Integration stark gedämpfter mechanischer Systeme mit Runge-Kutta-Verfahren

Dissertation

zur Erlangung des Grades eines Doktors der Naturwissenschaften
der Fakultät für Mathematik und Physik
der Eberhard-Karls-Universität Tübingen

vorgelegt von
Thomas Stumpp
aus Heudorf

Sommersemester 2004

Tag der mündlichen Prüfung: 7. Juni 2004

Dekan: Prof. Dr. Herbert Müther
Erster Berichterstatter: Prof. Dr. Christian Lubich
Zweiter Berichterstatter: Prof. Dr. Bernd Simeon (TU München)

An dieser Stelle möchte ich mich bei Prof. Dr. Christian Lubich für die exzellente Betreuung herzlich bedanken. Stets war er bereit, neue Ideen, offene Fragen und Probleme zu diskutieren. Thorsten Hans danke ich für seine hilfreichen Beiträge zum letzten Kapitel. Meinen Kolleginnen und Kollegen, insbesondere Tobias Jahnke, gebührt Dank für die Anregungen und Unterstützung beim Entstehen dieser Dissertation. Zuletzt gilt ein besonderes Dankeschön meinen Eltern, die mir das Studium der Mathematik ermöglicht haben.

Inhaltsverzeichnis

Einleitung

Mehrkörpersysteme sind in Forschung und industrieller Entwicklung von gleichermaßen hoher Bedeutung, da sie in so verschiedenen Gebieten wie Astrophysik, Molekulardynamik, Fahrzeugtechnik, Robotik oder Biomechanik auftreten. Ein Mehrkörpersystem beschreibt ein oder mehrere Objekte als Verbund von starren oder elastischen Körpern, zwischen denen bestimmte Kräfte wirken. Bei mechanischen Systemen sind die einzelnen Körper durch masselose Verbindungen wie Gelenke, Federn oder Dämpfer verknüpft.

Ein zentrales Problem, das bei der Modellbildung von Mehrkörpersystemen auftritt, ist die Beschreibung solcher Verbindungen. Dem wird in der Modellierung meist durch Zwangsbedingungen Rechnung getragen, die durch Einschränkung auf ein bestimmtes Koordinatensystem oder durch nichtlineare Nebenbedingungen eingeführt werden können. In Fahrzeugdynamik, Biomechanik und Robotik werden anstelle dieser Zwangsbedingungen häufig Feder-Dämpfer-Elemente verwendet, die oftmals durch große Dämpfungskonstanten gekennzeichnet sind. Auf der Ebene der mathematischen Formulierung führt dies auf *stark gedämpfte mechanische Systeme*, das heißt auf Bewegungsgleichungen, in denen starke Dämpfungskräfte gegenüber anderen Kräften dominieren.

Da es sich aufgrund der starken Dämpfung um steife Differentialgleichungen handelt, kommen bei der Wahl eines geeigneten Integrators zur numerischen Approximation der Lösung keine explizite Verfahren in Betracht, weil diese Verfahren nur bei winzigen Schrittweiten und entsprechend hohem Rechenaufwand eine akzeptable Genauigkeit erreichen würden. In numerischen Experimenten zeigt sich aber, dass gewisse implizite Methoden, beispielsweise RadauIIA-Verfahren, zur Erreichung einer vorgegebenen Genauigkeit eine Schrittweitenwahl unabhängig von der Größe des Dämpfungsparameters gestatten. Hierfür liegen jedoch keine theoretische Erkenntnisse vor. Deshalb wird in dieser Arbeit untersucht, welche Schwierigkeiten bei der numerischen Behandlung von stark gedämpften mechanischen Systemen auftreten. Anhand der Klasse der Runge-Kutta-Verfahren soll gezeigt werden, welche Bedingungen ein Verfahren erfüllen muss, um effizient für eine numerische Simulation eingesetzt werden zu können.

In Kapitel 1 stellen wir zunächst das allgemeine Umfeld stark gedämpfter mechanischer Systeme vor. Bei der Einführung der Bewegungsgleichung für unser Ausgangssystem, einer singulär gestörten gewöhnlichen Differentialgleichung zweiter Ordnung, sollen zentrale Fragestellungen für die analytische und numerische Behandlung konkretisiert werden. Nach der Bereitstellung zweier Transformationen dieser Differentialgleichung zu Beginn von Kapitel 2 charakterisieren wir glatte Lösungen durch eine asymptotische Entwicklung nach Potenzen des Kehrwerts des Dämpfungsparameters. Hierbei wird deutlich, dass sich die glatten Lösungen im Grenzfall sehr großer Dämpfungskonstanten den Lösungen eines differential-algebraischen Gleichungssystems vom Index 2 annähern. Die analytische Untersuchung des Ausgangssystems wird abgeschlossen durch eine Charakterisierung des qualitativen Verhaltens der Lösungen, das in der Existenz einer attraktiven invarianten Mannigfaltigkeit zum Ausdruck kommt.

Numerische Verfahren zur Zeitintegration unseres Problems werden in Kapitel 3 vorgestellt. Die impliziten Runge-Kutta-Verfahren werden durch bestimmte Voraussetzungen, die sich aus der Problemstruktur ergeben, auf spezielle Klassen von Methoden eingegrenzt. Bei der Fehleranalysis für die Ausgangsdifferentialgleichung sind numerische Lösungen der auftretenden differential-algebraischen Systeme maßgeblich beteiligt. Fehlerschranken für die Runge-Kutta-Approximationen dieser Systeme stellen wir mit Techniken aus [6] bereit.

Kapitel 5 beinhaltet Fehleranalysen der Runge-Kutta-Verfahren für das stark gedämpfte mechanische System. Die nötigen technischen Details, also Existenz und lokale Eindeutigkeit der Runge-Kutta-Lösungen, Einfluss von Rundungsfehlern sowie lokaler Fehler und Fehlerfortpflanzung werden in Kapitel 4 behandelt. Zur Berechnung der globalen Fehler wird eine asymptotische Entwicklung für die numerischen Lösungen benötigt, die zu Beginn von Kapitel 5 hergeleitet wird. Die Behandlung eines Anwendungsbeispiels aus der Biomechanik in Kapitel 6 schließt die Arbeit ab.

Kapitel 1

Stark gedämpfte mechanische Systeme

1.1 Problemfeld

Bei Modellbildungen in der Mehrkörperdynamik bieten sich verschiedene Möglichkeiten, um die in der Einleitung angesprochenen Verbindungsglieder zu modellieren. Werden die Gelenke oder Kopplungen durch Feder-Dämpfer-Elemente beschrieben, so ergibt sich eine gewöhnliche Differentialgleichung zweiter Ordnung. Sie hat die explizite Darstellung

$$M(y)\ddot{y} = f(y, \dot{y}) - \frac{1}{\varepsilon} D(y)\dot{y} - \frac{1}{\delta^2}\nabla U(y). \tag{1.1}$$

Hierbei steht die Zuordnung $t \mapsto y(t)$ für die Bewegung des mechanischen Systems, also die Änderung des Ortes als Funktion der Zeit. M bezeichnet eine symmetrische, positiv definite Massenmatrix. Auf der rechten Seite der Differentialgleichung sind in f Kräfte zusammengefasst, die sich etwa aus konservativen oder schwach gedämpften Kraftanteilen zusammensetzen können. Die dominanten Terme mit sehr kleinen, positiven Konstanten ε und δ enthalten eine symmetrische, positiv semidefinite Dämpfungsmatrix D und den Gradienten $\nabla U = (\partial U/\partial y)^T$ eines Potentials U.

Abhängig vom Größenverhältnis zwischen ε und δ führt (1.1) auf zwei problematische Fälle. Ist $\varepsilon \gg \delta$, so liegt ein *steifes oszillatorisches System* vor [13]. Der Fall $\varepsilon \ll \delta$ liefert ein stark gedämpftes mechanisches System, das in der vorliegenden Arbeit studiert werden soll. In der Praxis tritt oftmals auch der Fall auf, dass beide Parameter in etwa gleich groß sind, also $\varepsilon \approx \delta \ll 1$. Dieses Problem kann durch

Kombination der Ergebnisse aus [13] und den Resultaten dieser Arbeit behandelt werden.

In [13] untersucht Lubich die Integration steifer oszillatorischer mechanischer Systeme mit Runge-Kutta-Verfahren, wobei starke Dämpfungskräfte wie sie in (1.1) auftreten nicht berücksichtigt werden. Ausgangssystem ist eine Differentialgleichung zweiter Ordnung, gegeben durch

$$M(y)\ddot{y} = f(y, \dot{y}) - \frac{1}{\delta^2}\nabla U(y). \tag{1.2}$$

Für diesen Fall zeigt sich der Zusammenhang zwischen dem mechanischen System (1.2) und differential-algebraischen Gleichungen sehr schön an einem einfachen Beispiel. Ein Pendel, aufgebaut aus einem Massepunkt, der mit einer steifen, als masselos angenommenen Feder verbunden ist, lässt sich durch kartesische Koordinaten (y_1, y_2) beschreiben als (siehe [3], S. 134)

$$\begin{aligned} \ddot{y}_1 &= -\frac{1}{\delta^2}\frac{y_1}{\sqrt{y_1^2+y_2^2}}(y_1^2+y_2^2-1), \\ \ddot{y}_2 &= -\frac{1}{\delta^2}\frac{y_2}{\sqrt{y_1^2+y_2^2}}(y_1^2+y_2^2-1) - 1. \end{aligned} \tag{1.3}$$

Hierbei ist $\frac{1}{\delta^2}$ eine Federkonstante mit sehr kleinem, positiven Parameter δ. Weiter wird angenommen, dass die Feder in Ruheposition die euklidische Länge 1 und der materielle Punkt die Masse 1 besitzen. Die Gravitationskonstante wird ebenfalls gleich 1 gesetzt. Für Anfangswerte mit beschränkter Energie zeigt sich, dass die Bewegung des Systems (1.3) nahe an der Bewegung des mathematischen Pendels liegt ([16], S. 11 ff.). Formulieren wir das mathematische Pendel als mechanisches System mit der naheliegenden Zwangsbedingung, dass der euklidische Abstand des Massepunktes zum Aufhängepunkt gleich eins ist, so ergibt sich ein differential-algebraisches System

$$\begin{aligned} \ddot{y}_1 &= -y_1\lambda, \\ \ddot{y}_2 &= -y_2\lambda - 1, \\ 0 &= y_1^2 + y_2^2 - 1. \end{aligned} \tag{1.4}$$

Unter bestimmten Voraussetzungen lässt sich für die numerische Lösung von steifen mechanischen oszillatorischen Systemen durch implizite Runge-Kutta-Verfahren ein ähnliches Verhalten feststellen. Im Grenzübergang $\delta \to 0$ konvergieren Runge-Kutta-Lösungen von (1.2) mit Schrittweiten $h > \delta$ gegen die Runge-Kutta-Lösungen eines zugehörigen differential-algebraischen Systems vom Störungsindex 3. Dieses Verhalten ist auch der Grund für die Schwierigkeiten in der numerischen Behandlung von steifen oszillatorischen mechanischen Systemen. Im obigen Beispiel ist der Grenzfall zu (1.3) durch das Index-3-System (1.4) gegeben.

Ein anderer klassischer Zugang, Mehrkörpersysteme zu modellieren besteht, in der Verwendung mechanischer Systeme mit Zwangsbedingungen (*constrained mechanical systems*). Die Bewegung des mechanischen Systems wird dabei mit einem Satz gewöhnlicher Differentialgleichungen zweiter Ordnung beschrieben. Die Gelenke werden durch Zwangsbedingungen modelliert, die in nichtlinearen Nebenbedingungen realisiert sind. Dieser Zugang führt auf differential-algebraische Gleichungssysteme, die im Fall holonomer Zwangsbedingungen vom Störungsindex 3 sind. Ein Ansatz differential-algebraische Gleichungssysteme - auch mit Index kleiner als 3 - approximativ zu lösen beruht auf Regularisierungsmethoden (siehe [12], [2]). Hierbei werden mechanische Systeme mit Zwangsbedingungen in Systeme wie (1.1) umgeformt. Dies geschieht in der Absicht, dass die resultierenden steifen Differentialgleichungen durch geeignete implizite Zeitintegrationsverfahren behandelt werden können.

Zu Systemen wie (1.1) gibt es in der Mechanik eine ganze Reihe von Anwendungsbeispielen. Wir wollen hier allerdings Beispiele betrachten, bei denen die starken Dämpfungskräfte dominieren, der Term $\frac{1}{\delta^2}\nabla U(y)$ also nicht auftritt. Ein oft zitiertes Beispiel, in dem viele Feder-Dämpfer-Elemente zu simulieren sind, ist ein Lastwagenmodell, das sehr gut dokumentiert und auch in MATLAB [15] implementiert ist (etwa [3], [20], siehe Abbildung 1.2). In dieser Arbeit soll aber abschließend zur Illustration und Bedeutung von stark gedämpften mechanischen Systemen in Anwendungen ein Menschmodell unter biomechanischen Gesichtspunkten betrachtet werden, wie es in [10] vorgestellt wird (siehe Abbildung 1.3). Im vorliegenden Beispiel wird das biomechanische Verhalten des Menschmodells nach einem Unfall simuliert. Um die Herkunft von Bewegungsgleichungen wie (1.1) zu motivieren, sollen diese in Kapitel 6 anhand eines einfachen Falls hergeleitet werden.

1.2 Bewegungsgleichungen

Ein stark gedämpftes mechanisches System lässt sich durch eine Differentialgleichung zweiter Ordnung

$$M(y)\ddot{y} = f(y,\dot{y}) - \frac{1}{\varepsilon}D(y)\dot{y} \tag{1.5}$$

beschreiben. Hierbei ist ε eine sehr kleine, positive Konstante, wir setzen also generell

$$0 < \varepsilon \ll 1$$

voraus. Die physikalischen Effekte der Dämpfung sind in der Dämpfungsmatrix D wiedergegeben. Durch M sei eine Massenmatrix bezeichnet. Die Bewegung des stark gedämpften mechanischen Systems wird durch die Lösung $y(t) \in \mathbb{R}^d$ dieser Differentialgleichung für Zeiten t aus einem ε-unabhängigem Intervall $[0,T]$ beschrieben. Grundsätzlich nehmen wir an, dass die Funktionen $M : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, $f : \mathbb{R}^{2d} \to \mathbb{R}^d$

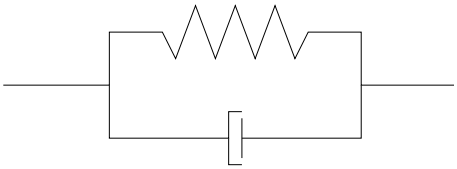

Abbildung 1.1: Symbolisierung von Feder-Dämpfer-Elementen in der Mechanik.

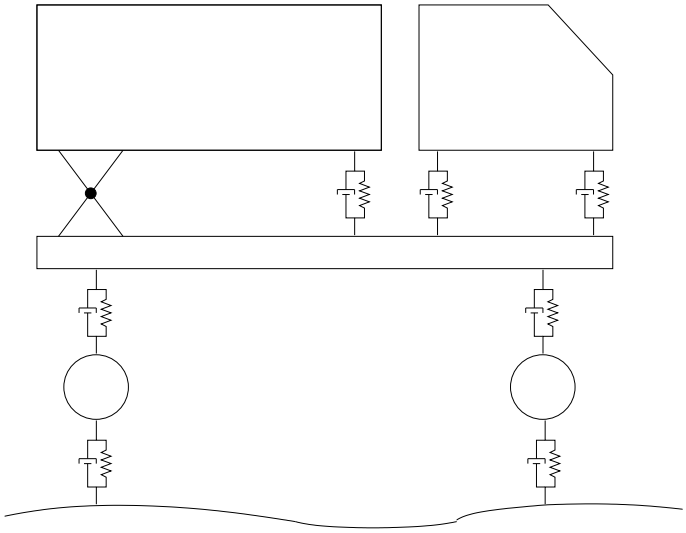

Abbildung 1.2: Lastwagenmodell aus der Fahrzeugmechanik, nach [3].

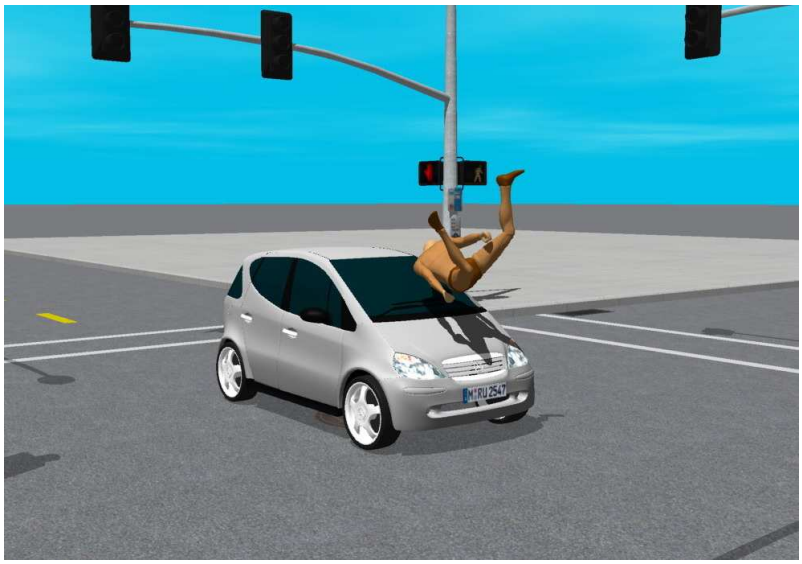

Abbildung 1.3: Menschmodell aus der Biomechanik bei der Simulation eines Unfalls, aus [10].

und $D : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ ausreichend viele beschränkte Ableitungen besitzen. Weitere Voraussetzungen, die im Laufe dieser Arbeit benötigt werden, sind:

Für alle $y \in \mathbb{R}^d$ sei $M(y)$ symmetrisch und positiv definit. (1.6a)

Für alle $y \in \mathbb{R}^d$ sei $D(y)$ symmetrisch, positiv semidefinit und von kostantem Rang m. (1.6b)

Die Matrix D besitzt einen nichttrivialen Kern, dessen Dimension generell mit l bezeichnet sei, es gilt also

$$m = d - l.$$

Wegen Voraussetzung (1.6b) verschwindet $D(y)\dot{y}$ auf einer $(d+l)$-dimensionalen Untermannigfaltigkeit

$$\mathcal{M}^0 = \{(y, v) \in \mathbb{R}^{2d} : D(y)v = 0\} \subseteq \mathbb{R}^{2d}. \tag{1.7}$$

Vom Standpunkt der singulären Störungstheorie ist (1.5) ein singulär singulär gestörtes Problem [18].

Bemerkung. Die Bewegung des mechanischen Systems wird durch eine zeitabhängige Funktion $t \mapsto y(t)$ beschrieben. Die Funktionen in (1.5) sind also streng genommen als zeitabhängige Funktionen zu formulieren, zum Beispiel $t \mapsto M(y(t))$ im Fall der Massenmatrix,. Um die Notation nicht zu überladen, wollen wir die Schreibweise aus (1.5) beibehalten und notieren folglich kurz $M(y(t)) = M(y)$, $f(t, y(t), \dot{y}(t)) = f(y, \dot{y})$ und $D(y(t)) = D(y)$.

Fragestellungen im kontinuierlichen Fall

Erinnern wir uns an die Beobachtungen am Beispiel des Federpendels (1.3), so stellt sich auch für die Bewegung stark gedämpfter mechanischer Systeme die Frage, inwieweit Lösungen von (1.5) mit Lösungen differential-algebraischer Systeme übereinstimmen. Diese Sichtweise hängt direkt mit der Problemstellung zusammen, glatte Lösungen von (1.5) zu approximieren. Ein Zugang liegt hierbei in der Konstruktion von ε-Entwicklungen.

Ein anderes Problem ist die mathematische Beschreibung des qualitativen Verhaltens der mechanischen Bewegung. Ein einfaches Beispiel um dieses Verhalten zu verdeutlichen, ist ein Pendel, bei dem ein Massepunkt über eine starre Verbindung mit einem Feder-Dämpfer-Gelenk verbunden ist, siehe Abbildung 1.4. Simulieren wir die Bewegung dieses Pendels und weisen dem Gelenk eine große Dämpfungskonstante zu, so nähert sich der Abstand zwischen den beiden Kopplungspunkten P_2

und P_3 des Gelenks schnell einem festen Wert. Die Relativgeschwindigkeit v zwischen P_2 und P_3 wird betragsmäßig schnell klein, wie in Abbildung 1.5 zu erkennen ist. Dies ist ein typisches Verhalten der Bewegung stark gedämpfter mechanischer Systeme. Ein Weg, dieses Verhaltensmuster mathematisch zu beschreiben, ist durch das Konzept attraktiver invarianter Mannigfaltigkeiten gegeben (siehe [17], [14]).

Ziele und Fragen der numerischen Behandlung

Eines der wichtigsten Ziele dieser Arbeit ist es, Runge-Kutta-Verfahren zu klassifizieren, die die Bewegungsgleichungen von stark gedämpften mechanischen Systemen effizient integrieren. Generelle Maßgabe ist dabei, Schrittweiten zu verwenden, die nicht durch den kleinen Dämpfungsparameter ε beschränkt werden. Diese Klassifikation ergibt sich im Rahmen einer Konvergenzanalyse für die Runge-Kutta-Approximationen von (1.5).

Die zu erwartenden analytischen Ergebnisse lassen die Vermutung zu, dass die Fehler der numerischen Lösungen der beteiligten differential-algebraischen Systeme einen entscheidenden Anteil zu den globalen Fehlerschranken beitragen. Daher sind zunächst Fehlerabschätzungen für diese differential-algebraischen Systeme von Interesse.

In Analogie zum kontinuierlichen Fall stellen wir uns auch die Frage, ob eine ε-Entwicklung für numerische Lösungen von (1.5) existiert. Mit dieser ε-Entwicklung könnte sofort ein Konvergenzresultat angegeben werden, das den Fall von Anfangswerten behandelt, die eine gleichmäßig in ε glatte Lösung gestatten. Aufgrund dieser rigorosen Bedingungen an die Startwerte stellt sich dann allerdings die Frage, ob noch ein Zusammenhang mit numerischen Lösungen zu allgemeineren Angangswerten hergestellt werden kann.

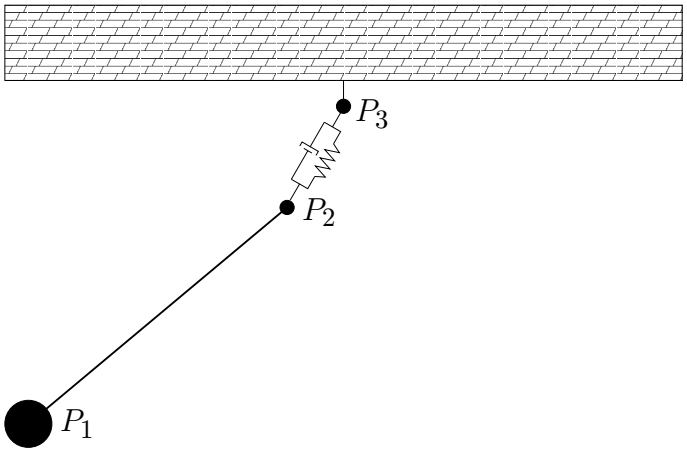

Abbildung 1.4: Pendel mit Kugel P_1 und Feder-Dämpfer-Element, das an zwei Punkte P_3 und P_2 gekoppelt ist.

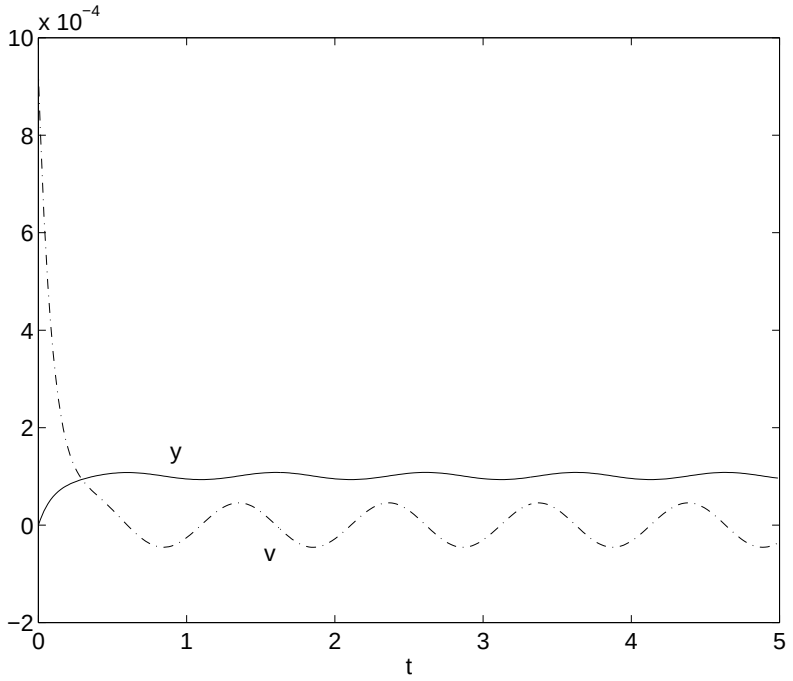

Abbildung 1.5: Qualitatives Verhalten des Abstandes y von P_2 und P_3 und der Relativgeschwindigkeit v von P_2.

Kapitel 2

Analytische Eigenschaften

2.1 Transformationen

Für die späteren Beweisführungen wird es sich als notwendig erweisen, die Differentialgleichung (1.5) in zwei andere Formulierungen zu überführen. Als wesentliche Voraussetzung benötigen wir hierzu die Symmetrie und die positive Definitheit beziehungsweise Semidefinitheit der Matrizen M und D. Beiden Transformationen liegt zu Grunde, dass unter der Voraussetzung (1.6b) für $D(y)$ eine Blockdiagonalisierung

$$Q^T(y)D(y)Q(y) = \begin{pmatrix} 0 & 0 \\ 0 & \widetilde{A}(y) \end{pmatrix} \tag{2.1}$$

existiert. $\widetilde{A}(y) \in \mathbb{R}^{m\times m}$ ist dann eine symmetrische, positiv definite Matrix; die Transformationsmatrix $Q(y) \in \mathbb{R}^{d\times d}$ ist orthogonal.

Lemma 1. *Unter den Voraussetzungen (1.6a) und (1.6b) ist (1.5) äquivalent zu einer gekoppelten Differentialgleichung der Ordnung 1*

$$\begin{aligned} \dot{y} &= S^{-T}(y)z, \\ \widetilde{M}(y)\dot{z} &= \widetilde{f}(y,z) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} z, \end{aligned} \tag{2.2}$$

wobei $\widetilde{M}(y)$ wieder symmetrisch und positiv definit ist. Insbesondere kann die Transformation $S(y)$ so gewählt werden, dass $\widetilde{M}$ und $\widetilde{f}$ beliebig oft differenzierbar und unabhängig von ε sind.

Beweis. Wir betrachten die Blockdiagonalisierung (2.1) und verwenden die gleichen Bezeichnungen. Die Abbildungen $\widetilde{A} : \mathbb{R} \to \mathbb{R}^d \to \mathbb{R}^{m\times m}$ und $Q : \mathbb{R} \to \mathbb{R}^d \to \mathbb{R}^{d\times d}$

können als Funktionen gewählt werden, die beliebig oft differenzierbar und unabhängig von ε sind (Theorem II.5.11 in [11], S. 115). Zu beachten ist hier, dass die Diagonalisierung von $\widetilde{A}(y)$ mit einer stetigen Transformationsmatrix nicht möglich ist, falls sich die Eigenwerte für ein $t \in [0,T]$ kreuzen (siehe Rellichs Beispiel in [11], S. 111). Zu $\widetilde{A}$ lässt sich jedoch eine Cholesky-Zerlegung $\widetilde{A}(y) = (LL^T)(y)$ finden, wobei $L(y) : \mathbb{R}^d \to R^{m\times m}$ als beliebig oft differenzierbare und von ε unabhängige Funktion gewählt werden kann. Setzen wir

$$S(y) = (QC)(y) \qquad \text{mit } C(y) = \begin{pmatrix} I_d & 0 \\ 0 & L(y) \end{pmatrix},$$

so ergibt sich

$$D(y) = S(y) \begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} S(y)^T. \tag{2.3}$$

Durch Einführung einer Variablen $\dot{y} = v$ ist (1.5) äquivalent zu einer Differentialgleichung erster Ordnung

$$\begin{aligned} \dot{y} &= v, \\ M(y)\dot{v} &= f(y,v) - \frac{1}{\varepsilon} D(y)v. \end{aligned} \tag{2.4}$$

Einsetzen der Transformation $v = S^{-T}(y)z$ und der entsprechenden Zeitableitung $\dot{v} = S^{-T}(y)\dot{z} + \mathcal{DS}(y,v)z$ in (2.4) liefert

$$\begin{aligned} \dot{y} &= S^{-T}(y)z, \\ (MS^{-T})(y)\dot{z} &= f(y, S^{-T}(y)z) - M(y)\mathcal{DS}(y, S^{-T}(y)z)z - \frac{1}{\varepsilon}(DS^{-T})(y)z. \end{aligned}$$

Dabei ist $\big(\mathcal{DS}(y,v)\big)\big|_{ij} = \langle \nabla_y \widetilde{s}_{ij}(y), v\rangle$, wobei $\widetilde{s}_{ij}(y)$ die Elemente von $S^{-T}(y)$ bezeichnet. Durch Multiplikation der zweiten Gleichung von links mit $S^{-1}(y)$ und mit den Bezeichnungen

$$\widetilde{M}(y) = (S^{-1}MS^{-T})(y), \quad \widetilde{f}(y,z) = S^{-1}\Big(f(y, S^{-T}(y)z) - M(y)\mathcal{DS}(y, S^{-T}(y)z)z\Big)$$

folgt (2.2). Dass sich die Symmetrie und die positive Definitheit von M auf $\widetilde{M}$ überträgt, folgt durch direktes Nachrechnen. Als Kompositionen beliebig oft differenzierbarer Funktionen sind $\widetilde{M}$ und $\widetilde{f}$ beliebig oft differenzierbar. Somit ist die Behauptung gezeigt. ∎

Mit (2.2) können nicht alle Resultate bewiesen werden, da die Massenmatrix in modifizierter Form immer noch auf der linken Seite der zweiten Differentialgleichung steht. Deshalb wird noch die folgende Umformulierung bereitgestellt und bewiesen.

Lemma 2. *Unter den Voraussetzungen (1.6a) und (1.6b) ist (1.5) äquivalent zu einer gekoppelten Differentialgleichung der Ordnung 1*

$$\begin{aligned} \dot{u} &= F(u,x), \\ \dot{x} &= -\frac{1}{\varepsilon}A(u)x + \varphi(u,x) \end{aligned} \tag{2.5}$$

mit einer positiv definiten Matrix $A(u)$. Die dabei auftretende Transformation lässt sich so konstruieren, dass $F : \mathbb{R}^{2d-m} \times \mathbb{R}^m \to \mathbb{R}^{2d-m}$, $A : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ und auch $\varphi : \mathbb{R}^{2d-m} \times \mathbb{R}^m \to \mathbb{R}^m$ als Funktionen gewählt werden können, die beliebig oft differenzierbar und unabhängig von ε sind.

Beweis. Aufgrund der Voraussetzung (1.6a) lässt sich die Massenmatrix in eine Zerlegung $M(y) = (M^{1/2}M^{1/2})(y)$ mit symmetrischen, positiv definiten Matrizen $M^{1/2}(y)$ aufspalten. Diese können als beliebig oft differenzierbare und von ε unabhängige Funktionen gewählt werden. Aus (2.4) erhalten wir nunmehr

$$\begin{aligned} \dot{y} &= v, \\ M^{1/2}(y)\dot{v} &= M^{-1/2}(y)f(y,v) - \frac{1}{\varepsilon}(\widehat{D}M^{1/2})(y)v, \end{aligned} \tag{2.6}$$

wobei $\widehat{D}(y) = (M^{-1/2}DM^{-1/2})(y)$ eine symmetrische, positiv semidefinite Matrix ist. Zu $\widehat{D}$ existiert wie in (2.1) eine Blockdiagonalisierung mit einer orthogonalen Matrix $\widehat{Q}(y)$ und einer positiv definiten Matrix $A(y)$. Diese Funktionen können ebenfalls als beliebig oft differenzierbare und von ε unabhängige Funktionen gewählt werden. Multiplizieren wir $\widehat{Q}^T$ von links an die zweite Gleichung von (2.6), so ist diese unter der Transformation $v = T(y)w$ mit $T(y) = (M^{-1/2}\widehat{Q})(y)$ und der entsprechenden Zeitableitung $\dot{v} = T(y)\dot{w} + \mathcal{DT}(y,v)w$ äquivalent zu

$$\begin{aligned} \dot{y} &= T(y)w, \\ T^{-1}(y)\big(T(y)\dot{w} + \mathcal{DT}(y,T(y)w)w\big) &= (\widehat{Q}^T M^{-1/2})(y)f(y,T(y)w) \\ &\quad -\frac{1}{\varepsilon}(\widehat{Q}^T\widehat{D}M^{1/2}T)(y)w. \end{aligned}$$

Dabei ist $\big(\mathcal{DT}(y,v)\big)\big|_{ij} = \langle \nabla_y t_{ij}(y), v\rangle$, wobei $t_{ij}(y)$ die Elemente von $T(y)$ bezeichnet. Diese Differentialgleichungen sind mit

$$\widehat{f}(y,w) = (\widehat{Q}^T M^{-1/2})(y)f(y,T(y)w) - T^{-1}(y)\mathcal{DT}(y,T(y)w)w$$

und der Blockdiagonalisierung von $\widehat{D}$ gleichwertig zu

$$\begin{aligned} \dot{y} &= T(y)w, \\ \dot{w} &= \widehat{f}(y,w) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & A(y) \end{pmatrix} w. \end{aligned} \tag{2.7}$$

Passend zum Format von A setzen wir

$$\widehat{f} = (\widehat{f}_1, \widehat{f}_2)^T, \quad w = (p, q)^T, \quad T(y) = (T_1(y), T_2(y))$$

und erhalten

$$\begin{aligned} \dot{y} &= T_1(y)p + T_2(y)q, \\ \dot{p} &= \widehat{f}_1(y, p, q), \\ \dot{q} &= -\frac{1}{\varepsilon}A(y)q + \widehat{f}_2(y, p, q). \end{aligned}$$

Mit

$$\begin{aligned} &u = (y, p)^T \in \mathbb{R}^{2d-m}, \ x = q \in \mathbb{R}^m, \\ &F(u, x) = \begin{pmatrix} T_1(y)p + T_2(y)q \\ \widehat{f}_1(y, p, q) \end{pmatrix} \text{ und } \varphi(u, x) := \widehat{f}_2(u, x) \end{aligned}$$

folgt (2.5). Die neu definierten Funktionen F und φ sind als Kompositionen von beliebig oft differenzierbaren Funktionen wieder beliebig oft differenzierbar und unabhängig von ε. Somit ist die Behauptung gezeigt. ∎

Bemerkungen. (1) Eine Umformung von (2.2) in eine Differentialgleichung erster Ordnung ohne Massematrix auf der linken Seite ist möglich, denn $\widetilde{M}(y)$ ist invertierbar (insbesondere wieder symmetrisch und positiv definit). Allerdings ergäbe sich bei dieser Vorgehensweise eine voll besetzte Matrix auf der rechten Seite. In späteren Beweisen wird allerdings ein äquivalentes System mit einer Matrix in Blockstruktur benötigt, wie es in (2.5) beziehungsweise (2.7) hergeleitet wurde.

(2) Die äquivalenten Formulierungen (2.2) und (2.5) werden im Folgenden ständig zur Herleitung von Abschätzungen gebraucht. Da numerische Verfahren unter diesen Transformationen nicht invariant sind, wird die Ausgangsgleichung (1.5) immer wieder in Betracht gezogen.

2.2 Asymptotische ε-Entwicklung der glatten Lösung

Im folgenden Theorem werden *glatte* Lösungen von (1.5) studiert, das heißt Lösungen, die genügend oft stetig differenzierbar und deren Ableitungen unabhängig von ε beschränkt sind. Von besonderem Interesse ist dabei, dass sich wie in den Fällen

von *singulär gestörten Problemen* (vergleiche [9], S.388 f.) und *steifen oszillatorischen mechanischen Systemen* (vergleiche [13], Theorem 2.2) die glatten Lösungen in einer ε-Entwicklung beziehungsweise ε^2-Entwicklung mit Koeffizienten $(y^k, \dot{y}^k)$, $k \geq 0$, darstellen lassen.
Im vorliegenden Problem treten bei der Konstruktion einer asymptotischen ε-Entwicklung Schwierigkeiten auf, insbesondere durch den Term $D(y)\dot{y}$, der von Position und Geschwindigkeit des stark gedäpften mechanischen Systems abhängt. Das standardgemäße Vorgehen (siehe etwa Theorem 2.2 in [13]) wäre, die Funktionen in (1.5) nach y^0 beziehungsweise $(y^0, \dot{y}^0)$ zu entwickeln, als existent angenommene ε-Entwicklungen für y und $\dot{y}$ einzusetzen und die Koeffizienten des resultierenden Systems nach Potenzen in ε zu vergleichen. Zur Konstruktion der Koeffizienten $(y^0, \dot{y}^0)$ ergäbe sich so als erstes System eine Ordnung-2-Differentialgleichung und eine algebraische Gleichung

$$\begin{aligned} M(y^0)\ddot{y}^0 &= f(y^0, \dot{y}^0) - D(y^0)\dot{y}^1 - \frac{\partial}{\partial y}(D(y)y^1)\Big|_{y=y^0} \cdot \dot{y}^1, \\ 0 &= D(y^0)\dot{y}^0. \end{aligned}$$

Würde in der Differentialgleichung dieses Systems nur eine Unbekannte y^1 oder $\dot{y}^1$ auftreten, so könnte sie auf kanonische Weise durch die Einführung eines Lagrange-Multiplikators mit der algebraischen Gleichung gekoppelt werden. Im obigen Fall ist dies aber nicht möglich. Aus diesem Grund ist es für das mechanische System (1.5) nötig, von der Umformulierung (2.2) auszugehen.

Bemerkung. In vielen der folgenden Beweise, auch in den weiteren Kapiteln, werden in den Abschätzungen immer wieder verschiedene reelle Konstanten auftreten. Diese sollen nicht durchnummeriert, sondern generell mit C bezeichnet werden.

Theorem 1. *Sei $N \geq 1$ eine beliebige natürliche Zahl und (1.6a) sowie (1.6b) erfüllt. Zu $0 < \varepsilon \leq \varepsilon_0$ existiert eine (d+l)-dimensionale Mannigfaltigkeit $\mathcal{M}^\varepsilon$, konstruiert durch eine Bijektion*

$$\Psi^\varepsilon : \mathcal{M}^0 \to \mathcal{M}^\varepsilon : (y^0, \dot{y}^0) \mapsto (y^\varepsilon, \dot{y}^\varepsilon)$$

mit $\Psi^\varepsilon = id + O(\varepsilon)$, und ein von ε unabhängiges Intervall $[0, T]$, so dass folgendes gilt: Ist $(y(0), \dot{y}(0)) \in \mathcal{M}^\varepsilon$, so ist für die Lösung $(y(t), \dot{y}(t))$ von (1.5) zu diesem Startwert

$$(y(t), \dot{y}(t)) \in \mathcal{M}^\varepsilon + O(\varepsilon^N) \qquad \text{für } 0 \leq t \leq T,$$

und es existieren asymptotische ε-Entwicklungen

$$\begin{aligned} y(t) &= y^0(t) + \varepsilon y^1(t) + \cdots + \varepsilon^N y^N(t) + O(\varepsilon^{N+1}), \\ \dot{y}(t) &= \dot{y}^0(t) + \varepsilon \dot{y}^1(t) + \cdots + \varepsilon^N \dot{y}^N(t) + O(\varepsilon^{N+1}), \end{aligned} \tag{2.8}$$

wobei $(y^0(t), \dot{y}^0(t)) \in \mathcal{M}^0$ *und* $(y^k(t), \dot{y}^k(t))$ *für* $0 \le k \le N$ *Lösungen von differential-algebraischen Gleichungen sind, die auf Seite 25 in (DAE 0),. . . ,(DAE k) formuliert sind. Die ersten* N *Ableitungen von* y *sind unabhängig von* ε *beschränkt.*

Beweis. Die Behauptungen folgen aus drei Beweisschritten. Aus den geschilderten Gründen betrachten wir die Umformulierung aus Lemma 1. Zunächst konstruieren wir die asymptotischen ε-Entwicklungen (2.8). In Beweisteil (b) zeigen wir, dass Lösungen von (1.5) mit Anfangswerten aus $\mathcal{M}^0$ für Zeiten $t \in [0, T]$ bis auf $O(\varepsilon^N)$ in $\mathcal{M}^\varepsilon$ liegen. Damit lässt sich die Beschränktheit der Ableitungen von y zeigen.

