Zur Theorie und zur numerischen Lösung von Anfangswertproblemen für differentiell-algebraische Systeme von höherem Index

Habilitationsschrift

vorgelegt der
Mathematisch–Naturwissenschaftlichen Fakultät
der Universität Rostock

von
Dr. rer. nat. Martin Arnold
geboren in Halle/S.

Gutachter: Prof. Dr. G. Mayer (Rostock)
Prof. Dr. P. Rentrop (Darmstadt)
Prof. Dr. K. Strehmel (Halle/S.)

Habilitationsschrift eingereicht am: 18. Dezember 1996
Tag der Probevorlesung und des Kolloquiums: 2. Juni 1997

Die vorliegende Arbeit wäre nicht denkbar gewesen ohne die zahllosen Anregungen und Hinweise aus schöpferischen Diskussionen mit vielen Fachkollegen. Ihnen allen sei an dieser Stelle herzlich gedankt.

Ganz besonders möchte ich Herrn Prof. Dr. K. Strehmel von der Martin–Luther–Universität Halle–Wittenberg danken, der mich vor nunmehr 10 Jahren an die damals noch junge Thematik der numerischen Lösung von differentiell-algebraischen Systemen heranführte und auf den auch die Anregung zur Beschäftigung mit der in Kapitel 2 dargestellten Störungstheorie zurückgeht.

Bei der Entwicklung und Implementierung der in Kapitel 3 besprochenen numerischen Verfahren für differentiell-algebraische Systeme konnte ich insbesondere von vielen interessanten Diskussionen mit den Herren Dr. C. Führer (Lund), Dr. A. Murua (San Sebastián) und Dr. B. Simeon (Darmstadt) profitieren.

Die in Kapitel 4 vorgestellten Modelle und Algorithmen zur dynamischen Simulation von Rad–Schiene–Systemen sind das Ergebnis der langjährigen engen Zusammenarbeit mit Herrn Dipl.-Ing. H. Netter (Oberpfaffenhofen).

Den Herren Prof. Dr. G. Mayer (Rostock), Prof. Dr. P. Rentrop (Darmstadt) und Prof. Dr. K. Strehmel (Halle/S.) bin ich dankbar für die Übernahme der Gutachten und für das Interesse, das sie der vorliegenden Arbeit entgegengebracht haben.

Schließlich — last, but not least — gilt mein Dank den Angehörigen des Fachbereichs Mathematik der Universität Rostock, insbesondere den Kollegen der Arbeitsgruppe von Herrn HDoz. Dr. K. Frischmuth. Mit ihnen verband mich in den letzten Jahren nicht nur die fachliche Zusammenarbeit, stets konnte ich darüberhinaus auf die für die wissenschaftliche Arbeit unverzichtbare Unterstützung und Hilfe zählen.

Teile der vorliegenden Arbeit erfreuten sich der finanziellen Förderung durch den Bundesminister für Bildung, Wissenschaft, Forschung und Technologie im Rahmen des gemeinschaftlich von der Universität Rostock und dem Institut für Robotik und Systemdynamik der DLR (Oberpfaffenhofen) getragenen Projekts „Differentialgleichungen und singuläre Mannigfaltigkeiten in der dynamischen Simulation von Rad–Schiene–Systemen“.

