Lehrstuhl für Numerische Mathematik

Numerische Integration steifer mechanischer Systeme mit impliziten Runge-Kutta-Verfahren

Meike Schaub

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. F. Bornemann
Prüfer der Dissertation:	
1.	Univ.-Prof. Dr. B. Simeon
2.	Univ.-Prof. Dr. M. Arnold, Martin-Luther-Universität Halle-Wittenberg

Die Dissertation wurde am 07.10.2003 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 26.01.2004 angenommen.

Vorwort

Ich möchte mich bei einigen Personen bedanken, die sehr zum Gelingen dieser Arbeit beigetragen haben. Zuallererst ist Prof. Dr. B. Simeon zu nennen, der durch die interessante Aufgabenstellung und engagierte Betreuung den Grundstein gelegt hat. Auch Prof. Dr. P. Rentrop verdanke ich viel, da er mich sehr freundlich in seine Arbeitsgruppen in Karlsruhe und München integriert und mich mit Rat und Tat unterstützt hat.

Prof. Dr. M. Arnold hat durch interessante Fragen und Anregungen einen wichtigen Beitrag zum Voranschreiten der Arbeit geleistet, und ich bin ihm sehr dankbar für die Übernahme eines Gutachtens. Desweiteren geht mein Dank an G. Hippmann vom DLR Oberpfaffenhofen für die Erstellung des Waschmaschinenmodells sowie die Beantwortung der zugehörigen Fragen.

Ich möchte mich auch bei Prof. Dr. L. Jay für die Bereitstellung des Fortran-Codes `SPARK3` und die damit verbundene Diskussionsbereitschaft bedanken, sowie bei Prof. Dr. C. Führer für die Einladung zum Forschungsaufenthalt in Lund und das Interesse an meiner Arbeit.

Den Kolleginnen und Kollegen am IWRMM in Karlsruhe und am Institut M2 in München danke ich für die nette Arbeitsatmosphäre, die dafür gesorgt hat, dass ich immer gerne die großen Entfernungen zurückgelegt habe. Außerdem konnte ich jederzeit auf den Rückhalt in meiner Familie und vor allem bei meinem Mann zurückgreifen, wofür ich ebenfalls sehr dankbar bin.

Bensheim, Oktober 2003 Meike Schaub

Inhaltsverzeichnis

Einleitung

Die numerische Simulation gewinnt seit der rasanten Computerentwicklung immer mehr an Bedeutung, da sie wichtige Vorabinformationen zum Ablauf von naturwissenschaftlichen Prozessen liefern kann. Dabei ist vor allem in der Fahrzeugdynamik und verwandten Gebieten die Modellierung durch Mehrkörpersysteme verbreitet. Oft enthalten diese Modelle steife Federn, die prinzipiell hochfrequente Oszillationen in das System einbringen. In der Realität werden diese Schwingungen nicht angeregt und spielen daher keine Rolle. In einer dynamischen Simulation aber bewirkt allein das Vorhandensein dieser Federn ähnliche Schwierigkeiten wie bei bei der Simulation klassischer steifer Systeme z.B. aus der Reaktionskinetik, was den Namen *steife mechanische Systeme* rechtfertigt.

Doch nicht nur durch die Existenz steifer Federn, sondern auch durch eine feine Ortsdiskretisierung von elastischen Körpern kann ein steifes mechanisches System entstehen, bei dem der Kehrwert der Ortsgitterweite die Rolle der Federsteifigkeit übernimmt.

Ziel dieser Arbeit ist es, die bei dieser Problemklasse auftretenden Schwierigkeiten zu analysieren und numerische Verfahren und Algorithmen für die Anwendung auf steife mechanische Systeme einzuführen und optimal auszulegen. Die Schwerpunkte liegen auf den Verfahren der Lobatto-Familie, der Schrittweitensteuerung und dem Newton-Verfahren.

Doch was bewirken steife Terme bei der numerischen Integration? Zunächst macht sich die Verwandtschaft zu differential-algebraischen Gleichungen bemerkbar, da dort ebenfalls das Phänomen der Ordnungsreduktion auftritt. Dies hat zur Folge, dass die Wahl der Integrationsverfahren auf solche mit höherer Ordnung und numerischer Dämpfung fallen muss.

Die durch die Ordnungsreduktion verringerte Ordnung hat aber auch Auswirkungen auf die Schrittweitensteuerung, die zur Kontrolle des Fehlers eingesetzt wird, da die verwendeten Fehlerschätzer nicht mehr zufriedenstellend arbeiten. Dies führt im Allgemeinen zu einer unnötigen Verkleinerung der Schrittweite. Ein weiteres Problem bei steifen mechanischen Systemen liegt in der Lösung der nichtlinearen Gleichungssysteme, die bei der Anwendung impliziter Verfah-

ren entstehen. Hohe Steifigkeitsterme sorgen dabei nicht nur für eine schlechte Kondition der Iterationsmatrix, auch die Kontraktivität der Fixpunktiteration ist gefährdet.

Die Klasse der impliziten Runge-Kutta-Verfahren ist aufgrund ihrer Struktur sehr gut für die Integration steifer mechanischer Systeme geeignet. In dieser Arbeit greifen wir das Radau IIA-Verfahren und die Familie der Lobatto-Verfahren heraus, die durch hervorragende Stabilitätseigenschaften bzw. hohe Flexibilität hervorstechen.

Die Arbeit ist in fünf Kapitel unterteilt. Das erste Kapitel befasst sich mit der Modellbildung, unterschiedlichen Darstellungen steifer mechanischer Systeme und anderen mechanischen Eigenschaften.

Implizite Runge-Kutta-Verfahren werden im zweiten Kapitel vorgestellt, wobei vor allem auf jene Kriterien eingegangen wird, die für die Anwendung auf steife mechanische Systeme relevant sind. Im weiteren Verlauf wird eine Rückwärtsanalyse durchgeführt, die Aussagen über den Energiefehler von Runge-Kutta-Verfahren erlaubt.

Im dritten Kapitel steht die Familie der Lobatto-Verfahren im Vordergrund. Es wird erläutert, wie die einzelnen Verfahren miteinander kombiniert werden können und welche Eigenschaften die dadurch entstehenden Verfahren besitzen. Die Vielseitigkeit der Verfahren ermöglicht somit die Anpassung der Eigenschaften an die Erfordernisse des Systems.

Das vierte Kapitel enthält eine Analyse der numerischen Algorithmen in Bezug auf steife mechanische Systeme. Unterschiedliche eingebettete Verfahren und Fehlerschätzer werden auf ihre Wirkung hin analysiert, der lokale Fehler als Erkennungsmerkmal steifer Komponenten eingeführt sowie die Konvergenz des Newton-Verfahrens für unterschiedliche Formulierungen untersucht. Zudem werden Anwendungsmöglichkeiten der Lobatto-Verfahren erläutert.

Die Ergebnisse der Simulationsbeispiele sind im fünften Kapitel aufgeführt. Neben einfacheren Beispielen zur genauen Betrachtung der vorgestellten Fehlerschätzer, Algorithmen und Formulierungen werden größere Beispiele wie ein Kurbeltrieb, eine Waschmaschine und die Ladefläche eines Lastwagens simuliert.

Kapitel 1

Mehrkörpersysteme

Zur Modellierung von mechanischen Systemen werden Mehrkörpersysteme verwendet, die eine Kopplung starrer Körper durch verschiedene Elemente wie Federn, Dämpfer oder Gelenke realisieren. Auch elastische Körper lassen sich in dieses Schema einbeziehen.

Je nach Wahl der Koordinaten und der Verbindungselemente als Gelenke oder Federn entstehen gewöhnliche Differentialgleichungen oder differential-algebraische Gleichungen, deren Vor- und Nachteile zu diskutieren sind. In diesem Zusammenhang wird der Begriff der *steifen mechanischen Systeme* eingeführt, der in der Arbeit eine zentrale Rolle spielt.

Auch bei Transformation auf ein System erster Ordnung ergeben sich verschiedene Formulierungen, je nachdem, ob Geschwindigkeiten oder Impulse als zusätzliche Größen verwendet werden.

Dieses Kapitel beginnt mit der Herleitung der Bewegungsgleichungen gefolgt von einer Analyse der entstandenen Gleichungen. Neben der Theorie der gewöhnlichen Differentialgleichungen gibt es Einblick in die Problematik differential-algebraischer Gleichungen und steifer mechanischer Systeme sowie in Eigenschaften von Hamilton-Systemen.

1.1 Herleitung der Bewegungsgleichungen

Zur mathematischen Beschreibung eines mechanischen Systems als *Mehrkörpersystem* geht man von n starren Körpern aus, die durch Gelenke oder auch durch Federn oder Dämpfer miteinander verbunden sind. Die einzelnen Körper sind dabei durch ihre Form und Dichte eindeutig bestimmt.

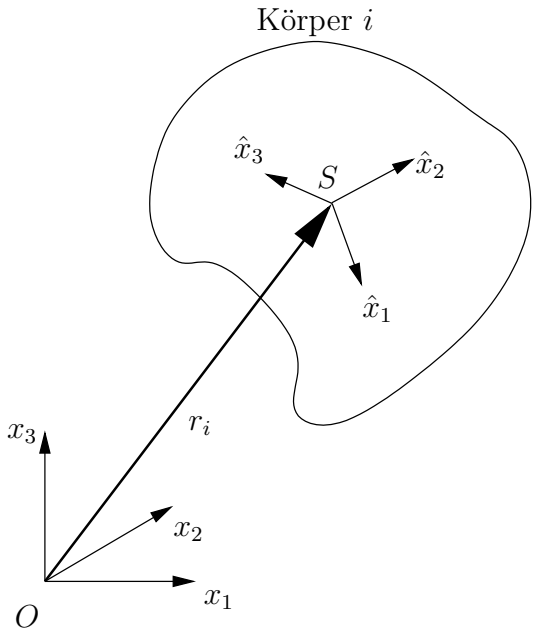


Abbildung 1.1: Absolutkoordinaten eines starren Körpers.

Eine Erweiterung dazu sind *flexible* Mehrkörpersysteme. Dabei werden einzelne Körper als elastisch angenommen, also nicht nur die Starrkörperbewegung, sondern auch die Verzerrungen innerhalb des Körpers mitberücksichtigt. Auch wenn die speziellen Problematiken dieser Klasse in dieser Arbeit nicht berücksichtigt werden, besitzen die durch Semidiskretisierung entstandenen Gleichungen dieselbe Struktur wie die Gleichungen starrer Körper.

Zur Herleitung der Bewegungsgleichungen stehen mehrere Prinzipien zur Verfügung, die jeweils zu äquivalenten Formulierungen führen. Wir verwenden die Methode nach Lagrange und betrachten dazu die Abb. 1.1.

Ausgehend von einem Inertialsystem wird die Lage eines starren Körpers i durch den Ortsvektor

$$r_i = (x_{i1}, x_{i2}, x_{i3})^T$$

und die Ausrichtung durch drei verallgemeinerte Koordinaten α_{i1}, α_{i2} und α_{i3}, i. A. Winkel, beschrieben. Für sich alleine betrachtet besitzt daher jeder starre Körper 6 Freiheitsgrade, die aber durch die Kopplung zu benachbarten Körpern eingeschränkt werden können.

Zum weiteren Vorgehen benötigen wir einen Ausdruck für die kinetische Energie.

Dieser lautet im Allgemeinen

$$T(q,\dot{q}) = \frac{1}{2}\sum_{i=1}^{n}\left(r_i^T m_i r_i + \omega_i^T S_i I_i S_i^T \omega_i\right),$$

wobei m_i die Masse, I_i den Trägheitstensor, $S_i(\alpha_{i1}, \alpha_{i2}, \alpha_{i3})$ den Drehtensor und ω_i die Winkelgeschwindigkeit des Körpers i beschreibt. Der Koordinatenvektor q setzt sich aus den Ortsvektoren r_i und den verallgemeinerten Koordinaten α_i zusammen. Mithilfe dieses Ausdrucks und dem Vektor $Q(q,\dot{q},t)$ der eingeprägten und äußeren Kräfte, die durch Freischneiden der einzelnen Körper ermittelt werden, können wir die Bewegungsgleichungen aufstellen.

Falls die Bewegungen nicht durch zusätzliche Zwangsbedingungen eingeschränkt werden, verwenden wir dazu die *Lagrangegleichungen 2. Art*

$$\frac{d}{dt}\left(\frac{\partial}{\partial\dot{q}}T(q,\dot{q})\right) - \frac{\partial}{\partial q}T(q,\dot{q}) = Q(q,\dot{q},t), \tag{1.1}$$

anderenfalls die *Lagrangegleichungen 1. Art*

$$\frac{d}{dt}\left(\frac{\partial}{\partial\dot{q}}T(q,\dot{q})\right) - \frac{\partial}{\partial q}T(q,\dot{q}) = Q(q,\dot{q},t) - G^T(q)\lambda, \tag{1.2a}$$

$$0 = g(q), \tag{1.2b}$$

die die Zwangsbedingungen als algebraische Gleichungen $g(q) = 0$ sowie als Zwangsmatrix $G(q) = \partial g(q)/\partial q$ enthalten. Letztere wird mit Hilfe der Lagrangemultiplikatoren λ angekoppelt.

Aus (1.2) erhalten wir mit der Massenmatrix

$$M(q) := \frac{\partial^2}{\partial\dot{q}^2}T(q,\dot{q})$$

und dem Kräfteausdruck

$$f(q,\dot{q},t) := Q(q,\dot{q},t) - \left(\frac{\partial^2}{\partial q\partial\dot{q}}T(q,\dot{q})\right)\dot{q} + \frac{\partial}{\partial q}T(q,\dot{q}) \tag{1.3}$$

durch Ausdifferenzieren die Bewegungsgleichung

$$M(q)\ddot{q} = f(q,\dot{q},t) - G^T(q)\lambda, \tag{1.4a}$$

$$0 = g(q). \tag{1.4b}$$

Aus (1.1) ergibt sich entsprechend die Bewegungsgleichung

$$M(q)\ddot{q} = f(q,\dot{q},t). \tag{1.5}$$

Wann welche Form der Bewegungsgleichungen verwendet wird, werden wir im nächsten Unterabschnitt diskutieren.

Zunächst betrachten wir den Kräfteausdruck (1.3) genauer. Schreiben wir die kinetische Energie als

$$T(q,\dot{q}) = \frac{1}{2}\dot{q}^T M(q)\dot{q}$$

und setzen dies in (1.3) ein, erhalten wir aus dem zweiten Term der rechten Seite den Ausdruck

$$-\left(\frac{\partial^2}{\partial q \partial \dot{q}} T(q,\dot{q})\right)\dot{q} = -\left(\frac{d}{dt} M(q)\right)\dot{q}$$

für die Corioliskräfte im System, die durch die Verwendung von *Relativkoordinaten* entstehen. Dies lässt auch eine alternative Darstellung nach Jay [24] der Bewegungsgleichungen zu, indem man diese Terme auf der linken Seite der Gleichungen einbindet. Wir erhalten im Fall ohne Zwangsbedingung die Gestalt

$$\frac{d}{dt}(M(q)\dot{q}) = \tilde{f}(q,\dot{q},t) \tag{1.6}$$

mit $\tilde{f}(q,\dot{q},t) = f(q,\dot{q},t) + \partial/\partial q\, M(q)\dot{q}$, analog mit Zwangsbedingungen.

Bemerkung 1.1 *In [28] leitet Lesser die Bewegungsgleichungen unter Verwendung von* `Maple V` *direkt aus dem Newtonschen Kraftgesetz her. Er verwendet lokale Koordinatensysteme und eliminiert die inneren Kräfte durch Projektion auf den Tangentialraum, die sogenannten* Gleichungen von Kane. *Diese Vorgehensweise ist aber weniger verbreitet.*

Konservative Systeme Bevor wir die unterschiedlichen Formulierungen der Bewegungsgleichungen genauer analysieren, möchten wir hier auf eine Vereinfachung eingehen, die für konservative Systeme, also energieerhaltende Systeme, vorgenommen werden kann.

In diesem Fall lassen sich die eingeprägten und äußeren Kräfte $Q(q,\dot{q},t)$ als Gradient eines Potentials $U(q)$ auffassen. Mit Hilfe der Lagrangefunktion $L(q,\dot{q})$

$$L(q,\dot{q}) := T(q,\dot{q}) - U(q)$$

und dem Hamilton-Prinzip

$$\delta \int_{t_0}^{t_1} L(q(t),\dot{q}(t))\mathrm{d}t = 0 \tag{1.7}$$

ergeben sich wiederum die Gleichungen (1.5), bzw. mit der um die Zwangsbedingungen und Lagrangemultiplikatoren erweiterten Lagrangefunktion

$$L(q,\dot{q}) = T(q,\dot{q}) - U(q) - g^T(q)\lambda \tag{1.8}$$

auch die Gleichungen (1.4). Hierbei wurden die Euler-Lagrange-Gleichungen der Variationsrechnung zugrunde gelegt.

Dieser Spezialfall der energieerhaltenden Systeme wird oft verwendet, um das qualitative Verhalten eines Systems ohne störende Einflüsse zu untersuchen. Obwohl dämpfende Komponenten durch Reibung in einem realen System immer vorhanden sind, sind sie schwer zu modellieren und werden daher oft vernachlässigt.

Flexible Mehrkörpersysteme Bisher haben wir ausschließlich Bewegungsgleichungen starrer Körper behandelt. Kommen elastische Körper hinzu, erhalten wir ein *flexibles Mehrkörpersystem*.

Die Bewegungsgleichung eines elastischen Körpers schreibt sich als

$$\rho(x)\ddot{u}(x,t) = \mathrm{div}\sigma(u(x,t)) + \beta(x,t) \tag{1.9}$$

mit Massendichte $\rho(x)$, Volumenkraft $\beta(x,t)$, Verschiebung $u(x,t)$ und Spannungstensor $\sigma(u(x,t))$, zur Herleitung der Gleichungen siehe z.B. [5]. Durch Überführung auf die schwache Form und geeignete Wahl von Ansatz- und Testfunktionen erhalten wir die semidiskretisierte Bewegungsgleichung

$$M_\triangle \ddot{d} = -K_\triangle d - D_\triangle \dot{d} + f_\triangle(d,\dot{d},t). \tag{1.10}$$

Die Gestalt der Massenmatrix $M_\triangle$, Steifigkeitsmatrix $K_\triangle$, Dämpfungsmatrix $D_\triangle$ und des Kraftvektors $f_\triangle$ wird in [42] hergeleitet.

In (1.10) liegt somit ein linearer Spezialfall von (1.5) vor. Auch im Fall von gekoppelten Systemen, die sowohl aus starren als auch elastischen Körpern bestehen, beschreiben obige Gleichungen die Situation. Wir erhalten ein System der Form

$$\begin{pmatrix} M(q) & C^T(q,d) \\ C(q,d) & M_\triangle \end{pmatrix} \cdot \begin{pmatrix} \ddot{q} \\ \ddot{d} \end{pmatrix} = \begin{pmatrix} f(q,\dot{q},d,\dot{d},t) \\ \tilde{f}_\triangle(q,\dot{q},d,\dot{d},t) \end{pmatrix}, \tag{1.11}$$

welches zusätzliche Kopplungsterme $C(q,d)$ enthält. Zu einer genaueren Analyse und Definition der einzelnen Terme siehe [42].

Die feste Form der Bewegungsgleichungen von starren und elastischen Körpern erlaubt eine gemeinsame Analyse sowohl des Systems als auch dessen Auswirkungen während einer Simulation in Bezug auf die Zeitintegration.

1.2 Zustandsform und Deskriptorform

Entscheidend für die Struktur der Gleichungen ist die Wahl der Koordinaten. Wählt man eine minimale Anzahl an freien Parametern, resultiert obige Vorgehensweise in einem System aus gewöhnlichen Differentialgleichungen (1.5), kurz ODE (ordinary differential equation).

Oft wird man dagegen auf zusätzliche Koordinaten zurückgreifen, in denen die Energieausdrücke einfacher aufzustellen sind. Über Zwangsbedingungen wird die Anzahl der Freiheitsgrade dann wieder reduziert. Es ergibt sich ein differential-algebraisches Gleichungssystem, kurz DAE (differential-algebraic equation), der Form (1.4). Die Form (1.5) nennt man die *Minimalform* oder *Zustandsform*, (1.4) die *Deskriptorform* des Systems.

Obwohl die Minimalform weniger komplex erscheint, birgt sie doch einige Nachteile. Zum Einen ist es zum Teil äußerst schwierig, überhaupt Minimalkoordinaten zu finden. Dies gilt besonders für den Fall geschlossener kinematischer Schleifen, siehe [42]. Zum Anderen zeichnen sich die Gleichungen (1.5) oft durch eine höhere Nichtlinearität aus, wie wir anhand eines Beispiels sehen werden.

Trotz zusätzlicher numerischer Probleme wird oft die Deskriptorform verwendet. Vor allem in Simulationspaketen, wo die Gleichungen automatisch erzeugt werden, kann man nicht auf sie verzichten. Die numerische Analyse von (1.4) ist daher von entscheidender Bedeutung.

1.2.1 Differential-algebraische Gleichungen

Die Gleichungen (1.4) bilden ein differential-algebraisches Gleichungssystem. Zur Klassifizierung von DAEs dient der Begriff des Indexes. In der Literatur werden verschiedene Indexbegriffe eingeführt. Wir gehen hier auf zwei Indexbegriffe ein, den *differentiellen Index* und den *Störungs- oder Pertubationsindex*.

Differentieller Index Der *differentielle Index* ist definiert als Anzahl der benötigten Ableitungen, um die DAE in eine ODE zu überführen. Je kleiner der differentielle Index, desto verwandter ist die DAE einer ODE und desto unproblematischer ist die numerische Behandlung.

Das System (1.4) besitzt den differentiellen Index 3. Um das zu sehen, differenzieren wir die Zwangsbedingung $0 = g(q)$ nach der Zeit t und erhalten die Geschwindigkeitsnebenbedingung

$$0 = G(q)\dot{q}. \tag{1.12}$$

Weiteres Differenzieren liefert die Beschleunigungsnebenbedingung

$$0 = G(q)\ddot{q} + \frac{\partial G(q)}{\partial q}(\dot{q}, \dot{q}). \tag{1.13}$$

Betrachtet man die erste Gleichung der DAE (1.4a) gemeinsam mit (1.13), so bilden diese ein Gleichungssystem für $\ddot{q}$ und λ. Um eine ODE zu erhalten, ist es

notwendig, dieses Gleichungssystem nach $\ddot{q}$ und λ aufzulösen. Dies ist gewährleistet, wenn die Matrix

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix} \tag{1.14}$$

regulär ist. Wir erhalten diese Voraussetzung im Fall einer regulären Matrix $M(q)$ und unabhängigen Zwangsbedingungen, da Letzteres die Regularität von G impliziert.

Nach dem Lösen des linearen Gleichungssystems differenzieren wir die Gleichung für λ und erhalten somit nach drei Differentiationen eine ODE. Da wir auf die Theorie gewöhnlicher Differentialgleichungen zurückgreifen können, ist auch die Existenz und Eindeutigkeit einer Lösung des Systems (1.4) gesichert.

Störungsindex Zur Definition des Störungsindexes betrachten wir die allgemeine Darstellung

$$F(y, y') = 0. \tag{1.15}$$

Im Gegensatz zum differentiellen Index gibt der Störungsindex an, wie sensibel die DAE auf kleine Störungen reagiert. Liegt eine gestörte Lösung mit einem Defekt $\delta(t)$ der Form

$$F(\hat{y}, \hat{y}') = \delta(t) \tag{1.16}$$

vor, dann besitzt das System (1.15) entlang einer Lösung y den Störungsindex m, falls für $0 < x \leq \bar{x}$ eine Abschätzung

$$||\hat{y}(t) - y(t)|| \leq C \left(||\hat{y}(0) - y(0)|| + \max_{0 \leq \xi \leq t} ||\delta(\xi)|| + \ldots + \max_{0 \leq \xi \leq t} ||\delta^{(m-1)}(\xi)|| \right) \tag{1.17}$$

existiert.

Enthält der Defekt $\delta(t)$ hochfrequente Anteile, gehen diese mit der $m-1$. Ableitung in die rechte Seite von (1.17) ein, die sehr groß werden kann. In diesem Fall wird bei einer Integration die Genauigkeit der numerischen Lösung beeinträchtigt.

Für die Gleichungen der Mehrkörperdynamik ergeben sich nach Arnold [1] Abschätzungen, die die zweite Ableitung der Störung enthalten, und somit auch der Störungsindex 3. Auffallend ist, dass die zweite Ableitung der Störung nicht in den Abschätzungen für alle Komponenten auftaucht, sondern nur bei den Lagrange-Multiplikatoren. Diese sind somit besonders anfällig, wenn hochfrequente Störungen im System auftreten.

Aufgrund dieses relativ hohen Indexes bereitet die Anwendung numerischer Verfahren auf die Deskriptorform mehr Probleme als die Zustandsform, da oft die Ordnung des Verfahrens für die Geschwindigkeits- und Lagrangekoordinaten reduziert wird. Dieses Phänomen wird in den Kap. 2 und 3 genauer untersucht.

1.2.2 Regularisierung

Um die Vorteile beider Formulierungen zu vereinen, werden oft die Zwangsbedingungen durch steife Federn und/oder Dämpfer ersetzt. Diese *Regularisierung* bewirkt dann, dass das System durch eine ODE der Form (1.5) beschrieben wird. Dabei wird allerdings übersehen, dass man durch das Erzeugen von steifen Federn ähnliche Probleme wie bei der Deskriptorform aufwirft.

Zu einer genauen Analyse extrahieren wir in der Zustandsform (1.5) die zusätzlichen Federn aus der rechten Seite, bezeichnen die Federkonstante mit $1/\epsilon^2$ und erhalten die Form

$$M(q)\ddot{q} = f_n(q,\dot{q}) - \frac{1}{\epsilon^2}\nabla U(q) \tag{1.18}$$

mit $\epsilon \ll 1$ nach Lubich [29]. Durch den letzten Term der rechten Seite wird eine singuläre Störung eingebracht, weswegen wir (1.18) auch als *singulär gestörtes Problem* bezeichnen. Zur Analyse setzen wir für die glatte Lösung q^ϵ die Entwicklung

$$q^\epsilon = q^0 + \epsilon^2 q^1 + \cdots + \epsilon^{2N} q^N + O(\epsilon^{2N+2}) \tag{1.19a}$$

$$\dot{q}^\epsilon = \dot{q}^0 + \epsilon^2 \dot{q}^1 + \cdots + \epsilon^{2N} \dot{q}^N + O(\epsilon^{2N+2}) \tag{1.19b}$$

mit Koeffizientenfunktionen $q^0, \dots q^N$ bzw. $\dot{q}^0, \dots \dot{q}^N$ an und bestimmen diese der Reihe nach durch Einsetzen von (1.19) in (1.18) und Taylorentwicklung. Für den Term zu ϵ^{-2} erhalten wir $\nabla U(q^0) = 0$ oder äquivalent dazu $g(q^0) = 0$. Zusammen mit dem Term zu ϵ^0

$$M(q^0)\ddot{q}^0 = f(q^0,\dot{q}^0) - \nabla^2 U(q^0)q^1$$

und der Bedingung

$$\nabla^2 U(q^0)q^1 = G^T(q^0)\lambda^0 \tag{1.20}$$

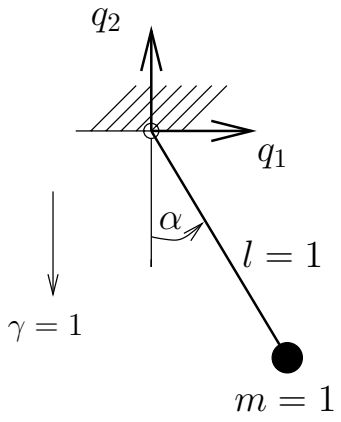
entspricht dies dem System (1.4) für den ersten Entwicklungsterm q^0. Die Bedingung (1.20) ersetzt die unbekannte Größe q^1 durch die Zwangskräfte λ^0 und erlaubt damit die Berechnung von q^0 unabhängig von den anderen Koeffizienten q^i, $i = 1, 2, \dots, N$.

Als Gleichungssystem für q^i mit $i \geq 1$ ergibt sich

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

mit ϕ^i linear in $\dot{q}^i$ und $\nabla^2 U(q^0) = G(q^0)^T K(q^0) G(q^0)$, zum Beweis siehe [29].

Die glatte Lösung des regularisierten Systems, an der wir normalerweise interessiert sind, ist somit der Lösung der DAE für kleine ϵ verwandt und wirft auch ähnliche numerische Probleme auf, wie wir im nächsten Kapitel sehen werden.



Minimalform:

$$\ddot{\alpha} = -\sin\alpha$$

Deskriptorform:

$$\begin{aligned} \ddot{q}_1 &= -q_1\lambda \\ \ddot{q}_2 &= -q_2\lambda - 1 \\ 0 &= q_1^2 + q_2^2 - 1 \end{aligned}$$

Regularisierte Bewegungsgleichungen:

$$\ddot{q}_1 = -\frac{1}{\epsilon^2}\frac{q_1}{\sqrt{q_1^2+q_2^2}}\left(\sqrt{q_1^2+q_2^2}-1\right)$$

$$\ddot{q}_2 = -\frac{1}{\epsilon^2}\frac{q_1}{\sqrt{q_1^2+q_2^2}}\left(\sqrt{q_1^2+q_2^2}-1\right) - 1$$

Abbildung 1.2: Pendel mit einem Freiheitsgrad.

1.2.3 Beispiel Pendel

Zur Veranschaulichung betrachten wir als einfaches Beispiel ein Pendel mit einem Freiheitsgrad, das in Abb. 1.2 dargestellt ist. Links ist die schematische Skizze abgebildet, rechts die Bewegungsgleichungen in den unterschiedlichen Formulierungen.

Die Minimalform besteht aus nur einer Gleichung und ist nichtlinear. Dagegen besteht die Deskriptorform aus insgesamt drei Gleichungen, wovon die beiden ersten Gleichungen linear sind. Die Nichtlinearität liegt allein in den Zwangsbedingungen vor.

Ersetzt man die Zwangsbedingung der Deskriptorformulierung durch eine Feder mit Steifigkeit $1/\epsilon^2$, ergibt sich der dritte Satz an Gleichungen. Erneut liegt eine hohe Nichtlinearität vor, aber die numerischen Probleme werden vor allem durch den Vorfaktor $1/\epsilon^2$ verursacht.

Wir haben in diesem Unterkapitel die Vor- und Nachteile unterschiedlicher Formulierungen kennengelernt. Um aber einige Eigenschaften von numerischen Verfahren zur Lösung der Bewegungsgleichungen zu analysieren, ist es zweckmäßig, noch auf eine andere Form einzugehen.

1.3 Hamilton-Systeme

Im folgenden Abschnitt werden Eigenschaften von Hamilton-Systemen diskutiert. Obwohl sie äquivalent zu den entsprechenden Gleichungen aus dem Lagrange-Formalismus sind, kann man anhand ihrer Struktur bestimmte zusätzliche Untersuchungen vornehmen.

Ein Hamilton-System schreibt sich als

$$\dot{p} = -\frac{\partial}{\partial q} H(p,q) \tag{1.21a}$$

$$\dot{q} = \frac{\partial}{\partial p} H(p,q) \tag{1.21b}$$

mit der Hamilton-Funktion

$$H(p,q) = p^T \dot{q} - L(q,\dot{q}) \tag{1.22}$$

bzw. mit der für konservative Mehrkörpersysteme äquivalenten Form

$$H(p,q) = \underbrace{1/2 \dot{q}^T M(q) \dot{q}}_{T(q,\dot{q})} + U(q) \tag{1.23}$$

und $\dot{q} = \dot{q}(p,q)$. Die Hamilton-Funktion entspricht damit der Gesamtenergie des Systems, wobei p und q die Impulse und Auslenkungen bezeichnen. Die spezielle Struktur der Hamiltonfunktion erlaubt uns einen einfachen Zugang zu Erhaltungseigenschaften von numerischen Verfahren.

Zur Umrechnung von den Geschwindigkeiten zu den Impulsen verwendet man

$$p = \frac{\partial}{\partial q} L(q,\dot{q}),$$

bzw. im Spezialfall der Mehrkörpersysteme $p = M(q)\dot{q}$. Die Transformation der Lagrange-Gleichungen zu Hamilton-Gleichungen kann zu erheblichen Problemen führen, da dazu die Inverse der Massenmatrix benötigt wird, welche schnell sehr kompliziert werden kann. Bei konstanter Massenmatrix sind allerdings beide Formulierungen auch bei der numerischen Behandlung äquivalent.

Zwei Eigenschaften solcher Systeme sind für uns von Bedeutung:

1. Energieerhaltung und
2. Symplektizität $\hat{=}$ Volumenerhaltung.

Wir werden diese im Folgenden erläutern.

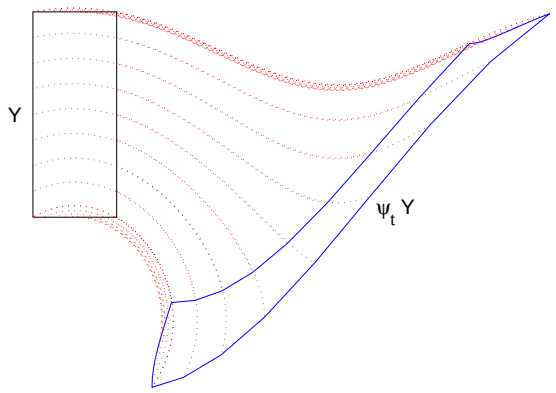


Abbildung 1.3: Transformation einer Menge $\mathcal{Y}$ unter dem Fluss ψ_t.

Energieerhaltung Wie oben bereits beschrieben, ist die Hamiltonfunktion für mechanische Systeme gleich der Gesamtenergie des Systems. Wegen (1.21) gilt

$$\frac{dH}{dt}(p,q) = H_p\dot{p} + H_q\dot{q} = 0,$$

daher ist die Energieerhaltung in Hamiltonsystemen entlang einer Lösung (p,q) direkt erfüllt und lässt sich daher leicht untersuchen.

Symplektizität Der Begriff der Symplektizität, der gleichbedeutend ist mit Volumenerhaltung, beinhaltet ein Konzept für den Einfluss der Anfangswerte auf die Lösung. Wir definieren dazu für die Differentialgleichung $y' = f(y,t)$ den Fluss $\psi_t(y_0)$ als

$$\psi_t(y_0) = y(t;t_0,y_0),$$

der die Anfangswerte auf die Lösung zum Zeitpunkt t abbildet. Erweitert man dies auf eine Menge von Anfangswerten $\mathcal{Y}$ ergibt sich

$$\psi_t\mathcal{Y} = \{y|\, y = y(t;t_0,y_0),\ y_0 \in \mathcal{Y}\},$$

siehe dazu Abb. 1.3.

Kommen wir auf die Symplektizität zurück. Um eine Differentialgleichung daraufhin zu untersuchen, berechnen wir das Volumen von $\mathcal{Y}$ unter Wirkung von ψ_t zu

$$\mathrm{Vol}(\psi_t\mathcal{Y}) = \int_{\psi_t\mathcal{Y}} \mathrm{d}y = \int_{\mathcal{Y}} \left|\det\left(\frac{\partial y}{\partial y_0}(t;t_0,y_0)\right)\right| \mathrm{d}y_0.$$

Nach [19, S. 99] ist dieser Ausdruck äquivalent zu

$$\mathrm{Vol}(\psi_t\mathcal{Y}) = \int_{\mathcal{Y}} \exp\left(\int_{t_0}^{t} \mathrm{Spur}\left(J_f(y(s;t_0,y_0))\right)\,\mathrm{d}s\right)\,\mathrm{d}y_0$$

mit J_f gleich der Jacobi-Matrix von f, was uns eine einfache Möglichkeit liefert, Symplektizität zu untersuchen.

Definition 1.1 *Der Fluss ψ_t heisst volumenerhaltend, falls gilt*

$$Vol(\psi_t \mathcal{Y}) = Vol(\mathcal{Y})$$

für alle $t \geq t_0$, wobei $Vol(\mathcal{Y})$ das Volumen bzw. die Fläche von $\mathcal{Y}$ bezeichnet. Speziell für Hamilton-Systeme nennt man diese Eigenschaft Symplektizität*, da sie sich in diesem Fall auf eine quadratische Form zurückführen lässt.*

Wie leicht zu sehen ist, gilt die Beziehung

$$\text{Spur}(J_f(y)) = 0 \quad \Rightarrow \quad \psi_t \text{ symplektisch},$$

und wegen

$$J_f(p,q) = \begin{pmatrix} -\frac{\partial^2}{\partial q \partial p} H(p,q) & -\frac{\partial^2}{\partial q^2} H(p,q) \\ \frac{\partial^2}{\partial p^2} H(p,q) & \frac{\partial^2}{\partial q \partial p} H(p,q) \end{pmatrix}$$

für das Hamilton-System (1.21) ist somit die Symplektizität ebenfalls direkt in die Hamilton-Formulierung integriert, so dass auf sie auch leicht zugegriffen werden kann.

Im folgenden Kapitel werden wir untersuchen, wie sich Runge-Kutta-Verfahren bei der Anwendung auf steife mechanische Systeme und DAEs verhalten und Bedingungen stellen, die ein Verfahren für die Anwendung auf solche Systeme auszeichnen. Im darauffolgenden Kapitel wird dann eine spezielle Klasse, die *konvex kombinierten Lobatto-Verfahren* eingeführt und daraufhin untersucht.

Kapitel 2

Implizite Runge-Kutta Verfahren

Zur Simulation von steifen mechanischen Systemen werden besondere Anforderungen an die numerischen Integrationsverfahren gestellt. Ähnlich wie bei den klassischen *steifen Systemen* z.B. aus der chemischen Reaktionskinetik neigen explizite Verfahren aufgrund von fehlender Stabilität zu sehr kleinen Schrittweiten, die aber bei der gewünschten Auflösung nicht notwendig sind. Wählt man dagegen ein implizites Verfahren geringer Ordnung wie z.B. den impliziten Euler, treten zwar keine Stabilitätsprobleme mehr auf, dafür sind nichtlineare Gleichungen zu lösen.

Die Klasse der Runge-Kutta-Verfahren, kurz RK-Verfahren, erlaubt durch die große Anzahl an freien Parametern das Bilden von Verfahren hoher Ordnung. Da man in vielen Fällen nicht auf die Implizitheit der Verfahren verzichten kann, ist man bestrebt, effiziente Verfahren zu konstruieren, die den zusätzlichen Rechenaufwand durch Stabilität ausgleichen.

Im Gegensatz zu den BDF-Verfahren [12] gehören RK-Verfahren zu den Einschrittverfahren, so dass keine separate Berechnung der ersten Schritte benötigt wird. Zudem ermöglicht das Einfügen von eingebetteten Verfahren durch die Wiederverwendung vorheriger Ergebnisse eine billige Schrittweitensteuerung.

Zunächst wollen wir die allgemeine Klasse der Runge-Kutta-Verfahren und ihre wichtigsten Eigenschaften in Bezug auf steife mechanische Systeme vorstellen, bzw. entsprechende Bedingungen an die Koeffizienten herleiten. Dazu werden grundlegende Eigenschaften definiert und die Stabilität und Konvergenz von steifen mechanischen Systemen analysiert. Zum Abschluss dieses Kapitels werden Eigenschaften von RK-Verfahren beschrieben, die auf Hamilton-Systeme aufbauen. Die Darstellung der meisten Eigenschaften in diesem Kapitel ist an [20] angelehnt.

2.1 Aufbau und Ordnung von RK-Verfahren

Zur Einführung der RK-Verfahren wenden wir uns zunächst wieder von der Formulierung als singulär gestörtes Problem (1.18) ab und betrachten die Gleichungen in der Standardformulierung als System 1. Ordnung

$$y' = f(y, t). \tag{2.1}$$

Dann schreibt sich ein s-stufiges RK-Verfahren als

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} f(Y_j, t_0 + c_j h), \quad i = 1, \ldots, s \tag{2.2a}$$

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i f(Y_i, t_0 + c_i h), \tag{2.2b}$$

wobei y_i jeweils die numerische Approximation zu $y(t_i)$ und h die Schrittweite darstellt. Die *Inkremente oder Zuwächse* Y_i kann man als Näherungen der exakten Lösung an den Knoten c_i interpretieren. Die *Koeffizienten* a_{ij}, *Gewichte* b_i und *Knoten* c_i legen die Eigenschaften des Verfahrens fest, die im Allgemeinen als sogenanntes *Butcher-Tableau*

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

mit $c = (c_1, \ldots, c_s)^T$, $b = (b_1, \ldots, b_s)^T$ und $A = (a_{ij})_{ij}$ angegeben werden. Als verkürzende Schreibweise verwenden wir auch (A, b, c).

Man unterscheidet aufgrund der Struktur der Koeffizientenmatrix A folgende Fälle:

- *Explizite RK-Verfahren*: $a_{ij} = 0$ für $i \leq j$,
- *Diagonalimplizite RK-Verfahren*: $a_{ij} = 0$ für $i < j$, mindestens ein $a_{ii} \neq 0$ und
- *Implizite RK-Verfahren*, kurz IRK-Verfahren, sonst.