(a) Sei also angenommen, dass eine ε-Entwicklung der exakten Lösung $(y(t), z(t))$ von (2.2) bekannt ist, das heißt

$$\begin{aligned} y(t) &= y^0(t) + \varepsilon y^1(t) + \cdots + \varepsilon^N y^N(t) + O(\varepsilon^{N+1}), \\ z(t) &= z^0(t) + \varepsilon z^1(t) + \cdots + \varepsilon^N z^N(t) + O(\varepsilon^{N+1}). \end{aligned}$$

Zunächst werden die Funktionen aus (2.2) im Hinblick auf die obigen ε-Entwicklungen um y^0, z^0 beziehungsweise (y^0, z^0) entwickelt. Wir erhalten

$$\begin{aligned} \widetilde{M}(y)\dot{z} &= \widetilde{M}(y^0)\dot{z} + \frac{\partial}{\partial y}\Big(\widetilde{M}(y)\dot{z}\Big)\Big|_{y=y_0} \cdot [y - y^0] + \dots \\ &= \widetilde{M}(y^0)(\dot{z}^0 + \cdots + \varepsilon^N \dot{z}^N + O(\varepsilon^{N+1})) + \\ &\qquad \frac{\partial}{\partial y}\Big(\widetilde{M}(y)(\dot{z}^0 + \cdots + \varepsilon^N \dot{z}^N + O(\varepsilon^{N+1}))\Big)\Big|_{y=y_0} \cdot \\ &\qquad\qquad [\varepsilon y^1 + \cdots + \varepsilon^N y^N + O(\varepsilon^{N+1})] + \dots, \end{aligned}$$

$$\begin{aligned} S^{-T}(y)z &= S^{-T}(y^0)z + \frac{\partial}{\partial y}\Big(S^{-T}(y)z\Big)\Big|_{y=y_0} \cdot [y - y^0] + \dots \\ &= S^{-T}(y^0)(z^0 + \cdots + \varepsilon^N z^N + O(\varepsilon^{N+1})) + \\ &\qquad \frac{\partial}{\partial y}\Big(S^{-T}(y)(z^0 + \cdots + \varepsilon^N z^N + O(\varepsilon^{N+1}))\Big)\Big|_{y=y_0} \cdot \\ &\qquad\qquad [\varepsilon y^1 + \cdots + \varepsilon^N y^N + O(\varepsilon^{N+1})] + \dots, \end{aligned}$$

$$\begin{aligned} \widetilde{f}(y, z) &= \widetilde{f}(y^0, z^0) + \widetilde{f}_y(y^0, z^0)[y - y^0] + \ \widetilde{f}_z(y^0, z^0)[z - z^0] + \dots \\ &= \widetilde{f}(y^0, z^0) + \widetilde{f}_y(y^0, z^0)[\varepsilon y^1 + \cdots + \varepsilon^N y^N + O(\varepsilon^{N+1})] \\ &\quad + \widetilde{f}_z(y^0, z^0)[\varepsilon z^1 + \cdots + \varepsilon^N z^N + O(\varepsilon^{N+1})] + \dots \quad . \end{aligned}$$

Nun setzen wir diese Taylor-Entwicklungen und entsprechende ε-Entwicklungen in (2.2) ein und führen einen Koeffizientenvergleich durch. Der Koeffizient mit ε^{-1} verschwindet genau dann, wenn

$$\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} z^0 = 0 \tag{2.9}$$

gilt. Die Koeffizienten mit ε^0 verschwinden genau dann, wenn die Gleichungen

$$\begin{aligned} \dot{y}^0 &= S^{-T}(y^0) z^0, \\ \widetilde{M}(y^0)\dot{z}^0 &= \widetilde{f}(y^0, z^0) - \begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} z^1 \end{aligned}$$

erfüllt sind. Da in diesem System z^1 nicht bestimmt werden kann, führen wir via

$$\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} z^1 = G^T \lambda^0 \tag{2.10}$$

einen Lagrangemultplikator λ^0 ein, wobei die von y^0 unabhängige Matrix G durch $G = [0\ I_m] \in \mathbb{R}^{m \times d}$ gegeben ist. Weil die ersten $d-m$ Zeilen der Matrizen nur Nullen als Einträge enthalten, kann gleichwertig

$$G z^1 = \lambda^0 \tag{2.11}$$

formuliert werden. Ebenso schreiben wir äquivalent zu (2.9)

$$G z^0 = 0. \tag{2.12}$$

Mit (2.10) ergibt sich

$$\begin{aligned} \dot{y}^0 &= S^{-T}(y^0) z^0, \\ \widetilde{M}(y^0)\dot{z}^0 &= \widetilde{f}(y^0, z^0) - G^T \lambda^0. \end{aligned} \tag{2.13}$$

Die Bewegungsgleichungen (2.13) und (2.12) stellen ein differential-algebraisches Gleichungssystem vom Index 2 in (y^0, z^0, λ^0) dar, das für alle konsistenten Anfangswerte $(y^0(0), z^0(0))$ eine eindeutige Lösung besitzt [6]. Genauer ist mit der ersten Gleichung aus (2.13) Nebenbedingung (2.9) äquivalent zu $D(y^0)S^{-1}(y^0)z^0 = 0$. Daraus wird ersichtlich, dass die Anfangswerte $(y^0(0), z^0(0))$ konsistent sind, falls die Bedingung $(y^0(0), S^{-1}(y^0(0))z^0(0)) \in \mathcal{M}^0$ erfüllt ist. Um dies anschaulich für das Ausgangssystem (1.5) noch besser zu verstehen, transformieren wir das differential-algebraische System zurück in die Variablen $(y^0, \dot{y}^0, \lambda^0)$. Aus der ersten Gleichung von (2.13) folgt $z^0 = S^T(y^0)\dot{y}^0$ und Differentiation nach der Zeit liefert nunmehr

$\dot{z}^0 = S^T(y^0)\ddot{y}^0 + \mathcal{D}\widetilde{\mathcal{S}}(y^0,\dot{y}^0)\dot{y}^0$. Dabei ist $\big(\mathcal{D}\widetilde{\mathcal{S}}(y,\dot{y})\big)\big|_{ij} = \langle \nabla_y s_{ij}(y), v\rangle$, wobei $s_{ij}(y)$ die Elemente von $S^T(y)$ bezeichnet.

Setzen wir diese Beziehungen in die zweite Gleichung von (2.13) ein, so folgt

$$\begin{aligned}(\widetilde{M}S^T)(y^0)\ddot{y}^0 &= \widetilde{f}(y^0, S^T\dot{y}^0) - \widetilde{M}(y^0)\mathcal{D}\widetilde{\mathcal{S}}(y^0,\dot{y}^0)\dot{y}^0 - G^T\lambda^0,\\ 0 &= GS^T(y^0)\ddot{y}^0.\end{aligned}$$

Mit den in Lemma 1 eingeführten Definitionen von $\widetilde{M}$ und $\widetilde{f}$ erhalten wir schließlich das differential-algebraische System

$$\begin{aligned}M(y^0)\ddot{y}^0 &= f^0(y^0,\dot{y}^0) - S(y^0)G^T\lambda^0\\ 0 &= GS^T(y^0)\dot{y}^0\end{aligned} \tag{2.14}$$

in den Variablen $y^0,\ \dot{y}^0,\ \lambda^0$. Dabei ist

$$f^0(y^0,\dot{y}^0) = f(y^0,\dot{y}^0) - M(y^0)\Big(\mathcal{D}\mathcal{S}(y^0,\dot{y}^0)S^T(y^0)\ + S^{-T}(y^0)\mathcal{D}\widetilde{\mathcal{S}}(y^0,\dot{y}^0)\Big)\dot{y}^0.$$

Die vorliegende Zwangsbedingung lässt sich auch als

$$\begin{pmatrix} 0 & 0\\ 0 & I_m\end{pmatrix} S^T(y^0)\dot{y}^0 = 0$$

formulieren; mit (2.3) ist dies gleichwertig zu $D(y^0)\dot{y}^0 = 0$. Also besitzt das differential-algebraische System (2.14) für Anfangswerte $(y^0(0),\dot{y}^0(0)) \in \mathcal{M}^0$ eine eindeutig bestimmte Lösung. Unter der durchgeführten Rücktransformation bleibt der Störungsindex invariant. Dieser wird in der Bemerkung, die sich an den Beweis anschließt, nachgerechnet.

Die Konstruktion der weiteren Koeffizienten erfolgt analog zur obigen Vorgehensweise. Koeffizienten mit ε^1 verschwinden genau dann, wenn

$$\begin{aligned}\dot{y}^1 &= S^{-T}(y^0)z^1 + \widetilde{H}^1(y^0,z^0,y^1),\\ \widetilde{M}(y^0)\dot{z}^1 &= \widetilde{\Phi}^1(y^0,z^0,\dot{z}^0,y^1,z^1) - \begin{pmatrix} 0 & 0\\ 0 & I_m\end{pmatrix} z^2\end{aligned}$$

gilt, wobei

$$\begin{aligned}\widetilde{\Phi}^1(y^0,z^0,\dot{z}^0,y^1,z^1) &= \widetilde{f}_y(y^0,z^0)y^1 + \widetilde{f}_z(y^0,z^0)z^1 - \frac{\partial}{\partial y}\Big(\widetilde{M}(y)\dot{z}^0\Big)\Big|_{y=y_0} y^1,\\ \widetilde{H}^1(y^0,z^0,y^1) &= \frac{\partial}{\partial y}\Big(S^{-T}(y)z^0\Big)\Big|_{y=y_0} y^1\end{aligned}$$

gesetzt werden. Da die Variable z^2 nicht bestimmt werden kann, führen wir erneut einen Lagrange-Multiplikator λ_1 durch

$$\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} z^2 = G^T \lambda^1 \tag{2.15}$$

ein. Diese Zwangsbedingung ist wiederum gleichwertig zu

$$Gz^2 = \lambda^1.$$

Somit gilt

$$\begin{aligned} \dot{y}^1 &= S^{-T}(y^0) z^1 + \widetilde{H}^1(y^0, z^0, y^1), \\ \widetilde{M}(y^0)\dot{z}^1 &= \widetilde{\Phi}^1(y^0, z^0, \dot{z}^0, y^1, z^1) - G^T \lambda^1. \end{aligned} \tag{2.16}$$

Die Gleichungen (2.16) und (2.11) bilden bei bekannten Variablen $y^0,\ z^0,\ \dot{z}^0$ wieder ein differential-algebraisches Gleichungssystem vom Index 2 in den Unbekannten $y^1,\ z^1,\ \lambda^1$. Der Anfangswert $y^1(0)$ ist frei wählbar, denn die Zwangsbedingungen (2.11) wirken nur auf die Geschwindigkeiten z^1, und $z^1(0)$ ist eindeutig bestimmt, falls es im Bild von $\widetilde{M}^{-1}(y^0(0))G^T$ liegt. Dies wird ersichtlich, wenn die Bedingung $z^1(0) \in Im(\widetilde{M}^{-1}(y^0(0))G^T)$ in (2.11) ausgenutzt wird, denn die Eindeutigkeit von $z^1(0)$ folgt dann sofort aufgrund der Invertierbarkeit von $G\widetilde{M}^{-1}(y^0(0))G^T$. Mit $Im(\cdot)$ sei das Bild der linearen Abbildung bezeichnet, welche die Matrix beschreibt.

Die Veranschaulichung in der Geometrie unseres Problems folgt erneut, indem wir das differential-algebraische System (2.16), (2.11) in die Variablen $y^1,\ \dot{y}^1,\ \lambda^1$ transformieren. Hierzu gehen wir vor wie für das vorige differential-algebraische Gleichungssystem. Zunächst liefert das Einsetzen von $z^0 = S^T(y^0)\dot{y}^0$ in die erste Gleichung von (2.16)

$$z^1 = S^T(y^0)\dot{y}^1 + z^1 + S^T(y^0)\widetilde{H}^1(y^0, S^T(y^0)\dot{y}^0, y^1).$$

Differentiation nach der Zeit ergibt

$$\dot{z}^1 = S^T(y^0)\ddot{y}^1 + \widetilde{\mathcal{DS}}(y^0, \dot{y}^0)\dot{y}^1 + D_t^1(y^0, \dot{y}^0, \ddot{y}^0, y^1, \dot{y}^1),$$

wobei D_t^1 die Zeitableitung des letzten Terms in der Gleichung von z^1 bezeichnet und $\widetilde{\mathcal{DS}}$ wie oben definiert ist. Setzen wir diese Resultate in die zweite Gleichung von (2.16) und die Bedingung für z^1 in (2.11) ein, so geht daraus das differential-algebraische Gleichungssystem

$$\begin{aligned} M(y^0)\ddot{y}^1 &= \Phi^1(y^0, \dot{y}^0, \ddot{y}^0, y^1, \dot{y}^1) - S(y^0)G^T\lambda^1, \\ 0 &= GS^T(y^0)\dot{y}^1 + GS^T(y^0)H^1(y^0, \dot{y}^0, y^1) - \lambda^0 \end{aligned} \tag{2.17}$$

in den Variablen $y^1,\ \dot{y}^1,\ \lambda^1$ hervor, wobei

$$\begin{aligned} \Phi^1 &= S\widetilde{\Phi}^1 - MS^{-T}(\widetilde{\mathcal{DS}}\dot{y}^1 - D_t^1), \\ H^1(y^0, \dot{y}^0, y^1) &= \widetilde{H}^1(y^0, S^T(y^0)\dot{y}^0, y^1) \end{aligned}$$

gesetzt wurde. Die Funktion Φ^1 hängt nur von den Variablen $y^0, \dot{y}^0, \ddot{y}^0, y^1, \dot{y}^1$ ab, wovon der Leser sich durch Betrachtung der auftretenden Funktionen leicht überzeugen kann. Dieses System ist eindeutig lösbar, wenn $y^1(0)$ frei wählbar ist. Dies ist hier der Fall, wie bei der Argumentation für das System (2.16) nachvollzogen wurde. Ist $\dot{y}^1(0)$ im Bild von $(M^{-1}S)(y^0(0))G^T$, so gilt für ein $\mu \in R^m$ die Gleichheit $\dot{y}^1(0) = (M^{-1}S)(y^0(0))G^T\mu$, wegen (2.11) also

$$G(S^T M^{-1} S)(y^0(0))G^T\mu = \lambda^0.$$

Da G vollen Rang hat, ist dieses lineare Gleichungssystem eindeutig lösbar, $\dot{y}^1(0)$ ist also eindeutig bestimmt. Das System (2.17) ist daher eindeutig lösbar, falls

$$\dot{y}^1(0) \in Im\Big((M^{-1}S)(y^0(0))G^T\Big).$$

Dieses Bild ist gleich dem orthogonalen Komplement von $Ker(DS^T(y^0(0)))$ bezüglich des $M(y^0(0))$-Skalarproduktes. Berechnen wir den Tangentialraum von $\mathcal{M}^0$ in $(y^0, \dot{y}^0)$, so ist klar, dass $Ker(\frac{\partial}{\partial y^0}(D(y^0)\dot{y}^0)) \times Ker(DS^T(y^0)) \subseteq T_{(y^0,\dot{y}^0)}\mathcal{M}^0$ gilt; $(0, \dot{y}^1(0))$ liegt also im $M(y^0)$-orthogonalen Komplemet dieses Tangentialraums.

Der Koeffizientenvergleich mit ε^2 liefert ein weiteres differential-algebraisches Gleichungssystem vom Index 2 in Unbekannten $y^2,\ z^2,\ \lambda^2$. Mit dieser Methodik werden die Koeffizientenfunktionen $y^k(t)$ und $z^k(t)$ beziehungsweise $\dot{y}^k(t)$ konstruiert. Setzen wir für $t \in [0, T]$

$$\begin{aligned} y^\varepsilon(t) &= y^0(t) + \varepsilon y^1(t) + \cdots + \varepsilon^N y^N(t), \\ z^\varepsilon(t) &= z^0(t) + \varepsilon z^1(t) + \cdots + \varepsilon^N z^N(t) \end{aligned} \tag{2.18}$$

mit den oben konstruierten, von ε unabhängigen Koeffizienten $y^k(t)$ und $z^k(t)$, so folgt eine abgebrochene ε-Entwicklung. Ein äquivalenter Übergang zu $\dot{y}^\varepsilon(t)$ ist durch die Transformation der differential-algebraischen Gleichungssysteme möglich.

Der Defekt dieser abgebrochenen ε-Entwicklung ist für beliebiges N von der Größenordnug $O(\varepsilon^N)$. Dies wird ersichtlich, wenn wir die Funktionen aus (2.2) um y^ε beziehungsweise $(y^\varepsilon, z^\varepsilon)$ entwickeln. Aus den resultierenden Taylor-Entwicklungen und mit den ε-Entwicklungen von y, $\dot{y}$, z und $\dot{z}$ erhalten wir

$$\begin{aligned} \dot{y}^\varepsilon &= S^{-T}(y^\varepsilon)z^\varepsilon + O(\varepsilon^{N+1}), \\ \widetilde{M}(y^\varepsilon)\dot{z}^\varepsilon &= \widetilde{f}(y^\varepsilon, z^\varepsilon) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} z^\varepsilon + \begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix} O(\varepsilon^N) + O(\varepsilon^{N+1}). \end{aligned} \tag{2.19}$$

Dass Ψ^ε bijektiv ist folgt aus der Existenz der Umkehrabbildung

$$(\Psi^\varepsilon)^{-1} : \mathcal{M}^\varepsilon \to \mathcal{M}^0,\ (y^\varepsilon, \dot{y}^\varepsilon) \mapsto (y^0, \dot{y}^0).$$

(b) Nun zeigen wir, dass für jede Lösung von (2.2) mit Anfangswerten, die

$$y(0) - y^\varepsilon(0) = O(\varepsilon^N), \quad z(0) - z^\varepsilon(0) = O(\varepsilon^N)$$

erfüllen, für Zeiten $t \in [0, T]$ auch

$$y(t) - y^\varepsilon(t) = O(\varepsilon^N), \quad z(t) - z^\varepsilon(t) = O(\varepsilon^N)$$

gilt. Hierzu subtrahieren wir (2.19) von (2.2). Dann ist

$$\begin{aligned} \dot{y} - \dot{y}^\varepsilon &= S^{-T}(y)z - S^{-T}(y^\varepsilon)z^\varepsilon + O(\varepsilon^{N+1}), \\ \widetilde{M}(y)\dot{z} - \widetilde{M}(y^\varepsilon)\dot{z}^\varepsilon &= \widetilde{f}(y,z) - \widetilde{f}(y^\varepsilon, z^\varepsilon) \\ &\quad - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix}(z - z^\varepsilon) + O(\varepsilon^N). \end{aligned} \tag{2.20}$$

Setzen wir $\Delta y = y - y^\varepsilon,\ \Delta z = z - z^\varepsilon,\ \Delta \dot{z} = \dot{z} - \dot{z}^\varepsilon$, so ist (2.20) mit der Lipschitzstetigkeit von S^{-T}, $\widetilde{M}$ und $\widetilde{f}$ äquivalent zu

$$\begin{aligned} \Delta \dot{y} &= S^{-T}(y)\Delta z + O(\|\Delta y\|) + O(\varepsilon^{N+1}), \\ \widetilde{M}(y)\Delta \dot{z} &= -\frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix}\Delta z + O(\|\Delta y\| + \|\Delta z\|) + O(\varepsilon^N). \end{aligned} \tag{2.21}$$

Da $\widetilde{M}(y)$ symmetrisch und positiv definit ist, benutzen wir wie in Lemma 1 eine Zerlegung $\widetilde{M}(y) = \widetilde{M}^{1/2}(y)\widetilde{M}^{1/2}(y)$ mit positiv definitem und symmetrischem $\widetilde{M}^{1/2}$. Als Konsequenz lässt sich (2.21) umformen zu

$$\begin{aligned} \Delta \dot{y} &= S^{-T}(y)\Delta z + O(\|\Delta y\|) + O(\varepsilon^{N+1}), \\ \widetilde{M}^{1/2}(y)\Delta \dot{z} &= -\frac{1}{\varepsilon}B(y)\widetilde{M}^{1/2}(y)\Delta z + O(\|\Delta y\| + \|\Delta z\|) + O(\varepsilon^N) \end{aligned}$$

mit einer Matrix

$$B(y) = \widetilde{M}^{-1/2}(y)\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix}\widetilde{M}^{-1/2}(y),$$

die symmetrisch und positiv semidefinit ist. Mit $\Delta w = \widetilde{M}^{1/2}(y)\Delta z$ folgt

$$\begin{aligned} \Delta \dot{y} &= S^{-T}(y)\widetilde{M}^{-1/2}(y)\Delta z + O(\|\Delta y\|) + O(\varepsilon^{N+1}), \\ \Delta \dot{w} &= -\frac{1}{\varepsilon}B(y)\widetilde{M}^{1/2}(y)\Delta w + O(\|\Delta y\| + \|\Delta w\|) + O(\varepsilon^N). \end{aligned} \tag{2.22}$$

Nun benutzen wir wie in Lemma 1 eine Blockdiagonalisierung, die die positiven Eigenwerte von den Nulleigenwerten von B trennt, also

$$\widetilde{Q}^T(y)B(y)\widetilde{Q}(y) = \begin{pmatrix} 0 & 0 \\ 0 & \widetilde{A}(y) \end{pmatrix}.$$

$\widetilde{Q}(y)$ und $\widetilde{A}(y)$ können mit derselben Argumentation wie im Beweis von Lemma 1 als glatte Funktionen gewählt werden. Hiermit und mit der Transformation $\widetilde{Q}^T(y)\Delta w = (\Delta w_1, \Delta w_2)^T$ spalten wir die zweite Zeile von (2.22) in die gekoppelte Differentialgleichung

$$\begin{aligned} \Delta\dot{w}_1 &= O(\|\Delta y\| + \|\Delta w_1\| + \|\Delta w_2\|) + O(\varepsilon^{N+1}), \\ \Delta\dot{w}_2 &= -\frac{1}{\varepsilon}\widetilde{A}(y)\Delta w_2 + b(t) \end{aligned} \tag{2.23}$$

auf, wobei $\|b(t)\| \leq C(\|\Delta y\| + \|\Delta w_1\| + \|\Delta w_2\|) + O(\varepsilon^N)$ ist. Nach diesen Umformungen ist es nun möglich, Abschätzungen für die einzelnen Größen herzuleiten. Für die Lösung der inhomogenen linearen Differentialgleichung gilt mit Variation der Konstanten

$$\Delta w_2(t) = R_\varepsilon(t)\Delta w_2(0) + \int_0^t R_\varepsilon(t)R_\varepsilon^{-1}(\tau)b(\tau)\, d\tau$$

wobei $R_\varepsilon(t)$ der Propagator der homogenen linearen Differntialgleichung

$$\Delta\dot{w}_2 = -\frac{1}{\varepsilon}\widetilde{A}(y(t))\Delta w_2 \tag{2.24}$$

ist. Eine Abschätzung für $\|R_\varepsilon(t)\|$ ergibt sich, indem wir (2.24) von links mit Δw_2^T multiplizieren. Es ist

$$\Delta w_2^T\Delta\dot{w}_2 = \|\Delta w_2\|\frac{d}{dt}\|\Delta w_2\| \leq -\frac{\widetilde{\alpha}}{\varepsilon}\|\Delta w_2\|^2,$$

denn wegen der positiven Definitheit von $\widetilde{A}(y)$ gilt $-x^T\widetilde{A}(y)x \leq -\widetilde{\alpha}\|x\|^2$ mit einem reellwertigen, positiven $\widetilde{\alpha}$. Division beider Seiten der obigen Gleichung mit $\|\Delta w_2\|$, Anwendung des Hauptsatzes der Differential- und Integralrechnung und zuletzt des Lemmas von Gronwall liefert die Ungleichung

$$\|\Delta w_2(t)\| \leq e^{-\frac{\widetilde{\alpha}}{\varepsilon}t}\|\Delta w_2(0)\|.$$

Damit und aus der Lösung $\Delta w_2(t) = R_\varepsilon(t)\Delta w_2(0)$ des homogenen Problems folgt sofort $\|R_\varepsilon(t)\| \leq e^{-\frac{\widetilde{\alpha}}{\varepsilon}t}$, für das Propagatorprodukt unter dem Integral gilt die Abschätzung $\|R_\varepsilon(t)R_\varepsilon^{-1}(\tau)\| \leq e^{-\frac{\widetilde{\alpha}}{\varepsilon}(t-\tau)}$. Somit folgt aus der Lösung für die inhomogene

lineare Differentialgleichung

$$\|\Delta w_2(t)\| \leq \int_0^t e^{-\frac{\widetilde{\alpha}}{\varepsilon}(t-\tau)} \Big(C(\|\Delta y(\tau)\| + \|\Delta w_1(\tau)\| + \|\Delta w_2(\tau)\|) \Big)\, d\tau + O(\varepsilon^N).$$

Wegen $\int_0^t e^{-\frac{\widetilde{\alpha}}{\varepsilon}(t-\tau)}\, d\tau \leq \frac{\varepsilon}{\widetilde{\alpha}}$ geht daraus

$$\|\Delta w_2(t)\| \leq \frac{\varepsilon}{\widetilde{\alpha}} C \max_{0\leq\tau\leq t} \Big(\|\Delta y(\tau)\| + \|\Delta w_1(\tau)\| + \|\Delta w_2(\tau)\| \Big) + O(\varepsilon^N) \tag{2.25}$$

mit einer von ε unabhängigen Konstanten C hervor. Eine Schranke für $\max \|\Delta w_2\|$ erhalten wir durch Bildung des Maximums über beide Seiten der obigen Gleichung. Es ergibt sich

$$\max_{0\leq\tau\leq t} \|\Delta w_2(\tau)\| \leq C\varepsilon \max_{0\leq\tau\leq t} \Big(\|\Delta y(\tau)\| + \|\Delta w_1(\tau)\| + \|\Delta w_2(\tau)\| \Big) + O(\varepsilon^N).$$

Auflösungen von Ungleichungen dieser Art werden an mehreren Stellen der Arbeit auftreten. Auch wenn es sich um eine einfache Rechnung handelt, soll sie an dieser Stelle einmal exemplarisch durchgeführt werden. Zunächst wenden wir die Dreiecksungleichung für das Maximum auf der rechten Seite an, um den Term $C\varepsilon \max_{0\leq\tau\leq t} \|\Delta w_2(\tau)\|$ zu isolieren und anschließend auf die linke Seite zu bringen. Dann ist

$$(1 - C\varepsilon) \max_{0\leq\tau\leq t} \|\Delta w_2(\tau)\| \leq C\varepsilon \max_{0\leq\tau\leq t} \Big(\|\Delta y(\tau)\| + \|\Delta w_1(\tau)\| \Big) + O(\varepsilon^N).$$

Division beider Seiten mit $1 - C\varepsilon > \frac{1}{2}$ ergibt $\frac{1}{1-C\varepsilon} = 1 + O(\varepsilon) = O(1)$ und mit einer Konstanten $C > 0$, die nicht von ε abängt folgt

$$\max_{0\leq\tau\leq t} \|\Delta w_2(\tau)\| \leq C\varepsilon \max_{0\leq\tau\leq t} \Big(\|\Delta y(\tau)\| + \|\Delta w_1(\tau)\| \Big) + O(\varepsilon^N).$$

Setzen wir dies in (2.25) ein, so ist

$$\|\Delta w_2(t)\| \leq C\varepsilon \max_{0\leq\tau\leq t} \Big(\|\Delta y(\tau)\| + \|\Delta w_1(\tau)\| \Big) + O(\varepsilon^N). \tag{2.26}$$

Multiplikation der ersten Gleichung (2.23) von links mit Δw_1^T und Ausnutzung der Gleichheit $\Delta w_1^T \Delta \dot{w}_1 = \|\Delta w_1\| \cdot \frac{d}{dt}\|\Delta w_1\|$ liefert

$$\frac{d}{dt}\|\Delta w_1\| = O(\|\Delta y\| + \|\Delta w_1\| + \|\Delta w_2\|) + O(\varepsilon^N).$$

Hieraus folgt durch Anwenden des Hauptsatzes der Integral- und Differentialrechnung und aus dem Lemma von Gronwall

$$\|\Delta w_1(t)\| \leq C \max_{0\leq\tau\leq t} \Big(\|\Delta y(\tau)\| + \|\Delta w_2(\tau)\| \Big) + O(\varepsilon^N).$$

Setzen wir die Abschätzung für $\max \|\Delta w_2\|$ ein, so liefern Dreiecksungleichung und Division mit $1 - C\varepsilon > \frac{1}{2}$ erneut

$$\|\Delta w_1(t)\| \leq C \max_{0\leq\tau\leq t} \|\Delta y(\tau)\| + O(\varepsilon^N). \tag{2.27}$$

Einsetzen von (2.27) in (2.26) ergibt

$$\|\Delta w_2(t)\| \leq C \max_{0\leq\tau\leq t} \|\Delta y(\tau)\| + O(\varepsilon^N).$$

Da aber $\widetilde{Q}^T(y)\Delta w = (\Delta w_1, \Delta w_2)^T$ gewählt war und $\widetilde{Q}(y)$ orthogonal und somit beschränkt ist, folgt mit geeigneter Konstante C unabhängig von ε

$$\|\Delta w(t)\| \leq C \max_{0\leq\tau\leq t} \|\Delta y(\tau)\| + O(\varepsilon^N). \tag{2.28}$$

Ganz analog verfahren wir mit Δy. Aus der ersten Gleichung von (2.23) erhalten wir

$$\Delta\dot{y} = O(\|\Delta w\|) + O(\|\Delta y\|) + O(\varepsilon^{N+1}).$$

Wiederum ergibt sich wegen $\Delta y^T \Delta\dot{y} = \|\Delta y\| \cdot \frac{d}{dt}\|\Delta y\|$ in der ersten Gleichung von (2.25)

$$\frac{d}{dt}\|\Delta y\| = O(\|\Delta w\|) + O(\|\Delta y\|) + O(\varepsilon^{N+1})$$

und mit dem Lemma von Gronwall und (2.28) folgt

$$\|\Delta y(t)\| \leq C \max_{0\leq\tau\leq t} \|\Delta y(\tau)\| + O(\varepsilon^N).$$

Die Berechnung von $\max \|\Delta y\|$ wie oben liefert nun

$$\|\Delta y(t)\| = \|y(t) - y^\varepsilon(t)\| = O(\varepsilon^N).$$

Wegen (2.28) ist auch $\|\Delta w(t)\|$ durch $O(\varepsilon^N)$ beschränkt und aus (2.22) folgt dann

$$\|\Delta\dot{y}(t)\| = \|\dot{y}(t) - \dot{y^\varepsilon}(t)\| = O(\varepsilon^N).$$

(c) Um zu zeigen, dass Lösungen mit Anfangswerten $(y^\varepsilon, \dot{y^\varepsilon})$ glatt sind, betrachten wir wieder (2.20). Mit der Lipschitzstetigkeit von S^{-T}, $\widetilde{M}$ und $\widetilde{f}$ folgt sofort

$$S^{-T}(y)z - S^{-T}(y^\varepsilon)z^\varepsilon = O(\varepsilon^{N+1}), \qquad \widetilde{f}(y,z) - \widetilde{f}(y^\varepsilon, z^\varepsilon) = O(\varepsilon^{N+1})$$

und hieraus ist ebenfalls ersichtlich, dass

$$z - z^\varepsilon = O(\varepsilon^{N+1})$$

gelten muss, was aber aus der ε-Entwicklung der Lösung a priori klar ist. Mit $\widetilde{M}(y)\dot{z} - \widetilde{M}(y^\varepsilon)\dot{z}^\varepsilon = \widetilde{M}(y^\varepsilon)(\dot{z} - \dot{z}^\varepsilon) + O(\varepsilon^{N+1})$ folgt

$$\dot{z} - \dot{z}^\varepsilon = O(\varepsilon^N).$$

Indem wir (2.19) und (2.2) nach der Zeit t differenzieren und die dadurch erhaltenen Systeme wieder subtrahieren, gilt mit der analogen Argumentationsweise

$$\ddot{y} - \ddot{y}^\varepsilon = O(\varepsilon^N), \quad \ddot{z} - \ddot{z}^\varepsilon = O(\varepsilon^{N-1}).$$

Fahren wir auf diese Weise fort, so lässt sich für allgemeines $1 \leq k \leq N$ induktiv folgern, dass

$$y^{(k)} - y^{\varepsilon^{(k)}} = O(\varepsilon^{N-k+2}), \quad z^{(k)} - z^{\varepsilon^{(k)}} = O(\varepsilon^{N-k+1})$$

ist. Damit sind alle Behauptungen gezeigt und der Beweis ist abgeschlossen. ∎

Zusammenfassung der differential-algebraischen Gleichungssysteme

Zum Überblick sollen nun nochmals die differential-algebraischen Systeme aufgeführt werden. Zur Konstruktion der Koeffizienten $(y^k(t), \dot{y}^k(t))$ der glatten Lösung von (1.5) erhalten wir die Koeffizienten für $k = 0$ aus

$$\begin{aligned} M(y^0)\ddot{y}^0 &= f^0(y^0, \dot{y}^0) - S(y^0)G^T\lambda^0, \\ 0 &= GS^T(y^0)\dot{y}^0. \end{aligned} \tag{DAE 0}$$

Für $k \geq 1$ ergeben sich $(y^k(t), \dot{y}^k(t))$ aus

$$\begin{aligned} M(y^0)\ddot{y}^k &= \Phi^k(y^0, \dot{y}^0, \ddot{y}^0, \ldots, y^{k-1}, \dot{y}^{k-1}, \ddot{y}^{k-1}, y^k, \dot{y}^k) - S(y^0)G^T\lambda^k, \\ 0 &= GS^T(y^0)\dot{y}^k + GS^T(y^0)H^k(y^0, \dot{y}^0, \ldots, y^{k-1}, \dot{y}^{k-1}, y^k) - \lambda^{k-1} \end{aligned} \tag{DAE k}$$

mit $G = [0\ I_m] \in \mathbb{R}^{m\times d}$, der Transformationsmatrix S aus Lemma 1 sowie Funktionen Φ^k und H^k in denen Terme zusammengefasst sind, die nur von den spezifizierten Variablen abhängen. Von den Koeffizienten der ε-Entwicklungen von $(y(t), z(t))$ erhalten wir $(y^0(t), z^0(t))$ aus

$$\begin{aligned} \dot{y}^0 &= S^{-T}(y^0)z^0, \\ \widetilde{M}(y^0)\dot{z}^0 &= \widetilde{f}(y^0, z^0) - G^T\lambda^0, \\ 0 &= Gz^0. \end{aligned} \tag{DAE 0'}$$

Für $k \geq 1$ ergeben sich $(y^k(t), z^k(t))$ aus

$$\begin{aligned} \dot{y}^k &= S^{-T}(y^0)z^k + \widetilde{H}^k(y^0, z^0, \ldots, y^{k-1}, z^{k-1}, y^k), \\ \widetilde{M}(y^0)\dot{z}^k &= \widetilde{\Phi}^k(y^0, z^0, \dot{z}^0, \ldots, \ldots, y^{k-1}, z^{k-1}, \dot{z}^{k-1}, y^k, z^k) - G^T\lambda^k, \\ 0 &= Gz^k - \lambda^{k-1} \end{aligned} \qquad \text{(DAE } k\text{')}$$

mit Funktionen $\widetilde{\Phi}^k$ und $\widetilde{H}^k$, wobei für gewisse Variablen Linearitäten berücksichtigt werden können. $\widetilde{H}^k$ ist beispielsweise linear in z^i für alle $i = 1 \ldots, k-1$, k-linear in y^1, $(k-1)$-linear in y^2.

Bemerkung. Der Übergang von $(y^\varepsilon, z^\varepsilon)$ zu $(y^\varepsilon, \dot{y}^\varepsilon)$ ist nicht nur über die Rücktransformation der differential-algebraischen Gleichungen möglich, sondern kann auch direkt vollzogen werden. Mit der Taylor-Entwicklung für $S^{-T}(y)z$ gilt

$$\begin{aligned} S^{-T}(y)z &= S^{-T}(y^0)(z^0 + \cdots + \varepsilon^N z^N + O(\varepsilon^{N+1})) + \sum_{k=1}^{N} \varepsilon^k \widetilde{H}^k \\ &= S^{-T}(y^0)z^0 + \varepsilon(S^{-T}(y^0)z^1 + \widetilde{H}^1) + \ldots \\ &\qquad + \varepsilon^N(S^{-T}(y^0)z^N + \widetilde{H}^N) + O(\varepsilon^{N+1}), \end{aligned}$$

was mit den Gleichheiten für $\dot{y}^0, \ldots, \dot{y}^k$ aus den differential-algebraischen Gleichungssystemen (DAE 1'), …, (DAE k') äquivalent zu der Existenz einer ε-Entwicklung von $\dot{y}(t)$ ist:

$$\dot{y}(t) = \dot{y}^0(t) + \varepsilon\dot{y}^1(t) + \cdots + \varepsilon^N\dot{y}^N(t) + O(\varepsilon^{N+1}).$$

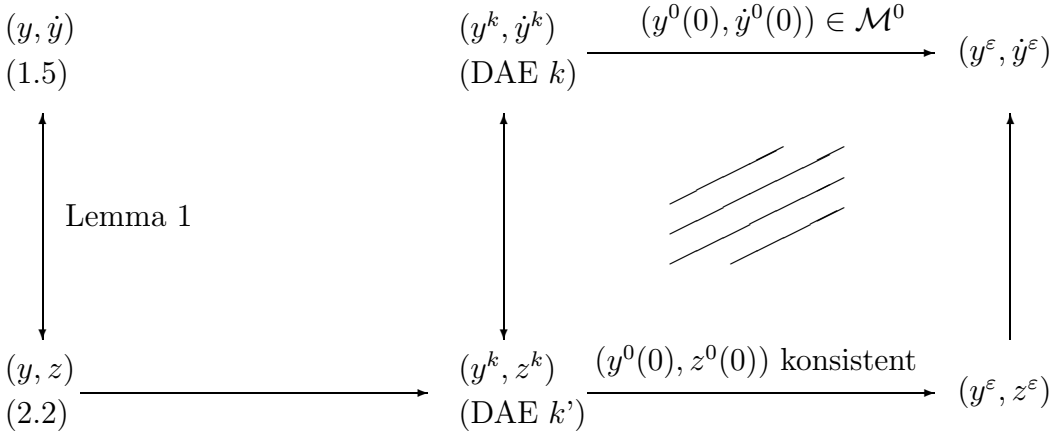

Abbildung 2.1: Beweisverfahren im Überblick

In Abbildung 2.1 sind die Zusammenhänge zur Veranschaulichung der Vorgehensweise im Beweis von Theorem 1 skizziert.

Bemerkung. Sowohl das differential-algebraische System (DAE 0) - (DAE k) als auch das System (DAE 0') - (DAE k') hat den Störungsindex $2+k$. Es ergibt sich also eine Folge von differential-algebraischen Gleichungssystemen mit Index $2, 3, 4, 5, \ldots$.

Beweis. Es genügt, den Störungsindex des Systems (DAE 0) - (DAE k) zu berechnen; für die differential-algebraischen Gleichungssysteme (DAE 0') - (DAE k') können wir ganz analog verfahren. Aus Gründen der Übersichtlichkeit soll auf die Angabe der Positions-, Geschwindigkeits- und Beschleunigungsvariablen teilweise verzichtet werden.

Die Bestimmung des Störungsindex folgt durch Induktion über k. Wir betrachten zunächst das System (DAE 0) und zeigen, dass der Störungsindex dieses differential-algebraischen Gleichungssystems 2 ist. Hierzu differenzieren wir die Zwangsbedingung nach der Zeit und erhalten

$$0 = GS^T(y^0)\ddot{y}^0 + G\dot{S}^T(y^0)\dot{y}^0. \tag{2.29}$$

Somit folgen $\ddot{y}^0$ und λ^0 aus

$$\begin{pmatrix} M & SG^T \\ GS^T & 0 \end{pmatrix} \begin{pmatrix} \ddot{y}^0 \\ \lambda^0 \end{pmatrix} = \begin{pmatrix} f \\ -G\dot{S}^T\dot{y}^0 \end{pmatrix}.$$

Einsetzen von $\ddot{y}^0 = M^{-1}(f - SG^T\lambda^0)$ in (2.29) liefert

$$\lambda^0 = (GS^TM^{-1}SG^T)^{-1}(GS^TM^{-1}f + G\dot{S}^T\dot{y}^0),$$

denn $GS^TM^{-1}SG^T$ ist invertierbar. Setzen wir die Lösung von λ^0 nun in die Gleichung für $\ddot{y}^0$ ein, so ergibt sich eine gewöhnliche Differentialgleichung zweiter Ordnung, die wir als Differentialgleichung erster Ordnung

$$\begin{aligned} \dot{y}^0 &= v^0, \\ \dot{v}^0 &= M^{-1}(f - P(GS^TM^{-1}f + G\dot{S}^T\dot{y}^0)) \end{aligned} \tag{2.30}$$

mit $P = SG^T(GS^TM^{-1}SG^T)^{-1}$ formulieren. Nun betrachten wir (DAE 0'), formuliert als System erster Ordnung, in einer gestörten Version

$$\begin{aligned} \dot{\widehat{y^0}} &= \widehat{v^0} + \gamma_1(t), \\ M(\widehat{y^0})\dot{\widehat{v^0}} &= f(\widehat{y^0}, \widehat{v^0}) - G^T\widehat{\lambda^0} + \theta_1(t), \\ 0 &= GS^T(\widehat{y^0})\widehat{v^0} + \delta_1(t), \\ 0 &= GS^T(\widehat{y^0})\dot{\widehat{v^0}} + G\dot{S}^T(\widehat{y^0})\widehat{\dot{y}^0} + \dot{\delta}_1(t), \end{aligned} \tag{2.31}$$

wobei durch $\gamma_1(t)$, $\theta_1(t)$ und $\delta_1(t)$ Störungen vorgegeben seien. Wie für (DAE 0') leiten wir auch für dieses System zwei gekoppelte Differentialgleichungen der Ordnung 1 her und erhalten

$$\begin{aligned} \dot{\widehat{y^0}} &= \widehat{v^0} + \gamma_1, \\ \dot{\widehat{v^0}} &= M^{-1}\Big(f - P(GS^TM^{-1}f + G\dot{S}^T\widehat{v^0} + GS^TM^{-1}\theta_1 + \dot{\delta}_1) + \theta_1\Big). \end{aligned}$$

Dieses System entspricht der Differentialgleichung (2.30), gestört durch

$$\delta = \begin{pmatrix} \gamma_1 \\ -M^{-1}(P(GS^TM\theta_1 + \dot{\delta}_1) + \theta_1) \end{pmatrix},$$

und wie in [3] folgt nun, dass

$$\begin{aligned} \|\widehat{y^0}(t) - y^0(t)\| &\leq C \max_{t_0\leq s\leq t} \int_{t_0}^{t} \gamma_1 \, ds, \\ \|\widehat{v^0}(t) - v^0(t)\| &\leq C \max_{t_0\leq s\leq t} \int_{t_0}^{t} -M^{-1}(P(GS^TM\theta_1 + \dot{\delta}_1) + \theta_1) \, ds \end{aligned}$$

gilt, das differential-algebraische Gleichungssystem (DAE 0) also den Störungsindex 2 hat.