Geisenbrunn, im November 1997 Martin Arnold

Inhaltsverzeichnis

Symbolverzeichnis

allgemein

$\|\cdot\|$	Norm in $\mathbb{R}^n$, 4.
α, β	Baumgarte–Koeffizienten, 55, 69.
$C^r([a,b],\mathbb{R}^n)$	Raum der r-mal stetig differenzierbaren Funktionen $f : [a,b] \to \mathbb{R}^n$, 3.
$\|\cdot\|_{C^r}$	Norm in $C^r([a,b],\mathbb{R}^n)$, 4, 21.
D_q	totale Ableitung bezüglich q, 75.
$D(t)$	Abkürzung für $\|\delta\|_{C^0} + \|\theta\|_{C^1}$, 24.
D_h	Abkürzung für $\delta + \frac{1}{h}\theta$, 34.
Δ_x	Ortsdiskretisierungsschrittweite, 15, 60.
δ_{jk}	Kronecker–Symbol, 61.
$\delta(t)$, δ_n	Störungen in den Differentialgleichungen, 24, 28, 33.
g_y, $\frac{\partial g}{\partial y}$	Jacobimatrix, 7.
$[g_y f]$, $[g_z^{-1} g_t]$	Produkt von partiellen Ableitungen, 7.
$\frac{\partial}{\partial y}[f_z(g_y f_z)^{-1}]$	Tensornotation (2.19), 25.
$\Gamma(q,v,\lambda)$	Jacobimatrix in MKS–Modellgleichungen, 43.
h	Integrationsschrittweite, 11.
$\mathfrak{M}$	Mannigfaltigkeit, 10.
μ	Schranke für $\|f_{zz}\|$ bzw. $\|f_{\lambda\lambda}\|$, 23, 44.
$\overline{\mu}$	$\overline{\mu} = \mu + 1/\Delta_z$, 24.
n_x, n_y, ...	Dimension von $x(t)$, $y(t)$, ..., 3.
$P(t)$	Projektor, 24.
$r^{(\mathrm{L})}$, $r^{(\mathrm{G})}$, $r^{(\mathrm{B})}$, ...	Funktionen in den MKS–Zwangsbedingungen, 67.
$S(t,\xi)$, $S(q)$	Projektor, 44.
TOL, ATOL, RTOL	Fehlerschranken, 18.
$T_y\mathfrak{M}$	Tangentialraum der Mannigfaltigkeit $\mathfrak{M}$, 10.
$\theta(t)$, θ_n	Störungen in den algebraischen Gleichungen, 24, 28, 33.
$\mathcal{U}$	Umgebung der analytischen Lösung, 19.
x', $\dot{x}$	Ableitung $x'(t) = \dot{x}(t) = \frac{d}{dt}x(t)$, 3, 15.

Abkürzungen

BDF	Backward differentiation formula(e), 12.
DA–System	Differentiell-algebraisches System, 3.
FSAL	„First same as last“, 102.
GGL–Formulierung	Gear–Gupta–Leimkuhler–Formulierung der MKS–Bewegungsgleichungen, 70.
HERK	Halb-explizite Runge–Kutta–Verfahren, 80.
MKS	mechanisches Mehrkörpersystem, 15.
PLMSV	Partitionierte lineare Mehrschrittverfahren, 88.

Diskretisierungsverfahren

a_{ij}, γ_{ij}, b_j, d_j	Parameter der HERK–Verfahren, 80.
α_j, β_j, $\hat{\alpha}_j$, $\hat{\beta}_j$	Parameter der PLMSV vom Adams–Typ, 88, 121.
c_i, $\hat{c}_{i+1}$	Knoten $c_i = \sum_j a_{ij}$, $\hat{c}_{i+1} = \sum_j \gamma_{ij}$, 81, 101.
$\delta y_h(t)$, $\delta z_h(t)$, $\delta\zeta_h(t)$	lokale Fehler, 83, 90.
$\kappa = \sum_i d_i w_{i1}$	Konstante in Kontraktivitätsbedingung, 84.
$\Psi(y_n, \ldots, z_{n+k}; f, g, h)$	Verfahrensfunktion von PLMSV, 88.
$\varrho(\xi)$, $\sigma(\xi)$, $\hat{\varrho}(\xi)$, $\hat{\sigma}(\xi)$	charakteristische Polynome, 90.
$\hat{s}$, s	Stufenzahl, 80, 102.
T_y, T_z	Mengen von Graphen im Baummodell, 97.
τ_j	Parameter der β-geblockten Verfahren, 121.
$W = (w_{ij})$	Matrix von Verfahrensparametern, 81.

