \mathbb{1} - \frac{d^2}{c} G(y) M(y)^{-1} G(y)^T & 0 \\ 0 & -\mathbb{1} \end{pmatrix}\right) \\
&= \max\left\{-1, \mu\left(\mathbb{1} - \frac{d^2}{c} G(y) M(y)^{-1} G(y)^T\right)\right\} \\
&= \max\left\{-1, 1 + \frac{d^2}{c}\mu\left(-G(y) M(y)^{-1} G(y)^T\right)\right\}.
\end{aligned}$$

Damit ergibt als Bedingung für unser Modellproblem in einer Lösungsumgebung:

$$\mu\left(\tilde{g}_{\tilde{\lambda}}\right) < 0 \quad \Leftrightarrow \quad -\mu\left(-GM^{-1}G^T\right) > \frac{c}{d^2}$$

Hiermit können wir Theorem 6 nutzen, um die Aussage aus [Dep13] mit Einschränkung an die Parameter auch für den Fall Principal Damping zu beweisen.

Korollar 4 *Konvergenzbeweis im Falle Principal Damping bei geeigneter Wahl der Parameter*
Gegeben sei das Modellproblem in der Form (3.57),(3.58) und erfülle alle Modellannahmen. Es seien Anfangswerte in der Form

$$\begin{aligned}
y(0) &= \sum_{i=0}^{\infty} \epsilon^{i/2} y_i^0 \\
\lambda(0) &= \sum_{i=0}^{\infty} \epsilon^{i/2} \lambda_i^0
\end{aligned}$$

gegeben. In einer von ϵ unabhängigen Umgebung um die Lösung y_0, λ_0 des reduzierten Problems gelte

$$-\mu\left(-GM^{-1}G^T\right) > \frac{c}{d^2}, \tag{3.61}$$

insbesondere liegen die Anfangswerte y_0^0, λ_0^0 in dieser Umgebung. Dann besitzt (3.57),(3.58) auf einem Zeitintervall $[0,\bar{t}]$ eine eindeutige Lösung der Form

$$\begin{aligned}
y(t) &= y_0(t) + \mathcal{O}(\sqrt{\epsilon}), \\
\lambda(t) &= \lambda_0(t) + \mathcal{O}(\sqrt{\epsilon}).
\end{aligned}$$

Beweis: Wie gezeigt wurde, ist das Problem (3.57),(3.58) äquivalent zu (3.60). Hierbei sind die Anfangswerte für θ aus (3.59) festgelegt, weiterhin ist das reduzierte Problem von (3.60) äquivalent zur ursprünglichen Differential-Algebraischen-Gleichung. Außerdem wurde gezeigt, dass das Erfüllen von (3.61) impliziert, dass $\tilde{g}_{\tilde{\lambda}}$ negativ ist. Mit Substitution von $\tilde{\epsilon}$ durch $\sqrt{\epsilon}$ sind damit alle Voraussetzungen für Theorem 6 mit $\tilde{\epsilon}$ erfüllt. Das Theorem liefert mit $N = 0$ die Existenz einer eindeutigen Lösungen der Form

$$\begin{aligned} y(t) &= y_0(t) + \mathcal{O}(\tilde{\epsilon}), \\ \lambda(t) &= \lambda_0(t) + \mathcal{O}(\tilde{\epsilon}), \\ \theta(t) &= \theta_0(t) + \mathcal{O}(\tilde{\epsilon}). \end{aligned}$$

Rücksubstitution von ϵ ergibt für y und λ die Aussage des Korollars. □

Bemerkung 3 *Die Erfüllung von* (3.61) *impliziert zunächst nur die Existenz einer negativen oberen Schranke für* $\mu(\tilde{g}_{\tilde{\lambda}})$. *Für Theorem 6 wird eigentlich die obere Schranke* -1 *benötigt. Dies lässt sich jedoch ganz analog zum Fall Strong Damping durch Neuskalierung des Problems* (3.60) *erreichen. Wichtig ist in diesem Kontext nur, dass es eine von* ϵ *unabhängige negative Schranke gibt.*

Weiterhin stellt Korollar 4 keine Verallgemeinerung von Theorem 6 dar, da die Darstellung der Nebenbedingung in der Form $0 = g(y) + \sqrt{\epsilon}\dot{g}(y)$ *entscheidend ist. Für Probleme in der allgemeinen Form* (3.57), (3.58) *und ohne die Eigenschaft* $\frac{d}{dt}g(y) = h(y, \lambda)$ *ist das Korollar 4 nicht anwendbar. Hierfür ist Korollar 3 verwendbar, allerdings unter deutlich schärferen Zusatzbedingungen.*

3.4 Runge-Kutta-Verfahren für Principal Damping und Strong Damping

Die bisher gezeigten Resultate für Principal Damping und Strong Damping beinhalten nur eine analytische Betrachtung der Lösung. Zwar wurde gezeigt, dass die Lösungen der regularisierten Probleme gegenüber der Lösung der ursprünglichen Differential-Algebraischen-Gleichung einen Fehler der Größenordnung $\mathcal{O}(\epsilon)$, respektive $\mathcal{O}(\sqrt{\epsilon})$ besitzen, jedoch führt die Regularisierung zu steifen Differentialgleichungen. Dies könnte bei einer numerischen Lösung der Probleme die gewonnenen Größenordnungen in ϵ zunichte machen. Allerdings gibt es Runge-Kutta-Verfahren, die diese Größenordnungen erhalten. Aussagen hierüber macht ein Theorem aus [Hai10], welches an dieser Stelle zitiert wird.

Hierfür wird allerdings noch die Definition von *A-Stabilität* benötigt.

Definition 3.5 *A-Stabilität*
Gegeben sei die Differentialgleichung $\dot{y} = \alpha y$, wobei α einen negativen Realteil besitze.
Ein numerisches Verfahren heißt *A-stabil*, falls dessen numerische Lösung dieser Differentialgleichung bei beliebiger aber konstanter Schrittweite beschränkt bleibt.

Damit lässt sich folgendes Theorem formulieren:

Theorem 7 *Korollar VI.3.10 aus [Hai10]*
Gegeben sei das Problem (3.1) *mit* $\mu\left(g_\lambda\right) \leq -1$. *Zu gegebenen Anfangswerten* $(y(0), \lambda(0))$ *existiere eine eindeutige, glatte Lösung* (y, λ). *Auf dieses werde das* s*-stufige, A-stabile Runge-Kutta-Verfahren*

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

mit Ordnung p und Stufenordnung $q < p$ angewendet, wobei A nur Eigenwerte mit positivem Realteil besitze. Weiterhin gelte für die Stabilitätsfunktion R des Runge-Kutta-Verfahrens, dass $|R(\infty)| < 1$. Dann gilt für den globalen Fehler nach n Schritten mit Schrittweite h:

$$y_n - y(nh) = \mathcal{O}\left(h^p\right) + \mathcal{O}\left(\epsilon h^{q+1}\right)$$
$$\lambda_n - \lambda(nh) = \mathcal{O}\left(h^{q+1}\right)$$

Ist die Methode weiterhin steif-genau, also $a_{si} = b_i$ für $i = 1, \ldots, s$, so folgt zusätzlich

$$\lambda_n - \lambda(nh) = \mathcal{O}\left(h^p\right) + \mathcal{O}\left(\epsilon h^q\right).$$

Bemerkung 4 *Mit diesem Theorem ist gezeigt, dass sich für hinreichend kleine Schrittweite h zunächst beim Fall Strong Damping die Fehlergrößenordnung ϵ von der analytischen Lösung auf die numerische Lösung überträgt. Im Falle Principal Damping muss hierfür wieder der Umweg über $\tilde{g}$ aus Kapitel 3.3.2 mit Substitution von ϵ gegangen werden. Damit folgt in diesem Fall die Größenordnung $\sqrt{\epsilon}$. Diese Größenordnungen sollten damit im nächsten Kapitel durch numerische Experimente mit Runge-Kutta-Verfahren verifiziert werden können.*

Kapitel 4

Numerische Experimente

In diesem Kapitel werden die Aussagen der vorherigen Kapitel durch ein vergleichsweise anschauliches numerisches Beispiel überprüft. Von besonderem Interesse sind die Konvergenzordnungen von Runge-Kutta-Verfahren für Differential-Algebraische-Gleichungen. Hierfür werden drei Kollokationsverfahren (implizites Euler-Verfahren, sowie das zwei- und dreistufige RadauIIA-Verfahren) auf die unregularisierte, Differential-Algebraische-Gleichung angewendet. Weiterhin wird das Fehlerverhalten der Regularisierungen Strong und Principal Damping untersucht. Zunächst wird das betrachtete numerische Modellbeispiel aufgestellt und auf die Erfüllung aller Modellannahmen überprüft.

4.1 Versuchsaufbau

4.1.1 Modellproblem

Als relativ einfaches Beispiel eines Mehrkörpersystems wird die bereits in [Dep13] betrachtete, reibungsfrei auf einer planen Ebene im Erdgravitationsfeld rollende Scheibe herangezogen, siehe Abb. 4.1.

Die Bewegung dieser rollenden Scheibe wird vollständig durch die 5 eingezeichneten verallgemeinerten Koordinaten $q := (x, y, \alpha, \beta, \gamma)$ beschrieben, wobei x und y in der $x_{\text{Lab}}y_{\text{Lab}}$-Ebene des Laborsystems liegen, x parallel zur Scheibenebene ist und y senkrecht zu x. Die Scheibe werde als zweidimensional mit homogener Massenverteilung betrachtet, ihr Radius sei R, ihre Masse m.
Mit diesen verallgemeinerten Koordinaten lassen sich die kinetische und potentielle Energie aufstellen. Damit ergeben sich, wie im Grundlagenkapitel beschrieben, mit dem Lagrangeformalismus die Bewegungsgleichungen. Zunächst stellen wir den Term für die kinetische Energie auf, diese setzt sich aus der kinetischen Energie des Masseschwerpunktes (Mittelpunkt der Scheibe, abgekürzt mit „CM“ für „Center of Mass“) und der Rotationsenergie der Scheibe im Schwerpunktssystem zusammen. Aus der Skizze lassen sich die Koordinaten des Masseschwerpunktes im Laborsystem ablesen:

$$\vec{r}_{\text{Lab}} := \begin{pmatrix} x_{\text{Lab}} \\ y_{\text{Lab}} \\ z_{\text{Lab}} \end{pmatrix}_{\text{CM}} = \begin{pmatrix} \cos(\alpha)x - \sin(\alpha)(y - \sin(\beta)R) \\ \sin(\alpha)x + \cos(\alpha)(y - \sin(\beta)R) \\ \cos(\beta)R \end{pmatrix}.$$

Entsprechend gibt sich für die kinetische Energie des Masseschwerpunktes

$$E_{\text{kin,CM}} = \frac{1}{2}\dot{\vec{r}}_{\text{Lab}}^{T} M \dot{\vec{r}}_{\text{Lab}}.$$

Die Rotationsenergie ist gegeben durch

$$E_{\text{rot}} = \frac{1}{2}\omega^T J \omega.$$

Hierbei ist J der Trägheitstensor, in unserem Fall (homogene Kreisscheibe) der Form

$$J = mR^2 \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$$

und ω ist der Vektor mit den Winkelgeschwindigkeiten um die jeweiligen Hauptträgheitsachsen. Diese lassen sich auch wieder aus der Skizze ablesen als

$$\omega = \begin{pmatrix} \dot{\beta} \\ \dot{\gamma} + \sin(\beta)\dot{\alpha} \\ \cos(\beta)\dot{\alpha} \end{pmatrix}.$$