Explizite Verfahren sind besonders effizient, da die Näherungslösung im neuen Zeitschritt direkt berechnet werden kann, ohne dass ein nichtlineares Gleichungssystem zu lösen ist. Wir werden aber sehen, dass sie aufgrund von Stabilitätsaussagen nicht für die Anwendung auf steife mechanische Systeme geeignet sind.

Im Fall von diagonalimpliziten Verfahren ist zwar ein nichtlineares Gleichungssystem zu lösen, dieses zerfällt aber in s Gleichungssysteme, die separat behandelt werden können. Owren/Simonsen [32] untersuchen den Spezialfall von Verfahren, bei denen die Diagonale der Koeffizientenmatrix jeweils mit demselben Eintrag besetzt ist, sogenannte SDIRK-Verfahren (Singly Diagonal Implicit RK). Dadurch kann man den Rechenaufwand weiter reduzieren.

Wir konzentrieren uns auf den allgemeinen Fall der impliziten RK-Verfahren, da diese in den meisten Fällen zur Berechnung von steifen mechanischen Systemen wegen ihrer Stabilitätseigenschaften bevorzugt werden.

Ordnungsbedingungen

Definition 2.1 *Ein numerisches Verfahren besitzt die* **klassische Ordnung** p*, wenn ein* $K > 0$ *existiert mit*

$$||y(t_0 + h) - y_1|| \leq Kh^{p+1}.$$

Sie beschreibt damit das Verhalten des lokalen Fehlers und ist eine der wichtigsten Eigenschaften von numerischen Verfahren.

Zur Herleitung der Ordnungsbedingungen an die freien Koeffizienten a_{ij}, b_i und c_i wollen wir auf die *B-Reihen-Theorie* zurückgreifen, die in Kap. 2.5 eingeführt wird. Sie ermöglicht eine prägnante und übersichtliche Darstellung. Wir werden sie daher erst nach Einführung dieser Theorie ausführen.

Die Bedingungen bis Ordnung $p = 4$ sind in Tab. 2.1 für den Spezialfall von autonomen Systemen

$$y' = f(y)$$

zusammengestellt. Nichtautonome Systeme kann man durch die Transformation

$$\begin{pmatrix} y \\ t \end{pmatrix}' = \begin{pmatrix} f(y,t) \\ 1 \end{pmatrix} \tag{2.3}$$

in ein solches umformen. Zusätzlich wird dann aber die Bedingung

$$\sum_{j=1}^{s} a_{ij} = c_i, \quad i = 1, \ldots, s \tag{2.4}$$

vorausgesetzt.

Im folgenden Abschnitt wollen wir vereinfachende Annahmen aufstellen, die nicht nur den Umgang mit den Ordnungsbedingungen erleichtern, sondern auch für sich alleine genommen von Bedeutung sind.

p	Bedingungen	p	Bedingungen
1	$\sum_{j=1}^{s} b_j = 1$	2	$\sum_{j,k=1}^{s} b_j a_{jk} = 1/2$
3	$\sum_{j,k,l=1}^{s} b_j a_{jk} a_{jl} = 1/3$ $\sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} = 1/6$	4	$\sum_{j,k,l,m=1}^{s} b_j a_{jk} a_{jl} a_{jm} = 1/4$ $\sum_{j,k,l,m=1}^{s} b_j a_{jk} a_{kl} a_{jm} = 1/8$ $\sum_{j,k,l,m=1}^{s} b_j a_{jk} a_{kl} a_{km} = 1/12$ $\sum_{j,k,l,m=1}^{s} b_j a_{jk} a_{kl} a_{lm} = 1/24$

Tabelle 2.1: Ordnungsbedingungen für RK-Verfahren bis Ordnung $p = 4$.

Vereinfachende Annahmen Betrachten wir im Folgenden die Bedingungen

$$\mathcal{B}(\xi): \qquad \sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q}, \qquad q = 1, \ldots, \xi, \tag{2.5a}$$

$$\mathcal{C}(\eta): \qquad \sum_{j=1}^{s} a_{ij} c_j^{q-1} = \frac{c_i^q}{q}, \qquad q = 1, \ldots, \eta, \quad i = 1, \ldots, s, \tag{2.5b}$$

$$\mathcal{D}(\zeta): \qquad \sum_{i=1}^{s} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1 - c_j^q), \quad q = 1, \ldots, \zeta, \quad j = 1, \ldots, s, \tag{2.5c}$$

die wir *vereinfachende Annahmen* nennen wollen.

Folgender Satz von Butcher bildet einen Zusammenhang zwischen den vereinfachenden Annahmen und der klassischen Ordnung.

Satz 2.1 *Ein RK-Verfahren besitzt die klassische Ordnung ξ, wenn die Koeffizienten die Bedingungen $\mathcal{B}(\xi)$, $\mathcal{C}(\eta)$ und $\mathcal{D}(\zeta)$ mit $\xi \leq \eta + \zeta + 1$ und $\xi \leq 2\eta + 2$ erfüllen.*

Beweis: Ein Vergleich von den Ordnungsbedingungen aus Tab. 2.1 mit den vereinfachenden Annahmen $\mathcal{C}(\eta)$ zeigt, dass sukzessives Anwenden von Letzterem auf die Ordnungsbedingungen zur linken Seite von (2.5a) führt, so dass die Bedingung $\mathcal{B}(\xi)$ dann die Gültigkeit der jeweiligen Ordnungsbedingung zur Folge hat.

Als Beispiel betrachten wir die vierte Bedingungsgleichung der Ordnung 4 aus Tab. 2.1. Es ergibt sich

$$\sum_{j,k,l,m=1}^{s} b_j a_{jk} a_{kl} a_{lm} \overset{\mathcal{C}(1)}{=} \sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} c_l \overset{\mathcal{C}(2)}{=} \frac{1}{2} \sum_{j,k=1}^{s} b_j a_{jk} c_k^2 \overset{\mathcal{C}(3)}{=} \frac{1}{6} \sum_{j=1}^{s} b_j c_j^3 \overset{\mathcal{B}(4)}{=} \frac{1}{24}.$$

Liegt $\eta < \xi - 1$ vor, kann man die höheren Bedingungen durch $\mathcal{D}(\zeta)$ mit $\zeta = \xi - \eta - 1$ ersetzen, die im Prinzip von der anderen Seite an die Bedingungsgleichungen herangehen. Dieses Ersetzen besitzt jedoch seine Grenzen, da $\mathcal{D}(\zeta)$ nicht bei Verzweigungen verwendet werden kann. Daher wird $\xi \leq 2\eta + 2$ vorausgesetzt.

Um die Grenzen von der Anwendbarkeit der Bedingungen $\mathcal{D}$ auszutesten, analysieren wir die zweite Bedingung vierter Ordnung der Tab. 2.1. Wir erhalten

$$\sum_{j,k,l,m=1}^{s} b_j a_{jk} a_{kl} a_{jm} \overset{\mathcal{C}(1)}{=} \sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} c_j \overset{\mathcal{D}(2)}{=} \frac{1}{2} \sum_{k,l=1}^{s} b_k (1 - c_k^2) a_{kl}$$
$$\overset{\mathcal{C}(1)}{=} \frac{1}{2} \sum_{k=1}^{s} b_l (1 - c_k^2) c_k \overset{\mathcal{B}(2),\mathcal{B}(4)}{=} \frac{1}{4} - \frac{1}{8} = \frac{1}{8}.$$

Die Bedingung $\mathcal{C}(1)$ ist also notwendig im ersten Schritt, da $\mathcal{D}(2)$ sonst nicht angewendet werden kann. Die zweite Verwendung von $\mathcal{C}(1)$ könnte auch durch $\mathcal{D}(3)$ ersetzt werden, welches aber nach den Voraussetzungen nicht gegeben sein muss. □

Sehen wir uns die Bedingungen einzeln an. $\mathcal{B}(\xi)$ hängt nur von b und c ab und kann somit als Ordnung der Knoten und Gewichte interpretiert werden. Bei den gebräuchlichen IRK-Verfahren stimmt ξ mit der klassischen Ordnung p überein, siehe obiger Satz.

Für steife mechanische Systeme und DAEs ist die *Stufenordnung* η von zentraler Bedeutung, die durch $\mathcal{C}(\eta)$ festgelegt wird. Sie beschreibt die Ordnung der Defekte Δ_0, die durch

$$y(t_0 + c_i h) = y(t_0) + h \sum_{j=1}^{s} a_{ij} y'(t_0 + c_j h) + \Delta_0$$

gegeben sind, also wenn die exakte Lösung in die RK-Formulierung eingesetzt wird. Um das zu sehen, entwickelt man einerseits die exakte Lösung $y(t_0 + c_i h)$ und andererseits die Ableitung derselben $y'(t_0 + c_j h)$ in eine Taylorreihe und vergleicht die Koeffizienten, nachdem Letztere oben eingesetzt wurde.

Aufgrund dieser Eigenschaft hat die Stufenordnung großen Einfluss auf die Konvergenzordnung von mechanischen Systemen, wie wir im Abschnitt zu DAEs näher erläutern werden.

Zuletzt analysieren wir die Bedingungen $\mathcal{D}(\zeta)$. Während $\mathcal{B}(\xi)$ und $\mathcal{C}(\eta)$ jeweils nur von der Koeffizientenmatrix A oder den Gewichten b abhängen, tauchen in $\mathcal{D}(\zeta)$ beide Größen auf. Die Bedingungen kann man daher als Bindeglied ansehen, damit aus Stufenordnung und Ordnung der Gewichte und Knoten die klassische Ordnung folgt. Für sich alleine genommen sind sie nicht von großer Bedeutung.

2.2 Stabilität von RK-Verfahren

2.2.1 Stabilitätsfunktion

In diesem Abschnitt stehen die Eigenschaften im Vordergrund, die sich aus der linearen Testgleichung

$$y' = \lambda y \tag{2.6}$$

mit $\lambda \in \mathbb{C}$ herleiten lassen. Dabei wird vor allem das qualitative Verhalten eines numerischen Verfahrens in Bezug zur exakten Lösung $y(t) = c \cdot \exp(\lambda t)$ untersucht.

Definition 2.2 *Die* **Stabilitätsfunktion** $\mathcal{R}(z)$ *berechnet sich aus*

$$y_1 = \mathcal{R}(z) y_0 \tag{2.7}$$

mit $z = \lambda h$. Für IRK-Verfahren besitzt sie die Form

$$\mathcal{R}(z) = 1 - z b^T (I - zA)^{-1} \mathbb{1}, \quad \mathbb{1} = (1, \ldots, 1)^T. \tag{2.8}$$

Bemerkung 2.1 *Im Fall von linearen ODE-Systemen*

$$y' = My$$

wird durch Hauptachsentransformation ein System der Form (2.6) gebildet. Dann entspricht λ gerade den Eigenwerten von M.

Wegen

$$y(t_0 + h) = \exp(z) y(t_0)$$

können wir die Stabilitätsfunktion (2.8) als Approximation an die Exponentialfunktion interpretieren. Dies erlaubt uns folgende Definition:

Definition 2.3 *Gilt für ein numerisches Verfahren*

$$\exp(z) - \mathcal{R}(z) = C z^{r_l + 1} + \mathcal{O}\left(z^{r_l + 2}\right), \quad C \neq 0,$$

dann besitzt es die **lineare Ordnung** r_l.

Schreiben wir die Stabilitätsfunktion als rationale Funktion

$$\mathcal{R}(z) = \frac{\mathcal{P}(z)}{\mathcal{Q}(z)}$$

mit $\mathcal{P} \in \Pi_k$ und $\mathcal{Q} \in \Pi_j$, besitzen die meisten gebräuchlichen RK-Verfahren die maximale lineare Ordnung, die für k und j erreichbar ist. In diesem Fall stimmt die Stabilitätsfunktion mit der eindeutigen (k, j)-*Padé-Approximation* überein.

Definition 2.4 *Die* (k, j)**-Padé-Approximation** *ist definiert durch*

$$\mathcal{R}_{kj}(z) = \frac{\mathcal{P}_{kj}(z)}{\mathcal{Q}_{kj}(z)}$$

mit

$$\mathcal{P}_{kj}(z) = 1 + \frac{k}{j+k}z + \ldots + \frac{k(k-1)\cdots 1}{(j+k)\cdots(j+1)} \cdot \frac{z^k}{k!}$$
$$\mathcal{Q}_{kj}(z) = 1 - \frac{j}{k+j}z - \ldots + (-1)^j \frac{j(j-1)\cdots 1}{(k+j)\cdots(k+1)} \cdot \frac{z^j}{j!} = \mathcal{P}_{jk}(-z).$$

RK-Verfahren mit $\mathcal{R}(z) = \mathcal{R}_{kj}(z)$ *besitzen die lineare Ordnung* $j + k$*.*

Für explizite Verfahren gilt im Allgemeinen $k > j$, was schlechte Stabilitätseigenschaften und damit auch Probleme bei steifen mechanischen Systemen nach sich zieht.

A- und L-Stabilität Für $Re(\lambda) < 0$ klingt die exakte Lösung ab und dieses Verhalten ist auch für die numerische Lösung wünschenswert. Da die Eigenwerte steifer mechanischer Systeme nahe an der imaginären Achse liegen, verlangen wir zusätzlich, dass Oszillationen durch das numerische Verfahren realistisch dargestellt werden.

Definition 2.5 *Die Menge*

$$\{z \in \mathbb{C} |\, |\mathcal{R}(z)| \leq 1\}$$

nennen wir das **Stabilitätsgebiet** *eines numerischen Verfahrens.*

Das Stabilitätsgebiet gibt an, in welchen Bereichen von λ wir eine numerische Lösung erhalten, die nicht aufschwingt. Soll ein RK-Verfahren ohne Einschränkung an die Schrittweite h auf ein System mit Eigenwerten $Re(\lambda) < 0$ oder $Im(\lambda) \gg 1$ angewendet werden, ist es sinnvoll, dass die gesamte linke Halbebene zum Stabilitätsgebiet gehört. Dies führt auf den Begriff der *A-Stabilität*:

Definition 2.6 *Ein numerisches heißt* **A-stabil***, falls*

$$|\mathcal{R}(z)| \leq 1 \quad \textit{für} \quad Re(z) \leq 0.$$

Wie wir später sehen werden, reicht dieser Begriff noch nicht aus, um eine stabile Integration zu gewährleisten. Dazu benötigen wir die

Definition 2.7 *Ein RK-Verfahren heißt* **L-stabil**, *falls*

$$|\mathcal{R}(z)| = 0 \quad \textit{für} \quad Re(z) \to -\infty.$$

Es lässt sich zeigen [20, S. 58], dass ein RK-Verfahren nur dann A- und L-stabil sein kann, wenn für die zugehörige (k, j)-Padé-Approximation die Bedingung $k \leq j$ gilt. Daher können explizite RK-Verfahren wegen $k > j$ nicht diesen Anforderungen genügen. Für die Anwendung beschränken wir uns daher auf implizite RK-Verfahren.

Aufgrund der Struktur der Stabilitätsfunktion liegt für L-stabile Verfahren abklingendes Verhalten auch für Systeme mit $Im(\lambda) \to \infty$ vor, und damit sind sie hervorragend für steife mechanische Systeme geeignet. Die Eigenschaft, dass numerische Verfahren abklingende Lösungskurven erzeugen, nennt man *numerische Dämpfung*. Ein Maß dafür liefert

Definition 2.8 *Die Größe* $\rho(\chi) := |\mathcal{R}(i\chi)|$ *mit* $\chi = \omega h$ *heißt der* **Spektralradius** *eines numerischen Verfahrens. Er gibt an, wie stark Eigenwerte auf der imaginären Achse numerisch gedämpft werden.*

Um zu entscheiden, wo numerische Dämpfung erwünscht und wo unerwünscht ist, teilen wir auf in die Bereiche $\chi < \pi$ und $\chi > \pi$, siehe [14]. Bei der numerischen Berechnung können wir davon ausgehen, dass Schwingungen mit $\chi < \pi$ in der numerischen Lösung zu sehen sind, da im Grenzfall die Schrittweite h gerade der halben Periodendauer entspricht. In diesem Fall ist numerische Dämpfung daher unerwünscht.

In dem zweiten Bereich $\chi > \pi$ geht es um Schwingungen, deren Frequenz oberhalb der höchsten auflösbaren Frequenz dieser Schrittweite liegt, so dass wir sie als Störungen interpretieren können. Daher wird dort maximale numerische Dämpfung verlangt.

Obige Definitionen sind alle auf die absolute Stabilität eines Verfahrens aufgebaut, und darauf ausgelegt, dass es eine abklingende bzw. zumindest keine aufklingende Lösungskurve hervorbringt. Der folgende Begriff betrachtet dagegen die Relation zur exakten Lösung $\exp(z)$:

Definition 2.9 (Relatives Stabilitätsgebiet) *Die Menge*

$$\{z \in \mathbb{C} |\, |\mathcal{R}(z)| < |e^z|\}$$

nennen wir das **Relative Stabilitätsgebiet** *eines numerischen Verfahrens, auch* **Ordnungssterne (engl. order stars)** *genannt.*

Die Bezeichnung *Ordnungssterne* wurde gewählt, da die Graphiken dieser Gebiete ein Sternmuster enthalten. Die Anzahl der Arme des Sterns gibt dabei gerade die lineare Ordnung des Verfahrens an. Wir werden in Kapitel 3 anhand von Beispielen näher darauf eingehen.

Dispersions- und Dissipationsordnung Die bisherigen Definitionen beruhen alle auf der Testgleichung (2.6) mit $\lambda \in \mathbb{C}$, wobei besonders der Grenzwert $Re(\lambda) \to -\infty$ berücksichtigt wird. Für mechanische Systeme interessiert uns aber mehr der Fall $Re(\lambda) \leq 0$ und $Im(\lambda) \gg 1$, da die Eigenwerte von Mehrkörpersystemen nahe der imaginären Achse liegen.

Daher gehen wir für die folgende Untersuchung auf die Testgleichung 2. Ordnung

$$\ddot{y} = -\omega^2 y \tag{2.9}$$

über, die eine Analyse für schwingende Systeme erlaubt und auf Owren/Simonsen [32] zurückzuführen ist. Wenden wir ein RK-Verfahren auf (2.9) an geschrieben als System 1. Ordnung, so erhalten wir

$$\begin{pmatrix} y_1 \\ y_1' \end{pmatrix} = \mathcal{R}(hK) \cdot \begin{pmatrix} y_0 \\ y_0' \end{pmatrix} \quad \text{mit} \quad K = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}$$

und der Stabilitätsfunktion $\mathcal{R}(z)$. Ein Abgleich mit der exakten Lösung von (2.9) zu den Anfangswerten $(y_0, y_0')^T = (1, 1)^T$

$$y(t) = \exp(i\omega t)$$

erlaubt uns, den Fehler des RK-Verfahrens getrennt nach Dispersion und Dissipation, also Imaginär- und Realteil von λ, zu untersuchen.

Dazu berechnen wir den Amplitudenfehler $d_a(\chi)$ und den Phasenfehler $d_p(\chi)$ aus dem Ansatz

$$\mathcal{R}(i\chi) = \exp[(-d_a(\chi) + i d_p(\chi))\, \chi] \tag{2.10}$$

und erhalten die Formeln

$$d_a(\chi) = -\frac{\ln[\mathcal{R}(i\chi) \cdot \mathcal{R}(-i\chi)]}{2\chi}, \tag{2.11a}$$

$$d_p(\chi) = \frac{\ln[\mathcal{R}(i\chi) \cdot \mathcal{R}(-i\chi)^{-1}]}{2i\chi}, \tag{2.11b}$$

die für $d_a = 0$ und $d_p = 1$ die exakte Lösung liefern. Um diese Fehlerterme auf ihre Ordnung zu untersuchen, benötigen wir die Taylorreihen

$$d_a(\chi) = c_a \chi^{r_a} + \mathcal{O}\left(\chi^{r_a+1}\right), \tag{2.12a}$$
$$d_p(\chi) - 1 = c_p \chi^{r_p} + \mathcal{O}\left(\chi^{r_p+1}\right), \tag{2.12b}$$

deren Existenz wegen $\mathcal{R}(0) = 1$ in einer Umgebung um $\chi = 0$ gesichert ist.

Definition 2.10 *Den Grad r_p der Taylorreihe von $d_p(\chi) - 1$ nennen wir die* **Dispersionsordnung** *des Verfahrens, entsprechend den Grad r_a von $d_a(\chi)$ die* **Dissipationsordnung**. *Die führenden Koeffizienten c_p und c_a bezeichnen wir mit* **Phasenverschiebung** *bzw.* **Amplitudenfaktor**.

Mit diesen Definitionen haben wir eine separate Ordnung für den Phasen- und Amplitudenfehler gefunden und können bei der Auswahl eines RK-Verfahrens darauf zurückgreifen. Selbstverständlich sind diese Begriffe nicht unabhängig von der klassischen Ordnung und der linearen Ordnung. Dazu gilt der folgende Satz:

Satz 2.2 *Zwischen den einzelnen Ordnungsbegriffen dieses Kapitels bestehen folgende Zusammenhänge:*

a) $min(r_a, r_p) = r_l$,

b) $p \leq r_l$.

Beweis:

a) Dazu setzen wir die Reihenentwicklungen (2.12a) und (2.12b) in die Formel (2.10) ein und erhalten

$$\begin{aligned}
\mathcal{R}(i\chi) &= \exp[(-d_a(\chi) + i d_p(\chi))\,\chi] \\
&= \exp[(c_a\chi^{r_a} + i(1 + c_p\chi^{r_p}))\,\chi] + \mathcal{O}\left(\chi^{min(r_a,r_p)+2}\right) \\
&= \exp[i\chi] + \tilde{c}\chi^{min(r_a,r_p)+1} + \mathcal{O}\left(\chi^{min(r_a,r_p)+2}\right) \\
&= \exp[i\chi] + \tilde{c}\chi^{r_l+1} + \mathcal{O}\left(\chi^{r_l+2}\right),
\end{aligned}$$

wobei in der letzten Zeile die Definition der linearen Ordnung r_l eingesetzt wurde.

b) Ergibt sich direkt aus den Definitionen, da die lineare Ordnung nach demselben Schema aber nur zu einer Untermenge an Systemen bestimmt wird.

2.3 Nichtlineare Eigenschaften von RK-Verfahren

Nach diesen Eigenschaften, die auf lineare Testgleichungen aufgebaut sind, wollen wir uns in diesem Abschnitt mit nichtlinearen Eigenschaften von impliziten RK-Verfahren beschäftigen. Zunächst führen wir drei weiterführende Stabilitätsbegriffe ein, nämlich AN-Stabilität, B-Stabilität und algebraische Stabilität. Daraufhin liefert uns die Koerzivität die Existenz und Eindeutigkeit von Lösungen.

2.3.1 Weiterführende Stabilitätsbegriffe

Wie lassen sich A- und L-Stabilität auf nichtlineare Systeme erweitern? Diese Frage kann man nicht so einfach beantworten. Sinnvoll wäre ein Stabilitätsbegriff, der auf Energieerhaltung beruht, da diese hinter vielen technischen Anwendungen steht. Dies stellt sich aber als zu schwierig heraus.

AN-Stabilität Ein einfacher Schritt zur Verallgemeinerung der oben eingeführten Begriffe ist der Übergang zu der nichtautonomen linearen Testgleichung

$$y' = \lambda(t)y.$$

Ähnlich wie zuvor können wir eine Stabilitätsfunktion definieren über

$$y_1 = \mathcal{K}(Z)y_0 \quad \text{mit} \quad \mathcal{K}(Z) = 1 + b^T Z(I - AZ)^{-1}\mathbb{1}$$

und $Z = \text{diag}(z_1, \ldots, z_s)$, $z_j = h\lambda(t_0 + c_j h)$.

Definition 2.11 *Ein RK-Verfahren heißt* **AN-stabil**, *falls*

$$|\mathcal{K}(Z)| \leq 1 \quad \textit{für} \quad Re(z_j) \leq 0, \quad j = 1, \ldots, s.$$

Wegen

$$\mathcal{K}(\text{diag}(z, \ldots, z)) = \mathcal{R}(z)$$

folgt sofort, dass AN-stabile Verfahren auch A-stabil sind.

Obwohl AN-Stabilität nur eine kleine Erweiterung darstellt, sind wir damit unserem Ziel schon deutlich näher gekommen. Wir gehen im Folgenden auf ein nichtlineares Stabilitätskonzept über.

B-Stabilität Wie oben schon angesprochen, eignet sich das Konzept der Energieerhaltung nicht, um einen weiterführenden Stabilitätsbegriff zu definieren. Wie in Kap. 1 gesehen, besitzen mechanische Systeme außerdem die Eigenschaft der *Volumenerhaltung*. Daraus lässt sich der Begriff der *B-Stabilität* ableiten.

Betrachten wir eine nichtlineare Differentialgleichung

$$y' = f(y,t),$$

die einer einseitigen Lipschitzbedingung

$$\langle f(y,t) - f(\hat{y},t), y - \hat{y}\rangle \leq \nu ||y - \hat{y}||_2^2 \tag{2.13}$$

mit einer einseitigen Lipschitzkonstanten ν genügt. Dann gilt für zwei beliebige Lösungen y und $\hat{y}$ mit $t \geq t_0$

$$||y(t) - \hat{y}(t)||_2 \leq ||y(t_0) - \hat{y}(t_0)||_2 \cdot e^{\nu(t-t_0)}.$$

Zwei beliebige Lösungen einer ODE mit $\nu < 0$ nähern sich mit der Zeit an, und diese Eigenschaft wünscht man sich auch für numerische Lösungen.

Definition 2.12 (B-Stabilität) *Ein RK-Verfahren heißt* **B-stabil**, *falls aus der einseitigen Lipschitzbedingung (2.13) mit* $\nu = 0$

$$\langle f(y,t) - f(\hat{y},t), y - \hat{y}\rangle \leq 0$$

für alle $h \geq 0$ *folgt*

$$||y_1 - \hat{y}_1||_2 \leq ||y_0 - \hat{y}_0||_2.$$

Dabei bezeichnen y_1 *und* $\hat{y}_1$ *die numerischen Approximationen nach einem Schritt mit Anfangswerten* y_0 *bzw.* $\hat{y}_0$.

Wie zuvor bei der AN-Stabilität folgt auch aus der B-Stabilität die A-Stabilität. Zum Beweis setzt man die lineare Testgleichung (2.6) in (2.13) ein und erhält nach Einsetzen der Stabilitätsfunktion die Behauptung.

Algebraische Stabilität Der Begriff der B-Stabilität ist etwas unhandlich und schwer zu überprüfen. Für RK-Verfahren mit disjunkten Knoten c_i ist er aber äquivalent zur *algebraischen Stabilität*.

Definition 2.13 (Algebraische Stabilität) *Ein Runge-Kutta-Verfahren heißt* **algebraisch stabil**, *falls*

a) $b_i \geq 0$ *für* $i = 1, \ldots, s$ *und*

b) $M = (m_{ij}) = (b_i a_{ij} + b_j a_{ji} - b_i b_j)_{ij=1}^{s}$ *positiv semidefinit ist.*

Zum Abschluss dieses Abschnitts wollen wir wieder einen Zusammenhang zwischen den einzelnen Stabilitätsbegriffen herstellen.

Satz 2.3 *Für RK-Verfahren mit disjunkten Knoten* c_i *sind die Konzepte von AN-Stabilität, B-Stabilität und algebraischer Stabilität äquivalent.*

Zum Beweis dieses Satzes siehe z.B. [20, S. 186].

Mit diesen für die in dieser Arbeit betrachteten RK-Verfahren äquivalenten Konzepten haben wir ein starkes Mittel in der Hand, um die Stabilität von numerischen Verfahren bei der Anwendung auf nichtlineare Differentialgleichungen zu untersuchen. Vorteilhaft ist hierbei auch, dass jeweils die A-Stabilität daraus folgt, so dass nicht etwas Neues definiert wurde, sondern man die nichtlineare Stabilität als Erweiterung der linearen Konzepte betrachten kann.

Damit möchten wir das Thema Stabilität vorerst abschließen und wenden uns nun dem zu lösenden nichtlinearen Gleichungssystem zu.

2.3.2 Existenz und Eindeutigkeit von IRK-Lösungen

Um die Lösung eines impliziten RK-Verfahrens zu erhalten, ist es notwendig, ein nichtlineares Gleichungssystem zu lösen. In diesem Abschnitt stehen Aussagen über die Existenz und Eindeutigkeit von dessen Lösung im Vordergrund. Dazu wird vorausgesetzt, dass die Differentialgleichung der einseitigen Lipschitzbedingung (2.13) genügt.

Zunächst führen wir den Begriff der *Koerzivität* ein.

Definition 2.14 *Der* Koerzivitätskoeffizient α_0 *ist gegeben durch*

$$\alpha_0(A^{-1}) = \sup_{D>0} \alpha_D(A^{-1}), \tag{2.14}$$

wobei $\alpha_D(A^{-1})$ *abhängig von einer Diagonalmatrix* D *mit positiven Einträgen* $d_i > 0$ *den kleinsten Eigenwert der Matrix*

$$\frac{1}{2}\left[D^{1/2}A^{-1}D^{-1/2} + (D^{1/2}A^{-1}D^{-1/2})^T\right]$$

bezeichnet.

Zum Teil ist auch eine äquivalente Definition hilfreich.

Lemma 2.4 *Der Koerzivitätskoeffizient $\alpha_D(A^{-1})$ berechnet sich als größtes α mit*

$$\langle u, A^{-1}u\rangle_D \geq \alpha\langle u, u\rangle_D \quad \textit{für alle} \quad u \in \mathbb{R}^s.$$

Dann gilt nach [20] folgendes Lemma

Lemma 2.5 *Sei f eine stetig differenzierbare Funktion, die die einseitige Lipschitzbedingung (2.13) erfüllt. Dann existiert eine eindeutige Lösung $(Y_1, \ldots, Y_s)$ von (2.2a), falls die Koeffizientenmatrix A invertierbar ist und die Bedingung*

$$h\nu < \alpha_0(A^{-1})$$

erfüllt ist.

Dies liefert uns die Existenz und Eindeutigkeit der IRK-Lösung in Abhängigkeit von der einseitigen Lipschitzkonstanten ν. Für die Anwendung auf steife mechanische Systeme sollte das Verfahren die Eigenschaft

$$\alpha_0(A^{-1}) \geq 0$$

besitzen, um ohne Einschränkung der Schrittweite eine eindeutige Lösung zu erhalten.

Die Berechnung von $\alpha_0(A^{-1})$ ist nicht immer einfach. Man erhält aber eine untere Schranke durch jedes $\alpha_D(A^{-1})$. Eine obere Schranke findet man, wenn man die äquivalente Definition aus Lemma 2.4 zugrundelegt. Dann liefert das Einsetzen von $u = e_i$ die Schranke

$$\alpha_0(A^{-1}) \leq \min_{i=1,\ldots,s} \omega_{ii} \tag{2.15}$$

mit $A^{-1} = (\omega_{ij})_{ij}$.

Die Einschränkung, dass die Koeffizientenmatrix A invertierbar sein muss, kann abgeschwächt werden, falls sie sich in der Form

$$A = \begin{pmatrix} 0 & 0 \\ a & \tilde{A} \end{pmatrix} \qquad \text{oder} \qquad A = \begin{pmatrix} \tilde{A} & 0 \\ a^T & 0 \end{pmatrix}$$

mit invertierbarer Matrix $\tilde{A}$ schreiben lässt. Dann gilt obiger Satz, nachdem $\alpha_0(A^{-1})$ durch $\alpha_0(\tilde{A}^{-1})$ ersetzt wurde.

2.4 Konvergenz

In diesem Kapitel werden die Ergebnisse aus [17] und [23] zusammengefasst, um die Konvergenz von RK-Verfahren angewandt auf differential-algebraische und steife Systeme zu untersuchen. Die Aussagen werden abhängig von den vereinfachenden Annahmen $\mathcal{B}, \mathcal{C}$ und $\mathcal{D}$ getroffen. Zusätzliche Vorteile bringen sogenannte *steifgenaue* Verfahren.

Definition 2.15 (steifgenau) *Wir nennen ein RK-Verfahren* **steifgenau**, *falls* $a_{si} = b_i$ *für* $i = 1, \ldots, s$.

Bei steifgenauen Verfahren stimmt die letzte Zeile der Koeffizientenmatrix mit dem Gewichtsvektor b überein. Desweiteren folgt zwingend $c_s = 1$ für autonome Systeme. Solche Verfahren werten demnach die numerische Lösung zum Zeitschritt $t_0 + h$ bereits bei der Berechnung der internen Stufen und nicht separat aus und sind daher effizienter. Es zeigt sich in den folgenden Untersuchungen, dass sie besonders für die Anwendung auf DAEs und damit auch für steife mechanische Systeme geeignet sind.

Konvergenz für differential-algebraische Gleichungen Aufgrund der Stabilitätsanforderungen, die steife mechanische Systeme an numerische Verfahren stellen, sind nur wenige Verfahren zu ihrer Integration geeignet. Ähnlich sieht es bei der Konvergenz aus, da wir mit dem Phänomen der Ordnungsreduktion rechnen müssen.

In [23, Th. 6.1] untersucht Jay die Konvergenzordnung für Runge-Kutta-Verfahren angewandt auf DAEs vom Index 3. Die Hauptaussage wird im folgenden Satz zusammengefasst, den wir ohne Beweis zitieren.

Satz 2.6 *Wir betrachten eine differential-algebraische Gleichung mit Index* 3 *der Form*

$$y' = f(y, z), \tag{2.16a}$$
$$z' = k(y, z, u), \tag{2.16b}$$
$$0 = g(y) \tag{2.16c}$$

mit konsistenten Anfangswerten (y_0, z_0, u_0) *zum Zeitpunkt* t_0 *und ein RK-Verfahren, dass auf (2.16) angewandt wird. Die RK-Koeffizienten genügen den Bedingungen* $\mathcal{B}(\xi)$, $\mathcal{C}(\eta)$ *mit* $\eta \geq 2$ *und* $\mathcal{D}(\zeta)$. *Zusätzlich sei die Koeffizientenmatrix*

invertierbar und steifgenau. Dann erhalten wir für $t_n - t_0 = nh \leq$ Konst *den globalen Fehler*

$$\begin{aligned} y_n - y(t_n) &= \mathcal{O}(h^{\min(\xi, 2\eta-2, \eta+\zeta)}) \\ z_n - z(t_n) &= \mathcal{O}(h^{\eta}) \\ u_n - u(t_n) &= \mathcal{O}(h^{\eta-1}). \end{aligned}$$

Gilt zusätzlich $k_{uu} = 0$ *in (2.16b), ergibt sich*

$$y_n - y(t_n) = \mathcal{O}(h^{\min(\xi, 2\eta-1, \eta+\zeta)}),$$

wobei auch auf die Bedingung $\mathcal{C}(2)$ *verzichtet werden kann.*

Vor allem in den z und u-Komponenten liegt demnach eine deutliche Ordnungsreduktion vor, die im Fall von Mehrkörpersystemen den Geschwindigkeiten und Lagrange-Multiplikatoren entsprechen. Wir werden im nächsten Abschnitt sehen, dass dies auch Auswirkungen auf die Konvergenzordnung von steifen mechanischen Systemen hat.

Wegen der auftretenden Probleme bei der direkten Anwendung auf Index 3 Probleme wird häufig der Index vorher reduziert, indem man nicht die Lagezwangsbedingung aus (1.4) verwendet, sondern die Geschwindigkeits- oder Beschleunigungsnebenbedingung (1.12) oder (1.13). Dies führt dann auf ein DAE-System vom Index 2 bzw. 1.

Bei diesem Vorgehen tritt aber ein anderes Problem auf, der sogenannte *Drift-off-Effekt*. Auch wenn man mit Anfangswerten startet, die die ursprüngliche Lagezwangsbedingung erfüllen, ist dies für spätere Zeitpunkte nicht mehr gewährleistet, da man nur die Einhaltung der differenzierten Bedingungen fordert.

Oft wird daher nach jedem Schritt auf die Mannigfaltigkeit projiziert, die durch die Lagezwangsbedingung aufgespannt wird, siehe z.B. [30]. Alternativ kann man auch mit zusätzlichen Lagrange-Multiplikatoren die ursprüngliche Nebenbedingung ankoppeln, die sogenannte GGL-Stabilisierung nach Gear, Gupta, Leimkuhler [13]. Es ergibt sich ein System der Form

$$\dot{q} = v + G^T(q)\mu, \tag{2.17a}$$

$$M(q)\dot{v} = f(q, v, t) - G^T(q)\lambda, \tag{2.17b}$$

$$0 = G(q)v, \tag{2.17c}$$

$$0 = g(q) \tag{2.17d}$$

mit Index 2, welches für den zusätzlichen Lagrange-Multiplikator $\mu = 0$ analytisch äquivalent zu (1.4) ist.

Konvergenz für steife mechanische Systeme Zum Abschluss dieser Sektion gehen wir wieder auf die Formulierung eines steifen mechanischen Systems als singulär gestörtes System (1.18) zurück. Wir wissen aus Kap. 1, dass die analytischen Lösungen von (1.18) und (1.4) nahe beieinander liegen. Es ist daher naheliegend, dass sich auch numerische Integrationsverfahren bei der Anwendung auf solche Systeme ähnlich verhalten. Eine detaillierte Aussage liefert uns der folgende Satz von Lubich [29].

Satz 2.7 *Ein RK-Verfahren erfülle die Bedingungen*

a) $\mathcal{C}(\eta)$,

b) Koeffizientenmatrix A *invertierbar und* $|\mathcal{R}(\infty)| < 1$,

c) Keine Eigenwerte von A *auf der imaginären Achse und* $|\mathcal{R}(i\chi)| < 1$ *für* $\chi \in \mathbb{R}\backslash\{0\}$.

Ferner gehen wir davon aus, dass die Anfangswerte $(q_0, \dot{q}_0)$ *auf der glatten Lösung von (1.18) liegen. Dann existiert für* $0 < \epsilon \leq h \leq h_0$ *mit* h_0 *hinreichend klein und unabhängig von* ϵ *eine eindeutige RK-Lösung (2.2) von (1.18) mit dem globalen Fehler*

$$q_n^\epsilon - q^\epsilon(t_n) = q_n^0 - q^0(t_n) + O(\epsilon^2 h^{\eta-2}) \tag{2.18a}$$
$$\dot{q}_n^\epsilon - \dot{q}^\epsilon(t_n) = \dot{q}_n^0 - \dot{q}^0(t_n) + O(\epsilon^2 h^{\eta-2}) \tag{2.18b}$$

für $0 \leq t_n \leq T$. *Dabei bezeichnen* q_n^ϵ *und* $\dot{q}_n^\epsilon$ *die RK-Lösungen des singulär gestörten Problems (1.18) und* q_n^0 *und* $\dot{q}_n^0$ *die RK-Lösungen der DAE (1.4).*

Beweisidee:

1. Zeige, dass die RK-Lösungen q_n^ϵ und $\dot{q}_n^\epsilon$ eine Entwicklung der Form

$$q_n^\epsilon = q_n^0 + \epsilon^2 q_n^1 + \cdots + \epsilon^{2k} q_n^k + r_n$$
$$\dot{q}_n^\epsilon = \dot{q}_n^0 + \epsilon^2 \dot{q}_n^1 + \cdots + \epsilon^{2k} \dot{q}_n^k + \dot{r}_n$$

   besitzen und $||r_n|| + ||\dot{r}_n|| = \mathcal{O}(\epsilon^{2k})$ für $\eta = 2k$ bzw. $||r_n|| + ||\dot{r}_n|| = \mathcal{O}(h\epsilon^{2k})$ für $\eta = 2k+1$, siehe [29, Th. 6.1].

2. Zeige die Fehlerabschätzungen

$$q_n^k - q^k(t_n) = \mathcal{O}(h^{\eta-2k}), \qquad \dot{q}_n^k - \dot{q}^k(t_n) = \mathcal{O}(h^{\eta-2k})$$

   für differential-algebraische Systeme vom Index $2k+3$, siehe [29, Th. 4.2].

3. Kombination dieser beiden Schritte ergibt

$$q_n^\epsilon - q^\epsilon(t_n) = \epsilon^0(q_n^0 - q^0(t_n)) + \cdots + \epsilon^{2k}(q_n^k - q^k(t_n)) + \mathcal{O}(h\epsilon^{2k})$$
$$\dot{q}_n^\epsilon - \dot{q}^\epsilon(t_n) = \epsilon^0(\dot{q}_n^0 - \dot{q}^0(t_n)) + \cdots + \epsilon^{2k}(\dot{q}_n^k - \dot{q}^k(t_n)) + \mathcal{O}(h\epsilon^{2k})$$

für $\eta = 2k+1$, $\eta = 2k$ analog. Da $\epsilon \ll h$ liefert dies zusammen mit der Existenz und Eindeutigkeit von RK-Lösungen die Behauptung. □

Dieser Satz zeigt, dass nicht nur analytisch ein Zusammenhang zwischen dem singulär gestörten Problem (1.18) und der Deskriptorform (1.4) besteht, sondern dass auch die numerische Integration ähnliche Probleme aufwirft. Geht man von $\epsilon \ll h$ aus, sind die Terme, die ϵ enthalten, vernachlässigbar gegenüber höheren Potenzen von h, so dass wir in diesem Fall für steife mechanische Systeme die gleichen Konvergenzeigenschaften erhalten wie für differential-algebraische Gleichungen vom Index 3. Nur der Fehler in den Lagrange-Multiplikatoren kann außer Acht gelassen werden.