Für den Induktionsschritt $(k{-}1) \to k$ sei als Voraussetzung bereits gezeigt, dass das differential-algebraische System (DAE 0 ... DAE $(k{-}1)$) den Index $2{\cdot}(k{-}1)$ hat. Betrachten wir den Induktionsanfang genau und schließen auf folgende Schritte, so ist klar, dass der Störterm $\delta_1(t)$ in jedem Schritt als Störung mit der höchsten Zeitableitung auftritt; $\delta_i(t)$ für $i > 1$ und auch $\theta_i(t)$ sowie $\gamma_i(t)$ kommen nur in niedrigeren Ableitungen vor. Um die Behauptung also für k zu zeigen, ist nachzuweisen, dass $\delta_1(t)$ in Abschätzungen für $\|\widehat{y^k}(t) - y(t)\|$, $\|\widehat{v^k}(t) - v(t)\|$ in seiner $(k{+}1)$-ten

Zeitableitung auftritt. Wir studieren also das System

$$\begin{aligned}
\dot{\widehat{y^0}} &= \widehat{v^0} + \gamma_1(t),\\
&\vdots\\
\dot{\widehat{y^k}} &= \widehat{v^k} + \gamma_k(t),\\
M(\widehat{y^0})\dot{\widehat{v^0}} &= f(\widehat{y^0},\widehat{v^0}) - S(\widehat{y^0})G^T\widehat{\lambda^0} + \theta_1(t),\\
&\vdots\\
M(\widehat{y^0})\dot{\widehat{v^k}} &= \Phi^k(\widehat{y^0},\widehat{v^0},\dot{\widehat{v^0}},\ldots,\widehat{y^k},\widehat{v^k}) - S(\widehat{y^0})G^T\widehat{\lambda^k} + \theta_k(t),\\
0 &= GS^T(\widehat{y^0})\dot{\widehat{v^0}} + G\dot{S}^T(\widehat{y^0})\widehat{v^0} + \dot{\delta}_1(t),\\
&\vdots\\
0 &= GS^T(\widehat{y^0})\dot{\widehat{v^k}} + G\dot{S}^T(\widehat{y^0})\widehat{v^k} - \widehat{\dot{\lambda}^{k-1}}\\
&\qquad + \frac{d}{dt}\Big(GS^T(\widehat{y^0})H^k(\widehat{y^0},\widehat{v^0},\ldots,\widehat{y^k})\Big) + \dot{\delta}_k(t).
\end{aligned}$$

Rechnerisch gehen wir wie beim Induktionsanfang vor. Die Differentialgleichung mit $\dot{\widehat{v^k}}$ multiplizieren wir von links mit $M^{-1}(y^0)$. Das Ergebnis setzen wir in die Zwangsbedingung ein, in der $\dot{\widehat{v^k}}$ auftritt, also in die Letzte der obigen Zwangsbedingungen. Das ergibt eine Darstellung von $\widehat{\lambda^k}$ (welche insbesondere linear von $\widehat{\dot{\lambda}^{k-1}}$ abhängt), die wiederum in die Gleichung für $\dot{\widehat{v^k}}$ einsetzt wird. So leiten wir für den gestörten Wert von $\dot{v}^k$ schließlich die Differentialgleichung

$$\dot{\widehat{v^k}} = M^{-1}\Big(\Phi^k - [GS^TM^{-1}\Phi^k + G\dot{S}^T\widehat{v^k} + \frac{d}{dt}\Big(GS^TH^k\Big) + GS^TM^{-1}\theta^k - \widehat{\dot{\lambda}^{k-1}} + \dot{\delta}_k] + \theta_k\Big)$$

her. Nur in dieser Gleichung tritt der Lagrangemultiplikator $\widehat{\dot{\lambda}^{k-1}}$ auf, der auch unter allen Variablen allein von $\delta_1^{(k+1)}$ abhängt. Die Gleichung für $\dot{\widehat{v}}^k$ ist also für den Index des Gesamtsystems entscheidend. Im Vergleich zu ihrer ungestörten Version ist sie unter der Voraussetzung, dass gestörtes und ungestörtes System von den gleichen Anfangswerten initialisiert werden, gestört durch

$$\delta = -M^{-1}\Big(P[GS^TM^{-1}\theta^k - \widehat{\dot{\lambda}^{k-1}} + \dot{\delta}_k] + \theta_k\Big).$$

Mit demselben Argument wie im Induktionsanfang ergeben sich nun entsprechende Abschätzungen. Die für den Index entscheidende Abschätzung ist durch

$$\|\widehat{v^k}(t) - v^k(t)\| \le C \max_{t_0\le s\le t} \int_{t_0}^{t} -M^{-1}\Big(P[GS^TM^{-1}\theta^k - \widehat{\dot{\lambda}^{k-1}} + \dot{\delta}_k] + \theta_k\Big)\, ds$$

gegeben. Da $\widehat{\dot{\lambda}^{k-1}}$ linear von $\delta_1^{(k+1)}$ abhängt, hat das differential-algebraische Gleichungssystem (DAE k … DAE 0) den Störungsindex $k+2$. ∎

2.3 Existenz einer attraktiven invarianten Mannigfaltigkeit

Vorbereitend formulieren wir ein Resultat über attraktive invariante Mannigfaltigkeiten, das auf Untersuchungen von Kirchgraber, Lasagni, Nipp und Stoffer basiert [17], [14].

Wir betrachten eine Abbildung $\Psi : X \to X$ auf einem Banachraum X. Das Paar (X, Ψ) bildet ein *diskretes dynamisches System*. Durch eine Folge $(a_n)_{n\in\mathbb{N}_0}$, wobei $a_{n+1} = \Psi(a_n)$ sei, ist ein *Orbit* gegeben. Er wird durch a_0 erzeugt. Eine Teilmenge $\mathcal{N} \subseteq X$ heißt *invariant*, falls $\Psi(\mathcal{N}) \subset \mathcal{N}$ gilt. $\mathcal{N}$ heißt *attraktiv*, falls jeder Orbit $(a_n)_{n\in\mathbb{N}_0}$ gegen $\mathcal{N}$ konvergiert, also $dist(a_n, \mathcal{N}) \to 0$ für $n \to \infty$ gilt.

Theorem 2. *Seien X, X' Banachräume und Y eine abgeschlossene, beschränkte Teilmenge von X'. Weiter sei eine Abbildung $\Psi : X \times Y \to X \times Y$ gegeben, die für $n \in \mathbb{N}_0$ als $\Psi(\xi_n, \eta_n) = (\xi_{n+1}, \eta_{n+1})$ geschrieben wird, wobei*

$$\begin{aligned} \xi_{n+1} &= \xi_n + \mathcal{F}(\xi_n, \eta_n), \\ \eta_{n+1} &= \mathcal{G}(\xi_n, \eta_n). \end{aligned} \tag{2.32}$$

Die Funktionen $\mathcal{F}$ und $\mathcal{G}$ seien Lipschitzstetig, wobei die Lipschitzkonstanten von $\mathcal{F}$ in den Variablen ξ_n und η_n mit $L_{\xi\xi}$ beziehungsweise $L_{\xi\eta}$ und die Lipschitzkonstanten von $\mathcal{G}$ auf analoge Art und Weise mit $L_{\eta\xi}$ und $L_{\eta\eta}$ bezeichnet seien. Erfüllen diese Lipschitzkonstanten

$$L_{\xi\xi} + L_{\eta\eta} + 2\sqrt{L_{\xi\eta}L_{\eta\xi}} < 1, \tag{2.33}$$

so gilt:

(i) Es gibt eine Lipschitzstetige Funktion $s : X \to Y$, so dass die folgende Implikation gilt:

$$\eta_0 = s(\xi_0) \quad \Rightarrow \quad \eta_n = s(\xi_n) \quad \text{für alle } n. \tag{2.34}$$

Die Menge $\mathcal{N} = \{(\xi, s(\xi)) : \xi \in X\} \subseteq X \times Y$ ist also eine invariante Mannigfaltigkeit für (2.32).

(ii) $\mathcal{N}$ ist attraktiv, denn für alle $(\xi_n, \eta_n) \in X \times Y$ gilt mit $\rho = \lambda L_{\xi\eta} + L_{\eta\eta} < 1$

$$\|\eta_{n+1} - s(\xi_{n+1})\| \leq \rho\|\eta_n - s(\xi_n)\|.$$

(iii) Es gilt die Eigenschaft der asymptotischen Phase, das heißt zu jedem Paar $(\xi_0, \eta_0) \in X \times Y$ existiert ein $(\xi_0^*, \eta_0^*) \in \mathcal{N}$, so dass die Lösungen (ξ_n, η_n) und (ξ_n^*, η_n^*) von (2.32), die von (ξ_0, η_0) beziehungsweise (ξ_0^*, η_0^*) erzeugt werden, mit

$$\|(\xi_n, \eta_n) - (\xi_n^*, \eta_n^*)\| \leq C_1\rho^n \cdot \|(\xi_0, \eta_0) - (\xi_0^*, \eta_0^*)\| \tag{2.35}$$

gegeneinander konvergieren. Weiter gilt

$$\|(\xi_0, \eta_0) - (\xi_0^*, \eta_0^*)\| \leq C_2 \cdot \|\eta_0 - s(\xi_0)\|. \tag{2.36}$$

Die Konstanten C_1 und C_2 hängen nur von den Größen in (2.33) ab.

$\mathcal{N}$ ist also eine attraktive invariante Mannigfaltigkeit bezüglich Ψ.

Beweis: [17] und [14].

In Abschnitt 1.2 wurde die geometrische Besonderheit des qualitativen Verhaltens der Lösungen von stark gedämpften mechanischen System schon beispielhaft beschrieben. Dabei fällt auf, dass auftretende Bewegungen oder Schwingungen von Körpern, die an Verbindungselemente mit starken Dämpfungskräften gekoppelt sind, sehr schnell abklingen. Das folgende Theorem zeigt, dass dieses Verhalten im kontinuierlichen Fall durch die schnelle Annäherung der Lösungen $(y, \dot{y})$ von (1.5) an eine attraktive invariante Mannigfaltigkeit erfasst werden kann.

Theorem 3. *(i) Erfüllen die Anfangswerte $(y(0), \dot{y}(0)) = (y_0, \dot{y}_0)$ von (1.5) die Bedingung $D(y_0)\dot{y}_0 = O(\varepsilon)$, so existiert zu dem stark gedämpften mechanischen System (1.5) eine attraktive invariante Mannigfaltigkeit $\mathcal{N}^\varepsilon$. Lösungen, die außerhalb von $\mathcal{N}^\varepsilon$ starten, nähern sich der Mannigfaltigkeit mit exponentieller Geschwindigkeit. Genauer gilt: Zu einer analytischen Lösung $(y(t), \dot{y}(t))$ von (1.5) mit Anfangswerten $(y_0, \dot{y}_0)$ gibt es ein Paar $(y^*(t), \dot{y}^*(t)) \in \mathcal{N}^\varepsilon$, so dass für $(y_0, \dot{y}_0) - (y_0^*, \dot{y}_0^*) = O(1)$ zur Zeit $t > 0$*

$$\|(y(t), \dot{y}(t)) - (y^*(t), \dot{y}^*(t))\| = O(e^{-C\frac{t}{\varepsilon}})$$

mit geeigneter Konstante C unabhängig von t und ε gilt.

(ii) Für beliebiges $N \geq 1$ liegen die attraktive invariante Mannigfaltigkeit $\mathcal{N}^\varepsilon$ und die Mannigfaltigkeit $\mathcal{M}^\varepsilon$ der abgebrochenen ε-Entwicklungen $O(\varepsilon^N)$-nahe beieinander, das heißt für alle $(y^\varepsilon, \dot{y^\varepsilon}) \in \mathcal{M}^\varepsilon$ und für alle $(y^*, \dot{y}^*) \in \mathcal{N}^\varepsilon$ gilt

$$dist\Big((y^\varepsilon, \dot{y^\varepsilon}), \mathcal{N}^\varepsilon\Big) = O(\varepsilon^N),\ dist\Big((y^*, \dot{y}^*), \mathcal{M}^\varepsilon\Big) = O(\varepsilon^N).$$

Lösungen von (1.5), die auf $\mathcal{N}^\varepsilon$ verlaufen, besitzen also ebenfalls ε-Entwicklungen, die bis auf $O(\varepsilon^N)$ eindeutig sind.

Beweis. Die Beweisidee liegt darin, für die Lösungen von (1.5) ein System wie (2.32) zu formulieren und die Voraussetzungen von Theorem 2 nachzuweisen. Der wesentliche Aufwand resultiert aus der Abschätzung der Lipschitzkonstanten, um die Bedingung (2.33) zu verifizieren. Auch hier erweist sich die Differentialgleichung

(1.5) als ungünstiges Ausgangssystem. Wir ziehen daher die Umformulierung aus Lemma 2 zur Untersuchung heran.

Wir betrachten (2.5) und fassen die Lösungen dieses Systems als von Zeit und Anfangswerten abhängige Funktionen auf, also

$$\begin{aligned} u &: [0,T] \times \mathbb{R}^{2d-m} \times \mathbb{R}^m \to \mathbb{R}^{2d-m},\ (t,u_0,x_0) \mapsto u(t,u_0,x_0), \\ x &: [0,T] \times \mathbb{R}^{2d-m} \times \mathbb{R}^m \to \mathbb{R}^m,\ (t,u_0,x_0) \mapsto x(t,u_0,x_0). \end{aligned}$$

Statt $u(t,u_0,x_0)$ und $x(t,u_0,x_0)$ schreiben wir im Folgenden kurz $u(t)$ und $x(t)$. Für die erste Gleichung von (2.5) wird der Hauptsatz der Differential und Integralrechnung, für die zweite Gleichung wird Variation der Konstanten auf einem Intervall $[0,h]$ angewandt. Das führt auf

$$\begin{aligned} u(h) &= u_0 + \int_0^h F(u(s),x(s))\ ds, \\ x(h) &= R_\varepsilon(h)x_0 + \int_0^h R_\varepsilon(h)R_\varepsilon^{-1}(s)\varphi(u(s),x(s))\ ds, \end{aligned} \tag{2.37}$$

wobei $R_\varepsilon(h)$ der Propagator zu der zweiten Gleichung von (2.5) sei. Mit

$$\begin{aligned} &\mathcal{F}(u,x) = \int_0^h F(u(s),x(s))\ ds, \\ &\mathcal{G}(u,x) = R_\varepsilon(h)x_0 + \int_0^h R_\varepsilon(h)R_\varepsilon^{-1}(s)\varphi(u(s),x(s))\ ds, \\ &\xi_{n+1} = u(h),\ \xi_n = u_0,\ \eta_{n+1} = x(h),\ \eta_n = x_0 \end{aligned}$$

ist dies ein System der Form (2.32). Um Theorem 2 anwenden zu können ist noch die Lipschitzstetigkeit der beteiligten Funktionen und die Bedingung (2.33) zu zeigen.

Hierzu differenzieren wir $\mathcal{F}$ nach den Anfangswerten. Mit der Notation $U_1 = \frac{\partial u}{\partial u_0}$, $U_2 = \frac{\partial u}{\partial x_0}$, $F_u = \frac{\partial F}{\partial u}$ und analog $X_1 = \frac{\partial x}{\partial u_0}$, $X_2 = \frac{\partial x}{\partial x_0}$, $F_x = \frac{\partial F}{\partial x}$ ergibt sich

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial u_0} &= \int_0^h F_u(s)X_1(s) + F_x(s)Z_1(s)\ ds, \\ \frac{\partial \mathcal{F}}{\partial x_0} &= \int_0^h F_u(s)X_2(s) + F_x(s)Z_2(s)\ ds. \end{aligned}$$

Schranken für die Lipschitzkonstanten von $\mathcal{F}$ folgen dann aus

$$\begin{aligned}\left\|\frac{\partial \mathcal{F}}{\partial u_0}\right\| &\leq \int_0^h \|F_u(s)\|\|U_1(s)\| + \|F_x(s)\|\|X_1(s)\|\, ds,\\ \left\|\frac{\partial \mathcal{F}}{\partial x_0}\right\| &\leq \int_0^h \|F_u(s)\|\|U_2(s)\| + \|F_x(s)\|\|X_2(s)\|\, ds.\end{aligned} \tag{2.38}$$

Es müssen also $\|U_1(s)\|$, $\|U_2(s)\|$, $\|X_1(s)\|$, $\|X_2(s)\|$ abgeschätzt werden, denn $\|F_u(s)\|$ und $\|F_x(s)\|$ sind lokal beschränkt, da F eine Funktion mit beschränkten Ableitungen ist. Um diese Abschätzungen vorzunehmen, differenzieren wir die erste Gleichung von (2.5) nach den Anfangswerten u_0 und z_0 und erhalten mit der eingeführten Notation

$$\begin{aligned}\dot{U}_1 &= F_u U_1 + F_x X_1,\\ \dot{U}_2 &= F_u U_2 + F_x X_2.\end{aligned} \tag{2.39}$$

Aus dem Hauptsatz der Integral- und Differentialrechnung und der Linearität des Integrals folgt sofort

$$\|U_1(h)\| \leq \|U_1(0)\| + \int_0^h \|F_u(s)\|\|U_1(s)\| + \|F_x(s)\|\|X_1(s)\|\, ds,$$

$$\|U_2(h)\| \leq \|U_2(0)\| + \int_0^h \|F_u(s)\|\|U_2(s)\| + \|F_x(s)\|\|X_2(s)\|\, ds,$$

und wegen $\|U_1(0)\| = 1$ sowie $\|U_2(0)\| = 0$ weiter

$$\|U_1(h)\| \leq 1 + \max_{0\leq s\leq h}(\|F_x(s)\|\|X_1(s)\|) + \int_0^h \|F_u(s)\|\|U_1(s)\|\, ds,$$

$$\|U_2(h)\| \leq \max_{0\leq s\leq h}(\|F_x(s)\|\|X_2(s)\|) + \int_0^h \|F_u(s)\|\|U_2(s)\|\, ds.$$

Auf diese Ungleichungen lässt sich das Lemma von Gronwall anwenden, und als Konsequenz erhalten wir mit $\|F_u(s)\| \leq C_u$ und $\|F_x(s)\| \leq C_x$

$$\begin{aligned}\|U_1(h)\| &\leq \Big(1 + C_x \max_{0\leq s\leq h}(\|X_1(s)\|)\Big)e^{hC_u},\\ \|U_2(h)\| &\leq C_x \max_{0\leq s\leq h}(\|X_2(s)\|)e^{hC_u}.\end{aligned} \tag{2.40}$$

$\|U_1\|$ und $\|U_2\|$ hängen also noch von $\|X_1\|$ und $\|X_2\|$ ab. Die Schranken für diese Normen ergeben sich aus der zweiten Gleichung von (2.37), denn $\|\frac{\partial \mathcal{G}}{\partial u_0}\| = \|X_1\|$ und $\|\frac{\partial \mathcal{G}}{\partial x_0}\| = \|X_2\|$. Analog zur obigen Vorgehensweise differenzieren wir nun die zweite Gleichung von (2.5) nach den Anfangswerten und erhalten

$$\begin{aligned} \dot{X}_1 &= -\frac{1}{\varepsilon}A(u)X_1 - \frac{1}{\varepsilon}B_1(u,x)U_1 + \varphi_u U_1 + \varphi_x X_1, \\ \dot{X}_2 &= -\frac{1}{\varepsilon}A(u)X_2 - \frac{1}{\varepsilon}B_2(u,x)U_2 + \varphi_u U_2 + \varphi_x X_2, \end{aligned} \tag{2.41}$$

wobei $B_1(u,x)U_1 := (\frac{\partial}{\partial u_0}(A(u)x))U_1$ und $B_2(u,x)U_2 := (\frac{\partial}{\partial x_0}(A(u)x))U_2$ gesetzt wurde und wir wieder kurz $B_i(u(t),x(t)) = B_i(u,x)$ schreiben. Mit der gleichen Argumentation wie im Beweis von Theorem 1 folgt aus (2.5) allgemein für beliebieges $t \in [0,T]$

$$\begin{aligned} \|x(t)\| &\leq e^{-\frac{\alpha}{\varepsilon}t}\|x(0)\| + \int_0^t e^{-\frac{\alpha}{\varepsilon}(h-s)}\|\varphi(u(s),x(s))\|\, ds \\ &\leq e^{-\frac{\alpha}{\varepsilon}t}\|x(0)\| + \frac{\varepsilon}{\alpha}\max_{0\leq s\leq t}\|\varphi(u(s),x(s))\|, \end{aligned}$$

und da die Funktion φ stetig ist, ist $\|\varphi(t)\|$ für $t \in [0,T]$ beschränkt. Mit der Voraussetzung $D(y_0)\dot{y}_0 = O(\varepsilon)$ ist $\|A(u(0))x(0)\| = O(\varepsilon)$, und somit gilt auch $\|x(0)\| = O(\varepsilon)$. Da die Funktionen B_1 und B_2 linear von $x(t)$ abhängen, lassen sich die Normen durch

$$\|B_1\| \leq C\varepsilon, \quad \|B_2\| \leq C\varepsilon,$$

beschränken, wobei C Konstanten sind, die nicht von ε und t abhängen. Mit (2.41) verfahren wir genauso wie zuvor mit den Differentialgleichungen für U_1 und U_2. Mit $-\frac{1}{\varepsilon}B_i(u,x)U_i + \varphi_u U_i + \varphi_x X_i =: b_i(t)$ für $i = 1,2$ geht daraus mit Variation der Konstanten die Abschätzung

$$\begin{aligned} \|X_1(h)\| &\leq \|e^{-\frac{\alpha}{\varepsilon}h}\| \cdot \|X_1(0)\| + \int_0^h \|e^{-\frac{1}{\varepsilon}(h-s)\alpha}\|\|b_1(s)\|\, ds, \\ \|X_2(h)\| &\leq \|e^{-\frac{\alpha}{\varepsilon}h}\| \cdot \|X_2(0)\| + \int_0^h \|e^{-\frac{1}{\varepsilon}(h-s)\alpha}\|\|b_2(s)\|\, ds \end{aligned}$$

hervor. Wegen $\|X_1(0)\| = 0$ und $\|X_2(0)\| = 1$ schließen wir

$$\begin{aligned}\|X_1(h)\| &\leq \int_0^h e^{-\frac{\alpha}{\varepsilon}(h-s)}\|b_1(s)\|\,ds,\\ \|X_2(h)\| &\leq e^{-\frac{\alpha}{\varepsilon}h} + \int_0^h e^{-\frac{\alpha}{\varepsilon}(h-s)}\|b_2(s)\|\,ds.\end{aligned}$$

Nutzen wir $\int_0^h e^{-\frac{\alpha}{\varepsilon}(h-s)}\,ds < \frac{\varepsilon}{\alpha}$ aus, so folgt

$$\begin{aligned}\|X_1(h)\| &\leq \frac{\varepsilon}{\alpha}\max_{0\leq s\leq h}\Big\| -\frac{1}{\varepsilon}B_1(t)U_1(s) + \varphi_u(s)U_1(s) + \varphi_x(s)X_1(s)\Big\|,\\ \|X_2(h)\| &\leq e^{-\frac{\alpha h}{\varepsilon}} + \frac{\varepsilon}{\alpha}\max_{0\leq s\leq h}\Big\| -\frac{1}{\varepsilon}B_2(t)U_2(s) + \varphi_u(s)U_2(s) + \varphi_x(s)X_2(s)\Big\|.\end{aligned}$$

Ungleichungen dieser Art wurden bereits im Beweis von Theorem 1 behandelt. Wir gehen hier genau gleich vor und isolieren in der ersten Ungleichung zunächst $\|X_1\|$ auf der rechten Seite durch Verwenden der Dreiecksungleichung. Bilden wir das Maximum über beide Seiten, fassen die Terme mit $\max\|X_1\|$ zusammen und teilen dann durch $(1 - C\varepsilon) > \frac{1}{2}$, so gilt

$$\max_{0\leq s\leq h}\|X_1(s)\| \leq C\varepsilon\max_{0\leq s\leq h}\Big\| -\frac{1}{\varepsilon}B_1(s)U_1(s) + \varphi_u(s)U_1(s)\Big\|,$$

und durch Einsetzen folgt

$$\|X_1(h)\| \leq C(\varepsilon+\varepsilon^2)\max_{0\leq s\leq h}\Big\| -\frac{1}{\varepsilon}B_1(s)U_1(s) + \varphi_u(s)U_1(s)\Big\|.$$

Analog verfahren wir mit der Ungleichung für X_2 und erhalten

$$\max_{0\leq s\leq h}\|X_2(s)\| \leq Ce^{-\frac{\alpha}{\varepsilon}h} + C\varepsilon\max_{0\leq s\leq h}\Big\| -\frac{1}{\varepsilon}B_2(s)U_2(s) + \varphi_x(s)U_2(s)\Big\|$$

und daraus

$$\|X_2(h)\| \leq (1+C\varepsilon)e^{-\frac{\alpha}{\varepsilon}h} + C(\varepsilon+\varepsilon^2)\max_{0\leq s\leq h}\Big\| -\frac{1}{\varepsilon}B_2(s)U_2(s) + \varphi_x(s)U_2(s)\Big\|.$$

Somit haben wir für $\|X_1(h)$ und $\|X_2(h)\|$ Abschäztungen gewonnen, die von $\|U_1\|$ und $\|U_2\|$ abhängen. Mit Anfangswerten $(y_0, \dot{y}_0)$, für die $D(y_0)\dot{y}_0 = O(\varepsilon)$ gilt, ist $\|B_1(u,x)\| = O(\varepsilon)$, $\|B_2(u,x)\| = O(\varepsilon)$, und mit (2.40) ergibt sich mit der gleichen Argumentationsweise wie für $\|X_1(h)\|$ und $\|X_2(h)\|$

$$\max_{0\leq s\leq h}\|U_1(s)\| \leq Ce^{hC_u}, \quad \max_{0\leq s\leq h}\|U_2(s)\| \leq Ce^{hC_u}e^{-\frac{\alpha}{\varepsilon}h}$$

und damit

$$\|U_1(h)\| \leq Ce^{hC_u}(1+\varepsilon+\varepsilon^2), \quad \|U_2(h)\| \leq Ce^{hC_u}(e^{-\frac{\alpha}{\varepsilon}h}+\varepsilon+\varepsilon^2+\varepsilon^3).$$

Nun lassen sich mit geeigneten Konstanten C auch $\|X_1(h)\|$ und $\|X_2(h)\|$ nach oben abschätzen. Es gilt

$$\|X_1(h)\| \leq Ce^{hC_u}(\varepsilon+\varepsilon^2) \quad \text{und} \quad \|X_2(h)\| \leq Ce^{hC_u}e^{-\frac{\alpha}{\varepsilon}h}(1+\varepsilon+\varepsilon^2+\varepsilon^3).$$

Mit diesen Abschätzungen können nun Schranken für die Lipschitzkonstanten berechnet werden, denn es ist

$$\begin{aligned}\left\|\frac{\partial \mathcal{F}}{\partial u_0}\right\| &\leq C_u \int_0^h \|U_1(s)\|\, ds + C_x \int_0^h \|X_1(s)\|\, ds \\ &\leq hC_u \max_{0\leq s\leq h}\|U_1(s)\| + hC_x \max_{0\leq s\leq h}\|X_1(s)\| \\ &\leq Ch\end{aligned}$$

und

$$\begin{aligned}\left\|\frac{\partial \mathcal{F}}{\partial x_0}\right\| &\leq hC_u \max_{0\leq s\leq h}\|U_2(s)\| + hC_u \max_{0\leq s\leq h}\|X_2(s)\| \\ &\leq Ch\end{aligned}$$

für h aus einem beschränkten, von ε unabhängigem Zeitintervall $[0,h_0] \subseteq [0,T]$, wobei h_0 klein genug sei. Für die Lipschitzkonstanten

$$L_{\xi\xi} = \max_{0\leq\tau\leq h}\left\|\frac{\partial \mathcal{F}}{\partial u_0}(u(\tau),x(\tau))\right\| \text{ und } L_{\xi\eta} = \max_{0\leq\tau\leq h}\left\|\frac{\partial \mathcal{F}}{\partial x_0}(u(\tau),x(\tau))\right\|$$

gelten somit die Abschätzungen

$$L_{\xi\xi} \leq Ch, \quad L_{\xi\eta} \leq Ch.$$

Für $\|\frac{\partial \mathcal{G}}{\partial u_0}\|$ und $\|\frac{\partial \mathcal{G}}{\partial x_0}\|$ gilt

$$\left\|\frac{\partial \mathcal{G}}{\partial u_0}\right\| \leq \|X_1(h)\| \leq Ce^{hC_u}(\varepsilon+\varepsilon^2) \leq C\varepsilon$$

und

$$\begin{aligned}\left\|\frac{\partial \mathcal{G}}{\partial x_0}\right\| \leq \|X_2(h)\| &\leq (1+C\varepsilon)e^{-\frac{\alpha}{\varepsilon}h} + Ce^{hC_u}e^{-\frac{\alpha}{\varepsilon}h}(\varepsilon+\varepsilon^2+\varepsilon^3) \\ &\leq e^{-\frac{\alpha}{\varepsilon}h}(1+O(\varepsilon))\end{aligned}$$

für $h \in [0, h_0] \subseteq [0, T]$. Für die Lipschitzkonstanten von $\mathcal{G}$ erhalten wir also

$$L_{\eta\xi} \leq C\varepsilon \text{ und } L_{\eta\eta} \leq \rho_0 + O(h)$$

mit einem ρ_0 strikt kleiner als 1 für $\varepsilon < h$. Somit ist

$$L_{\xi\xi} + L_{\eta\eta} + 2\sqrt{L_{\xi\eta}L_{\eta\xi}} = Ch + C\varepsilon + 2\sqrt{C\varepsilon h} < 1,$$

also kann Theorem 2 angewendet werden. Es gibt eine Funktion $s : \mathbb{R}^{2d-m} \to \mathbb{R}^m$ mit Lipschitzkonstante $\lambda = 2O(\varepsilon)/(1 - O(h) - O(\varepsilon))$, so dass

$$\mathcal{N}_\varepsilon = \{(u_0, s(u_0)) : u \in \mathbb{R}^{2d-m}\}$$

invariant unter Ψ ist. Ψ ist die Funktion, die die analytischen Lösungen des stark gedämpften mechanischen Systems darstellt: $\Psi(u_0, x_0) = (u(h), x(h))$ mit

$$\begin{aligned} u(h) &= u_0 + \int_0^h F(u(s), x(s))\, ds \quad \text{und} \\ x(h) &= R_\varepsilon(h)x_0 + \int_0^h R_\varepsilon(h)R_\varepsilon^{-1}(s)\varphi(u(s), x(s))\, ds. \end{aligned}$$

Orbits von Ψ werden mit dem Faktor $\rho = \lambda O(h) + O(\varepsilon) = O(\varepsilon) < 1$ angezogen, das heißt für alle $(u_0, x_0) \in \mathbb{R}^{2d-m} \times \mathbb{R}^m$ gilt

$$\|x(h) - s(u(h))\| \leq \rho\|x_0 - s(u_0)\|.$$

$\mathcal{N}_\varepsilon$ ist also eine attraktive invariante Mannigfaltigleit. Für Orbits

$$\begin{aligned} (y(t), \dot{y}(t)) &= (u(nh), x(nh)) \in \mathbb{R}^{2d-m} \times \mathbb{R}^m \text{ und} \\ (y^*(t), \dot{y}^*(t)) &= (u^*(nh), s(u^*(nh))) \in \mathcal{N}^\varepsilon \end{aligned}$$

gilt nach n Schritten, also zum Zeitpunkt $t = nh$ mit Theorem 2 weiter

$$\begin{aligned} \|u(t) - u^*(t)\| &= \|u(nh) - u^*(nh)\| \leq C_1\rho^n\|x_0 - s(u_0^*)\|, \\ \|x(t) - s(u^*(t))\| &= \|x(nh) - s(u^*(nh))\| \leq C_2\rho^n\|x_0 - s(u_0^*)\|. \end{aligned}$$

Der Parameter ρ ist in der Größenordnung von $O(\varepsilon)$. Daraus folgt, dass sich ρ^n wie $O(e^{-C\frac{t}{\varepsilon}})$ verhält, wobei C eine Konstante ist, für die $\varepsilon < Ch$ gilt. Mit fortschreitender Zeit nähern sich Lösungen mit beliebigen Startwerten der Mannigfaltigkeit $\mathcal{N}_\varepsilon$ also mit exponentieller Geschwindigkeit.

(*ii*) Um die zweite Aussage zu beweisen, betrachten wir eine abgebrochene ε-Entwicklung $(y^\varepsilon(t), \dot{y}^\varepsilon(t)) \in \mathcal{M}^\varepsilon$ und zeigen, dass sich durch ein Zurückgehen in der Zeit über die differential-algebraischen Gleichungssysteme (DAE k) und einem Vorwärtsschreiten in der Zeit mit der Differentialgleichung (1.5) ein Paar $(y^*(t), \dot{y}^*(t)) \in \mathcal{N}^\varepsilon$

finden lässt, das $O(\varepsilon^N)$-nahe an $(y^\varepsilon(t), \dot{y}^\varepsilon(t))$ liegt. Mit einem Dimensionsargument folgt dann die Behauptung.

Sei also für $t \in [0,T]$ eine abgebrochene ε-Entwicklung $(y^\varepsilon(t), \dot{y}^\varepsilon(t))$ vorgegeben, deren Anfangswerte $O(\varepsilon^{N+1})$ - nahe an $\mathcal{M}^\varepsilon$ liegen. Ist $(y(t), \dot{y}(t))$ eine Lösung von (1.5), die eine ε-Entwicklung (2.8) besitzt, so ist $(y^\varepsilon(0), \dot{y}^\varepsilon(0)) - (y(0), \dot{y}(0))$ von der Größenordnung $O(\varepsilon^{N+1})$. Mit Theorem 1 folgt weiter, dass dann

$$(y^\varepsilon(t), \dot{y}^\varepsilon(t)) - (y(t), \dot{y}(t)) = O(\varepsilon^N)$$

für $t \in [0,T]$ gilt. Die abgebrochenen ε-Entwicklungen

$$\begin{aligned} y^\varepsilon &= y^0(t) + \varepsilon y^1(t) + \cdots + \varepsilon^N y^N(t), \\ \dot{y}^\varepsilon &= \dot{y}^0(t) + \varepsilon \dot{y}^1(t) + \cdots + \varepsilon^N \dot{y}^N(t) \end{aligned}$$

sind nach Theorem 1 bis auf $O(\varepsilon^N)$ eindeutig; das heißt für $t \in [0,T]$ gilt

$$(y^\varepsilon(t), \dot{y}^\varepsilon(t)) \in \mathcal{M}^\varepsilon + O(\varepsilon^N).$$

Die Koeffizienten dieser abgebrochenen ε-Entwicklung wurden für $k = 0, \ldots, N$ aus der differential-algebraischen Gleichung (DAE k) bestimmt, die sich für $t \in [0,T]$ in der Gestalt

$$\begin{aligned} \ddot{y}^k(t) &= f^k\big(y^k(t), \dot{y}^k(t), \lambda^k(t)\big), \\ 0 &= g^k\big(y^k(t), \dot{y}^k(t)\big) \end{aligned}$$

formulieren lässt, wobei wir kurz f^k und g^k für die entsprechenden Funktionen auf der rechten Seite von (DAE k) schreiben. Wollen wir für diese differential-algebraischen Gleichungen nun in der Zeit zurück gehen, so werden für ein festes $t \in [0,T]$ durch das System

$$\begin{aligned} \ddot{y}^k(t-s) &= f^k\big(y^k(t-s), \dot{y}^k(t-s), \lambda^k(t-s)\big), \\ 0 &= g^k\big(y^k(t-s), \dot{y}^k(t-s)\big) \end{aligned}$$

für $s \in [0,t]$ und konsistente Anfangswerte $y^k(t)$, $\dot{y}^k(t)$ eindeutige Lösungen $y^k(s)$, $\dot{y}^k(s)$ beschrieben. Ist etwa $s = t - t_0$ für ein $t_0 \in [0,t]$, so erhalten wir eindeutig bestimmte Lösungen $y^k(t_0)$, $\dot{y}^k(t_0)$, die von ε unabhängige Koeffizienten zu einer abgebrochenen ε-Entwicklung $(y^\varepsilon(t_0), \dot{y}^\varepsilon(t_0))$ darstellen. Daraus folgt, dass auch

$$(y^\varepsilon(t_0), \dot{y}^\varepsilon(t_0)) \in \mathcal{M}^\varepsilon + O(\varepsilon^N),$$

gilt; das heißt bezüglich der exakten Lösung $(y(t), \dot{y}(t))$ gilt

$$(y^\varepsilon(t_0), \dot{y}^\varepsilon(t_0)) - (y(t_0), \dot{y}(t_0)) = O(\varepsilon^N)$$

für $t_0 \in [0,t]$. Nun propagieren wir von $(y(t_0), \dot{y}(t_0))$ via (1.5) N Schritte der Länge h bis zum Zeitpunkt $t \in [0,T]$. Mit dem ersten Teil des Theorems folgt dann, dass es ein $(y^*(t), \dot{y}^*(t)) \in \mathcal{N}^\varepsilon$ gibt, so dass

$$(y(t), \dot{y}(t)) - (y^*(t), \dot{y}^*(t)) = O(e^{-\frac{Nh}{\varepsilon}}) = O(\varepsilon^N),$$

wobei $(y(t), \dot{y}(t))$ wieder die exakte Lösung bezeichnet. Somit gilt für $t \in [0,T]$

$$\begin{aligned} & \|(y^\varepsilon(t), \dot{y}^\varepsilon(t)) - (y^*(t), \dot{y}^*(t))\| \\ \leq\ & \|(y^\varepsilon(t), \dot{y}^\varepsilon(t)) - (y(t), \dot{y}(t))\| + \|(y(t), \dot{y}(t)) - (y^*(t), \dot{y}^*(t))\| \\ \leq\ & C\varepsilon^N \end{aligned}$$

mit einer geeigneten Konstanten C, die nicht von h und ε abhängt. Daraus folgt, dass zu fest vorgegebenem $(y^\varepsilon, \dot{y}^\varepsilon) \in \mathcal{M}^\varepsilon$

$$dist\Big((y^\varepsilon, \dot{y}^\varepsilon), \mathcal{N}_\varepsilon\Big) = O(\varepsilon^N)$$

gilt. Weil die beiden Mannigfaltigkeiten $\mathcal{M}^\varepsilon$ und $\mathcal{N}^\varepsilon$ dieselbe Dimension haben, besitzen alle $(y^*(t), \dot{y}^*(t)) \in \mathcal{N}^\varepsilon$ bis auf $O(\varepsilon^N)$ eindeutige ε-Entwicklungen. Also ist (ii) gezeigt und der Beweis abgeschlossen. ∎

Kapitel 3

Zeitintegration

3.1 Runge-Kutta-Verfahren

In diesem Abschnitt soll kurz in die Anwendung der Runge-Kutta Verfahren auf die vorliegenden Probleme eingeführt werden. Für technische Details sei auf [8] und [9] verwiesen. Eine Spezifikation der benötigten Voraussetzungen schränkt die Runge-Kutta-Verfahren auf eine bestimmte Klasse von Verfahren ein. Um die Konvergenz dieser Verfahrensklasse für die differential-algebraischen Systeme nachzuweisen, benötigen wir Resultate aus [6].