Mehrkörpersysteme mit Kontaktbedingungen

a, ϱ	Parameter im Modellproblem, 133.
$\mathfrak{C}$	Kurve auf der Radoberfläche, 135.
$d(\xi_x, \xi_y)$	Abstand zweier undeformierter Starrkörper, 149.
$\Delta(\xi; q)$, $\Delta(s; q_{\mathrm{rel}}, x)$	Abstandsfunktion, 133, 136.
δ	elastische Annäherung im elastischen Modell, 150.
$\delta^{(\nu)}$	Annäherung der Körper im quasi-elastischen Modell, 153.
E, $E^{(W)}$, $E^{(R)}$	Elastizitätsmoduln, 139, 150.
e_3	Einheitsvektor in $\boldsymbol{R}$, 135.
$\mathcal{E}$	Tangentialebene im Kontaktpunkt, 148.
$F(s)$	Profilfunktion Rad, 134.
F_N	Zwangskraft, 23, 52, 139.
F_R	Reibungskraft, 52, 139.
F_y, F_z	Kräfte in y- bzw. z-Richtung, 139.

$G(v; x)$	Profilfunktion Schiene, 134.
$\mathfrak{K}$	Teilstück der Kurve $\mathfrak{C}$, 153.
$\kappa_x^{(\mathrm{W})}$, $\kappa_y^{(\mathrm{W})}$, $\kappa_x^{(\mathrm{R})}$, $\kappa_y^{(\mathrm{R})}$	Krümmungen von Paraboloiden, 156.
$L(q, \dot{q}, \lambda)$	Lagrange–Funktion, 130.
ν_{Stahl}, ν_{W}, ν_{R}	Querkontraktionszahlen, 139, 150.
P_{W}, P_{R}	Punkte auf dem Rad bzw. der Schiene, 135.
P_{W}^*	Kontaktpunkt im Starrkörperkontaktmodell, 138.
$p(\xi_x', \xi_y')$	Normaldruckverteilung, 148.
$q_{\mathrm{rel}} = (\xi_v, \xi_w, \varphi, \vartheta, \psi)^T$	relative Lage des Rades zur Schiene, 135.
$\boldsymbol{R}$	Koordinatensystem Schiene („**R**ail“), 134.
s_*	Kontaktpunktkoordinate, 138.
$\tilde{s}$	$P_{\mathrm{W}}(\tilde{s})$ ist Angriffspunkt der Reibungskraft, 156.
$\mathrm{smax}_s^{(\nu)}\zeta$	Regularisierung von $\max_s \zeta$, 151.
$\mathrm{smax}_s^{(\nu,h)}\zeta$	Diskretisierung von $\mathrm{smax}_s^{(\nu)}\zeta$, 162.
$\overline{u}_z^{(W)}$, $\overline{u}_z^{(R)}$	vertikale Verschiebung der Oberflächen, 149.
V_0	Geschwindigkeit der geführten Bewegung, 139.
$\boldsymbol{W}$	Koordinatensystem Rad („**W**heel“), 134.
$\zeta(\xi; q)$, $\zeta(s; q_{\mathrm{rel}}, x)$	Hilfsfunktion $\zeta = q_2 - \Delta$ bzw. $\zeta = -\xi_w - \Delta$, 133, 136.
$\Omega \subset \mathcal{E}$	Kontaktfläche im elastischen Modell, 148.

Vorwort

Das seit Beginn der achtziger Jahre außerordentlich gewachsene Interesse an der analytischen Untersuchung und an der numerischen Lösung von differentiell-algebraischen Systemen ist untrennbar mit der Entwicklung von immer umfangreicheren Modellen in praktischen Anwendungen in Naturwissenschaft und Technik verbunden. Traditionell versucht man — z. B. in der klassischen Mechanik — Differentialgleichungsmodelle in Minimalkoordinaten zu formulieren. Diese Minimalkoordinaten sind voneinander unabhängig, ihre zeitliche Ableitung wird durch ein System

$$y'(t) = f(t, y(t)) \tag{0.1}$$

von gewöhnlichen Differentialgleichungen beschrieben. Dem Vorteil einer niedrigeren Dimension der Modellgleichungen steht als Nachteil gegenüber, daß die Bestimmung von Minimalkoordinaten für größere Modelle, die z. B. mit Unterstützung des Computers aufgestellt werden, häufig kompliziert ist und einen unverhältnismäßig hohen numerischen Aufwand erfordert.