2.5 B-Reihen-Theorie

Wendet man RK-Verfahren auf Hamilton-Systeme an, stellt sich die Frage, inwiefern sich die in Kap. 1 erwähnten Eigenschaften auf die numerische Lösung übertragen. Von besonderem Interesse sind *symplektische Verfahren*, die nicht nur die Symplektizität des mechanischen Flusses erhalten, sondern auch positive Eigenschaften bezüglich der Energieerhaltung aufweisen. Um Letzteres zu sehen, folgen wir der Darstellung aus [16] und führen eine Rückwärtsanalyse durch.

Definition 2.16 *Wir betrachten ein RK-Verfahren angewandt auf ein Hamilton-System mit genügend glatter rechter Seite. Dann heißt es* **symplektisch**, *falls der dazugehörige diskrete Fluss φ aus*

$$y_1 = \varphi(y_0)$$

symplektisch ist.

Ausgehend von dieser Definition ergibt dies für die Koeffizienten eines RK-Verfahren den

Satz 2.8 *Erfüllen die Koeffizienten eines RK-Verfahrens die Bedingung*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \quad \textit{für alle} \quad i,j = 1,\ldots,n, \tag{2.19}$$

dann ist es symplektisch.

Beweis siehe [19, S. 316].

Interessant ist hierbei, dass sich die Bedingung an dieselbe Matrix M knüpft wie bei der Definition der algebraischen Stabilität. Dort waren die Eigenwerte der Matrix M von Bedeutung, während hier das Verschwinden jeder Komponente verlangt wird.

2.5.1 Grundlagen der B-Reihen-Theorie

Die symplektischen Integratoren haben nicht nur die Eigenschaft, dass sie die Volumen erhalten. Die uns interessierende Aussage von Hairer [16] lautet, dass der Energiefehler beschränkt bleibt, wenn Hamilton-Systeme mit symplektischen Verfahren integriert werden. Da wir die Inhalte von [16] später spezifizieren wollen, ist es notwendig, die einzelnen Schritte des Beweises genauer zu betrachten.

Nach einer Idee von Butcher [7] kann die Taylorreihe einer RK-Lösung in Bezug auf die Schrittweite h auch geschrieben werden als

$$y_1 = y_0 + \sum_{u \in \mathcal{T}} \frac{h^{\rho(u)}}{\rho(u)} \alpha(u) \left(\gamma(u) \sum_{i=1}^{s} b_i \phi_i(u) \right) F(u)(y_0)$$

mit

$\mathcal{T}$ Menge der Wurzelbäume,

$\rho(u)$ Anzahl der Knoten des Baumes u (Ordnung des Baumes),

$\alpha(u)$ Anzahl der möglichen monotonen Indizierungen von u,

$\gamma(u)$ ein Integer-Koeffizient,

$\phi_i(u)$ ein Ausdruck, der von den RK-Koeffizienten a_{ij} abhängt,

$F(u)$ das elementare Differential abhängig von $f(y)$.

Bezeichne τ den einzigen Baum der Ordnung 1 und

$$u = [u_1, \ldots, u_m]$$

den Baum mit einer Wurzel und m Zweigen, an denen die Bäume $u_1, \ldots, u_m$ angehängt werden. Die Größen $\mu_1, \mu_2, \ldots$ zählen die Anzahl der gleichen Bäume in $u_1, \ldots, u_m$. Jede der von u abhängigen Größen wird mit den Startwerten

$$\rho(\tau) = 1, \qquad \alpha(\tau) = 1, \qquad \gamma(\tau) = 1, \qquad \phi_i(\tau) = 1, \qquad F(\tau)(y) = f(y)$$

initialisiert und über

$$\rho(u) = 1 + \rho(u_1) + \ldots + \rho(u_m), \tag{2.20a}$$

$$\alpha(u) = \begin{pmatrix} \rho(u) - 1 \\ \rho(u_1), \ldots, \rho(u_m) \end{pmatrix} \cdot \alpha(u_1) \cdot \alpha(u_m) \cdot \frac{1}{\mu_1! \mu_2! \cdots}, \tag{2.20b}$$

$$\gamma(u) = \rho(u) \cdot \gamma(u_1) \cdot \ldots \cdot \gamma(u_m), \tag{2.20c}$$

$$\phi_i(u) = \sum_{j_1,\ldots,j_m} a_{ij_1}\phi_{j_1}(u_1) \cdot \ldots \cdot a_{ij_m}\phi_{j_m}(u_m), \tag{2.20d}$$

$$F(u)(y) = f^{(m)}(y) \cdot (F(u_1)(y), \ldots, F(u_m)(y)) \tag{2.20e}$$

rekursiv definiert.

Definition 2.17 *Eine Reihe der Form*

$$y_1 = y_0 + \sum_{u \in \mathcal{T}} \frac{h^{\rho(u)}}{\rho(u)} \alpha(u) \mathrm{a}(u) F(u)(y_0) \tag{2.21}$$

bezeichnet man als **B-Reihe**.

Im unserem Fall wird die B-Reihe durch

$$\mathrm{a}(u) = \gamma(u) \sum_{i=1}^{s} b_i \phi_i(u) \tag{2.22}$$

gegeben.

Beispiele In Tab. 2.2 sind sämtliche Bäume bis Ordnung 5 zusammen mit $\alpha(u)$ und $\gamma(u)$ dargestellt. Letztere kann man direkt mit der Rekursionsformel bestimmen, für $\alpha(u)$ zählt man die Anzahl der monotonen Indizierungen von u, d.h., man verteilt eine Indexmenge, z.B. $\{i, j, k, l\}$ so, dass abgehende Bäume jeweils einen höheren Index erhalten, Sprünge zu höheren Indizes sind dabei erlaubt. Allerdings werden Bäume mit zwei gleichen Teilbäumen nur einfach gezählt.

Zur Bestimmung von $\phi_i(u)$ nimmt man sich eine der monotonen Indizierungen vor, multipliziert für jeden Zweig von Knoten i nach Knoten j den Term a_{ij} und summiert über alle Indizes der Indexmenge außer der Wurzel. Es ergibt sich z.B.

$$\phi_i \left(\begin{array}{ccc} & \mathrm{k} & \\ \mathrm{j} & & \mathrm{l} \\ & \mathrm{i} & \end{array} \right) = \sum_{j,k,l} a_{ij} a_{jk} a_{il}.$$

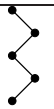
Ordnung									
1	u	•							
	$\alpha(u)$	1							
	$\gamma(u)$	1							
2	u								
	$\alpha(u)$	1							
	$\gamma(u)$	2							
3	u								
	$\alpha(u)$	1	1						
	$\gamma(u)$	6	3						
4	u								
	$\alpha(u)$	1	1	3	1				
	$\gamma(u)$	24	12	8	4				
5	u									
	$\alpha(u)$	1	1	3	1	4	3	4	6	1
	$\gamma(u)$	120	60	40	20	30	20	15	10	5

Tabelle 2.2: Bäume bis Ordnung 5.

Satz 2.9 *Ein RK-Verfahren besitzt die klassische Ordnung p genau dann, wenn die Beziehung*

$$\sum_{j=1}^{s} b_j \phi_j(u) = \frac{1}{\gamma(u)} \quad bzw. \quad \mathrm{a}(u) = 1 \tag{2.23}$$

für alle Bäume u mit $\rho(u) \leq p$ erfüllt ist.

Beweis: Die exakte Lösung von (2.1) lautet

$$y^{(m)}(t_0) = \sum_{u \in \mathcal{T}, \rho(u)=m} \alpha(u) F(u)(y_0).$$

Um das zu sehen, betrachten wir zunächst obige Gleichung für $m = 1$. Es ergibt sich daraus gerade die Differentialgleichung (2.1). Für höhere Ableitungen sind in Abb. 2.1 die Bäume zusammen mit ihren elementaren Differentialen dargestellt.

Die Ableitung $m = k$ leitet sich jeweils aus $m = k - 1$ her, indem man an jedem Knoten des Baumes einen zusätzlichen Zweig anbringt. Daraus ergeben sich aber gerade die Bäume der Ordnung k, wenn man gleiche Bäume mit unterschiedlichen

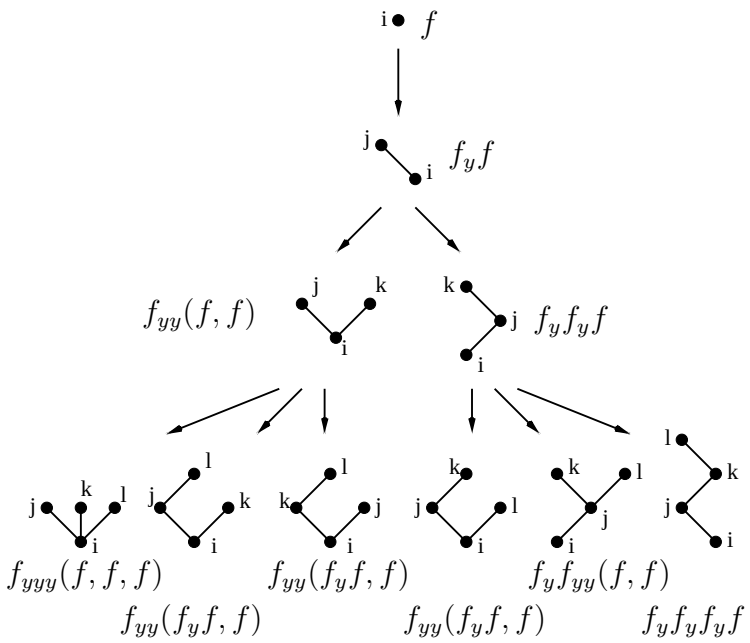


Abbildung 2.1: Ableitungen der exakten Lösung.

Indizierungen zusammenfasst. Dies wird durch den Faktor $\alpha(u)$ wieder ausgeglichen. Die Ordnungsbedingungen ergeben sich durch Koeffizientenvergleich der numerischen Lösung mit der exakten Lösung. □

Partitionen Um später die oben erwähnten Aussagen über symplektische Verfahren treffen zu können, benötigen wir die

Definition 2.18 *Eine* **Partition** (u, S) *von* $u \in \mathcal{T}$ *in* k *Unterbäume* $s_1, \ldots, s_k$ *ist die Menge* S *mit* $k-1$ *Zweigen von* u*, so dass die Bäume* $s_1, \ldots, s_k$ *entstehen, wenn man die Zweige von* S *aus* u *entfernt. Mit* $\alpha(u, S)$ *bezeichnen wir die Anzahl der monotonen Indizierungen von* u *unter der zusätzlichen Einschränkung, dass die Knoten jedes Unterbaums aufeinanderfolgend indiziert sind.*

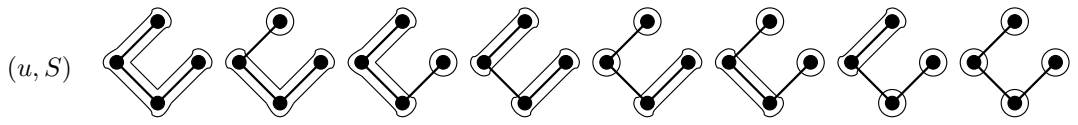


Abbildung 2.2: Alle Partitionen von $u = [[\tau], \tau]$ und Werte $\alpha(u, S)$.

In Abb. 2.2 sind die Partitionen des Baumes $u = [[\tau], \tau]$ mit jeweils dem Wert für $\alpha(u, S)$ aufgelistet. Wie man leicht sieht, gilt

$$\alpha(u, S) \leq \alpha(u),$$

da durch die zusätzliche Einschränkung bei der Partitionierung nicht alle monotonen Indizierungen mitgezählt werden. Mit diesen Begriffen haben wir alles zusammen, was wir für die folgende Rückwärtsanalyse benötigen.

2.5.2 Rückwärtsanalyse

Um Eigenschaften von numerischen Lösungen zu untersuchen, ist es sinnvoll, die Differentialgleichung zu betrachten, die von der numerischen Lösung exakt gelöst wird. Dieses Vorgehen nennt man *Rückwärtsanalyse*. Eine dazugehörige Aussage liefert uns der folgende Satz.

Satz 2.10 *Sei* $\mathrm{a} : \mathcal{T} \to \mathbb{R}$ *gegeben und die rechte Seite der ODE (2.1)* N*-mal stetig differenzierbar. Dann gilt für die numerische Lösung* y_1*, die durch das RK-Verfahren (2.2) definiert wird,*

$$y_1 = \tilde{y}(t_0 + h) + \mathcal{O}(h^{N+1}),$$

wobei $\tilde{y}(t)$ *die exakte Lösung der gestörten Differentialgleichung*

$$\tilde{y}' = \sum_{\rho(u) \leq N} \frac{h^{\rho(u)-1}}{\rho(u)!} \alpha(u) \mathrm{b}(u) F(u)(\tilde{y}) \tag{2.24}$$

ist. Die Koeffizienten $\mathrm{b}(u)$ *sind implizit definiert über*

$$\mathrm{a}(u) = \sum_{k=1}^{\rho(u)} \frac{1}{k!} \sum_{(u,S)} \binom{\rho(u)}{\rho(s_1), \ldots, \rho(s_k)} \frac{\alpha(u, S)}{\alpha(u)} \mathrm{b}(s_1) \cdot \ldots \mathrm{b}(s_k). \tag{2.25}$$

Dabei wird in der zweiten Summe über alle Partitionen von u *in* k *Unterbäume* $s_1, \ldots, s_k$ *summiert.*

Zum Beweis dieses Satzes siehe [16].

Die Formeln (2.25) bis $\rho(u) = 3$ lauten

$$\begin{aligned}
\mathrm{a}(\bullet) &= \mathrm{b}(\bullet), \\
\mathrm{a}(\bullet\!\!-\!\!\bullet) &= \mathrm{b}(\bullet\!\!-\!\!\bullet) + \mathrm{b}(\bullet)^2, \\
\mathrm{a}(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= \mathrm{b}(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + 3/2 \cdot \mathrm{b}(\bullet\!\!-\!\!\bullet)\mathrm{b}(\bullet) + \mathrm{b}(\bullet)^3, \\
\mathrm{a}(\bullet\!\!<\!\!\bullet) &= \mathrm{b}(\bullet\!\!<\!\!\bullet) + 3 \cdot \mathrm{b}(\bullet\!\!-\!\!\bullet)\mathrm{b}(\bullet) + \mathrm{b}(\bullet)^3.
\end{aligned}$$

Da in der Gleichung zu $\mathrm{a}(u)$ außer dem Term $\mathrm{b}(u)$ nur Bäume geringerer Ordnung involviert sind, ist die Lösbarkeit der Gleichungen garantiert.

Bemerkung 2.2 *Für Methoden der Ordnung p gilt*

$$\mathrm{b}(u) = \begin{cases} 1 & \textit{für} \quad u = \tau, \\ 0 & \textit{für} \quad 2 \le \rho(u) \le p, \\ \mathrm{a}(u) - 1 & \textit{für} \quad \rho(u) = p + 1. \end{cases} \tag{2.26}$$

Daher ist die Differenz der rechten Seiten von (2.24) und (2.1) von der Größenordnung $\mathcal{O}(h^p)$ und der führende Term entspricht dem Term für den lokalen Abschneidefehler.

Die Gleichungen (2.25) stellen einen direkten Zusammenhang zwischen den Koeffizienten $\mathrm{a}(u)$ der B-Reihe und der zugehörigen gestörten ODE mit den Koeffizienten $\mathrm{b}(u)$ her. Die einfache Berechnung von $\mathrm{b}(u)$ ist entscheidend für die folgenden Aussagen.

Eigenschaften symplektischer Integratoren Um die Eigenschaften symplektischer Integratoren in Bezug auf die Rückwärtsanalyse zu untersuchen, ist es sinnvoll, die Symplektizitätsbedingung für RK-Verfahren (2.19) in Abhängigkeit von $\mathrm{a}(u)$ und anderen Größen der B-Reihen-Theorie darzustellen.

Satz 2.11 *Ein RK-Verfahren ist symplektisch genau dann, wenn die Koeffizienten der B-Reihe (2.21) die Bedingung*

$$\frac{\mathrm{a}(v \circ w)}{\gamma(v \circ w)} + \frac{\mathrm{a}(w \circ v)}{\gamma(w \circ v)} = \frac{\mathrm{a}(v)}{\gamma(v)} \cdot \frac{\mathrm{a}(w)}{\gamma(w)}, \quad \textit{für} \quad v, w \in \mathcal{T} \tag{2.27}$$

erfüllen. Dabei bezeichnet $v \circ w$ den Baum v, an dessen Wurzel die Wurzel von w angehängt wird, also

$$v \circ w = [v_1, \ldots, v_m, w].$$

Beweis: Wir zeigen nur die Hinrichtung. Dazu gehen wir von der Gleichung (2.19)

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \quad i, j = 1, \ldots, n,$$

aus und multiplizieren mit $\phi_i(v) \cdot \phi_j(w)$. Nach Summation über $i, j = 1, \ldots, s$ erhalten wir die Beziehung

$$\sum_{i=1}^{s} b_i \phi_i(v \circ w) + \sum_{j=1}^{s} b_j \phi_j(w \circ v) - \left(\sum_{i=1}^{s} b_i \phi_i(v) \right) \left(\sum_{j=1}^{s} b_j \phi_j(w) \right) = 0, \quad v, w \in \mathcal{T}.$$

Mithilfe von (2.22) ist dies äquivalent zu (2.27). □

Im nächsten Satz wollen wir die Aussage von Satz 2.10 noch erweitern. Wir gehen nun davon aus, dass unsere Differentialgleichung als Hamilton-System vorliegt, und charakterisieren diejenigen RK-Verfahren, die durch Rückwärtsanalyse wieder auf ein Hamilton-System führen.

Bevor wir diesen Satz aufstellen und beweisen können, sind noch einige Vorarbeiten notwendig. Wir benötigen neben den elementaren Differentialen (2.20e) auch entsprechende *elementare Hamilton-Funktionen*, um aus der rechten Seite von (2.24) die gestörte Hamilton-Funktion konstruieren zu können. Da wir diese in Abhängigkeit von den elementaren Differentialen ausdrücken wollen, müssen wir zwischen

$$f_1(y,t) = -\frac{\partial H}{\partial q}(p,q) \quad \text{und} \quad f_2(y,t) = \frac{\partial H}{\partial p}(p,q)$$

unterscheiden. Daher bezeichnen wir mit $F(v)(p,q)$ die Differentiale mit Wurzel f_1 und mit $F(w)(p,q)$ die mit Wurzel f_2.

Auch bei der Darstellung als Bäume müssen wir diese Unterscheidung berücksichtigen und führen daher *schwarze* ($\bullet$) und *weiße* ($\circ$) Knoten für f_1 bzw. f_2 ein. Den Baum v intepretieren wir dann jeweils als Baum mit schwarzer Wurzel und schreiben $v \in \mathcal{T}_1$, entsprechend $w \in \mathcal{T}_2$. Damit können wir folgende Aussage aufstellen:

Satz 2.12 *Die Differentialgleichung (2.24) mit den elementaren Differentialen (2.20e) eines Hamilton-Systems*

$$f_1(y,t) = -\frac{\partial H}{\partial q}(p,q)$$
$$f_2(y,t) = \frac{\partial H}{\partial p}(p,q)$$

mit $y = (p,q)^T$ ist ein Hamilton-System genau dann, wenn

$$\frac{\mathrm{b}(v \circ w)}{\gamma(v \circ w)} + \frac{\mathrm{b}(w \circ v)}{\gamma(w \circ v)} = 0, \quad v \in \mathcal{T}_1, w \in \mathcal{T}_2. \tag{2.28}$$

Der Beweis folgt auf Seite 41, nachdem zuvor noch wichtige Definitionen und Lemmata eingeführt werden.

Definition 2.19 *Für eine gegebene Funktion $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ definieren wir*

die **elementaren Hamilton-Funktionen** $H(u) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *durch*

$$H(\tau)(p,q) = H(p,q) \tag{2.29a}$$

$$\begin{aligned} H(u)(p,q) = \frac{\partial^{m+l} H(p,q)}{\partial^m p \partial^l q} \big[&(-1)^{\delta(v_1)} F(v_1), \ldots, (-1)^{\delta(v_m)} F(v_m), \\ &(-1)^{\delta(w_1)} F(w_1), \ldots, (-1)^{\delta(w_l)} F(w_l)\big], \end{aligned} \tag{2.29b}$$

wobei $u = [v_1, \ldots, v_m, w_1, \ldots, w_l]$ *mit* $v_i \in \mathcal{T}_1$ *und* $w_i \in \mathcal{T}_2$*. Die Größen* $\delta(v_i)$ *und* $\delta(w_i)$ *zählen die Anzahl der schwarzen Knoten in* v_i *bzw.* w_i*.*

Bevor wir Beispiele betrachten, möchten wir Bäume, die dieselben elementaren Hamilton-Funktionen liefern, zusammenfassen. Dabei hilft uns die

Definition 2.20 *Wir bezeichnen mit* $\sim$ *die* **Äquivalenzrelation** *mit*

$$u^v \sim u^w \Leftrightarrow \begin{cases} u^v \text{ und } u^w \text{ identisch bis auf die Wurzel} \\ \text{oder} \\ u^v = v \circ w \text{ und } u^w = w \circ v \text{ mit } v \in \mathcal{T}_1, w \in \mathcal{T}_2. \end{cases} \tag{2.30}$$

Lemma 2.13 *Für äquivalente Bäume stimmen die elementaren Hamilton-Funktionen überein.*

Beweis: Nach Definition hängen die elementaren Hamilton-Funktionen nicht vom Typ der Wurzel ab, daher brauchen wir nur

$$H(v \circ w)(p,q) = H(w \circ v)(p,q)$$

zu zeigen. Sei $v = [v_1, \ldots, v_m] \in \mathcal{T}_1$, $w = [w_1, \ldots, w_l] \in \mathcal{T}_2$ und μ, λ die Anzahl der Bäume $v_1, \ldots, v_m$ bzw. $w_1, \ldots, w_l$ in $\mathcal{T}_1$. Unter Vernachlässigung des Arguments (p,q) erhalten wir

$$H(v \circ w) = \pm \sum_{J=1}^{n} \frac{\partial^{m+1} H}{\partial^\mu p \partial^{m-\mu} q \partial p^J} \left(F(v_1), \ldots, F(v_m)\right) \cdot F^J(w), \tag{2.31}$$

wobei der obere Index J die J-te Komponente eines Vektors bezeichnet. Einsetzen des Ausdrucks

$$F^J(w) = \frac{\partial^{l+1} H}{\partial^\lambda p \partial^{l-\lambda} q \partial p^J} \left(F(w_1), \ldots, F(w_m)\right)$$

in (2.31) liefert eine Formel, die symmetrisch ist in Bezug auf v und w. Da das Vorzeichen von $H(v \circ w)$ immer positiv ist, ist damit die Behauptung bewiesen. □

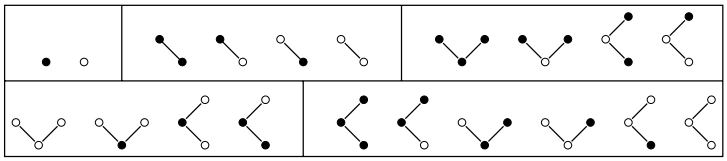


Abbildung 2.3: Äquivalenzklassen für $\rho(u) \leq 3$.

Beispiel In Abb. 2.3 sind die Äquivalenzklassen für $\rho(u) \leq 3$ zusammengestellt. Die dazugehörigen elementaren Hamilton-Funktionen lauten

$$H(\bullet) = H, \qquad H(\bullet\!\!-\!\!\bullet) = H_p H_q,$$
$$H(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) = H_{pp} H_q^2, \qquad H(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) = H_{pq} H_q H_p, \qquad H(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) = H_{qq} H_p^2.$$

Für den Beweis von Satz 2.12 benötigen wir neben den elementaren Hamilton-Funktionen zusätzlich Ausdrücke ihrer partiellen Ableitungen nach p und q. Dies liefert uns das

Lemma 2.14 *Für einen Baum $u \in \mathcal{T}$ gilt*

$$-\frac{\partial H(u)(p,q)}{\partial q} = \sum_{v \in \mathcal{A}(u) \cap \mathcal{T}_1} \alpha(v)\beta(v) F(v)(p,q) \tag{2.32a}$$

$$\frac{\partial H(u)(p,q)}{\partial p} = \sum_{w \in \mathcal{A}(u) \cap \mathcal{T}_2} \alpha(w)\beta(w) F(w)(p,q) \tag{2.32b}$$

mit der Äquivalenzklasse

$$\mathcal{A}(u) = \{v \in \mathcal{T} | v \sim u\}.$$

Die Koeffizienten $\beta : \mathcal{T} \to \mathbb{R}$ sind gegeben durch die Bedingungen (2.28) mit $\beta(u)$ anstelle von $\mathrm{b}(u)$ und durch

$$\rho(u) = \sum_{v \in \mathcal{A}(v) \cap \mathcal{T}_1} (-1)^{\delta(v)+1} \alpha(v)\beta(v) = \sum_{w \in \mathcal{A}(w) \cap \mathcal{T}_2} (-1)^{\delta(w)} \alpha(w)\beta(w).$$

Zum Beweis siehe [16].

Mit diesen Definitionen und Lemmata haben wir alle Hilfsmittel zusammen, um Satz 2.12 beweisen zu können. Zur besseren Übersicht beschränken wir uns wieder auf die uns interessierende Richtung, die Rückrichtung findet man in [16].

Beweis von Satz 2.12: Unter der Voraussetzung (2.28) wollen wir zeigen, dass (2.24) ein Hamilton-System ist. Dazu suchen wir nach einer gestörten Hamilton-Funktion der Form

$$\tilde{H}(p,q) = H_1(p,q) + \frac{h}{2!} H_2(p,q) + \ldots + \frac{h^{N-1}}{N!} H_N(p,q), \tag{2.33}$$

wobei

$$H_k(p,q) = \sum_{u\in\mathcal{V},\rho(u)=k} \mathrm{c}(u)H(u)(p,q) \tag{2.34}$$

mit reellen Koeffizienten $\mathrm{c}(u)$ und der Untermenge $\mathcal{V}$ von $\mathcal{T}$, die genau eine Ausprägung jeder Äquivalenzklasse von $\mathcal{T}/_\sim$ enthält. Damit (2.24) das Hamilton-System in Bezug auf (2.33) ist, muss gelten

$$-\sum_{u\in\mathcal{V},\rho(u)=k} \mathrm{c}(u)\frac{\partial H(u)(p,q)}{\partial q} = \sum_{v\in\mathcal{T}_1,\rho(v)=k} \alpha(v)\mathrm{b}(v)F(v)(p,q), \tag{2.35a}$$

$$\sum_{u\in\mathcal{V},\rho(u)=k} \mathrm{c}(u)\frac{\partial H(u)(p,q)}{\partial p} = \sum_{w\in\mathcal{T}_2,\rho(w)=k} \alpha(w)\mathrm{b}(w)F(w)(p,q). \tag{2.35b}$$

Einsetzen der Formeln (2.32) liefert die Bedingungen

$$\mathrm{c}(u)\beta(v) = \mathrm{b}(v), \qquad \mathrm{c}(u)\beta(w) = \mathrm{b}(w) \tag{2.36}$$

für $v \in \mathcal{A}(v) \cap \mathcal{T}_1$ und $w \in \mathcal{A}(w) \cap \mathcal{T}_2$. Die Koeffizienten $\beta(v)$ und $\beta(w)$ sind ungleich Null, außerdem erfüllen sowohl $\beta(u)$ als auch $\mathrm{b}(u)$ die Bedingungen (2.28), so dass die Koeffizienten $\mathrm{c}(u)$ wohl definiert sind. □

Mit diesem Satz steht uns ein starkes Mittel zur Verfügung, um RK-Verfahren anhand ihrer Koeffizienten auf die Anwendbarkeit von Hamilton-Systemen zu untersuchen. Zum Abschluss dieses Abschnitts bilden wir einen Zusammenhang zwischen symplektischen Verfahren und Hamilton-Systemen. Dazu verwenden wir das

Lemma 2.15 *Wir betrachten Abbildungen* $\mathrm{a} : \mathcal{T} \to \mathbb{R}$ *und* $\mathrm{b} : \mathcal{T} \to \mathbb{R}$*, die über (2.25) in Beziehung stehen. Dann ist (2.27) äquivalent zu (2.28).*

Zur Vereinfachung des Beweises zitieren wir ein Lemma von Calvo und Sanz-Serna [8].

Lemma 2.16 *Seien $F(u)(p,q)$ elementare Differentiale zu (1.21) und (2.21) eine B-Reihe (original P-Reihe). Falls (2.21) eine symplektische Abbildung für allgemeine $H(p,q)$ darstellt, genügen die Koeffizienten* $\mathrm{a}(u)$ *der Beziehung (2.27).*

Beweis von Lemma 2.15: Die Hinrichtung, also (2.27) impliziert (2.28), können wir mit vollständiger Induktion beweisen. Wir gehen davon aus, dass (2.27) gilt, und zeigen, dass

$$\frac{\mathrm{b}(v\circ w)}{\gamma(v\circ w)} + \frac{\mathrm{b}(w\circ v)}{\gamma(w\circ v)} = 0, \quad \text{für} \quad \rho(v)+\rho(w) \le k. \tag{2.37}$$

Für $k = 0$ ist dies trivialerweise erfüllt. Unter der Annahme (2.37) betrachten wir die Differentialgleichung

$$\hat{p}' = \sum_{\rho(v)\le k} \frac{h^{\rho(v)-1}}{\rho(v)!}\alpha(v)\mathrm{b}(v)F(v)(\hat{p},\hat{q}) \tag{2.38a}$$

$$\hat{q}' = \sum_{\rho(w)\le k} \frac{h^{\rho(w)-1}}{\rho(w)!}\alpha(w)\mathrm{b}(w)F(w)(\hat{p},\hat{q}). \tag{2.38b}$$

Nach der Induktionshypothese und Satz 2.12 ist dies ein Hamilton-System. Also ist die exakte Lösung von (2.38)

$$\hat{p}(t_0+h) = \sum_{v\in\mathcal{T}_1} \frac{h^{\rho(v)}}{\rho(v)!}\alpha(v)\mathrm{a}_k(v)F(v)(p_0,q_0)$$

$$\hat{q}(t_0+h) = \sum_{w\in\mathcal{T}_2} \frac{h^{\rho(w)}}{\rho(w)!}\alpha(w)\mathrm{a}_k(w)F(w)(p_0,q_0)$$

eine symplektische Transformation, und die Koeffizienten $\mathrm{a}_k(u)$ genügen daher mit obigem Lemma den Bedingungen (2.27). Darüber hinaus sind sie auch über Formel (2.25) mit den Ausdrücken $\mathrm{b}_k(u)$ verbunden, wobei

$$\mathrm{b}_k(u) = \begin{cases} \mathrm{b}(u) & \text{für} \quad \rho(u)\le k, \\ 0 & \text{für} \quad \rho(u) > k. \end{cases}$$

Daraus folgt für $\mathrm{a}_k(u)$

$$\mathrm{a}_k(u) = \begin{cases} \mathrm{a}(u) & \text{für} \quad \rho(u)\le k, \\ \mathrm{a}(u)-\mathrm{b}(u) & \text{für} \quad \rho(u) = k+1 \end{cases}$$

und weiterhin

$$\mathrm{b}(v\circ w) = \mathrm{a}(v\circ w) - \mathrm{a}_{k+1}(v\circ w)$$

für $\rho(v)+\rho(w) = k+1$. Subtraktion von (2.27) und der analogen Formel für $\mathrm{a}_k(u)$ liefert die Behauptung. □

Mit diesem Satz haben wir einen Zusammenhang zwischen symplektischen Integrationsverfahren und gestörten Hamilton-Systemen aufgestellt und können somit die Hauptaussage dieses Unterkapitels über B-Reihen-Theorie treffen. Der Beweis dieser Aussage besteht aus der Zusammenfassung der bisherigen Überlegungen.

Satz 2.17 *Die Abbildung* $\mathrm{a} : \mathcal{T} \to \mathbb{R}$ *erfülle (2.27) und die rechte Seite des Hamilton-Systems (1.21) sei* N*-mal stetig differenzierbar. Die numerische Lösung* $y_1 = (p_1, q_1)^T$ *von (1.21) definiert durch (2.2) ergibt dann*

$$p_1 = \tilde{p}(t_0+h) + \mathcal{O}(h^{N+1}), \qquad q_1 = \tilde{q}(t_0+h) + \mathcal{O}(h^{N+1}),$$

wobei $\tilde{p}(t)$ und $\tilde{q}(t)$ die exakte Lösung des gestörten Hamilton-Systems

$$\tilde{p}' = -\frac{\partial \tilde{H}}{\partial q}(\tilde{p}, \tilde{q}), \qquad \tilde{q}' = \frac{\partial \tilde{H}}{\partial p}(\tilde{p}, \tilde{q})$$

sind. Die gestörte Hamilton-Funktion ist gegeben durch (2.33) und (2.34), die Koeffizienten $\mathrm{c}(u)$, $\mathrm{b}(u)$ *und* $\beta(u)$ *durch (2.36), (2.25) sowie Lemma 2.14.*

Beweis: Da $\mathrm{a} : \mathcal{T} \to \mathbb{R}$ die Bedingungen (2.27) erfüllt, können wir Lemma 2.15 und anschließend Satz 2.12 anwenden. Damit ist (2.24) ein Hamilton-System. □

Aufgabe dieses Kapitels war es, wichtige Eigenschaften von RK-Verfahren zu definieren und in Zusammenhang zu bringen. Dabei wurde besonders die Anwendbarkeit auf steife mechanische Systeme berücksichtigt.

L-Stabilität garantiert numerische Dämpfung im linearen Fall sowie B-Stabilität die Volumenreduzierung. Desweiteren erhalten wir für steifgenaue Verfahren mit invertierbarer Koeffizientenmatrix und hoher Stufenordnung bestmögliche Konvergenzeigenschaften, während symplektische Verfahren besondere Aussagen über Hamilton-Systeme erlauben. Im nächsten Kapitel werden einzelne IRK-Verfahren daraufhin untersucht.

Kapitel 3

Konvex kombinierte Lobatto-Verfahren

Neben den RK-Verfahren, die im vorherigen Kapitel eingeführt wurden, sind besonders in der Strukturmechanik die Newmark-Verfahren interessant, siehe dazu [21]. Obwohl sie maximal Ordnung $p = 2$ besitzen, wird durch das Vorhandensein zweier Parameter ein hohes Maß an Flexibilität eingeräumt, insbesondere in Bezug auf numerische Dämpfung, welche für die Integration steifer mechanischer Systeme von Vorteil ist.

Im Gegensatz dazu steht das Radau IIA-Verfahren, das das am häufigsten verwendete IRK-Verfahren ist. Es bietet eine hohe Ordnung und Stabilität, ist aber durch diese Anforderungen eindeutig festgelegt.

Im Vordergrund dieses Kapitels steht die Familie der Lobatto-Verfahren[1]. Wie auch das Radau IIA-Verfahren gehören sie zu den IRK-Verfahren, allerdings eröffnet sich durch die Kombination der Verfahren untereinander die Möglichkeit, ähnlich wie beim Newmark-Verfahren die numerische Dämpfung und andere Eigenschaften zu variieren.

Dazu werden zunächst die Lobatto-Verfahren eingeführt und die Flexibilität genauer erläutert. In den nächsten beiden Abschnitten wird dann jeweils auf die Eigenschaften der linearen und nichtlinearen Theorie der neu entstandenen Verfahrensklasse eingegangen, wobei zum Vergleich oft das Radau IIA-Verfahren herangezogen wird. Die Eigenschaften, die sich auf die Anwendung von Hamilton-Systemen beziehen, sind am Ende des Kapitels aufgeführt.

[1] Rehuel Lobatto, 1797-1860, war Prof. für Mathematik an der Technischen Universität Delft und erster jüdischer Professor der Niederlande. Er verfasste zahlreiche Werke hauptsächlich zur Differential- und Integralrechnung und sprach fließend Holländisch, Spanisch, Portugiesisch, Englisch, Französisch und Latein.

Verfahren	ξ	η	ζ	sonst.
Gauß	$2s$	s	s	
Radau IA	$2s-1$	$s-1$	s	
Radau IIA	$2s-1$	s	$s-1$	steifgenau
Lobatto IIIA	$2s-2$	s	$s-2$	steifgenau
Lobatto IIIB	$2s-2$	$s-2$	s	
Lobatto IIIC	$2s-2$	$s-1$	$s-1$	steifgenau
Lobatto IIID	$2s-2$	$s-1$	$s-1$	symplektisch

Tabelle 3.1: Grundlegende Eigenschaften impliziter RK-Verfahren.

3.1 Aufbau und Konstruktion

3.1.1 Implizite RK-Verfahren

Zurückgehend auf Butcher [6] konstruierte man IRK-Verfahren, die auf Quadraturformeln beruhen. Neben den Gauß-Knoten, die sich als Nullstellen des geschifteten Legendre-Polynoms

$$\frac{d^s}{dx^s}\left(x^s(x-1)^s\right)$$

ergeben, liefern die Polynome

$$\frac{d^{s-1}}{dx^{s-1}}\left(x^s(x-1)^{s-1}\right) \quad \text{und} \quad \frac{d^{s-1}}{dx^{s-1}}\left(x^{s-1}(x-1)^s\right)$$

die linken bzw. rechten Radau-Knoten und entsprechend

$$\frac{d^{s-2}}{dx^{s-2}}\left(x^{s-1}(x-1)^{s-1}\right)$$

die Lobatto-Knoten. Wir erhalten $c_1 = 0$ für die linken Radau- und die Lobatto-Knoten sowie $c_s = 1$ für die rechten Radau- und die Lobatto-Knoten.

Um daraus ein implizites RK-Verfahren zu konstruieren, werden die Gewichte b_i, $i = 1, \ldots, s$, so gewählt, dass $\mathcal{B}(2s)$ für die Gauß-Gewichte, $\mathcal{B}(2s-1)$ für die Radau-Gewichte und $\mathcal{B}(2s-2)$ für die Lobatto-Gewichte erfüllt ist. Abhängig von den vereinfachenden Annahmen $\mathcal{B}(\xi)$, $\mathcal{C}(\eta)$ und $\mathcal{D}(\zeta)$ werden die übrigen Koeffizienten so festgelegt, dass die Eigenschaften gemäß der Tab. 3.1 zutreffen. Die klassische Ordnung der Verfahren ist nach Satz 2.1 jeweils identisch mit ξ.

Für die Lobatto-Verfahren ergeben sich anhand der Konstruktionsvorschriften die Butcher-Tableaus der Tab. 3.2 für $s = 2$ bzw. Tab. 3.3 für $s = 3$ Stufen.

Während man bei den steifgenauen Verfahren Lobatto IIIA und IIIC die Gleichheit der letzten Matrixzeile mit dem Gewichtsvektor beobachtet, fällt bei Lobatto

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
A	$\frac{1}{2}$	$\frac{1}{2}$

0	$\frac{1}{2}$	0
1	$\frac{1}{2}$	0
B	$\frac{1}{2}$	$\frac{1}{2}$

0	$\frac{1}{2}$	$-\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$
C	$\frac{1}{2}$	$\frac{1}{2}$

0	$\frac{1}{4}$	$-\frac{1}{4}$
1	$\frac{3}{4}$	$\frac{1}{4}$
D	$\frac{1}{2}$	$\frac{1}{2}$

Tabelle 3.2: Lobatto IIIA-B-C-D Verfahren mit $s = 2$ Stufen.

0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
A	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

0	$\frac{1}{6}$	$-\frac{1}{6}$	0
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$	0
1	$\frac{1}{6}$	$\frac{5}{6}$	0
B	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

0	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{6}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
C	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

0	$\frac{1}{12}$	$-\frac{1}{6}$	$\frac{1}{12}$
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{12}$	$\frac{5}{6}$	$\frac{1}{12}$
D	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Tabelle 3.3: Lobatto IIIA-B-C-D Verfahren mit $s = 3$ Stufen.

IIIA und IIIB auf, dass eine komplette Zeile bzw. Spalte zu Null wird. Bei Lobatto IIIB und IIIC ist zu sehen, dass $a_{1i} = b_1$ für $i = 1, \ldots, s$ gilt.

Analyse der IRK-Verfahren

Gauß Das Gauß-Verfahren ist das IRK-Verfahren mit der höchsten Ordnung und ist damit für nichtsteife Systeme sehr gut geeignet. Es stellt sich allerdings heraus, dass für die Anwendung auf steife mechanische Systeme andere Eigenschaften wie L- und B-Stabilität sinnvoller sind.

Radau IA Ähnlich wie das Gauß-Verfahren mangelt es dem Radau IA-Verfahren an Stabilität, und es wird daher keine große Rolle in unserem Kontext spielen.

Radau IIA Das Radau IIA-Verfahren ist der bekannteste und weit verbreiteter Vertreter dieser Klasse. Durch seine relativ hohe klassische Ordnung sowie Stufenordnung kombiniert mit L- und B-Stabilität legt es einen hohen Standard fest.