In der numerischen Mathematik sind Runge-Kutta-Verfahren zunächst zur approximativen Lösung von explizit gegebenen Ordnung-1-Differentialgleichungen

$$\dot{y} = f(t, y(t)) \tag{3.1}$$

entwickelt worden. Als Einschrittverfahren liefern sie ausgehend von einer bekannten Näherung y_n Approximationen y_{n+1} an die exakte Lösung $y(t_{n+1})$ zu einem Zeitpunkt t_{n+1} aus einem Zeitintervall $[t_0, t_{end}]$ durch

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i \dot{Y}_{ni} \tag{3.2a}$$

mit inneren Stufen

$$Y_{ni} = y_n + h \sum_{j=1}^{s} a_{ij} \dot{Y}_{nj}, \tag{3.2b}$$

die für $i = 1, \ldots, s$

$$\dot{Y}_{ni} = f(t_n + c_i h, Y_{ni}) \tag{3.2c}$$

erfüllen müssen. Die Koeffizienten a_{ij}, b_i, c_i bestimmen das Verfahren. Oft werden sie zur Darstellung des Verfahrens in einem *Butcher-Tableau* zusammengefasst. Die natürliche Zahl s bezeichnet die Anzahl der Stufen; h sei von nun an generell die Schrittweite des Runge-Kutta-Verfahrens. Falls $a_{ij} = 0$ für $i \leq j$ ist, sprechen wir von einem *expliziten*, sonst von einem *impliziten* Verfahren.

Um Runge-Kutta-Verfahren (3.2) auf gewöhnliche Differentialgleichungen und differential-algebraische Gleichungssysteme anwenden zu können, wollen wir statt des explizit gegebenen, nichtautonomen Systems (3.1) im Fall der Ordnung 1 autonome, implizit formulierte Systeme

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

betrachten. In der Formulierung des Runge-Kutta-Verfahrens führt dies statt der Bedingung (3.2c) auf

$$F(Y_{ni}, \dot{Y}_{ni}) = 0. \tag{3.2d}$$

Der Vorteil dieser Schreibweise liegt darin, dass das Verfahren in der Formulierung (3.2a), (3.2b), (3.2d) auch auf Systeme wie (DAE k') anwendbar ist.

Viele Differentialgleichungen treten in zweiter Ordnung auf. Zwar lässt sich eine Differentialgleichung beliebiger Ordnung immer als System der Ordnung 1 formulieren, dennoch wollen wir für den Ordnung-2-Fall ein Runge-Kutta-Verfahren angeben. Wir betrachten dazu ebenfalls die implizite Darstellung

$$F(y, \dot{y}, \ddot{y}) = 0.$$

Das stark gedämpfte mechanische System (1.5) oder auch die differential-algebraischen Systeme (DAE k) genügen dieser Darstellung. Ein Runge-Kutta-Verfahren angewandt auf obige Gleichung liefert Approximationen $(y_{n+1}, \dot{y}_{n+1})$ an die exakte Lösung $(y(t_{n+1}), \dot{y}(t_{n+1}))$ zu Zeitpunkten $t_{n+1} \in [t_0, t_{end}]$ via

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i \dot{Y}_{ni}, \quad \dot{y}_{n+1} = \dot{y}_n + h \sum_{i=1}^{s} b_i \ddot{Y}_{ni} \tag{3.3a}$$

mit inneren Stufen

$$Y_{ni} = y_n + h \sum_{j=1}^{s} a_{ij} \dot{Y}_{nj}, \quad \dot{Y}_{ni} = h \sum_{j=1}^{s} a_{ij} \ddot{Y}_{nj} \tag{3.3b}$$

die für $i = 1, \ldots, s$

$$F(Y_{ni}, \dot{Y}_{ni}, \ddot{Y}_{ni}) = 0 \tag{3.3c}$$

erfüllen müssen.

Die Bewegungsgleichung (1.5) ist wegen den starken Dämpfungskräften eine steife Differentialgleichung. Aufgrund ihres Stabilitätsverhaltens sind also implizite Verfahren expliziten Methoden vorzuziehen. Die Konstruktion impliziter Runge-Kutta-Verfahren stützt sich wesentlich auf die folgenden Bedingungen:

$$\sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k} \text{ für } k = 1, \ldots, p \text{ und } i = 1, \ldots, s. \tag{3.4}$$

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \text{ für } k = 1, \ldots, q \text{ und } i = 1, \ldots, s. \tag{3.5}$$

$$\sum_{i=1}^{s} b_i a_{ij} c_i^{k-1} = \frac{b_j}{k}(1 - c_j^k) \text{ für } k = 1, \ldots, r \text{ und } j = 1, \ldots, s. \tag{3.6}$$

Die Bedingung (3.4) bedeutet, dass das Runge-Kutta-Verfahren die *Ordnung* p hat, (3.5) definiert die sogenannte *Stufenordnung* q des Verfahrens.

Aufgrund der speziellen Problemstruktur werden sich weitere Voraussetzungen als notwendig erweisen. Hierzu ist es nötig, die *Stabilitätsfunktion* R eines Runge-Kutta-Verfahrens zu betrachten. Wir erhalten sie aus der Anwendung des Verfahrens auf die lineare Testgleichung $\dot{y} = \lambda y$ mit $z = h\lambda$, Re $\lambda < 0$, als

$$R(z) = 1 + zb^T(I - z\mathcal{A})^{-1}\mathbb{1}$$

mit $b^T = (b_1, \ldots, b_s)$, $\mathcal{A} = (a_{ij})_{i,i=1}^{s}$ und dem s-Tupel $\mathbb{1} = (1, \ldots, 1)^T$ (siehe [9]). Somit können wir zwei weitere Bedingungen formulieren, die sich als ganz zentral erweisen werden:

$$\mathcal{A} \text{ ist invertierbar und es gilt } |R(\infty)| = |1 - b^T\mathcal{A}^{-1}\mathbb{1}| < 1. \tag{3.7}$$

$$\begin{aligned}&\mathcal{A} \text{ hat keine Eigenwerte auf der negativen reellen Halbachse,}\\ &\text{und für alle reellen } \omega > 0 \text{ gilt } |R(-\omega)| < 1.\end{aligned} \tag{3.8}$$

Generell wird vorausgesetzt, dass die Ordnung p des Verfahrens, wenn es auf nichtsteife Probleme angewandt wird, größer als die Stufenordnung q ist, das heißt es gelte $p \geq q + 1$. Desweiteren sei durchweg $q \geq 1$ vorausgesetzt.

Die wichtigste Klasse von numerischen Verfahren, die diese Bedingungen erfüllen, sind RadauIIA-Verfahren. Sie basieren auf Radau-Quadraturformeln. Für Details zur genauen Konstruktionsweise sei an dieser Stelle auf [9], S. 72 ff. verwiesen. In den Tabellen 3.1 und 3.2 sind die Koeffizienten der RadauIIA-Verfahren für die Ordnungen 3 und 5 angegeben.

Konvergenzresultate für differential-algebraische Systeme vom Index 2

Nun sollen Konvergenzresultate für differential-algebraische Systeme vom Index 2, wie sie in [6] hergeleitet und bewiesen wurden, aufgeführt werden, denn im folgenden Abschnitt wird auf diese Resultate mehrmals zurückgegriffen.

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

Abbildung 3.1: RadauIIA-Verfahren der Ordnung 3

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

Abbildung 3.2: RadauIIA-Verfahren der Ordnung 5

Wir formulieren das differential-algebraische Gleichungssystem allgemein als

$$\begin{aligned} \dot{y} &= f(y,z), \\ 0 &= g(y), \end{aligned} \tag{3.9}$$

wobei f und g genügend oft differenzierbar seien. Ist

$$\|g_y(y) f_z(y,z))\| \leq const. \tag{3.10}$$

in einer Umgebung der exakten Lösung, so ist (3.9) ein System vom Störungsindex 2. Weiterhin geben wir konsistente Anfangswerte y_0, z_0 für (3.9) vor; es muss also

$$g(y_0) = 0, \qquad g_y(y_0) f(y_0, z_0) = 0 \tag{3.11}$$

gelten. Auf das System (3.9) lässt sich das Runge-Kutta-Verfahren (3.2) anwenden. Wir erhalten Approximationen y_{n+1} und z_{n+1} an die exakte Lösung durch (3.2a) mit entsprechenden inneren Stufen Y_{ni} und Z_{ni} via (3.2b). Für diese inneren Stufen muss (3.2d) gelten, also

$$\begin{aligned} \dot{Y}_{ni} &= f(Y_{ni}, Z_{ni}), \\ 0 &= g(Y_{ni}). \end{aligned}$$

Beim Studium der Existenz und Eindeutigkeit von Runge-Kutta-Lösungen für (3.9) genügt es, Anfangswerte zu betrachten, die die Konsistenzbedingungen (3.11) mit

Störungen der Größenordnung $O(h^2)$ beziehungsweise $O(h)$ erfüllen, denn exakte Lösungen werden durch y_{n+1}, z_{n+1} lediglich angenähert, so dass auch (3.9) nur bis auf leichten Störungen erfüllt ist.

Theorem 4 (Existenz und lokale Eindeutigkeit der Runge-Kutta-Lösung). *Falls für die Anfangswerte* (y_0, z_0)

$$g(y_0) = O(h^2), \quad g_y(y_0)f(y_0, z_0) = O(h)$$

gilt, die Bedingung (3.10) in einer h-unabhängigen Umgebung von (y_0, z_0) *erfüllt ist und die Runge-Kutta-Matrix* $\mathcal{A}$ *invertierbar ist, so besitzt das nichtlineare Gleichungssystem*

$$\begin{aligned} Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_j, Z_j), \\ 0 &= g(Y_i) \end{aligned} \tag{3.12}$$

für $h \leq h_0$ *eine Lösung. Diese ist lokal eindeutig und erfüllt*

$$Y_i - y_0 = O(h), \qquad Z_i - z_0 = O(h).$$

Beweis: [6] S. 31-32.

Um letztlich globale Fehlerschranken für die approximativen Lösungen ableiten zu können ist es unumgänglich, den Einfluss von Störungen auf die Lösungen des Systems (3.12) zu untersuchen. Von großem Nutzen ist, dass die Störungen des folgenden Resultats auf verschiedene Arten interpretierbar sind, beispielsweise als Rundungsfehler oder Fehler in den iterativen Lösungen des nichtlinearen Gleichungssystems. In den Konvergenzresultaten für (DAE k) werden wir entscheidend ausnutzen, dass Störungen auch als Konsistenzdefekt interpretiert werden können, der entsteht, wenn die exakte Lösung in das Verfahren eingesetzt wird.

Zu bemerken ist weiterhin, dass die Abschätzungen des folgenden Theorems das diskrete Anaolgon zu Abschätzungen sind, wie sie in der Berechnung des Störungsindex von (3.9) auftreten (vergleiche mit Beweis der Bemerkung auf Seite 27). Dem Auftreten einer Zeitableitung im kontinuierlichen Fall entspricht hier die Division durch h.

Theorem 5 (Einfluss von Störungen). *Seien* Y_i, Z_i *wie in (3.12) und gestörte Werte* $\widehat{Y}_i$, $\widehat{Z}_i$ *vorgegeben, die dem System*

$$\begin{aligned} \widehat{Y}_i &= \widehat{y}_0 + h\sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j, \widehat{Z}_j) + h\delta_i, \\ 0 &= g(\widehat{Y}_i) + \theta_i \end{aligned}$$

genügen. Unter den Voraussetzungen von Theorem 4 und falls

$$\widehat{y}_0 - y_0 = O(h^2), \quad \delta_i = O(h), \quad \theta_i = O(h^2)$$

gilt, folgen für $h \le h_0$ die Abschätzungen

$$\begin{aligned} \|\widehat{Y}_i - Y_i\| &\le C\Big(\|\widehat{y}_0 - y_0\| + h\|\delta\| + \|\theta\|\Big), \\ \|\widehat{Z}_i - Z_i\| &\le \frac{C}{h}\Big(\|g_y(y_0)(\widehat{y}_0 - y_0)\| + h\|\widehat{y}_0 - y_0\| + h\|\delta\| + \|\theta\|\Big). \end{aligned}$$

Hierbei ist $\delta = (\delta_1, \ldots, \delta_s)$ und $\theta = (\theta_1, \ldots, \theta_s)$.

Beweis: [6] S. 33-34.

Nun muss nur noch der lokale Fehler abgeschätzt werden, um die Konvergenzresultate abzuleiten. Dazu betrachten wir wieder ungestörte Angangswerte $y_0 = y(t_0)$, $z_0 = z(t_0)$ und wenden einen Schritt des Runge-Kutta-Verfahrens (3.2) auf (3.9) an. Das folgende Lemma liefert Schranken für diesen lokalen Fehler.

Lemma 3 (Lokaler Fehler). *Falls das Runge-Kutta-Verfahren (3.2) für $p \ge q+1$ und $q \ge 1$ die Bedingungen (3.4) und (3.5) erfüllt, so gilt für den lokalen Fehler*

$$\begin{aligned} &y_1 - y(t_0 + h) = O(h^{q+1}), \;\; P(t_0)(y_1 - y(t_0 + h)) = O(h^{q+2}), \\ &z_1 - z(t_0 + h) = O(h^q). \end{aligned}$$

Falls $p = q$ ist, gelten in der y-Komponente die Abschätzungen

$$y_1 - y(t_0 + h) = O(h^{q+1}), \quad P(t_0)(y_1 - y(t_0 + h)) = O(h^{q+1}).$$

P ist hier die durch

$$P(t) = I - (f_z(g_y f_z)^{-1} g_y)(y(t), z(t))$$

gegebene Projektion.

Beweis: [6] S. 35.

Nun stehen alle Mittel bereit, um globale Fehlerschranken anzugeben.

Theorem 6 (Kovergenz in der y-Komponente). *Sei die Bedingung (3.10) in einer Umgebung der Lösung $(y(t), z(t))$ von (3.9) erfüllt und seien die Anfangswerte konsistent. Unter der Voraussetzung (3.7) und falls der lokale Fehler*

$$y_1 - y(t_0 + h) = O(h^r), \quad P(t_0)(y_1 - y(t_0 + h)) = O(h^{r+1}) \tag{3.13}$$

mit oben definiertem P erfüllt, konvergieren die Lösungen y_n des Runge-Kutta-Verfahrens (3.2) mit Ordnung r, für $t_n = t_0 + nh \in [t_0, t_{end}]$ gilt also

$$y_n - y(t_n) = O(h^r).$$

Falls $y_1 - y(t_0 + h) = O(h^{r+1})$ ist, gilt insbesondere $g(y_n) = O(h^{r+1})$.

Beweis: [6] S. 36-39.

Bemerkung. (1) Mit Lemma 3 folgt, dass der globale Fehler in der y-Komponente mindestens Ordnung $q + 1$ besitzt. Für viele wichtige Verfahren kann die Voraussetzung (3.13) noch verschärft werden, vergleiche Theorem 5.9 in [6].

(2) Falls $a_{si} = b_i$ für $i = 1, \ldots, s$ gilt, erhalten wir sogar $g(y_n) = g(Y_{ns}) = 0$.

Theorem 7 (Kovergenz in der z-Komponente). *Sei die Bedingung (3.10) in einer Umgebung der Lösung $(y(t), z(t))$ von (3.9) erfüllt und seien die Anfangswerte konsistent. Weiter sei die Voraussetzung (3.7) angenommen, der globale Fehler der y-Komponente von der Größenordnung $O(h^k)$, sowie $g(y_n) = O(h^{k+1})$. Der lokale Fehler in der z-Komponente sei*

$$z_1 - z(t_0 + h) = O(h^k).$$

Dann konvergieren die Lösungen z_n des Runge-Kutta-Verfahrens (3.2) mit Ordnung k, für $t_n = t_0 + nh \in [t_0, t_{end}]$ gilt also

$$z_n - z(t_n) = O(h^k).$$

Beweis: [6] S. 40.

Bemerkung. Falls die Bedingungen (3.4) und (3.5) mit $p \geq q$ erfüllt sind, gilt als Konsequenz von Lemma 3 und Theorem 6 also $z_n - z(t_n) = O(h^k)$.

3.2 Fehler des Runge-Kutta-Verfahrens für die differential-algebraischen Systeme

Theorem 8 (Fehlerabschätzung für das differential-algebraische System vom Index 2). *Seien die Bedingungen (3.4), (3.5) und (3.7) für das Runge-Kutta-Verfahren (3.2) erfüllt und seien konsistente Anfangswerte y_0^0, $\dot{y}_0^0$, λ_0^0 zu (DAE 0)*

gegeben. Wenden wir das Verfahren auf (DAE 0) an, so ergeben sich für den globalen Fehler die Schranken

$$y_n^0 - y^0(t_n) = O(h^{q+1}),\ \dot{y}_n^0 - \dot{y}^0(t_n) = O(h^{q+1}),\ \lambda_n^0 - \lambda^0(t_n) = O(h^q) \qquad (3.14)$$

im Sinne gleichmäßiger Konvergenz für $0 \le t_n \le T$.

Beweis. Mit

$$\begin{aligned} \check{f}(y^0, \dot{y}^0, \lambda^0) &= \begin{pmatrix} \dot{y}^0 \\ M^{-1}(y^0)(f^0(y^0, \dot{y}^0) - S(y_0)G^T\lambda^0) \end{pmatrix}, \\ \check{g}(y^0, \dot{y}^0) &= GS^T(y_0)\dot{y}^0 \end{aligned}$$

ist (DAE 0) mit den Variablen $\eta = (y^0, \dot{y}^0)$ beziehungsweise $\xi = \lambda^0$ ein System der Form

$$\begin{aligned} \dot{\eta} &= \check{f}(\eta, \xi), \\ 0 &= \check{g}(\eta). \end{aligned}$$

Für diese differential-algebraischen Gleichungssysteme vom Index 2 ist

$$\begin{aligned} \|(\check{g}_\eta \check{f}_\xi)^{-1}\| &= \left\| \left([\frac{\partial}{\partial y^0}(GS^T(y_0)\dot{y}^0)), GS^T(y^0)] \cdot \begin{bmatrix} 0 \\ M^{-1}(y^0)S(y^0)G^T \end{bmatrix} \right)^{-1} \right\| \\ &= \| \Big(-G(S^T M^{-1} S)(y^0)G^T \Big)^{-1} \| = \left\| \left(-[0\ I_m]\widetilde{M}^{-1}(y^0) \begin{bmatrix} 0 \\ I_m \end{bmatrix} \right)^{-1} \right\| \\ &= \|\widetilde{\widetilde{M}}_{22}(y^0)\| \le const, \end{aligned}$$

wobei die Massenmatrix in

$$\widetilde{M}^{-1} = \begin{pmatrix} \widetilde{\widetilde{M}}_{11} & \widetilde{\widetilde{M}}_{12} \\ \widetilde{\widetilde{M}}_{21} & \widetilde{\widetilde{M}}_{22} \end{pmatrix}$$

aufgespalten wurde. Der Block $\|\widetilde{\widetilde{M}}_{22}(y^0)\|$ ist beschränkt, da M nach Voraussetzung und S wie in Lemma 1 gesehen beliebig oft differenzierbar und von ε unabhängig sind.

Mit dem Nachweis dieser Grundvoraussetzung sind die Resultate aus Kapitel 4 in [6] anwendbar. Aus Lemma 3 erhalten wir die lokalen Fehler

$$y_1^0 - y^0(t_1) = O(h^{q+1}),\ P(t)(y_1^0 - y^0(t_1)) = O(h^{q+2}),\ \lambda_1^0 - \lambda^0(t_1) = O(h^q),$$

aus den Theoremen 6 und 7 ergeben sich die gewünschten globalen Fehler. ∎

Bemerkung. Falls $b_i = a_{si}$ für alle $i = 1, \ldots, s$ sowie die Bedingungen (3.4) - (3.6) mit $p \leq 2q$ und $p \leq r + q + 1$ erfüllt sind, gilt für die lokalen Fehler

$$y_1^0 - y^0(t_1) = O(h^{p+1}), \quad \dot{y}_1^0 - \dot{y}^0(t_1) = O(h^{p+1})$$

und das Resultat für den globalen Fehler verbessert sich entsprechend, das heißt wir erhalten durch Anwendung von Theorem 6

$$y_n^0 - y^0(t_n) = O(h^p), \quad \dot{y}_n^0 - \dot{y}^0(t_n) = O(h^p).$$

Theorem 9 (Fehlerabschätzung für das differential-algebraische System vom Index 2+k). *Das Runge-Kutta-Verfahren (3.2) erfülle (3.7) und besitze die Stufenordnung $q \geq 1$. Seien weiter konsistente Anfangswerte y_0^0, $\dot{y}_0^0$, λ_0^0 für das differential-algebraische System (DAE 0, ... DAE k) gegeben. Wenden wir das Runge-Kutta-Verfahren auf (DAE 0,...,DAE k) an, so ergibt sich für den globalen Fehler für $k \leq q$*

$$\begin{aligned} &y_n^k - y^k(t_n) = O(h^{q+1-k}), \quad \dot{y}_n^k - \dot{y}^k(t_n) = O(h^{q+1-k}), \\ &\lambda_n^k - \lambda^k(t_n) = O(h^{q-k}) \end{aligned} \tag{3.15}$$

im Sinne gleichmäßiger Konvergenz für $0 \leq t_n \leq T$.

Beweis. Um den Beweis zu führen, betrachten wir den Fall $k = 1$. Der Nachweis für allgemeines k kann durch ein Induktionsargument erbracht werden. Die Vorgehensweise erfolgt dabei analog zum vorgestellten Fall.

Der Fall $k = 0$ wurde bereits in Theorem 8 bewiesen. Um das Resultat für das System differential-algebraischer Gleichungen (DAE 0, DAE 1) zu beweisen, werden wiederum die Resultate aus Kapitel 4 von [6] verwendet. Der Beweis ist in zwei wesentliche Schritte unterteilt. Zunächst leiten wir mit Theorem 6 Schranken für $\|(y^1(c_i h), \dot{y}^1(c_i h)^T - (Y_i^1, \dot{Y}_i^1))^T\|$ und $\|\lambda^1(c_i h) - \Lambda_i^1\|$ her, wobei die exakten Lösungen $(y^1(c_i h), \dot{y}^1(c_i h))^T$ und $\lambda^1(c_i h)$ in der Rolle von $\widehat{Y}_i$ beziehungsweise $\widehat{\dot{Y}}_i$ und die Runge-Kutta-Approximationen $(Y_i^1, \dot{Y}_i^1)^T$ und Λ_i^1 in der Rolle von Y_i beziehungsweise $\dot{Y}_i$ sind. Sind die Defekte δ_i und θ_i sowie die Fehler in den Anfangswerten nachgerechnet, so lässt sich Theorem 6 anwenden und wir erhalten die gewünschten Schranken. Die Berechnung der lokalen Fehler erfolgt mit den gleichen Techniken wie im Beweis von Lemma 3. Im zweiten Beweisschritt leiten wir mit Theorem 6 und Theorem 7 Abschätzungen für die globalen Fehler $(y_n^1, \dot{y}_n^1) - (y^1(t_n), \dot{y}^1(t_n))$ beziehungsweise $\lambda_n^1 - \lambda^1(t_n)$ durch Nachweisen der entsprechenden Voraussetzungen der Theoreme her.

(a) Beim Studium der lokalen Fehler ist es angesichts der Notation angenehm, zunächst nur einen Schritt des Runge-Kutta-Verfahrens zu betrachten, das heißt

im Hinblick auf (3.2) setzen wir $n = 1$. Um die lokalen Fehler bestimmen zu können, sind zunächst die Defekte δ_i und θ_i zu berechnen. Hierzu betrachten wir einerseits (DAE 1), die sich äquivalent umformulieren lässt zu

$$\begin{aligned} \begin{pmatrix} \dot{y}^1(t) \\ \ddot{y}^1(t) \end{pmatrix} &= F^1(t, y^1(t), \dot{y}^1(t), \lambda^1(t)), \\ 0 &= g^1(y^1(t), \dot{y}^1(t)), \end{aligned} \tag{3.16}$$

wobei $F^1 = M^{-1}(\Phi^1 - SG^T\lambda^1))$ und $g^1 = GS^T\dot{y}^1 + GS^TH^1 - \lambda^0$ seien. Die exakten Lösungen $y^0(t), \dot{y}^0(t), \ddot{y}^0(t)$ sind aus (DAE 0) zu jedem Zeitpunkt $t \in [0, T]$ bekannt. Somit sind die analytischen Lösungen $y^1(t), \dot{y}^1(t)$ zu gegebenen Anfangswerten wohlbestimmt. Für die Approximationen der Lösungen von (DAE 1) durch das Runge-Kutta-Verfahren (3.2) gilt allerdings

$$\begin{pmatrix} y_1^1 \\ \dot{y}_1^1 \end{pmatrix} = \begin{pmatrix} y_0^1 \\ \dot{y}_0^1 \end{pmatrix} + h \sum_{i=1}^{s} b_i \begin{pmatrix} \dot{Y}_i^1 \\ \ddot{Y}_i^1 \end{pmatrix}, \tag{3.17a}$$

$$\begin{pmatrix} Y_i^1 \\ \dot{Y}_i^1 \end{pmatrix} = \begin{pmatrix} y_0^1 \\ \dot{y}_0^1 \end{pmatrix} + h \sum_{i=1}^{s} a_{ij} \begin{pmatrix} \dot{Y}_j^1 \\ \ddot{Y}_j^1 \end{pmatrix}. \tag{3.17b}$$

Die inneren Stufen müssen

$$\begin{aligned} M(Y_i^0)\ddot{Y}_i^1 &= \Phi^1(Y_i^0, \dot{Y}_i^0, \ddot{Y}_i^0, Y_i^1, \dot{Y}_i^1) - S(Y_i^0)G^T\Lambda_i^1) \\ &=: \widetilde{F}^1(Y_i^0, \dot{Y}_i^0, \ddot{Y}_i^0, Y_i^1, \dot{Y}_i^1, \Lambda_i^1), \\ 0 &= GS^T(Y_i^0)\dot{Y}_i^1 + GS^T(Y_i^0)H^1(Y_i^0, \dot{Y}_i^0, \dot{Y}_i^1) - \Lambda_i^0 \end{aligned} \tag{3.17c}$$

genügen. Für die eingeführten Funktionen F^1 und $\widetilde{F}^1$ ist insbesondere die Beziehung

$$F^1\big(t, y^1(t), \dot{y}^1(t), \lambda^1(t)\big) = \widetilde{F}^1\big(y^0(t), \dot{y}^0(t), \ddot{y}^0(t), y^1(t), \dot{y}^1(t), \lambda^1(t)\big) \tag{3.18}$$

erfüllt.

Um die Defekte zu berechnen, benötigen wir für die Runge-Kutta-Approximationen $Y_i^0,\ \dot{Y}_i^0,\ \Lambda_i^0$ aus Lemma 3 die Fehlerabschätzungen

$$y^0(c_ih) - Y_i^0 = O(h^{q+1}),\ \dot{y}^0(c_ih) - \dot{Y}_i^0 = O(h^{q+1}),\ \lambda(c_ih) - \Lambda_i^0 = O(h^q). \tag{3.19}$$

Um $\ddot{y}^0(c_ih) - \ddot{Y}_i^0$ abzuschätzen, betrachten wir die Bedingung (3.2c) für (DAE 0), also

$$\begin{aligned} M(Y_i^0)\ddot{Y}_i^0 &= f(Y_i^0, \dot{Y}_i^0) - S(Y_i^0)G^T\Lambda_i^0, \\ 0 &= GS^T(Y_i^0)\dot{Y}_i^0 \end{aligned} \tag{3.20}$$

Da $\ddot{y}^0(c_ih)$ die exakte Lösung von (DAE 0) zum Zeitpunkt $t = c_ih$ ist, gilt insbesondere

$$M(y^0(c_ih))\ddot{y}^0(c_ih) = f(y^0(c_ih), \dot{y}^0(c_ih)) - S(y^0(c_ih))G^T\lambda^0(c_ih).$$

Entwickeln wir die glatten beschränkten Funktionen in dieser Gleichung um Y_i^0 beziehungsweise $(Y_i^0, \dot{Y}_i^0)$, so ergibt sich aus der Subtraktion mit der ersten Gleichung von (3.20) und mit (3.19)

$$\ddot{y}^0(c_ih) - \ddot{Y}_i^0 = O(h^q).$$

Durch geeignete Interpretation der exakten beziehungsweise der Runge-Kutta-Lösungen lassen sich über die beiden Systeme (3.16) und (3.17) die Defekte berechnen. Wie bei Theorem 5 schon angesprochen, fassen wir die exakte Lösung $(y^1, \dot{y}^1)$ als gestörte Runge-Kutta-Approximation auf. Also ist zunächst der Defekt zu berechnen, der sich ergibt, wenn $(y^1(c_ih), \dot{y}^1(c_ih))$ in (3.17) eingesetzt wird. Wir haben daher

$$\begin{pmatrix} y^1(c_ih) \\ \dot{y}^1(c_ih) \end{pmatrix} = \begin{pmatrix} y^1(0) \\ \dot{y}^1(0) \end{pmatrix} + \sum_{i=1}^{s} a_{ij} \begin{pmatrix} \dot{y}^1(c_jh) \\ \ddot{y}^1(c_jh) \end{pmatrix} + h\widehat{\delta}_i. \tag{3.21}$$

Setzen wir

$$\begin{pmatrix} y^1(c_ih) \\ \dot{y}^1(c_ih) \end{pmatrix} = \begin{pmatrix} y^1(0) \\ \dot{y}^1(0) \end{pmatrix} + c_ih \begin{pmatrix} \dot{y}^1(0) \\ \ddot{y}^1(0) \end{pmatrix} + \cdots + \frac{c_i^{q+1}h^{q+1}}{(q+1)!} \begin{pmatrix} y^{1^{(q+1)}}(0) \\ \dot{y}^{1^{(q+1)}}(0) \end{pmatrix} + O(h^{q+2})$$

und

$$\begin{pmatrix} \dot{y}^1(c_ih) \\ \ddot{y}^1(c_ih) \end{pmatrix} = \begin{pmatrix} \dot{y}^1(0) \\ \ddot{y}^1(0) \end{pmatrix} + c_ih \begin{pmatrix} \ddot{y}^1(0) \\ \dddot{y}^1(0) \end{pmatrix} + \cdots + \frac{c_i^q h^q}{q!} \begin{pmatrix} y^{1^{(q+1)}}(0) \\ \dot{y}^{1^{(q+1)}}(0) \end{pmatrix} + O(h^{q+1})$$

in (3.21) ein und fassen die Terme mit Ableitungen gleicher Ordnung zusammen, so ergibt sich mit der Voraussetzung für die Stufenordnug des Verfahrens für den Defekt

$$h\widehat{\delta}_i = \frac{1}{q!}c_i^q h^{q+1} \begin{pmatrix} y^{1^{(q+1)}}(0) \\ \dot{y}^{1^{(q+1)}}(0) \end{pmatrix} \left(\frac{c_i^{q+1}}{q+1} - \sum_{i=1}^{s} a_{ij}c_j^q \right) + O(h^{q+2}).$$

Der Defekt in (3.17b) ist also nicht größer als $O(h^q)$ und der Defekt in (3.17a) und (3.17b) ist von der Größenordnung $O(h^q)$, wenn die exakte Lösung eingesetzt wird.

Kehren wir also wieder zu (3.21) zurück, wo wir die exakten Lösungen in (3.17b) eingesetzt hatten. Gleichwertig zu (3.21) ist

$$\begin{aligned} \begin{pmatrix} y^1(c_ih) \\ \dot{y}^1(c_ih) \end{pmatrix} &= \begin{pmatrix} y^1(0) \\ \dot{y}^1(0) \end{pmatrix} + \sum_{i=1}^{s} a_{ij}\widetilde{F}^1(Y_j^0, \dot{Y}_j^0, \ddot{Y}_j^0, y^1(c_jh), \dot{y}^1(c_jh), \lambda^1(c_jh)) + h\widehat{\delta}_i \\ &= \begin{pmatrix} y^1(0) \\ \dot{y}^1(0) \end{pmatrix} + \sum_{i=1}^{s} a_{ij}F^1(c_jh, y^1(c_jh), \dot{y}^1(c_jh), \lambda^1(c_jh)) + h\widetilde{\delta}_i + h\widehat{\delta}_i, \end{aligned}$$

wobei $\widetilde{\delta}_i$ durch

$$\begin{aligned}\widetilde{\delta}_i &= \sum_{i=1}^{s} a_{ij}\Big(\widetilde{F}^1\big(Y_j^0, \dot{Y}_j^0, \ddot{Y}_j^0, y^1(c_jh), \dot{y}^1(c_jh), \lambda^1(c_jh)\big) \\ &\qquad -F^1\big(c_jh, y^1(c_jh), \dot{y}^1(c_jh), \lambda^1(c_jh)\big)\Big)\end{aligned}$$

gegeben ist. Indem wir die Gleichheit (3.18) benutzen, lässt sich obige Differenz durch Linearisierung der Funktionen detailliert untersuchen. Mit elementarer Rechenarbeit ergibt sich ohne weitere Schwierigkeiten

$$\widetilde{\delta}_i = O\big(\|y^0(c_jh) - Y_j^0\| + \|\dot{y}^0(c_jh) - \dot{Y}_j^0\| + \|\ddot{y}^0(c_ih) - \ddot{Y}_i^0\|\big) = O(h^q).$$

Der Gesamtdefekt ist durch die Summe der berechneten Teildefekte gegeben,

$$h\delta_i = h\widehat{\delta}_i + h\widetilde{\delta}_i,$$

es folgt also

$$\delta_i = O(h^q).$$

Setzen wir $y^1(c_ih)$, $\dot{y}^1(c_ih)$ in die zweite Gleichung von (3.17c) ein, so ergibt sich mit (3.19) der Defekt $\theta_i = O(h^q)$, denn unter Ausnutzung der Zwangsbedingung aus (DAE 1) zum Zeitpunkt $t = c_ih$ gilt

$$\begin{aligned}& GS^T(Y_i^0)\dot{y}^1(c_ih) + GS^T(Y_i^0)H^1(Y_i^0, \dot{Y}_i^0, y^1(c_ih)) - \Lambda_i^0 \\ = \;& \big(GS^T(Y_i^0) - GS^T(y^0(c_ih))\big)\dot{y}^1(c_ih) \\ & + GS^T(Y_i^0)H^1(Y_i^0, \dot{Y}_i^0, y^1(c_ih)) - GS^T(y^0(c_ih))H^1(y^0(c_ih), \dot{y}^0(c_ih), y^1(c_ih)) \\ & - (\Lambda_i^0 - \lambda^0(c_ih)) \\ & + GS^T(y^0(c_ih))\dot{y}^1(c_ih) + GS^T(y^0(c_ih))H^1(y^0(c_ih), \dot{y}^0(c_ih), y^1(c_ih)) - \lambda^0(c_ih) \\ = \;& O\big(\|y^0(c_ih) - Y_i^0\| + \|\dot{y}^0(c_ih) - \dot{Y}_i^0\| + \|\lambda^0(c_ih) - \Lambda_i^0\|\big).\end{aligned}$$

Als Voraussetzung zur Anwendung des Runge-Kutta-Verfahrens sind konsistente Anfangswerte zu wählen. Das heißt die Konsistenzbedingungen auf (3.11) sind für Startwerte $y^1(t_0) = y_0^1$, $\dot{y}^1(t_0) = \dot{y}_0^1$ und $\lambda^1(t_0) = \lambda_0^1$ für $t_0 \in [0, T]$ zu erfüllen, also

$$GS^T(y_0^0)\dot{y}_0^1 + GS^T(y_0^0)H^1(y_0^0, \dot{y}_0^0, y_0^1) - \lambda_0^0 = 0$$

und

$$GS^T(y_0^0)\Big(M^{-1}(y_0^0)(\Phi^1(y_0^0, \dot{y}_0^0, \ddot{y}_0^0, y_0^1, \dot{y}_0^1) - S(y_0^0)G^T\lambda_0^1)\Big) = 0.$$

Wegen (3.16) gilt dies aber exakt. Die Startwerte weisen also keine Defekte auf, so dass

$$y_0^1 - y^1(t_0) = 0, \quad \dot{y}_0^1 - \dot{y}^1(t_0) = 0.$$

Somit lässt sich Theorem 6 anwenden und wir erhalten

$$\begin{aligned} \left\| \begin{pmatrix} y^1(c_i h) \\ \dot{y}^1(c_i h) \end{pmatrix} - \begin{pmatrix} Y_i^1 \\ \dot{Y}_i^1 \end{pmatrix} \right\| &\leq C(h\|\delta\| + \|\theta\|) = O(h^q), \\ \|\lambda^1(c_i h) - \Lambda_i^1\| &\leq \frac{C}{h}(h\|\delta\| + \|\theta\|) = O(h^{q-1}) \end{aligned} \tag{3.22}$$

mit Konstanten C unabhängig von h. Die Größen δ und θ sind durch $\delta = (\delta_1, \ldots, \delta_s)^T$ beziehungsweise $\theta = (\theta_1, \ldots, \theta_s)^T$ gegeben.

Für Schritte $n \geq 1$ erfolgt die Abschätzung der lokalen Fehler analog zur obigen Vorgehensweise, wobei benötigt wird, dass die Fehler der inneren Stufen wie in (3.19) gegeben sind, also durch

$$\begin{aligned} &Y_{ni}^0 - y^0(t_n + c_i h) = O(h^{q+1}), \quad \dot{Y}_{ni}^0 - \dot{y}^0(t_n + c_i h) = O(h^{q+1}), \\ &\Lambda_{ni}^0 - \lambda^0(t_n + c_i h) = O(h^q). \end{aligned} \tag{3.23}$$

Die Argumentation hierfür erfolgt wieder mit Techniken aus [6]. Wie im Beweis von Theorem 6 betrachten wir zur Untersuchung der Fehlerfortpflanzung zwei Runge-Kutta-Lösungen $(\widehat{y}, \widehat{\dot{y}})$ und $(\widetilde{y}, \widetilde{\dot{y}})$, wobei wir letztere als exakte Lösungen der differential-algebraischen Gleichung (DAE 0) zum Zeitpunkt $t = nh$ interpretieren. Dabei sei $(\widehat{y}, \widehat{\dot{y}})$ in der Rolle des hier angewandten Runge-Kutta-Verfahrens. Hierfür lässt sich wie im Beweisteil c) und d) von Theorem 6 zeigen, dass

$$\left\| \begin{pmatrix} \Delta y_n^0 \\ \Delta \dot{y}_n^0 \end{pmatrix} \right\| \leq C_1 h^{q+1}, \quad \|GS^T(y_n^0)\dot{y}_n^0\| \leq C_2 h^{q+1} \tag{3.24}$$

vorausgesetzt werden kann, wobei $(\Delta y_n^0, \Delta \dot{y}_n^0)^T = (y_n^0 - y^0(nh), \dot{y}_n^0 - \dot{y}^0(nh))^T$ gesetzt wird. Wenden wir Theorem 6 hierauf an, wobei die auftretenden Defekte beide in der Größenordnung von h^{q+1} liegen, so ergeben sich Abschätzungen

$$\begin{aligned} \left\| \begin{pmatrix} Y_{ni}^0 - y^0(t_n + c_i h) \\ \dot{Y}_{ni}^0 - \dot{y}^0(t_n + c_i h) \end{pmatrix} \right\| &\leq C\Big(\left\| \begin{pmatrix} \Delta y_n^0 \\ \Delta \dot{y}_n^0 \end{pmatrix} \right\| + h^{q+2} \Big), \\ \|\Lambda_{ni}^0 - \lambda^0(t_n + c_i h)\| &\leq \frac{C}{h}\Big(\Big\| g_y(y_n^0, \dot{y}_n^0) \begin{pmatrix} \Delta y_n^0 \\ \Delta \dot{y}_n^0 \end{pmatrix} \Big\| \\ &\qquad + h \left\| \begin{pmatrix} \Delta y_n^0 \\ \Delta \dot{y}_n^0 \end{pmatrix} \right\| + h^{q+2} \Big), \end{aligned}$$

wobei

$$g_y(y_n^0, \dot{y}_n^0) = \Big(\frac{\partial}{\partial y^0} \big(GS^T(y^0)\dot{y}^0 \big), GS^T(y^0) \Big) \Big|_{y^0 = y_n^0,\ \dot{y}^0 = \dot{y}_n^0}$$

ist. Mit (3.24) liefert dies (3.23).

(b) Nun folgen die Schranken (3.15) für den globalen Fehler für $k = 1$ wie im Beweis von Theorem 6, der die weiteren Abschätzungen

$$Y_{ni}^1 - y^1(t_n + c_i h) = O(h^q), \quad \dot{Y}_{ni}^1 - \dot{y}^1(t_n + c_i h) = O(h^q),$$
$$\Lambda_{ni}^1 - \lambda^1(t_n + c_i h) = O(h^{q-1}).$$

für den nächsten Induktionsschritt $k = 2$ liefert. ■

Kapitel 4

Fehleranalysis

In diesem Kapitel sollen die nötigen Hilfsmittel bereitgestellt werden, um Konvergenzresultate für die globalen Fehler der Runge-Kutta-Approximationen des stark gedämpften mechanischen Systems beweisen zu können. Bevor wir uns dem Einfluß von Störungen, den lokalen Fehlern und der Fehlerfortpflanzung zuwenden, wollen wir die Existenz und lokale Eindeutigkeit der Runge-Kutta-Lösungen nachweisen.