Die gesamte kinetische Energie ist dann $E_{\text{kin,CM}} + E_{\text{rot}}$. Die Potentielle Energie ist ausschließlich durch die Gravitation gegeben, sie berechnet sich aus der Höhe des Masseschwerpunktes und mit der Gravitationskonstante $g = 9.81\text{N/kg}$, also

$$E_{\text{pot}} = mgR\cos(\beta).$$

Setzt man dies in (1.6) ein, ergibt sich nach längerer Rechnung die ebenfalls in [Dep13] aufgestellte Bewegungsgleichung $M(q)\ddot{q} = F(q, \dot{q})$, wobei

$$M(q) = m \begin{pmatrix} 1 & 0 & R\sin(\beta) - y & 0 & 0 \\ 0 & 1 & x & -\cos(\beta)R & 0 \\ R\sin(\beta) - y & x & x^2 - \frac{5R^2}{4}\cos(\beta)^2 + y^2 + \frac{3R^2}{2} - 2Ry\sin(\beta) & -x\cos(\beta)R & \frac{R^2}{2}\sin\beta \\ 0 & -\cos(\beta)R & -x\cos(\beta)R & \frac{5}{4}R^2 & 0 \\ 0 & 0 & \frac{R^2}{2}\sin(\beta) & 0 & \frac{R^2}{2} \end{pmatrix}$$

und

$$F(q, \dot{q}) = m \begin{pmatrix} \dot{\alpha}(2\dot{y} - 2\dot{\beta}R\cos(\beta) + \dot{\alpha}x) \\ -R(\dot{\alpha}^2 + \dot{\beta}^2)\sin(\beta) - 2\dot{\alpha}\dot{x} + \dot{\alpha}^2 y \\ -\frac{1}{2}(5\sin(\beta)\dot{\alpha}R + R\dot{\gamma} - 4\dot{\alpha}y)\dot{\beta}R\cos(\beta) + R(2\dot{\alpha}\dot{y} - \dot{\beta}^2 x)\sin(\beta) - 2\dot{\alpha}(x\dot{x} + y\dot{y}) \\ R(\frac{1}{4}(5\sin(\beta)\dot{\alpha}R - 4\dot{\alpha}y + 8\dot{x} + 2R\dot{\gamma})\dot{\alpha}\cos(\beta) + g\sin(\beta)) \\ -\frac{R^2}{2}\dot{\alpha}\cos(\beta)\dot{\beta}. \end{pmatrix}$$

Nun zu den noch fehlenden Zwangsbedingungen: Ein „reibungsfreies“ Rollen bedeutet hier, dass die Geschwindigkeit des Auflagepunktes (beschrieben durch α, x und y) im Laborsystem mit der Bewegung der Scheibe übereinstimmt. Die Zwangsbedingung, dass die Scheibe immer auf der Ebene bleibt, diese nie verlässt oder durchdringt, ist durch die Wahl der verallgemeinerten Koordinaten bereits gewährleistet. Bestenfalls könnte noch verlangt werden, dass $\beta \in [-\pi, \pi]$). Dies führt zu den Bedingungen

$$\dot{x} = R\dot{\gamma} + y\dot{\alpha}$$
$$\dot{y} = -x\dot{\alpha}.$$

Anschaulich dürfen sich x und y nur genau so ändern, wie es durch Rollen ($R\dot{\gamma}$) oder Drehen der Scheibe im Winkel α ($y\dot{\alpha}$ und $-x\dot{\alpha}$) erreicht wird. Somit lässt sich die Matrix $G = G(q)$ aufstellen als

$$G(q) = \begin{pmatrix} 1 & 0 & -y & 0 & -R \\ 0 & 1 & x & 0 & 0 \end{pmatrix},$$

da damit $G(q)\dot{q}$ gerade obige Bedingungen ergibt.

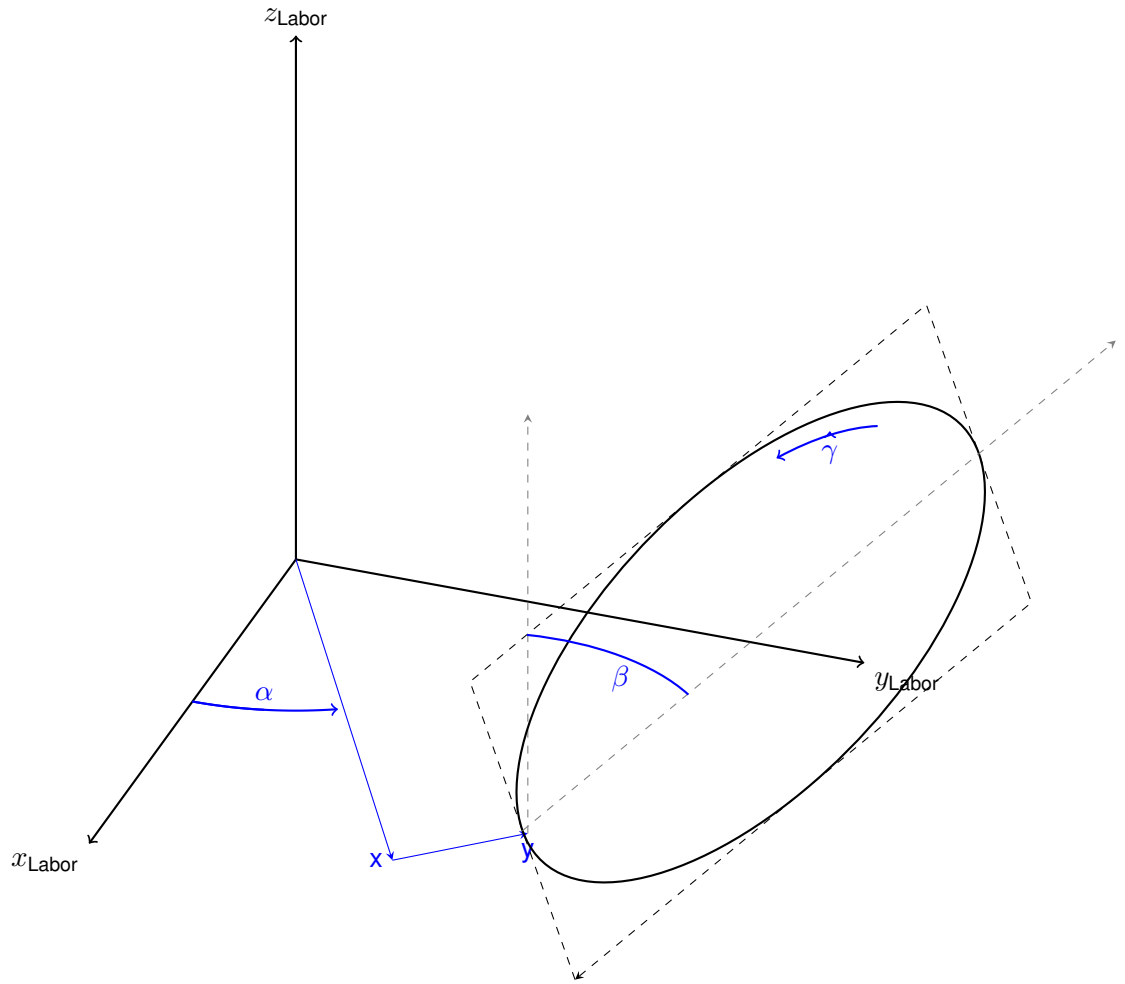

Abb. 4.1: Rollende Platte

Nun zum Überprüfen der Modellannahmen 1 bis 5. Als Erstes gilt es zu verifizieren, dass die Matrix $M(q)$ in hinreichender Lösungsumgebung symmetrisch und positiv definit ist. Aus physikalischer Sicht ist die positive Definitheit offensichtlich, da die kinetische Gesamtenergie des Systems durch $\dot{y}^T M(y)\dot{y}$ berechnet wird und die kinetische Energie immer positiv ist. Dennoch wird es hier zum Test überprüft. Die Symmetrie der Matrix ist offensichtlich, für die positive Definitheit kann das Hauptminorenkriterium genutzt werden. Die fünf ausgerechneten und vereinfachten Hauptminoren lauten

1. m,
2. m,
3. $m\frac{R^2}{4}\left(2-\cos(\beta)^2\right)$,
4. $m\frac{R^4}{16}\left(2-\cos(\beta)^2\right)\left(5-4\cos(\beta)^2\right)$,
5. $m\frac{R^6}{32}\cos(\beta)^2\left(5-4\cos(\beta)^2\right)$.

Diese sind unabhängig von den eingesetzten Werten positiv, damit ist $M(q)$ in jeder beliebigen Lösungsumgebung positiv definit. Weiterhin hat $G(q)$ dank der ersten beiden Spalten ebenfalls un-

abhängig von q den vollen Rang 2. Für Modellannahme 3 gilt, dass

$$\begin{aligned}\left(\frac{\partial}{\partial(q,\dot{q})}G(q)\dot{q}\right)\frac{\partial}{\partial\lambda}\left(\dot{q},M(q)^{-1}\left(F(q,\dot{q})-G(q)^T\lambda\right)\right)^T &= \left(\frac{\partial G(q)\dot{q}}{\partial q},G(q)\right)\left(0,-M(q)^{-1}G(q)^T\right)^T\\ &= -G(q)M(q)^{-1}G(q)^T.\end{aligned}$$

Dies ist, wie in (3.43) gezeigt, negativ definit, damit eine quadratische Matrix von vollem Rang, also invertierbar. Weiterhin sind die Abhängigkeiten von F, M und G von q und $\dot{q}$ polynomiell, damit hinreichend glatt, beschränkt und lokal Lipschitz-stetig in jeder beschränkten Umgebung. Daraus folgt insbesondere $M(q+\delta q) = M(q) + \delta M$ mit einem δM, für das die Abschätzung $\|\delta M\| \leq c\,\|\delta q\|$ mit einer auf jeder beschränkten Umgebung von q endlichen c gilt. Nutzt man Satz 4.50 aus [Pla10], folgt außerdem für $M(q)^{-1}$:

$$\begin{aligned}\left\|\frac{\partial M(q)^{-1}}{\partial q}\right\| &\leq \limsup_{\delta q\to 0}\frac{\left\|M(q+\delta q)^{-1}-M(q)\right\|}{\|\delta q\|}\\ &= \limsup_{\delta q\to 0}\frac{\left\|(M(q)+\delta M))^{-1}-M(q)\right\|}{\|\delta q\|}\\ &\overset{\text{Satz 4.50}}{\leq} \limsup_{\delta q\to 0}\frac{2\,\|\delta M\|\left\|M(q)^{-1}\right\|^2}{\|\delta q\|}\\ &\overset{\text{s.o.}}{\leq} \limsup_{\delta q\to 0}\frac{2c\,\|\delta q\|\left\|M(q)^{-1}\right\|^2}{\|\delta q\|}\\ &\leq 2c\left\|M(q)^{-1}\right\|^2 < \infty\end{aligned}$$

Damit folgen die restlichen Modellannahmen.
Weiterhin kann man $-G(q)M(q)^{-1}G(q)^T$ berechnen um Bedingungen für die Parameter c und d für den Fall Principal Damping zu erhalten. Es ergibt sich nach Vereinfachung der anfallenden Terme

$$-G(q)M(q)^{-1}G(q)^T = \begin{pmatrix}-3 & 0\\ 0 & \frac{-5}{5-4\cos(\beta)^2}\end{pmatrix}.$$

Der maximal erreichbare Eigenwert hiervon ist -1, damit folgt als Bedingung $\frac{d^2}{c} > 1$. Im Folgenden wird $c = d = 2$ gewählt. Weiterhin sei $m = 1kg$ und $R = 10cm$.