Lobatto IIIA Dieses Verfahren kann fehlende L- und B-Stabilität durch Steifgenauigkeit und hohe Stufenordnung ausgleichen. Es sticht zudem als Kollokationsverfahren dieser Familie ins Auge.

Lobatto IIIB Dies ist der unbedeutendste Vertreter der Lobatto-Familie, da weder B- oder L-Stabilität noch Steifgenauigkeit noch hohe Stufenordnung vorliegen.

Lobatto IIIC Es besitzt wie das Radau IIA-Verfahren B- und L-Stabilität und ersetzt eine Ordnungsstufe durch stärkere numerische Dämpfung. Damit liegt hier der für steife mechanische Systeme am besten geeignete Vertreter der Lobatto-Familie vor.

Lobatto IIID Das Verfahren Lobatto IIID führt zwar keine numerische Dämpfung ein, dafür kann man durch die Symplektizität Einiges erreichen.

3.1.2 Konvex kombinierte Lobatto-Verfahren

Mithilfe der konvex kombinierten Lobatto-Verfahren ist es möglich, zusätzliche Flexibilität in die numerische Integration einzubringen. Dies gilt nicht nur für die Konstruktion neuer Verfahren, sondern auch vor allem für die gezielte Anwendung von Verfahren mit speziellen Eigenschaften auf einzelne Komponenten. Ermöglicht wird dies durch die verschiedenen Verfahren der Lobatto-Familie, die mit völlig unterschiedlichen Eigenschaften ausgestattet sind, aber trotzdem miteinander verknüpft werden können.

Die sogenannten SPARK-Methoden (Super Partitioned Additive Runge-Kutta), siehe Jay [24], erlauben die Kombination von Verfahren der gleichen Familie. Dabei wird die rechte Seite der Differentialgleichung $f(y)$ aufgespalten zu

$$y' = f(y) = \sum_{m=1}^{4} f_m(y)$$

und jeweils unterschiedliche Verfahren einer Familie auf die einzelnen Summanden angewendet. Im Speziellen sind das die Lobatto IIIA-B-C-D-SPARK Methoden.

Die Idee hinter den *konvex kombinierten* Lobatto-Verfahren, kurz *BL*-Verfahren wegen englisch *Blended Lobatto*, beinhaltet die Vermischung von zwei dieser Verfahren über einen Parameter θ, so dass für jeden Wert $\theta \in (0, 1)$ ein neues IRK-Verfahren entsteht. Für $\theta = 0$ und $\theta = 1$ erhalten wir die Originalverfahren.

Zur Konstruktion verwenden wir die namensgebende konvexe Aufspaltung

$$f(y) = \underbrace{\theta f(y)}_{f_X} + \underbrace{(1-\theta) f(y)}_{f_Y},$$

und dies führt konkret auf die Iterationsvorschrift

$$Y_i = y_0 + h \sum_{j=1}^{s} \underbrace{\left[a_{ij}^X \theta + a_{ij}^Y (1-\theta)\right]}_{a_{ij}^\theta} f(Y_j), \tag{3.1a}$$

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i f(Y_i), \tag{3.1b}$$

wobei a_{ij}^X bzw. a_{ij}^Y jeweils die Koeffizienten der beiden Lobatto-Verfahren bezeichnen, die miteinander kombiniert werden, im Folgenden *Randverfahren* genannt.

Welche klassische Ordnung p und Stufenordnung η besitzen die Verfahren mit $0 < \theta < 1$? Diese Frage beantworten uns die folgenden beiden Sätze:

Satz 3.1 *Gegeben seien zwei RK-Verfahren mit den zugehörigen Koeffizientenmatrizen $A^X = (a_{ij}^X)$ und $A^Y = (a_{ij}^Y)$ sowie identischen Knoten- und Gewichtsvektoren $c = c^X = c^Y$, $b = b^X = b^Y$. Ferner gelte für das Verfahren (A^X, b, c) die vereinfachenden Annahmen $\mathcal{B}(\xi^X)$, $\mathcal{C}(\eta^X)$ und $\mathcal{D}(\zeta^X)$ sowie für (A^Y, b, c) entsprechend $\mathcal{B}(\xi^Y)$, $\mathcal{C}(\eta^Y)$ und $\mathcal{D}(\zeta^Y)$. Dann erfüllen die Verfahren aus (3.1) mit $\theta \in (0, 1)$ die Bedingungen $\mathcal{B}(\xi^\theta)$, $\mathcal{C}(\eta^\theta)$ und $\mathcal{D}(\zeta^\theta)$ mit*

i) $\xi^\theta = \xi^X = \xi^Y$,

ii) $\eta^\theta = \min\{\eta^X, \eta^Y\}$ *und*

iii) $\zeta^\theta = \min\{\zeta^X, \zeta^Y\}$.

Beweis: Die Aussage i) ist wegen $c = c^X = c^Y$ und $b = b^X = b^Y$ trivial. Zum Beweis von ii) setzen wir $a_{ij}^\theta = a_{ij}^X\theta + a_{ij}^Y(1-\theta)$ in die Definition von $\mathcal{C}(\eta)$ (2.5b) ein und erhalten für $q = 1, \ldots, \eta^\theta$

$$\begin{aligned}\sum_{i=1}^{s} a_{ij}^\theta c_i^{q-1} &= \sum_{i=1}^{s} \left[a_{ij}^X\theta + a_{ij}^Y(1-\theta)\right] c_i^{q-1} \\ &= \theta\left[\sum_{i=1}^{s} a_{ij}^X c_i^{q-1}\right] + (1-\theta)\left[\sum_{i=1}^{s} a_{ij}^Y c_i^{q-1}\right] \\ &\overset{\mathcal{C}(\min\{\zeta^X,\zeta^Y\})}{=} \theta\frac{c_i^q}{q} + (1-\theta)\frac{c_i^q}{q} = \frac{c_i^q}{q}\end{aligned}$$

für $q = 1, \ldots, \min\{\eta^X, \eta^Y\}$, $i = 1, \ldots, s$. Analog folgt aus der Linearität von $\mathcal{D}(\zeta)$ in A die Behauptung iii).

□

Satz 3.2 *Zusätzlich zu den Voraussetzungen des Satzes 3.1 gelte*

$$\xi^\theta \le \eta^X + \zeta^Y + 2, \tag{3.2a}$$
$$\xi^\theta \le \eta^Y + \zeta^X + 2, \tag{3.2b}$$
$$\xi^X \le 2\eta^X + 2, \tag{3.2c}$$
$$\xi^Y \le 2\eta^Y + 2, \tag{3.2d}$$
$$\xi^X \le \eta^X + \zeta^X + 1, \tag{3.2e}$$
$$\xi^Y \le \eta^Y + \zeta^Y + 1. \tag{3.2f}$$

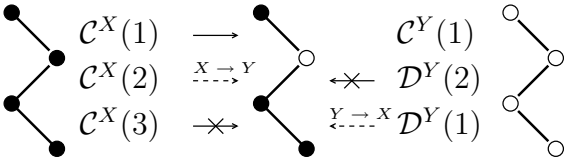
Dann ist ξ^θ die klassische Ordnung des Verfahrens mit Koeffizienten (A^θ, b, c).

Beweis: Aus (3.2c) bis (3.2f) folgt nach Satz 2.1, dass beide Randverfahren die klassische Ordnung ξ besitzen.

Für weitere Aussagen setzen wir zunächst wie im vorherigen Satz die Definition von a_{ij}^θ in die Ordnungsbedingungen ein und nutzen die Linearität aus, um die einzelnen Terme auseinander zu ziehen. Es entstehen Ordnungsbedingungen für partitionierte Systeme, die anstelle von a_{ij} jede Kombination aus a_{ij}^X und a_{ij}^Y enthalten. Die Behauptung folgt, wenn wir jeden einzelnen gemischten Term mit Hilfe der vereinfachenden Annahmen reduzieren können.

Um Letzteres zu sehen, interpretieren wir die Bedingungen $\mathcal{C}$ und $\mathcal{D}$ als Hilfsmittel, die die Ordnungsbedingungen von rechts bzw. links abarbeiten, siehe Beweis zu Satz 2.1 auf Seite 18. In der B-Reihen-Theorie kann man dies auch als Abarbeiten der Bäume von oben mit $\mathcal{C}$ und von unten mit $\mathcal{D}$ betrachten, wobei ähnlich wie in Abschnitt 2.5 von zwei unterschiedlichen Knoten für a_{ij}^X (schwarz) und a_{ij}^Y (weiß) ausgegangen werden muss. O. B. d. A. sei $\eta^X > \eta^Y$ und $\zeta^X < \zeta^Y$.

Beispiel:



Solange sowohl $\mathcal{C}(\eta^X)$ als auch $\mathcal{C}(\eta^Y)$ bzw. sowohl $\mathcal{D}(\zeta^X)$ als auch $\mathcal{D}(\zeta^Y)$ erfüllt sind, können die Bedingungen wie im Beweis zu Satz 2.1 reduziert werden, siehe normale Pfeile. Ein Problem taucht in dem Bereich auf, in dem beide Randverfahren unterschiedliche vereinfachende Annahmen zur Reduktion verwenden. Wenn in einem Baum in diesem Bereich ein weißer Knoten über einem schwarzen Knoten liegt, kann weder von oben $\mathcal{C}$ noch von unten $\mathcal{D}$ angewendet werden, siehe durchgestrichene Pfeile. In diesem Fall benötigen wir die Bedingungen (3.2a) und (3.2b), um die Lücke zu schließen, was durch die gestrichelten Pfeile dargestellt wird.

Im Vergleich zu Satz 2.1 kann auf $\xi^\theta \leq \eta^X + \zeta^Y + 1$ und $\xi^\theta \leq \eta^Y + \zeta^X + 1$ verzichtet werden, da ein einzelner Knoten jeweils von den Bedingungen der Randverfahren bearbeitet werden kann. □

Die Familie der Lobatto-Verfahren besitzt daher die klassische Ordnung $2s - 2$, da wegen $\eta^X + \zeta^Y \geq 2s - 4$ für alle Kombinationen von Lobatto-Verfahren obiger Satz mit $\xi^\theta = 2s - 2$ angewendet werden kann.

Paarungen Bisher haben wir davon gesprochen, wie man zwei Lobatto-Verfahren miteinander kombinieren kann und welche Eigenschaften sich daraus ergeben. Doch welche Paarungen sind sinnvoll? Wir untersuchen folgende Kombinationen:

a) IIIA mit IIIB, genannt IIIAB

b) IIIA mit IIIC, genannt IIIAC

c) Das Verfahren aus a) mit $\theta = 0.5$ mit IIIC, genannt IIIABC

d) IIID mit IIIC, genannt IIIDC, und

e) Das Verfahren aus a) mit $\theta = 0.5$ mit IIID, genannt IIIABD.

In den folgenden Abschnitten, in denen diese Klassen genauer untersucht werden sollen, konzentrieren wir uns auf die IIIAC- und IIIDC oder IIIABC-Verfahren, da diese das höchste Anwendungspotential für steife mechanische Systeme besitzen. Für Bilder, explizite Formeln etc. werden $s = 3$ Stufen verwendet, die abgeleiteten Aussagen gelten meistens für alle s, Ausnahmen werden angegeben.

3.2 Eigenschaften von BL-Verfahren

3.2.1 Lineare Stabilitätskonzepte

Als Grundlage für eine lineare Analyse dient die Stabilitätsfunktion. Sie lautet für die IIIAC-Verfahren

$$\mathcal{R}(z) = 2\frac{12 + 3z + 3z\theta + z^2\theta}{24 - 18z + 6z\theta - 4z^2\theta + 6z^2 + z^3\theta - z^3}$$

und für IIIDC-Verfahren

$$\mathcal{R}(z) = \frac{48 + 12z\theta + 12z + 6z^2\theta + z^3\theta}{48 + 12z\theta - 36z - 6z^2\theta + 12z^2 + z^3\theta - 2z^3}.$$

Daraus ergeben sich die Stabilitätsgebiete aus Abb. 3.1, wobei jeweils die Bereiche außerhalb der gezeichneten Umrandung stabil sind. Aus dieser Darstellung lässt sich direkt die A-Stabilität für alle Werte von θ sowohl für IIIAC- als auch für IIIDC-Verfahren ablesen.

Um die L-Stabilität zu analysieren, betrachten wir jeweils die Polynomgrade der Stabilitätsfunktion im Zähler und im Nenner. Für die IIIAC-Methoden ist der Polynomgrad des Zählers jeweils kleiner als der Polynomgrad des Nenners mit Ausnahme von $\theta = 1$, während für IIIDC-Methoden beide Grade bis auf $\theta = 0$ übereinstimmen. Daher sind die IIIAC-Verfahren L-stabil im Gegensatz zu den anderen Kombinationen.

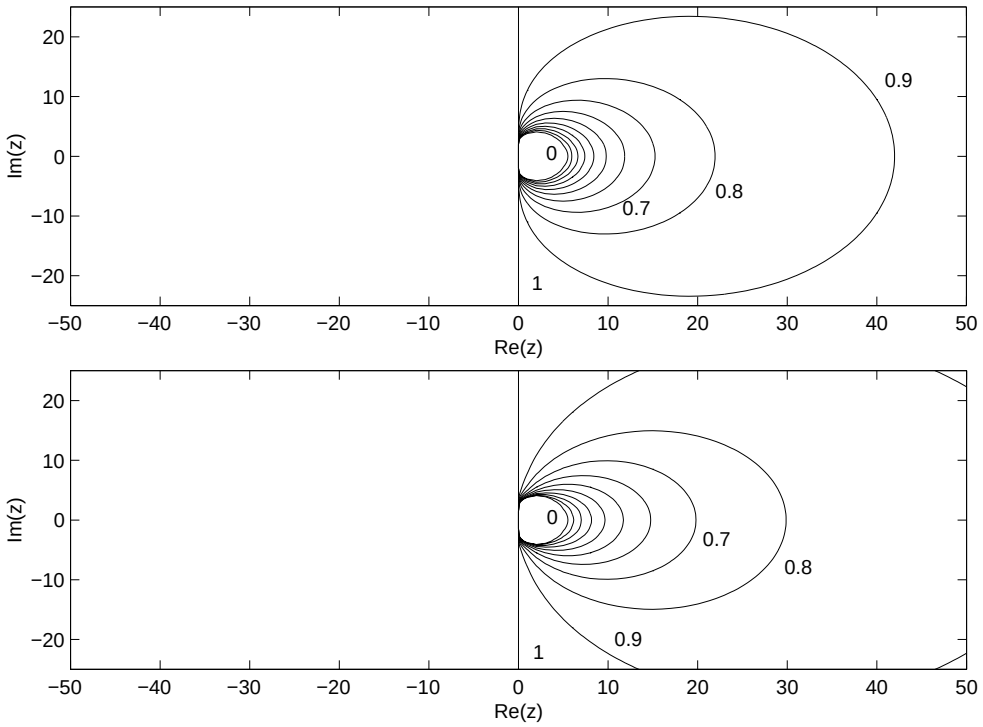


Abbildung 3.1: Ränder der Stabilitätsgebiete für IIIAC- (oben) und IIIDC-Verfahren (unten) mit $s = 3$ Stufen für verschiedene Werte von θ.

Neben der L-Stabilität interessiert uns im Besonderen der Betrag der numerischen Dämpfung, also die Antwort auf die Frage: Welche Frequenzen werden abhängig von der Schrittweite wie stark gedämpft? Dazu betrachten wir die Spektralradien der IIIAC- und IIIDC-Verfahren aus Abb. 3.2.

Wie in Kap. 2 erläutert, wollen wir möglichst wenig Dämpfung für $\chi < \pi$ und möglichst viel Dämpfung für $\chi > \pi$ erhalten. Der Spektralradius für Lobatto IIIC, also $\theta = 0$ in beiden Klassen, liefert dazu schon eine gute Vorlage, auch für $s > 3$ Stufen. Durch Erhöhung des Parameters θ kann man die zwar nicht sichtbare aber dennoch vorhandene Dämpfung im vorderen Bereich abmildern.

Man erkennt, dass die IIIDC-Verfahren trotz fehlender L-Stabilität im hinteren Bereich sehr stark dämpfen, so dass sie für die Anwendung auf steife mechanische Systeme auch sehr gut geeignet sind. Dadurch, dass sich die Linien zum Teil kreuzen, kann man durch geeignete Wahl von θ die numerische Dämpfung maximieren.

Für Lobatto IIIABC liegt ebenfalls starke numerische Dämpfung trotz fehlender L-Stabilität vor. Das Überschneiden der Kurven für verschiedene Werte von θ ist aber nicht zu beobachten. Die Anwendungsmöglichkeiten der einzelnen Klassen

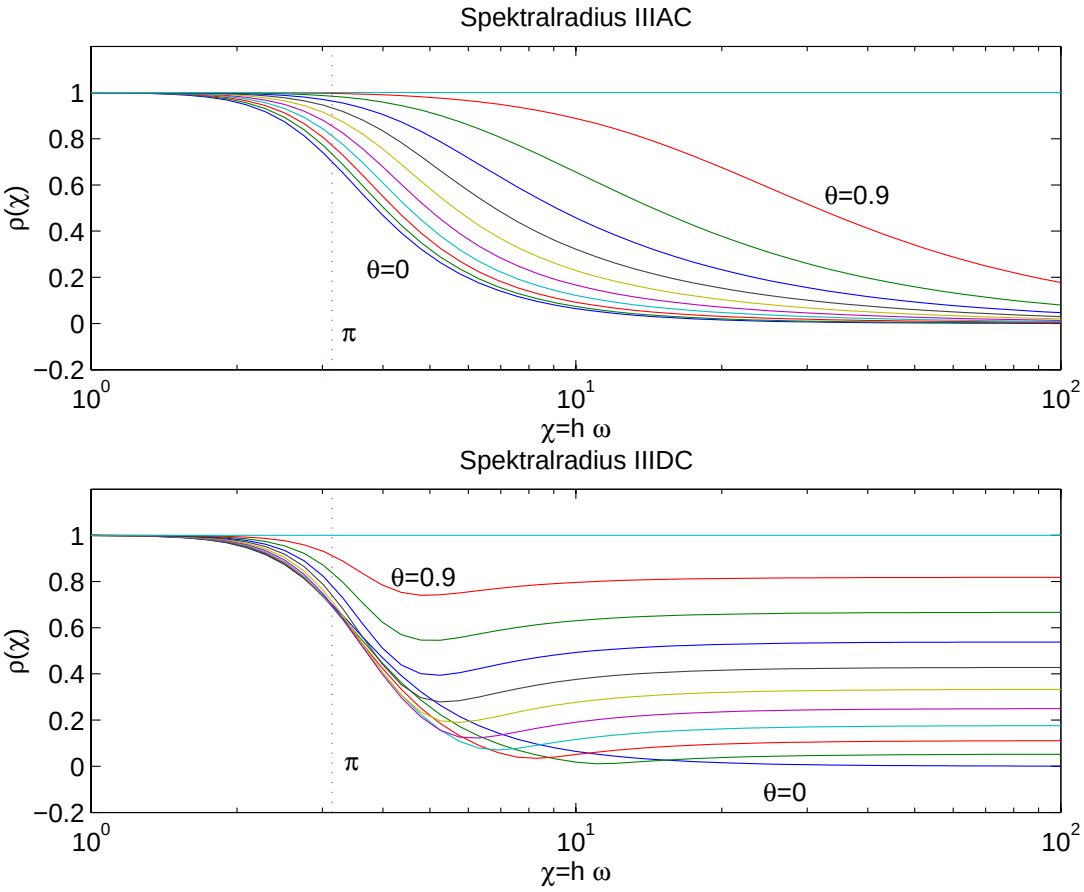


Abbildung 3.2: Spektralradien für IIIAC- und IIIDC-Verfahren für verschiedene Werte von θ.

werden wir in Kap. 4 genauer untersuchen.

Dispersions- und Dissipationsordnung Wenden wir uns nun den Eigenschaften zu, die auf der Testgleichung (2.9) für Systeme 2. Ordnung beruhen. Die Reihenentwicklungen, die für die Dispersions- und Dissipationsordnung eine Rolle spielen, lauten für die IIIAC-Verfahren

$$\begin{aligned} d_a(\chi) = &\left(\frac{1}{1152}\theta^2 - \frac{1}{576}\theta + \frac{1}{1152}\right)\chi^5 \\ &+ \left(-\frac{1}{18432}\theta^4 + \frac{1}{6912}\theta^3 - \frac{5}{27648}\theta^2 + \frac{1}{6912}\theta - \frac{1}{18432}\right)\chi^7 \\ &+ \mathcal{O}(\chi^9), \end{aligned} \tag{3.3}$$

Paarung	θ-Bereich	Disp.-Ord.	Diss.-Ord.
IIIAC	$\frac{s}{2s-1} \neq \theta < 1$	$2s-2$	$2s-1$
IIIAC	$\theta = \frac{s}{2s-1}$	$2s$	$2s-1$
IIIAC	$\theta = 1$	$2s-2$	∞
IIIDC	$\theta < 1$	$2s-2$	$2s-1$
IIIDC	$\theta = 1$	$2s-2$	∞
IIIABD	$\theta \neq 2 - \frac{2}{2s-1}\sqrt{(2s-1)(s-1)}$	$2s-2$	∞
IIIABD	$\theta = 2 - \frac{2}{2s-1}\sqrt{(2s-1)(s-1)}$	$2s$	∞

Tabelle 3.4: Dispersions- und Dissipationsordnung für IIIAC, IIIDC und IIIABD.

und

$$\begin{aligned} d_p(\chi) - 1 = &\left(-\frac{1}{288}\theta + \frac{1}{480}\right)\chi^4 \\ &+ \left(\frac{1}{4608}\theta^3 - \frac{7}{13824}\theta^2 + \frac{1}{1536}\theta - \frac{1}{3584}\right)\chi^6 + \mathcal{O}(\chi^8) \end{aligned} \tag{3.4}$$

bzw. für die IIIDC-Verfahren

$$\begin{aligned} d_a(\chi) = &\left(-\frac{1}{1152}\theta + \frac{1}{1152}\right)\chi^5 \\ &+ \left(\frac{1}{18432}\theta^3 - \frac{1}{6144}\theta^2 + \frac{1}{6144}\theta - \frac{1}{18432}\right)\chi^7 + O(\chi^9), \end{aligned} \tag{3.5a}$$

$$d_p(\chi) = \frac{1}{480}\chi^4 + \left(-\frac{1}{4608}\theta^2 + \frac{1}{2304}\theta - \frac{1}{3584}\right)\chi^6 + \mathcal{O}(\chi^8). \tag{3.5b}$$

In diesem Fall ist auch die Kombination IIIABD interessant. Die Reihe für d_p lautet dann

$$\begin{aligned} d_p(\chi) = &\left(\frac{1}{1152}\theta^2 - \frac{1}{288}\theta + \frac{1}{480}\right)\chi^4 \\ &+ \left(\frac{1}{110592}\theta^4 - \frac{1}{13824}\theta^3 + \frac{1}{6912}\theta^2 - \frac{1}{16128}\right)\chi^6 + \mathcal{O}\left(\chi^8\right), \end{aligned}$$

während die für $d_a(\chi)$ identisch verschwindet.

Daraus ergeben sich die Dispersions- und Dissipationsordnungen der Tab. 3.4. Wie zu erkennen ist, existieren drei Werte von θ, die ins Auge fallen. Zum Einen verschwindet der Amplitudenfehler für Lobatto IIIA und IIID, also $\theta = 1$, komplett, siehe Dissipationsordnung ∞, zum Anderen ist für IIIAC-Verfahren mit $\theta = \frac{s}{2s-1}$ die Dispersionsordnung um 2 höher als für andere Werte von θ, was auch die lineare Ordnung um 1 erhöht. Der Grund dafür ist, dass die Stabilitätsfunktion dann mit der des gleichstufigen Radau IIA-Verfahrens übereinstimmt, welches die klassische und auch die lineare Ordnung $2s-1$ besitzt.

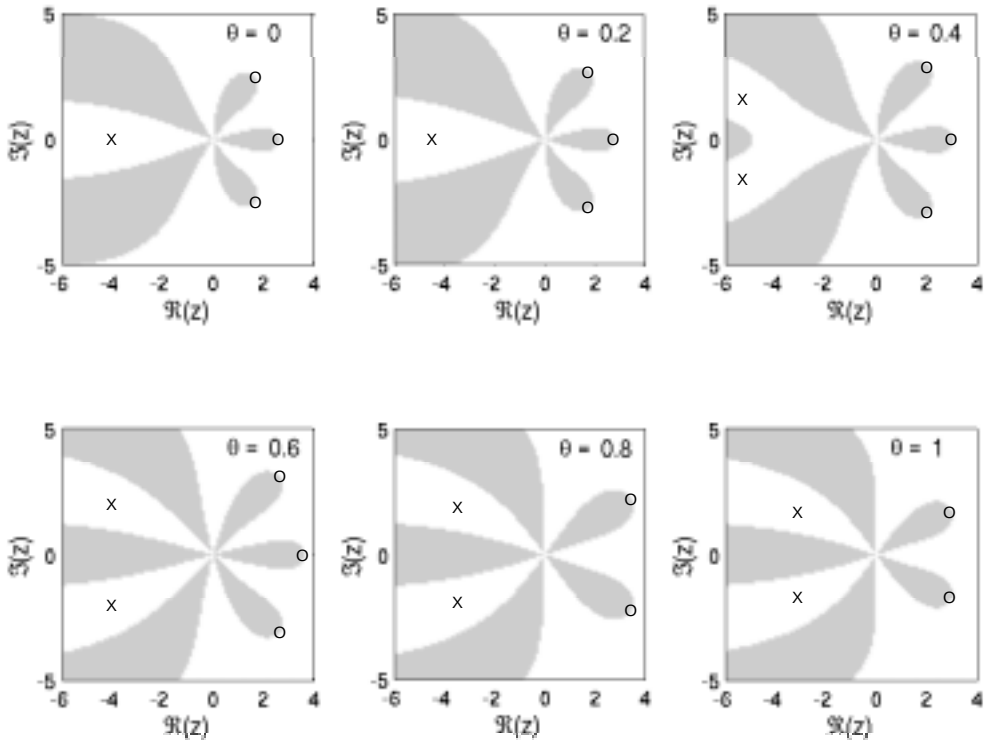


Abbildung 3.3: Ordnungssterne für IIIAC-Verfahren mit $s = 3$ Stufen: $\circ$ bezeichnet die Pole, $\times$ die Nullstellen von $\mathcal{R}(z)$.

Damit stimmen für diesen Wert alle linearen Eigenschaften mit dem Radau IIA-Verfahren überein, z.B. auch die Form des Spektralradius aus Abb. 3.2. Man erkennt daraus, dass die numerische Dämpfung des Radau IIA-Verfahrens deutlich schwächer ist als die des Lobatto IIIC-Verfahrens.

Ähnlich wie für die IIIAC-Verfahren mit $\theta = \frac{s}{2s-1}$ fällt auch für die IIIABD-Verfahren mit $\theta = 2 - \frac{2}{2s-1}\sqrt{(2s-1)(s-1)}$ die Stabilitätsfunktion mit der des Gauß-Verfahrens zusammen, was die lineare Ordnung $2s$ zur Folge hat. Beide Werte von θ wurden für $s = 2, 3, 4$ berechnet und für allgemeine Stufenzahlen s extrapoliert.

Diesen Wechsel zu einer höheren linearen Ordnung kann man auch anhand der Ordnungssterne beobachten, die in Abb. 3.3 für die IIIAC-Verfahren dargestellt sind. Die Anzahl der Arme bei Rotation um den Nullpunkt um eins erniedrigt gibt die lineare Ordnung wieder. Man erkennt auch die unterschiedliche Anzahl der Pole und Nullstellen, deren Gesamtanzahl für $0 < \theta < 1$ um eins höher ist als für die Randverfahren. Zur weiterführenden Theorie der Ordnungssterne siehe [19].

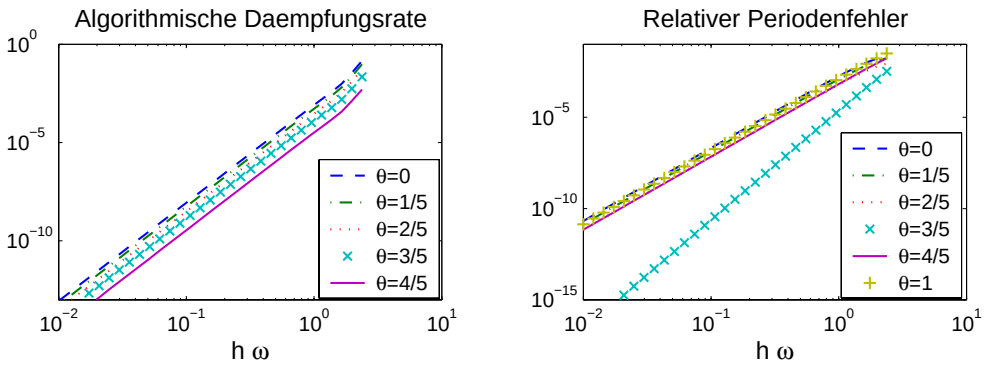


Abbildung 3.4: Algorithmische Dämpfungsrate (links) und relativer Periodenfehler (rechts) für IIIAC mit $s = 3$ Stufen.

Nach diesem Ausflug in die Theorie der Ordnungssterne kommen wir nun wieder zurück zu den Größen d_a und d_p. Vor allem in der Strukturmechanik werden noch zwei weitere Eigenschaften von numerischen Verfahren gegenübergestellt, nämlich der algorithmische Dämpfungsfaktor $\bar{\xi}$ und der relative Periodenfehler $\bar{\gamma}$.

Ersterer ist definiert als

$$\bar{\xi}(\chi) = -\frac{\ln|\mathcal{R}(i\chi)|}{\arctan\left(\frac{\Im[\mathcal{R}(i\chi)]}{\Re[\mathcal{R}(i\chi)]}\right)}$$

und nach Einsetzen der Definitionen von d_a und d_p (2.11) ist dies äquivalent zu

$$\overline{\xi}(x) = -\frac{\ln\left(e^{-d_a(\chi)\chi}\right)}{\arctan\left(\frac{\sin(d_p(\chi)\chi)}{\cos(d_p(\chi)\chi)}\right)} = \frac{d_a(\chi)}{d_p(\chi)}.$$

Die Funktion $\bar{\xi}$ ist in Abb. 3.4 links für verschiedene Werte von θ der Lobatto IIIAC-Familie dargestellt. Wegen $d_p(\chi) \approx 1$ ist der Nenner vernachlässigbar und die Dissipationsordnung dominiert. Zudem nimmt die algorithmische Dämpfungsrate für steigende Werte von θ wegen $d_a(\chi) \equiv 0$ für Lobatto IIIA zu.

Im Gegensatz zu obigem Kriterium liefert der relative Periodenfehler

$$\bar{\gamma}(\chi) = \frac{\bar{T} - T}{T}$$

eine Interpretation der Dispersionsordnung. Mit den Periodenlängen $T = 2\pi/\omega$, $\bar{T} = 2\pi/\bar{\omega}$ und $\bar{\omega} = d_p(\chi)\,\omega$ erhalten wir

$$\bar{\gamma}(\chi) = \frac{1}{d_p(\chi)} - 1.$$

Die Funktion $\bar{\gamma}(\chi)$ ist in Abb. 3.4 rechts dargestellt. Wie auch aufgrund der höheren Dispersionsordnung zu erwarten, ist die Steigung für $\theta = 3/5$ höher als die für andere Werte von θ.

Damit möchten wir die lineare Analyse der BL-Verfahren abschließen. Auf den Vergleich mit den Verfahren höherer Ordnung wie Radau IIA und Gauß konnte hier verzichtet werden, da sie, was die lineare Theorie betrifft, jeweils als Spezialfälle einzelner Paarungen auftreten. Als besonders geeignet hat sich in diesem Abschnitt wegen ihrer L-Stabilität die Klasse der IIIAC-Verfahren herausgestellt.

3.2.2 Eigenschaften der nichtlinearen Theorie

Die nichtlineare Theorie stellt zusätzliche Anforderungen an ein RK-Verfahren. So ist die B-Stabilität zwar die Erweiterung zur A-Stabilität, aber eine nichtlineare Ergänzung zur L-Stabilität existiert nicht. Auch die Koerzivität, die zur Existenz- und Eindeutigkeit von RK-Lösungen führt, ist völlig unabhängig von den Stabilitätskonzepten. Daher werden hier andere Kombinationen von Lobatto-Verfahren eine Rolle spielen.

Symplektizität und B-Stabilität Die Symplektizität und B-Stabilität können gut zusammen untersucht werden, da sie auf dieselbe Matrix aufbauen, wenn man die zur B-Stabilität in unserem Fall äquivalente algebraische Stabilität zugrundelegt. Zunächst können wir aber einen Satz aus [20, Th. 12.8] zitieren, um die B-Stabilität der IIIDC-Verfahren zu beweisen. Er lautet

Satz 3.3 *Gegeben sei ein RK-Verfahren mit disjunkten Knoten c_i und positiven Gewichten b_i, welches den vereinfachenden Annahmen $\mathcal{B}(2s-2)$, $\mathcal{C}(s-1)$ und $\mathcal{D}(s-1)$ genügt. Dann ist diese Methode genau dann algebraisch stabil, wenn für die Stabilitätsfunktion $\mathcal{R}(z)$ die Bedingung $|\mathcal{R}(\infty)| \leq 1$ erfüllt ist.*

Der Beweis dieses Satzes erfolgt mit Hilfe der W-Transformation und verwendet $\eta = \zeta$. Aus der A-Stabilität der Lobatto IIIDC-Familie folgt mit diesem Satz deren B-Stabilität.

Für die IIIABC-Verfahren gilt für $\theta > 0$ nur $\mathcal{C}(s-2)$, so dass dieser Satz nicht angewendet werden kann. Stattdessen hilft uns der folgende Satz weiter:

Satz 3.4 *Seien A^X und A^Y die Koeffizientenmatrizen von zwei RK-Verfahren mit identischem Knotenvektor c und Gewichtsvektor b. Dann gilt*

i) (A^X, b, c) und (A^Y, b, c) symplektisch $\Rightarrow$ (A^θ, b, c) symplektisch,

ii) (A^X, b, c) *B-stabil und* (A^Y, b, c) *symplektisch* $\Rightarrow$ (A^θ, b, c) *B-stabil,*

iii) (A^X, b, c) *und* (A^Y, b, c) *B-stabil* $\Rightarrow$ (A^θ, b, c) *B-stabil.*

Beweis: Allgemein folgt aus der Linearität von

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j$$

in a_{ij} und $a_{ij}^\theta = \theta a_{ij}^X + (1-\theta) a_{ij}^Y$ die Beziehung

$$M^\theta = \theta M^X + (1-\theta) M^Y$$

mit der Matrix $M^\theta = (b_i a_{ij}^\theta + b_j a_{ji}^\theta - b_i b_j)_{ij}$ und analog M^X und M^Y. Die Behauptung i) folgt direkt, da für symplektische Verfahren M^X und M^Y der Nullmatrix entsprechen. Da die Summe aus zwei positiv semidefiniten Matrizen wieder positiv semidefinit ist, folgen ii) und iii). □

Mit diesem Satz haben wir neben der B-Stabilität von IIIABC und IIIDC auch die Symplektizität von IIIABD gezeigt. In Bezug auf nichtlineare Eigenschaften erweisen sich also diese drei Klassen als besonders brauchbar, während es im linearen Fall eher die Klasse IIIAC ist. Im Kap. 4 werden wir diese Unterschiede bei den Anwendungen ausnutzen.

Koerzivität Gehen wir nun zur Koerzivität über, die uns Aussagen über die Existenz und Eindeutigkeit der IRK-Lösungen (2.2a) liefert.

Lemma 3.5 *Wir erhalten für die Lobatto IIIA-B-C-D-Verfahren sowie das Radau IIA-Verfahren die Koerzivitätskoeffizienten*

$$\textit{Lobatto IIIA:}\quad \alpha_0(\tilde{A}^{-1}) = \begin{cases} 2 & \textit{für} \quad s = 2, \\ c_{s-1}^{-1} & \textit{für} \quad s > 2, \end{cases}$$

$$\textit{Lobatto IIIB:}\quad \alpha_0(\tilde{A}^{-1}) = \begin{cases} 2 & \textit{für} \quad s = 2, \\ (1-c_2)^{-1} & \textit{für} \quad s > 2, \end{cases}$$

$$\textit{Lobatto IIIC:}\quad \alpha_0(A^{-1}) = \begin{cases} 1 & \textit{für} \quad s = 2, \\ 0 & \textit{für} \quad s > 2, \end{cases}$$

$$\textit{Lobatto IIID:}\quad \alpha_0(A^{-1}) = 0.$$

$$\textit{Radau IIA:}\quad \alpha_0(A^{-1}) = \begin{cases} 1 & \textit{für} \quad s = 1, \\ \frac{1}{2c_{s-1}} & \textit{für} \quad s > 1, \end{cases}$$

Beweisskizze: Für Lobatto IIIA-B-C und Radau IIA werden diese Koeffizienten bereits in [20, Th.IV.14.8] bewiesen. Dazu wird jeweils eine Diagonalmatrix D erzeugt, so dass

$$M^D = DA^{-1} + (DA^{-1})^T$$

Diagonalgestalt besitzt. Daraus folgt, dass diese Diagonalmatrix D diejenige ist, für die das Supremum in (2.14) angenommen wird, also $\alpha_0(A^{-1}) = \alpha_D(A^{-1})$ mit $\alpha_D(A^{-1})$ gleich dem kleinsten Diagonaleintrag von M^D.

Für das Lobatto IIID-Verfahren ist diese Vorgehensweise nicht möglich, da M für keine Diagonalmatrix D diagonal ist. Allerdings führt die Wahl von $D = B := diag(b)$ auf

$$M^B = (e_1 + e_s)\,(e_1 + e_s)^T$$

mit dem einfachen Eigenwert 2 und dem $s-1$-fachen Eigenwert 0 von M^B. Damit ist $\alpha_0(A^{-1}) \geq 0$. Da die Diagonalelemente der Inversen von A größer oder gleich null sind, folgt mit (2.15) die Behauptung. □

Welche Auswirkungen haben diese Unterschiede auf das numerische Verhalten der einzelnen Verfahren? Für $s = 2$ Stufen besitzen alle Verfahren bis auf Lobatto IIID eine Koerzivität $\alpha_0(A^{-1}) > 0$. Dies bedeutet nach Lemma 2.4, dass für jedes ν eine Schrittweite h existiert, so dass das IRK-Verfahren eine eindeutige Lösung besitzt.

Für $s > 2$ Stufen des Lobatto IIIC Verfahrens bzw. Lobatto IIID generell gilt nur $\alpha_0(A^{-1}) = 0$. In diesem Fall erhält man nur für $\nu < 0$ sicher eine eindeutige Lösung, für $\nu \geq 0$ können wir mit diesem Lemma keine Aussage treffen.

Da es in diesem Fall sehr schwierig ist, allgemeine Aussagen über konvex kombinierte Lobatto-Verfahren zu treffen, untersuchen wir nur den Fall für $s = 3$ Stufen und Lobatto IIIAC, IIIABC und IIIDC.

Lemma 3.6 *Für $s = 3$ Stufen ergeben sich mit $0 \leq \theta \leq 1$ die Koerzivitätskoeffizienten*

$$\begin{aligned}
&\textit{Lobatto IIIAC:} && \alpha_0(A^{-1}) = \frac{4}{3}\frac{\theta^2}{\theta^2 - 8\theta + 8},\\
&\textit{Lobatto IIIABC:} && \alpha_0(A^{-1}) = 0,\\
&\textit{Lobatto IIIDC:} && \alpha_0(A^{-1}) = 0.
\end{aligned}$$

Beweis: Die notwendigen Berechnungen wurden mit Hilfe des Computeralgebraprogramms `Maple V` durchgeführt. Dabei führten die Diagonalmatrizen

$$\begin{aligned}
&\text{Lobatto IIIAC:} && D = \operatorname{diag}(1, 4(1-\theta), 1-\theta),\\
&\text{Lobatto IIIABC:} && D = \operatorname{diag}(f(\theta)^2, -4f(\theta), 1) \quad \text{mit} \quad f(\theta) := \frac{\theta^2 + 12\theta - 24}{\theta^2 - 24\theta + 24},\\
&\text{Lobatto IIIDC:} && D = B
\end{aligned}$$

jeweils auf eine Form von M^D mit Diagonaleinträgen und zusätzlichen Einträgen in m^D_{1s} und m^D_{s1}. Wiederum folgt aus dem Vergleich mit den Diagonaleinträgen der Inversen die Behauptung. □

Bis auf die Klasse Lobatto IIIABC besitzen die BL-Verfahren also die Koerzivität $\alpha_0(A^{-1}) = 0$. Dies ist im Vergleich zum Radau IIA-Verfahren nicht überzeugend und eventuell mit ein Grund für die Newton-Konvergenzprobleme, die in Kap. 4.2 thematisiert werden, vgl. dazu in [38] das Konzept der B-Konvergenz. Betrachtet man (2.2a) als Fixpunktiteration, erhält man mit dem Banachschen Fixpunktsatz die Existenz und Eindeutigkeit einer Lösung nur unter starker Einschränkung an die Schrittweite h.