Dazu betrachten wir zunächst einen Schritt des Runge-Kutta-Verfahrens (3.3) angewandt auf (1.5). Die Runge-Kutta-Gleichungen lassen sich auch schreiben als

$$y_1 = y_0 + h\dot{y}_0 + h^2 \sum_{i,j=1}^{s} b_i a_{ij} \ddot{Y}_j, \quad \dot{y}_1 = \dot{y}_0 + h \sum_{i=1}^{s} b_i \ddot{Y}_i \tag{4.1a}$$

mit inneren Stufen

$$Y_i = y_0 + c_i h\dot{y}_0 + h^2 \sum_{j,k=1}^{s} a_{ij} a_{jk} \ddot{Y}_k, \quad \dot{Y}_i = \dot{y}_0 + h \sum_{j=1}^{s} a_{ij} \ddot{Y}_j, \tag{4.1b}$$

die für $i = 1, \ldots, s$

$$M(Y_i)\ddot{Y}_i = f(Y_i, \dot{Y}_i) - \frac{1}{\varepsilon} D(Y_i)\dot{Y}_i \tag{4.1c}$$

erfüllen müssen.

Für (4.1) verwenden wir Transformationen, wie sie in Lemma 2 auftreten. Unter diesen ist das Runge-Kutta-Verfahren zwar nicht invariant, doch die entscheidende Idee dieser Vorgehensweise liegt darin, dass dominierende Terme exakt transformiert werden können und die übrigen Funktionen nur in unbedeutenden Größenordnungen abweichen. Genauer bedeutet dies, dass der Übergang von $\frac{1}{\varepsilon} D(y)\dot{y}$ zu $\frac{1}{\varepsilon} A(u)x$ in den inneren Stufen des Runge-Kutta-Verfahrens exakt vollzogen werden kann. Damit können die angesprochenen Resultate zur Fehleranalysis mit grundlegenden Techniken aus [6] nachgewiesen werden.

4.1 Existenz und lokale Eindeutigkeit

Im folgenden Lemma beweisen wir die Existenz und lokale Eindeutigkeit der Runge-Kutta-Lösungen (4.1).

Lemma 4. *Falls die Runge-Kutta-Matrix $\mathcal{A}$ invertierbar ist, keine Eigenwerte auf der negativen reellen Halbachse besitzt und für die Anfangswerte $D(y_0)\dot{y}_0 = O(h)$ gilt, so existiert zu (4.1) für $0 < \varepsilon \le h \le h_0$ lokal eine eindeutige Lösung, wobei h_0 genügend klein gewählt werden muss, aber unabhängig von ε ist.*

Beweis. Um den Beweis zu führen, formulieren wir (4.1) zunächst mit den Schritten um, die wir schon bei der Koordinatentransformation in Lemma 2 durchgeführt haben. Als Konsequenz ergibt sich ein System der Form (2.5) in den inneren Stufen (U_i, X_i) der transformierten Runge-Kutta-Lösungen. Durch eine Homotopie für diese inneren Stufen lässt sich dann die Behauptung beweisen.

(a) Es ist leicht zu zeigen, dass das Runge-Kutta-Verfahren unter der Transformation des Systems (4.1) in ein System erster Ordung invariant ist. Deshalb kann wie im Beweis von Lemma 2 die Namensgebung geändert werden. Mit den Bezeichnungen $\dot{y}_1 = v_1,\ \ddot{Y}_i = \dot{V}_i,\ \dot{Y}_i = V_i$ ergibt sich das zu (4.1) äquivalente Verfahren

$$y_1 = y_0 + hv_0 + h^2 \sum_{i,j=1}^{s} b_i a_{ij} \dot{V}_j, \quad v_1 = v_0 + h \sum_{i=1}^{s} b_i \dot{V}_i \tag{4.2a}$$

mit

$$Y_i = y_0 + c_i h v_0 + h^2 \sum_{j,k=1}^{s} a_{ij} a_{jk} \dot{V}_k, \quad V_i = v_0 + h \sum_{j=1}^{s} a_{ij} \dot{V}_j, \tag{4.2b}$$

wobei diese inneren Stufen für $i = 1, \ldots, s$

$$\begin{aligned} \dot{Y}_i &= V_i, \\ M(Y_i)\dot{V}_i &= f(Y_i, V_i) - \frac{1}{\varepsilon} D(Y_i) V_i \end{aligned} \tag{4.2c}$$

erfüllen müssen.

Mit der Zerlegung der Massenmatrix $M(Y_i) = M^{1/2}(Y_i)M^{1/2}(Y_i)$ und einer Blockdiagonalisierung für $\widehat{D}(Y_i) = (M^{-1/2}DM^{-1/2})(Y_i)$ wie in (2.1) ist (4.2c) gleichwertig zu

$$\begin{aligned} \dot{Y}_i &= V_i, \\ (\widehat{Q}^T M^{1/2})(Y_i)\dot{V}_i &= (\widehat{Q}^T M^{-1/2})(Y_i) f(Y_i, V_i) - \frac{1}{\varepsilon} (\widehat{Q}^T \widehat{D} M^{1/2})(Y_i) V_i. \end{aligned} \tag{4.2d}$$

Runge-Kutta-Verfahren sind im allgemeinen nicht invariant unter positionsabhängigen, aber invariant unter konstanten Transformationen. Um den Übergang von $\frac{1}{\varepsilon}D(y)\dot{y}$ zu $\frac{1}{\varepsilon}A(u)x$ jedoch exakt zu beschreiben, verwenden wir für (4.2d) die Transformationen

$$V_i = T(Y_i)W_i, \quad \dot{V}_i = T(y_0)\dot{W}_i, \tag{4.3}$$

wobei die Matrix T wieder durch $T(y) = (M^{-1/2}\widehat{Q})(y)$ gegeben sei. Durch Einsetzen von (4.3) in (4.2) ergibt sich das Runge-Kutta-Verfahren

$$y_1 = y_0 + hT(y_0)w_0 + h^2\sum_{i,j=1}^{s} b_i a_{ij} T(y_0)\dot{W}_j, \quad w_1 = w_0 + h\sum_{i=1}^{s} b_i\dot{W}_i \tag{4.4a}$$

mit inneren Stufen

$$Y_i = y_0 + c_i hT(y_0)w_0 + h^2\sum_{j,k=1}^{s} a_{ij}a_{jk}T(y_0)\dot{W}_k, \quad W_i = w_0 + h\sum_{j=1}^{s} a_{ij}\dot{W}_j \tag{4.4b}$$

für die

$$\begin{aligned} \dot{Y}_i &= T(Y_i)W_i, \\ T^{-1}(Y_i)T(y_0)\dot{W}_i &= \bar{f}(Y_i, W_i) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & A(Y_i) \end{pmatrix} W_i \end{aligned} \tag{4.4c}$$

mit $\bar{f}(Y_i, W_i) = (\widehat{Q}M^{-1/2})(Y_i)f(Y_i, B(Y_i)W_i)$ gilt. Die Matrix $T^{-1}(Y_i)T(y_0)$ besitzt die Inverse $T^{-1}(y_0)T(Y_i)$, die von der Größenordnung $I + O(h)$ ist. Wir formulieren (4.4c) daher als

$$\begin{aligned} \dot{Y}_i &= T(Y_i)W_i, \\ \dot{W}_i &= \big(I + O(h)\big)\big(\bar{f}(Y_i, W_i) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & A(Y_i) \end{pmatrix} W_i\big). \end{aligned}$$

Mit derselben Variablentrennung wie im Beweis von Lemma 2 ergeben sich durch $\bar{f} = (\bar{f}_1, \bar{f}_2)^T$, $w = (p, q)^T$, $T(y) = (T_1, T_2)$ in passenden Dimensionen die Runge-Kutta-Approximationen

$$\begin{aligned} y_1 &= y_0 + hT(y_0)\begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + h^2\sum_{i,j=1}^{s} b_i a_{ij} T(y_0)\begin{pmatrix} \dot{P}_j \\ \dot{Q}_j \end{pmatrix}, \\ p_1 &= p_0 + h\sum_{i=1}^{s} b_i\dot{P}_i, \quad q_1 = q_0 + h\sum_{i=1}^{s} b_i\dot{Q}_i \end{aligned}$$

mit inneren Stufen

$$\begin{aligned} Y_i &= y_0 + c_i hT(y_0)\begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + h^2\sum_{j,k=1}^{s} a_{ij}a_{jk} T(y_0)\begin{pmatrix} \dot{P}_k \\ \dot{Q}_k \end{pmatrix}, \\ P_i &= p_0 + h\sum_{j=1}^{s} a_{ij}\dot{P}_j, \quad Q_i = q_0 + h\sum_{j=1}^{s} a_{ij}\dot{Q}_j, \end{aligned}$$

die den Gleichungen

$$\begin{aligned}
\dot{Y}_i &= (I + O(h))T(y_0)\begin{pmatrix} \dot{P}_i \\ \dot{Q}_i \end{pmatrix}, \\
\dot{P}_i &= \bar{f}_1(Y_i, P_i, Q_i) + O(h), \\
\dot{Q}_i &= \bar{f}_2(Y_i, P_i, Q_i) - \frac{1}{\varepsilon}A(Y_i)W_i + O(\frac{h}{\varepsilon})A(Y_i)W_i + O(h)
\end{aligned} \tag{4.5}$$

genügen. Für $\dot{Y}_i = T(y_0)(P_i, Q_i)^T$ erhalten wir aus der inneren Stufe für Y_i von obigem Verfahren wieder

$$\begin{aligned}
Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij}T(y_0)\Big((p_0, q_0)^T + h\sum_{k=1}^{s} a_{jk}(\dot{P}_k, \dot{Q}_k)^T\Big) \\
&= y_0 + h\sum_{j=1}^{s} a_{ij}T(y_0)(P_j, Q_j)^T = y_0 + h\sum_{j=1}^{s} a_{ij}\dot{Y}_j.
\end{aligned}$$

Mit

$$\begin{aligned}
&\widetilde{F}(Y_i, P_i, Q_i) = (T(y^0)(P_i, Q_i)^T + O(h), \bar{f}_1(Y_i, P_i, Q_i) + O(h))^T, \\
&\widetilde{\varphi}(Y_i, P_i, Q_i) = \bar{f}_2(Y_i, P_i, Q_i) + O\Big(\frac{h}{\varepsilon}\Big)A(Y_i)W_i + O(h) \text{ und} \\
&U_i = (Y_i, P_i),\ X_i = Q_i
\end{aligned}$$

gemäß der Notation von Lemma 2 ergibt sich das gewünschte Runge-Kutta-Verfahren

$$u_1 = u_0 + h\sum_{i=1}^{s} b_i\dot{U}_j, \quad x_1 = x_0 + h\sum_{i=1}^{s} b_i\dot{X}_i \tag{4.6a}$$

mit inneren Stufen

$$U_i = u_0 + h\sum_{j=1}^{s} a_{ij}\dot{U}_j, \quad X_i = x_0 + h\sum_{j=1}^{s} a_{ij}\dot{X}_j, \tag{4.6b}$$

wobei für diese

$$\begin{aligned}
\dot{U}_i &= \widetilde{F}(U_i, X_i), \\
\dot{X}_i &= -\frac{1}{\varepsilon}A(U_i)X_i + \widetilde{\varphi}(U_i, X_i)
\end{aligned} \tag{4.6c}$$

erfüllt sein muss. Insbesondere sei hier darauf verwiesen, dass $\widetilde{\varphi}$ von ε (und h) abhängt.

Nun betrachten wir via

$$(\dot{U}_i, \dot{X}_i) \longrightarrow (\dot{Y}_i, \dot{W}_i) \longrightarrow (\dot{Y}_i, \dot{V}_i) \longrightarrow \ddot{Y}_i \longrightarrow (Y_i, \dot{Y}_i)$$

alle unbekannten Runge-Kutta-Lösungen als Funktionen in $(\dot{U}_i, \dot{X}_i)$.

(b) Um die Existenz und lokale Eindeutigkeit zu beweisen, gehen wir vor wie in [6] oder [9]. Für (U_i, X_i) aus (4.6) definieren wir eine Homotopie

$$\begin{aligned} U_i(\tau) &= u_0 + h\sum_{j=1}^{s} a_{ij}\widetilde{F}(U_i(\tau), X_i(\tau)) + (\tau-1)h\sum_{j=1}^{s} a_{ij}\widetilde{F}(u_0, x_0), \\ X_i(\tau) &= x_0 + h\sum_{j=1}^{s} a_{ij}\Upsilon(U_i(\tau), X_i(\tau)) + (\tau-1)h\sum_{j=1}^{s} a_{ij}\Upsilon(u_0, x_0) \end{aligned} \tag{4.7}$$

mit $0 \le \tau \le 1$. Statt der zweiten Gleichung in (4.6c) schreiben wir aus Gründen der Übersichtlichkeit

$$\dot{X}_i = \Upsilon(U_i, X_i)$$

mit $\Upsilon(U_i, X_i) = -\frac{1}{\varepsilon}A(U_i)X_i + \widetilde{\varphi}(U_i, X_i)$.

Hierbei ist zu beachten, dass für $\tau = 0$ die Lösung durch $U_i = u_0$, $X_i = x_0$ gegeben ist, für $\tau = 1$ ist das System äquivalent zu (4.6). Differenzieren wir (4.7) nach τ und bezeichnen mit $'$ nun die Ableitung nach τ, so folgt

$$\begin{aligned} U_i' &= h\sum_{j=1}^{s} a_{ij}(\widetilde{F}_u(U_i, X_i)U_j' + \widetilde{F}_x(U_i, X_i)X_j') + h\sum_{j=1}^{s} a_{ij}\widetilde{F}(u_0, x_0), \\ X_i' &= h\sum_{j=1}^{s} a_{ij}\frac{\partial}{\partial\tau}\Upsilon(U_j, X_j) + h\sum_{j=1}^{s} a_{ij}\Upsilon(u_0, x_0), \end{aligned} \tag{4.8}$$

wobei

$$\frac{\partial}{\partial\tau}\Upsilon(U_j, X_j) = -\frac{1}{\varepsilon}\Big(A(U_i)X_i' + B(U_i, X_i)U_i'\Big) + \widetilde{\varphi}_u(U_i, X_i)U_i' + \widetilde{\varphi}_x(U_i, X_i)X_i'$$

ist und B den Tensor

$$B(U_i, X_i)U_i' = \Big(\frac{\partial}{\partial u}A(u)\Big|_{u=U_i} X_i\Big)U_i'$$

bezeichnet. Mit den Notationen

$$\begin{aligned} U &= (U_1,\ldots,U_s)^T,\ X=(X_1,\ldots,X_s)^T,\ \mathbb{1}=(1,\ldots,1)^T,\\ \{\widetilde{F}_u\} &= \text{blockdiag}(\widetilde{F}_u(U_1,X_1),\ldots,\widetilde{F}_u(U_s,X_s)),\\ \{\widetilde{F}_x\} &= \text{blockdiag}(\widetilde{F}_x(U_1,X_1),\ldots,\widetilde{F}_x(U_s,X_s)),\\ \{\widetilde{\varphi}_u\} &= \text{blockdiag}(\widetilde{\varphi}_u(U_1,X_1),\ldots,\widetilde{\varphi}_u(U_s,X_s)),\\ \{\widetilde{\varphi}_x\} &= \text{blockdiag}(\widetilde{\varphi}_x(U_1,X_1),\ldots,\widetilde{\varphi}_x(U_s,X_s)),\\ \{B\} &= \text{blockdiag}(B(U_1,X_1),\ldots,B(U_s,X_s)),\\ \{A\} &= \text{blockdiag}(A(U_1),\ldots,A(U_s)) \end{aligned}$$

erhalten wir aus (4.8) für $i=1,\ldots,s$

$$\begin{aligned} &\begin{pmatrix} I-h(\mathcal{A}\otimes I)\{\widetilde{F}_u\} & -h(\mathcal{A}\otimes I)\{\widetilde{F}_x\} \\ h(\mathcal{A}\otimes I)(-\frac{1}{\varepsilon}\{B\}+\{\widetilde{\varphi}_u\}) & I-h(\mathcal{A}\otimes I)(-\frac{1}{\varepsilon}\{A\}+\{\widetilde{\varphi}_x\}) \end{pmatrix}\begin{pmatrix} U' \\ X' \end{pmatrix}\\ &=\begin{pmatrix} h(\mathcal{A}\mathbb{1}\otimes\widetilde{F}(u_0,x_0)) \\ h[\mathcal{A}\mathbb{1}\otimes\big(-\frac{1}{\varepsilon}A(u_0)x_0+\widetilde{\varphi}(u_0,x_0)\big)] \end{pmatrix}. \end{aligned}$$

Das Symbol $\otimes$ bezeichnet das Kronecker-Produkt. Liegen die Anfangswerte $(y_0,\dot{y}_0)$ in einer Umgebung der glatten Lösung, das heißt $D(y_0)\dot{y}_0 = O(h)$, so ist auch $A(u_0)x_0 = O(h)$. Skalieren wir nun den unteren Block des obigen Systems mit $\frac{\varepsilon}{h}$, so ergibt sich

$$\begin{aligned} &\begin{pmatrix} I-h(\mathcal{A}\otimes I)\{\widetilde{F}_u\} & -h(\mathcal{A}\otimes I)\{\widetilde{F}_x\} \\ (\mathcal{A}\otimes I)(\{B\}+\varepsilon\{\widetilde{\varphi}_u\}) & \frac{\varepsilon}{h}I-(\mathcal{A}\otimes I)(-\{A\}+\varepsilon\{\widetilde{\varphi}_x\}) \end{pmatrix}\begin{pmatrix} U' \\ X' \end{pmatrix}\\ &=\begin{pmatrix} h(\mathcal{A}\mathbb{1}\otimes\widetilde{F}(u_0,x_0)) \\ h[\mathcal{A}\mathbb{1}\otimes\big(-\frac{1}{h}A(u_0)x_0+\frac{\varepsilon}{h}\widetilde{\varphi}(u_0,x_0)\big)] \end{pmatrix} \end{aligned} \tag{4.9}$$

mit einer rechten Seite, die von der Größenordnung $O(h)$ ist. Falls die Runge-Kutta-Matrix $\mathcal{A}$ invertierbar ist und keine Eigenwerte auf der negativen reellen Halbachse besitzt, existiert zu der Matrix aus diesem linearen Gleichungssystem eine beschränkte Inverse der Form

$$\begin{pmatrix} I+O(h) & O(h) \\ O(1) & O(1) \end{pmatrix}. \tag{4.10}$$

Dabei ist weiter vorauszusetzen, dass alle (U_i,X_i) in einer (h-unabhängigen) Umgebung $\mathcal{W}$ von (u_0,x_0) liegen, wobei h ausreichend klein sei, etwa $h\le h_0$. Dann folgt, dass (4.9) mit Anfangswerten $U(0)=\mathbb{1}u_0$ beziehungsweise $X(0)=\mathbb{1}x_0$ eine eindeutige Lösung in $\mathcal{W}$ auf einem nichtleeren Intervall $[0,\tau^\star]$ besitzt, das so groß gewählt werden kann, bis die Lösung $\mathcal{W}$ verlässt. Aus (4.9) und (4.10) folgt

$$U' = O(h) \tag{4.11}$$

und wegen $U_i(\tau) = u_0 + \int_0^\tau U_i'(t)\,dt$ gilt für $\tau \le \tau^\star$ auch $U_i(\tau) = u_0 + O(\tau h)$. Ebenso liefern (4.9) und (4.10)

$$X' = O(h) \tag{4.12}$$

und für $\tau \le \tau^\star$ gilt auch $X_i(\tau) = x_0 + O(\tau h)$.

Es folgt also $(U_i(\tau), X_i(\tau)) \in \mathcal{W}$ für alle $\tau \le 1$ falls h genügend klein gewählt ist, das heißt das Anfangswertproblem (4.9) mit Startwerten wie in der Voraussetzung gegeben besitzt eine Lösung für $0 \le \tau \le 1$. Die Lösung von (4.1) existiert also und es gilt

$$U_i - u_0 = O(h), \quad X_i - x_0 = O(h).$$

Die lokale Eindeutigkeit folgt aus Lemma 5 mit Störungen $\delta_i = \theta_i = 0$.

4.2 Einfluss von Störungen

Nun betrachten wir zusätzlich zu (4.1) ein gestörtes System

$$\widehat{y}_1 = \widehat{y}_0 + h\widehat{\dot{y}}_0 + h^2 \sum_{i,j=1}^{s} b_i a_{ij} \widehat{\ddot{Y}}_j, \quad \widehat{\dot{y}}_1 = \widehat{\dot{y}}_0 + h \sum_{i=1}^{s} b_i \widehat{\ddot{Y}}_i \tag{4.13a}$$

mit inneren Stufen

$$\widehat{Y}_i = \widehat{y}_0 + c_i h\widehat{\dot{y}}_0 + h^2 \sum_{j,k=1}^{s} a_{ij} a_{jk} \widehat{\ddot{Y}}_k, \quad \widehat{\dot{Y}}_i = \widehat{\dot{y}}_0 + h \sum_{j=1}^{s} a_{ij} \widehat{\ddot{Y}}_j, \tag{4.13b}$$

die für $i = 1, \ldots, s$

$$M(\widehat{Y}_i)\widehat{\ddot{Y}}_i = f(\widehat{Y}_i, \widehat{\dot{Y}}_i) - \frac{1}{\varepsilon} D(\widehat{Y}_i)\dot{Y}_i + d_i \tag{4.13c}$$

genügen. Für diese beiden Systeme von Runge-Kutta-Approximationen gilt mit den Bezeichnungen $\Delta y_0 = y_0 - \widehat{y}_0$, $\Delta \dot{y}_0 = \dot{y}_0 - \widehat{\dot{y}}_0$, $\Delta Y_i = Y_i - \widehat{Y}_i$, $\Delta \dot{Y}_i = \dot{Y}_i - \widehat{\dot{Y}}_i$, $\Delta \ddot{Y}_i = \ddot{Y}_i - \widehat{\ddot{Y}}_i$ sowie vorliegenden Störungen in den Anfangswerten Δy_0, $\Delta \dot{y}_0$ und einer Störung d_i das folgende Resultat.

Lemma 5. *Seien die Voraussetzungen von Lemma 4 gegeben, sei $D(y_0)\dot{y}_0 = O(h)$ und seien in den Anfangswerten Störungen $\Delta y_0 = O(h)$, $\Delta \dot{y}_0 = O(h)$ vorgegeben.*

Dann gibt es Matrizen $T = T(\widehat{y}_0)$ *und* $S_i = S(\widehat{y}_0, \widehat{Y}_i)$*, so dass die transformierten Variablen*

$$\begin{pmatrix} \Delta\dot{U}_i \\ \Delta\dot{X}_i \end{pmatrix} = \begin{pmatrix} \Delta\dot{Y}_i \\ T^{-1}\Delta\ddot{Y}_i \end{pmatrix}, \quad \begin{pmatrix} \Delta u_0 \\ \Delta x_0 \end{pmatrix} = \begin{pmatrix} \Delta y_0 \\ T^{-1}\Delta\dot{y}_0 \end{pmatrix}$$

beziehungsweise die transformierten Störungen $\delta_i = ((S_i d_i)_j)_{j=1,\ldots,n-m}$ *und* $\theta_i = ((S_i d_i)_j)_{j=n-m+1,\ldots,n}$ *den Abschätzungen*

$$\begin{aligned} \|\Delta\dot{U}_i\| &\le C\Big(\|\Delta u_0\| + \|\Delta x_0\| + \delta + \varepsilon\theta\Big), \\ \|\Delta\dot{X}_i\| &\le C\Big(\frac{1}{h}(\|\Delta u_0\| + \|\Delta x_0\|) + h\delta + \frac{\varepsilon}{h}\theta\Big) \end{aligned} \tag{4.14}$$

genügen, wobei $\delta = max_i\|\delta_i\|$, $\theta = max_i\|\theta_i\|$ *ist und* C *Konstanten sind, die von* ε *und* h *unabhängig sind,* $0 < \varepsilon \le h < h_0$*.*

Beweis. Wir zeigen die Behauptung wieder in zwei Schritten. Die Idee, den Einfluss der Störungen zu berechnen, liegt darin, über die Differenz der transformierten Runge-Kutta-Verfahren (4.1) und (4.13) zu einer Abschätzung für $\Delta\dot{U}_i$, $\Delta\dot{X}_i$ zu kommen, die nur von den vorliegenden Störungen abhängt, siehe Abbildung 4.1. Weil $D(y)$ singulär ist, bietet sich wieder der Weg über die Umformulierungen von Lemma 2 an. Im ersten Beweisschritt transformieren wir (4.13) so, dass $\frac{1}{\varepsilon}D(\widehat{Y}_i)\dot{Y}_i$ ebenfalls exakt wiedergegeben wird. Unter Berücksichtigung dieses Übergangs fassen wir die Differenz der Transformationen von (4.1) und (4.13) geeignet zusammen. Im zweiten Beweisschritt lassen sich dann die Runge-Kutta-Verfahren in (u, x) und $(\widehat{u}, \widehat{x})$ subtrahieren. Nun müssen nur noch spezielle Terme in Abhängigkeit von Δu_0 und Δx_0 abgeschätzt werden.

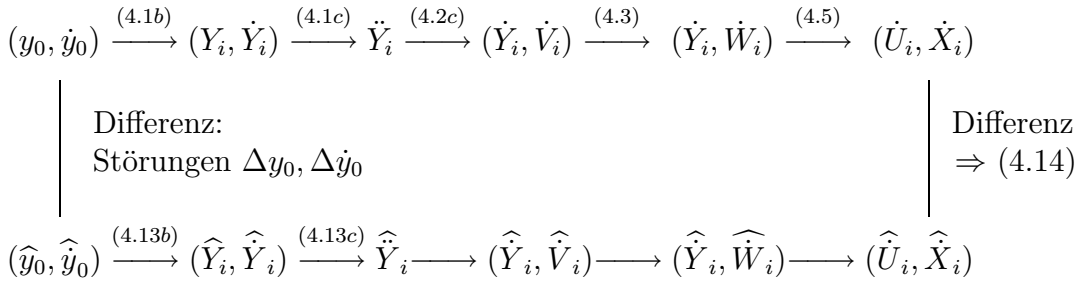

Abbildung 4.1: Beweisverfahren im Überblick

(a) Für die Größen in (4.13) benutzen wir also analoge Transformationen wie im Beweis von Lemma 4. Zunächst überführen wir (4.13c) wieder in eine Differentialgleichung der Ordnung 1 und erhalten Runge-Kutta-Gleichungen wie in (4.2a) und

(4.2b) für $\widehat{y}_1$, $\widehat{v}_1$, $\widehat{Y}_i$, $\widehat{V}_i$. Diese sollen in den Zwischenschritten nun nicht mehr aufgeführt werden. Mit denselben Matrixzerlegungen für M und $\widehat{D}$ müssen die inneren Stufen der gestörten Runge-Kutta-Gleichungen das System

$$\begin{aligned} \dot{\widehat{Y}}_i &= \widehat{V}_i, \\ (\widehat{Q}^T M^{1/2})(\widehat{Y}_i)\dot{\widehat{V}}_i &= (\widehat{Q}^T M^{-1/2})(\widehat{Y}_i) f(\widehat{Y}_i, \widehat{V}_i) \\ &\quad -\frac{1}{\varepsilon}(\widehat{Q}^T \widehat{D} M^{1/2})(\widehat{Y}_i)\widehat{V}_i + (\widehat{Q}^T M^{-1/2})(\widehat{Y}_i) d_{i2} \end{aligned}$$

erfüllen. Zu beachten ist hierbei, dass $\widehat{D}$ und $\widehat{Q}$ nicht etwa gestörte Matrizen sind, sondern in Lemma 2 so bezeichnet wurden. Mit Transformationen

$$\widehat{V}_i = T(\widehat{Y}_i)\widehat{W}_i, \quad \dot{\widehat{V}}_i = T(\widehat{y}_0)\dot{\widehat{W}}_i, \tag{4.15}$$

wobei wiederum $T(y) = (M^{-1/2}\widehat{Q})(y)$ ist, ergibt sich ein Runge-Kutta-Verfahren, das bis auf Störungen (4.4) entspricht. Die Gleichung für die inneren Stufen ist insbesondere durch

$$\begin{aligned} \dot{\widehat{Y}}_i &= T(\widehat{Y}_i)\widehat{W}_i, \\ T^{-1}(\widehat{Y}_i)T(\widehat{y}_0)\dot{\widehat{W}}_i &= \bar{f}(\widehat{Y}_i, \widehat{W}_i) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & A(\widehat{Y}_i) \end{pmatrix}\widehat{W}_i + \widehat{Q}^T M^{-1/2}(\widehat{Y}_i) d_i \end{aligned}$$

gegeben und lässt sich mit $(T^{-1}(\widehat{Y}_i)T(\widehat{y}_0))^{-1} = I + O(h)$ umformulieren zu

$$\begin{aligned} \dot{\widehat{Y}}_i &= T(\widehat{Y}_i)\widehat{W}_i, \\ \dot{\widehat{W}}_i &= \big(I + O(h)\big)\Big(\bar{f}(\widehat{Y}_i, \widehat{W}_i) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & A(\widehat{Y}_i) \end{pmatrix}\widehat{W}_i + \widehat{Q}^T M^{-1/2}(\widehat{Y}_i) d_i\Big). \end{aligned}$$

Trennen wir die Variablen wie in Lemma 2, so folgt ein Verfahren

$$\begin{aligned} \dot{\widehat{Y}}_i &= T(\widehat{Y}_i)(\widehat{P}_i, \widehat{Q}_i)^T, \\ \dot{\widehat{P}}_i &= \bar{f}_1(\widehat{Y}_i, \widehat{P}_i, \widehat{Q}_i) + d_{i1} + O(h), \\ \dot{\widehat{Q}}_i &= \bar{f}_2(\widehat{Y}_i, \widehat{P}_i, \widehat{Q}_i) - \frac{1}{\varepsilon}A(\widehat{Y}_i)\widehat{W}_i + O\Big(\frac{h}{\varepsilon}\Big)A(\widehat{Y}_i)\widehat{W}_i + d_{i2} + O(h), \end{aligned}$$

wobei wir

$$(d_{i1},\ d_{i2})^T = S_i d_i \quad \text{mit } S_i = S(\widehat{y}_0, \widehat{Y}_i) = T^{-1}(\widehat{y}_0)(T\widehat{Q}M^{1/2})(\widehat{Y}_i)$$

setzen. Mit Definitionen von $\widetilde{F}$ und $\widetilde{\varphi}$ wie in Lemma 4 und Variablensubstitutionen $\widehat{U}_i = (\widehat{Y}_i,\ \widehat{P}_i)$, $\widehat{X}_i = \widehat{Q}_i$ erhalten wir schließlich das Runge-Kutta-Verfahren

$$\widehat{u}_1 = \widehat{u}_0 + h\sum_{i=1}^{s} b_i \dot{\widehat{U}}_j, \quad \widehat{x}_1 = \widehat{x}_0 + h\sum_{i=1}^{s} b_i \dot{\widehat{X}}_i \tag{4.16a}$$

mit inneren Stufen

$$\widehat{U}_i = \widehat{u}_0 + h\sum_{j=1}^{s} a_{ij}\widehat{\dot{U}}_j, \quad \widehat{X}_i = \widehat{x}_0 + h\sum_{j=1}^{s} a_{ij}\widehat{\dot{X}}_j, \tag{4.16b}$$

die

$$\begin{aligned} \widehat{\dot{U}}_i &= \widetilde{F}(\widehat{U}_i, \widehat{X}_i) + \delta_i, \\ \widehat{\dot{X}}_i &= -\frac{1}{\varepsilon}A(\widehat{U}_i)\widehat{X}_i + \widetilde{\varphi}(\widehat{U}_i, \widehat{X}_i) + \theta_i \end{aligned} \tag{4.16c}$$

mit $\delta_i = d_{i1}$ und $\theta_i = d_{i2}$ genügen. Um die Differenz der Runge-Kutta-Verfahren (4.16) und (4.6) betrachten zu können, müssen genau diese Umformungen durchgeführt werden. Daher können wir die folgenden Transformationen für die Differenzen von (4.13) und (4.1) ansetzen. Es ist

$$\begin{pmatrix} u_0 \\ x_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ p_0 \\ q_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ w_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ T^{-1}(y_0)v_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ T^{-1}(y_0)\dot{y}_0 \end{pmatrix}$$

und analog

$$\begin{pmatrix} \widehat{u}_0 \\ \widehat{x}_0 \end{pmatrix} = \begin{pmatrix} \widehat{y}_0 \\ \widehat{p}_0 \\ \widehat{q}_0 \end{pmatrix} = \begin{pmatrix} \widehat{y}_0 \\ \widehat{w}_0 \end{pmatrix} = \begin{pmatrix} \widehat{y}_0 \\ T^{-1}(\widehat{y}_0)\widehat{v}_0 \end{pmatrix} = \begin{pmatrix} \widehat{y}_0 \\ T^{-1}(\widehat{y}_0)\widehat{\dot{y}}_0 \end{pmatrix},$$

als Differenz ergibt sich also

$$\begin{pmatrix} u_0 - \widehat{u}_0 \\ x_0 - \widehat{x}_0 \end{pmatrix} = \begin{pmatrix} \Delta y_0 \\ T^{-1}(\widehat{y}_0)\Delta\dot{y}_0 + O(\|\Delta y_0\|) \end{pmatrix}.$$

Der 'O-Term' spielt für die nachfolgenden Beweisschritte keine weitere Rolle. Dies motiviert die Einführung der Größen

$$\begin{pmatrix} \Delta u_0 \\ \Delta x_0 \end{pmatrix} = \begin{pmatrix} \Delta y_0 \\ T^{-1}(\widehat{y}_0)\Delta\dot{y}_0 \end{pmatrix}, \quad \begin{pmatrix} \Delta\dot{U}_i \\ \Delta\dot{X}_i \end{pmatrix} = \begin{pmatrix} \Delta\dot{Y}_i \\ T^{-1}(\widehat{y}_0)\Delta\ddot{Y}_i \end{pmatrix}. \tag{4.17}$$

Die Gleichheit zwischen den Vektoren darf nicht darüber hinweg täuschen, dass $\Delta u_0 \in \mathbb{R}^{d+m}$ sowie $\Delta x_0 \in \mathbb{R}^m$, aber $\Delta y_0,\ \Delta\dot{y}_0 \in \mathbb{R}^d$ sind.

Die entscheidenden Transformationen zur exakten Wiedergabe von $\frac{1}{\varepsilon}A(u)x$ waren $V_i = T(Y_i)W_i$ beziehungsweise $\widehat{V}_i = T(\widehat{Y}_i)\widehat{W}_i$; hier müssen wir also genauer vorgehen. Wie oben ist

$$\begin{pmatrix} U_i \\ X_i \end{pmatrix} = \begin{pmatrix} Y_i \\ T^{-1}(Y_i)\dot{Y}_i \end{pmatrix}, \quad \begin{pmatrix} \widehat{U}_i \\ \widehat{X}_i \end{pmatrix} = \begin{pmatrix} \widehat{Y}_i \\ T^{-1}(\widehat{Y}_i)\widehat{\dot{Y}}_i \end{pmatrix}$$

und daher

$$\begin{pmatrix} \Delta U_i \\ \Delta X_i \end{pmatrix} = \begin{pmatrix} \Delta Y_i \\ T^{-1}(Y_i)\dot{Y}_i - T^{-1}(\widehat{Y}_i)\dot{\widehat{Y}}_i \end{pmatrix}.$$

Nun folgt eine längere Rechnung, um die Differenz in der zweiten Zeile des Blockvektors präzise genug abzuschätzen. Die Abschätzung nehmen wir für den späteren Gebrauch gleich in Abhängigkeit von Δu_0, Δx_0 und $\Delta\dot{U}_i$, $\Delta\dot{X}_i$ vor. Wir betrachten zunächst die Entwicklungen

$$\begin{aligned} T^{-1}(Y_i)\dot{Y}_i &= T^{-1}(\widehat{y}_0)\dot{Y}_i + \ell(\widehat{y}_0)(Y_i - \widehat{y}_0)\dot{Y}_i, \\ &\qquad + \mathcal{R}[\widehat{y}_0, Y_i](Y_i - \widehat{y}_0, Y_i - \widehat{y}_0)\dot{Y}_i \\ T^{-1}(\widehat{Y}_i)\dot{\widehat{Y}}_i &= T^{-1}(\widehat{y}_0)\dot{\widehat{Y}}_i + \ell(\widehat{y}_0)(\widehat{Y}_i - \widehat{y}_0)\dot{\widehat{Y}}_i \\ &\qquad + \mathcal{R}[\widehat{y}_0, \widehat{Y}_i](\widehat{Y}_i - \widehat{y}_0, \widehat{Y}_i - \widehat{y}_0)\dot{\widehat{Y}}_i \end{aligned} \tag{4.18}$$

mit

$$\begin{aligned} \ell(\widehat{y}_0)(Y-y)\dot{Y} &= \frac{\partial}{\partial y}(T^{-1}(y)\dot{Y})\Big|_{y=\widehat{y}_0}(Y-y), \\ \mathcal{R}[y,Y](Y-y,Y-y)\dot{Y} &= \int_0^1 \mathrm{BLF}(y+s(Y-y))(Y-y,Y-y)\dot{Y}\,ds \qquad \text{und} \\ \mathrm{BLF}(\zeta)(Y-y,Y-y) &= \frac{\partial^2}{\partial y^2}\Big(T^{-1}(y)\Big)\Big|_{y=\zeta}(Y-y,Y-y), \end{aligned}$$

wobei BLF für eine Bilinearform steht, die in $Y-y$ bilinear ist und von ζ abhängt. Die Variable ζ ist ein Punkt auf der Verbindungsgeraden zwischen y und Y. Die Differenz der beiden Terme aus (4.18) liefert unter Ausnutzung der Rechenregeln für Linear- und Bilinearformen und der Gleichheit $Y_i = \widehat{Y}_i + \Delta Y_i$

$$\begin{aligned} T^{-1}(Y_i)\dot{Y}_i \;-\; & T^{-1}(\widehat{Y}_i)\dot{\widehat{Y}}_i = \\ & T^{-1}(\widehat{y}_0)\Delta\dot{Y}_i + \ell(\widehat{y}_0)(\widehat{Y}_i - \widehat{y}_0)\Delta\dot{Y}_i + \ell(\widehat{y}_0)(\Delta Y_i)\dot{Y}_i \\ +\; & \mathcal{R}[\widehat{y}_0, Y_i](\widehat{Y}_i - \widehat{y}_0, \widehat{Y}_i - \widehat{y}_0)\dot{Y}_i - \mathcal{R}[\widehat{y}_0, \widehat{Y}_i](\widehat{Y}_i - \widehat{y}_0, \widehat{Y}_i - \widehat{y}_0)\dot{\widehat{Y}}_i \\ +\; & \big(\mathcal{R}[\widehat{y}_0, Y_i](\widehat{Y}_i - \widehat{y}_0, \Delta Y_i) + \mathcal{R}[\widehat{y}_0, Y_i](\Delta Y_i, \widehat{Y}_i - \widehat{y}_0) \\ +\; & \mathcal{R}[\widehat{y}_0, Y_i](\Delta Y_i, \Delta Y_i)\big)\dot{Y}_i. \end{aligned}$$

Da $\dot{Y}_i = O(1)$, $\widehat{Y}_i - \widehat{y}_0 = O(h)$, $\Delta Y_i = O(h)$ und die Linear- und Bilinearformen durch ihre linearen Anteile beschränkt sind, gilt mit $\Delta\dot{Y}_i = \Delta\dot{y}_0 + h\sum_{j=1}^s a_{ij}\Delta\ddot{Y}_j$

$$\begin{aligned} T^{-1}(Y_i)\dot{Y}_i - T^{-1}(\widehat{Y}_i)\dot{\widehat{Y}}_i &= T^{-1}(\widehat{y}_0)\Delta\dot{y}_0 + h\sum_{j=1}^s a_{ij}T^{-1}(\widehat{y}_0)\Delta\ddot{Y}_j \\ &\qquad + O\big(\|\Delta Y_i\| + h\|\Delta\dot{Y}_i\| + h^2\|\dot{Y}_i\| + h^2\|\dot{\widehat{Y}}_i\|\big). \end{aligned}$$

Wir erhalten also mit (4.17), $\dot{Y}_i = O(1)$, $\widehat{\dot{Y}}_i = O(1)$ und $\Delta\ddot{Y} = (\Delta\ddot{Y}_i)_{i=1}^s$ insgesamt

$$\begin{aligned}\begin{pmatrix} \Delta U_i \\ \Delta X_i \end{pmatrix} &= \begin{pmatrix} \Delta y_0 + c_i h \Delta\dot{y}_0 + h^2 \sum_{j,k=1}^s a_{ij} a_{jk} \Delta\ddot{Y}_k \\ T^{-1}(\widehat{y}_0)\Delta\dot{y}_0 + h\sum_{j=1}^s a_{ij} T^{-1}(\widehat{y}_0)\Delta\ddot{Y}_j + R_0 \end{pmatrix} \\ &= \begin{pmatrix} \Delta u_0 \\ \Delta x_0 \end{pmatrix} + h\sum_{j=1}^s a_{ij} \begin{pmatrix} \Delta\dot{U}_j \\ \Delta\dot{X}_j \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 \\ O\big(\|\Delta u_0\| + h\|\Delta x_0\| + h^2(\|\Delta\dot{U}\| + \|\Delta\dot{X}\|)\big) \end{pmatrix},\end{aligned} \tag{4.19}$$

wobei $R_0 = O\big(\|\Delta y_0\| + h\|\Delta\dot{y}_0\| + h^2\|\Delta\ddot{Y}\|\big)$.