Somit ist das Problem in der Form (1.7), (1.9) mit den Annahmen konform und kann wie beschrieben nach Regularisierung oder direkt nach Umformung auf ein System erster Ordnung gelöst werden. Als Anfangswerte werden zunächst q und $\dot{q}$ gesetzt als

$$\begin{aligned}q(0) &= (0,0,0,\frac{5\pi}{180},0),\\ \dot{q}(0) &= (R\pi,0,0,0,\pi),\end{aligned}$$

und schließlich $\lambda(0)$ aus der Differential-Algebraischen-Gleichung berechnet (einmaliges Ableiten der algebraischen Bedingung und Nutzen von Modellannahme 3 liefert eine nach λ auflösbare Gleichung).

4.1.2 Die verwendeten Verfahren

Als Runge-Kutta-Verfahren mit der höchsten hier verwendeten Konvergenzordnug betrachten wir die *RadauIIA5*-Methode, ein 3-stufiges implizites Runge-Kutta-Verfahren der Ordnung 5 mit dem Butcher-

Tableau

$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

Nachrechnen ergibt, dass die Verfahrensmatrix nur Eigenwerte mit positiven Realteilen hat. Da dieses Runge-Kutta-Verfahren äquivalent zu einer Kollokationsmethode ist (vgl. [Hoc12], Kapitel 10), hat es nach Korollar 1 also für unsere Differential-Algebraische-Gleichung mindestens die Ordnung 3. Direktes Nachrechnen der Bedingungen aus Theorem 2 verifiziert dies, insbesondere sind $B(1)$ bis $B(5)$ erfüllt, allerdings nur $C(1)$ bis $C(3)$. Bei $C(4)$ gilt etwa

$$\sum_{j=1}^{3} a_{1j} c_1^3 \approx 0.0069 > \frac{c_1^4}{4} \approx 0.00014.$$

Dennoch folgt aus Theorem 2 eine Konvergenzordnung von 5 bzgl. der y-Komponente und nach wie vor mindestens Ordnung 3 für die λ-Komponente, da dieses Verfahren eine Kollokationsmethode mit Konvergenzordnung 5 ist.

Dieses Runge-Kutta-Verfahren wurde, wie in [HLR89] und [Hai10] vorgestellt, bereits in Fortran-Code mit Schrittweitensteuerung und einigen Optionen für Differential-Algebraische-Gleichungen implementiert und von Ch. Engstler in Matlab-Code übertragen. Diese Matlab-Implementierung wurde für die folgenden numerischen Experimente übernommen, sowohl für die ursprüngliche Differential-Algebraische-Gleichung, als auch für die regularisierten Probleme. Für Tests der Konvergenzordnung musste nur eine kleine Modifikation vorgenommen werden, um konstante Schrittweiten zu ermöglichen.

Weiterhin untersuchen wir die beiden zu Kollokationsverfahren äquivalenten Verfahren RadauIIA1 (implizites Euler-Verfahren) und RadauIIA3, also implizite 1 bzw. 2-stufige Runge-Kutta-Verfahren mit Butchertableaus

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}.$$

und

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}.$$

Die Realteile der Eigenwerte ergeben sich als 1 bzw. $1/3$ und sind damit alle positiv. Wiederum kann man die Bedingungen B und C nachrechnen und erhält Mindestordnungen von 1 und 2. Eine Verbesserung durch Äquivalenz zu einem Kollokationsverfahren ergibt sich im Falle des impliziten Eulerverfahrens nicht, da damit bereits die volle Verfahrensordnung 1 ausgeschöpft ist und Korollar 2 nur maximal eine Ordnung für Differential-Algebraische-Gleichungen liefert, die der gewöhnlichen Konvergenzordnung entspricht. Für das RadauIIA3 Verfahren ergibt sich jedoch eine Verbesserung auf Ordnung 3 für die y-Komponente. Weiterhin sind alle RadauIIA-Verfahren nach [Hoc12] A-stabil, somit ist Theorem 7 anwendbar und die Verfahren sollten die ϵ- bzw. $\sqrt{\epsilon}$-Abhängigkeit der Fehler der regularisierten Lösungen wiedergeben.

Um die Verfahren auf Differential-Algebraische-Gleichungen anzuwenden, wurde dieser modifizierte RadauIIA5-Code mit konstanter Schrittweite und jeweils angepasster Verfahrensmatrix genutzt.

4.2 Ergebnisse

Nachdem alle Grundlagen und Theorie besprochen und das Modellbeispiel formuliert wurde, können die numerischen Experimente durchgeführt werden. Zunächst die Lösung der unregularisierten Differential-Algebraischen-Gleichung via Radau5-Code mit variabler Schrittweitensteuerung und sehr geringer Fehlertoleranz (absolute und relative Toleranz 10^{-12}). Dies wird im Folgenden als Referenzlösung dienen.

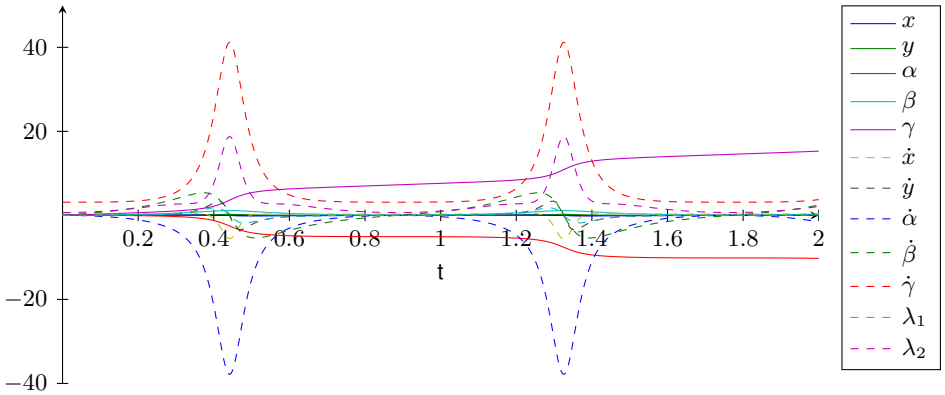

Abb. 4.2: Lösung der Differential-Algebraischen-Gleichung

Hierbei sind alle Variablen geplottet. Man erkennt bei einigen eine periodische Bewegung, was zu erwarten war für eine Platte, die sich auf der Oberfläche dreht. Um dieses etwas unübersichtliche Ergebnis etwas anschaulicher zu machen und um zu testen, ob dieses Ergebnis physikalisch Sinn machen kann, bietet sich als Test die kinetische und potentielle Energie an. Da die Differential-Algebraische-Gleichung eine gleitreibungsfreie Bewegung darstellt und im Modellbeispiel keine nichtkonservativen äußeren Kräfte wirken, sollte die Gesamtenergie zeitlich konstant bleiben.

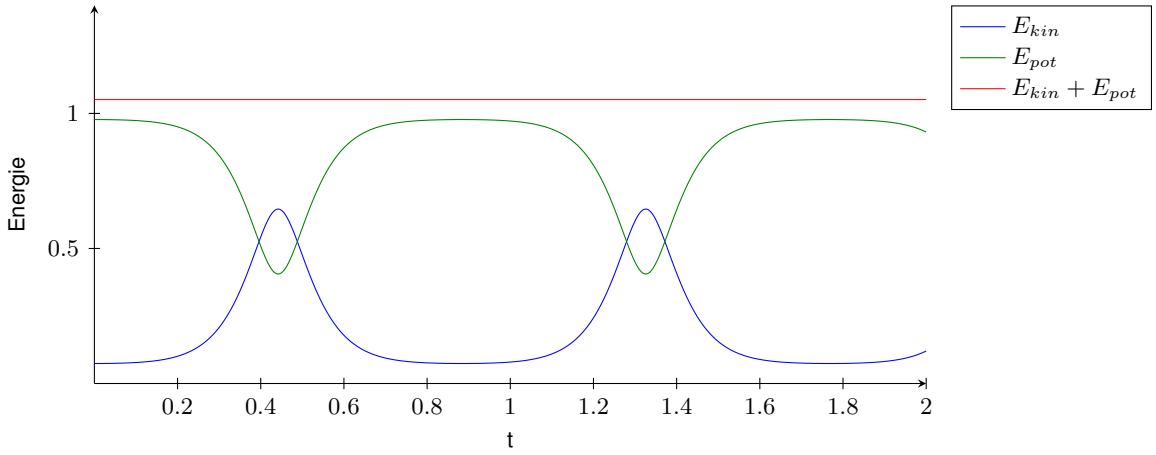

Abb. 4.3: Erhaltung der Gesamtenergie

Wie erwartet ist die Gesamtenergie des Systems konstant. Weiterhin erkennt man auch hier das *Taumeln* der Platte, die sich abwechselnd aufrichtet (hohe Potentielle Energie) und dann wieder kippt (niedrige Potentielle, aber hohe kinetische Energie).

Nun kann man überprüfen, ob die Zwangsbedingung von der numerischen Lösung auch tatsächlich eingehalten wird. Bei der Lösung der unregularisierten und unreduzierten Differential-Algebraischen-Gleichung ist dies nach Konstruktion zu erwarten, da in jedem Schritt ein nichtlineares Gleichungssystem inklusive der Zwangsbedingung gelöst wird, wie man in Abbildung 4.4 sieht:

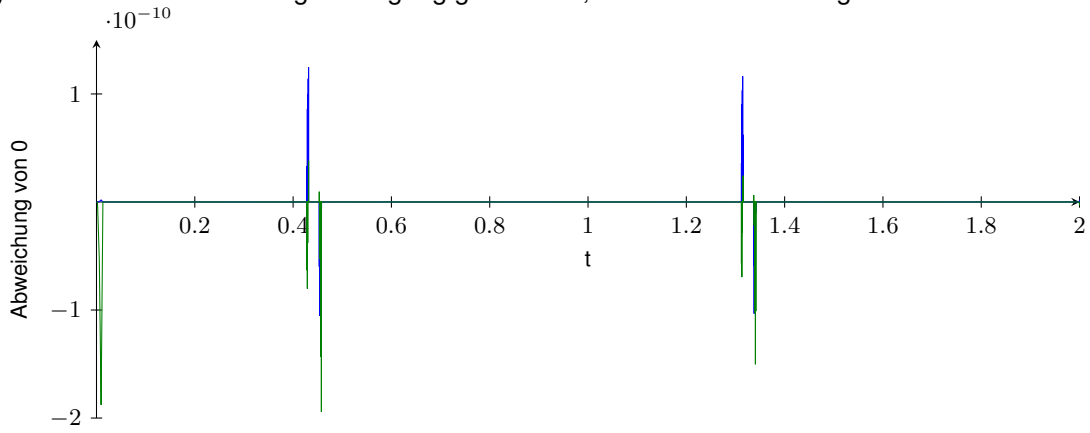

Abb. 4.4: Verletzung der Zwangsbedingung bei Integration der Differential-Algebraischen-Gleichung. Zu sehen sind beide Komponenten der Zwangsbedingung.