Konvergenz Kommen wir vom Konvergenzverhalten des Newton-Verfahrens zu dem des RK-Verfahrens an sich zurück. In Satz 2.4 auf Seite 30 haben wir unter relativ starken Voraussetzungen für das allgemeine DAE-System vom Index 3 Aussagen über den globalen Fehler eines IRK-Verfahrens getroffen, die wir auf die Lobatto IIIAC-Familie mit $\theta < 1$ und das Radau IIA-Verfahren anwenden können. Es ergibt sich

$$\begin{aligned}
y_n - y(t_n) &= \mathcal{O}(h^{2s-3}),\\
z_n - z(t_n) &= \mathcal{O}(h^{s-1}),\\
u_n - u(t_n) &= \mathcal{O}(h^{s-2})
\end{aligned}$$

für Lobatto IIIAC mit $s \geq 3$ und $\theta < 1$ und

$$\begin{aligned}
y_n - y(t_n) &= \mathcal{O}(h^{2s-1}),\\
z_n - z(t_n) &= \mathcal{O}(h^{s}),\\
u_n - u(t_n) &= \mathcal{O}(h^{s-1})
\end{aligned}$$

für Radau IIA mit $s \geq 2$ für den uns interessierenden Spezialfall $k_{uu} = 0$.

Während in letzterem Verfahren keine Ordnungsreduktion in der y-Komponente vorliegt, ist die Konvergenzordnung für Lobatto IIIC um eins niedriger als die klassische Ordnung. In der z-Komponente macht sich außerdem die höhere Stufenordnung bemerkbar.

Für andere Lobatto-Verfahren ist Satz 2.4 nicht anwendbar, da sie entweder nicht steifgenau sind wie Lobatto IIID oder die Koeffizientenmatrix nicht invertierbar ist wie bei Lobatto IIIA und IIIB. Für Lobatto IIIA als Kollokationsverfahren wurde der lokale und globale Fehler allerdings separat in [2, Th.4] bewiesen. Es ergibt sich wiederum für $k_{uu} = 0$ und $s \geq 3$

$$\begin{aligned} y_n - y(t_n) &= \mathcal{O}(h^{2s-4}), \\ z_n - z(t_n) &= \mathcal{O}(h^{s-1}), \\ u_n - u(t_n) &= \mathcal{O}(h^{s-3}), \end{aligned}$$

so dass für DAE's vom Index 3 und $s = 3$ keine Konvergenz in der u-Komponente vorliegt. Für steife mechanische Systeme sind aber nur die y- und z-Komponente von Bedeutung, so dass mit Lobatto IIIA für $s \geq 3$ ein weiteres Verfahren aus der Lobatto-Familie zur Verfügung steht, das auf steife mechanische Systeme angewendet werden kann. Dasselbe erwarten wir auch für die ganze Klasse Lobatto IIIAC.

Bevor wir auf spezielle Eigenschaften bei Anwendung auf Hamilton-Systeme eingehen, wollen wir kurz die Vor- und Nachteile der einzelnen Klassen zusammenfassen:

a) Die Variante IIIAB kombiniert zwei A-stabile Verfahren miteinander. Wegen der Differenz von 2 in der Stufenordnung sind die erzielten Verfahren nicht von großer Bedeutung, allerdings erhalten wir für $\theta = 0.5$ ein symplektisches Verfahren, welches in den Klassen c) und e) zum Kombinieren herangezogen wird. Dieses nennen wir auch das *Lobatto IIIAB*-Verfahren.

b) Durch die Steifgenauigkeit beider Randverfahren von IIIAC sind die Verfahren dieser Klasse steifgenau, außerdem erhalten wir L-Stabilität für $\theta < 1$ und einen Koerzivitätskoeffizienten $\alpha_0(A^{-1}) > 0$ für $\theta > 0$. Die B-Stabilität von IIIC geht verloren.

c) Durch die Symplektizität von IIIAB und B-Stabilität von IIIC entsteht durch Kombination eine Klasse von B-stabilen Verfahren.

d) Lobatto IIIDC ist ebenfalls eine B-stabile BL-Familie, besitzt allerdings die Stufenordnung $s - 1$ im Gegensatz zu c) mit Stufenordnung $s - 2$. Wir werden daher bei den Anwendungen eher auf diese Paarung zurückgreifen.

e) Die Kombination von zwei symplektischen Verfahren IIIAB und IIID führt auf eine ganze Klasse symplektischer Verfahren, die von Chan [9] eingeführt wurde. Für die Anwendung auf steife mechanische Systeme fehlt es dieser Klasse an Stabilität.

3.3 Eigenschaften von Hamilton-Systemen

In Kap. 2 haben wir in Anlehnung an [16] bewiesen, dass für symplektische Verfahren die aus der Rückwärtsanalyse hervorgegangene gestörte Differentialgleichung wieder ein Hamilton-System bildet. In diesem Abschnitt wollen wir untersuchen, inwieweit man diese Aussage auf nichtsymplektische Verfahren übertragen kann.

Genauer gesagt möchten wir eine Aussage darüber treffen, wie sich der Energiefehler verhält, indem wir ihn in Bezug zu einem gestörten Hamilton-System stellen. Liegt bei der gestörten Differentialgleichung (2.24), die von einem numerischen Verfahren anstelle des ursprünglichen Systems gelöst wird, ein Hamilton-System vor, so bleibt der Energiefehler beschränkt. Uns interessiert die Ordnung der Terme von (2.24), die nicht zu einem Hamilton-System zusammengefasst werden können. Diese beschreiben dann das Wegtriften des Energiefehlers von einer beschränkten Größe.

3.3.1 Symplektizität der Ordnung σ

Definition 3.1 *Wir nennen ein RK-Verfahren symplektisch der Ordnung σ genau dann, wenn die Koeffizienten der B-Reihe (2.21) die Bedingung*

$$\frac{\mathrm{a}(v \circ w)}{\gamma(v \circ w)} + \frac{\mathrm{a}(w \circ v)}{\gamma(w \circ v)} = \frac{\mathrm{a}(v)}{\gamma(v)} \cdot \frac{\mathrm{a}(w)}{\gamma(w)}, \quad v, w \in \mathcal{T} \quad \textit{mit} \quad \rho(v \circ w) \leq \sigma \tag{3.6}$$

erfüllen.

Wir wollen untersuchen, welche Eigenschaften symplektische Verfahren der Ordnung σ besitzen.

Lemma 3.7 *Wir betrachten Abbildungen* $\mathrm{a} : \mathcal{T} \to \mathbb{R}$ *und* $\mathrm{b} : \mathcal{T} \to \mathbb{R}$*, die über (2.25) in Beziehung stehen. Dann ist (2.27) äquivalent zu (2.28) unter der Einschränkung $\rho(v \circ w) \leq \sigma$.*

Beweis: Der Beweis zu Lemma 2.15 ist induktiv aufgebaut, so dass wir prinzipiell zum Beweisen dieses Lemmas genauso vorgehen. Der Unterschied besteht darin, dass wir für $\rho(v \circ w) = \sigma + 1$ nicht mehr auf (2.27) zurückgreifen können. Damit endet die Induktion und wir erhalten (2.28) ebenfalls nur bis $\rho(v \circ w) \leq \sigma$. □

Ausgehend von Lemma 3.7 können wir auch schon eine wichtige Eigenschaft für symplektische Verfahren der Ordnung σ benennen. Zuvor benötigen wir noch die

Definition 3.2 *Eine Differentialgleichung (2.24) heißt ein* **Hamilton-System bis Ordnung** σ*, falls eine Hamilton-Funktion* $\tilde{H}(p,q)$ *existiert mit*

$$\tilde{p}' = -\frac{\partial \tilde{H}}{\partial q}(\tilde{p},\tilde{q}) + \mathcal{O}\left(h^{\sigma+1}\right),$$
$$\tilde{q}' = \frac{\partial \tilde{H}}{\partial p}(\tilde{p},\tilde{q}) + \mathcal{O}\left(h^{\sigma+1}\right).$$

Es gilt dann folgender

Satz 3.8 *Die gestörte Differentialgleichung (2.24) mit den elementaren Differentialen (2.20e) eines Hamilton-Systems*

$$f_1(y,t) = -\frac{\partial H}{\partial q}(p,q)$$
$$f_2(y,t) = \frac{\partial H}{\partial p}(p,q)$$

mit $y = (p,q)^T$ *ist ein Hamilton-System bis Ordnung* σ*, wenn*

$$\frac{\mathrm{b}(v \circ w)}{\gamma(v \circ w)} + \frac{\mathrm{b}(w \circ v)}{\gamma(w \circ v)} = 0, \quad v \in \mathcal{T}_1, w \in \mathcal{T}_2 \quad \textit{mit} \quad \rho(v \circ w) \leq \sigma. \tag{3.7}$$

Beweis: Zum Beweis dieses Satzes gehen wir die einzelnen Schritte des ähnlichen Satzes 2.12 auf Seite 39 durch und zeigen, dass das Fehlen der Bedingungen (3.7) für $\rho(v \circ w) > \sigma$ keine Auswirkung auf die Aussagen bis $\rho(v \circ w) = \sigma$ hat.

Die Lemmata 2.13 und 2.14 auf den Seiten 40 und 41 hängen nur von der rechten Seite der Differentialgleichung und nicht von dem verwendeten numerischen Verfahren ab, so dass beide unverändert übernommen werden können.

Der eigentliche Beweis von Satz 2.28 stellt die Größen $\beta(v)$, $\beta(w)$ und $\mathrm{b}(v)$, $\mathrm{b}(w)$ gegenüber und berechnet neue Terme $\mathrm{c}(u)$ aus dem Vergleich für jedes $k = \rho(v \circ w)$. Da diese Gleichungen für alle $k \in \mathbb{N}$ unabhängig sind, können wir dieses Vorgehen bis $k = \sigma$ übernehmen. Für $k > \sigma$ erfüllen die Ausdrücke $\mathrm{b}(u)$ nicht mehr die Bedingung (3.7) und können daher keine elementaren Hamilton-Funktionen mehr bilden. Diese Terme lassen sich dann in $\mathcal{O}(h^{\sigma+1})$ zusammenfassen. □

Mit diesem Satz haben wir die Hauptaussage dieses Abschnitts bewiesen. Ähnlich wie in Satz 2.17 können wir diese Aussage mit Lemma 3.7 verknüpfen und erhalten, dass symplektische Verfahren der Ordnung σ angewandt auf ein Hamilton-System als gestörte Differentialgleichung (2.24) ein Hamilton-System bis Ordnung σ zur Folge haben. Doch welche Verfahren besitzen welche symplektische Ordnung?

Lemma 3.9 *Numerische Verfahren mit klassischer Ordnung p besitzen eine symplektische Ordnung $\sigma \geq p$.*

Beweis: Für Verfahren mit klassischer Ordnung p gilt $\mathrm{a}(u) = 1$ für $\rho(u) \leq p$. Die Rekursionsformel für $\gamma(u)$ (2.20c) liefert daher für alle $v \in \mathcal{T}_1$ und $w \in \mathcal{T}_2$ mit $\rho(v \circ w) \leq p$ die Beziehung

$$\begin{aligned}\frac{\mathrm{a}(v \circ w)}{\gamma(v \circ w)} + \frac{\mathrm{a}(w \circ v)}{\gamma(w \circ v)} &= \frac{1}{\gamma(v \circ w)} + \frac{1}{\gamma(w \circ v)} \\ &= \frac{1}{\rho(v \circ w)\gamma(v)\gamma(w_1)\cdots\gamma(w_m)} + \frac{1}{\rho(w \circ v)\gamma(w)\gamma(v_1)\cdots\gamma(v_l)} \\ &= \frac{\rho(w)}{\rho(v \circ w)\gamma(v)\gamma(w)} + \frac{\rho(v)}{\rho(w \circ v)\gamma(w)\gamma(v)} \\ &= \frac{1}{\gamma(v)} \cdot \frac{1}{\gamma(w)} = \frac{\mathrm{a}(v)}{\gamma(v)} \cdot \frac{\mathrm{a}(w)}{\gamma(w)}\end{aligned}$$

mit $v = [v_1, \ldots v_l]$, $w = [w_1, \ldots, w_m]$ und $\rho(v \circ w) = \rho(w \circ v) = \rho(v) + \rho(w)$. □

Damit bilden für numerische Verfahren mit klassischer Ordnung p jeweils die gestörten Differentialgleichungen (2.24) ein Hamilton-System bis Ordnung p. Wegen $\mathrm{b}(u) = 0$ für $1 < \rho(u) \leq p$ und $\mathrm{b}(u) = 1$ für $u = \tau$ folgt für Verfahren mit $\sigma = p$, dass die gestörte Hamilton-Funktion aus Satz 3.2 $\tilde{H}(p, q)$ identisch ist mit der ursprünglichen Hamilton-Funktion $H(p, q)$.

Für die beiden B-stabilen Familien der BL-Verfahren erhalten wir

Lemma 3.10 *Die Familien Lobatto IIIABC und IIIDC sind für $s = 3$ Stufen symplektisch der Ordnung $\sigma = 5 = p + 1$.*

Beweis: Wegen Lemma 3.9 reicht es aus zu zeigen, dass die Bedingung (2.27) an die Koeffizienten $\mathrm{a}(v \circ w)$ für $\rho(v \circ w) = 5$ erfüllt ist. Da $v \circ w$ und $w \circ v$ äquivalente Bäume bezeichnen, suchen wir zunächst nach den Paaren v und w, die untersucht werden müssen.

Nummeriert man die Bäume der Ordnung 5 aus Tab. 2.2 von links nach rechts mit $u_1, \ldots, u_9$, ergeben sich die Paare

$$u_1 - u_5, \quad u_2 - u_7, \quad u_3 - u_8, \quad u_4 - u_9, \quad u_5 - u_6, \quad \text{und} \quad u_7 - u_8.$$

Die Größen $\mathrm{a}(u_i)$ der Lobatto IIIDC-Familie berechnen sich zu

$$\begin{aligned}\mathrm{a}(u_1) &= \frac{5}{4}, \quad \mathrm{a}(u_2) = \frac{5}{4}, \quad \mathrm{a}(u_3) = \frac{5}{6}, \\ \mathrm{a}(u_4) &= \frac{5}{6}, \quad \mathrm{a}(u_5) = \frac{15}{16}, \quad \mathrm{a}(u_6) = \frac{25}{24}, \\ \mathrm{a}(u_7) &= \frac{15}{16}, \quad \mathrm{a}(u_8) = \frac{25}{24}, \quad \mathrm{a}(u_9) = \frac{25}{24}\end{aligned}$$

und mit den gegebenen $\gamma(u)$ aus Tab. 2.2 und $\mathrm{a}(u) = 1$ für $\rho(u) \leq 4$ erhalten wir mit Hilfe der oben genannten Paarungen die Behauptung, ganz analog für Lobatto IIIABC. □

Die höhere symplektische Ordnung der B-stabilen BL-Verfahren macht sich bemerkbar, sobald man den Energiefehler eines konservativen Systems in Hamilton-Form betrachtet. In Kap. 5 werden wir diesen anhand eines einfachen Beispiels untersuchen und einen Vergleich zwischen Lagrange- und Hamilton-Formulierung durchführen.

3.3.2 Gemischte Lagrange-Hamilton-Systeme

Da die Überführung von Lagrange- in Hamilton-Koordinaten umständlich sein kann, sind wir auch an gemischten Lagrange-Hamilton-Systemen interessiert. Diese liegen bereits dann vor, wenn wir auf gewohnte Weise ein konservatives Lagrange-System aufstellen und lineare Subsysteme vorliegen, die z.B. durch Semidiskretisierung entstanden sein können.

Formal schreiben wir ein solches System als

$$\dot{q}_H = \frac{\partial}{\partial p_H} H(q_H, p_H) \tag{3.8a}$$

$$\dot{p}_H = -\frac{\partial}{\partial q_H} H(q_H, p_H) \tag{3.8b}$$

$$\dot{q}_L = v_L \tag{3.8c}$$

$$M(q_H, q_L)\dot{v}_L = f(q_H, q_L, p_H, v_L), \tag{3.8d}$$

wobei q_H und p_H die Koordinaten des Hamilton-Subsystems und q_L und v_L die des Lagrange-Subsystems kennzeichnen.

Entstehen können solche Systeme, wenn beim Aufstellen der kinetischen und potentiellen Energie eines konservativen Systems eine Aufspaltung der Form

$$\begin{aligned} T(q_H, q_L, p_H, v_L) &= T_H(q_H, p_H) + T_L(q_H, q_L, p_H, v_L) \\ U(q_H, q_L) &= U_H(q_H) + U_L(q_H, q_L) \end{aligned}$$

mit

$$H(q_H, p_H) := T_H(q_H, p_H) + U_H(q_H) = \text{konst.}$$

existiert. Dann erhält man (3.8a) und (3.8b) direkt durch Anwendung auf obige Hamilton-Funktion sowie (3.8c) und (3.8d) durch den Lagrange-Formalismus mit

$$L(q_H, q_L, p_H, v_L) = T_L(q_H, q_L, p_H, v_L) - U_L(q_H, q_L),$$

wobei nur die partiellen Ableitungen nach q_L und $v_L = \dot{q}_L$ herangezogen werden.

Liegt ein System der Form (3.8) vor, gelten für den separaten Hamilton-Teil der Gleichungen dieselben Aussagen wie für komplette Hamilton-Systeme, da die Gleichungen in q_H und p_H unabhängig von den Lagrange-Koordinaten q_L und v_L sind und damit getrennt berechnet werden können.

Eine Verallgemeinerung auf Systeme der Form

$$\dot{q}_H = \frac{\partial}{\partial p_H} H(q_H, q_L, p_H, v_L) \tag{3.9a}$$

$$\dot{p}_H = -\frac{\partial}{\partial q_H} H(q_H, q_L, p_H, v_L) \tag{3.9b}$$

$$\dot{q}_L = v_L \tag{3.9c}$$

$$M(q_H, q_L)\dot{v}_L = f(q_H, q_L, p_H, v_L), \tag{3.9d}$$

wobei die Hamilton-Funktion H zusätzlich von q_L und v_L abhängen darf, ist nicht sinnvoll. Für Hamilton-Systeme wird stets die Forderung

$$\frac{d}{dt} H(q, p) = 0$$

vorausgesetzt, die auch direkt die Formulierung beeinflusst. Im Fall von (3.9a) und (3.9b) ergäbe sich aber der Ausdruck

$$\frac{d}{dt} H(q, p) = H_{q_H} \dot{q}_H + H_{q_L} \dot{q}_L + H_{p_H} \dot{p}_H + H_{v_L} \dot{v}_L \tag{3.10a}$$

$$= H_{q_H} H_{p_H} + H_{q_L} v_L - H_{p_H} H_{q_H} + H_{v_L} M^{-1} f \tag{3.10b}$$

$$= H_{q_L} v_L + H_{v_L} M^{-1} f, \tag{3.10c}$$

welcher i.A. nicht identisch verschwindet. Aus diesem Grund könnten die allgemeinen Aussagen über Hamilton-Systeme nicht angewendet werden und die Formulierung (3.9) brächte keine neuen Erkenntnisse.

In diesem Kapitel haben wir verschiedene konvexe Kombinationen von Verfahren der Lobatto-Familie kennengelernt. Aufgefallen ist dabei die Vielseitigkeit an Verfahren, die dadurch erzielt werden können. Neben der Wahl zwischen einer L-stabilen und zwei B-stabilen Familien können auch die lineare Ordnung bzw. Dispersions- oder Dissipationsordnung durch ein bestimmte Parameterwahl erhöht werden.

Vorteilhaft ist auch die Tatsache, dass sich die meisten Eigenschaften der Randverfahren ohne größere Verluste auf die kombinierten Verfahren übertragen. Dies gilt nicht nur für die Eigenschaften der linearen Theorie, sondern auch für Konvergenzaussagen sowie Existenz und Eindeutigkeit von RK-Lösungen.

Kapitel 4

Praktische Aspekte

In diesem Kapitel steht die Umsetzung impliziter RK-Verfahren im Vordergrund. Bei der Anwendung auf steife mechanische Systeme treten zusätzliche Probleme auf, die bei einer Implementierung berücksichtigt werden müssen. Im Einzelnen sind dies die auftretende Ordnungsreduktion und Konvergenzprobleme des vereinfachten Newton-Verfahrens. Die geringere Ordnung einzelner Komponenten sorgt für eine Überschätzung des Fehlers und zieht somit zu kleine Schrittweiten nach sich. Aber auch die Wahl des Fehlerschätzers ist entscheidend.

Abgesehen von diesen Schwierigkeiten treten Schrittweiteneinbußen bei der Lösung des nichtlinearen Gleichungssystems auf. Die Iterationsmatrix des vereinfachten Newton-Verfahrens ist bei steifen Systemen im Allgemeinen schlecht konditioniert und die zugehörige Fixpunktiteration nur bei sehr guten Anfangswerten konvergent, so dass die Schrittweite verkleinert werden muss, um einen genaueren Startwert zu erhalten.

Neben dieser Problembehandlung erlaubt uns die Kombination von Lobatto-Verfahren, eine zusätzliche Dimension in die Integration einzubringen, indem die Parameter des Verfahrens an die Erfordernisse des Systems angepasst werden.

Der erste Abschnitt befasst sich mit der Schrittweitensteuerung, wobei unterschiedliche eingebettete Verfahren und die Ordnungsreduktion steifer mechanischer Systeme analysiert werden. Anschließend wollen wir unterschiedliche Formulierungen in Bezug auf das Newton-Konvergenzverhalten von impliziten RK-Verfahren vergleichen. Den Abschluss dieses Kapitels bildet eine Untersuchung der BL-Familien und ihrer Anwendungsmöglichkeiten.

4.1 Schrittweitensteuerung

Die Schrittweitensteuerung spielt bei vielen technischen Anwendungen eine große Rolle. Nicht nur, weil sie eine Anpassung der Schrittweite an den Lösungsverlauf erlaubt, sondern auch, weil man dadurch eine Fehlertoleranz vorgeben kann, die einen Hinweis auf die Genauigkeit der numerischen Lösung liefert.

Für die Klasse der IRK-Verfahren stellt sich die Berechnung eines Fehlerschätzers, welcher eine Grundlage für die Schrittweitensteuerung darstellt, als besonders effizient heraus. Grund dafür sind die sogenannten *eingebetteten* Verfahren, welche die gleiche Koeffizientenmatrix A und den gleichen Knotenvektor c wie das ursprüngliche Verfahren aber einen neuen Gewichtsvektor $\hat{b}$ besitzen. Da der Hauptaufwand der IRK-Verfahren in der Lösung des nichtlinearen Gleichungssystems (2.2a) liegt, welches für beide, ursprüngliches und eingebettetes Verfahren, identisch ist, ist der zusätzliche Aufwand für die Berechnung des Fehlerschätzers gering.

Umformulierung Bevor wir die Schrittweitensteuerung erläutern, sind zur Implementierung noch einige Umformulierungen sinnvoll. Wie in [20, S.118] beschrieben, werden die Inkremente Y_i nicht direkt berechnet, sondern die kleineren Größen

$$z_i = Y_i - y_0,$$

welche weniger anfällig für Rundungsfehler sind. Damit schreibt sich das zu lösende Gleichungssystem (2.2a) als

$$z_i = h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, y_0 + z_j), \quad i = 1, \ldots, s$$

und wegen

$$Z = hAF$$

mit $Z = (z_1, \ldots, z_s)^T$ und $F = (f(t_0 + c_1 h, y_0 + z_1), \ldots, f(t_0 + c_s h, y_0 + z_s))^T$ können wir die RK-Lösung (2.2b) in der Form

$$y_1 = y_0 + \sum_{i=1}^{s} d_i z_i$$

darstellen. Die neuen Koeffizienten d_i berechnen sich aus $d = A^{-T} b$, welches im Fall von steifgenauen Verfahren wegen $b_i = a_{si}$ gerade dem s-ten Einheitsvektor entspricht.

Genauso kann man auch das eingebettete Verfahren formulieren und erhält

$$\hat{y}_1 = y_0 + \sum_{i=1}^{s} \hat{d}_i z_i$$

bzw.

$$\Delta\mu := \hat{y}_1 - y_1 = \sum_{i=1}^{s} e_i z_i \tag{4.1}$$

mit $\hat{d} = A^{-T}\hat{b}$ und $e_i = \hat{d}_i - d_i$.

Standardvorgehen Da für den steifen Fall der einfache Fehlerschätzer $\Delta\mu$ unbeschränkt ist, verwendet man meistens den modifizierten Schätzer

$$\Delta\nu := (I - h\gamma_0 J)^{-1}(\hat{y}_1 - y_1) \tag{4.2}$$

mit der Jacobi-Matrix J von f, siehe z.B. [20, 10]. Die Berechnung von $\Delta\nu$ ist sehr billig, da aus dem vereinfachten Newton-Verfahren bereits eine LR-Zerlegung von $(I - h\gamma_0 J)$ für einen festen Parameter γ_0 bekannt ist, für Details siehe [20]. Dieser Fehlerschätzer ist zudem beschränkt für $h\lambda \to \infty$.

Bemerkung 4.1 *Für steifgenaue Verfahren wie Radau IIA und Lobatto IIIC stimmt der Fehlerschätzer $\Delta\nu$ bis auf einen konstanten Faktor mit den in [10] eingeführten impliziten Fehlerschätzer*

$$\hat{y}_{n+1} = y_n + h\left(\hat{b}_0 f(y_n) + (\hat{b}^T \otimes I)F(Y) + \gamma_0 f(\hat{y}_{n+1})\right)$$

überein, wenn zur Auflösung nach $\hat{y}_{n+1}$ genau ein Newton-Schritt durchgeführt wird. Dieser erzeugt den Faktor $(I - h\gamma_0 J)^{-1}$ während der Newton-Iteration und wegen der impliziten Berechnung des eingebetteten Verfahrens besitzt dieser Fehlerschätzer wie das Originalverfahren gute Stabilitätseigenschaften.

Mit $\Delta\nu$ haben wir einen guten Schätzer für den lokalen Fehler $y(t_1) - y_1$ vorliegen. Um zu entscheiden, ob der Fehler klein genug ist, geben wir eine absolute $Atol_i$ und relative Toleranz $Rtol_i$ für jede Komponente y_i vor und überprüfen, ob

$$||err|| := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\Delta\nu_i}{sc_i}\right)} < 1$$

mit $\Delta\nu_i$ gleich der i-ten Komponente von $\Delta\nu$ und

$$sc_i = Atol_i + \max(|y_{0,i}|, |y_{1,i}|)Rtol_i$$

erfüllt ist. Ist dies nicht der Fall, wird der Schritt verworfen und mit einer kleineren Schrittweite wiederholt. Die neue Schrittweite h_{new} berechnet sich aus

$$h_{new} = fac \cdot h \cdot ||err||^{-1/(\hat{p}+1)}$$

mit der klassischen Ordnung $\hat{p} < p$ des eingebetteten Verfahrens und einem Sicherheitsfaktor fac. Obige Formel basiert auf dem Fehlerterm

$$\hat{y}_1 - y_1 = C_1 \cdot h^{\hat{p}+1} \tag{4.3}$$

unter der Annahme, dass die Konstante C_1 sich von einem Schritt zum nächsten kaum verändert, also

$$C_n \approx C_{n-1} \tag{4.4}$$

mit den Konstanten C_n aus dem Fehlerterm für den lokalen Fehler im n-ten Schritt.

Verwendet man statt (4.4) die Relation

$$\frac{C_{n+1}}{C_n} \approx \frac{C_n}{C_{n-1}},$$

gelangt man zum *Vorhersageregler (predictive controller)* mit

$$h_{new} = fac \cdot h_n \cdot ||err_{n+1}||^{-1/(\hat{p}+1)} \cdot \left(\frac{||err_n||}{||err_{n+1}||} \right)^{1/(\hat{p}+1)},$$

siehe [20, S. 124]. Weitere alternative Ansätze gewinnt man aus der Kontrolltheorie, siehe z.B. Söderlind [45].

4.1.1 Konstruktion eingebetteter Verfahren

Das eingebettete Verfahren mit Koeffizienten $(A, \hat{b}, c)$ besitzt normalerweise eine geringere Ordnung als das Verfahren (A, b, c). Dies hat zur Folge, dass die Differenz $\hat{y}_1 - y_1$ nicht immer das erwartete Konvergenzverhalten zeigt.

Neben der Effizienz des eingebetteten Verfahrens bildet die Forderung, dass der durch das eingebettete Verfahren entstandene Fehlerschätzer das Konvergenzverhalten des globalen Fehlers widerspiegelt, das Hauptkriterium zur Auswahl eines geeigneten Fehlerschätzers. Ziel ist es, dass bei vorgegebenen Toleranzen $Atol_i$ und $Rtol_i$ der tatsächliche Fehler sich innerhalb dieses Rahmens bewegt.

Um dies zu erreichen, ist es möglich, ohne viel zusätzlichen Aufwand das Butcher-Tableau des Originalverfahrens um eine zusätzliche Stufe zu erweitern und diese für das eingebettete Verfahren heranzuziehen. Falls alle Knoten $c_i \neq 0$ sind, bietet sich $c_0 = 0$ als zusätzliche Stufe an.

Für das Radau IIA-Verfahren mit $s = 3$ Stufen erhält man wie in [20] mit

$$\hat{y}_1 - y_1 = \gamma_0 h f(t_0, y_0) + \sum_{i=1}^{3} e_i z_i$$

0	0	0	0	0
$\frac{4-\sqrt{6}}{10}$	0	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	0	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	0	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
$(\text{IIA})\widehat{}$	γ_0	$b_1 - \frac{2+3\sqrt{6}}{6}\,\gamma_0$	$b_2 - \frac{2-3\sqrt{6}}{6}\,\gamma_0$	$b_3 - \frac{1}{3}\,\gamma_0$

Tabelle 4.1: Butcher-Tableau für Radau $(\text{IIA})\widehat{}$.

und

$$(e_1, e_2, e_3) = \gamma_0(-13 - 7\sqrt{6}, -13 + 7\sqrt{6}, -1)$$

ein eingebettetes Verfahren der Ordnung $\hat{p} = 3$, im Folgenden Radau $(\text{IIA})\widehat{}$ genannt, während ohne zusätzliche Stufe nur Ordnung $\hat{p} = 2$ möglich wäre. Dabei ist keine zusätzliche Funktionsauswertung notwendig, da $f(t_0, y_0)$ wegen der Steifgenauigkeit aus der letzten Stufe des letzten Schrittes bekannt ist. Das eingebettete Verfahren entspricht dem erweiterten Butcher-Tableau der Tab. 4.1 mit $s = 4$ Stufen, wobei die untere 3×3-Matrix dem Standard Radau IIA-Verfahren entnommen ist.

Wir wollen versuchen, diese Idee auf die einzelnen BL-Familien zu übertragen, wobei wir möglichst einen Fehlerschätzer unabhängig vom Parameter θ im Auge haben. Falls man Letzteres nicht berücksichtigt, würde man für verschiedene Werte von θ in einem System verschiedene Fehlerkoeffizienten benötigen. Diese Umsetzung wäre zwar prinzipiell möglich, würde aber den Programmieraufwand für den Fehlerschätzer erheblich erhöhen.

Betrachten wir zunächst die Familie Lobatto IIIAC. Ein Problem hierbei besteht darin, dass die oben hinzugenommene Stufe mit Nulleinträgen bereits bei Lobatto IIIA vorliegt. Dies hat zum Einen zur Folge, dass die zusätzliche Stufe etwas komplizierter sein muss, zum Anderen ist die Koeffizientenmatrix nicht invertierbar und dadurch wird die Darstellung der Form (4.1) nicht immer möglich sein. Für das Lobatto IIIA-Verfahren ist Letzteres kein Problem, da wegen der Steifgenauigkeit der Vektor b im Bild von A^T liegt. Für mögliche eingebettete Verfahren muss dies aber nicht der Fall sein.

Mit dem Ansatz

$$(\hat{a}_{1j})_j = (0, \alpha_1, \alpha_2, \alpha_3)$$

für die zusätzliche Koeffizientenzeile des eingebetteten Verfahrens erhalten wir

0	0	$-\frac{1}{4}(\theta-2)\alpha_0$	$\frac{1}{2}(\theta-2)\alpha_0$	$-\frac{1}{4}(\theta-2)\alpha_0$
0	0	$\frac{1}{6}(1-\theta)$	$-\frac{1}{3}(1-\theta)$	$\frac{1}{6}(1-\theta)$
$\frac{1}{2}$	0	$\frac{1}{24}(4+\theta)$	$\frac{1}{12}(5-\theta)$	$-\frac{1}{24}(2-\theta)$
1	0	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
$(\text{IIIAC})\hat{}$	γ_0	$\frac{1}{6}-\gamma_0$	$\frac{2}{3}$	$\frac{1}{6}$

Tabelle 4.2: Butcher-Tableau für Lobatto (IIIAC)$\hat{}$.

unter Berücksichtigung der Ordnungsbedingungen mit Hilfe von `Maple V` ein eingebettetes Verfahren der Ordnung $\hat{p}=3$ mit dem Butcher-Tableau aus Tab. 4.2.

Es hängt von zwei Parametern γ_0 und α_0 ab, wobei Ersterer die gleiche Rolle spielt wie für das Radau IIA-Verfahren. Der Parameter α_0 sollte so gewählt werden, dass für keinen Wert von θ die zusätzliche Stufe mit der ersten Stufe des Lobatto IIIAC-Verfahrens übereinstimmt. Dies ist genau dann der Fall, wenn

$$\theta = 2\,\frac{3\alpha_0-1}{3\alpha_0-2}$$

erfüllt ist. Für $0 \le \alpha_0 \le 1/3$ erhalten wir einen Wert $0 \le \theta \le 1$, so dass dieser Bereich ausgeschlossen werden muss. Als Standardwert verwenden wir $\alpha_0 = 1/4$.

Um dieses eingebettete Verfahren in der Form (4.1) schreiben zu können, benötigen wir einen Trick, da die Koeffizientenmatrix von Lobatto IIIA nicht invertierbar ist. Dazu schreiben wir

$$Z = h\tilde{A}F - hF_1$$

mit $F_1 = (f(t_0+c_1h, y_0+z_1), 0, \ldots, 0)^T$ und der Matrix $\tilde{A} = (\tilde{a}_{ij})$ mit Koeffizienten $\tilde{a}_{11} = a_{11}+1$ und $\tilde{a}_{ij} = a_{ij}$ sonst. Daraus ergibt sich mit den Koeffizienten

$$\beta = \tilde{A}^{-T}\hat{b} = \alpha_0\left(\frac{3}{4}, -3, \frac{3}{4}\right)^T,$$

welche wie gewünscht unabhängig von θ sind, die Beziehung

$$z_0 = \beta_1\left(z_1 + hf(t_0+c_1h, y_0+z_1)\right) + \beta_2 z_2 + \beta_3 z_3$$

mit den Einträgen β_i von β.

Für die eigentliche Berechnung des Fehlerschätzers $\Delta\mu$ verzichten wir auf obige Umformulierung, um die Unabhängigkeit von θ zu gewährleisten. Wir erhalten

$$\Delta\mu = \hat{y}_1 - y_1 = h\gamma_0\left[f(t_0, y_0+z_0) - f(t_0+c_1h, y_0+z_1)\right] \tag{4.5}$$

0	0	$-\alpha_0(2-\theta)$	$2\alpha_0(2-\theta)$	$-\alpha_0(2-\theta)$
0	0	$\frac{1}{12}(2-\theta)$	$-\frac{1}{6}(2-\theta)$	$\frac{1}{12}(2-\theta)$
$\frac{1}{2}$	0	$\frac{1}{24}(4+\theta)$	$\frac{1}{12}(5-\theta)$	$-\frac{1}{24}(2-\theta)$
1	0	$\frac{1}{12}(2-\theta)$	$\frac{1}{6}(4+\theta)$	$\frac{1}{12}(2-\theta)$
$(\text{IIIDC})\widehat{\ }$	γ_0	$\frac{1}{6}-\gamma_0$	$\frac{2}{3}$	$\frac{1}{6}$

Tabelle 4.3: Butcher-Tableau für Lobatto $(\text{IIIDC})\widehat{\ }$.

mit nur einer zusätzlichen Funktionsauswertung für $f(t_0, y_0 + z_0)$, da der Term $f(t_0 + c_1 h, y_0 + z_1)$ bereits im Newton-Verfahren berechnet wird.

Für das Lobatto IIIDC-Verfahren ist es nicht notwendig, die Koeffizientenmatrix A abzuändern, da sie für alle $0 \leq \theta \leq 1$ invertierbar ist. Als eingebettetes Verfahren ergibt sich analog zu Lobatto IIIAC das Butcher-Tableau der Tab. 4.3.

Auch wenn sich hier die zusätzliche Stufe gerade als ein Vielfaches der ersten Stufe herausstellt, ist die klassische Ordnung des eingebetteten Verfahrens $\hat{p} = 3$. Als Standardwert wählen wir $\alpha_0 = 1$. Umformulierung in die Form (4.1) liefert

$$\beta = \alpha_0 \left(-12, 0, 0\right)^T,$$

bzw. die zusätzliche Stufe

$$z_0 = \beta_1 z_1.$$

Als Fehlerschätzer ergibt sich wie bei den Lobatto IIIAC-Verfahren die Beziehung (4.5), wobei wieder nur eine zusätzliche Funktionsauswertung $f(t_0, y_0 + z_0)$ vonnöten ist. Erneut sind die Fehlerkoeffizienten unabhängig von θ wählbar.

Damit haben wir für Lobatto IIIAC und IIIDC eingebettete Verfahren der Ordnung $\hat{p} = 3$ gefunden. Für Lobatto IIIABC stellt sich die Suche nach einem solchen Verfahren als schwieriger heraus. Ein eingebettetes Verfahren der Ordnung $\hat{p} = 3$ existiert zwar auch für diese Familie, aber die Fehlerkoeffizienten sind nicht mehr unabhängig vom Parameter θ, sondern hängen sogar nichtlinear davon ab. Auch bei Rückgang auf Ordnung $\hat{p} = 2$ ändert sich daran nicht viel. Neben der höheren Stufenordnung ist dies ein weiterer Grund, die Familie Lobatto IIIDC der Familie IIIABC vorzuziehen.

4.1.2 Analyse eingebetteter Verfahren

Die beiden eingebetteten Verfahren von Lobatto IIIAC und IIIDC mit Ordnung $\hat{p} = 3$ wollen wir im Folgenden mit einem eingebetteten Verfahren vergleichen, welches von Jay in `SPARK3` für die ganze Klasse an Lobatto-Verfahren vorgeschlagen wird. Es berechnet sich nach

$$\hat{d} = \left(\frac{4}{5}, \frac{2}{5}, \frac{4}{5}\right)$$

ohne zusätzlich eingeführte Stufe und besitzt Ordnung $\hat{p} = 1$. Wir wollen die zugehörigen Fehlerschätzer aufgrund folgender Kriterien untersuchen:

a) Effizienz,

b) korrekte Wiedergabe des zu schätzenden Fehlers bzgl. der Ordnung und

c) gleichmäßiges Verhalten von vorgegebener zu erhaltener Toleranz.

Die letzten beiden Punkte sollen anhand einer linearen Testgleichung mit unterschiedlich steifen Anteilen verglichen werden.

Effizienz und Implementierung Das eingebettete Verfahren mit Ordnung $\hat{p} = 1$ ist sehr effizient in der Berechnung, weil es im Vergleich zu den Verfahren mit Ordnung $\hat{p} = 3$ auf zusätzliche Funktionsauswertungen verzichten kann. Es liefert als Fehlerschätzer nur einen Term der Größenordnung $\mathcal{O}(h^2)$, da der lokale Fehler zugrundeliegt. Um den globalen Fehler des verwendeten Verfahrens zu schätzen, wird daher die Differenz aus Original- und eingebettetem Verfahren zunächst quadriert, bevor mit der Matrix $(I - hJ)^{-1}$ multipliziert wird.

Für das eingebettete Verfahren dritter Ordnung ist dies nicht notwendig, da der lokale Fehler bereits einen Term der Größenordnung $\mathcal{O}(h^4)$ liefert. Dafür kommt eine Funktionsauswertung pro Schritt hinzu.