Ein typisches Beispiel ist die separate Modellierung einzelner Baugruppen eines mechanischen Mehrkörpersystems. Fügt man diese Teilmodelle zu einem Differentialgleichungsmodell für das vollständige Mehrkörpersystem zusammen, so können die Lage- und Geschwindigkeitskoordinaten der einzelnen Teilmodelle i. allg. nicht unabhängig voneinander gewählt werden, weil die Baugruppen im Mehrkörpersystem gekoppelt sind. Dieser Kopplung entsprechen in den Modellgleichungen Zwangsbedingungen an die Koordinaten x und an ihre Ableitungen x', so daß die Modellierung hier zunächst nicht auf ein System gewöhnlicher Differentialgleichungen, sondern auf Gleichungen der allgemeinen Form

$$F(t, x(t), x'(t)) = 0 \tag{0.2}$$

führt. Solche *differentiell-algebraischen Systeme* treten in zahlreichen Anwendungsgebieten als Modellgleichungen auf, Schwerpunkte sind neben der dynamischen Simulation von mechanischen Mehrkörpersystemen u. a. Anwendungen in der Schaltkreissimulation, in der chemischen Reaktionskinetik und in Problemen der optimalen Steuerung.

Oft wäre es prinzipiell möglich, zum differentiell-algebraischen System zumindest lokal eine äquivalente Beschreibung in Minimalkoordinaten, d. h. ein äquivalentes System gewöhnlicher Differentialgleichungen anzugeben. Die Entwicklung der letzten Jahre hat jedoch gezeigt, daß es sowohl aus Sicht der Theorie als auch aus Sicht der effizienten numerischen Lösung günstiger ist, direkt das differentiell-algebraische System zu untersuchen. Neben der praktischen Anwendung der neu entwickelten numerischen Verfahren konzentriert man sich dabei auf zwei Aspekte:

1. Analysis und numerische Lösung von differentiell-algebraischen Systemen der allgemeinen Struktur (0.2) (hierzu zählen u. a. die Klassifikation der differentiell-algebraischen Systeme nach ihrem Index und Aussagen zur Existenz und Eindeutigkeit der Lösung von Anfangs- und Randwertproblemen),

2. Konstruktion, Untersuchung und Implementierung von effizienten numerischen Verfahren für bestimmte Klassen von differentiell-algebraischen Systemen, die häufig in praktischen Anwendungen auftreten.

Die vorliegende Arbeit ist dem 2. Schwerpunkt zuzurechnen und behandelt im wesentlichen drei Themen.

Den Anfang bildet die Darstellung einer Störungstheorie für einfache Modellprobleme (nichtlineare differentiell-algebraische Systeme vom Index 2 und 3 in Hessenbergform), die bis ins Detail den Einfluß kleiner Störungen auf die Lösung von Anfangswertproblemen erklärt. Interpretiert man die bei der numerischen Lösung auftretenden Diskretisierungs- und Rundungsfehler als solche Störungen der Gleichungen des differentiell-algebraischen Systems, so ergibt sich als Spezialfall der Störungstheorie der Nachweis der Konvergenz und der numerischen Stabilität von Diskretisierungsverfahren.

Im Mittelpunkt des zweiten Teils steht die Lösung von Anfangswertproblemen für differentiell-algebraische Systeme mit numerischen Verfahren, die an bekannte Verfahren aus der Theorie der gewöhnlichen Differentialgleichungen angelehnt sind. Durch analytische Transformationen (Indexreduktion) kann das differentiell-algebraische System so umgeformt werden, daß eine zuverlässige und robuste numerische Integration möglich ist. Speziell für Index-2-Systeme in Hessenbergform werden durch neue Verfahrensansätze (halb-explizite Runge–Kutta–Verfahren mit expliziter Stufe, partitionierte lineare Mehrschrittverfahren) Diskretisierungsverfahren konstruiert, die die aus der Literatur bekannten Verfahren sinnvoll ergänzen und insbesondere für nicht-steife Systeme außerordentlich effektiv sind.