Von einigen wenigen Abweichungen in der Größenordnung 10^{-10} abgesehen, geht die Verletzung der Zwangsbedingung im Rundungsfehler unter. Als Kontrast hierzu kann man die Differential-Algebraische-Gleichung vor der Integration durch Indexreduktion, respektive Elimination der λ-Komponente nach (3.17) auf eine explizite Form bringen. Damit wird die Zwangsbedingung beim Lösen des nichtlinearen Gleichungssystems innerhalb der Schritte des Runge-Kutta-Verfahrens nicht mehr explizit berücksichtigt. Die Verletzung der Zwangsbedingungen ist in Abbildung 4.5 zu sehen.

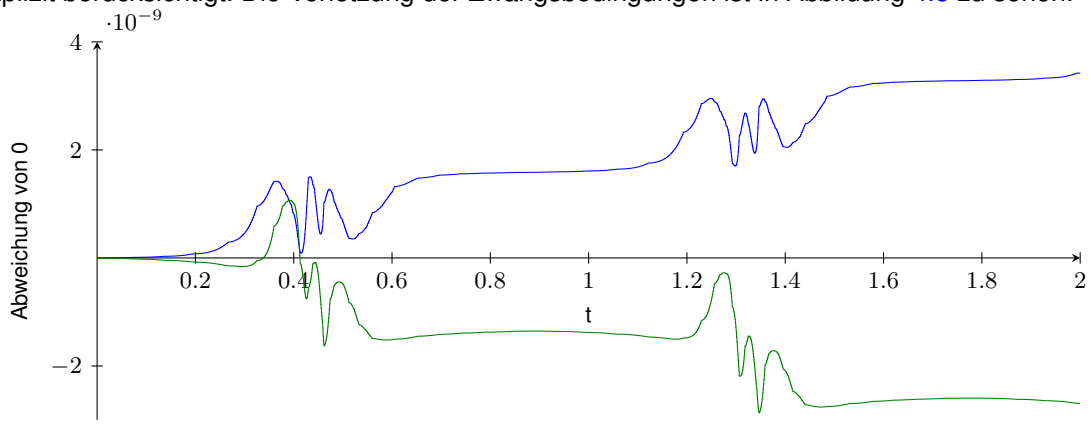

Abb. 4.5: Verletzung der Zwangsbedingungen nach Indexreduktion, beide Komponenten.

Es ergibt sich ein numerischer Drift, dank Verfahrensfehler und Rundungsungenauigkeit, durch welche die Zwangsbedingung mehr und mehr verletzt wird, es also eine immer höhere Relativgeschwindigkeit der Auflageflächen zueinander gibt. Dies ist aus physikalischer Sicht verheerend, da das der Differential-Algebraischen-Gleichung zu Grunde liegende physikalische Modell explizit keine Relativgeschwindigkeit erlaubt. Die numerische Lösung wäre damit bestenfalls Lösung einer Differentialgleichung eines gänzlich anderen Typs (Gleitreibung).
Weiterhin interessant ist für die folgenden numerischen Versuche noch der Schrittweitenverlauf der Referenzlösung, um eine untere Schranke für die Schrittweite bei den Konvergenzuntersuchungen zu erhalten. Es ergibt sich folgender Verlauf:

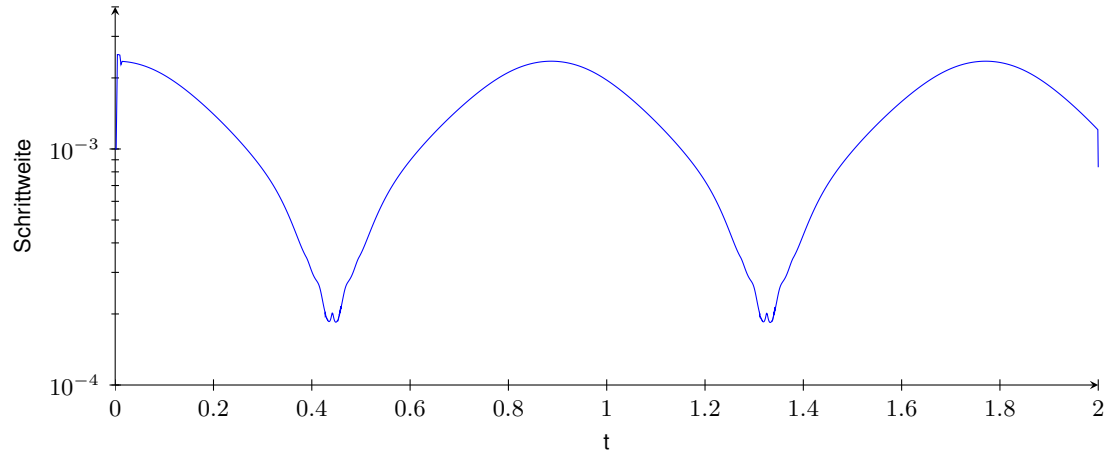

Abb. 4.6: Schrittweitenverlauf der Referenzlösung

Das Verfahren benötigt bei sehr geringer Toleranz also Schrittweiten zwischen 10^{-4} und 10^{-3}, damit ist die untere Grenze für einen Konvergenztest festgelegt. Die obere Grenze ergibt sich aus praktischen Gründen als größtmögliche Schrittweite, bei der das nichtlineare Gleichungssystem des Verfahrens noch mit überschaubarer Schrittzahl des vereinfachten Newtonverfahrens gelöst wird. Dies ist, wie sich zeigt, etwa bei $10^{-1.5}$ der Fall, da hier das Verfahren mehr als 200 Schritte der vereinfachten Newton-Iteration benötigt (Ein üblicher default-Wert mit aktiver Schrittweitensteuerung ist 7). Damit können wir die Konvergenzordnung des Radau5-Codes anhand der Modell-Differential-Algebraischen-Gleichung testen. Hierzu vergleichen wir das Ergebnis des Radau5-Codes bei verschiedenen konstanten Schrittweiten mit dem Ergebnis der Referenzlösung in Abb. 4.2. Es ergibt sich für die y-Komponente Abb. 4.7. Die in den doppellogarithmisch aufgetragenen Plot eingepasste Gerade hat hierbei die Steigung 4.82. Dies bestätigt die vorhergesagte Konvergenzordnung 5. Führt man Selbes für die λ-Koordinate durch, zu sehen in Abbildung 4.8, erhält man für die Ausgleichsgerade eine Steigung von 4.79. Zu erwarten wäre eigentlich die Konvergenzordnung 3 gewesen, damit übertrifft der Radau5-Code bei diesem Modellbeispiel sogar die zu erwartende Konvergenzordnung für die λ-Komponente.

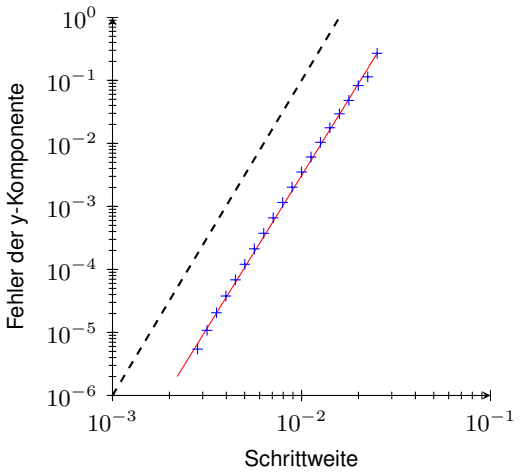

Abb. 4.7: Radau5 Konvergenzordnung der y-Komponente

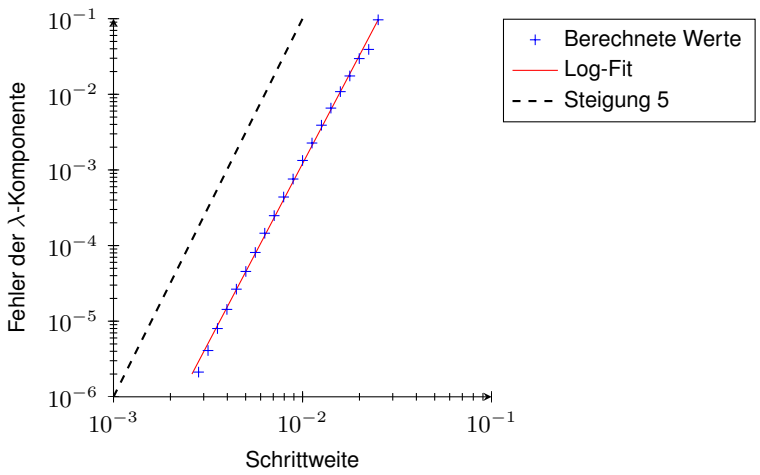

Abb. 4.8: Radau5 Konvergenzordnung der λ-Komponente

Als zweites Runge-Kutta-Verfahren wird das implizite Euler-Verfahren getestet, zu erwarten sind Konvergenzordnung 1 für y- und λ-Komponente.

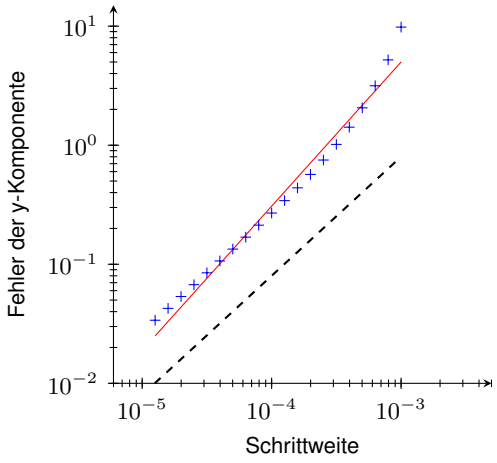

Abb. 4.9: Impliziter Euler: Konvergenzordnung der y-Komponente

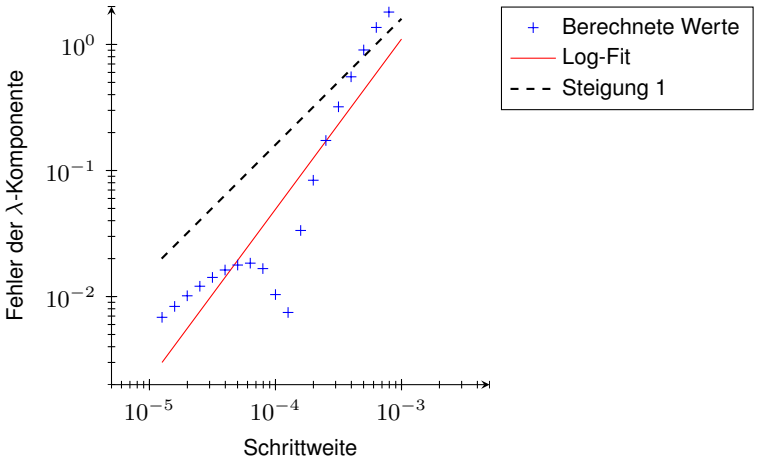

Abb. 4.10: Impliziter Euler: Konvergenzordnung der λ-Komponente

Es ergibt sich eine Steigung der Ausgleichsgerade von 1.198 bei der y- und 1.379 bei der λ-Komponente, also in etwa das Erwartete. Bemerkenswert ist in diesem Kontext allerdings das seltsame Verhalten des Fehlers der λ-Komponente im Bereich der Schrittweite 10^{-4}, bei der ein im Vergleich niedriger Fehler auftritt. Unabhängig davon ist der Fehler allgemein um einige Größenordnungen größer als beim RadauIIA5-Verfahren, was ja aber auch zu erwarten war.
Nun zum RadauIIA3-Verfahren, zu erwarten sind Konvergenzordnung 3 für die y- und 2 für die λ-Komponente.