Wiedergabe des zu schätzenden Fehlers Für die Lobatto IIIDC-Familie ergibt sich bei Anwendung auf ein lineares System der Form

$$y' = Jy + u(t)$$

mit

$$J = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} \quad \text{und} \quad u(t) = \begin{pmatrix} 0 \\ \omega^2 \cos(2\Omega t) - 4\Omega^2 \cos(2\Omega t) \end{pmatrix}$$

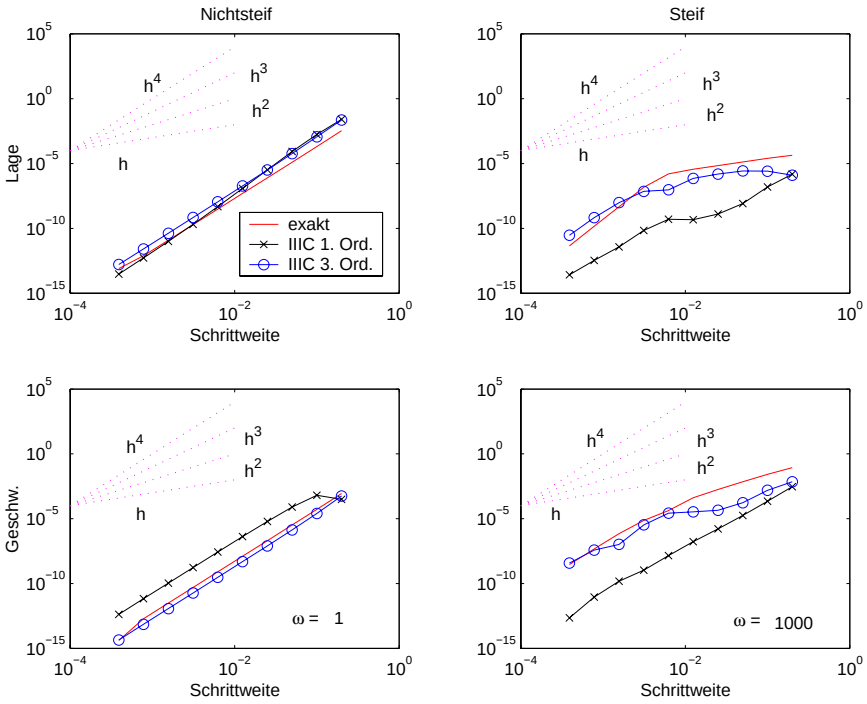


Abbildung 4.1: Zu schätzender globaler Fehler (exakt), Fehlerschätzer erster Ordnung und Fehlerschätzer dritter Ordnung der Lobatto IIIDC-Familie für $\theta = 0$.

das Verhalten aus Abb. 4.1. Dabei wird jeweils eine konstante Schrittweite vorgegeben.

Für nichtsteife Differentialgleichungen zeigen beide Fehlerschätzer ein brauchbares Ergebnis, aber im Fall $\omega = 10^3$ zeigt sich, dass das Verfahren höherer Ordnung das Phänomen der Ordnungsreduktion besser repräsentiert. Für Lobatto IIIAC zeigt sich ein ähnliches Bild, so dass man diesen Punkt als Vorteil für die Fehlerschätzer höherer Ordnung verbuchen kann.

Verhalten von vorgegebener zu erhaltener Toleranz Als letzten Punkt auf der Vergleichsliste möchten wir das Verhalten von vorgegebener zu erhaltener Toleranz untersuchen. Dazu führen wir ähnlich wie zur Ordnungsbestimmung Testläufe durch, diesmal aber mit unterschiedlichen vorgegebenen Toleranzen. Anschließend vergleichen wir die vorgegebenen Toleranzen mit dem entstandenen Fehler, der wegen der Existenz einer analytischen Lösung exakt zur Verfügung steht. Wir erhalten die Ergebnisse aus Abb. 4.2.

Für steife ODEs fällt der große Fehler des Verfahrens erster Ordnung im Bereich kleiner Toleranzen sofort ins Auge. Dieser liegt teilweise um Faktor 10^5 über

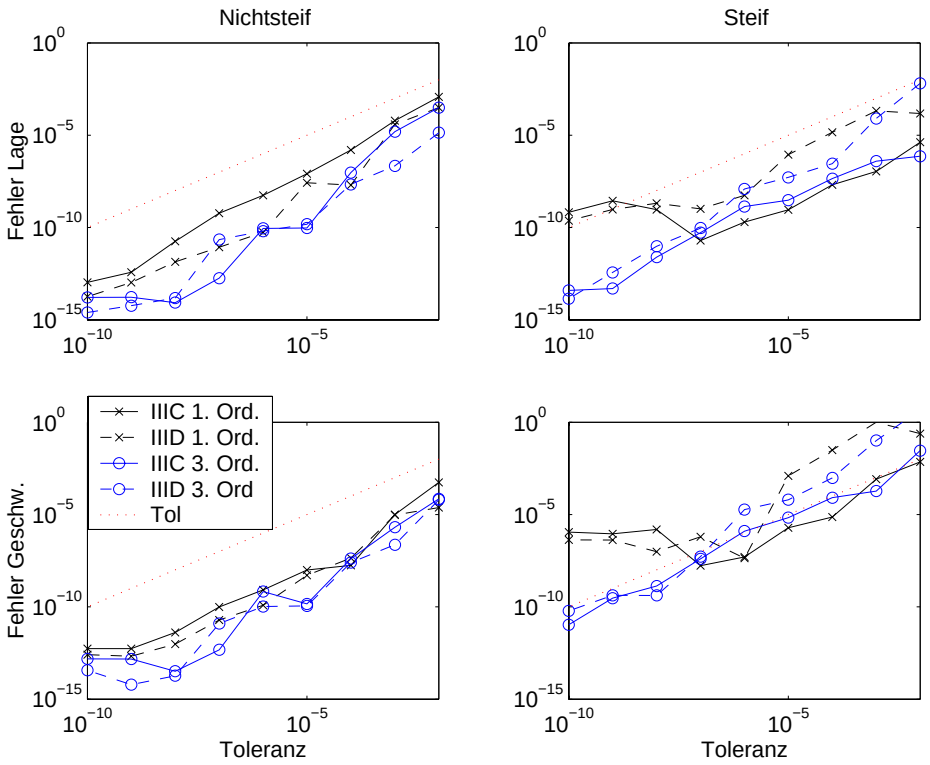


Abbildung 4.2: Verhältnis von vorgegebener und erhaltener Toleranz für die Fehlerschätzer erster und dritter Ordnung von Lobatto IIIC und IIID.

der vorgegebenen Toleranz, die durch die gepunktete Linie dargestellt ist. Das Verfahren höherer Ordnung zeigt dagegen einen relativ gleichmäßigen Verlauf sowohl für Lobatto IIIC als auch Lobatto IIID. Das gleiche Verhalten liegt auch bei Lobatto IIIAC vor.

Aufgrund dieses Vergleichs und unserem Interesse an steifen mechanischen Systemen ziehen wir für numerischen Berechnungen die eingebetteten Verfahren höherer Ordnung vor. Die zusätzlichen Funktionsauswertungen nehmen wir zugunsten eines präziseren Fehlerschätzers in Kauf.

4.1.3 Stabilisierung der Schrittweitensteuerung

Betrachten wir die Abb. 4.1 erneut, stellt sich die Frage, wie man das Verhalten des lokalen Fehlers bei steifen Systemen erklären kann. Wie in Kap. 2 gesehen, wird das steife mechanische System von der Ordnungsreduktion der zugrundeliegenden DAE beeinflusst. Im Folgenden wollen wir anhand einer linearen Testgleichung das Konvergenzverhalten für steife mechanische Systeme direkt unter-

suchen, also ohne auf die Verwandtschaft zu DAEs einzugehen. Ziel dieser Vorgehensweise ist eine Stabilisierung der Schrittweitensteuerung, da bei der Integration steifer mechanischer Systeme oft starke Oszillationen im Schrittweitenverlauf zu beobachten sind.

Analyse des lokalen Fehlers Die folgende Darstellung richtet sich weitestgehend nach Simeon [42]. Dort wird der globale und lokale Fehler anhand von kleinen Abweichungen der Anfangswerte von der glatten Lösung diskutiert. Für L-stabile Verfahren zeigt sich, dass der lokale Fehler über den globalen Fehler dominiert. Wir beschränken uns daher auf eine Untersuchung des lokalen Fehlers, was auch dadurch begründet werden kann, dass für die Schrittweitensteuerung ebenfalls nur dieser herangezogen wird.

Um auf Unterschiede in den Lage- und Geschwindigkeitskomponenten eingehen zu können, wird die Testgleichung

$$\ddot{q} = -\omega^2(q - \varphi) + \ddot{\varphi}$$

verwendet, welche auf Prothero/Robinson zurückgeht. Dabei ist eine glatte Bewegung φ und eine Frequenz ω vorgegeben. Wir gehen außerdem davon aus, dass wir auf der glatten Lösung starten, und wählen daher die Anfangswerte

$$q(0) = \varphi(0), \quad \dot{q}(0) = \dot{\varphi}(0).$$

Geschrieben als System 1. Ordnung folgt

$$\dot{y} = J(y - \psi) + \dot{\psi} \quad \text{mit} \quad J = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}, \tag{4.6}$$

wobei $y = (q, \dot{q})^T$ und $\psi = (\varphi, \dot{\varphi})^T$ gesetzt wird. Wendet man ein IRK-Verfahren auf (4.6) an, ergibt sich

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} \left[J\left(Y_j - \psi(t_0 - c_j h)\right) + \dot{\psi}(t_0 - c_j h) \right], \quad i = 1, \ldots, s, \tag{4.7a}$$

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i \left[J\left(Y_i - \psi(t_0 - c_i h)\right) + \dot{\psi}(t_0 - c_i h) \right]. \tag{4.7b}$$

Um einen Ausdruck für den lokalen Fehler zu erhalten, setzen wir anstelle der rechten Seite von (4.6) die exakte Lösung ψ in das RK-Verfahren ein und erhalten

$$\psi(t_0 + c_i h) = \psi(t_0) + h \sum_{j=1}^{s} a_{ij} \dot{\psi}(t_0 + c_j h) + \Delta_i, \quad i = 1, \ldots, s, \tag{4.8a}$$

$$\psi(t_0 + h) = \psi(t_0) + h \sum_{i=1}^{s} b_i \dot{\psi}(t_0 + c_i h) + \Delta_0, \tag{4.8b}$$

wodurch die Defekte Δ_i und Δ_0 definiert werden. Subtraktion der jeweiligen Gleichungen aus (4.7) und (4.8) liefert nach Elimination von Y_i mit $y_0 = \psi(t_0)$ die Beziehung

$$\delta_0 := y_1 - \psi(t_0 + h) = -(b^T \otimes hJ)(I - A \otimes hJ)^{-1}\Delta_Y - \Delta_0 \tag{4.9}$$

für den lokalen Fehler mit dem Kroneckerprodukt $\otimes$ und $\Delta_Y = (\Delta_1, \ldots, \Delta_s)^T$.

Aufgrund der besonderen Struktur der Jacobi-Matrix J kann man sich diesen Ausdruck detailliert anschauen. Zunächst ist eine Umsortierung in Δ_Y nach Lage- und Geschwindigkeitskomponenten sinnvoll. Dies wird gerade durch eine Vertauschung der Faktoren im Kroneckerprodukt bewerkstelligt. Es ergibt sich

$$\delta_0 = -(hJ \otimes b^T)(I - hJ \otimes A)^{-1}\Delta_{q,\dot{q}} - \Delta_0$$

mit $\Delta_{q,\dot{q}} = (\Delta_{q,1}, \ldots, \Delta_{q,s}, \Delta_{\dot{q},1}, \ldots, \Delta_{\dot{q},s})^T = (\Delta_q, \Delta_{\dot{q}})^T$. Eine binomische Erweiterung mit $(I + hJ \otimes A)$ liefert

$$\delta_0 = -(hJ \otimes b^T)\underbrace{(I - h^2J^2 \otimes A^2)^{-1}}_{=(I+h^2\omega^2 I \otimes A^2)^{-1}}(I + hJ \otimes A)\Delta_{q,\dot{q}} - \Delta_0,$$

was ein Auflösen des Kroneckerprodukts mit Hilfe der Jacobimatrix aus (4.6) erlaubt. Wir erhalten

$$\delta_0 = \begin{pmatrix} hb^T(I + h^2\omega^2A^2)^{-1}(h\omega^2A\Delta_q - \Delta_{\dot{q}}) \\ -h\omega^2b^T(I + h^2\omega^2A^2)^{-1}(-\Delta_q - hA\Delta_{\dot{q}}) \end{pmatrix} - \Delta_0.$$

Diesen Ausdruck wollen wir für den Grenzfall $h \to 0$ unter der Nebenbedingung $h\omega \gg 1$, $\epsilon = 1/\omega$ untersuchen. Die Defekte Δ_q und $\Delta_{\dot{q}}$ besitzen als Größenordnung die Stufenordnung η, und für die zu invertierende Matrix setzen wir

$$(I + h^2\omega^2A^2)^{-1} = \mathcal{O}(\epsilon^2h^{-2}).$$

Wir erhalten daraus eine Konsistenzordnung von

$$\begin{aligned} \hat{e}_q &:= \hat{q}_1 - \varphi(t_1) = \mathcal{O}(\epsilon^2h^\eta) + \mathcal{O}(h^{\eta+1}), \\ \hat{e}_{\dot{q}} &:= \dot{\hat{q}}_1 - \dot{\varphi}(t_1) = \mathcal{O}(h^\eta), \end{aligned}$$

wobei die klassische Ordnung des Terms $-\Delta_0$ als vernachlässigbar gegenüber der Stufenordnung η angenommen wird. Die Ordnungsreduktion, die durch die Verwandtschaft zu DAEs bereits nachgewiesen wurde, tritt also auch bei dieser direkten Analyse der steifen mechanischen Systeme auf.

Im Spezialfall von steifgenauen Verfahren lässt sich (4.9) vereinfachen zu

$$\delta_0 = -(e_s^T \otimes I)(I - A \otimes hJ)^{-1}\Delta_{0,Y}.$$

Analog zu dem allgemeinen Fall folgt daraus die Ordnung

$$e_q := q_1 - \varphi(t_1) = \mathcal{O}(\epsilon^2 h^{\eta-1}),$$
$$e_{\dot{q}} := \dot{q}_1 - \dot{\varphi}(t_1) = \mathcal{O}(\epsilon^2 h^{\eta-1}) + \mathcal{O}(h^\eta),$$

welche in den Lagegrößen ein völlig unterschiedliches Verhalten an den Tag legt als im allgemeinen Fall.

Welche Auswirkungen hat dieser Zusammenhang für die numerische Integration? In beiden Fällen, also sowohl für steifgenaue Verfahren als auch im allgemeinen Fall, ist eine deutliche Ordnungsreduktion festzustellen. Während im allgemeinen Fall die Terme, die ϵ enthalten, jeweils vernachlässigbar im Vergleich zu reinen $\mathcal{O}(h^k)$-Termen sind, trifft dies bei steifgenauen Verfahren nicht zu. Dort ist zwar auch eine Ordnungsreduktion in den Lagekoordinaten zu beobachten, die entsprechenden Terme enthalten aber zusätzlich den Faktor ϵ^2, der den Fehler nach unten skaliert.

Dies hat zur Folge, dass bei steifen mechanischen Systemen der Fehler in den Lagekomponenten kleiner und in den Geschwindigkeitskomponenten größer ist als der von nichtsteifen Systemen, vgl. Abb. 4.1. Aber auch die Schrittweitensteuerung ist davon betroffen. Geht man davon aus, dass das Originalverfahren steifgenau ist und das eingebettete Verfahren nicht, was z.B. für Lobatto IIIC und Radau IIA der Fall ist, ergibt sich für den Fehlerschätzer $\Delta\mu = (\Delta\mu_q, \Delta\mu_{\dot{q}})^T$ die Beziehung

$$\Delta\mu_q := \hat{q}_{n+1} - q_{n+1} = \mathcal{O}(\epsilon^2 h^{\eta-1}) + \mathcal{O}(h^{\eta+1}),$$
$$\Delta\mu_{\dot{q}} := \hat{\dot{q}}_{n+1} - \dot{q}_{n+1} = \mathcal{O}(\epsilon^2 h^{\eta-1}) + \mathcal{O}(h^\eta).$$

Als Schätzer für den lokalen Fehler von steifgenauen Verfahren ist er damit ungeeignet, da sich die Ordnung des nichtsteifgenauen eingebetteten Verfahrens auf den Schätzer $\Delta\mu$ überträgt. Im Gegensatz dazu liefert der modifizierte Schätzer

$$\Delta\nu_q = \mathcal{O}(\epsilon^2 h^{\eta-1}),$$
$$\Delta\nu_{\dot{q}} = \mathcal{O}(\epsilon^2 h^{\eta-2}) + \mathcal{O}(h^\eta),$$

was genau der Ordnung der steifgenauen Verfahren entspricht. Dieser Schätzer ist somit nicht nur aufgrund seiner Beschränktheit für $h\lambda \to \infty$, sondern auch wegen obigen Überlegungen eine gute Wahl.

Auffinden steifer Komponenten und h-Skalierung Obige Untersuchung der Fehlerschätzer $\Delta\mu$ und $\Delta\nu$ lässt noch weitere Schlussfolgerungen zu. Durch die bei steifen Komponenten reduzierte Ordnung in den Geschwindigkeiten ist der Fehler dieser Komponenten deutlich größer als der nichtsteifer Komponenten. Die Schrittweitensteuerung reagiert darauf mit einer Verkleinerung der Schrittweite.

Lobatto IIIC	Lagek. (steif)	Geschw. (steif)	nichtsteif
$\Delta\mu_i$	$O(h^{\eta+1})$	$O(h^{\eta})$	$O(h^{p})$
$\Delta\nu_i$	$O(\epsilon_i^2 h^{\eta-1})$	$O(h^{\eta})$	$O(h^{p})$
ζ_i	$2\log\epsilon_i - 2\log h$	0	0

Tabelle 4.4: Vergleich der Fehlerschätzer $\Delta\mu$ und $\Delta\nu$.

Sind wir an der glatten Lösung interessiert, was oft der Fall ist, ist diese Verkleinerung allerdings unerwünscht, da die kleinen Schrittweiten nicht notwendig sind, um die sichtbaren nichtsteifen Lösungsanteile gemäß der vorgegebenen Toleranz aufzulösen.

Wie man anhand des Spektralradius sehen kann, führt eine Verkleinerung der Schrittweite zudem dazu, dass die numerische Dämpfung reduziert wird. Damit werden hochoszillatorische Störungen weniger gedämpft und die Wahrscheinlichkeit für weitere Schrittweitenverkleinerungen steigt.

Um diesen Effekt zu verhindern, ist es sinnvoll, den geschätzten Fehler von im Vorhinein bekannten steifen Komponenten zusätzlich mit der Schrittweite h zu skalieren, die sogenannte *h-Skalierung*, siehe [41, 20]. Sie bewirkt, dass der Einfluss dieser Komponenten auf die Schrittweitensteuerung reduziert wird und die nichtsteifen Komponenten die neue Schrittweite festlegen.

Mit Hilfe obiger Fehlerschätzer $\Delta\mu$ und $\Delta\nu$ ist es möglich, aufgrund eines Vergleichs steife Komponenten zu detektieren. Der Hintergrund dazu liegt im unterschiedlichen Verhalten beider Fehlerschätzer für steife mechanische Systeme. Zur Analyse definieren wir die neue Größe $\zeta = (\zeta_1, \ldots, \zeta_{2n})^T = (\zeta_q, \zeta_{\dot{q}})^T$ als

$$\zeta_i := \log(\Delta\nu_i / \Delta\mu_i),$$

siehe Tab. 4.4.

Während $\zeta_{q,i} \approx 0$ mit $\zeta_q = (\zeta_{q,1}, \ldots, \zeta_{q,n})^T$ für nichtsteife Komponenten ist, gilt im steifen Fall $\zeta_{q,i} \approx 2\log\epsilon_i - 2\log h$, welches für $\epsilon \ll h$ viel kleiner als Null ist. Deswegen nennen wir eine Komponente *steif*, falls $\zeta_{q,i} < \vartheta < 0$ mit einem negativen Schwellwert ϑ ist.

Auf Komponenten, die auf diese Weise als steif befundenen werden, kann wie oben beschrieben die h-Skalierung angewendet werden. Dazu verwenden wir als neuen Schätzer

$$\widetilde{\Delta\nu}_{\dot{q},i} := h \cdot \Delta\nu_{\dot{q},i}$$

für $i \in \mathcal{S}$ mit der Indexmenge $\mathcal{S}$ der steifen Komponenten. Dies hat zur Folge, dass der Schrittweitenverlauf von den weniger steifen Komponenten festgelegt wird.

Bemerkung 4.2

a) *Eine Untersuchung von ζ_q während einer Integration macht klar, dass sich die Werte von einem Integrationsschritt zum nächsten stark unterscheiden können, aber im Durchschnitt die notwendige Information extrahiert werden kann. Aus diesem Grund vermeiden wir die Berechnung von ζ_q in jedem Schritt, sondern definieren steife und nichtsteife Komponenten während der ersten n_ζ Schritte. Dabei berechnet der Zähler $\texttt{count}_i$ die Anzahl der Schritte mit $\zeta_{q,i} < \vartheta$ für jede Komponente und skaliert die zugehörige Geschwindigkeitskomponente, falls gilt $\texttt{count}_i > n_\zeta/3$.*

   Der Wert von ϑ legt die Schranke zwischen steifen und nichtsteifen Komponenten fest. Obwohl es so aussieht, dass dies ein sehr wichtiger Faktor ist, reagiert der Algorithmus vergleichsweise stabil auf kleine Änderungen, und nur mittelmäßig steife Komponenten sind davon betroffen. Für einen Erfolg des Algorithmus spielen die Anfangsschrittweite h_0 und die Größe n_ζ eine bedeutendere Rolle. Für die meisten Mehrkörpersysteme sind Werte $-1 < \vartheta < -0.4$ sinnvoll, was $\epsilon = 10^{-0.5}h, \ldots, 10^{-0.2}h$ entspricht.

b) *Oft formen die Lösungskomponenten mechanischer Systeme Gruppierungen, wobei jedes Mitglied einer Gruppierung ähnliche Eigenschaften aufweist. In diesem Fall könnte durch die scharfe Grenze ϑ eine Trennung dieser Komponenten entstehen, so dass nur einige Mitglieder skaliert werden, obwohl die Steifheit sehr nahe beieinander liegt. Um dies zu verhindern, führen wir eine Funktion ein, die solche Gruppen lokalisiert und die Entscheidung, ob skaliert wird oder nicht, für die Gruppe gemeinsam trifft. Diese Vorgehensweise verhindert außerdem, dass alle Komponenten als steif erkannt werden.*

   Anstatt von ϑ werden neue Parameter ϑ_g und $\Delta\vartheta$ eingeführt. Dabei definiert $\Delta\vartheta$ einen minimalen Abstand zwischen Gruppierungen, so dass geringere Abstände von zwei Komponenten sie als zu einer Gruppe zugehörig ausweisen. Der Mittelpunkt der Lücken zwischen zwei Gruppen entscheidet über die Skalierung, wobei wieder ein Schwellwert ϑ_g verwendet wird.

c) *Im nichtlinearen Fall hat Lubich [29] RK-Verfahren angewandt auf steife mechanische Systeme (1.18) studiert. Aus seinen Ergebnissen folgt, dass die Ordnungsreduktion dort prinzipiell die gleiche Struktur besitzt wie für die lineare Testgleichung (4.6). Aus diesem Grund erwarten wir, dass unser Ansatz auch im nichtlinearen Fall gute Ergebnisse liefert.*

d) *Implementiert wird dieser Algorithmus sowohl in RADAU5 der Autoren Hairer/Wanner als auch in SPARK3 von Jay. In beiden Codes ist der Programmier- und Rechenaufwand aufgrund der bereits vorhandenen Fehlerschätzer gering.*

e) Aufgrund der Tatsache, dass Komponenten entweder ganz oder überhaupt nicht skaliert werden, ist die Entscheidung, ob eine Komponente steif oder nichtsteif ist, nicht invariant gegenüber einer Koordinatentransformation. Wird eine solche durchgeführt, verteilen sich die steifen Komponenten evtl. auf alle Komponenten des neuen Koordinatensystems. Eine Skalierung aller Komponenten erscheint aber weniger erstrebenswert, da wir nur dort skalieren wollen, wo auch Ordnungsreduktion auftritt.

Bei einer Finite-Element-Diskretisierung bietet sich dieser Algorithmus speziell für ein modales System an, in Verbindung mit starren Komponenten kann auch eine Skalierung aller Knotenvariablen sinnvoll sein. Da die Charakterisierung in steife und nichtsteife Komponenten abhängig vom numerischen Verhalten getroffen wird, werden sie nur dann skaliert, wenn sie numerische Probleme bereiten, was auch das Ziel der Untersuchung ist.

4.2 Newton-Konvergenz

In diesem Abschnitt beschäftigen wir uns mit dem zu lösenden Gleichungssystem (2.2a), welches bei IRK-Verfahren entsteht. Wie zuvor auch konzentrieren wir uns auf steife mechanische Systeme der Form (1.18), wobei ein Vergleich unterschiedlicher Formulierungen der Bewegungsgleichungen vorgenommen wird. Als Lösungsverfahren für das nichtlineare Gleichungssystem beschränken wir uns auf das vereinfachte Newton-Verfahren, da dieses aufgrund der teuren Berechnung der Jacobi-Matrix sinnvoll ist und auch in den später verwendeten Implementierungen enthalten ist. Zunächst legen wir die Euklidischen Norm zugrunde.

Die Formulierungen, die im Folgenden betrachten werden, sind:

a) das IRK-Verfahren direkt angewandt auf

$$M(q)\ddot{q} = f_n(q, \dot{q}) - \frac{1}{\epsilon^2}\nabla U(q) \tag{4.10}$$

als System 2. Ordnung,

b) das IRK-Verfahren angewandt auf

$$\dot{q} = v, \tag{4.11a}$$

$$M(q)\dot{v} = f_n(q, v) - \frac{1}{\epsilon^2}\nabla U(q) \tag{4.11b}$$

als System 1. Ordnung und

c) die von Lubich [29] für das steife Pendel vorgeschlagene und anschließend erweiterte Formulierung

$$\dot{q} = v, \tag{4.12a}$$
$$M(q)\dot{v} = f_n(q,v) - G^T(q)\lambda, \tag{4.12b}$$
$$\epsilon^2\lambda = g(q), \tag{4.12c}$$

im Folgenden $\epsilon^2\lambda$-Formulierung genannt, welche für $\nabla U(q) = G^T(q)g(q)$ äquivalent zu den anderen beiden Formulierungen ist.

Bemerkung 4.3 *Im Gegensatz zu den ersten beiden Varianten (4.10) und (4.11) stellt (4.12) keine ODE, sondern eine DAE mit Index 1, bzw. im Grenzfall $\epsilon = 0$ mit Index 3 dar. Steht nur ein Integrator für ODE's zur Verfügung, kann statt der Nebenbedingung (4.12c) die differenzierte Nebenbedingung*

$$\epsilon^2\dot{\lambda} = G(q)v$$

verwendet werden.

Ziel ist es, Aussagen über die Konvergenz des vereinfachten Newton-Verfahrens zu gewinnen, wozu die Kontraktivität der zugehörigen Fixpunktiteration und die Kondition der Iterationsmatrix abhängig vom Parameter ϵ untersucht wird.

Weil uns bei dieser Analyse nur der Einfluss von ϵ und h interessiert und die Koeffizientenmatrix A unabhängig davon ist, können wir o.B.d.A. den impliziten Euler als Integrationsverfahren heranziehen.

Die Formulierung (4.10) wurde bereits in [29] genauer untersucht. Dort bezeichnet die Funktion $\mathcal{F}(\ddot{Q})$ das zu lösende nichtlineare Gleichungssystem

$$\mathcal{F}(\ddot{Q}) = M(Q)\ddot{Q} - f(Q,\dot{Q}) + \frac{1}{\epsilon^2}\nabla U(Q),$$

wobei die Iterationsvorschrift des impliziten Eulers angewendet auf ein System 2. Ordnung

$$Q = q_0 + h\dot{q}_0 + h^2\ddot{Q} \quad \text{und} \quad \dot{Q} = \dot{q}_0 + h\ddot{Q}$$

eingesetzt wird. Danach ergibt sich mit der approximierten Jacobi-Matrix

$$\mathcal{J}(q_0) = M(q_0) + \frac{h^2}{\epsilon^2}\nabla^2 U(q_0) + \mathcal{O}(h)$$

von $\mathcal{F}(\ddot{Q})$ an der Stelle q_0 die Fixpunktiteration

$$\ddot{Q}^{(k+1)} = \phi(\ddot{Q}^{(k)}) \quad \text{mit} \quad \phi(\ddot{Q}) = \ddot{Q} - \mathcal{J}(q_0)^{-1}\mathcal{F}(\ddot{Q})$$

für das vereinfachte Newton-Verfahren. Um Konvergenz zu gewährleisten, benötigen wir nach dem Banachschen Fixpunktsatz

$$\|\phi'(\ddot{Q})\| < 1,$$

was wegen

$$\phi'(\ddot{Q}) = \mathcal{J}(q_0)^{-1}\left[\mathcal{J}(q_0) - \mathcal{J}(\ddot{Q})\right] \tag{4.13}$$

sowie $\|\mathcal{J}(q_0)^{-1}\| = \mathcal{O}(1)$ und $\nabla^2 U(Q) - \nabla^2 U(q_0) = \mathcal{O}(h)$ auf die Schrittweitenbeschränkung $h \leq c\,\epsilon^{2/3}$ führt, falls das Potential U nicht quadratisch ist. Zur Darstellung der glatten Lösung ist aber $h \gg \epsilon$ ausreichend, so dass sich diese Bedingung sehr nachteilig auf die Integration auswirken kann.

Neben der Einschränkung durch die Konvergenz der Fixpunktgleichung kommt noch eine schlechte Kondition der Jacobi-Matrix $\mathcal{J}(q_0)$ hinzu, da sie für $\epsilon \to 0$ unbeschränkt ist. Dies führt zu zusätzlichen Problemen bei der Lösung des linearen Gleichungssystems.

Im Folgenden wollen wir diese Ergebnisse mit denen der anderen Formulierungen vergleichen. Wünschenswert wäre eine Formulierung, in der die Fixpunktiteration ohne Einschränkung an die Schrittweite h durch den Steifheitsparameter ϵ konvergiert und die Iterationsmatrix eine niedrige Konditionszahl besitzt.

Betrachten wir das System 1. Ordnung (4.11) mit der Diskretisierung

$$\begin{aligned} Q &= q_0 + hV \\ M(Q)V &= M(Q)v_0 + h\left[f_n(Q,V) - \frac{1}{\epsilon^2}\nabla U(Q)\right]. \end{aligned}$$

Als Iterationsmatrix $\mathcal{J}$ erhalten wir daraus

$$\mathcal{J}(q_0,v_0) = \begin{pmatrix} I & -hI \\ \epsilon^{-2}h\nabla^2 U(q_0) - hf_{n,q}(q_0,v_0) + \mathcal{O}(1) & M(q_0) - hf_{n,v}(q_0,v_0) \end{pmatrix}$$

mit $f_{n,q}$, $f_{n,v}$ als partiellen Ableitungen von f_n nach q bzw. v, welche für $\epsilon \to 0$ ebenfalls eine unbeschränkte Konditionszahl besitzt. Zur Untersuchung der Konvergenz schreiben wir $\phi'(Q,V)$ analog zu (4.13) als

$$\phi'(Q,V) = \mathcal{J}(q_0,v_0)^{-1}\left[\mathcal{J}(q_0,v_0) - \mathcal{J}(Q,V)\right].$$

Dies liefert wegen $\|\mathcal{J}(q_0,v_0)^{-1}\| = \mathcal{O}(1)$ und

$$\epsilon^{-2}h\left[\nabla^2 U(q_0) - \nabla^2 U(Q)\right] = \mathcal{O}(\epsilon^{-2}h^2)$$

die Bedingung $h \leq c\,\epsilon$, welche sogar noch strenger ist als die entsprechende Bedingung für ein System 2. Ordnung. Mit der Formulierung (4.11) haben wir im

Vergleich zu (4.10) demnach nichts erreicht; die Konvergenz der Fixpunktiteration wird sogar noch schlechter.

Verwenden wir stattdessen die $\epsilon^2\lambda$-Formulierung (4.12), führen wir zusätzliche Lagrange-Multiplikatoren λ ein und der steife Anteil $\nabla U(q)/\epsilon^2$ wird aufgesplittet. Als Iterationsmatrix erhalten wir für diese Formulierung

$$\mathcal{J}(q_0, v_0, \lambda_0) = \begin{pmatrix} I & -hI & 0 \\ -hf_{n,q}(q_0, v_0) - h\,\partial/\partial q\,(G^T(q_0))\lambda_0 & M(q_0) - hf_{n,v}(q_0, v_0) & G^T(q_0) \\ G(q_0) & 0 & -\epsilon^2 I \end{pmatrix},$$

welche bis auf den ϵ^2-Term mit der Iterationsmatrix für DAEs vom Index 3 übereinstimmt.

In [17] wurden Verfahren für differential-algebraische Gleichungen auf ihre Konvergenzeigenschaften untersucht. Für Systeme vom Index 3 ergibt sich

$$\phi'(Q, V, \Lambda) = \begin{pmatrix} \mathcal{O}(h) & \mathcal{O}(h^2) & \mathcal{O}(h^3) \\ \mathcal{O}(1) & \mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1/h) & \mathcal{O}(1) & \mathcal{O}(h) \end{pmatrix},$$

und legen wir die Norm

$$||y||_h := ||q||_2 + h||v||_2 + h^2||\lambda||_2$$

für $y = (q, v, \lambda)^T$ zugrunde, die wir in Zukunft *h-Norm* nennen wollen, erhalten wir $||\phi'||_h = \mathcal{O}(h)$ ohne Einschränkung durch ϵ. In der Euklidischen Norm liegt wegen des Terms $\mathcal{O}(1/h)$ Konvergenz der Fixpunktiteration nur unter weiteren Einschränkungen an die Anfangswerte vor. Da ein solches System auch im Grenzfall $\epsilon \to 0$ für die $\epsilon^2\lambda$-Formulierung (4.12) vorliegt, können wir diese Ergebnisse für $\epsilon \ll h$ darauf übertragen.

Wie sieht die Konvergenz in der h-Norm für die ersten beiden Formulierungen aus? In der Formulierung zweiter Ordnung (4.10) fällt sie mit der Euklidischen Norm zusammen, weswegen wir dieselben Ergebnisse erhalten. Für (4.11) als System erster Ordnung ergibt sich dagegen eine Verbesserung von $\mathcal{O}(\omega^2 h^2)$ auf $\mathcal{O}(\omega^2 h^3)$, welches somit auf die etwas schwächere und zur ersten Formulierung äquivalenten Bedingung $h \leq c\,\epsilon^{2/3}$ führt.

Diesen Vorteil der h-Norm gegenüber der Euklidischen Norm wollen wir ausnutzen, um die Anzahl der Konvergenztestfehler während der Integration zu reduzieren. Dazu wählen wir die h-Norm als Norm in den Abbruchkriterien des vereinfachten Newton-Verfahren und evtl. iterativer Gleichungslöser. Wie bei der Schrittweitensteuerung auch skalieren wir nur solche Komponenten mit h, die

vorher als steif erkannt worden sind. Die verwendete gemischte h-Norm lautet daher

$$||y||_{h,2} := ||q||_2 + ||v_n||_2 + h||v_s||_2 + h^2||\lambda||_2,$$

wobei v_n die nichtsteifen und v_s die steifen Geschwindigkeitskomponenten bezeichnet.

Die gerade eingeführte Norm liefert einige Vorteile bei der Integration von steifen mechanischen Systemen. Wie in Kap. 4.1.3 anhand der dazu äquivalenten h-Skalierung verdeutlicht, ermöglicht sie einerseits, die Ordnungsreduktion der Geschwindigkeitskomponenten auszugleichen, andererseits erlaubt sie auch bei der Untersuchung von Fixpunktiterationen bessere Konvergenzaussagen.

4.3 Anwendungen der BL-Verfahren

Kommen wir auf die BL-Verfahren, die wir in Kap. 3 untersucht haben, und ihre Anwendungsmöglichkeiten auf steife mechanische Systeme zurück. Der Vorteil dieser Verfahrensklasse liegt darin, dass man die einzelnen Verfahren beliebig kombinieren kann. Wir möchten diese Flexibilität hauptsächlich dadurch nutzen, dass wir gezielt Verfahren mit bestimmten Eigenschaften auf einzelne Komponenten anwenden.

Als geeignetes Kriterium zur Auswahl der Verfahren hat sich der Energiefehler herauskristallisiert. Bei der Anwendung eines einzelnen Verfahrens auf ein ganzes System ist numerische Dämpfung notwendig, sobald ein hinreichend steifes System vorliegt. In diesem Fall kann aber keine Rücksicht mehr auf den Energiefehler genommen werden.

Allgemeine Ansätze zur Energie- und Impulserhaltung werden z.B. in [3, 4] dargestellt oder zur Erhaltung der Passivität in [31]. Bei steifen Problemen jedoch beinhalten diese Verfahren Stabilitätsprobleme aufgrund fehlender numerischer Dämpfung.

Wendet man stattdessen unterschiedliche Verfahren auf steife und nichtsteife Komponenten an, erlaubt uns dies, den Energiefehler der nichtsteifen Komponenten niedrig zu halten. Um diese Aussage genauer zu spezifizieren, unterscheiden wir im Folgenden zwischen linearen und nichtlinearen Systemen. Der Grund liegt darin, dass die Energie im Allgemeinen eine nichtlineare Invariante ist. RK-Verfahren können aber nur im linearen Fall energieerhaltend sein, Beweis siehe [46].

4.3.1 L-stabile konvex kombinierte Lobatto-Verfahren

Beschäftigen wir uns zunächst mit dem linearen Fall. Wie gerade angedeutet stellt die Energie dann eine Invariante dar, die quadratisch von den Zustandsgrößen abhängt. Solche Invarianten werden von symplektischen Verfahren erhalten, also Verfahren ohne numerische Dämpfung, siehe [46].

Im Fall eines linearen steifen mechanischen Systems benötigen wir im Gegensatz dazu L-stabile Verfahren, um eine stabile Integration zu gewährleisten. Da wir mit den IIIAC-Verfahren L- und A-Stabilität verknüpfen können, wählen wir Lobatto IIIC für die steifen Komponenten und Lobatto IIIA sonst. Als Unterscheidung zwischen steifen und nichtsteifen Komponenten steht uns aus vorheriger Analyse des lokalen Fehlers bereits ein Algorithmus zur Verfügung.

Welche Aussagen können wir über dieses gemischte System treffen? Wie in Unterkapitel 3.3.2 gesehen lässt die allgemeine Form nur schwer eine Analyse zu. Für lineare Systeme kann man allerdings über die Stabilitätsfunktion $\mathcal{R}(h\lambda)$ die Energieabnahme bestimmen. Wir betrachten zuerst den nichtskalaren Fall mit einheitlicher Koeffizientenmatrix und anschließend mit unterschiedlichen Koeffizientenmatrizen.

Wir schreiben die Hamiltonfunktion als quadratische Invariante der Form

$$H(y) = \frac{1}{2} y^T S y$$

mit $y = (p, q)^T$ und

$$S = \begin{pmatrix} M^{-1} & 0 \\ 0 & K \end{pmatrix},$$

wobei M die Massenmatrix und K die Steifigkeitsmatrix bezeichnet. Das dazugehörige Hamilton-System lautet

$$y' = Jy \quad \text{mit} \quad J = \begin{pmatrix} 0 & K \\ M^{-1} & 0 \end{pmatrix}.$$

Dann gilt für einen Schritt eines IRK-Verfahrens die Vorschrift

$$y_1 = \mathcal{R}(hJ) y_0$$

mit $\mathcal{R} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ gegeben durch

$$\mathcal{R}(hJ) = I + (b^T \otimes hJ)(I - A \otimes hJ)^{-1}(\mathbb{1} \otimes I). \tag{4.14}$$

Man kann zeigen, dass die Eigenwerte der Matrix $\mathcal{R}(hJ)$ durch die komplex konjugierten Paare $\mathcal{R}(\pm i\omega_j h)$ gegeben sind, wobei ω_j die Eigenwerte des verallgemeinerten Eigenwertproblems

$$K\tilde{y} = \omega^2 M\tilde{y} \tag{4.15}$$

darstellt.

Der Absolutbetrag der Eigenwerte $\mathcal{R}(\pm i\omega_j h)$ legt das numerische Dämpfungsverhalten für das System fest, was für lineare Systeme gleichbedeutend mit dem Energieverlust ist. Abhängig von $\chi = \omega h$ kann man Letzteren also direkt aus dem Spektralradius $\rho(\chi) = |\mathcal{R}(i\chi)|$ ablesen.

Anders sieht es aus, wenn wir verschiedene Koeffizientenmatrizen innerhalb eines Systems verwenden. Statt (4.14) erhalten wir die Funktion

$$\tilde{\mathcal{R}}(hJ) = I + (b^T \otimes hJ) \left(I - \sum_{i=1}^{N} \left[A^{\theta_i} \otimes hE_i J \right] \right)^{-1} (\mathbb{1} \otimes I)$$

mit $E_m = (e_{ij}^{(m)})$ gleich der Nullmatrix außer $e_{mm}^{(m)} = 1$. Äquivalente Aussagen über die Eigenwerte von $\tilde{\mathcal{R}}(hJ)$ gelten ausschließlich, wenn die Subsysteme, die zu verschiedenen Werten von θ gehören, entkoppelt sind. Ist dies nicht der Fall, wissen wir zwar noch, dass die Eigenwerte von $\tilde{\mathcal{R}}(hJ)$ den Energieverlust bestimmen, diese sind aber nicht mehr im Vorhinein bekannt.