Um die in der Regel harten Anforderungen an Rechengeschwindigkeit und Robustheit der Integrationsverfahren in industriellen Anwendungen zu erfüllen, reicht es oft nicht aus, vorhandene Integrationssoftware direkt auf gegebene Modellgleichungen anzuwenden. Im dritten Teil der Arbeit wird am Beispiel der dynamischen Simulation von Rad–Schiene–Systemen gezeigt, daß die numerischen Lösungsverfahren für differentiell-algebraische Systeme auch diesen Anforderungen der industriellen Praxis gerecht werden können, wenn das technische System geeignet modelliert wird und die vorhandene Software an die spezielle Struktur der Modellgleichungen angepaßt wird. Das zentrale Problem ist dabei die geometrische Modellierung des Rad–Schiene–Kontakts, denn die zunächst naheliegende Verwendung eines Starrkörperkontaktmodells führt zu Singularitäten im differentiell-algebraischen System („Impasse points“). Für die dynamische Simulation von Rad–Schiene–Systemen im Rahmen des Programmpakets SIMPACK wurde ein regularisiertes Kontaktmodell aufgestellt und implementiert.

Die Untersuchung und Verbesserung von numerischen Lösungsverfahren für die Modellgleichungen von mechanischen Mehrkörpersystemen mit Zwangsbedingungen gab den Anstoß zu der vorliegenden Arbeit. In den numerischen Testrechnungen konzentrieren wir uns auf diese für differentiell-algebraische Systeme typische Anwendung. Gegenstand der theoretischen Untersuchungen sind dagegen nicht nur die Modellgleichungen für dieses Teilgebiet, sondern beliebige differentiell-algebraische Systeme vom Index ≤ 3, wobei wir uns vorwiegend auf semi-explizite Systeme beschränken.

Kapitel 1

Einleitung

Gegenstand der vorliegenden Arbeit sind die analytischen Eigenschaften und die numerische Lösung von Anfangswertproblemen für differentiell-algebraische Systeme (DA–Systeme):

$$F(t, x(t), x'(t)) = 0 \ , \quad (\, t \in [0,T]\,) \ , \quad x(0) = x_0 \ . \tag{1.1}$$

Hier bezeichnet $x : [0,T] \to \mathbb{R}^{n_x}$ die (gesuchte) Lösung des Anfangswertproblems, die auf dem endlichen Zeitintervall $[0,T]$ definiert ist[1].

Definition 1 Die Funktion $x \in C^1([0,T], \mathbb{R}^{n_x})$ heißt *Lösung* des Anfangswertproblems (1.1), wenn für alle $t \in [0,T]$ gilt $F(t, x(t), x'(t)) = 0$ und die Anfangsbedingung $x(0) = x_0$ erfüllt ist. ($C^r([a,b], \mathbb{R}^n)$ bezeichnet den Raum der r-fach stetig differenzierbaren Funktionen $f : [a,b] \to \mathbb{R}^n$. Ist die Dimension n des Bildraums aus dem Kontext ersichtlich, so schreiben wir kurz $C^r([a,b])$).

Die Theorie der differentiell-algebraischen Systeme hat sich in den vergangenen Jahren rasch entwickelt, so daß ein vollständiger Überblick weit über den Umfang dieser Einleitung hinausgehen würde. Neben den klassischen Monographien von Griepentrog und März ([76]) und von Hairer, Lubich und Roche ([81]) sei hier vor allem auf die im Jahr 1996 erschienenen überarbeiteten 2. Auflagen der Bücher von Brenan, Campbell und Petzold ([39]) und von Hairer und Wanner ([84]) verwiesen. In diesem einleitenden Kapitel werden einige grundlegende Begriffe der Theorie der DA–Systeme zusammengestellt, und es wird ein Überblick über die in der vorliegenden Arbeit untersuchten Fragestellungen gegeben.