(b) Nun subtrahieren wir die Runge-Kutta-Verfahren (4.16) und (4.6). Dies liefert

$$\Delta u_1 = \Delta u_0 + h\sum_{i=1}^s b_i \Delta\dot{U}_j, \quad \Delta x_1 = \Delta x_0 + h\sum_{i=1}^s b_i \Delta\dot{X}_i \tag{4.20a}$$

mit

$$\Delta U_i = \Delta u_0 + h\sum_{j=1}^s a_{ij} \Delta\dot{U}_j, \quad \Delta X_i = x_0 + h\sum_{j=1}^s a_{ij} \Delta\dot{X}_j, \tag{4.20b}$$

wobei für die inneren Stufen

$$\begin{aligned}\Delta\dot{U}_i &= \widetilde{F}(U_i, X_i) - \widetilde{F}(\widehat{U}_i, \widehat{X}_i) - \delta_i, \\ \Delta\dot{X}_i &= -\frac{1}{\varepsilon}(A(U_i)X_i - A(\widehat{U}_i)\widehat{X}_i) + \widetilde{\varphi}(U_i, X_i) - \widetilde{\varphi}(\widehat{U}_i, \widehat{X}_i) - \theta_i\end{aligned} \tag{4.20c}$$

gilt. Die Differenzen der Funktionen $\widetilde{F}$ und $\widetilde{\varphi}$ sind durch Taylor-Entwicklungen um $(\widehat{U}_i, \widehat{X}_i)$ einfach abzuschätzen. Es ergibt sich

$$\begin{aligned}\Delta\dot{U}_i &= O\big(\| \begin{pmatrix} \Delta U_i \\ \Delta X_i \end{pmatrix} \|\big) - \delta_i, \\ \Delta\dot{X}_i &= -\frac{1}{\varepsilon}\big(A(U_i)X_i - A(\widehat{U}_i)\widehat{X}_i\big) + O\big(\frac{h}{\varepsilon}\| \begin{pmatrix} \Delta U_i \\ \Delta X_i \end{pmatrix} \|\big) - \theta_i.\end{aligned} \tag{4.21}$$

Also bleibt der Term $A(U_i)X_i - A(\widehat{U}_i)\widehat{X}_i$ in der zweiten Gleichung abzuschätzen. Hierzu betrachten wir erneut eine Taylorentwicklung von $A(U_i)X_i$ beziehungsweise $A(\widehat{U}_i)\widehat{X}_i$ um $\widehat{u}_0$, die nach den ersten Entwicklungsgliedern mit entsprechenden Resttermen abgebrochen werden kann. Wie bei der Abschätzung der auf Seite 65 behandelten Differenz $T^{-1}(Y_i)\dot{Y}_i - T^{-1}(\widehat{Y}_i)\widehat{\dot{Y}}_i$ folgt durch einfache Rechenarbeit

$$A(U_i)X_i - A(\widehat{U}_i)\widehat{X}_i = A(\widehat{u}_0)\Delta X_i + O(h\|\Delta X_i\|) + O(h\|\Delta U_i\|), \tag{4.22}$$

denn unter der Voraussetzung $D(y_0)\widehat{y}_0 = O(h)$ folgt aus (4.6) für $0 < \varepsilon < h$, dass $\|X_i\| = O(h)$ gilt.

Aus (4.19) ergibt sich direkt

$$\begin{aligned}\| \begin{pmatrix} \Delta U_i \\ \Delta X_i \end{pmatrix} \| \;&=\; \| \begin{pmatrix} \Delta u_0 \\ \Delta x_0 \end{pmatrix} + h(\mathcal{A} \otimes I) \begin{pmatrix} \Delta U \\ \Delta X \end{pmatrix} \\ &\qquad + \begin{pmatrix} 0 \\ O(\|\Delta u_0\| + h\|\Delta x_0\| + h^2(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)) \end{pmatrix} \|\end{aligned}$$

und somit

$$\| \begin{pmatrix} \Delta U_i \\ \Delta X_i \end{pmatrix} \| = O(\|\Delta u_0\| + \|\Delta x_0\| + h(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)).$$

Setzen wir dies sowie (4.22) in (4.21) ein, so ist

$$\begin{aligned}\Delta \dot{U}_i \;&=\; O(\|\Delta u_0\| + \|\Delta x_0\| + h(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)) - \delta_i, \\ \Delta \dot{X}_i + \frac{1}{\varepsilon} A(\widehat{u}_0)\Delta X_i \;&=\; O\Big(\frac{h}{\varepsilon}(\|\Delta u_0\| + \|\Delta x_0\|)\Big) \\ &\qquad + O\Big(\frac{h^2}{\varepsilon}(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)\Big) - \theta_i.\end{aligned} \tag{4.23}$$

Mit dem Eintrag aus (4.19) für ΔX_i folgt für die zweite Zeile dieses Systems

$$\begin{aligned}\Delta \dot{X}_i \;+\; &\frac{1}{\varepsilon} A(\widehat{u}_0)\Big(\Delta x_0 + h\sum_{j=1}^{s} a_{ij}\Delta \dot{X}_j \\ &\qquad + O\big(\|\Delta u_0\| + h\|\Delta x_0\| + h^2(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)\big)\Big) \\ &= O\Big(\frac{h}{\varepsilon}\big(\|\Delta u_0\| + \|\Delta x_0\| + h(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)\big)\Big) - \theta_i.\end{aligned}$$

Nach Skalierung mit $\frac{\varepsilon}{h}$ erhalten wir für $0 < \varepsilon \leq h < h_0$ und wegen der Beschränktheit von A

$$\begin{aligned}\frac{\varepsilon}{h}\Delta \dot{X}_i \;+\; &A(\widehat{u}_0)\sum_{j=1}^{s} a_{ij}\Delta \dot{X}_j \\ &= O(h\|\Delta \dot{U}\| + h\|\Delta \dot{X}\|) + O\Big(\frac{1}{h}\|\Delta u_0\| + \frac{1}{h}\|\Delta x_0\|\Big) - \frac{\varepsilon}{h}\theta_i.\end{aligned}$$

Mit derselben Notation wie in Lemma 4 (siehe Seite 60) ist dieses Gleichungssystem in $\Delta \dot{X} = (\Delta \dot{X}_1, \ldots, \Delta \dot{X}_s)^T$ gleichwertig zu

$$\begin{aligned}&(\frac{\varepsilon}{h} I_s \otimes I_m + \mathcal{A} \otimes A)\Delta \dot{X} = \\ &\qquad O(h\|\Delta \dot{U}\| + h\|\Delta \dot{X}\|) + O\Big(\frac{1}{h}(\|\Delta u_0\| + \|\Delta x_0\|)\Big) - \frac{\varepsilon}{h}\theta_i.\end{aligned} \tag{4.24}$$

Für $0 < \varepsilon \leq h$ ist $\frac{\varepsilon}{h} I_s \otimes I_m$ invertierbar, und weil $A(\widehat{u}_0)$ und $\mathcal{O}$ keine negativen Eigenwerte haben, ist die Matrix auf der linken Seite invertierbar. Ihre Inverse ist von der Größenordnung $O(1)$, also folgt aus (4.24)

$$\Delta\dot{X} = O(h\|\Delta\dot{U}\|) + O(h\|\Delta\dot{X}\|) + O\Big(\frac{1}{h}(\|\Delta u_0\| + \|\Delta x_0\|)\Big) - C\frac{\varepsilon}{h}\theta_i.$$

Zur Auflösung dieser Gleichung nach $\|\Delta\dot{X}\|$ argumentieren wir wieder wie im Beweis von Theorem 1 und mit $\theta = \max_{i=1,\ldots,s} \|\theta_i\|$ erhalten wir schließlich

$$\|\Delta\dot{X}\| = O(h\|\Delta\dot{U}\|) + O\Big(\frac{1}{h}(\|\Delta u_0\| + \|\Delta x_0\|) + \frac{\varepsilon}{h}\theta\Big). \tag{4.25}$$

Eingesetzt in die erste Gleichung von (4.23) folgt

$$\Delta\dot{U}_i = O(h\|\Delta\dot{U}\| + \|\Delta u_0\| + \|\Delta x_0\| + \varepsilon\theta) - \delta_i,$$

und mit $\delta = \max_{i=1,\ldots,s} \|\delta_i\|$ ergibt sich hieraus

$$\|\Delta\dot{U}\| = O(\|\Delta u_0\| + \|\Delta x_0\| + \delta + \varepsilon\theta).$$

Nutzen wir diese Abschätzung in (4.25) aus, so ist

$$\|\Delta\dot{X}\| = O\Big(\frac{1}{h}(\|\Delta u_0\| + \|\Delta x_0\|) + h\delta + \frac{\varepsilon}{h}\theta\Big).$$

Aus diesen Schranken folgen die Abschätzungen von $\|\Delta\dot{U}_i\|$ und $\|\Delta\dot{X}_i\|$ für Indizes $i = 1, \ldots, s$. Somit ist die Behauptung des Lemmas bewiesen. ∎

4.3 Fehler nach einem Schritt

Wir bestimmen zunächst den Fehler nach einem Schritt. Mit der gleichen Argumentation und einem Induktionsargument lässt sich dann auf den Fehler von Schritt n zu Schritt $n+1$ schließen.

Lemma 6. *Mit den Voraussetzungen von Lemma 5 gilt für die transformierten Variablen nach einem Schritt*

$$\begin{aligned}
\Delta u_1 &= \Delta u_0 + O\Big(h(\|\Delta u_0\| + \|\Delta x_0\|) + h\delta + h\varepsilon\theta\Big), \\
\Delta x_1 &= R\Big(-\frac{h}{\varepsilon}A\Big)\Delta x_0 + O\Big(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta + \varepsilon\theta\Big).
\end{aligned}$$

Hierbei bezeichnet R die Stabilitätsfunktion des Runge-Kutta-Verfahrens.

Beweis. Aus der ersten Gleichung von (4.20a) und mit (4.14) folgt sofort die Behauptung für Δu_1 mit der Schranke für $\|\Delta \dot{U}_i\|$ aus (4.14), denn

$$\Delta u_1 = \Delta u_0 + h \sum_{i=1}^{s} b_i \Delta \dot{U}_i = \Delta u_0 + O(h(\|\Delta u_0\| + \|\Delta x_0\| + \delta + \varepsilon\theta)).$$

Mit den Abschätzungen für $\|\Delta \dot{U}_i\|$ und $\|\Delta \dot{X}_i\|$ ergibt sich aus der zweiten Zeile von (4.23) mit $\theta = \max_{i=1,\ldots,s} \|\theta_i\|$

$$\Delta \dot{X}_i = -\frac{1}{\varepsilon} A(\widehat{u}_0)\Delta X_i + O\Big(\frac{h}{\varepsilon}(\|\Delta u_0\| + \|\Delta x_0\| + h\delta) + \theta\Big). \tag{4.26}$$

Mit (4.14) errechnen wir aus der Gleichheit in der zweiten Komponente der Formel (4.19) weiter, dass

$$\Delta X_i = \Delta x_0 + h \sum_{j=1}^{s} a_{ij} \Delta \dot{X}_j + O(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta + h\varepsilon\theta) \tag{4.27}$$

gilt und nach einem Schritt ist

$$\Delta x_1 = \Delta x_0 + h \sum_{i=1}^{s} b_i \Delta \dot{X}_i + O(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta + h\varepsilon\theta). \tag{4.28}$$

Ignorieren wir zunächst die Störterme in (4.26), (4.27) und (4.28), so ist

$$\begin{aligned} \Delta x_1 &= \Delta x_0 + h \sum_{i=1}^{s} b_i \Delta \dot{X}_i, \\ \Delta X_i &= \Delta x_0 + h \sum_{j=1}^{s} a_{ij} \Delta \dot{X}_j, \end{aligned}$$

wobei

$$\Delta \dot{X}_i = -\frac{1}{\varepsilon} A(\widehat{u}_0)\Delta X_i$$

erfüllt sein muss. Dies sind Runge-Kutta-Gleichungen der linearen Differentialgleichung

$$\dot{w} = -\frac{1}{\varepsilon} A w.$$

Die Runge-Kutta-Lösung dieser Differentialgleichung ist nach einem Schritt durch

$$w_1 = R\Big(-\frac{h}{\varepsilon}A\Big) w_0$$

gegeben, wobei dann R die Stabilitätsfunktion des Runge-Kutta-Verfahrens bezeichnet. Ziehen wir die Störterme in Betracht und setzen (4.27) in (4.26) ein, so

ergibt sich mit der Notation wie in Lemma 4 und $\Delta\dot{X} = (\Delta\dot{X}_1, \dots \Delta\dot{X}_s)$ sowie $b = (b_1, \dots, b_s)^T$

$$\Delta\dot{X} \;=\; \big(I \otimes A\big)\Big(I - (\mathcal{A} \otimes (-\frac{h}{\varepsilon})A)\Big)^{-1}\Big(-\frac{1}{\varepsilon} \cdot \mathbb{1} \otimes \Delta x_0 + O(\beta)\Big) \tag{4.29}$$

mit $\beta = \frac{1}{\varepsilon}(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta) + \theta$. Auch (4.28) formulieren wir als

$$\Delta x_1 = \Delta x_0 + h(b^T \otimes I)\Delta\dot{X} + O(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta + h\varepsilon\theta)$$

und Einsetzen von (4.29) liefert

$$\begin{aligned}\Delta x_1 \;&=\; \Delta x_0 + h\big(b^T \otimes A\big)\Big(I - (\mathcal{A} \otimes (-\frac{h}{\varepsilon})A)\Big)^{-1}\Big(-\frac{1}{\varepsilon} \cdot \mathbb{1} \otimes \Delta x_0 + O(\beta)\Big)\\ &=\; R\Big(-\frac{h}{\varepsilon}A\Big)\Delta x_0 + h\big(b^T \otimes A\big)\Big(I - (\mathcal{A} \otimes (-\frac{h}{\varepsilon})A)\Big)^{-1} \cdot O(\beta)\\ &\quad +O(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta + h\varepsilon\theta).\end{aligned}$$

Nun ist noch der zweite Summand der obigen Gleichung abzuschätzen, und dazu bestimmen wir die Größenordnung von $(I - (\mathcal{A} \otimes (-\frac{h}{\varepsilon})A))^{-1}$. Es ist

$$\Big(I - (\mathcal{A} \otimes (-\frac{h}{\varepsilon})A)\Big) = -\frac{h}{\varepsilon}\Big(-\mathcal{A} \otimes A\Big)\Big(-\frac{\varepsilon}{h}(\mathcal{A} \otimes A)^{-1} + I\Big)$$

und daraus folgt

$$\Big(I - \frac{h}{\varepsilon}(-\mathcal{A} \otimes A)\Big)^{-1} = \Big(I - \frac{\varepsilon}{h}(-\mathcal{A} \otimes A)^{-1}\Big)^{-1}\Big(-\frac{h}{\varepsilon}(-\mathcal{A} \otimes A)\Big)^{-1} = O\Big(\frac{\varepsilon}{h}\Big),$$

denn für $0 < \varepsilon \le h < h_0$ mit ausreichend kleinem h_0 liefert die Neumannsche Reihe, dass $\big(I - \frac{\varepsilon}{h}(-\mathcal{A} \otimes A)^{-1}\big)^{-1} = I + O(\frac{\varepsilon}{h})$ und $\big(-\frac{h}{\varepsilon}(-\mathcal{A} \otimes A)\big)^{-1} = O(\frac{\varepsilon}{h})$. Also können die Restterme zusammengefasst werden und es ergibt sich

$$\Delta x_1 = R\Big(-\frac{h}{\varepsilon}A\Big)\Delta x_0 + O\Big(\|\Delta u_0\| + h\|\Delta x_0\| + h^2\delta + \varepsilon\theta\Big).$$

Somit ist die Behauptung gezeigt und der Beweis abgeschlossen. ∎

Aus den Beweisen der bisherigen Lemmata dieses Paragraphen lässt sich durch analoge Vorgehensweisen der Fehler von Schritt n zu Schritt $n+1$ rekonstruieren.

Folgerung (Fehler von Schritt n zum Schritt $n+1$). Die Runge-Kutta-Matrix $\mathcal{A}$ besitze keine Eigenwerte auf der negativen reellen Halbachse und sei invertierbar. Weiter setzen wir voraus, dass die ungestörten beziehungsweise gestörten Runge-Kutta-Approximationen, gegeben durch $(y_n, \dot{y}_n)$ und $(\widehat{y}_n, \dot{\widehat{y}}_n)$, für $0 \le nh \le T$ existieren und dass

$$D(y_n)\dot{y}_n = O(h), \quad D(y_n)\dot{\widehat{y}}_n = O(h) \quad \text{für } 0 \le nh \le T \tag{4.30a}$$

gilt. Für die Anfangswerte setzen wir wieder

$$\begin{aligned} D(y_0)\dot{y}_0 = O(h)\ ,\ D(y_0)\widehat{\dot{y}}_0 = O(h) \quad \text{und} \\ y_0 - \widehat{y}_0 = O(h)\ ,\ \dot{y}_0 - \widehat{\dot{y}}_0 = O(h) \end{aligned} \tag{4.30b}$$

voraus. Weiter seien Transformationsmatrizen $T = T(\widehat{y}_n)$ und $S_{ni} = S(\widehat{y}_n, \widehat{Y}_{ni})$ wie in Lemma 5 gegeben. Damit betrachten wir die transformierten Variablen

$$\begin{pmatrix} \Delta\dot{U}_{ni} \\ \Delta\dot{X}_{ni} \end{pmatrix} = \begin{pmatrix} \Delta\dot{Y}_{ni} \\ T^{-1}\Delta\ddot{Y}_{ni} \end{pmatrix}, \quad \begin{pmatrix} \Delta u_n \\ \Delta x_n \end{pmatrix} = \begin{pmatrix} \Delta y_n \\ T^{-1}\Delta\dot{y}_n \end{pmatrix}, \tag{4.30c}$$

wobei wir $\Delta y_n = y_n - \widehat{y}_n$, $\Delta\dot{y}_n = \dot{y}_n - \widehat{\dot{y}}_n$, $\Delta\dot{Y}_{ni} = \dot{Y}_{ni} - \widehat{\dot{Y}}_{ni}$, $\Delta\ddot{Y}_{ni} = \ddot{Y}_{ni} - \widehat{\ddot{Y}}_{ni}$ setzen. Die Störung d_{ni} des Verfahrens (4.13) im n-ten Schritt wird zu

$$\delta_{ni} = ((S_{ni}d_{ni})_j)_{j=1,\ldots,n-m} \quad \text{und} \quad \theta_{ni} = ((S_{ni}d_{ni})_j)_{j=n-m+1,\ldots,n} \tag{4.30d}$$

transformiert. Dabei bezeichnen δ und θ nun die Schranken

$$\|\delta_{ni}\| \le \delta, \quad \|\theta_{ni}\| \le \theta \quad \text{für } 0 \le nh \le T \text{ und alle } i = 1, \ldots, s. \tag{4.30e}$$

Diese Schranken der Störung seien von der Größenordnung

$$\delta = O(h), \quad \varepsilon\theta = O(h). \tag{4.30f}$$

Unter diesen Voraussetzungen gilt für den Fehler von Schritt n nach Schritt $n+1$ mit $A_n = A(\widehat{u}_n)$

$$\begin{aligned} \Delta u_{n+1} &= \Delta u_n + O\Big(h(\|\Delta u_n\| + \|\Delta x_n\|) + h\delta + h\varepsilon\theta\Big), \\ \Delta x_{n+1} &= R\Big(-\frac{h}{\varepsilon}A_n\Big)\,\Delta x_n + O\Big(\|\Delta u_n\| + h\|\Delta x_n\|) + h^2\delta + \varepsilon\theta\Big). \end{aligned} \tag{4.31}$$

Beweis. Weil der Beweis analog zu entsprechenden Teilen der Beweise von Lemma 4, Lemma 5 und Lemma 6 geführt wird, werden nur die wesentlichen Punkte nochmals zusammenfassend dargestellt. Das Runge-Kutta-Verfahren im Schritt $n+1$ wird formuliert als

$$y_{n+1} = y_n + h\dot{y}_n + h^2\sum_{i,j=1}^{s} b_i a_{ij}\ddot{Y}_{nj}, \quad \dot{y}_{n+1} = \dot{y}_n + h\sum_{i=1}^{s} b_i\ddot{Y}_{ni}$$

mit inneren Stufen

$$Y_{ni} = y_n + c_i h\dot{y}_n + h^2\sum_{j,k=1}^{s} a_{ij}a_{jk}\ddot{Y}_{nk}, \quad \dot{Y}_{ni} = \dot{y}_n + h\sum_{j=1}^{s} a_{ij}\ddot{Y}_{nj}$$

die für $i = 1, \ldots, s$

$$M(Y_{ni})\ddot{Y}_{ni} = f(Y_{ni}, \dot{Y}_{ni}) - \frac{1}{\varepsilon} D(Y_{ni})\dot{Y}_{ni}$$

erfüllen. Das zugehörige gestörte Verfahren in den Variablen $\widehat{y}_n$, $\widehat{\dot{y}}_n$, $\widehat{Y}_{ni}$, $\widehat{\dot{Y}}_{ni}$, $\widehat{\ddot{Y}}_{ni}$ und einer Störung d_{ni} wird formuliert wie in (4.13). Wir transformieren beide Runge-Kutta-Verfahren wie im Beweis von Lemma 4 beziehungsweise Lemma 5, und mit derselben Argumentation wie im Beweis von Lemma 5 setzen wir

$$\begin{pmatrix} \Delta u_n \\ \Delta x_n \end{pmatrix} = \begin{pmatrix} \Delta y_n \\ T^{-1}(\widehat{y}_n)\Delta \dot{y}_n \end{pmatrix}, \quad \begin{pmatrix} \Delta \dot{U}_{ni} \\ \Delta \dot{X}_{ni} \end{pmatrix} = \begin{pmatrix} \Delta \dot{Y}_{ni} \\ T^{-1}(\widehat{y}_n)\Delta \ddot{Y}_{ni} \end{pmatrix} \tag{4.32}$$

und

$$\begin{aligned} \begin{pmatrix} \Delta U_{ni} \\ \Delta X_{ni} \end{pmatrix} &= \begin{pmatrix} \Delta u_n \\ \Delta x_n \end{pmatrix} + h \sum_{j=1}^{s} a_{ij} \begin{pmatrix} \Delta \dot{U}_{nj} \\ \Delta \dot{X}_{nj} \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 \\ O\big(\|\Delta u_n\| + h\|\Delta x_n\| + h^2(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)\big) \end{pmatrix}, \end{aligned} \tag{4.33}$$

mit $\Delta \dot{U} = (\Delta \dot{U}_{n1}, \ldots, \Delta \dot{U}_{ns})^T$ und $\Delta \dot{X} = (\Delta \dot{X}_{n1}, \ldots, \Delta \dot{X}_{ns})^T$. Via Subtraktion der beiden transformierten Runge-Kutta-Verfahren und mit (4.32) und (4.33) ergibt sich ein Runge-Kutta-Verfahren analog zu (4.20)

$$\Delta u_{n+1} = \Delta u_n + h \sum_{i=1}^{s} b_i \Delta \dot{U}_{nj}, \quad \Delta x_{n+1} = \Delta x_n + h \sum_{i=1}^{s} b_i \Delta \dot{X}_{ni} \tag{4.34a}$$

mit inneren Stufen

$$\Delta U_{ni} = \Delta u_n + h \sum_{j=1}^{s} a_{ij} \Delta \dot{U}_{nj}, \quad \Delta X_{ni} = x_n + h \sum_{j=1}^{s} a_{ij} \Delta \dot{X}_{nj} \tag{4.34b}$$

wobei für diese

$$\begin{aligned} \Delta \dot{U}_{ni} &= \widetilde{F}(U_{ni}, X_{ni}) - \widetilde{F}(\widehat{U}_{ni}, \widehat{X}_{ni}) - \delta_{ni}, \\ \Delta \dot{X}_{ni} &= -\frac{1}{\varepsilon}(A(U_{ni})X_{ni} - A(\widehat{U}_{ni})\widehat{X}_{ni}) \\ &\quad + \widetilde{\varphi}(U_{ni}, X_{ni}) - \widetilde{\varphi}(\widehat{U}_{ni}, \widehat{X}_{ni}) - \theta_{ni} \end{aligned} \tag{4.34c}$$

gilt. Mit den Voraussetzungen (4.30a) und transformierten Variablen gemäß (4.32) und (4.33) fassen wir die Funktionen auf der rechten Seite über Taylor-Entwicklungen zusammen (vergleiche (4.20c), (4.21), (4.22)) und gelangen so zu einer Formulierung

$$\begin{aligned} \Delta \dot{U}_{ni} &= O\Big(\|\Delta u_n\| + \|\Delta x_n\| + h(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|)\Big) - \delta_{ni}, \\ \Delta \dot{X}_{ni} + \frac{1}{\varepsilon} A(\widehat{U}_{ni})\Delta X_{ni} &= O\Big(\frac{h}{\varepsilon}(\|\Delta u_n\| + \|\Delta x_n\| + h(\|\Delta \dot{U}\| + \|\Delta \dot{X}\|))\Big) - \theta_{ni}, \end{aligned}$$

aus der sich mit $\delta_n = \max_i \|\delta_{ni}\|$, $\theta_n = \max_i \|\theta_{ni}\|$ und für $0 < \varepsilon \leq h < h_0$ eine Abschätzung wie (4.14) ergibt:

$$\begin{aligned} \|\Delta \dot{U}_{ni}\| &\leq C\Big(\|\Delta u_n\| + \|\Delta x_n\| + \delta_n + \varepsilon\theta_n\Big), \\ \|\Delta \dot{X}_{ni}\| &\leq C\Big(\frac{1}{h}(\|\Delta u_n\| + \|\Delta x_n\|) + h\delta_n + \frac{\varepsilon}{h}\theta_n\Big). \end{aligned}$$

Analoge Argumentation zu Lemma 6 liefert schließlich die Gleichungen aus der Behauptung. ∎

4.4 Fehlerfortpflanzung

Mit der Folgerung von Abschnitt 4.3 können wir nun ein Lemma über die Fehlerfortpflanzung formulieren.

Lemma 7. *Falls das Runge-Kutta-Verfahren die Voraussetzungen (4.30a) - (4.30f) und die Stabilitätsbedingungen (3.7) sowie (3.8) erfüllt, existiert die Runge-Kutta-Lösung $(y_n, \dot{y}_n)$ für $0 \leq nh \leq T$, $0 < \varepsilon < h < h_0$, und für die transfromierten Variablen gilt*

$$\begin{aligned} \|\Delta u_n\| &\leq \|\Delta u_0\| + C\Big(h\|\Delta x_0\| + \delta + \varepsilon\theta\Big), \\ \|\Delta x_n\| &\leq C\Big(\|\Delta u_0\| + (\rho^n + h)\|\Delta x_0\| + \delta + \varepsilon\theta\Big). \end{aligned}$$

Beweis. Die Existenz der Runge-Kutta-Lösung folgt mit einem induktiven Argument. Dies zeigen wir im Anschluss an Theorem 13. Aus der Folgerung zu Lemma 6 ist bekannt, dass nach einem Schritt des Runge-Kutta-Verfahrens für den Fehler

$$\begin{aligned} \|\Delta u_{n+1}\| &\leq \|\Delta u_n\| + O\Big(h(\|\Delta u_n\| + \|\Delta x_n\|) + h\delta + h\varepsilon\theta\Big), \\ \|\Delta x_{n+1}\| &\leq \|R\Big(-\frac{h}{\varepsilon}A_n\Big)\|\,\|\Delta x_n\| + O\Big(\|\Delta u_n\| + h\|\Delta x_n\| + h^2\delta + \varepsilon\theta\Big) \end{aligned}$$

gilt. Zunächst ist $\|R(-\frac{h}{\varepsilon}A_n)\|$ abzuschätzen. Weil A_n symmetrisch und positiv definit ist, gibt es eine orthogonale Matrix $S \in \mathbb{R}^{m\times m}$ mit $S^T A_n S = \Lambda$, wobei die Diagonalmatrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_l) \in \mathbb{R}^{m\times m}$ die Eigenwerte λ_i von A_n enthält, die alle positiv sind. Wegen Voraussetzung (3.8) ist folglich

$$\begin{aligned} \|R\Big(-\frac{h}{\varepsilon}A_n\Big)\| &= \|S^T R\Big(-\frac{h}{\varepsilon}\Lambda\Big)S\| = \|R\Big(-\frac{h}{\varepsilon}\Lambda\Big)\| \\ &= \max_{i=1,\ldots,l} \|R\Big(-\frac{h}{\varepsilon}\lambda_i\Big)\| \leq \rho < 1. \end{aligned}$$

Als Konsequenz hiervon ergibt sich

$$\begin{pmatrix} \|\Delta u_{n+1}\| \\ \|\Delta x_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1+O(h) & O(h) \\ O(1) & \rho+O(h) \end{pmatrix} \begin{pmatrix} \|\Delta u_n\| \\ \|\Delta x_n\| \end{pmatrix} + \begin{pmatrix} O(h\delta+h\varepsilon\theta) \\ O(h^2\delta+\varepsilon\theta) \end{pmatrix}.$$

Um $(\|\Delta u_n\|, \|\Delta x_n\|)^T$ abzuschätzen, transformieren wir die $(2\times2)-$Matrix dieses Systems in Diagonalgestalt. Dies lässt sich durchführen, denn die Eigenwerte sind ohne Einschränkung wegen $\rho \neq 1$ verschieden und lassen sich auf elementare Art und Weise als $\mu_1 = 1+O(h)$ und $\mu_2 = \rho+O(h)$ bestimmen. Mit einer Transformationsmatrix X der Gestalt

$$X = \begin{pmatrix} 1 & O(h) \\ O(1) & 1 \end{pmatrix}$$

ist

$$\begin{pmatrix} 1+O(h) & O(h) \\ O(1) & \rho+O(h) \end{pmatrix} = X^{-1} \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} X.$$

Hiermit haben wir

$$\begin{pmatrix} \|\Delta u_n\| \\ \|\Delta x_n\| \end{pmatrix} \leq X^{-1} \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} X \begin{pmatrix} \|\Delta u_{n-1}\| \\ \|\Delta x_{n-1}\| \end{pmatrix} + \begin{pmatrix} O(h\delta+h\varepsilon\theta) \\ O(h^2\delta+\varepsilon\theta) \end{pmatrix}$$

und weil für $(\|\Delta u_{n-1}\|, \|\Delta x_{n-1}\|)^T$ die analoge Abschätzung gilt, ist

$$\begin{aligned} \begin{pmatrix} \|\Delta u_n\| \\ \|\Delta x_n\| \end{pmatrix} \leq\ & X^{-1} \begin{pmatrix} \mu_1^2 & 0 \\ 0 & \mu_2^2 \end{pmatrix} X \begin{pmatrix} \|\Delta u_{n-2}\| \\ \|\Delta x_{n-2}\| \end{pmatrix} \\ & + X^{-1} \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} X \begin{pmatrix} O(h\delta+h\varepsilon\theta) \\ O(h^2\delta+\varepsilon\theta) \end{pmatrix} + \begin{pmatrix} O(h\delta+h\varepsilon\theta) \\ O(h^2\delta+\varepsilon\theta) \end{pmatrix}. \end{aligned}$$

Mutatis mutandis erhalten wir

$$\begin{aligned} \begin{pmatrix} \|\Delta u_n\| \\ \|\Delta x_n\| \end{pmatrix} \leq\ & X^{-1} \begin{pmatrix} \mu_1^n & 0 \\ 0 & \mu_2^n \end{pmatrix} X \begin{pmatrix} \|\Delta u_0\| \\ \|\Delta x_0\| \end{pmatrix} \\ & + \sum_{j=1}^{n} X^{-1} \begin{pmatrix} \mu_1^{n-j} & 0 \\ 0 & \mu_2^{n-j} \end{pmatrix} X \begin{pmatrix} O(h\delta+h\varepsilon\theta) \\ O(h^2\delta+\varepsilon\theta) \end{pmatrix}. \end{aligned}$$

Mit den bereits berechneten Eigenwerten μ_1 und μ_2 ist

$$X^{-1} \begin{pmatrix} \mu_1^n & 0 \\ 0 & \mu_2^n \end{pmatrix} X \begin{pmatrix} \|\Delta u_0\| \\ \|\Delta x_0\| \end{pmatrix} = \begin{pmatrix} \|\Delta u_0\| + O(h\|\Delta x_0\|) \\ O(\|\Delta u_0\|) + (\rho^n + O(h))\|\Delta x_0\| \end{pmatrix},$$

und die Restterme ergeben sich durch

$$\begin{aligned}
& \sum_{j=1}^{n} X^{-1} \begin{pmatrix} \mu_1^{n-j} & 0 \\ 0 & \mu_2^{n-j} \end{pmatrix} X \begin{pmatrix} O(h\delta + h\varepsilon\theta) \\ O(h^2\delta + \varepsilon\theta) \end{pmatrix} \\
= & \sum_{j=1}^{n} \begin{pmatrix} 1 + O(h) & O(h) \\ O(1) & \rho^{n-j} + O(h) \end{pmatrix} \begin{pmatrix} O(h\delta + h\varepsilon\theta) \\ O(h^2\delta + \varepsilon\theta) \end{pmatrix} \\
= & \begin{pmatrix} nO(h\delta + h\varepsilon\theta) \\ nO(h\delta + h\varepsilon\theta) + \sum_{j=1}^{n} \rho^{n-j} O(h^2\delta + \varepsilon\theta) \end{pmatrix} \\
\leq & \begin{pmatrix} \frac{T}{h} O(h\delta + h\varepsilon\theta) \\ \frac{T}{h} O(h\delta + h\varepsilon\theta) + O(h^2\delta + \varepsilon\theta) \end{pmatrix} \\
= & \begin{pmatrix} O(\delta + \varepsilon\theta) \\ O(\delta + \varepsilon\theta) \end{pmatrix},
\end{aligned}$$

denn $n \leq \frac{T}{h}$ und $\sum_{j=1}^{n} \rho^{n-j} = O(1)$ für $\rho < 1$. Somit folgt die Behauptung. ∎

Kapitel 5

Konvergenzresultate

Nachdem nun die wesentlichen technischen Details aus Kapitel 4 zur Verfügung stehen, sind wir in der Lage die Hauptresultate der Arbeit - globale Fehlerschranken für die numerischen Lösungen von (1.5) - zu formulieren und zu beweisen. Da wir diese Schranken für die ε-Entwicklung $(y^\varepsilon, \dot{y}^\varepsilon)$ formulieren, ist zunächst noch zu zeigen, dass auch die numerische Lösung $(y_n, \dot{y}_n)$ in Analogie zur analytischen Lösung eine ε-Entwicklung besitzt.

5.1 Asymptotische ε-Entwicklung der numerischen Lösung

Theorem 10. *Sei ein Runge-Kutta-Verfahren (3.3) mit Stufenordnung q gegeben, das den Stabilitätsbedingungen (3.7) und (3.8) genügt. Weiter sei der Anfangswert $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ auf der Mannigfaltigkeit $\mathcal{M}^\varepsilon$ aus Theorem 1, das heißt die exakte Lösung von (1.5) mit Startwerten $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ ist glatt. Dann besitzt das stark gedämpfte mechanische System (1.5) für $0 < \varepsilon < h < h_0$ eine eindeutig bestimmte Runge-Kutta-Lösung $(y_n^\varepsilon, \dot{y}_n^\varepsilon)$ gemäß (3.3). Die Approximationen $(y_n^\varepsilon, \dot{y}_n^\varepsilon)$ lassen sich für $nh \in [0, T]$ als ε-Entwicklungen*

$$\begin{aligned} y_n^\varepsilon &= y_n^0 + \varepsilon y_n^1 + \cdots + \varepsilon^q y_n^q + O(\varepsilon^{q+1}), \\ \dot{y}_n^\varepsilon &= \dot{y}_n^0 + \varepsilon \dot{y}_n^1 + \cdots + \varepsilon^q \dot{y}_n^q + O(\varepsilon^{q+1}) \end{aligned} \tag{5.1}$$

darstellen, wobei y_n^k, $\dot{y}_n^k$ für $k = 0, \ldots, q$ Runge-Kutta-Lösungen der differential-algebraischen Systeme (DAE 0) - (DAE k) sind. Die Anfangswerte $(y_0^k, \dot{y}_0^k)$ für die Runge-Kutta-Verfahren zu diesen differential-algebraischen Systemen werden als Koeffizienten von ε^k in der ε-Entwicklung von $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ gewählt.

Beweis. Die Behauptung folgt aus zwei Beweisschritten. Zunächst berechnen wir analog zum analytischen Fall den Defekt, der entsteht, wenn abgebrochene ε-Entwicklungen der Runge-Kutta-Lösung in das Runge-Kutta-Verfahren (3.3) eingesetzt werden. Im zweiten Schritt folgt dann die Behauptung des Theorems durch Anwendung von Lemma 7.

(a) Wir betrachten also abgebrochene ε-Entwicklungen

$$\begin{aligned} \widehat{y}_n &= y_n^0 + \varepsilon y_n^1 + \cdots + \varepsilon^q y_n^q, \\ \widehat{\dot{y}}_n &= \dot{y}_n^0 + \varepsilon \dot{y}_n^1 + \cdots + \varepsilon^q \dot{y}_n^q \end{aligned} \tag{5.2}$$

der numerischen Lösung sowie abgebrochene Entwicklungen für die inneren Stufen

$$\begin{aligned} \widehat{Y}_{ni} &= Y_{ni}^0 + \varepsilon Y_{ni}^1 + \cdots + \varepsilon^q Y_{ni}^q, \\ \widehat{\dot{Y}}_{ni} &= \dot{Y}_{ni}^0 + \varepsilon \dot{Y}_{ni}^1 + \cdots + \varepsilon^q \dot{Y}_{ni}^q, \\ \widehat{\ddot{Y}}_{ni} &= \ddot{Y}_{ni}^0 + \varepsilon \ddot{Y}_{ni}^1 + \cdots + \varepsilon^q \ddot{Y}_{ni}^q, \end{aligned}$$

wobei $y_n^k,\ \dot{y}_n^k, Y_{ni}^k,\ \dot{Y}_{ni}^k,\ \ddot{Y}_{ni}^k$ Koeffizienten aus dem Runge-Kutta-Verfahren für das System (DAE 0) - (DAE k) sind. Wir setzen diese Entwicklungen in (3.3a) beziehungsweise (3.3b) ein und vergleichen das Resultat für $k = 0, \ldots, q$ nach Koeffizienten ε^k mit ε-Entwicklungen für den nächsten Zeitschritt, die (5.2) entsprechen. Wegen der Linearität in (3.3a) und (3.3b) folgt sofort, dass in diesen Runge-Kutta-Gleichungen kein Defekt auftritt. Allerdings ergibt sich ein Defekt d_{ni} durch Einsetzen in (3.3c). Diese Bedingung für die inneren Stufen ist hier wieder gegeben durch

$$M(\widehat{Y}_{ni})\widehat{\ddot{Y}}_{ni} = f(\widehat{Y}_{ni}, \widehat{\dot{Y}}_{ni}) - \frac{1}{\varepsilon} D(\widehat{Y}_{ni})\dot{Y}_{ni} + d_{ni}. \tag{5.3}$$

Der Defekt d_{ni} wird im nachfolgenden Lemma berechnet.