Ausgehend von dieser Analyse ist es möglich, durch die bereits bei der h-Skalierung durchgeführte Trennung nach steifen und nichtsteifen Komponenten Erstere mit einem numerisch dämpfenden Verfahren, z.B. Lobatto IIIC, und die nichtsteifen Komponenten mit einem energieerhaltenden Verfahren, z.B. Lobatto IIIA, zu integrieren.

Wir können die Teilenergie des so entstandenen nichtsteifen Subsystems mit Hilfe des A-stabilen Verfahrens erhalten, falls es von den steifen Komponenten entkoppelt ist, falls also ein System der Form (3.8) vorliegt, wobei das Hamilton-Subsystem mit dem nichtsteifen Subsystem gleichzusetzen ist. Die zur stabilen Integration notwendige numerische Dämpfung wird nur auf solche Komponenten angewandt, die eine deutliche Ordnungsreduktion während der Integration aufzeigen.

Liegt keine Entkopplung von steifen und nichtsteifen Komponenten vor, wird die Teilenergie des nichtsteifen Subsystems nicht mehr konserviert. Die Aufsplittung der Komponenten und die Anwendung unterschiedlicher Verfahren ist dennoch sinnvoll, da der Energiefehler dadurch deutlich geringer ausfällt, als wenn alle Komponenten numerisch gedämpft würden.

4.3.2 B-stabile konvex kombinierte Lobatto-Verfahren

Ganz ähnlich sieht die Situation für nichtlineare Systeme aus. Wie in Kap. 3 bereits erläutert, gehen wir dann auf eine B-stabile Verfahrensklasse über, und

wegen verschiedener Vorteile im Vergleich zu Lobatto IIIABC verwenden wir Lobatto IIIDC. Die Aufsplittung in steife und nichtsteife Komponenten erfolgt genauso wie im linearen Fall über den Quotienten der beiden Fehlerschätzer $\Delta\mu$ und $\Delta\nu$.

Im Gegensatz zum linearen Fall ist es aber nicht mehr möglich, die Teilenergie eines Subsystems zu erhalten. Aufgrund der Arbeiten von Hairer [16] können wir aber von der Beschränktheit des Energiefehler ausgehen, wenn wir ein symplektisches Verfahren zur Integration heranziehen.

Diese Aussage lässt sich auch wieder auf ein gesplittetes System (3.8) erweitern, falls eine Enkoppelung vorliegt. Für gekoppelte Systeme können wir auf die Ergebnisse aus Abschnitt 3.3 zurückgreifen, dass die symplektische Ordnung σ der Lobatto IIIDC-Verfahren für $s = 3$ Stufen um eins höher ist als die klassische Ordnung, was sich positiv auf den Energiefehler auswirkt.

Als ein Problem dieser Verfahrensklasse könnte sich die fehlende Steifgenauigkeit des Lobatto IIID-Verfahrens herausstellen. Da es aber nur auf die nichtsteifen Komponenten angewendet wird, erwarten wir keine Verschlechterung des Konvergenzverhaltens.

Maximierung der numerischen Dämpfung Speziell die Lobatto IIIDC-Verfahren mit $s = 3$ Stufen lassen noch eine weitere Anwendung zu, nämlich die Maximierung der numerischen Dämpfung. Dazu wird der Spektralradius dieser Verfahrensklasse, der von $\chi = \omega h$ und θ abhängt, nach θ differenziert und die Extremwerte untersucht. Als optimalen Wert von θ erhalten wir

$$\theta(\chi) = \max\left\{0, 1 - \frac{\sqrt{\chi^6 - 12\chi^4 + 2304}}{\sqrt{\chi^6 + 12\chi^4 + 144\chi^2}}\right\}. \tag{4.16}$$

Diese Funktion ist für alle Werte von $\chi \geq 0$ kleiner als $\theta_{\max} \approx 0{,}4126401258$, und für $\chi < 2{,}692016116$ verschwindet sie identisch. Die Spektralradien von Lobatto IIIC und des durch diese Minimierung entstandene Verfahren, im Folgenden Lobatto $\text{IIIDC}_{\min}$ genannt, sind in Abb. 4.3 dargestellt.

Man erkennt, dass sich die beiden Spektralradien vor allem in dem Bereich $\chi > \pi$ unterscheiden. Das Verfahren Lobatto $\text{IIIDC}_{\min}$ dämpft in diesem Bereich deutlich stärker als Lobatto IIIC.

Bemerkung 4.4 *Um Lobatto IIIDC$_{\min}$ anwenden zu können, muss ein Wert für $\chi = \omega h$ angegeben werden. Dies kann sich zum Teil als schwierig herausstellen, da die steifen Frequenzen dazu bekannt sein müssen. Ist dies nicht der Fall, kann man mit Hilfe des Terms $\zeta_{q,i}$ die Frequenz schätzen. Man setzt dazu*

$$\zeta_{q,i} = 2\log\epsilon_i - 2\log h,$$

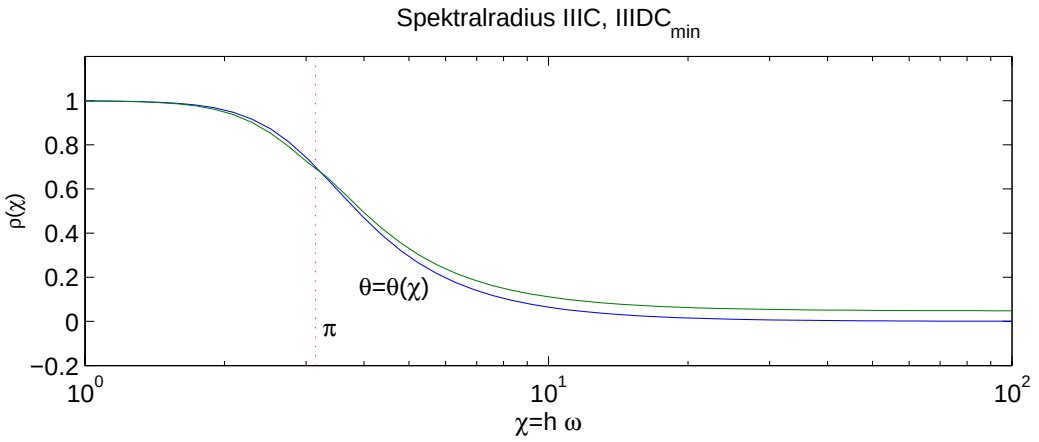


Abbildung 4.3: Vergleich der Spektralradien von Lobatto IIIC und IIIDC$_{\text{min}}$.

und Auflösen nach ϵ_i liefert

$$\omega_i = \frac{1}{\epsilon_i} = \frac{10^{-\zeta_{q,i}/2}}{h},$$

was anstelle der exakten Eigenfrequenz in (4.16) eingesetzt werden kann.

Fassen wir zum Abschluss dieses Abschnitts nochmal die verschiedenen Möglichkeiten zusammen. Die IIIAC-Verfahren besitzen außer der L-Stabilität den Vorteil, dass sie steifgenau sind, und werden daher für lineare Systeme bevorzugt. Durch Separierung der Komponenten eines Systems in steife und nichtsteife Komponenten ist es möglich, Teile des Systems energieerhaltend zu integrieren, während die Schwierigkeiten bereitenden steifen Komponenten trotzdem numerisch gedämpft werden.

Für nichtlineare Systeme bevorzugen wir Lobatto IIIDC oder IIIABC-Verfahren wegen der B-Stabilität. Das Aufsplitten der Komponenten funktioniert in gleicher Weise wie bei linearen Systemen, allerdings können wir nur die Beschränktheit des Energiefehlers gewährleisten, nicht die Energieerhaltung. Zusätzlich erlaubt uns diese Lobatto-Familie die Maximierung der numerischen Dämpfung, wodurch das Verfahren Lobatto IIIDC$_{\text{min}}$ entsteht.

Kapitel 5

Simulationsbeispiele

In diesem Kapitel werden die durchgeführten Simulationen erläutert. Als Beispiele dienen zum Einen kleinere Systeme wie ein lineares Testbeispiel, das steife Pendel und ein Doppelpendel mit Drehfeder, welche aufgrund ihrer überschaubaren Größe eine detaillierte Analyse erlauben. Zum Anderen werden die Techniken auch auf größere Beispiele wie einen Kurbeltrieb, ein Waschmaschinenmodell und die Ladefläche eines LKWs übertragen.

Zur Modellbildung wurde oftmals `MapleV` hinzugezogen, während die Integration mit Hilfe der `FORTRAN`-Codes `RADAU5` und `SPARK3` vorgenommen wurde. `RADAU5` von Hairer/Wanner [20] enthält eine Implementierung des Radau IIA-Verfahrens und verwendet LU-Faktorisierung zur Lösung der linearen Gleichungssysteme, wobei die Iterationsmatrix geeignet in Blöcke zerlegt wird.

`SPARK3` von Jay [25] stellt eine Implementierung der in Abschnitt 3.1.2 eingeführten SPARK-Methoden zur Verfügung. Neben den Lobatto-Verfahren enthält es zusätzlich die Radau IIA- und Gauss-Koeffizienten, so dass diese Verfahren unabhängig von der Implementierung verglichen werden können.

Wie in Kap. 1 bereits erwähnt, bezieht sich Jay auf die Formulierung (1.6), welche die Corioliskräfte auf der linken Seite enthält. Ein weiterer Unterschied in der Implementierung besteht in der Lösung der linearen Gleichungssysteme. Diese werden in `SPARK3` nach Vorkonditionierung und W-Transformation durch iterative Verfahren gelöst. Zur Verfügung stehen das GMRES-Verfahren und das Verfahren von Richardson, wobei wir auf Ersteres zurückgreifen.

In beide vorhandene Codes wurde zusätzlich die Berechnung der Werte $\zeta_{q,i}$ zur Detektierung steifer Komponenten implementiert. Die beiden Fehlerschätzer $\Delta\mu$ und $\Delta\nu$ stehen dazu bereits zur Verfügung. Zur Einführung der h-Skalierung ist in `SPARK3` eine zusätzliche Routine einzubauen, während in `RADAU5` die bereits für differential-algebraische Gleichungen verwendete Skalierung bei der Schritt-

weitensteuerung erweitert werden kann. Die zusätzliche Verwendung der h-Norm im Newton-Verfahren und bei iterativen Gleichungslösern bereitet keine großen Probleme.

Zur Verbesserung des Fehlerschätzers wurden die Koeffizienten der in Kap. 4 eingeführten eingebetteten Verfahren und die zugehörigen Erweiterungen zur Berechnung der zusätzlichen Stufe eingefügt. Bei den Simulationen werden in `SPARK3` falls nicht anders angegeben bei linearen Systemen die L-stabilen BL-Verfahren mit dem eingebetteten Verfahren (IIIAC)$\hat{}$ und bei nichtlinearen Systemen die IIIDC-Verfahren mit (IIIDC)$\hat{}$ angewendet.

5.1 Detaillierte Analyse

Die Funktionsweise der h-Skalierung wird zunächst an einem linearen Beispiel untersucht. Anschließend folgt eine Analyse des steifen Pendels, wobei vor allem das Konvergenzverhalten des Newton-Verfahrens von Bedeutung ist. Ein weiteres nichtlineares Beispiel mit zwei Freiheitsgraden verdeutlicht die Unterschiede von Lagrange- und Hamilton-Formulierung bezüglich des Energiefehlers.

Zur Integration von DAEs unterscheiden wir zwischen der Schreibweise in Standardform

$$\dot{q} = v + G^T(q)\,\mu, \tag{5.1a}$$
$$M(q)\,\dot{v} = f_n(q,v) - \frac{1}{\epsilon^2}\nabla U(q) - G^T(q)\,\lambda, \tag{5.1b}$$
$$0 = G(q)\,v, \tag{5.1c}$$
$$0 = g(q) \tag{5.1d}$$

und der in semiexpliziten Form

$$\dot{q} = v + G^T(q)\,\mu, \tag{5.2a}$$
$$\dot{v} = a, \tag{5.2b}$$
$$0 = M(q)\,a - f_n(q,v) + \frac{1}{\epsilon^2}\nabla U(q) + G^T(q)\,\lambda, \tag{5.2c}$$
$$0 = G(q)\,v, \tag{5.2d}$$
$$0 = g(q) \tag{5.2e}$$

mit der zusätzlichen Beschleunigungskoordinate a, wobei in beiden Formulierungen zusätzlich GGL-Stabilisierung nach Gear/Gupta/Leimkuhler [13] für die Indexreduktion zur Anwendung kommt.

Skal.	NSTEPS	NACCPT	NREJCT
-	266	215	51
v_1	146	137	8

Tabelle 5.1: Integrationsstatistik der linearen DAE in Standardform.

5.1.1 Lineare Gleichungen

Anhand von einer linearen DAE wollen wir die Wirkung der h-Skalierung in beiden Formulierungen (5.1) und (5.2) mit `RADAU5` untersuchen. Aufgrund der Linearität wird die gemischte h-Norm nur bei der Schrittweitensteuerung eingesetzt. Wir wählen $n_q = 3$ und $n_\lambda = 1$ sowie

$$g(q) = q_2 - q_3, \qquad f_n(q, v) = -Kq - u,$$
$$\nabla U(q) = (q_1 - \cos \Omega t, 0, 0)^T, \qquad \epsilon = \frac{1}{\omega_1}$$

mit

$$K = \begin{pmatrix} \omega_2^2 & -\omega_2^2 & 0 \\ -\omega_2^2 & \omega_2^2 + \omega_3^2 & -\omega_3^2 \\ 0 & -\omega_3^2 & \omega_3^2 \end{pmatrix}, \quad u = \begin{pmatrix} \Omega^2 + \omega_2^2 \\ 4\Omega^2 - \omega_2^2 \\ 0 \end{pmatrix} \cos \Omega t,$$

und da die glatte Lösung $q_1 = \cos \Omega t$, $q_2 = q_3 = 2 \cos \Omega t$ bekannt ist, können wir die Anfangswerte auf der glatten Lösung mit

$$y = (q; v; \lambda, \mu)^T = (1, 2, 2; 0, 0, 0; 0, 0)^T$$

direkt angeben.

Die Abb. 5.1 zeigt den Schrittweitenverlauf sowie den relativen Fehler der Lage- und Geschwindigkeitskomponenten für die Werte $(\omega_i) = (10^3, 1, 1)^T$ und $\Omega = 20$ bei Integration der Standardform (5.1) mit Toleranzen $Atol = Rtol = 10^{-5}$. Dabei wird zwischen der Simulation ohne und mit h-Skalierung unterschieden. Man erkennt, dass der Schrittweitenverlauf mit h-Skalierung deutlich glatter ist als der ohne h-Skalierung, was auf einen verringerten Einfluss des Fehlers in der steifen Geschwindigkeitskomponente v_1 zurückzuführen ist, zu sehen jeweils als oberste Kurve im Fehlerdiagramm.

Die dazugehörige Integrationsstatistik ist in Tab. 5.1 zusammengefasst. Die Anzahl der verworfenen Schritte `NREJCT` wird durch h-Skalierung deutlich reduziert und die Gesamtschrittzahl `NSTEPS` verringert sich ca. um die Hälfte. `NACCPT` bezeichnet die Anzahl der akzeptierten Schritte.

Abb. 5.2 enthält die Ergebnisse aus der Integration der semiexpliziten Form mit den gleichen Werten wie zuvor. Auffällig ist, dass die h-Skalierung zwar eine

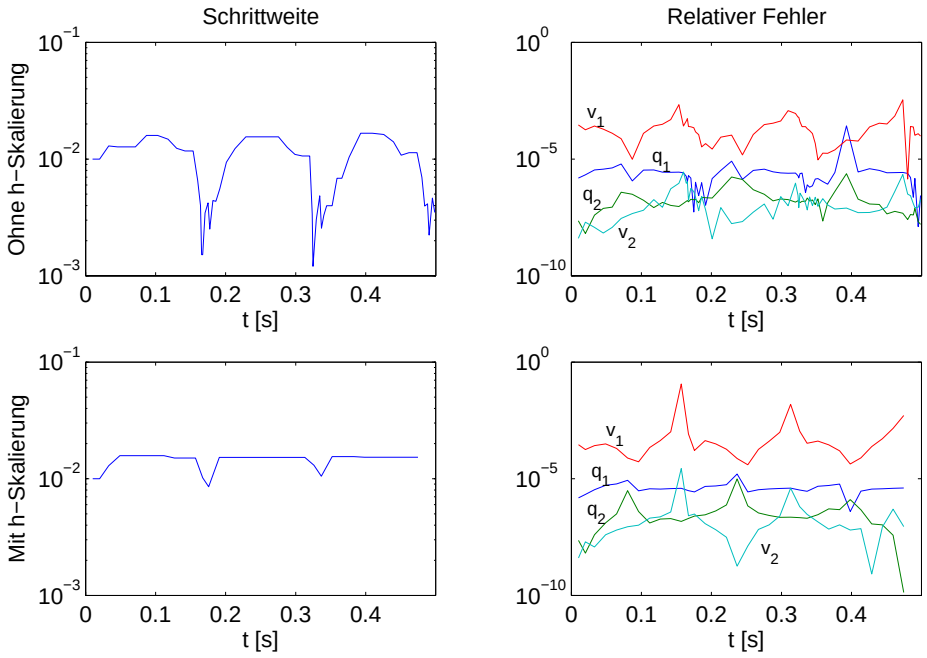


Abbildung 5.1: Schrittweitenverlauf und relativer Fehler ohne und mit h-Skalierung für lineare DAE in Standardform mit steifen Komponenten q_1, v_1 und nichtsteifen Komponenten $q_2 = q_3$, $v_2 = v_3$.

Glättung des Schrittweitenverlaufs zur Folge hat, aber immer noch starke Oszillationen zu sehen sind. Bei einer Betrachtung des relativen Fehlers fällt auf, dass der Fehler in der steifen Beschleunigungskomponente a ebenfalls sehr groß ist. Erst eine zusätzliche Skalierung der Beschleunigungskomponenten bringt daher den erhofften Erfolg, siehe Tab. 5.2.

Obige Simulationen haben somit gezeigt, dass h-Skalierung ein effektives Mittel zur Glättung des Schrittweitenverlaufs darstellt. Der Fehler in den Geschwindigkeitskomponenten wird dadurch zwar weitestgehend vernachlässigt, aber der Vergleich der Schrittanzahl zeigt, dass dies zugunsten einer schnelleren Integration zugelassen werden sollte. Bei der semiexpliziten Formulierung ist außerdem eine Skalierung der Beschleunigungskomponenten notwendig, um ähnliche Ergebnisse wie bei der Standardformulierung zu erzielen.

5.1.2 Steifes Pendel

Ein weit verbreitetes Beispiel zur Untersuchung steifer mechanischer Systeme ist das steife Pendel, siehe [29], welches wir in Abb. 1.2 auf Seite 11 bereits in verschiedenen Formulierungen kennengelernt haben. Wir konzentrieren uns bei

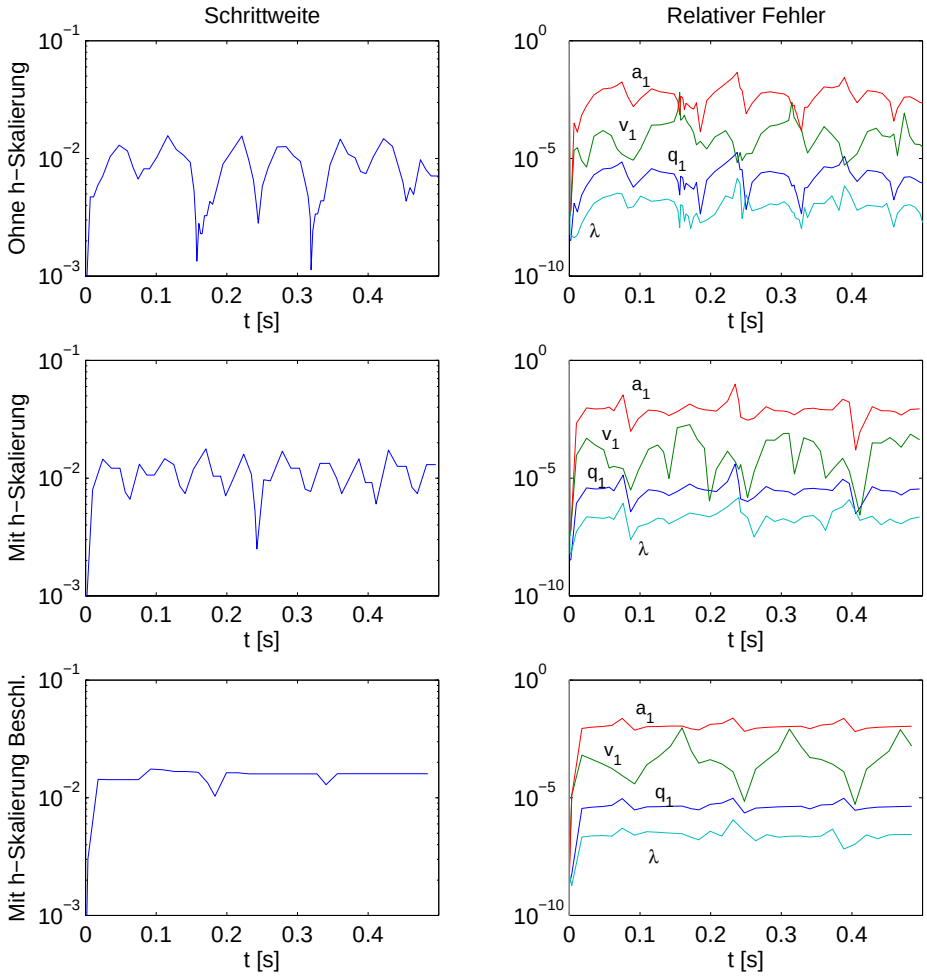


Abbildung 5.2: Schrittweitenverlauf und relativer Fehler der steifen Komponenten ohne und mit h-Skalierung sowie mit zusätzlicher Skalierung der Beschleunigungskomponenten a für lineare DAE in semiexpliziter Form.

Skal.	NSTEPS	NACCPT	NREJCT
-	342	284	57
v_1	227	193	33
v_1,a_1	142	135	2

Tabelle 5.2: Integrationsstatistik der linearen DAE in semiexpliziter Form.

Verf.	$\epsilon = 10^{-2}$				$\epsilon = 10^{-4}$			
	NS	NA	NR	NC	NS	NA	NR	NC
IIIC	1062	1056	6	4	5767	4174	1593	62
IIID	956	937	19	20	96591	65929	30662	29542
RIIA	501	325	176	116	8827	4268	4559	420
$\text{IIIDC}_{\text{min}}$	1064	1058	6	2	5705	4097	1608	92

Tabelle 5.3: Integrationsstatistik für steifes Pendel in Standardformulierung.

der numerischen Simulation auf die regularisierte Bewegungsgleichungen

$$\ddot{q}_1 = -\frac{1}{\epsilon^2}\frac{q_1}{\sqrt{q_1^2+q_2^2}}\left(\sqrt{q_1^2+q_2^2}-1\right), \tag{5.3a}$$

$$\ddot{q}_2 = -\frac{1}{\epsilon^2}\frac{q_2}{\sqrt{q_1^2+q_2^2}}\left(\sqrt{q_1^2+q_2^2}-1\right)-1 \tag{5.3b}$$

und die zugehörige $\epsilon^2\lambda$-Formulierung

$$\ddot{q}_1 = -\frac{q_1}{\sqrt{q_1^2+q_2^2}}\lambda, \tag{5.4a}$$

$$\ddot{q}_2 = -\frac{q_2}{\sqrt{q_1^2+q_2^2}}\lambda - 1, \tag{5.4b}$$

$$\epsilon^2\lambda = \sqrt{q_1^2+q_2^2}-1. \tag{5.4c}$$

Verwendet werden sowohl der Integrator `RADAU5` als auch `SPARK3`. Hauptaugenmerk dieser Analyse ist der Vergleich der Konvergenztestfehler im Newton-Verfahren, allerdings kommt auch das die numerische Dämpfung maximierende Verfahren Lobatto $\text{IIIDC}_{\text{min}}$ zum Zuge.

Die Ergebnisse, die mit dem Code `SPARK3` und der Formulierung (5.3) als System 1. Ordnung sowie $Atol = Rtol = 10^{-6}$ erzielt wurden, sind in Tab. 5.3 zusammengefasst. Dabei bezeichnet `NS` die Anzahl der Schritte, `NA` die Anzahl akzeptierter Schritte, `NR` die Anzahl verworfener Schritte und `NC` die Anzahl der Konvergenztestfehler; das Radau IIA-Verfahren innerhalb von `SPARK3` wird mit RIIA abgekürzt.

Für $\epsilon = 10^{-2}$, also nur geringer Steifigkeit, dominiert das Radau IIA-Verfahren aufgrund seiner höheren Ordnung, die drei untersuchten Lobatto Verfahren IIIC, IIID und $\text{IIIDC}_{\text{min}}$ zeigen ein relativ ähnliches Verhalten. Die Anzahl der Konvergenztestfehler ist allerdings bei Radau IIA am größten.

Diese Abfolge ändert sich für $\epsilon = 10^{-4}$, da die numerische Dämpfung in diesem Fall von herausragender Bedeutung ist. Das führt dazu, dass die Rechenzeit bei

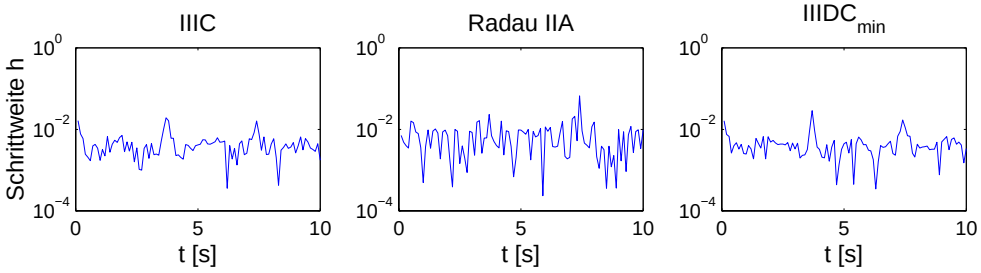


Abbildung 5.3: Schrittweitenverlauf für steifes Pendel in Standardformulierung.

Verf.	Skal.	$\epsilon = 10^{-2}$				$\epsilon = 10^{-3}$			
		NS	NA	NR	NC	NS	NA	NR	NC
IIIC	-	1735	1497	238	1	25311	22209	3102	1
RIIA	-	1652	1344	308	7	21106	18539	2567	66
IIIC	λ	1223	1155	68	0	7024	6789	235	0
RIIA	λ	724	595	129	0	174	138	36	2

Tabelle 5.4: Integrationsstatistik für steifes Pendel in $\epsilon^2\lambda$-Formulierung mit differenzierter Zwangsbedingung mit und ohne Skalierung von λ.

Lobatto IIIDC$_{\text{min}}$ am kürzesten ist, während bei Radau IIA sehr viele verworfene Schritte hinzukommen. Lobatto IIIC kann noch sehr gut mithalten, während Lobatto IIID aufgrund fehlender numerischer Dämpfung deutlich mehr Schritte benötigt.

Die Schrittweitenverläufe mit $\epsilon = 10^{-4}$ sind für Lobatto IIIC, IIIDC$_{\text{min}}$ und Radau IIA in Abb. 5.3 dargestellt. Die größere Anzahl der verworfenen Schritte führt bei Letzterem zu starken Oszillationen, bei den beiden Lobatto-Verfahren fallen diese deutlich geringer aus.

In Tab. 5.4 ist die Integrationsstatistik für die $\epsilon^2\lambda$-Formulierung mit differenzierter Zwangsbedingung aufgeführt. Schon bei $\epsilon = 10^{-2}$ benötigen sowohl Lobatto IIIC als auch Radau IIA deutlich mehr Schritte als zuvor, das Lobatto IIIDC$_{\text{min}}$-Verfahren fällt mit Lobatto IIIC zusammen. Der Grund für dieses Verhalten liegt darin, dass die numerisch bestimmte Zwangskraft λ stark oszilliert. Dies lässt sich auf das im Grenzfall vorliegende Index 2-Problem zurückführen, bei dem die algebraischen Variablen λ mit geringerer Ordnung, genauer Ordnung s für Radau IIA und $s-1$ für Lobatto IIIC, integriert werden. Ein Skalieren dieser Variablen führt aus diesem Grund auf deutlich bessere Resultate, was auch der Tab. 5.4 zu entnehmen ist.

Diese Ergebnisse mit dem Integrator `SPARK3` sind überraschend, da bei der Stan-

Verf.	$\epsilon = 10^{-2}$				$\epsilon = 10^{-3}$			
	NS	NA	NR	NC	NS	NA	NR	NC
Std.	292	220	2	70	1251	642	0	609
$\epsilon^2\lambda$	263	214	9	40	257	204	0	53

Tabelle 5.5: Integrationsstatistik für steifes Pendel, `RADAU5`.

dardformulierung kaum Newton-Konvergenzprobleme auftauchen und durch ausreichende numerische Dämpfung ein glatter Schrittweitenverlauf erzielt wird. Zudem benötigt der Integrator bei der $\epsilon^2\lambda$-Formulierung, welche nur mit differenzierter Zwansbedingung integriert werden kann, deutlich mehr Schritte als bei der Standardformulierung, was auf die zugrundeliegende Index-2-Formulierung zurückgeführt werden kann. Nur zusätzliche Skalierung der Lagrange-Multiplikatoren schafft Abhilfe.

Im Folgenden betrachten wir dieses Beispiel im Integrator `RADAU5`, der im Gegensatz zu `SPARK3` die direkte Anwendung der $\epsilon^2\lambda$-Formulierung erlaubt. Die Tab. 5.5 zeigt die Integrationsstatistik für $\epsilon = 10^{-2}$ und $\epsilon = 10^{-3}$ sowie $Atol = Rtol = 10^{-5}$ in beiden Formulierungen. Man erkennt deutlich, dass bei der Standardformulierung und $\epsilon = 10^{-3}$ sehr viele Konvergenztestfehler auftreten. Durch die Transformation auf die $\epsilon^2\lambda$-Formulierung erzielt man dagegen wesentlich bessere Resultate, auch die Ordnungsreduktion in λ macht sich nicht bemerkbar. Dieses Verhalten stimmt also mit den Überlegungen aus Kap. 4.2 überein.

Bei diesem Beispiel stellt sich heraus, dass sich die beiden Integratoren `SPARK3` und `RADAU5` trotz vieler Gemeinsamkeiten im Verhalten unterscheiden. Während Ersterer entgegen den Erwartungen kaum Konvergenzprobleme aufweist, dafür aber mit der $\epsilon^2\lambda$-Formulierung Probleme hat, entsprechen die Ergebnisse mit `RADAU5` der Theorie.

Das Lobatto IIIDC$_{\text{min}}$-Verfahren erzielt etwas bessere Resultate als Lobatto IIIC alleine. Weil zur Anwendung die vorherige Kenntnis der Steifigkeiten notwendig ist, bietet es sich besonders bei einer modalen Analyse an.

5.1.3 Doppelpendel mit Drehfeder

Um die Unterschiede zwischen Lagrange- und Hamilton-Formulierung in Bezug auf den Energiefehler zu untersuchen, betrachten wir ein starres Doppelpendel mit einer Drehfeder am oberen Gelenk. Die Bewegungsgleichungen lauten als Lagrange-System

$$M(q)\ddot{q} = f(q,\dot{q}) \quad \text{mit} \quad M(q) = \begin{pmatrix} J & -\frac{1}{2}ml^2\sin(q_2-q_1) \\ -\frac{1}{2}ml^2\sin(q_2-q_1) & J+ml^2 \end{pmatrix}$$

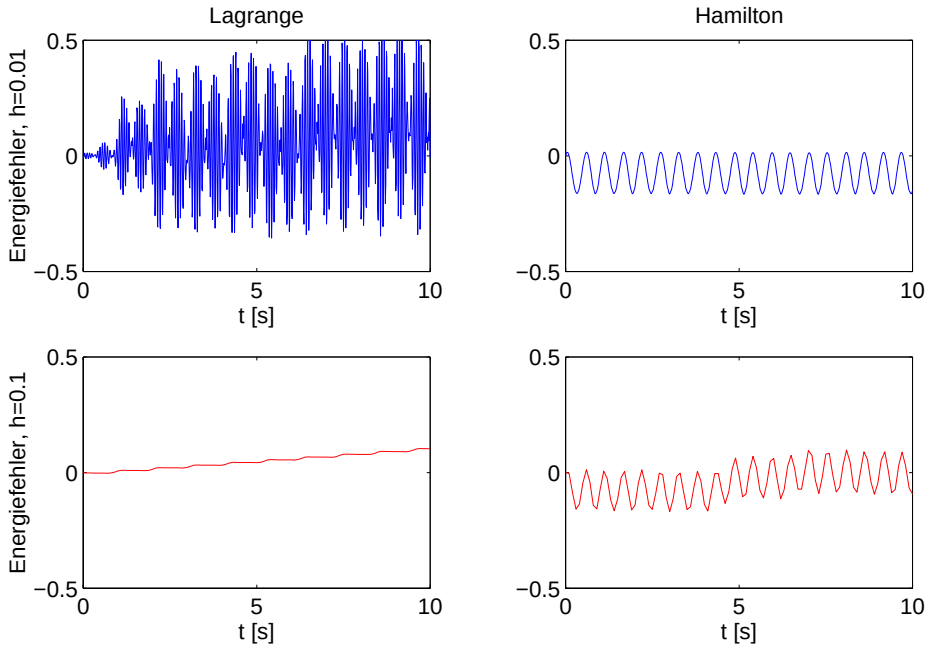


Abbildung 5.4: Absoluter Energiefehler des Doppelpendels mit Drehfeder für Lagrange- und Hamilton-Formulierung und unterschiedliche Schrittweiten.

und

$$f(q,\dot{q}) = \begin{pmatrix} -\frac{3}{2}ml\gamma\sin(q_1) - kq_1 + \frac{1}{2}ml^2\dot{q}_1\dot{q}_2\cos(q_1-q_2) \\ -\frac{1}{2}ml\gamma\sin(q_2) - \frac{1}{2}ml^2\dot{q}_1\dot{q}_2\cos(q_1-q_2) \end{pmatrix}.$$

Aufgrund der Nichtlinearität der Massenmatrix $M(q)$ ist zur Darstellung als Hamilton-System eine nichtlineare Transformation $p = M^{-1}(q)\dot{q}$ notwendig. Diese wurde mit `Maple V` durchgeführt und das Ergebnis mit Hilfe der eingebauten Funktion `fortran` in Fortran 77-Notation überführt. Die Komplexität der Gleichungen nimmt durch diese Transformation deutlich zu.

Die Abb. 5.4 zeigt den absoluten Energiefehler $||H(p(t_n),q(t_n)) - H(p_n,q_n)||$ für die Lagrange- und die Hamilton-Formulierung für unterschiedliche Schrittweiten h. Dabei wurde $k = 1/\epsilon^2 = 10^4$, $\gamma = 9.81\,\mathrm{m/s}$, $l = 1\,\mathrm{m}$ und $m = 1\,\mathrm{kg}$ gewählt und das Lobatto IIIC-Verfahren auf die steife Komponente q_1 und Lobatto IIID auf q_2 angewandt. Die konstante Schrittweite ist notwendig, da die Ergebnisse der Rückwärtsanalyse auf diesen Fall eingeschränkt sind.

Sowohl die Lagrange- als auch die Hamilton-Formulierung zeigen bei $h = 0.01$ im Lösungsverlauf zusätzliche Oszillationen, welche auf die fehlende numerische Dämpfung von Lobatto IIID zurückzuführen sind. Die Auswirkungen auf den Energiefehler sind aber völlig verschieden. Während sich die Störungen in der Lösung beim Lagrange-System im Energiefehler widerspiegeln, was an den star-

ken Oszillationen zu sehen ist, ist dies beim Hamilton-System nicht der Fall. Dort macht sich noch nicht einmal die teilweise Anwendung von Lobatto IIIC bemerkbar, sondern nur die gestörte Hamilton-Funktion ist sichtbar.

Bei großen Schrittweiten $h = 0.1$ verschwinden die Oszillationen in der Lösung, wodurch der Energiefehler der Lagrange-Formulierung einen glatten Verlauf erhält. Im Hamilton-System wird dabei die Abweichung zur gestörten Hamilton-Funktion sichtbar.

Bemerkung 5.1 *Bei dem Versuch, dieses Beispiel vollständig mit Lobatto IIID anzugehen, lässt der Integrator* `SPARK3` *keine konstanten Schrittweiten zu, sondern verringert sie, um Konvergenzprobleme zu vermeiden.*

Die Entwicklung des Energiefehlers macht deutlich, dass beide Formulierungen ihre Vor- und Nachteile besitzen. Während die Lagrange-Formulierung leichter aufzustellen ist, erhält man im Hamilton-System auch mit Störungen im Lösungsverlauf einen glatten und beschränkten Energiefehler. Bei großen Schrittweiten bzw. nicht vorhandenen oszillatorischen Störungen erzielt aber auch die Lagrange-Formulierung einen zufriedenstellenden Energiefehlerverlauf.

Bei nichtkonstanter Massenmatrix ist daher aufgrund des hohen Aufwands die Transformation auf Hamilton-Formulierung nur eingeschränkt sinnvoll, zumal die Lagrange-Formulierung durch Stabilisierungstechniken wie h-Skalierung unterstützt werden kann. Empfehlenswert ist auf jeden Fall eine Modellierung in absoluten Koordinaten, da dann eine konstante Massenmatrix entsteht, siehe [12]. Dies ermöglicht eine Vereinigung der Vorteile beider Formulierungen.

5.2 Größere Anwendungen

5.2.1 Kurbeltrieb

Der Kurbeltrieb aus Abb. 5.5 wurde bereits in sehr vielen unterschiedlichen Formulierungen für die elastische Pleuelstange untersucht, siehe z.B. [43, 42, 40]. Wir wählen zwei Eigenformen in vertikaler Richtung und einen quadratischen Ansatz in longitudinaler Richtung zur Diskretisierung des elastischen Balkens analog zu [42, S. 75]. Die Zeitintegration erfolgt in der semiexpliziten Form (5.2) mit `RADAU5`. Die Frequenzen

$$\omega_1 = 203.2\,\mathrm{Hz}, \quad \omega_2 = 812.8\,\mathrm{Hz}, \quad \omega_3 = 4216.7\,\mathrm{Hz}, \quad \omega_4 = 15171.3\,\mathrm{Hz}$$

reflektieren das Verhalten eines elastischen Balkens, da die Frequenzen der Längsschwingungen $\omega_{3/4}$ deutlich größer sind als die der Querschwingungen $\omega_{1/2}$. Ein

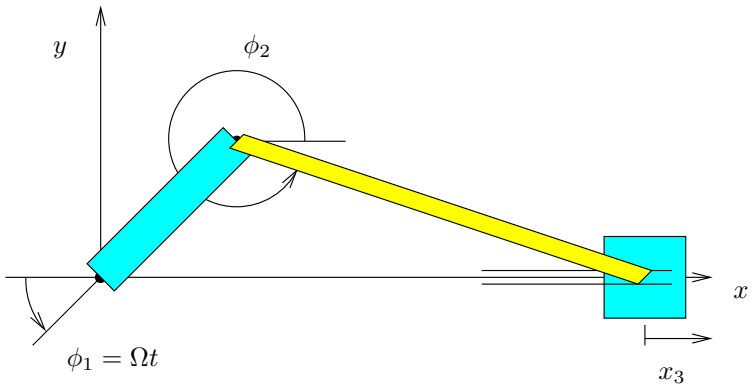


Abbildung 5.5: Kurbeltrieb mit Starrkörperkoordinaten $q_1 = \phi_1$, $q_2 = \phi_2$, $q_3 = x_3$ und elastischen Koordinaten $q_{4/5}$ in Quer- und $q_{6/7}$ in Längsrichtung.

Scal.	NSTEPS	NACCPT	NREJCT
-	420	338	79
q_7	233	194	38
$q_{6/7}$	126	105	20

Tabelle 5.6: Integrationsstatistik des Kurbeltriebs mit glatten Anfangswerten.

Skalieren der Komponenten q_6 and q_7 der Längsschwingungen mit den Frequenzen ω_3 und ω_4 erscheint daher sinnvoll. Die Frage ist nun, ob der Algorithmus zum Detektieren steifer Komponenten zu demselben Ergebnis führt.

Die Antwort ist Ja, was anhand der Abb. 5.6 beobachtet werden kann. Zur Integration wurden gemäß Bem. 4.2 b) die Parameter $\vartheta_g = -0.2$ als Grenzwert, $\Delta\vartheta_g = 0.2$ als Abstand, $n_\zeta = 7$ Schritte und Anfangsschrittweite $h = 0.1$ gewählt. Verglichen werden die Berechnungen ohne und mit h-Skalierung sowie die Version ohne Gruppierung der Komponenten, welches auf die alleinige Skalierung von q_7 führt. Wie beim linearen Beispiel auch lässt sich durch h-Skalierung der Gesamtaufwand deutlich reduzieren, siehe Tab. 5.6.

Weitaus größere Einsparungen können bei Integrationen mit gestörten Anfangswerten erzielt werden, siehe Tab. 5.7. Um dies zu erreichen, verwenden wir den

Scal.	NSTEPS	NACCPT	NREJCT
-	6336	4996	1337
$q_{6/7}$	157	130	25

Tabelle 5.7: Integrationstatistik des Kurbeltriebs mit gestörten Anfangswerten.