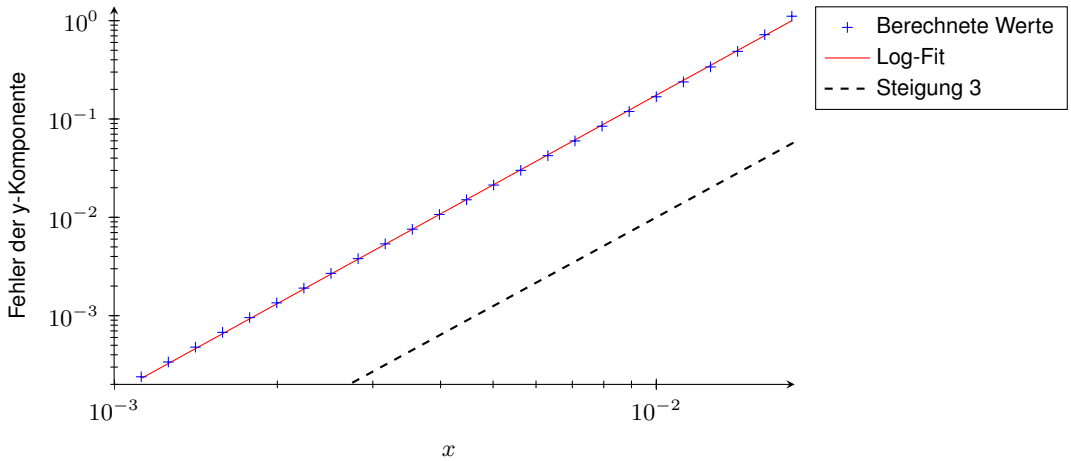

Abb. 4.11: RadauIIA3: Konvergenzordnung der y-Komponente

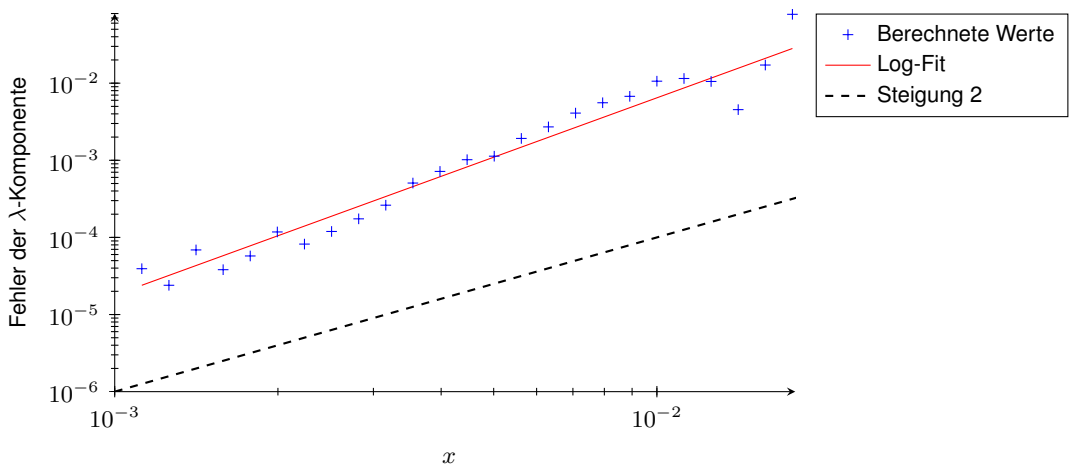

Abb. 4.12: RadauIIA3: Konvergenzordnung der λ-Komponente

Die Ausgleichsgeraden haben Steigungen von 3.016 bei der y- und 2.661 bei der λ-Komponente. Damit ist die Vorhersage aus Korollar 2 recht gut erfüllt. Der Versuch zeigt zwar wieder eine leicht bessere Konvergenzordnung bezüglich der λ-Komponente, wie schon bei RadauIIA5 gesehen, jedoch deutlich geringer ausgeprägt.

Schließlich zur Betrachtung der numerischen Ergebnisse der Regularisierung. Die Eigenschaft als neues physikalisches Modell kann hierbei jedoch nicht untersucht werden, da eine Referenzlösung fehlt. An dieser Stelle müsste man einen reellen Versuch starten. Allerdings kann man testen, ob –wie vorhergesagt– der Grenzfall für $\epsilon \to 0$ wieder die ursprüngliche Differential-Algebraische-Gleichung ist. Hierfür wird analog zu oben die Abweichung der bei ebenfalls geringer Fehlertoleranz mit dem Radau5-Code gelösten, regularisierten Differentialgleichung in der Form (3.7), (3.8) mit der Lösung der unregularisierten Differential-Algebraischen-Gleichung verglichen. Dies geschieht für verschiedene ϵ.

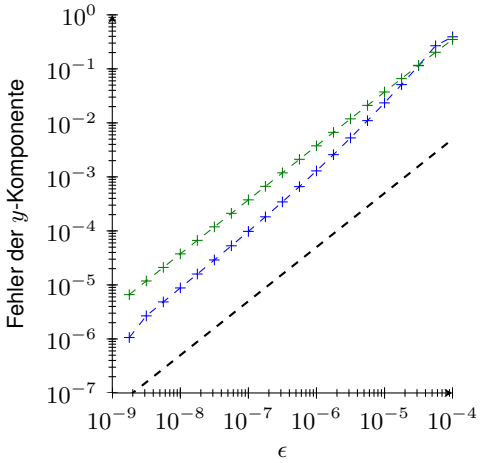

Abb. 4.13: Konvergenz der y-Komponente

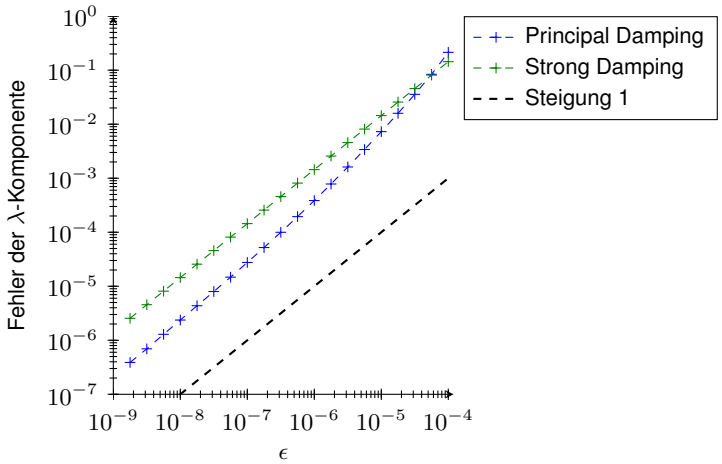

Abb. 4.14: ... und der λ-Komponente bzgl. ϵ

Hierbei ergeben die linearen Fits an die doppellogarithmisch aufgetragenen Fehlerplots Steigungen von 1.17 für die y- und 1.19 für die λ-Komponente im Falle Principal Damping, sowie 0.99 für die y- und 1.00 für die λ-Komponente im Falle Strong Damping.
Der Fall Strong Damping ist hiermit bestätigt, es war eine Abweichung zur Differential-Algebraischen-Gleichung der Größenordnung $\mathcal{O}(\epsilon)$ in jeder Komponente vorhergesagt, dieses trifft mit sehr geringer

Abweichung zu. Der Fall Principal Damping jedoch weicht stark von der Vorhersage ab, statt $\mathcal{O}(\sqrt{\epsilon})$ liegt ebenfalls eine $\mathcal{O}(\epsilon)$-Abhängigkeit vor. Betrachtet man nun noch die Verletzung der Zwangsbedingung in Abhängigkeit von ϵ, zeigt sich ein weiteres Problem für die Anwendung von Korollar 3:

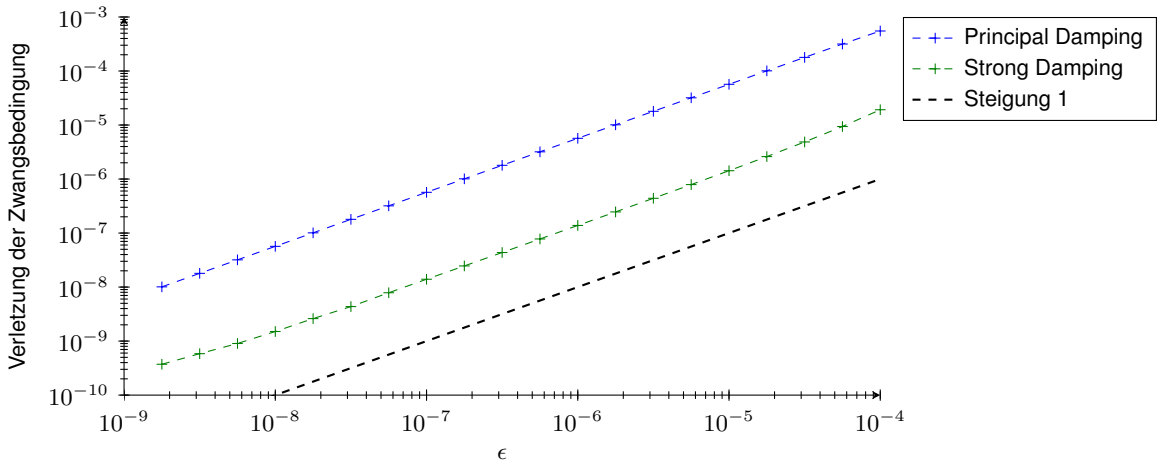

Abb. 4.15: Verletzung der Zwangsbed. der regularisierten Differential-Algebraischen-Gleichung

Als Steigungen der Log-Fits ergeben sich 0.99 für den Fall Strong Damping und 0.99 bei Principal Damping. Da damit bei Principal Damping sowohl $g(y) = G(q)\dot{q}$ in obiger Notation (also y_0 als Lösung der Differential-Algebraischen-Gleichung), als auch der Fehler der y-Komponente $y - y_0$ von der Größenordnung $\mathcal{O}(\epsilon)$ sind, folgt hieraus $g(y) = \mathcal{O}(y - y_0)$ und nicht $g(y) = \mathcal{O}(\sqrt{\epsilon}(y - y_0))$. Damit sind die Voraussetzungen für Satz 3 nicht erfüllt. Zwar folgt aus Korollar 4 dennoch die Konvergenz von der Ordnung $\sqrt{\epsilon}$, es ergibt sich aber eine sogar noch bessere Konvergenzordnung 1 statt 0.5. Insbesondere wird der Ansatz (3.50),(3.51) in Frage gestellt, da bei einem Fehler der Ordnung $\mathcal{O}(\epsilon)$ zumindest die Terme y_1 und λ_1 Null sein müssen. Des Weiteren scheint allgemein die Konvergenz bezüglich ϵ unabhängig von κ zu sein, denn wenn man obiges numerisches Experiment für verschiedene Werte von κ zwischen Null und Eins durchführt, ergibt sich stets das gleiche Bild:

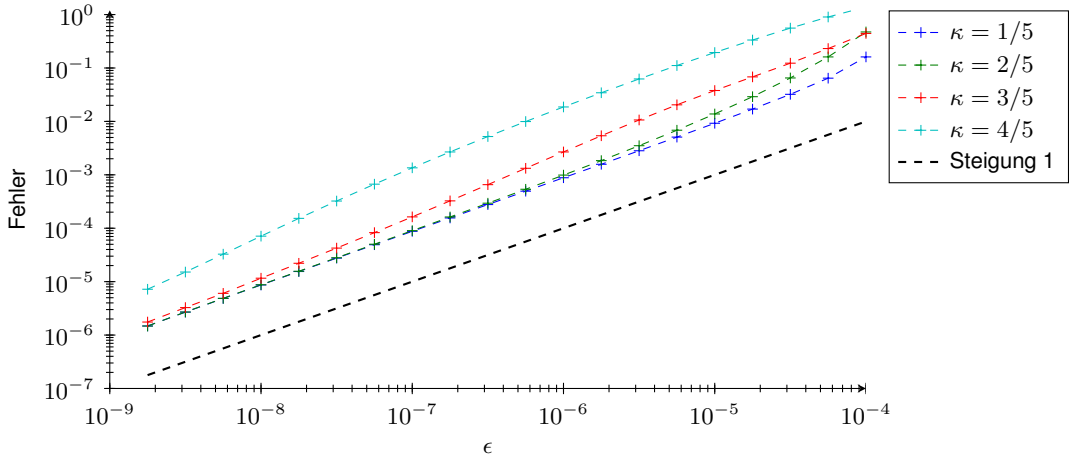

Abb. 4.16: Konvergenz bzgl- ϵ bei verschiedenen κ

Die Ausgleichsgeraden (der Übersichtlichkeit halber nicht eingezeichnet) haben bei allen κ, bis auf kleine Abweichungen, die Steigung 1.

Die gezeigten Plots für Principal Damping wurden alle unter Verwendung der nicht nach Kapitel 3.3.2

transformierten Differentialgleichung erstellt. Strenggenommen sind die Aussagen nach Theorem 7 nur für das transformierte Problem anwendbar, da die Existenz- und Fehleraussagen der numerischen Lösung durch das verwendete Runge-Kutta-Verfahren nur für diese Form getroffen wurden. Interessant ist nun, ob es bei numerischer Rechnung einen relevanten Unterschied zwischen der Lösung von Principal Damping und der transformierten Form von Principal Damping gibt. Hierfür ein Test über eine Sekunde bei $\epsilon = 10^{-6}$ und konstanter Schrittweite. Berechnet wird der Unterschied zwischen beiden Lösungen. Es ergibt sich, zu sehen in Abbildung 4.17, eine Abweichung deutlich geringer als der Fehler der Regularisierung selbst. Dieser liegt nach Abbildungen 4.13 und 4.14 deutlich unterhalb des Fehlers der Regularisierung, diese liegt für $\epsilon = 10^{-6}$ bei etwa 10^{-4}. Auch weitere Tests mit anderen Werten von ϵ zeigen das gleiche Bild. Damit sind die aus den Plots folgenden Aussagen über Konvergenz und Fehlerverhalten übertragbar.

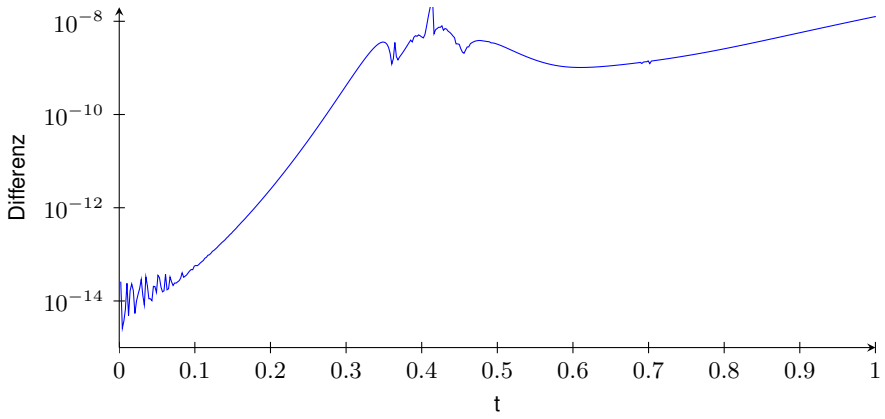

Abb. 4.17: Norm der Differenz der Lösungen von Principal Damping in transformierter und untransformierter Form

Schließlich zum Phänomen des Boundary-Layers. In den bisherigen Versuchen wurde für die λ-Komponente stets ein mit den Zwangsbedingungen konformer Anfangswert benutzt, d.h. der Anfangswert wurde aus den Anfangswerten für y berechnet. Insbesondere nach Theorem 6 mit $N = 0$, also im Falle *Strong Damping*, ist zu erwarten, dass bei gestörten Anfangswerten für λ die Lösung dennoch gegen die ungestörte Lösung konvergiert, indem die Störung exponentiell abnimmt. Im Falle *Principal Damping* gilt dies nach Korollar 4. Zum Test wird in beiden Fällen die Lösung der Differential-Algebraischen-Gleichung als Referenz verwendet. Als gestörte Anfangswerte für λ wird der Vektor $(1, 1)$ verwendet und $\epsilon = 10^{-8}$ gesetzt. Die Ergebnisse werden bei konstanter Schrittweite 10^{-6} verglichen.

Im Falle *Strong Damping*, siehe Abbildung 4.18, fällt die Störung sehr schnell. Trotz sehr kleiner Zeitschritte ist der Fehler bereits nach vier Schritten, damit nach $4 \cdot 10^{-6}$ Sekunden, weitestgehend ausgedämpft und auf einem konstantem Niveau von etwa 10^{-8}. Dies ist noch deutlich unterhalb des resultierenden Fehlers von etwa 10^{-5} nach Abbildung 4.13. Der Fehler der ungestörten Lösung ist zwar noch im Bereich der Rechentoleranz, dies wird sich allerdings durch Verfahrensfehler noch angleichen. Insgesamt resultiert ein gestörter Anfangswert also in einer schnellen Dämpfung auf Niveau des Verfahrensfehlers. Damit ist dieses Verfahren relativ robust bezüglich Störung in der λ-Komponente und man müsste die Anfangswerte dafür nicht einmal gesondert bestimmen.
Der Fall *Principal Damping* im Vergleich dazu, zu sehen bei Abbildung 4.19, zeigt eine deutlich höhere Empfindlichkeit bezüglich der Störung. Es benötigt eine deutlich längere *Einschwingzeit*, bis der Fehler auf konstantem Niveau des Verfahrensfehlers und vergleichbar zum Fehler mit konsistenten Anfangswerten liegt (etwa $1.2 \cdot 10^{-3}$). Ein weiterer Test mit größerem ϵ verifiziert dies: $\epsilon = 10^{-6}$ führt bei *Strong Damping* (Abb. 4.20) immer noch zu einem sehr schnellen Abfall der Störung. Die

Einschwingzeit ist mit etwa $4 \cdot 10^{-6}$ im Bereich von $\epsilon = 10^{-8}$. Dies legt den Verdacht nahe, dass die genaue Dauer der Einschwingzeit beim numerischen Experiment noch stark vom Verfahrensfehler abhängt. Im Vergleich dazu ergibt der Fall *Principal Damping* (Abb. 4.21) nun eine Einschwingzeit von etwa $8 \cdot 10^{-3}$. *Principal Damping* ist also, zumindest beim vorliegenden Versuchsaufbau, erheblich anfälliger für Störungen als *Strong Damping*, bzw. dämpft Störungen langsamer. Das Phänomen des Boundary-Layers existiert aber unabhängig davon bei beiden Fällen, denn Störungen der λ-Komponente klingen sowohl bei Strong als auch bei Principal Damping nach kurzer Einschwingzeit ab. Selbst bei $\epsilon = 10^{-6}$ im Falle *Principal Damping* liegt diese immerhin bei nur etwa 10^{-2} Sekunden.

Weitere interessante Eigenschaften der numerischen Lösung der regularisierten Differential-Algebraischen-Gleichung sind der Verlauf der Verletzung der Zwangsbedingung, sowie Rechenzeit und Schrittweitenverlauf der Schrittweitensteuerung des RadauIIA5-Codes.

Zunächst zur Verletzung der Zwangsbedingung, zu sehen in den Abbildungen 4.22 und 4.23. Hierbei wurde bei $\epsilon = 10^{-8}$ die Norm der Verletzung der Zwangsbedingungen über die Zeit aufgetragen. Es zeichnet sich ein periodischer Ablauf ohne wesentliche Verstärkung der Verletzung über die Zeit ab. Aufgrund des Modells (Verletzung der Zwangsbedingung modelliert mit Feder und Dämpfung) war zu erwarten, dass es periodische Verletzungen der Zwangsbedingungen gibt, jedoch keinen numerischen Drift wie bei einfacher Indexreduktion etwa bei Abbildung 4.5.

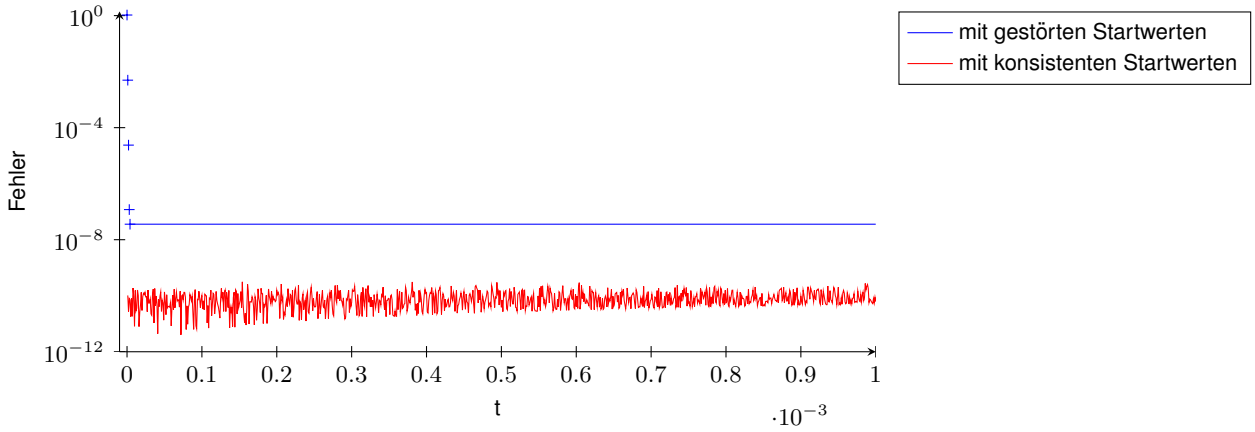

Abb. 4.18: Boundary-Layer bei Strong Damping, die ersten Zeitschritte als $+$, $\epsilon = 10^{-8}$