(b) Nun lassen sich die Defekte in (5.1) durch Anwendung von Lemma 7 berechnen. Da die Anfangswerte $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ nach Voraussetzung auf der Mannigfaltigkeit $\mathcal{M}^\varepsilon$ liegen, folgt sofort, dass

$$y_0^\varepsilon - \widehat{y}_0 = O(\varepsilon^{q+1}), \quad \dot{y}_0^\varepsilon - \widehat{\dot{y}}_0 = O(\varepsilon^{q+1})$$

gilt. Die Restterme in (5.1) ergeben sich nun durch Betrachtung von $r_n = y_n^\varepsilon - \widehat{y}_n$ beziehungsweise $\dot{r}_n = \dot{y}_n^\varepsilon - \widehat{\dot{y}}_n$. Wir wenden Lemma 7 an; r_n und $\dot{r}_n$ sind in der Rolle von Δy_n beziehungsweise $\Delta \dot{y}_n$. Für Stufenordnung q erhalten wir mit der Kenntnis von d_{ni} aus Lemma 8 sowie (4.30d) und (4.30e)

$$\delta = O(\varepsilon^{q+1}), \quad \theta = O(\varepsilon^q).$$

Also folgt

$$r_n \le C\varepsilon^{q+1}, \quad \dot{r}_n \le C\varepsilon^{q+1}$$

und damit die Behauptung des Theorems. ∎

Lemma 8 (Defekt der abgebrochenen Entwicklungen). *Der Defekt d_{ni} aus (5.3) ist gegeben durch*

$$d_{ni} = \varepsilon^q S(Y_{ni}^0) G^T \Lambda_{ni}^q + O(\varepsilon^{q+1}),$$

wobei $S(Y_{ni}^0) = S(\widehat{Y}_{ni}) + O(\varepsilon)$ gilt. S ist definiert wie in Lemma 1 und G^T ist wieder durch $G = [0\ I_m]$ erklärt. Λ_{ni}^k bezeichnet die Runge-Kutta-Approximation an $\lambda^k(t_n + c_i h)$ aus (DAE k') und ist für $k \le q$ von der Größenordnung

$$\Lambda_{ni}^k = O(1).$$

Beweis. Wie schon gesehen sind die Runge-Kutta-Gleichungen (3.3a) und (3.3b) linear mit Koeffizienten, die nicht von ε abhängen, es tritt also kein Defekt auf. Desweiteren ist (3.3c) von derselben Form wie die entsprechende Differentialgleichung (1.5) beziehungsweise die differntial-algebraischen Systeme (DAE 0') - (DAE k') und (DAE 0) - (DAE k). Statt (5.3) betrachten wir also

$$\begin{aligned} \widehat{\dot{Y}}_{ni} &= S^{-T}(\widehat{Y}_{ni})\widehat{Z}_{ni}, \\ \widetilde{M}(\widehat{Y}_{ni})\widehat{\dot{Z}}_{ni} &= \widetilde{f}(\widehat{Y}_{ni}, \widehat{Z}_{ni}) - \frac{1}{\varepsilon}\begin{pmatrix} 0 & 0 \\ 0 & I_m \end{pmatrix}\widehat{Z}_{ni} + S^{-1}(\widehat{Y}_{ni}) d_{ni} \end{aligned}$$

mit Bezeichnungen gemäß Lemma 1. Wir berechnen Taylor-Entwicklungen von $\widetilde{M}$ und $\widetilde{f}$ um Y_{ni}^0 beziehungsweise (Y_{ni}^0, Z_{ni}^0) und setzen die abgebrochenen ε-Entwicklungen ein. Ordnen der Koeffizienten nach Potenzen von ε^k für $k = 0, \ldots, q$ liefert aufgrund der Runge-Kutta-Relation (3.2d) für (DAE 0') - (DAE k') die Gleichung

$$S^{-1}(\widehat{Y}_{ni}) d_{ni} = \varepsilon^q G^T \Lambda_{ni}^q + O(\varepsilon^{q+1}).$$

Dies ist dieselbe Verfahrensweise, wie sie im Beweis von Theorem 1 zur Konstruktion der Koeffizienten der ε-Entwicklungen angewandt wurde. Mit $S(\widehat{Y}_{ni}) = S(Y_{ni}^0) + O(\varepsilon)$ gilt also die erste Behauptung. Ganz analog wie im Beweis von Theorem 9 berechnen wir für $k = 0$ und $k = 1$ die Schranken

$$\Lambda_{ni}^0 - \lambda^0(t_n + c_i h) = O(h^q), \quad \Lambda_{ni}^1 - \lambda^1(t_n + c_i h) = O(h^{q-1})$$

und induktiv folgt $\Lambda_{ni}^k - \lambda^k(t_n + c_i h) = O(h^{q-k})$. Für $k \le q$ ist Λ_{ni}^k also durch eine Konstante beschränkt. ∎

Bemerkung. In Abbildung 5.1 sind die Beziehungen zwischen den wichtigsten Variablen des analytischen Teils und ihren entsprechenden Runge-Kutta-Approximationen in einem Diagramm dargestellt.

5.2 Globale Fehlerabschätzungen

Nun lässt sich mit Theorem 10 und den Fehlerabschätzungen für die differential-algebraischen Gleichungssysteme (DAE 0,...,DAE k) aus Theorem 9 das folgende Konvergenzresultat beweisen.

Theorem 11. *Sei ein Runge-Kutta-Verfahren (3.3) mit Stufenordnung q gegeben, das den Stabilitätsbedingungen (3.7) und (3.8) genügt. Weiter sei der Anfangswert $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ auf der Mannigfaltigkeit $\mathcal{M}^\varepsilon$ aus Theorem 1, das heißt die exakte Lösung von (1.5) mit Startwerten $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ ist glatt. Dann besitzt das stark gedämpfte mechanische System (1.5) für $0 < \varepsilon < h < h_0$ eine eindeutig bestimmte Runge-Kutta-Lösung $(y_n^\varepsilon, \dot{y}_n^\varepsilon)$ gemäß (3.3) und für den globalen Fehler gilt*

$$\begin{aligned} y_n^\varepsilon - y^\varepsilon(t_n) &= y_n^0 - y^0(t_n) + O(\varepsilon h^q), \\ \dot{y}_n^\varepsilon - \dot{y}^\varepsilon(t_n) &= \dot{y}_n^0 - \dot{y}^0(t_n) + O(\varepsilon h^q) \end{aligned} \tag{5.4}$$

gleichmäßig für $\varepsilon \le h \le h_0$ und $0 \le t_n \le T$. Hierbei sind y_n^0, $\dot{y}_n^0$ und $y^0(t)$, $\dot{y}^0(t)$ die Runge-Kutta-Lösungen beziehungsweise exakten Lösungen der differential-algebraischen Gleichung (DAE 0) vom Index 2, wobei die Anfangswerte $(y_0^0, \dot{y}_0^0)$ aus der ε-Entwicklung von $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ gegeben sind.

Beweis. Mit ε-Entwicklungen von $(y^\varepsilon(t), \dot{y}^\varepsilon(t))$ sowie $(y_n^\varepsilon, \dot{y}_n^\varepsilon)$ aus den Theoremen 1 und 10 gilt

$$\begin{aligned} y_n^\varepsilon - y^\varepsilon(t_n) &= (y_n^0 - y^0(t_n)) + \varepsilon(y_n^1 - y^1(t_n)) + \dots \\ &\quad + \varepsilon^q(y_n^q - y^q(t_n)) + O(\varepsilon^{q+1}), \\ \dot{y}_n^\varepsilon - \dot{y}^\varepsilon(t_n) &= (\dot{y}_n^0 - \dot{y}^0(t_n)) + \varepsilon(\dot{y}_n^1 - \dot{y}^1(t_n)) + \dots \\ &\quad + \varepsilon^q(\dot{y}_n^q - \dot{y}^q(t_n)) + O(h\varepsilon^q). \end{aligned} \tag{5.5}$$

Mit Theorem 9 gilt für $k = 1, \dots, q$

$$\begin{aligned} \varepsilon^k(y_n^k - y^k(t_n)) &= O(\varepsilon^k h^{q+1-k}), \\ \varepsilon^k(\dot{y}_n^k - \dot{y}^k(t_n)) &= O(\varepsilon^k h^{q+1-k}). \end{aligned}$$

Für $\varepsilon \le h$ sind daher alle Summanden in (5.5) sowohl in den Positions- als auch Geschwindigkeitstermen durch $O(\varepsilon h^q)$ beschränkt. Da noch das Restglied $O(\varepsilon^{q+1})$ auftritt, folgt die Behauptung. ∎

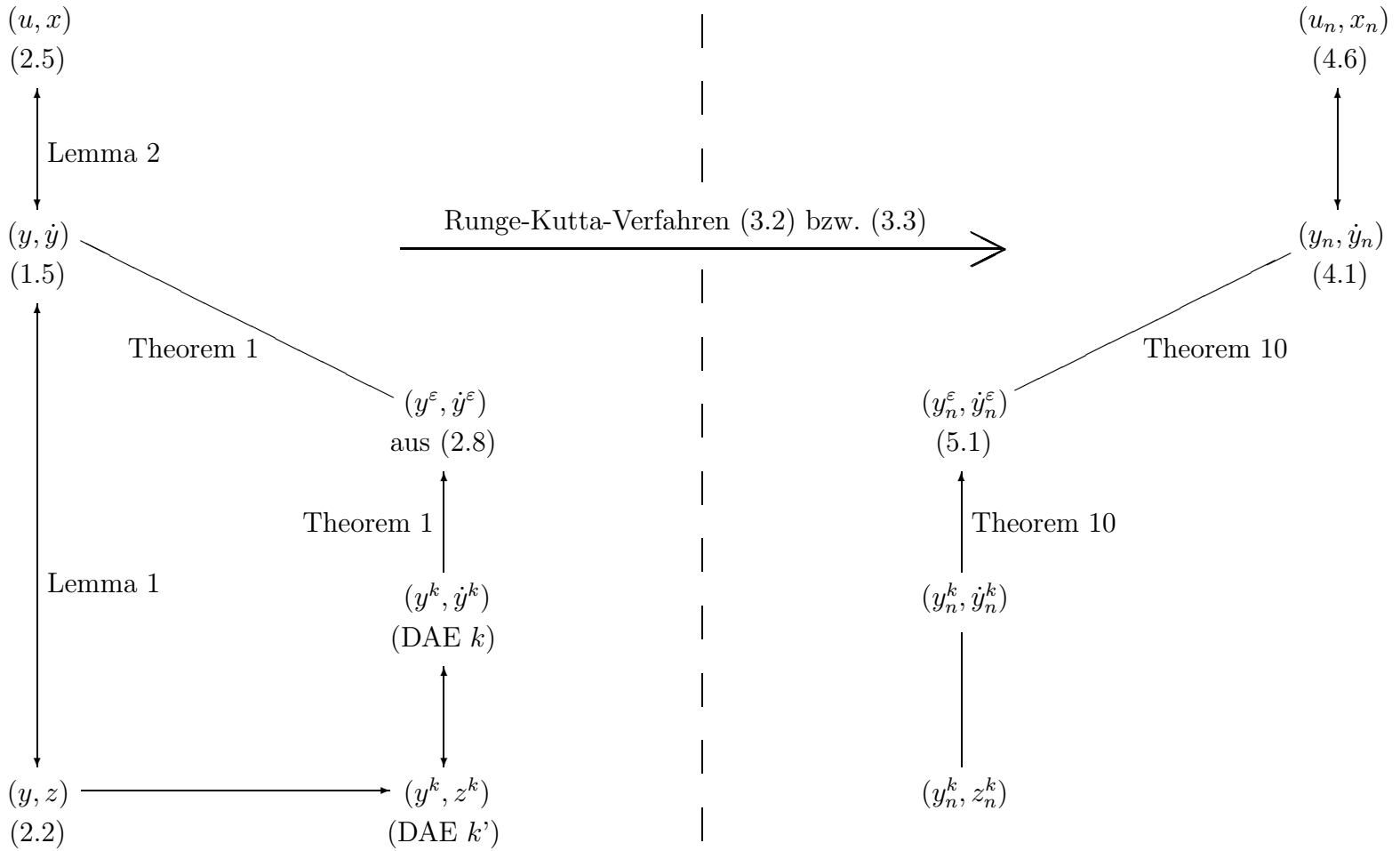

Abbildung 5.1: Zuammenhang der Variablen im Kontinuierlichen und Übergang zu diskreten Lösungen durch die geeigneten Runge-Kutta-Verfahren

Theorem 12. *Mit den Voraussetzungen von Theorem 11 gilt für s-stufige RadauIIA-Verfahren für den globalen Fehler*

$$(y_n^\varepsilon, \dot{y}_n^\varepsilon) - (y^\varepsilon(t_n), \dot{y}^\varepsilon(t_n)) = O(h^{2s-1}) + O(\varepsilon h^s)$$

Beweis. Für RadauIIA-Verfahren sind Ordnung und Stufenordnung durch die natürlichen Zahlen $p = 2s-1$ beziehungsweise $q = s$ gegeben (siehe [9], S. 73). RadauIIA-Verfahren erfüllen auch alle Voraussetzungen aus der Bemerkung zu Theorem 8, es gilt also

$$(y_n^0, \dot{y}_n^0) - (y^0(t_n), \dot{y}^0(t_n)) = O(h^{2s-1}).$$

Aus (5.4) folgt die Behauptung. ∎

5.3 Existenz einer attraktiven invarianten Mannigfaltigkeit für die numerischen Lösungen

Die Anfangswerte $(y_0^\varepsilon, \dot{y}_0^\varepsilon)$ liegen nur in Ausnahmefällen auf der Mannigfaltigkeit $\mathcal{M}^\varepsilon$. Eher zu erwarten sind Anfangswerte, die nicht auf $\mathcal{M}^\varepsilon$, aber genügend nahe an dieser Mannigfaltigkeit liegen.

Eine ähnliche Situation haben wir bereits im kontinuierlichen Fall betrachtet. In Theorem 3 konnten wir beobachten, dass für Anfangswerte, die $O(\varepsilon)$ - nahe an der Mannigfaltigkeit $\mathcal{M}^0$ aus Voraussetzung (1.7) liegen, eine attraktive invariante Mannigfaligkeit $\mathcal{N}^\varepsilon$ existiert, an die sich die kontinuierlichen Lösungen mit exponentieller Geschwindigkeit annähern. Zudem stimmt $\mathcal{N}^\varepsilon$ für beliebiges N bis auf $O(\varepsilon^N)$ mit $\mathcal{M}^\varepsilon$ überein. Ein diskretes Analogon hierzu zeigen wir in Theorem 13. Um dieses Resultat zu beweisen, benötigen wir ebenfalls Theorem 2.

Theorem 13. *Sei ein Runge-Kutta-Verfahren (3.3) mit Stufenordnung q gegeben, das den Stabilitätsbedingungen (3.7) und (3.8) genügt. Falls der Anfangswert $(y_0, \dot{y}_0)$ die Voraussetzung $D(y_0)\dot{y}_0 = O(h)$ erfüllt, existiert ein $(y_0^\varepsilon, \dot{y}_0^\varepsilon) \in \mathcal{M}^\varepsilon$, so dass die Runge-Kutta-Lösungen $(y_n, \dot{y}_n)$ und $(y_n^\varepsilon, \dot{y}_n^\varepsilon)$ zu den entsprechenden Anfangswerten für $0 < \varepsilon < h < h_0$ und $0 \le t_n \le T$*

$$(y_n, \dot{y}_n) - (y_n^\varepsilon, \dot{y}_n^\varepsilon) = O(h\rho^n + \varepsilon^{q+1}) \tag{5.6}$$

mit $\rho < 1$ unabhängig von h, ε und n erfüllen.

Beweis. Der Beweis verläuft im Wesentlichen analog zum Beweis von Theorem 3.2, [13]. Der vorgegebene Startwert sei mit $(\bar{y}_0, \bar{\dot{y}}_0)$ bezeichnet, um ihn von anderen

Startwerten zu unterscheiden, die im Laufe des Beweises auftreten. Nach Voraussetzung gilt

$$D(\bar{y}_0)\bar{\dot{y}}_0 = O(h), \tag{5.7}$$

der Anfangswert $(\bar{y}_0, \bar{\dot{y}}_0)$ ist daher $O(h)$ - nahe an der Mannigfaltigkeit $\mathcal{M}^\varepsilon$. Es gibt also ein $(y(0), \dot{y}(0)) \in \mathcal{M}^\varepsilon$, so dass $(y(0), \dot{y}(0)) - (\bar{y}_0, \bar{\dot{y}}_0) = O(h)$ gilt. Wir benutzen $(y(0), \dot{y}(0))$ als Anfangswert einer glatten Lösung von (1.5) und betrachten nun Runge-Kutta-Verfahren mit Näherungslösungen $(y_n, \dot{y}_n)$ zu Anfangswerten $(y_0, \dot{y}_0)$, die (5.7) erfüllen und die $O(h)$ - nahe an $(\bar{y}_0, \bar{\dot{y}}_0)$ liegen. Aus Lemma 7 und Theorem 11 ist bekannt, dass solche Runge-Kutta-Lösungen

$$y_n - y(t_n) = O(h), \quad \dot{y}_n - \dot{y}(t_n) = O(h) \tag{5.8}$$

gleichmäßig für $\varepsilon \le h \le h_0$ und $0 \le t_n \le T$ erfüllen. Im Hinblick hierauf interpretieren wir die exakte Lösung $(y(t_n), \dot{y}(t_n))$ nun als Störung der Runge-Kutta-Lösungen $(y_n, \dot{y}_n)$ und befinden uns somit in der Situation von Lemma 6 beziehungsweise der anschließenden Folgerung. Deshalb führen wir wie in Lemma 5 transformierte Variablen

$$\begin{pmatrix} u_n \\ x_n \end{pmatrix} = \begin{pmatrix} y_n \\ T^{-1}(y(t_n))\dot{y}_n \end{pmatrix}$$

ein, wobei $T(y) = (M^{-1/2}\widehat{Q})(y)$ wie in (4.15) gegeben ist. In diesen Variablen lässt sich das Runge-Kutta-Verfahren für (1.5), also (4.6), als eine Rekursion der Form (2.32) schreiben. Wir formulieren das Runge-Kutta-Verfahren (4.6) im Schritt $n+1$ als

$$\begin{aligned} u_{n+1} &= u_n + hb^T \widetilde{F}_n(U_n(t_n, u_n, x_n), X_n(t_n, u_n, x_n)), \\ x_{n+1} &= x_n + hb^T\Big(-\frac{1}{\varepsilon}\{A_n\}X_n(t_n, u_n, x_n) + \\ &\qquad\qquad \widetilde{\varphi}_n\big(U_n(t_n, u_n, x_n), X_n(t_n, u_n, x_n)\big)\Big). \end{aligned} \tag{5.9}$$

Hierbei interpretieren wir die inneren Stufen, die in $U_n = (U_{n1}, \ldots, U_{ns})^T$ und $X_n = (X_{n1}, \ldots, X_{ns})^T$ zusammengefasst sind, als Funktionen, die von t_n, u_n sowie x_n abhängen. Weiter sind

$$\begin{aligned} \widetilde{F}_n(U_n, X_n) &:= (\widetilde{F}(U_{n1}, X_{n1}), \ldots, \widetilde{F}(U_{ns}, X_{ns}))^T, \\ \widetilde{\varphi}_n(U_n, X_n) &:= (\widetilde{\varphi}(U_{n1}, X_{n1}), \ldots, \widetilde{\varphi}(U_{ns}, X_{ns}))^T, \\ \{A_n\} &:= \text{blockdiag}(A(U_{n1}), \ldots, A(U_{ns})), \\ b &:= (b_1, \ldots, b_s)^T. \end{aligned}$$

Die Gleichungen (5.9) beschreiben ein nichtautonomes System. Deshalb setzen wir

$$\begin{aligned} &\xi_n = (t_n, u_n), \quad \eta_n = x_n, \\ &\mathcal{F} = hb^T\widetilde{F}_n, \\ &\mathcal{G} = \eta_n + hb^T\Big(-\frac{1}{\varepsilon}\{A_n\}X_n(\xi_n, \eta_n) + \widetilde{\varphi}_n\Big) \end{aligned}$$

und somit ist (5.9) eine Rekursion der Form (2.32). Die Abschätzungen für die Lipschitzkonstanten ergeben sich hier wesentlich einfacher als im kontinuierlichen Fall. Für Runge-Kutta-Lösungen (u_n, x_n) und $(\widehat{u}_n, \widehat{x}_n)$ gilt mit der Definition der Lipschitzstetigkeit

$$\begin{aligned}
\|\xi_{n+1} - \widehat{\xi}_{n+1}\| &\leq \|\xi_n - \widehat{\xi}_n\| + \|\mathcal{F}(\xi_n, \eta_n) - \mathcal{F}(\widehat{\xi}_n, \widehat{\eta}_n)\| \\
&\leq \|\xi_n - \widehat{\xi}_n\| + L_{\xi\xi}\|\xi_n - \widehat{\xi}_n\| + L_{\xi\eta}\|\eta_n - \widehat{\eta}_n\|, \\
\|\eta_{n+1} - \widehat{\eta}_{n+1}\| &= \|\mathcal{G}(\xi_n, \eta_n) - \mathcal{G}(\widehat{\xi}_n, \widehat{\eta}_n)\| \\
&\leq L_{\eta\xi}\|\xi_n - \widehat{\xi}_n\| + L_{\eta\eta}\|\eta_n - \widehat{\eta}_n\|.
\end{aligned}$$

Andererseits sind wir mit $\Delta u_{n+1} = u_{n+1} - \widehat{u}_{n+1}$ und $\Delta x_{n+1} = x_{n+1} - \widehat{x}_{n+1}$ in der Situation von der Folgerung zu Lemma 6 mit Störungen $\delta = \theta = 0$, mit (4.31) gilt nunmehr

$$\begin{aligned}
\Delta u_{n+1} &= \Delta u_n + O(h\|\Delta u_n\| + h\|\Delta x_n\|), \\
\Delta x_{n+1} &= R\Big(-\frac{h}{\varepsilon}A_n\Big)\,\Delta x_n + O(\|\Delta u_n\|) + O(h\|\Delta x_n\|).
\end{aligned}$$

Hieraus lassen sich die Lipschitzkonstanten direkt ablesen. Wir haben

$$L_{\xi\xi} = O(h),\ L_{\xi\eta} = O(h),\ L_{\eta\xi} = O(1),\ L_{\eta\eta} = \rho_0 + O(h),$$

wobei ρ_0 eine obere Schranke von $\|R(-\frac{h}{\varepsilon}A_n)\|$ bezeichnet. Dieses ρ_0 ergibt sich mit der analogen Argumentation wie in Lemma 7. Zunächst gelten diese Lipschitzkonstanten nur lokal entlang $y(t)$ für $t \in [0, T]$, was aus den Voraussetzungen für die verwendete Folgerung sofort ersichtlich ist. Außerhalb dieser Umgebung können wir $\mathcal{F}$ und $\mathcal{G}$ so modifizieren und fortsetzen, dass (2.33) mit globalen Lipschitzkonstanten gilt. Wir können also Theorem 2 auf (5.9) mit den definierten Größen ξ_n, η_n, $\mathcal{F}$ und $\mathcal{G}$ anwenden. Wegen (i) gilt: Es gibt eine Lipschitzstetige Funktion $s_h : \mathbb{R}^{d_\xi} \to \mathbb{R}^{d_\eta}$, so dass für alle n

$$\eta_0 = s_h(\xi_0) \quad \Rightarrow \quad \eta_n = s_h(\xi_n)$$

gilt. Wegen $\xi = (t, u)$ erhalten wir also für $0 \leq t \leq T$ die $(d{+}l)$-dimensionale Mannigfaltigkeiten $M_h^\varepsilon(t) = \{(\xi, s_h(\xi)) : \xi \in \mathbb{R}^{2d-m+1}\}$. In den Variablen $(y, \dot{y})$ gilt somit für alle n

$$(y_0, \dot{y}_0) \in M_h^\varepsilon(0) \quad \Rightarrow \quad (y_n, \dot{y}_n) \in M_h^\varepsilon(t_n).$$

Aus Teil (iii) von Theorem 2 folgt, dass zu jedem $(y_0, \dot{y}_0)$ ein $(y_0^*, \dot{y}_0^*) \in M_h^\varepsilon(0)$ existiert, so dass die Lösungen $(y_n, \dot{y}_n)$ und $(y_n^*, \dot{y}_n^*)$ von (2.32) zu den entsprechenden Anfangswerten gegeneinander konvergieren:

$$\|(y_n, \dot{y}_n) - (y_n^*, \dot{y}_n^*)\| \leq C_1\rho^n\|(y_0, \dot{y}_0) - (y_0^*, \dot{y}_0^*)\|,$$

wobei $\rho = \rho_0 + O(h) < 1$ ist. Für alle $(y_0, \dot{y}_0)$, die $O(h)$ - nahe an $(\bar{y}_0, \bar{\dot{y}}_0)$ liegen und (5.7) erfüllen, gibt es daher ein $(y_0^*, \dot{y}_0^*) \in M_h^\varepsilon(0)$ für das mit (2.36)

$$(y_n, \dot{y}_n) = (y_n^*, \dot{y}_n^*) + O(\rho^n h) \tag{5.10}$$

gilt.

Um auf das Endresultat zu schließen wird nun auf Theorem 10 zurückgegriffen. Wir betrachten abgebrochene ε-Entwicklungen (5.2), die wir wieder mit $(\widehat{y}_n, \widehat{\dot{y}}_n)$ bezeichnen. Ihre Koeffizienten $(y_n^k, \dot{y}_n^k)$ seien wiederum als Runge-Kutta-Lösungen der differential-algebraischen Systeme (DAE 0) - (DAE q) gegeben, wobei die Anfangswerte $(y_0^k, \dot{y}_0^k)$ aus der ε-Entwicklung von $(y_0^\varepsilon, \dot{y}_0^\varepsilon) \in \mathcal{M}^\varepsilon$ vorgegeben seien, es gilt also

$$(\widehat{y}_0, \widehat{\dot{y}}_0) \in \mathcal{M}^\varepsilon + O(\varepsilon^{q+1}). \tag{5.11}$$

Die Menge $\mathcal{M}_{n,h}^\varepsilon$ aller abgebrochenen ε-Entwicklungen $(\widehat{y}_n, \widehat{\dot{y}}_n)$ ist eine $(d{+}l)$-dimensionale Mannigfaltigkeit. Anwendung von Theorem 11 mit Startwerten $(\widehat{y}_0, \widehat{\dot{y}}_0)$ wie in (5.11) liefert unter Beachtung der Störung $O(\varepsilon^{q+1})$ in Lemma 7

$$(\widehat{y}_n, \widehat{\dot{y}}_n) - (y^\varepsilon(t_n), \dot{y}^\varepsilon(t_n)) = O(\varepsilon h^q),$$

die Mannigfaltigkeiten $\mathcal{M}^\varepsilon$ und $\mathcal{M}_{n,h}^\varepsilon$ liegen also $O(\varepsilon h^q)$ - nahe beieinander. Für $(y_0^\varepsilon, \dot{y}_0^\varepsilon) \in \mathcal{M}^\varepsilon$ folgt aus (5.1) direkt

$$(y_n^\varepsilon, \dot{y}_n^\varepsilon) = (\widehat{y}_n, \widehat{\dot{y}}_n) + O(\varepsilon^{q+1}), \tag{5.12}$$

somit gilt

$$(y_n^\varepsilon, \dot{y}_n^\varepsilon) \in \mathcal{M}_{n,h}^\varepsilon + O(\varepsilon^{q+1}).$$

Weil $(\bar{y}_0, \bar{\dot{y}}_0)$ $O(h)$ - nahe an $\mathcal{M}^\varepsilon$ liegt, gilt mit (5.10) insbesondere für alle Paare $(y_0^\varepsilon, \dot{y}_0^\varepsilon) \in \mathcal{M}^\varepsilon$, die in einer $O(h)$-Umgebung von $(\bar{y}_0, \bar{\dot{y}}_0)$ liegen, dass

$$(y_n^\varepsilon, \dot{y}_n^\varepsilon) = (y_n^*, \dot{y}_n^*) + O(\rho^n h), \tag{5.13}$$

also

$$(y_n^\varepsilon, \dot{y}_n^\varepsilon) \in M_h^\varepsilon(t_n) + O(\rho^n h).$$

Da (5.12) und (5.13) gleichmäßig für $\varepsilon \le h \le h_0$ und $0 \le t_n \le T$ gelten, folgt

$$(\widehat{y}_n, \widehat{\dot{y}}_n) - (y_n^*, \dot{y}_n^*) = O(\rho^n h + \varepsilon^{q+1}).$$

Die Mannigfaltigkeiten $M_h^\varepsilon(t_n)$ und $\mathcal{M}_{n,h}^\varepsilon$ liegen $O(\rho^n h + \varepsilon^{q+1})$ - nahe beieinander.

Der Beweisschluss folgt nun mit demselben Dimensionsargument wie im Beweis von Theorem 3. Weil beide Mannigfaltigkeiten dieselbe Dimension $(d{+}l)$ haben, ist ein Umkehrschluss möglich. Zu jedem Startwert $(y_0^*, \dot{y}_0^*) \in M_h^\varepsilon(0)$ existiert ein

$(y_0^\varepsilon, \dot{y}_0^\varepsilon) \in \mathcal{M}^\varepsilon = \mathcal{M}_{0,h}^\varepsilon + O(\varepsilon^{q+1})$, so dass Runge-Kutta-Lösungen $(y_n^*, \dot{y}_n^*)$ und $(y_n^\varepsilon, \dot{y}_n^\varepsilon)$ zu den entsprechenden Anfangswerten

$$(y_n^*, \dot{y}_n^*) = (y_n^\varepsilon, \dot{y}_n^\varepsilon) + O(\rho^n h + \varepsilon^{q+1})$$

erfüllen. Mit dieser Gleichheit und (5.10) folgt

$$(y_n, \dot{y}_n) - (y_n^\varepsilon, \dot{y}_n^\varepsilon) = O(\rho^n h + \varepsilon^{q+1}).$$

Dies ist die Behauptung des Theorems. ∎

Mit den Theoremen 11 und 13 kann nun der Nachweis für die Existenz von Runge-Kutta-Lösungen $(y_n, \dot{y}_n)$ aus Lemma 7 nachgetragen werden. Aus dem Beweis ergibt sich gleichzeitig die Gültigkeit von (4.30a) für alle $(y_n, \dot{y}_n)$ mit $0 \le nh \le T$.

Folgerung. Falls für die Anfangswerte $(y_0, \dot{y}_0)$ des Runge-Kutta-Verfahrens (4.1) $D(y_0)\dot{y}_0 = O(h)$ gilt und das Runge-Kutta-Verfahren die Voraussetzungen von Lemma 4 erfüllt, so existieren Runge-Kutta-Lösungen $(y_n, \dot{y}_n)$ für alle $0 \le nh \le T$ und sie erfüllen $D(y_n)\dot{y}_n = O(h)$ für alle $n \in \mathbb{N}_0$.

Beweis. Wir führen den Beweis durch Induktion über n. Für $n = 0$ haben wir bereits in Lemma 4 gezeigt, dass unter den dort angegebenen Voraussetzungen eine Lösung $(y_1, \dot{y}_1)$ existiert und lokal eindeutig ist.

Setzen wir voraus, dass $(y_n, \dot{y}_n)$ existiert, lokal eindeutig ist und

$$\|D(y_n)\dot{y}_n\| \le C_0 h \tag{5.14}$$

mit einer festen Konstanten C_0 gilt, so ist zu zeigen, dass auch $(y_{n+1}, \dot{y}_{n+1})$ existiert und lokal eindeutig ist. Dazu genügt es nachzuweisen, dass $\|D(y_{n+1})\dot{y}_{n+1}\| \le C_0 h$ ist, denn mit dieser Voraussetzung können wir mit derselben Argumentation wie im Beweis von Lemma 4 auf die Existenz und lokale Eindeutigkeit schließen.

Wegen $D(y_0)\dot{y}_0 = O(h)$ gilt (5.6). Zusammen mit (5.4) und der Abschätzung aus (3.14) für (DAE 0) gilt für $q \ge 1$

$$\|(y_n, \dot{y}_n) - (y(t_n), \dot{y}(t_n))\| \le C_1(h^2 + h\rho^n), \tag{5.15}$$

wobei C_1 nicht von C_0 abhängt. Unter der Voraussetzung (5.14) gelten dann für die Runge-Kutta-Lösungen $(y_{n+1}, \dot{y}_{n+1})$ die Abschätzungen

$$\|(y_{n+1}, \dot{y}_{n+1}) - (y(t_{n+1}), \dot{y}(t_{n+1}))\| \le C_0 C_1 C_2 (h^2 + h\rho^n) \tag{5.16}$$

dabei hängt C_2 nicht von C_0 und C_1 ab. Die Abschätzungen (5.16) ergeben sich, indem der Beweis der Folgerung von Seite 70 mit den Voraussetzungen (5.14) und

(5.15) unter strikter Verwendung der Konstanten C_0 und C_1 nachvollzogen wird. Mit Theorem 1 ist klar, dass für alle $t \in [0, T]$

$$\|D(y(t))\dot{y}(t)\| \leq C_3\varepsilon$$

gilt. Damit, mit der Lipschitzstetigkeit der Funktion $(y(t), \dot{y}(t)) \mapsto D(y(t))\dot{y}(t)$ und mit (5.16) erhalten wir

$$\|D(y_{n+1})\dot{y}_{n+1}\| \leq c_{n+1}h$$

mit einer Rekursion $c_{n+1} = C(\rho^n + \alpha)c_n + C_3$ und einem $\alpha < 1$, wobei $c_1 = (C\rho + \alpha)c_0$ und $\|D(y_0)\dot{y}_0\| \leq c_0 h$ gilt. Diese Rekursion konvergiert gegen einen Grenzwert c. Setzen wir $C_0 := c$, so folgt die Behauptung. ∎

Kapitel 6

Ein Anwendungsbeispiel

In der Einleitung und im ersten Kapitel wurde die Bedeutung stark gedämpfter mechanischer Systeme für Anwendungen aus einigen Disziplinen der Physik- und Ingenieurwissenschaften bereits angedeutet. Nun soll diese anhand eines mechanischen Menschmodells von Hans [10] illustriert werden. Dabei wird insbesondere auf eine möglichst realitätsgetreue Modellierung der Gelenke geachtet. Die Zeitintegration von zwei einfachen Beispielen schließt die Arbeit ab.

Für den Biomechaniker stellt sich zunächst die generelle Frage, nach welchem Prinzip er die Bewegungsgleichungen aufstellt. Ein häufig verwendeter Ansatz liegt darin, das System mit einem maximalen Satz von Koordinaten darzustellen und Verbindungen zwischen den starren Körpern, wie etwa menschliche Gelenke, durch Zwangsbedingungen zu formulieren. Die Bewegung des mechanischen Systems wird dann durch gewöhnliche Differentialgleichungen zweiter Ordnung beschrieben. Zwangsbedingungen liegen in Form von algebraischen, im allgemeinen nichtlinearen Gleichungen, vor.
Ein anderer Ansatz ist die Verwendung von Minimalkoordinaten. Bei diesem Konzept werden starre Körper so mit einem minimalen Satz an Koordinaten beschrieben, dass die Zwangsbedingungen automatisch erfüllt sind. Betrachten wir beispielsweise ein System von zwei starren Körpern im dreidimensionalen Raum, die über ein Scharniergelenk miteinander verbunden sind, so existieren insgesamt sieben Freiheitsgrade. Eine mögliche Beschreibung durch Minimalkoordinaten wäre, die Orientierung eines Körpers im Raums durch die Festlegung von drei kartesischen Koordinaten und drei Winkeln zu beschreiben und die Lage des anderen Körpers durch die Vorgabe eines Winkels für das Scharniergelenk festzulegen. Somit ist das System durch die Festlegung von sieben Koordinaten eindeutig beschrieben.

Aus der Sicht der Biomechanik liegt bei beiden beschriebenen Ansätzen ein wesentlicher Nachteil in der Modellierung menschlicher Gelenke über Zwangsbedingungen.

Perfekte Zwangsbedingungen beschreiben exakte Bewegungen, wie sie etwa durch ein Scharnier- oder Kugelgelenk vorgegeben werden. Durch die Bandapparate in menschlichen Gelenken ist eine große Zahl von Bewegungen möglich, die grob dem Bewegungsspektrum von mechanischen Kugel- oder Scharniergelenken folgen. Im Detail der Bewegungsabläufe sind vom biomechanischen Standpunkt aus jedoch wesentliche Unterschiede festzustellen.

Um auch diese Details besser simulieren zu können, beschreibt Hans [10] alle Gelenke seines Menschmodells durch Feder-Dämpfer-Elemente. Für menschliche Gelenke, die in der Biomechanik als Scharniergelenk charakterisiert werden, werden beispielsweise zwei Feder-Dämpfer-Verbindungen, für Kugelgelenke nur eine verwendet. Mit geeigneten Kraft- und Dämpfungskonstanten lassen sich so auch kompliziertere Gelenke, wie etwa das menschliche Knie, realitätsgetreu modellieren. Ein weiterer Vorteil dieses Konzepts liegt in der Aufstellung von Bewegungsgleichungen, wie aus dem Beispiel des Stabpendels im nächsten Abschnitt ersichtlich wird. Diese lassen sich einerseits sehr schnell und einfach aufstellen. Andererseits ist die Größe des Systems unabhängig von der Anzahl der Zwangsbedingungen, denn zusätzliche Bedingungen erweitern das System nur um Feder-Dämpfer-Elemente, ändern aber an der Anzahl der Freiheitsgrade nichts. Alle Zwangsbedingungen, mit denen nicht nur Gelenke, sondern auch Kollisionen simuliert werden können, sind implizit in den auftretenden Kräften und Drehmomenten formuliert.
Die große Flexibilität, die von diesem Prinzip ausgeht, wird an den in [10] studierten Anwendungen deutlich, von denen zwei Simulationsbeispiele in den Abbildungen 6.1 und 6.2 aufgeführt sind.

Die Bildersequenz 6.1 zeigt die Simulation einer Reckturnübung, die Abbildung 6.2 das Verhalten des Menschmodells bei einem Verkehrsunfall. Es sei bemerkt, dass für die Zeitintegration der Bewegungsgleichungen RadauIIA-Verfahren verwendet wurden, wie sie in dieser Arbeit analysiert werden. Anstatt die komplizierten Modelle für diese Simulationen zu betrachten, sollen im folgenden Abschnitt exemplarisch Bewegungsgleichungen für ein Stab-Pendel mit Feder-Dämpfer-Element aufgestellt werden, das in Abbildung 6.3 nochmals dargestellt ist. Feder-Dämpfer-Verbindungen dieses Typs werden bei den obigen Simulationsbeispielen mehrfach eingesetzt, und das Aufstellen der Bewegungsgleichungen erfolgt nach dem selben Prinzip.
Ein physikalisches Problem, das in der Simulation 6.2 auftritt, ist die Kollisionserkennung. Der Lösungsanstaz, der hier verwendet wird, nutzt ebenfalls starke Dämpfungskräfte aus und wird im nächsten Abschnitt anhand eines einfachen Beispiels erklärt.

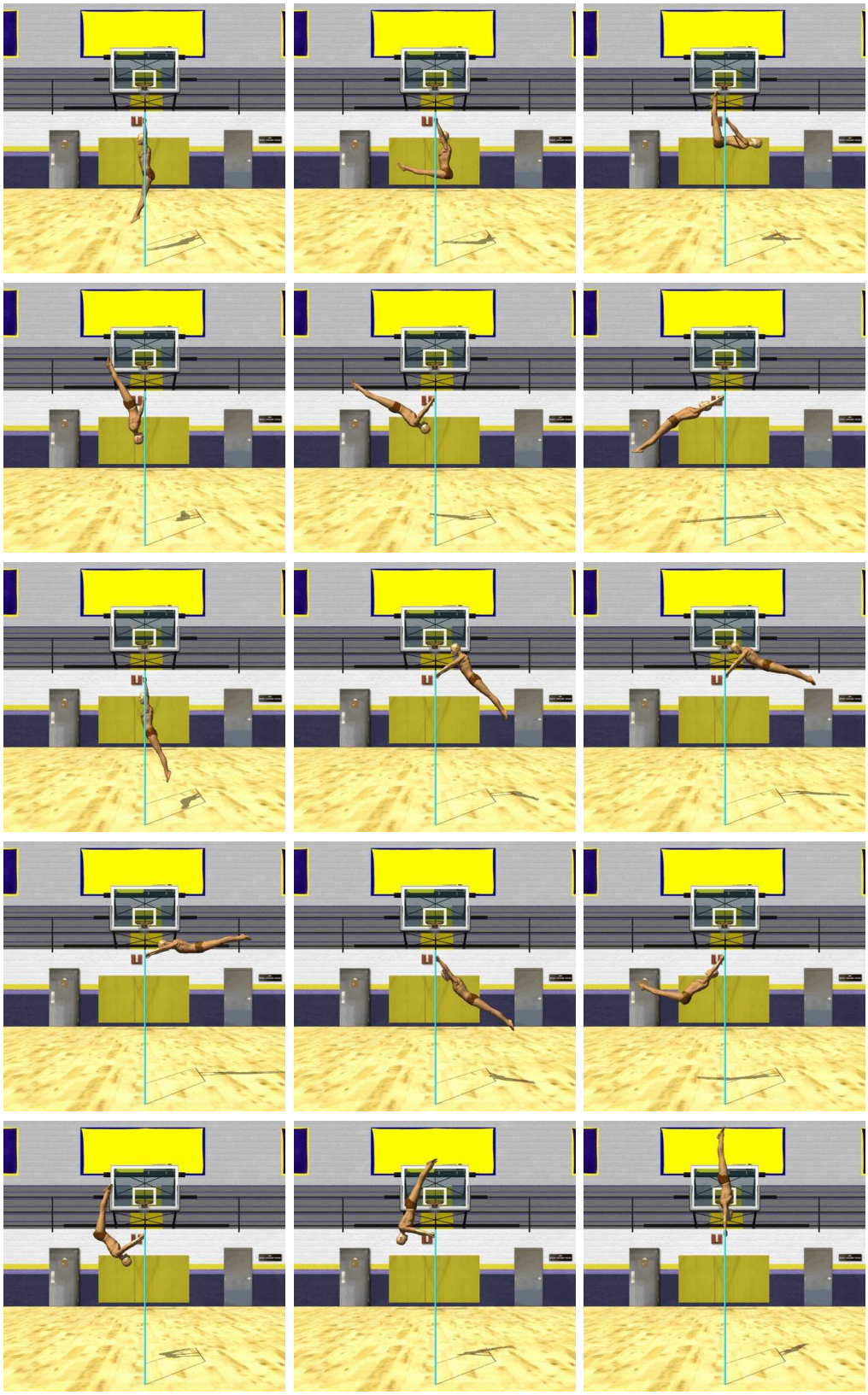

Abbildung 6.1: Menschmodell aus der Biomechanik bei der Simulation einer Reckturnübung, aus [10].