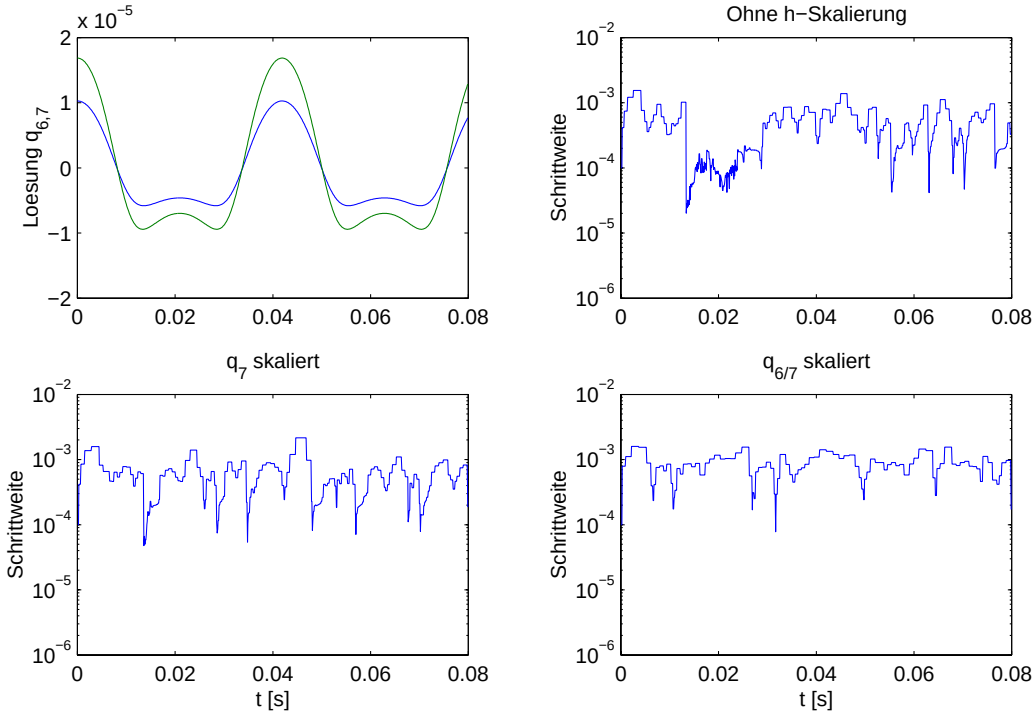


Abbildung 5.6: Lösung der steifen Komponenten und Schrittweitenverlauf für den Kurbeltrieb mit glatten Anfangswerten.

Parameter $n_\zeta = 3$, wodurch wir die Auswirkungen der großen Anfangsschrittweite maximieren können. Dadurch wird die Steifigkeitsgröße ζ_q nur während der anfänglichen Reduzierung der Schrittweite berechnet und nicht bei der später verwendeten sehr kleinen Schrittweite, was die Wahrscheinlichkeit für das Finden steifer Komponenten deutlich erhöht.

Abb. 5.7 zeigt die Schrittweiten und jeweils einen Ausschnitt der Lösung bei der Integration mit gestörten Anfangswerten. Man erkennt, dass die gestörten Anfangswerte ohne h-Skalierung zu kleinen Oszillationen im Lösungsverlauf führen, die eine sehr kleine Schrittweite zur Folge haben. Mit h-Skalierung gelingt es, diese als Störungen zu erkennen und entsprechend zu dämpfen.

Bemerkung 5.2

1. *Dieses Beispiel wurde zusätzlich wie die lineare DAE mit skalierten Beschleunigungen berechnet. Im Fall von glatten Anfangsdaten lassen sich dadurch keine weiteren Verbesserungen feststellen, bei gestörten Anfangsdaten reduziert sich die Anzahl der Schritte auf* 147.

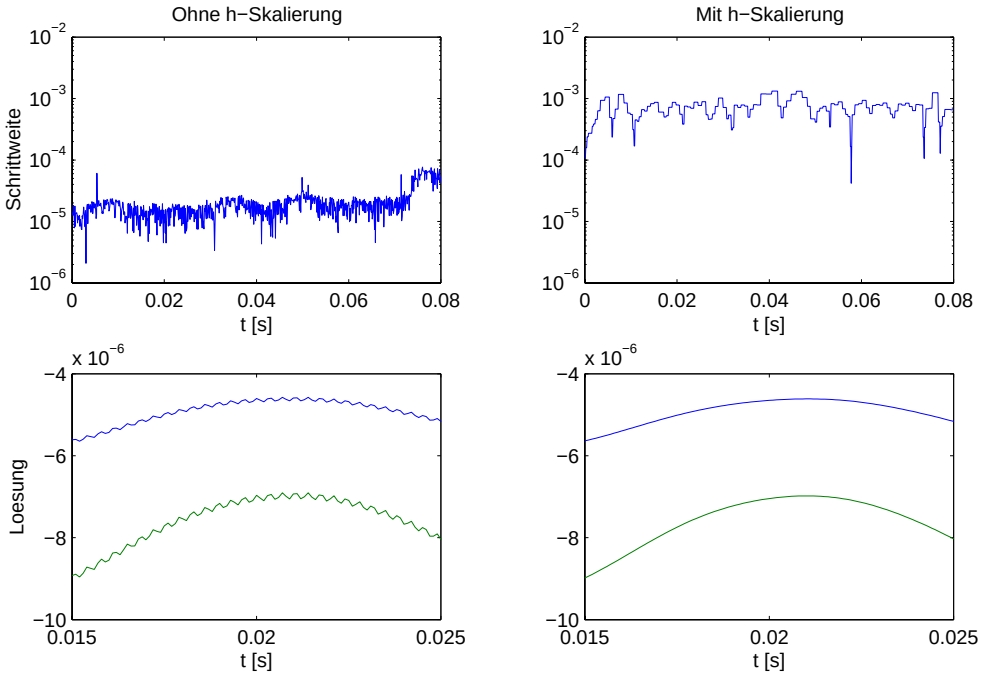


Abbildung 5.7: Schrittweiten (oben) und Ausschnitt der Lösung (unten) des Kurbeltriebs mit gestörten Anfangswerten.

2. *Die Berechnung von glatten Anfangswerten kann bei gekoppelten Systemen aus starren und elastischen Körpern sehr aufwendig sein. Oft ist die statische Gleichgewichtslage zumindest eine hinreichend gute Näherung für Anfangswerte auf der glatten Lösung. Im Fall dieses Kurbeltriebs wurden die Anfangswerte mit Hilfe einer entkoppelten quasistatischen Analyse ermittelt, die durch eine Partitionierung der starren und elastischen Komponenten eine Trajektorie der Starrkörperlösung berechnet, siehe [42, S. 118] für Details.*

3. *Um den Kurbeltrieb auch mit* `SPARK3` *integrieren zu können, wurde die Zwangsbedingung in q_3 vernachlässigt und die weiteren Zwangsbedingungen eliminiert. Das entstehende elastische Doppelpendel wurde in* `SPARK3` *sowohl in der Standard- als auch in der $\epsilon^2\lambda$-Formulierung mit diskretisierter Nebenbedingung integriert, wobei die Komponenten q_6 und q_7 als steif behandelt wurden.*

   Während in der Standardformulierung die Konvergenzprobleme so schwerwiegend waren, dass die Simulation abgebrochen wurde, wurden bei der $\epsilon^2\lambda$-Formulierung mit Radau IIA und Lobatto IIIC einschließlich h-Skalierung eine Gesamtschrittzahl von `NS`$= 670$ *bzw.* `NS`$= 2411$ *erreicht.*

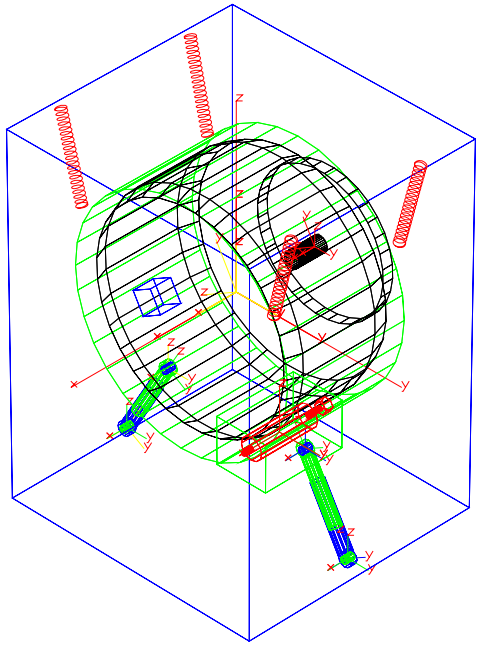


Abbildung 5.8: Modell einer Waschmaschine, erstellt mit SIMPACK.

5.2.2 Waschmaschine

Das nächste Modell, eine Waschmaschine, wurde mit dem Simulationspacket SIMPACK von G. Hippmann erstellt, siehe Abb. 5.8. Ursprünglich zur Dämpferoptimierung gedacht, um ein Anschlagen der Trommel am Gehäuse zu vermeiden, enthält es im Original $n_q = 19$ Freiheitsgrade und $n_\lambda = 6$ Zwangsbedingungen. Zur Integration mit SPARK3 dient eine ODE-Variante, bei der die Zwangsbedingungen regularisiert wurden.

Das Erzeugen von symbolischen Code erlaubt das Extrahieren des Modells aus SIMPACK, indem das Residuum der Gleichungen als Fortran90-Routine zur Verfügung gestellt wird. Diese kann dann in verschiedene Codes integriert werden.

Als Anregung wird die Winkelgeschwindigkeit der Trommel vorgegeben, die in Abb. 5.9 für Lobatto IIIC mit einer typischen Lösungskomponente und dem Schrittweitenverlauf in Euklidischer und gemischter h-Norm dargestellt ist, wobei Toleranzen $Atol = Rtol = 10^{-6}$ gewählt wurden.

Beim Detektieren steifer Komponenten verzichten wir ab diesem Beispiel auf die Gruppierung von Komponenten, da die Abstände in ζ_q zu klein werden. Trotzdem stellen wir fest, dass die jeweils symmetrisch zueinander liegenden Komponenten unter Verwendung der Standardwerte $n_\zeta = 7$, $\vartheta = -0.4$ und Anfangsschrittweite

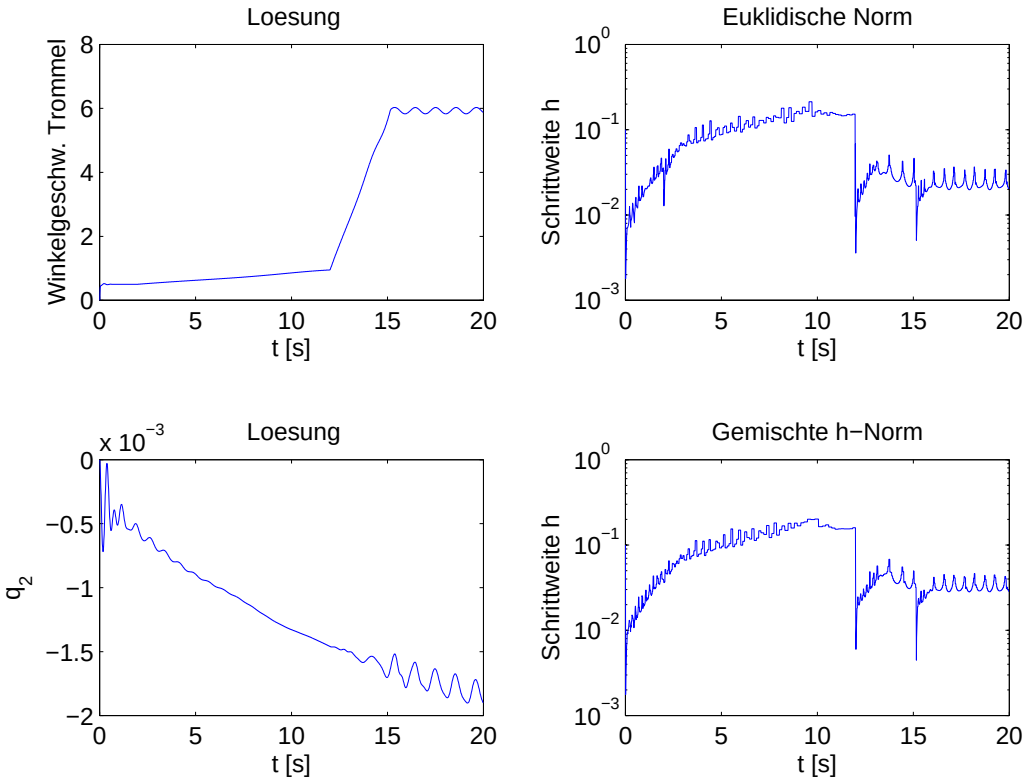


Abbildung 5.9: Anregung, Lösung und Schrittweitenverlauf zur Waschmaschine mit Lobatto IIIC mit Euklidischer und gemischter h-Norm, Fehlerschätzer (IIIAC)$\hat{}$.

$h = 0.1$ gleichartig behandelt werden. Bei dem Fehlerschätzer (IIIDC)$\hat{}$ werden insgesamt 11 Komponenten, bei (IIIAC)$\hat{}$ dagegen nur 10 Komponenten skaliert, wobei sich der Hauptanteil an steifen Komponenten erwartungsgemäß in den beiden Dämpfern befindet. Obwohl am Schrittweitenverlauf kaum ein Unterschied zwischen den Rechnungen in den beiden Normen zu bemerken ist, ist der Gesamtaufwand mit Skalierung deutlich geringer, siehe Tab. 5.8.

Mit den Radau IIA-Koeffizienten berechnet sich die Lösung am effizientesten, obwohl der Algorithmus zum Detektieren steifer Komponenten keine solchen entdeckt. Die Simulation mit Lobatto IIIC liefert teilweise genauere Ergebnisse, was aufgrund der vorgegebenen Toleranzen $Atol = Rtol = 10^{-6}$ auch wünschenswert ist.

Um die Auswirkungen der h-Skalierung bei verschiedenen vorgegebenen Toleranzen zu beobachten, zeigt Abb. 5.10 das Verhältnis von Fehler und Anzahl an Funktionsauswertungen für Rechnungen mit und ohne h-Skalierung für das Lobatto IIIC-Verfahren mit dem Fehlerschätzer (IIIAC)$\hat{}$. Zur besseren Unter-

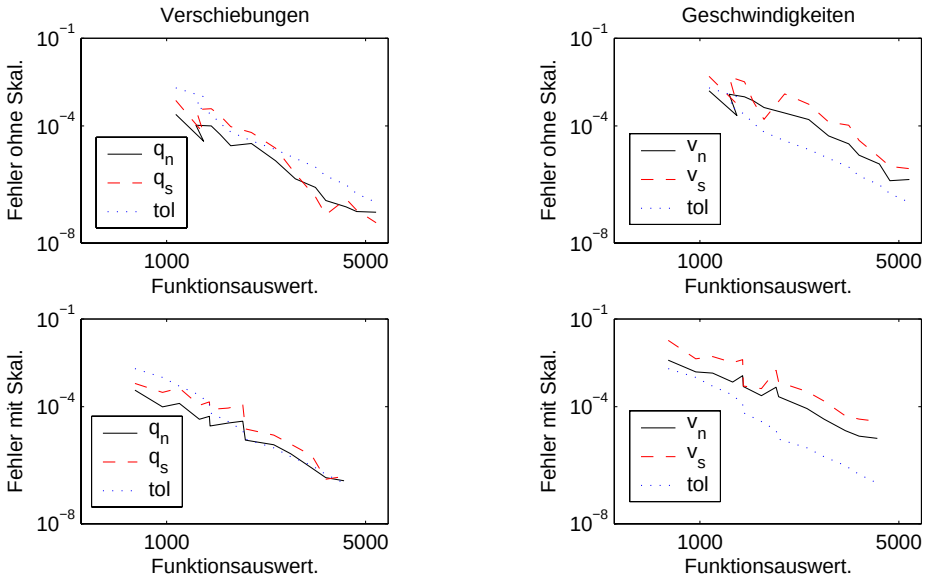


Abbildung 5.10: Absoluter Fehler in Abhängigkeit von der Anzahl der Funktionsauswertungen für Rechnung mit und ohne h-Skalierung aufgesplittet nach steifen und nichtsteifen Komponenten, Lobatto IIIC mit (IIIAC)^.

scheidung sind Lage- und Geschwindigkeits- sowie steife und nichtsteife Komponenten getrennt dargestellt, wobei Letztere jeweils nach den zu den Toleranzen $Atol = Rtol = 10^{-6}$ detektierten Werten gesplittet werden.

Im Vergleich zur vorgegebenen Toleranz schneidet der Fehler der skalierten Version etwas schlechter ab, liegt aber in den Lagekoordinaten noch meistens darunter. Dafür benötigen die Rechnungen ohne h-Skalierung mehr Funktionsauswertungen, so dass sich insgesamt ein ähnliches Bild ergibt. Auffallend ist, dass sich der Fehler in den steifen und nichtsteifen Komponenten nur geringfügig unterscheidet, was darauf hindeutet, dass alle Komponenten steife Anteile enthalten. Da keine Entkopplung vorliegt und die Gleichungen nichtlinear sind, vermischen sich

Verf.	Skal.	NSTEP	NACCPT	NREJCT	NCTF	$\|(y_n - \tilde{y}_n)/\tilde{y}_n\|_2$
IIIC,(IIIDC)^	nein	556	509	47	0	$0.038335 \cdot 10^{-2}$
IIIC,(IIIDC)^	ja	431	394	37	0	$0.098734 \cdot 10^{-2}$
IIIC,(IIIAC)^	nein	346	322	24	36	$0.18160 \cdot 10^{-2}$
IIIC,(IIIAC)^	ja	255	229	26	0	$2.0073 \cdot 10^{-2}$
RIIA	nein	190	149	41	3	$0.92099 \cdot 10^{-2}$

Tabelle 5.8: Integrationsstatistik der Waschmaschine in SPARK3, $\tilde{y}$ Referenzlösung Radau IIA mit Toleranzen $Atol = Rtol = 10^{-12}$.

die Steifheitseffekte.

Bemerkung 5.3 *Bei der Integration mit einer um Faktor* 10 *schnelleren Anregung der Trommel treten schwerwiegende Konvergenzprobleme auf, die durch die Nichtlinearität des steifen Potentialgradienten* $\nabla U(q)$ *verursacht werden. Da die Bewegungsgleichungen nur als automatisch generierter Code vorliegen, ist es nicht möglich, die störenden Steifheitsterme in die* $\epsilon^2\lambda$*-Formulierung zu überführen. Alleinige Transformation der Zwangsbedingungen liefert keine besseren Ergebnisse, auch nicht in Kombination mit der h-Norm, da die Schrittweitenbeschränkung* $h < \epsilon^{2/3}$ *bestehen bleibt.*

5.2.3 Ladefläche

Als letztes und größtes Beispiel betrachten wir die Ladefläche eines LKWs, welche mit Hilfe der `MATLAB PDE-Toolbox` und einer Erweiterung für quadratische Finite Elemente, siehe [42, S. 107], modelliert wird. Das entstehende Finite-Element-Gitter enthält 322 Knoten, die jeweils einen Freiheitsgrad in x- und y-Richtung besitzen, siehe Abb. 5.11. Der Knoten 18 der Ladefläche wird als Auflagepunkt zum Chassis als fest angenommen, und am Knoten 28 wird eine Anregung der Form

$$u(t) = \frac{1}{100}\left(\sum_{l=0}^{10} u_i^c \cos(2\pi lkt) + \sum_{l=1}^{10} u_i^s \sin(2\pi lkt)\right) \cdot \exp\left(-\frac{1}{10}(t-7)^2\right)$$

mit Fourierkoeffizienten u_i^c und u_i^s und einer Konstanten k als vorgegebene Auslenkung in y-Richtung vorgegeben. Zusätzlich wird eine Ladung $G = l^3\varrho$ mit $l \approx 1{,}5\,\mathrm{m}$ und $\varrho = 5000\,\mathrm{kg/m^3}$ zwischen den Knoten 49 und 53 aufgebracht, siehe [44].

Um den Einfluss der Schrittweite auf das Detektieren steifer Komponenten zu untersuchen, transformieren wir das entstehende System auf Eigenformen. Die ersten drei davon sind in Abb. 5.12 dargestellt, wobei die festen Knoten 18 und 28 mit einem $\times$ markiert sind.

Für unterschiedliche Schrittweiten h wird eine Anlaufrechnung bis `NACCPT`$= n_\zeta$ durchgeführt und dabei die durch den Algorithmus detektierten steifen Komponenten notiert. Diese Vorgehensweise wird mit einer um Faktor 100 beschleunigten Anregung $u(100\,t)$ wiederholt. Eine Übersicht über steife (weiß) und nichtsteife Eigenformen (grau) befindet sich in Abb. 5.13.

Bei der Betrachtung dieser Abbildung fällt sofort die Regelmäßigkeit ins Auge, mit der die Anzahl der nichtskalierten Komponenten mit wachsender Schrittan-

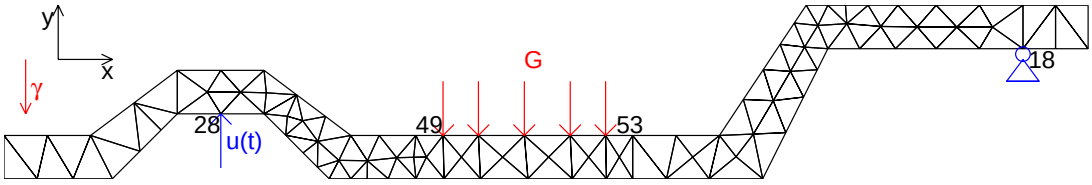


Abbildung 5.11: Modell einer Ladefläche bei LKWs.

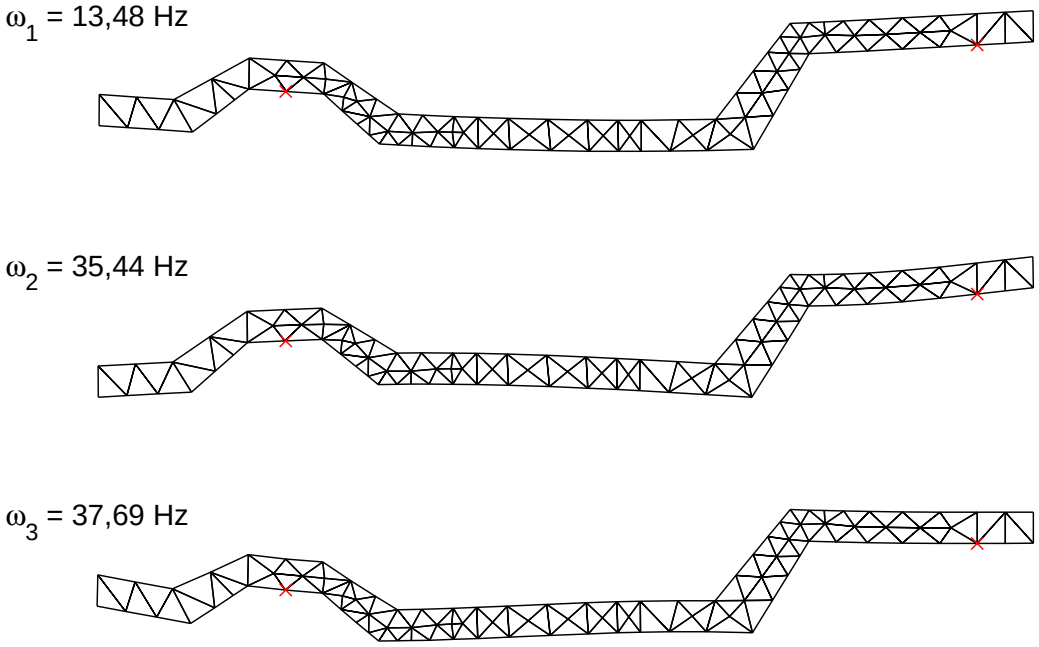


Abbildung 5.12: Eigenwerte der Ladefläche.

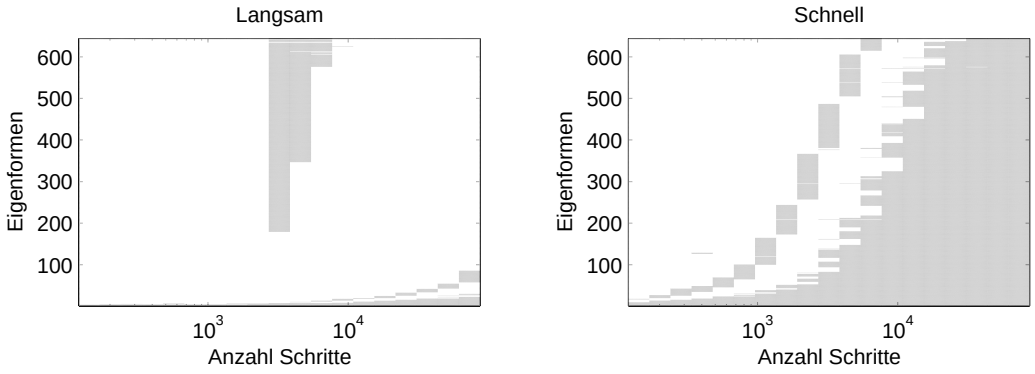


Abbildung 5.13: Skalierte (weiß) und nicht skalierte Eigenformen (grau) bei langsamer und schneller Anregung.

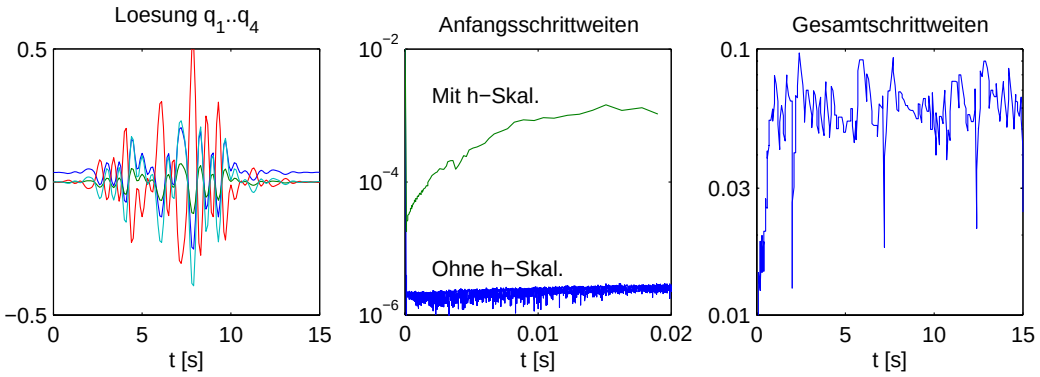


Abbildung 5.14: Lösungskomponenten q_1 bis q_4 und Schrittweitenverlauf der Ladefläche als System aus Eigenformen mit Lobatto IIIC mit und ohne h-Skalierung.

zahl zunimmt. Dieses Verhalten erklärt sich durch die Approximation der Steifigkeitswerte

$$\zeta_{q,i} = \log(\Delta\nu_{q,i}/\Delta\mu_{q,i}) \approx 2\log\epsilon_i - 2\log h,$$

die wir in Kap. 4 hergeleitet haben. Je kleiner die Schrittweite, desto größer wird $\zeta_{q,i}$ für ein festes i und nur Eigenformen mit höheren Frequenzen, also kleinerem ϵ_i, werden skaliert. Die Approximation erweist sich daher als zutreffend. Dass bei einigen Läufen der langsameren Anregung auch höhere Frequenzen teilweise nicht skaliert werden, lässt sich dadurch erklären, dass bei der Berechnung des Logarithmusterms $\log(\Delta\nu_{q,i}/\Delta\mu_{q,i})$ Komponenten mit $\Delta\mu_{q,i} < 10^{-15}$ ausgeschlossen werden, um Division durch Null zu vermeiden.

Der Vergleich zwischen beiden Anregungsgeschwindigkeiten zeigt, dass bei der langsamen Anregung viel mehr Eigenformen als steif erkannt werden als bei der schnellen Version. Der Grund liegt darin, dass die Steifheit einer Komponente nicht alleine von ihrer Frequenz abhängt, sondern im Zusammenhang zwischen eigener und angeregter Frequenz zu sehen ist. Entscheidend für die Steifheit ist das Verhältnis dieser beiden Frequenzen, da schnellere Anregfrequenzen Ω wegen $h \sim 1/\Omega$ kleinere Schrittweiten nach sich ziehen.

Zum Abschluss dieses Kapitels zeigt die Abb. 5.14 einige Lösungskomponenten und den Schrittweitenverlauf der langsamen Anregung mit den steifen Komponenten zur Schrittweite $h = 0.1$ und Toleranzen $Atol = 10^{-12}$ und $Rtol = 10^{-4}$.

Als Anfangsdaten wird die statische Gleichgewichtslage herangezogen. Da diese nicht exakt auf der glatten Lösung liegt, benötigt die Simulation einige Zeit, um mit Hilfe von numerischer Dämpfung und h-Skalierung eine große Schrittweite zu erreichen. Ohne h-Skalierung gelingt dies nicht, wie anhand der Anfangsschrittweiten in Abb. 5.14 zu erkennen ist. Daher wird nur die skalierte Version fortgeführt, deren Schrittweitenverlauf nach der Anfangsphase weitestgehend glatt

ist. Die Schwingungen der Eigenformen stimmen der Form nach mit der Anregungsfunktion $u(t)$ überein, nur die Amplitude variiert.

Bemerkung 5.4

1. *Zur Verstärkung der gemischten h-Norm als Konvergenzkriterium werden die Parameter* `RWORK(4)` $=$ `RWORK(5)` $= 10$ *gesetzt, welche das Verhältnis von den Toleranzen bei Newton- bzw. GMRES/Richardson-Iterationen und Schrittweitensteuerung angeben. Im Gegensatz zur üblichen Vorgehensweise verlangen wir damit eine geringere Genauigkeit bei den Iterationen als beim Fehlerschätzer an sich. Diese Wahl der Parameter alleine ohne h-Norm bringt keine Vorteile.*

   Desweiteren führt die Angabe `IWORK(4)` $= 1$ *auf etwas bessere Resultate, welche als Anfangswerte der Newton-Iteration interpolierte Werte im Gegensatz zu Null-Anfangswerten verwendet. Eine genaue Untersuchung der Kombination aus Schrittweitensteuerung und Kontrolle des Konvergenzfehlers wird in [15] durchgeführt.*

2. *Die Simulation wurde auch mit Lobatto IIIDC*$_{\min}$ *durchgeführt. Das Verhalten ist sehr ähnlich zu Lobatto IIIC und wird daher nicht extra dargestellt.*

Anhand der gezeigten Beispiele kann man die vorher diskutierten Phänomene sehr gut beobachten. Stabilitätseinbußen treten sowohl bei den Testbeispielen als auch bei den größeren Anwendungen auf, können aber mit Hilfe von h-Skalierung in Schrittweitensteuerung und den Abbruchkriterien minimiert werden. Der dazu verwendete Algorithmus zum Detektieren steifer Komponenten erzielt sehr gute Resultate vor allem bei der modalen Analyse der Ladefläche.

Die Anwendung von BL-Verfahren, insbesondere Lobatto IIIC, erweist sich vor allem bei sehr steifen Systemen wie dem vorgeführten Pendel als vorteilhaft gegenüber Radau IIA, da dort die numerische Dämpfung entscheidend ist. Auch das Lobatto IIIDC$_{\min}$-Verfahren bringt Vorteile, allerdings ist es nur zu empfehlen, falls die Eigenfrequenzen des Systems bekannt sind.

Das Doppelpendel mit Drehfeder zeigt, dass die BL-Verfahren sehr gute Eigenschaften in Bezug auf Energieerhaltung besitzen. Während bei kleinen Schrittweiten die Integration als Hamilton-System noch sinnvoll ist, liefert bei großen Schrittweiten das Lobatto IIIC-Verfahren auch in Lagrange-Formulierung vernünftige Ergebnisse. Die Transformation ist daher nur bei kleinen Systemen praktikabel, bei größeren sollte man versuchen, durch globale Koordinaten die numerische Äquivalenz beider Formulierungen zu erreichen.

Zusammenfassung

Beim Aufstellen der Bewegungsgleichungen von Mehrkörpersystemen tauchen häufig große Steifigkeitsterme auf, die bei der numerischen Simulation zu Problemen führen. Sie können z.B. bei vorhandenen steifen Federn oder auch bei einer feinen Ortsdiskretisierung vorliegen. Durch die Verwandtschaft zu differentialalgebraischen Gleichungen muss man bei solchen steifen mechanischen Systemen mit Ordnungsreduktion rechnen, aber auch die Schrittweitensteuerung und das Newton-Konvergenzverhalten werden beeinflusst.

Die Wahl der numerischen Verfahren fällt auf implizite Verfahren, die genügend Stabilität besitzen, um diesen Problemen entgegenzuwirken. Im Besonderen erlaubt die Familie der Lobatto-Verfahren durch Kombination von Verfahren mit sehr verschiedenen Eigenschaften die gezielte Anpassung der Integration an die Erfordernisse eines Systems. So wird die Kombination aus stabilen mit energieerhaltenden Verfahren oder die Maximierung numerischer Dämpfung möglich.

Unterstützt werden die Integrationsverfahren von der Wahl geeigneter eingebetteter Verfahren bzw. Fehlerschätzer und einem neuen Algorithmus zum Detektieren steifer Komponenten. Letzterer liefert einen Ansatzpunkt für die Wahl einer skalierten Norm und gezielte Umformulierungen, die nicht nur eine stabile Schrittweitensteuerung, sondern auch bessere Konvergenz und Kondition des vereinfachten Newton-Verfahrens nach sich ziehen.

Die Beispiele verdeutlichen, dass der Algorithmus die Separierung der Komponenten sinnvoll und in nachvollziehbarer Weise durchführt. Durch die gesonderte Behandlung der nichtsteifen Komponenten und die positiven Eigenschaften der Lobatto-Verfahren wird der Energiefehler reduziert. Zudem zieht die Anwendung des numerisch sehr stark dämpfenden Verfahrens Lobatto IIIC verknüpft mit einer Skalierung der steifen Komponenten eine Glättung des Schrittweitenverlaufs und eine Verringerung der Newton-Konvergenzfehler nach sich, die dem Lösungsverhalten angemessen sind. Der Gesamtaufwand verringert sich dadurch erheblich, vor allem, wenn mit gestörten Anfangswerten gerechnet wird, was anhand der Beispiele eines Kurbeltriebs und der Ladefläche eines LKWs zu erkennen ist.

Literaturverzeichnis

[1] M. Arnold. A perturbation analysis for the dynamical simulation of mechanical multibody systems. *App. Numer. Math.*, 18:37–56, 1995.

[2] N. Biehn, S. Campbell, L. Jay, and T. Westbrook. Some comments on dae theory for irk methods and trajectory optimization. *J. Comput. Appl. Math.*, 120:109–131, 2000.

[3] M. Borri, C. L. Botasso, and L. Trainelli. A novel momentum-preserving energy-decaying algorithm for finite-element multibody procedures. *Computer Assisted Mechanics and Engineering Sciences*, 9:315–340, 2002.

[4] C. L. Botasso, O. A. Bauchau, and J.-Y. Choi. An energy decaying scheme for nonlinear dynamics of shells. *Eingereicht in Computer Methods in Applied Mechanics and Engineering.*

[5] D. Braess. *Finite Elemente, Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.* Springer, 1992.

[6] J. Butcher. Implicit runge-kutta processes. *Math. Comput.*, 18:50–64, 1964.

[7] J.C. Butcher, editor. *The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods.* John Wiley, 1987.

[8] M. P. Calvo and J. M. Sanz-Serna. Canonical b-series. *Numer. Math.*, 67:161–175, 1994.

[9] R. P. K. Chan. On symmetric runge-kutta methods of high order. *Computing*, 45:301–309, 1990.

[10] J. de Swart and G. Söderlind. On the construction of error estimators for implicit runge-kutta methods. *J. Comp. and Appl. Math.*, 86:347–358, 1997.

[11] P. Deuflhard and F. Bornemann. *Numerische Mathematik II. Gewöhnliche Differentialgleichungen.* 2. Auflage. de Gruyter, Berlin, New York, 2002.

[12] E. Eich-Soellner and C. Führer. *Numerical Methods in Multibody Dynamics.* Teubner, Stuttgart, 1998.

[13] C.W Gear, G.K. Gupta, and B.J. Leimkuhler. Automatic integration of the euler-lagrange equations with constraints. *J. Comp. Appl. Math.* , 12&13:77–90, 1985.

[14] M. Günther. Simulating digital circuits numerically - a charge-oriented row-approach. *Numer. Math.*, 79(2):203–212, 1998.

[15] K. Gustafsson and G. Söderlind. Control strategies for the iterative solution of nonlinear equations in ode solvers. *SISC*, 18(1):23–40, 1997.

[16] E. Hairer. Backward analysis of numerical integrators and symplectic methods. *Annals of Numerical Mathematics*, 1:107–132, 1994.

[17] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Springer, 1989.

[18] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*. Springer, Berlin, Heidelberg, New York, 2002.

[19] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations I, Nonstiff Problems*. 2. Auflage. Springer, Berlin, Heidelberg, New York, 1993.

[20] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*. 2. Auflage. Springer, Berlin, Heidelberg, 1996.

[21] T. J. Hughes. *The Finite Element Method*. Prentice Hall, Englewood Cliffs, 1987.

[22] L. Jay. Runge-kutta type methods for index three differential-algebraic equations with applications to hamiltonian systems. Ph.d. thesis, University of Geneva, Department of Mathematics, 1994.

[23] L. Jay. Convergence of runge-kutta methods for differential-algebraic systems of index 3. *Applied Numerical Mathematics*, 17:97–118, 1995.

[24] L. Jay. Structure preservation for constrained dynamics with super partitioned additive runge-kutta methods. *SIAM J. Sci. Comput.*, 20:416–446, 1999.

[25] L. Jay. Inexact simplified newton iterations for implicit runge-kutta methods. *SIAM J. Numer. Anal.*, 38(4):1369–1388, 2000.

[26] L. Jay. Iterative solution of nonlinear equations for spark methods applied to daes. *Numer. Algorithms*, 31:171–191, 2002.

[27] L. Jay. Solution of index 2 implicit differential-algebraic equations by lobatto runge-kutta methods. *BIT*, 43:91–104, 2003.

[28] M. Lesser. *The analysis of complex nonlinear mechanical systems, a computer algebra assisted approach*. World Scientific Publishing, Singapore, 1995.

[29] C. Lubich. Integration of stiff mechanical systems by runge-kutta methods. *ZAMP*, 44:1022–1053, 1993.

[30] R. März and C. Tischendorf. Recent results in solving index-2 differential-algebraic equations in circuit simulation. *SIAM J. Sci. Comput.*, 18(1):139–159, 1997.

[31] K. Ochs. Passive integration methods: Fundamental theory. *Int. J. Electron. Commun.*, 55(3):153–163, 2001.

[32] B. Owren and H.H. Simonsen. Alternative integration methods for problems in structural dynamics. *Comput. Methods Appl. Mech. Engrg.*, 122:1–10, 1995.

[33] W. C. Rheinboldt and B. Simeon. On computing smooth solutions of dae's for elastic multibody systems. *Computers Math. Applic.*, 37:69–83, 1999.

[34] M. Schaub and B. Simeon. Automatic h-scaling for the efficient time integration of stiff mechanical systems. *Multibody System Dynamics*, 8:329–345, 2002.

[35] M. Schaub and B. Simeon. Blended lobatto methods in multibody dynamics. *ZAMM*, 83(10):720–728, 2003.

[36] F. A. Scheck. *Mechanik*. 5. Auflage. Springer, 1996.

[37] W. O. Schiehlen, editor. *Multibody System Handbook*. Springer, Berlin, 1990.

[38] J. Schneid. B-convergence of lobatto iiic formulas. *Numer. Math.*, 51:229–235, 1987.

[39] B. Simeon. Mbspack - numerical integration software for constrained mechanical motion. *Surv. on Math. in Ind.*, 5:169–202, 1995.

[40] B. Simeon. Modelling a flexible slider crank mechanism by a mixed system of daes and pdes. *Math. Modelling of Systems*, 2:1–18, 1996.

[41] B. Simeon. Order reduction of stiff solvers at elastic multibody systems. *Applied Numer. Math.*, 28:459–475, 1998.

[42] B. Simeon. *Numerische Simulation gekoppelter Systeme von partiellen und differential-algebraischen Gleichungen in der Mehrkörperdynamik*. Fortschritt-Berichte VDI Reihe 20 Nr. 315. VDI-Verlag, Düsseldorf, 2000.

[43] B. Simeon. Numerical analysis of flexible multibody systems. *Multibody System Dynamics*, 6:305–325, 2001.

[44] B. Simeon, F. Grupp, C. Führer, and P. Rentrop. A nonlinear truck model and its treatment as a multibody system. *J. Comp. Appl. Math.*, 50:523–532, 1994.

[45] G. Söderlind. Automatic control and adaptive time-stepping. *Numerical Algorithms*, 31:281–310, 2002.

[46] A. Zanna. On the numerical solution of isospectral flows. Dissertation, Unversity of Cambridge, 1998.