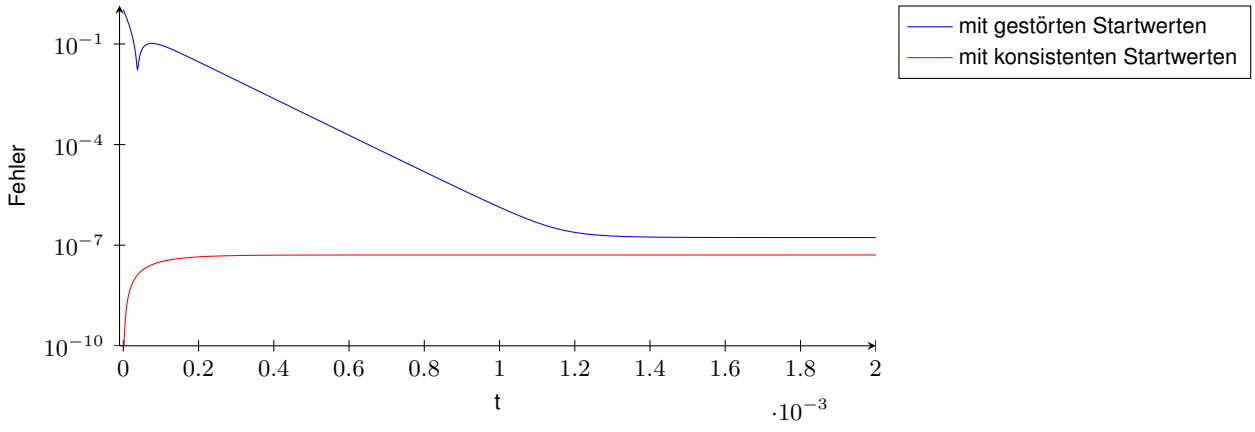

Abb. 4.19: Boundary-Layer bei Principal Damping, $\epsilon = 10^{-8}$

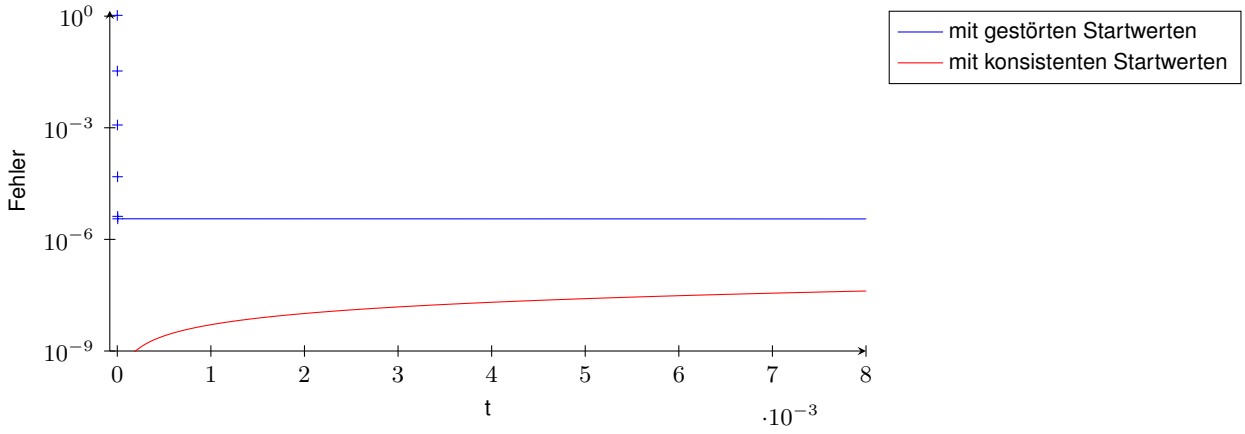

Abb. 4.20: Boundary-Layer bei Strong Damping, die ersten Zeitschritte als $+$, $\epsilon = 10^{-6}$

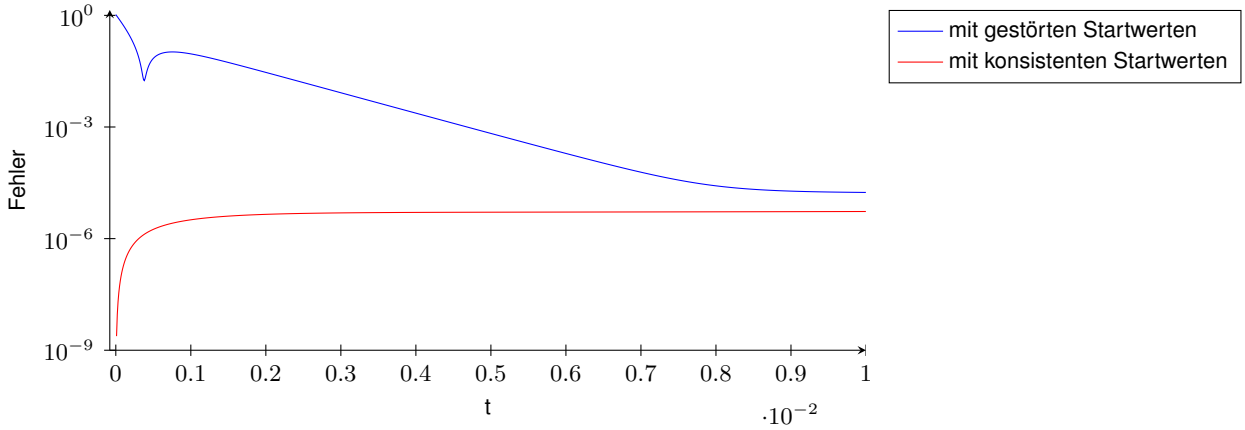

Abb. 4.21: Boundary-Layer bei Principal Damping, $\epsilon = 10^{-6}$

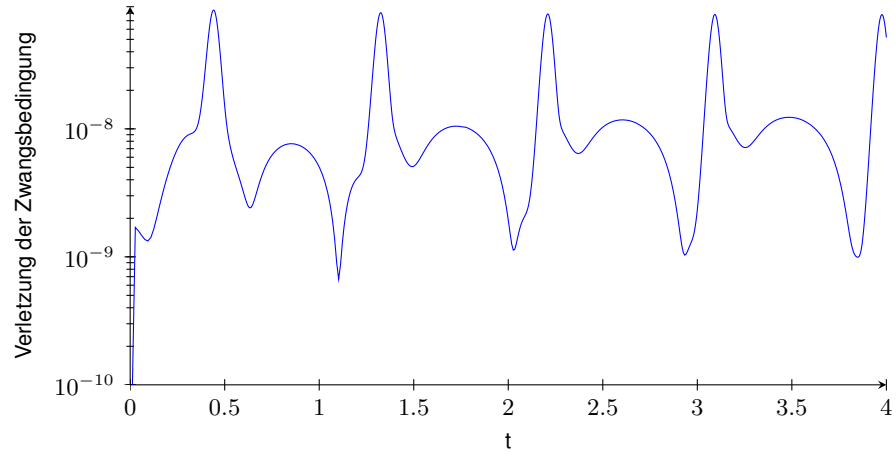

Abb. 4.22: Verletzung der Zwangsbedingung bei Strong Damping, $\epsilon = 10^{-8}$

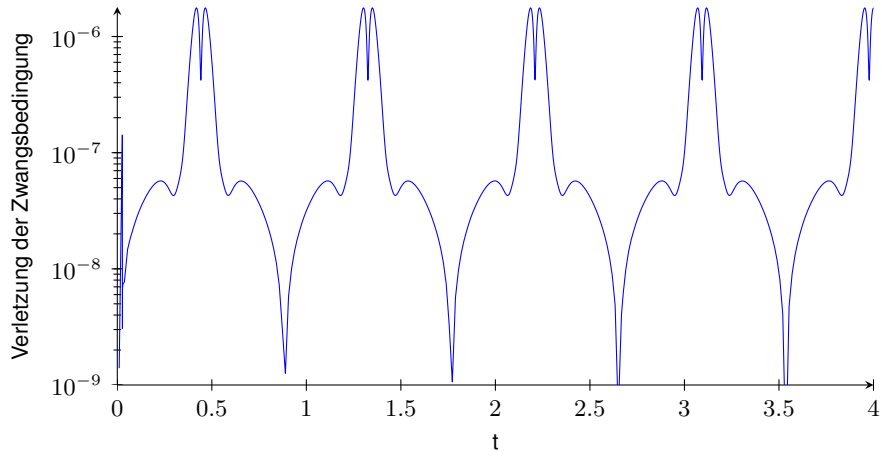

Abb. 4.23: Verletzung der Zwangsbedingung bei Principal Damping, $\epsilon = 10^{-8}$

Nun zum Schrittweitenverlauf, ebenfalls bei $\epsilon = 10^{-8}$, zu sehen in Abbildungen 4.24 und 4.25. Sie zeigen gewisse Abweichungen zur Referenzlösung (Abbildung 4.6), insbesondere gibt es an einzelnen Stellen stärkere Schwankungen. Die allgemeinen Verläufe sind jedoch in etwa vergleichbar.

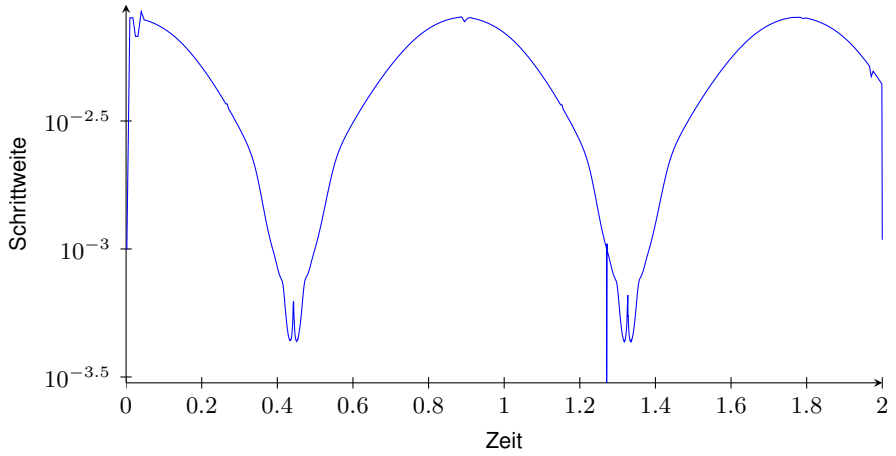

Abb. 4.24: Schrittweitenverlauf bei Strong Damping, $\epsilon = 10^{-8}$

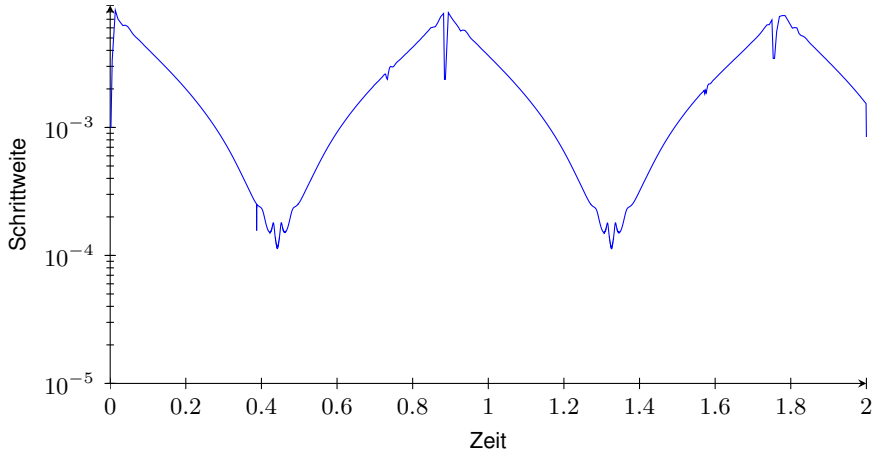

Abb. 4.25: Schrittweitenverlauf bei Principal Damping, $\epsilon = 10^{-8}$

Zum Schluss noch zur Rechenzeit* in Abhängigkeit von ϵ bei Principal und Strong Damping, sowie der nach Kapitel 3.3.2 transformierten Version von Principal Damping. Dies ist zu sehen in Abbildung 4.26. Hierzu wurde absolute und relative Toleranz auf $2 \cdot 10^{-8}$ gesetzt und für verschiedene ϵ die Rechendauer gemessen. Als Referenzwert bei gleicher Toleranz wurde die Differential-Algebraische-Gleichung direkt gelöst, dies dauerte 0.6852 Sekunden. In allen Fällen wurde der beschriebene RadauIIA5 Code mit Schrittweitensteuerung genutzt.
Das Lösen der regularisierten Differential-Algebraischen-Gleichung mit dem verwendeten Code benötigt hier weniger Rechenzeit als das Lösen der Regularisierungen, ist aber von der Größenordnung mit diesen in etwa vergleichbar. Auffallend ist, dass die Abhängigkeit von ϵ bei Principal Damping ohne Transformation erheblich ausgeprägter ist als bei Strong Damping oder Principal Damping nach Transformation. Insbesondere bei Strong Damping bleibt die benötigte Zeit auch bei sehr kleinen ϵ nahezu konstant, beim transformierten Principal Damping ist die Abhängigkeit zumindest deutlich geringer ausgeprägt. Interessant ist ebenfalls, dass die transformierte Version von Principal Damping um etwa den Faktor 3 rechenzeiteffizienter ist als die ursprüngliche Form.