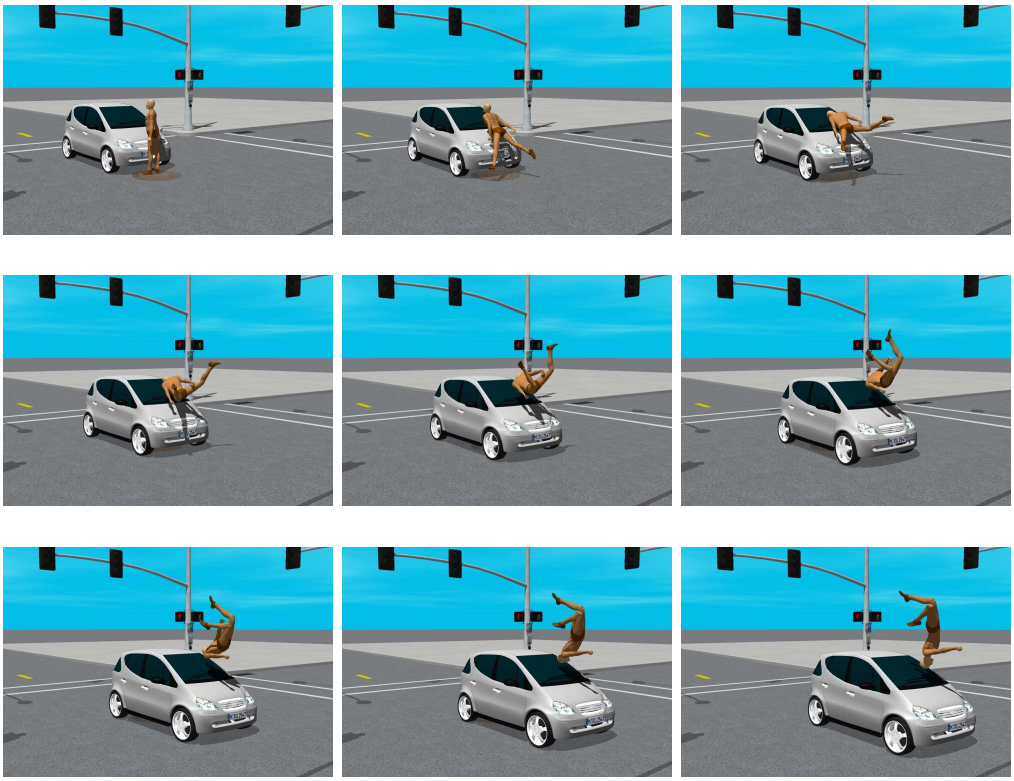

Abbildung 6.2: Menschmodell aus der Biomechanik bei der Simulation eines Unfalls, aus [10].

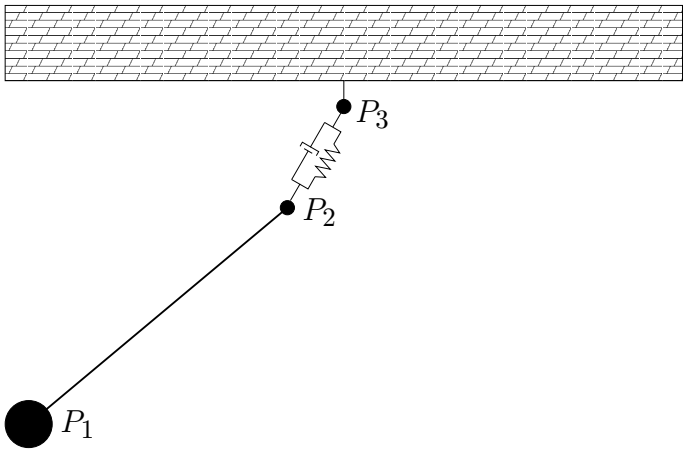

Abbildung 6.3: Pendel mit Kugel P_1 und Feder-Dämpfer-Element, das an zwei Punkte P_3 und P_2 gekoppelt ist.

6.1 Simulationsbeispiele

Bewegungsgleichungen des Stabpendels mit Feder-Dämpfer-Element

In diesem Abschnitt wollen wir die Bewegungsgleichungen für das Stabpendel aus 6.3 gemäß der in [10] verwendeten Vorgehensweise aufstellen. Wir betrachten dazu ein körperfestes Inertialsystem (KS), dessen Ursprung wir im Schwerpunkt des Körpers P_1 wählen, und ein raumfestes Inertialsystem (RS) mit Ursprung im Aufhängepunkt P_3 des Pendels, siehe Abbildung 6.4. Dabei vermittelt die Transformationsmatrix

$$O(t) = \begin{pmatrix} o_{11}(t) & o_{12}(t) & o_{13}(t) \\ o_{21}(t) & o_{22}(t) & o_{23}(t) \\ o_{31}(t) & o_{32}(t) & o_{33}(t) \end{pmatrix}$$

zwischen Vektoren des körperfesten und des raumfesten Inertialsystems. Zum Zeitpunkt t gilt etwa für den ersten kanonischen Basisvektor von KS die Transformation

$$O(t) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}_{KS} = \begin{pmatrix} o_{11}(t) \\ o_{21}(t) \\ o_{31}(t) \end{pmatrix}_{RS} .$$

Im Falle des Pendels betrachten wir ein beschleunigtes Bezugssystem mit rotatorischer, ebener Bewegung um die z-Achse von RS. Die Transformationsmatrix $O(t)$ ist dann durch

$$O(t) = \begin{pmatrix} \cos\varphi t & \sin\varphi t & 0 \\ -\sin\varphi t & \cos\varphi t & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

erklärt. Um den Zustand des starren Körpers, in unserem Fall der massebehafteten Kugel P_1, vollständig zu beschreiben, ist weiter die Kenntnis seiner Position y_1, seiner Geschwindigkeit v_1 und seines Drehimpulses L_1 nötig. Zu Zwecken der Zeitintegration fasst man alle Größen zu einem Vektor

$$Y(t) = (y_1(t),\ v_1(t),\ o_1(t),\ o_2(t),\ o_3(t), L_1(t))^T$$

zusammen, wobei $o_i(t)$ für $i = 1, 2, 3$ die Spalten der Matrix $O(t)$ bezeichnen. Zur Beschreibung der Bewegung von P_1 ist die zeitliche Änderung $\frac{d}{dt}Y(t)$ des Zustandsvektors zu berechnen. Hat der starre Körper P_1 die Masse m, so ist diese zeitliche Änderung für y_1, v_1 und L_1 einfach anzugeben. Die Bewegungsgleichungen für einen starren Körper sind

$$\begin{aligned} M\dot{v}_1(t) &= f(t, y(t), v(t)), \\ \dot{L}_1(t) &= T(t), \end{aligned}$$

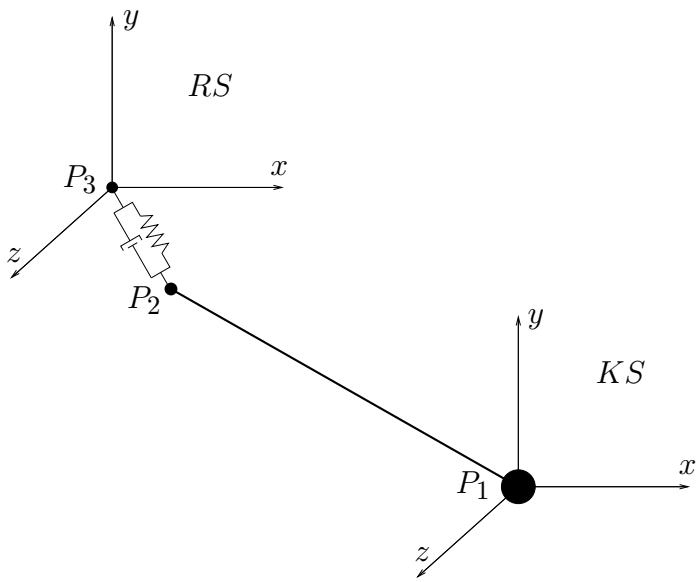

Abbildung 6.4: Raumfestes (RS) und körperfestes (KS) Inertialsystem des Pendels mit Ursprung in P_3 beziehungsweise P_1.

wobei in diesem Fall die Massenmatrix durch $M = m \cdot I_3$ gegeben ist. In f ist die Summe aller Kräfte zusammengefasst und in T sind alle Drehmomente aufsummiert, die zum Zeitpunkt t auf den starren Körper wirken.

Die Kräfte, die auf P_1 ausgeübt werden, ergeben sich aus der Kraftkonstanten k beziehungsweise der Dämpfungskonstanten d des Feder-Dämpfer-Elements. Diese wirken auf den Punkt P_2 und wir erhalten

$$f = k \cdot y_2(t) + d \cdot v_2(t).$$

Dabei bezeichnen $y_2(t)$ und $v_2(t)$ Position und Geschwindigkeit des Punktes P_2. Der Vektor y_2 lässt sich zum Zeitpunkt t als Vektorzug $y_2 = y_1 + a$ darstellen, wobei $a(t) = O(t) \cdot (0, l, 0)^T$ den Vektor von P_1 nach P_2 bezüglich RS darstellt und l die Länge des Stabes sei.
Das Kreuzprodukt eines Vektors mit der Winkelgeschwindigkeit zum Zeitpunkt t gibt seine zeitliche Änderung an. Mit der noch zu bestimmenden Winkelgeschwindigkeit $\omega(t)$ ist die zeitliche Änderung von v_2 dann durch $v_2(t) = v_1(t) + \omega(t) \times a(t)$ gegeben. Somit ergibt sich für f schließlich

$$f(t, y_1(t), v_1(t)) = k(y_1(t) + a(t)) + d(v_1(t) + \omega(t) \times a(t)).$$

Das auf P_1 wirkende Drehmoment ist

$$T(t) = a(t) \times f(t).$$

Für die reine Schwerpunktbewegung ergeben sich also keine weiteren Probleme. Die rotatorische Bewegung ist etwas schwieriger zu bestimmen. Hierzu benötigen wir zunächst den Trägheitstensor I_{KS} von P_1, der wie die Masse als konstant angenommen wird. Im vorliegenden Beispiel ist er als Trägheitstensor einer Kugel mit Radius r durch $I_{KS} = \frac{2}{5}r^2 \cdot I_3$ gegeben. Im raumfesten Inertialsystem berechnet sich die Darstellung des Trägheitstensors I_{RS} via

$$I_{RS}(t) = O(t)I_{KS}O(t)^T,$$

denn der Ursprung von KS wurde im Körperschwerpunkt gewählt. Dann ist die Winkelgeschwindigkeit mit dem Drehimpuls von P_1 durch

$$\omega(t) = I_{RS}^{-1}(t)L_1(t)$$

gegeben. Damit kann nun die zeitliche Änderung der Spaltenvektoren von $O(t)$ durch

$$\dot{o}_i(t) = \omega(t) \times o_i(t)$$

für $i = 1, 2, 3$ angegeben werden. Also ist $\dot{Y}(t)$ und somit die Bewegung von P_1 vollständig bestimmt und kann zusammengefasst werden:

$$\dot{Y}(t) = \begin{pmatrix} v_1(t) \\ M^{-1}f(t, y_1(t), v_1(t)) \\ \omega(t) \times o_1(t) \\ \omega(t) \times o_2(t) \\ \omega(t) \times o_3(t) \\ T_1(t) \end{pmatrix}.$$

In den ersten beiden Komponeneten tritt ein System der Form (1.1) auf, denn mit $\dot{y}_1(t) = v_1(t)$ haben wir

$$M\ddot{y}_1(t) = k(y_1(t) + a(t)) + d(\dot{y}_1(t) + \omega(t) \times a(t)).$$

Wir erhalten daraus ein stark gedämpftes Feder-Pendel, indem wir d in Relation zu k sehr groß wählen.

Modellierung der Kollisionserkennung

Die Modellierung von Kollisionen in [10] wird anhand eines einfachen, eindimensionalen Beispiels ersichtlich. Wir betrachten einen materiellen Punkt P, der aus einer Anfangshöhe y_0 mit einer Anfangsgeschwindigkeit v_0 auf die Erde aufprallt, siehe Abbildung 6.5. Dabei sei eine y-Achse entgegengesetzt zur Fallrichtung vorgegeben, die Erdoberfläche wird durch den 0-Punkt simuliert.

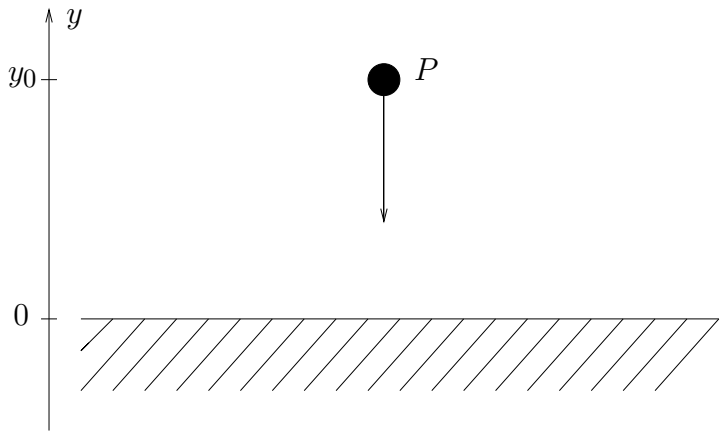

Abbildung 6.5: Modell zur Kollisionserkennung beim freien Fall eines materiellen Punktes P

Besitzt der materielle Punkt P die Masse 1, so wird dieses einfache Modell durch eine Differentialgleichung

$$\begin{aligned} \dot{y}(t) &= v(t), \\ \dot{v}(t) &= -g \end{aligned}$$

beschrieben, wobei g hier die Gravitationskonstante bezeichnet. Falls $y(t) < 0$ ist, liegt eine Kollision vor und wir ersetzen $-g$ durch eine Kraft

$$f = k \cdot y^2(t) + d \cdot y(t) \cdot v(t)$$

mit einer Kraftkonstanten k und einer Dämpfungskonstanten d. Wie beim Stabpendel erreichen wir eine starke Dämpfung, indem wir d in Relation zu k groß wählen.

6.2 Zeitintegration

Wir wollen abschließend die Kollisionserkennung und das Stabpendel mit Feder-Dämpfer-Element aus dem vorhergehenden Abschnitt mit zwei Runge-Kutta-Verfahren integrieren. Beide Anwendungsbeispiele wurden als Code in MATLAB programmiert. Der zu integrierende Vektor hat für die Kollisionserkennung die Länge 2 und für das Stabpendel die Länge 18. Als implizites Runge-Kutta-Verfahren, das die Bedingungen (3.7) und (3.8) erfüllt, benutzen wir einen in MATLAB implementierten Code von RADAU5 [4]. Dieser wurde in der Tabelle 3.2 als RadauIIA-Verfahren mit Ordnung 5 vorgestellt. Stellvertretend für die expliziten Runge-Kutta-Verfahren verwenden wir die MATLAB-Routine ODE45. Dieses Verfahren ist ein Zeitintegrator, der in einer Arbeit von Dormand und Prince [1] vorgestellt wird. In [19] ist eine

genaue Beschreibung der Routine zu finden. Mit diesen beiden Verfahren simulieren wir beide Beispiele mit unterschiedlich großen Dämpfungskonstanten.

In den vorigen Kapiteln konnte gezeigt werden, dass nur die Klasse impliziter Runge-Kutta-Verfahren, die insbesondere den Voraussetzungen (3.7) und (3.8) genügt (und zu der beispielsweise RADAU5 zählt), zur Integration stark gedämpfter mechanischer Systeme geeignet ist. Bei Anwendung impliziter Verfahren, die nicht in diese Klasse fallen, verlieren die bewiesenen Konvergenzresultate ihre Gültigkeit.
Dieser Unterschied lässt sich allerdings an den beiden Modellen aufgrund der kleinen Systemgrößen nicht demonstrieren. Hierzu wäre ein komplexeres Modell, wie etwa die Crash-Simulation 6.2 zu betrachten. Hans stellt in [10] fest, dass dieses Beispiel nur mit RADAU5 und mit keinem der anderen impliziten Verfahren, die in MATLAB als Routinen vorliegen, effizient zu simulieren ist. Die komplexe Unfallsimulation 6.2 einzuführen und numerisch zu behandeln würde aber den Rahmen dieser Arbeit sprengen.

Zeitintegration der Kollisionserkennung

Zunächst betrachten wir eine Kollision mit relativ moderaten Dämpfungskräften. Wir simulieren den Aufprall eines 1 kg schweren materiellen Punktes aus einer Anfangshöhe y_0 von einem Meter. Die Anfangsgeschwindigkeit v_0 betrage $10m/s$, die Kraft- und Dämpfungskonstanten wählen wir als $k = 4 \cdot 10^4 N/m^2$ beziehungsweise $d = 4 \cdot 10^3 Ns/m^2$. Das Ergebnis einer Zeitintegration über 2 Sekunden ist in Abbildung 6.6 dargestellt.

Aufgrund der relativ schwachen Dämpfung wird eine verhältnismäßig große Eindringtiefe in den Untergrund - sie beträgt fast 10 cm - erreicht. Auch die Höhe von nahezu 40 cm, die der Punkt vom Boden zurückspringt, lässt Rückschlüsse auf die milde Dämpfung zu. Nach einer Sekunde ist ein Stillstand erreicht und der materielle Punkt liegt auf der Erdoberfläche.

Zum Vergleich der vorgestellten Runge-Kutta-Verfahren betrachten wir eine Kollision mit höherer Ausgangsgeschwindigkeit und wählen $v_0 = 100\ m/s$. Die Kraftkonstante setzen wir $k = 10^5 N/m^2$ und die Dämpfungskonstante steigern wir je Simulation von $d = 10^4\ Ns/m^2$ bis $d = 5 \cdot 10^7 Ns/m^2$ in Schritten mit Faktor 5 beziehungsweise 2. Diese Simulationen werden je mit ODE45 und RADAU5 über eine Zeit von 0.6 Sekunden durchgeführt. Für beide Integrationsverfahren wählen wir gemäß des ODEFILE-Standards in MATLAB die Toleranzen $RelTol = 10^{-6}$ und $AbsTol = 10^{-8}$. Das Lösungsverhalten ist in Abbildung 6.7 in einem Abschnitt von 0 bis 0.1 Sekunden dargestellt. Die beträchtliche Energie des Systems geht durch die sehr starke Dämpfung sofort verloren und jegliche Bewegung kommt direkt beim

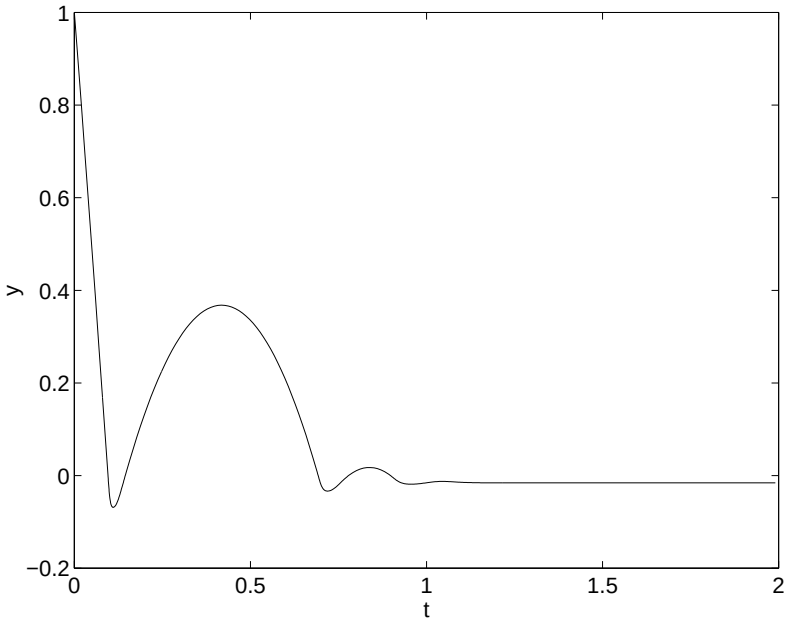

Abbildung 6.6: Aufprall eines Massepunktes mit $y_0 = 1m$, $v_0 = 10m/s$, $k = 40000N/m^2$ und $d = 4000Ns/m^2$. Zeit t gegen Position y.

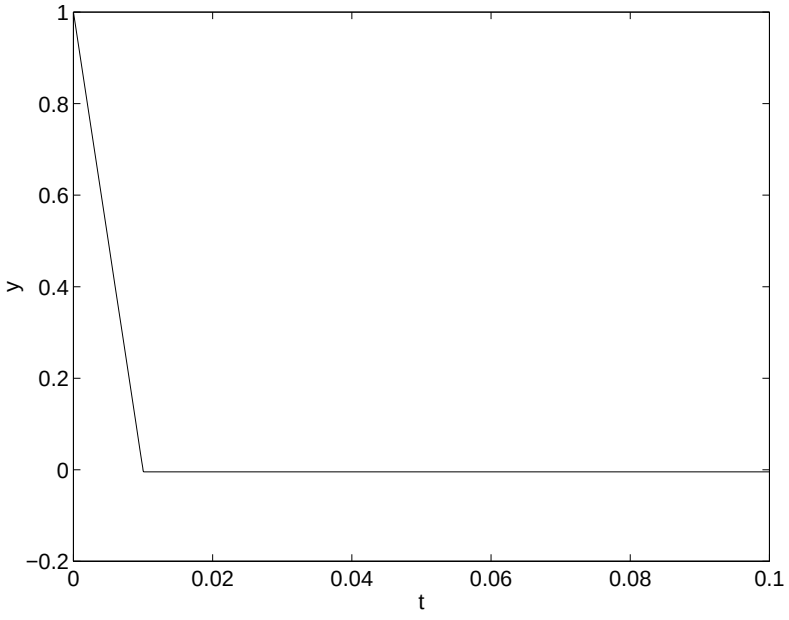

Abbildung 6.7: Aufprall eines Massepunktes mit $y_0 = 1m$, $v_0 = 100m/s$, $k = 10^5N/m^2$ und $d = 10^7Ns/m^2$. Zeit t gegen Position y.

Aufprall zum Erliegen. In einem noch begrenzteren Ausschnitt wird auch deutlich, dass die Lösung zur Zeit $t = 0.01$ hinreichend oft stetig differenzierbar ist.

Um explizites und implizites Verfahren zu vergleichen, tragen wir in Abbildung 6.8 die durchschnittliche Schrittweite von RADAU5 und ODE45 gegen die jeweils verwendeteten Dämpfungskonstanten auf. Wie zu erwarten ist, müssen bei den expliziten Verfahren mit steigenden Dämpfungskonstanten immer kleinere Zeitschritte gewählt werden, um die exakte Lösung mit einem akzeptablen Fehler zu approximieren. RADAU5 hingegen bewältigt alle Simulationen mit relativ großen Schrittweiten. In Abbildung 6.9 ist zu erkennen, dass ODE45 auch die Vorzüge des deutlich geringeren Rechenaufwandes sehr schnell verliert. Für die Dämpfungskonstante 10^4 ist die Anzahl der Fließpunktoperationen bei RADAU5 noch doppelt so groß wie bei ODE45, bei $d = 5 \cdot 10^5$ ist das implizite Verfahren bereits im Vorteil. Zur Absicherung dieser Ergebnisse vergleichen wir die berrechneten Lösungen in Abhängigkeit von den Dämpfungskonstanten mit einer Referenzlösung. Beide Verfahren erreichen dabei bis auf nicht erwähnenswerte Unterschiede dieselben Genauigkeiten, die in der Größenordnung der gewählten Toleranz $AbsTol = 10^{-8}$ liegen.

Zusammenfassend bleibt also festzuhalten, dass explizite Verfahren Lösungen der Bewegungsgleichungen der Kollisionserkennung unter Verwendung starker Dämpfungskräfte nur mit sehr kleiner Schrittweite exakt genug approximieren können. Auch die Tatsache, dass bei expliziten Methoden keine nichtlinearen Gleichungssysteme gelöst werden müssen, zahlt sich beim Rechenaufwand nicht aus.

Zeitintegration des Stabpendels

Um komplexe menschliche Gelenke, beispielsweise das Knie, in Simulationen wie sie in den Abbildungen 6.1 und 6.2 dargestellt sind zu modellieren, wurden in [10] Feder-Dämpfer-Elemente mit großen Dämpfungskonstanten verwendet. Ein Feder-Dämpfer-Element dieses Typs wollen wir nun für eine Simulation mit dem Stabpendel verwenden und wiederum die Dämpfungskonstante zum Vergleich von ODE45 und RADAU5 variieren.

Die Bewegung des im vorigen Abschnitt beschriebenen Stabpendels simulieren wir zunächst über die Zeit von 10 Sekunden mit einer Kraftkonstanten von $k = 10^5 N/m^2$ und einer Dämpfung mit $d = 10^4 Ns/m^2$. Dem materiellen Punkt ordnen wir einen Radius von 0.01 m und eine Masse von 1 kg zu. Der Stab wird als masselos mit Länge 1 angenommen. Die Anfangslänge der Feder wird idealisiert als 0 gesetzt, der Anfangswinkel der Auslenkung des Stabes beträgt $\frac{2}{9}\pi$. In Abbildung 6.10 ist die Zeit gegen den Abstand des Punktes P_2 zu P_3 und die Relativgeschwindigkeit v des Punktes P_2 aufgetragen. Sie berechnet sich aus $v = \langle v_2, r/\|r\| \rangle$, wobei r den Vektor von P_2 nach P_3 bezeichnet. Die starke Dämpfung wird dadurch ersichtlich, dass

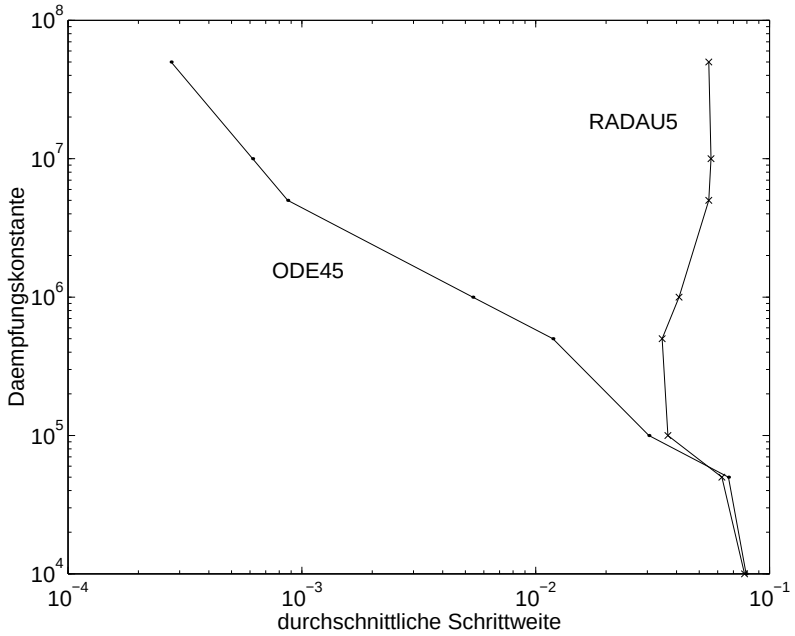

Abbildung 6.8: Durchschnittliche Schrittweite - Dämpfungskonstante d.

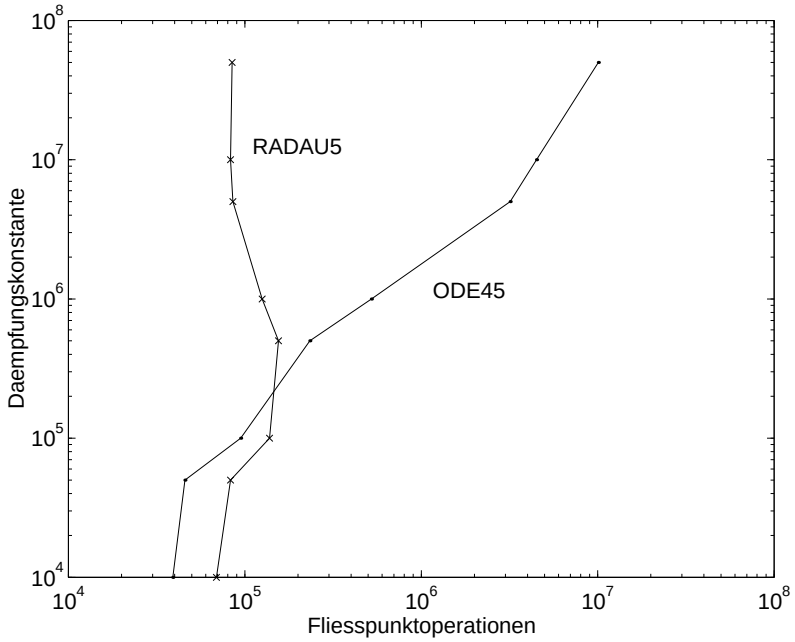

Abbildung 6.9: Fließpunktoperationen - Dämpfungskonstante d.

sich die Länge der Feder sehr schnell auf ein kleines Intervall beschränkt. Erhöht man die verwendete Dämpfungskonstante, so wird dieses Intervall immer kleiner. Die Relativgeschwindigkeit, die zu Beginn der Simulation noch groß ist, nähert sich sehr schnell einem Wert nahe 0 an. Hier treten positive und negative Werte auf, denn das Feder-Dämpfer-Element wird aufgrund der Schwingung des Stabpendels abwechselnd zusammengezogen und gestreckt. Hat der Stab vertikale Position erreicht, so ist die Auslenkung der Feder am größten, denn die Zuglast des materiellen Punktes ist in dieser Position am höchsten.

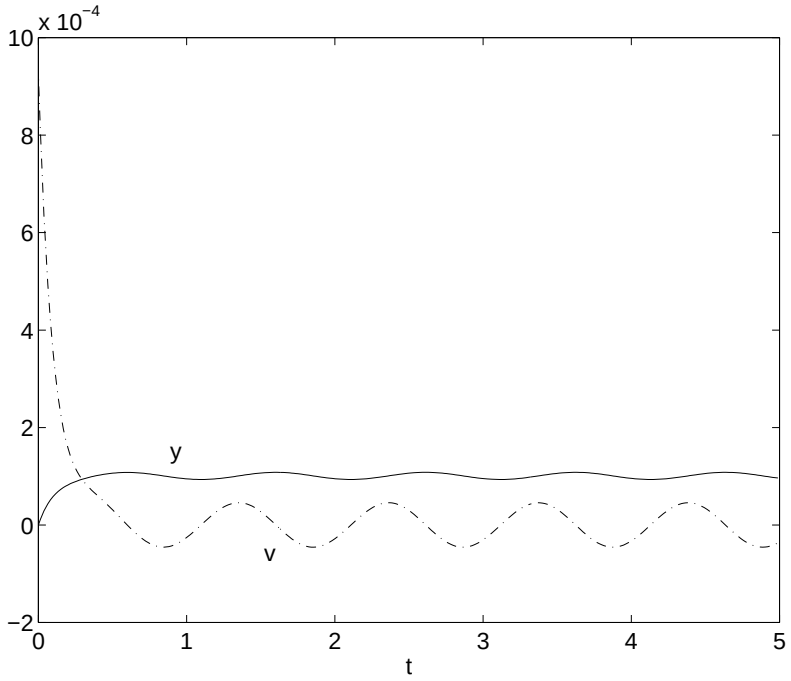

Abbildung 6.10: Simulationsergebnis des Stabpendels: qualitatives Verhalten des Abstandes y von P_2 und P_3 und der Relativgeschwindigkeit v von P_2.

Die Verwendung von hohen Kraft- und Dämpfungskonstanten führt uns hier sehr nahe an den Grenzfall dieses Stabpendels. Das System verliert kaum Energie und die Bewegung des materiellen Punktes ist fast identisch zu der Bewegung eines mathematischen Pendels mit entsprechenden Ausgangsdaten. Diese Beobachtung steht in Einklang mit Theorem 11.

Zum Vergleich des expliziten und impliziten Verfahrens vergrößern wir wieder die Dämpfungskonstanten von 10^1 bis 10^8 in Schritten mit Faktor 10. Wir integrieren die jeweiligen Pendelmodelle über eine Zeit von einer Sekunde und wählen die restlichen Größen wie für das Modell aus Abbildung 6.10. Absolute und relative Toleranz für

die Integrationsverfahren werden wieder durch $AbsTol = 10^{-8}$ und $RelTol = 10^{-6}$ festgelegt.

Dämpfungskonstante d in Ns/m^2	RADAU5		ODE45	
	Fließpunkt-operationen	Schritte	Fließpunkt-operationen	Schritte
10^1	$6.51 \cdot 10^6$	209	$3.39 \cdot 10^8$	75027
10^2	$1.68 \cdot 10^6$	46	$3.4 \cdot 10^9$	753265
10^3	$1.67 \cdot 10^6$	37	-	-
10^4	$3.79 \cdot 10^6$	60	-	-
10^5	$8.91 \cdot 10^6$	128	-	-
10^6	$1.26 \cdot 10^7$	185	-	-
10^7	$2.18 \cdot 10^7$	291	-	-
10^8	$4.43 \cdot 10^7$	583	-	-

In der Tabelle sind Fließpunktoperationen und die Anzahl der Schritte unter Bezug auf die jeweils eingesetzte Dämpfungskonstante zusammengefasst. Bei RADAU5 wächst der Aufwand in etwa proportional zu größer werdenden Dämpfungskonstanten (ab $d = 10^3$). Der höhere Rechenaufwand für $d = 10$ ist durch den großen Unterschied zur Kraftkonstanten der Feder ($k = 10^5$) zu erklären. Dieses Simulationsbeispiel fällt in die Klasse der steifen oszillatorischen mechanischen Systeme (1.2). Die verschiedenen Beispiele gehen also mit wachsenden Dämpfungskonstanten von steifen oszillatorischen mechanischen Systeme in stark gedämpfte mechanische Systeme über.

ODE45 konnte nur für kleine Dämpfungskonstanten ein Ergebnis in einer akzeptablen Zeit liefern. Die Simulationen für $d = 10^3$ bis $d = 10^8$ wurden abgebrochen, da das explizite Verfahren zu kleine Schrittweiten wählen musste.
Um einen Vergleich zu erhalten, berechnen wir aus den ersten 30000 Schritten von ODE45 die durchschnittliche Schrittweite. Die Ergebnisse wurden in Abbildung 6.11 gegen die zugehörigen Dämpfungskonstanten aufgetragen und mit den entsprechenden Resultaten von RADAU5 verglichen. RADAU5 integriert die Beispiele mit nicht nennenswerter Steigerung der Schrittweite während diese bei ODE45 proportional zur Dämpfungskonstanten verringert werden muss, um die exakte Lösung genau genug zu approximieren.

Fassen wir die Ergebnisse aus beiden Beispielen nochmals zusammen, so konnten bei der Kollisionserkennung noch alle Simulationen mit dem expliziten Verfahren in einem vernünftigen Zeitrahmen durchgeführt werden. Bereits beim Stabpendel ist dies nicht mehr möglich, wogegen die in dieser Arbeit untersuchten impliziten Verfahren vom RadauIIA-Typ unabhängig von den Dämpfungskonstanten gleichmäßig gute Ergebnisse liefern.

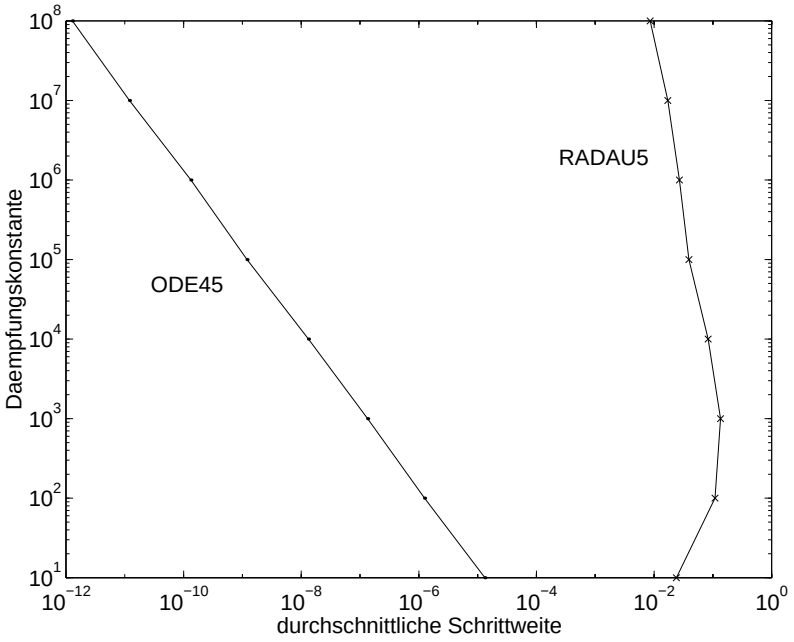

Abbildung 6.11: Durchschnittliche Schrittweite - Dämpfungskonstante d.

Literaturverzeichnis

[1] J. R. Dormand, P. J. Prince, *A family of embedded Runge-Kutta formulae.* J. Comp. Appl. Math., Vol. 6, 19-26 (1980).

[2] E. Eich, M. Hanke, *Regularization methods for constrained mechanical multibody systems.* Z. Angew. Math. Mech. 75, No. 10, 761-773 (1995).

[3] E. Eich-Soellner, C. Führer, *Numerical Methods in Multibody Dynamics.* Teubner, Stuttgart 1998.

[4] C. Engstler, *Code zur MATLAB-Implementierung von RADAU5.* Verfügbar unter `http://na.uni-tuebingen.de/na/software.shtml`.

[5] E. Hairer, Ch. Lubich, M. Roche, *Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations.* BIT 29, No. 1, 77-90 (1989).

[6] E. Hairer, Ch. Lubich, M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods.* Springer LNM 1409, 1989.

[7] E. Hairer, Ch. Lubich, G. Wanner, *Geometric Numerical Integration.* Springer-Verlag, Berlin, Heidelberg, New York 2001.

[8] E. Hairer, S. P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I.* Springer-Verlag, Berlin 1987.

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

[10] T. Hans, *Interaktive Simulation biomechanischer Bewegungsabläufe.* Dissertation, Universität Tübingen, Tübingen 2004.

[11] T. Kato, *Perturbation Theory for Linear Operators.* 2nd Ed., Grundlehren math. Wiss. 132, Springer-Verlag, Berlin, Heidelberg, New York 1980.

[12] P. Lötstedt, *On a penalty function method for the simulation of mechanical systems subject to constraints.* Report TRITA-NA 7919, Royal Inst. of Technology, Stockholm 1979.

[13] Ch. Lubich, *Integration of Stiff Mechanical Systems by Runge-Kutta methods.* Z. Angew. Math. Phys. 44, No. 6, 1022-1053, (1993).

[14] U. Kirchgraber, F. Lasagni, K. Nipp, D. Stoffer, *On the application of invariant manifold theory, in particular to numerical analysis.* Int. Ser. Num. Math. 97, 189-197 (1991).

[15] Mathworks, *Using Matlab.* User-Manual, Version 5.3 (Release 11) 1999.

[16] D. Morgenstern und I. Szabó, *Vorlesungen über theoretische Mechanik.* Grundlehren math. Wiss. 112, Springer-Verlag, Berlin, Göttingen, Heidelberg 1961.

[17] K. Nipp, D. Stoffer, *Attractive invariant manifolds for maps: Existence, smoothness, and continuous dependence on the map.* Report 92-11, SAM, ETH Zürich 1992.

[18] R. E. O'Malley, Jr., *Singular Perturbation Methods for Ordinary Differential Equations.* Springer-Verlag, Berlin, Heidelberg, New York 1991.

[19] L. F. Shampine, M. W. Reichelt, *The MATLAB ODE Suite.* SIAM J. Sci. Comput. 18, No. 1, 1-22 (1997).

[20] B. Simeon, F. Grupp, C. Führer, P. Rentrop, *A nonlinear truck model and its treatment as a multibody system.* J. Comput. Appl. Math. 50, No. 1-3, 523-532 (1994).

Lebenslauf

	Thomas Stumpp
29.5.1972	geboren in Riedlingen
1978 - 1982	Grundschule in Scheer
1982-1991	Gymnasium Mengen
17.6.1991	Abitur
1991 - 1992	Zivildienst
1993 - 2000	Studium der Mathematik und Geographie an der Eberhard-Karls Universität in Tübingen
31.5.2000	Wissenschaftliche Prüfung für das Lehramt an Gymnasien
seit Juni 2000	Doktorand und wissenschaftlicher Angestellter an der Universiät Tübingen

Meine akademischen Lehrer in Mathematik waren die Herren Professoren und Dozenten

R. Bödi, U. Felgner, Ch. Hering, H. Heyer, M. Hochbruck, W. Knapp, P. Leinen, F. Loose, Ch. Lubich, F. Räbiger, U. Riese, H. Salzmann, P. Schmid, H. Yserentant.