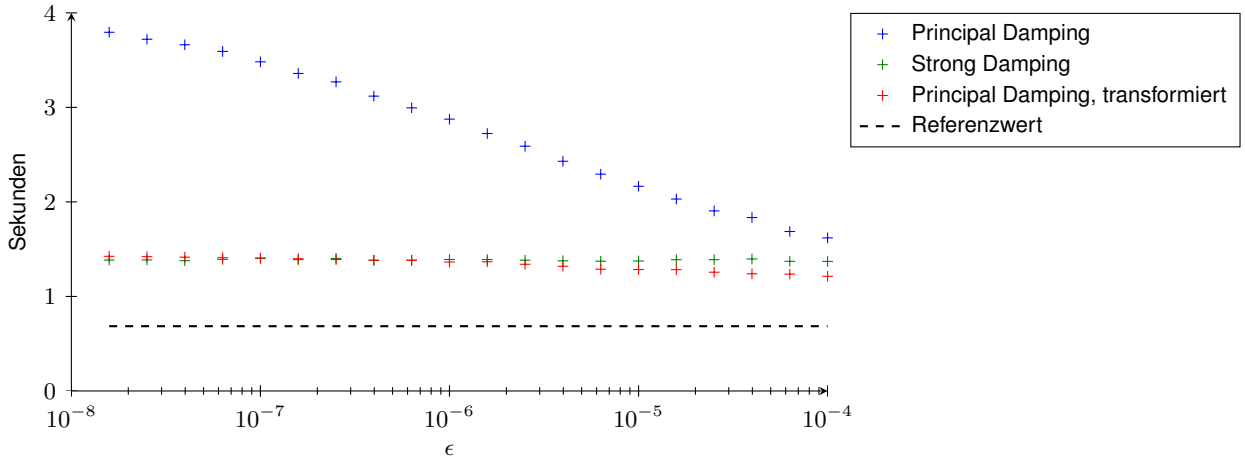

Abb. 4.26: Rechenzeit bei verschiedenen ϵ

*Das Referenzsystem arbeitet mit einem Intel i5-3450 Prozessor und 8 GB DDR3 RAM

Fazit

Das Fazit eines großen Teils der vorangegangenen Untersuchungen zur numerischen Behandlung Differential-Algebraischer-Gleichungen könnte man stark vereinfacht damit zusammenfassen, dass Runge-Kutta-Verfahren sehr gut zur numerischen Lösung von Differential-Algebraischen-Gleichungen und abgewandelten Systemen geeignet sind.

Zum einen lässt sich als sehr erfreuliches Ergebnis zeigen, dass es numerische Lösungsverfahren hoher Konvergenzordnung zur direkten Lösung Differential-Algebraischer-Gleichungen vom Index 2 gibt. Implizite Runge-Kutta-Methoden, die zu Kollokationsverfahren äquivalent sind, behalten ihre volle Konvergenzordnung zumindest bei den Variablen von Index 0 bei. Etwa Methoden vom Typ RadauIIA haben damit die Ordnung $2s - 1$ bezüglich der Index-0 Variablen und immerhin noch Ordnung s bei den Index-2 Variablen. Damit gibt es bei diesen, zumindest was die im Beispiel wichtige Index-0 Variable betrifft, letztlich keinen großen Unterschied zur Lösung einer gewöhnlichen, expliziten Differentialgleichung, von Details wie der in Kapitel 2.4 beschriebenen und gelösten Probleme bei der Implementierung abgesehen. Insbesondere die bereits existierende Implementierung des RadauIIA5-Codes in Fortran oder Matlab bietet damit ein sehr gutes Werkzeug zur numerischen Lösung Differential-Algebraischer-Gleichungen.

Zum anderen erbrachte die Untersuchung der Regularisierungen nach [Dep13], dass diese wie erhofft Lösungen in einer ϵ-Umgebung (respektive $\sqrt{\epsilon}$) um die Lösung der Differential-Algebraischen-Gleichung besitzen. Diese sind ebenfalls, nach Theorem 7, sehr gut durch gewisse Runge-Kutta-Verfahren zu lösen. Damit ist die Grundlage geschaffen, diese Regularisierungen als Bestandteil eines ganzheitlicheren Modells zur Beschreibung von Roll- und Haftreibungsvorgängen mit effizienter Rechnung zu nutzen. Dies war immerhin die eigentliche Motivation zur Formulierung dieser Regularisierungen.

Die numerischen Experimente bestätigen soweit diese Aussagen, die verwendeten RadauIIA-Verfahren besitzen bezüglich aller Variablen die volle erwartete Konvergenzordnung, teilweise für die Index-2 Komponente λ sogar noch mehr. Dies dürfte allerdings dem vergleichsweise simplen Modellproblem geschuldet sein.

Die Konvergenzresultate der als Principal Damping regularisierten Differential-Algebraischen-Gleichung erweisen sich als besser als erwartet. War ursprünglich durch Theorem 4 eine Konvergenz in der Größenordnung $\sqrt{\epsilon}$ vorhergesagt, zeigen die Ergebnisse sogar die Ordnung ϵ. Dies spricht gegen die angenommene Wurzel-Reihenentwicklung der Lösung. Weitere Versuche zeigen, dass dies sogar unabhängig vom verwendeten Parameter κ der Fall zu sein scheint. Allerdings kann auch dies wieder eine besondere Eigenschaft des Modellproblems sein. Dies wäre Motivation weiterführender Untersuchungen. Dass das verwendete numerische Verfahren dies überhaupt auflösen kann, obwohl nach Theorem 7 zunächst nur ein Verfahrensfehler der Ordnung $\sqrt{\epsilon}$ zu erwarten ist, lässt sich aus der Literatur erklären: In den Bemerkungen zu Theorem 7 in [Hai10] wird erwähnt, dass RadauIIA-Verfahren einen um eine ϵ-Potenz geringeren Fehler besitzen, sich also hier die Fehlerordnung ϵ ergibt.

Insgesamt ist die Regularisierungen nach [Dep13] zur Modellierung mechanischer Systeme in einem

größeren Modell nutzbar und mit dem verwendeten RadauIIA5-Code bietet sich ein Verfahren, diese effizient und zuverlässig zu lösen. Im Falle Principal Damping gilt dies insbesondere nach Transformation nach Kapitel 3.3.2. Durch diese wird das Problem in eine Form überführt, bei der Konvergenz und Existenz einer numerischen Lösung gewährleistet werden kann und sich nach den numerischen Experimenten sogar eine Rechenzeitersparnis ergibt. Damit kann das Augenmerk auf die weiteren Bestandteile dieses größeren, ganzheitlichen Modells gerichtet werden. Hierbei ist insbesondere der Zustand der Gleitreibung und vor allem der Übergang zwischen Haft- und Gleitreibung von Interesse.

Literaturverzeichnis

[AE06] Herbert Amann and Joachim Escher. Analysis I, 2006.

[AE08] Herbert Amann and Joachim Escher. Analysis II, 2008.

[Ber08] Prof. Dr. Ing. A. Bertram. Vorlesungsmanuskript zur Festigkeitslehre I+II, 2007-2008. Vorlesungsbegleitendes Skriptum zu den Vorlesungen Festigkeitslehre von Prof. Dr. A. Bertram, Institut für Mechanik, Universität Magdeburg.

[Dep13] Fidlin Deppler. Regularization of Nonholonomic Constraints. Vorläufiges Paper von J. Deppler and A. Fidlin, Karlsruher Institut für Technologie, Institut für Technische Mechanik, Abteilung Dynamik / Mechatronik, Karlsruhe, Germany, 2013.

[Fli09] Torsten Fliessbach. Mechanik: Lehrbuch zur Theoretischen Physik I, 2009.

[HA09] Weimin Han and Kendall E. Atkinson. Theoretical Numerical Analysis : A Functional Analysis Framework, 2009.

[Hai] Ernst Hairer. Fortran and Matlab Codes. http://www.unige.ch/ hairer/software.html.

[Hai10] Ernst Hairer. Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems, 2010.

[HLR89] Ernst Hairer, Christian Lubich, and Michel Roche. *The numerical solution of differential algebraic systems by Runge-Kutta methods*. Lecture notes in mathematics ; 1409. Springer, Berlin, 1989.

[HNWXX] Ernst Hairer, Syvert P. Norsett, and Gerhard Wanner. Solving ordinary differential equations, 19XX. Bd. 2 nur verf. von E. Hairer und G. Wanner.

[Hoc12] Prof. Dr. Marlis Hochbruck. Skriptum zur Vorlesung Numerik 1-4. Vorlesungsbegleitendes Skriptum zu den Vorlesungen Numerik 1 bis 4 von Prof. Dr. Marlis Hochbruck, Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie, 2011-2012.

[Hop68] Hoppenstaedt. Asymptotic stability in singular perturbation problems. *Journal of Differential Equations*, 1968.

[Lö85] Per Lötstedt. On the relation between singular perturbation problems and differential-algebraic equations, 1985.

[Pla10] Robert Plato. Numerische Mathematik kompakt : Grundlagenwissen für Studium und Praxis, 2010.

[Tes12] Gerald Teschl. *Ordinary differential equations and dynamical systems*. Graduate studies in mathematics ; 140. American Mathematical Society, Providence, RI, 2012. Includes bibliographical references and index.

Danksagungen

An dieser Stelle möchte ich all jenen danken, ohne die diese vorliegende Diplomarbeit in dieser Form nicht möglich gewesen wäre.
Ich möchte mich bei meiner Betreuerin, Frau Prof. Dr. Marlis Hochbruck, für ihre aufgewendete Zeit und die grundlegenden Anstöße bedanken. Weiterhin bei Herrn Dipl.-Ing. Jens Deppler und Herrn Prof. Dr.-Ing. Alexander Fidlin vom Institut für Technische Mechanik des KIT für ihre Unterstützung beim Beweis im Falle Principal Damping, bei meinen Kommilitonen für ihr sorgfältiges Korrekturlesen und schließlich bei meinen Eltern, die mich während der ganzen Zeit unterstützt haben.