Der Index eines differentiell-algebraischen Systems

Schon einfache lineare Beispiele zeigen, daß sich das Lösungsverhalten von DA–Systemen grundlegend von den Eigenschaften der Lösung eines Systems gewöhnlicher Differentialgleichungen unterscheiden kann:

Beispiel 1 Gegeben sei das lineare DA–System

$$N x'(t) + x(t) = q(t) \ , \quad (t \in [0,T]) \tag{1.2}$$

[1] Hier wie im folgenden gibt n_x die Dimension des Vektors $x(t) = (\, x_1(t), \ldots, x_{n_x}(t)\,)^T$ an, entsprechend n_y die des Vektors $y(t)$ usw., d. h. $x(t) \in \mathbb{R}^{n_x}$, $y(t) \in \mathbb{R}^{n_y}$, $\lambda(t) \in \mathbb{R}^{n_\lambda}$, ...

mit $x(0) = x_0$, $q : [0,T] \to \mathbb{R}^{n_x}$ und der nilpotenten Matrix

$$N = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{pmatrix} \in \mathbb{R}^{m\times m} \quad \text{mit} \quad m := n_x \, .$$

In Komponentendarstellung ergibt sich

$$\begin{aligned} x_k(t) + x'_{k+1}(t) &= q_k(t) \, , \quad (k = 1, \ldots, m-1) \, , \\ x_m(t) &= q_m(t) \, . \end{aligned}$$

Sind die Komponenten $q_k(t)$ des inhomogenen Terms $q(t)$ hinreichend oft differenzierbar ($q_k \in C^{k-1}([0,T],\mathbb{R})$), so erhält man die allgemeine Lösungsdarstellung

$$\begin{aligned} x_m(t) &= q_m(t) \, , \\ x_{m-1}(t) &= q_{m-1}(t) - q'_m(t) \, , \\ &\vdots \\ x_1(t) &= \sum_{k=1}^{m} (-1)^{k+1} \frac{d^{k-1}}{dt^{k-1}} q_k(t) \end{aligned} \tag{1.3}$$

von (1.2). Wesentliche Unterschiede zur Lösung eines linearen Systems gewöhnlicher Differentialgleichungen sind u. a.:

1. Die Lösung $x(t)$ ist durch das DA–System bereits eindeutig bestimmt. Das Anfangswertproblem $x(0) = x_0 = (x_{1,0}, \ldots, x_{m,0})^T$ hat genau dann eine Lösung, wenn die Anfangswerte konsistent zum DA–System sind, d. h., wenn gilt $x_{k,0} = x_k(0)$, $(k = 1, \ldots, m)$ mit den in (1.3) definierten Funktionen $x_k(t)$.

2. Wenn $m > 1$ ist, dann hängt die Lösung $x(t)$ bezüglich der Maximumnorm $\|q\|_{C^0} = \max_{t\in[0,T]} \|q(t)\|$ nicht stetig von den Eingangsgrößen $q(t)$ ab[2]. Die Lösung $x(t)$ enthält Ableitungen von $q(t)$ bis einschließlich $(m-1)$-ter Ordnung.

3. Das DA–System hat nur für hinreichend oft differenzierbare Funktionen $q(t)$ eine klassische Lösung. Dabei ergeben sich unterschiedlich starke Forderungen an die Differenzierbarkeit der einzelnen Komponenten von q (z. B. $q_1 \in C([0,T])$ aber $q_m \in C^{m-1}([0,T])$).

Wendet man für $m > 1$ ein Diskretisierungsverfahren direkt auf (1.2) an, so wird die Funktion $q(t)$ numerisch differenziert, dies führt zu instabilen numerischen Algorithmen (vgl. z. B. [49, Abschnitt 2.3]). Die Schwierigkeiten bei der numerischen Lösung wachsen für größeres m rasch an. Der aus der linearen Algebra ([69, §XII.5]) bekannte Nilpotenzindex m des Matrixbüschels $\{ I + \lambda N : \lambda \in \mathbb{C} \}$ gab diesem (groben) Maß für die Probleme, die die numerische Lösung von (1.2) verursachen kann, den Namen: m heißt *Index* des linearen DA–Systems (1.2).

[2] Wegen der Normäquivalenz im $\mathbb{R}^{n_x}$ gilt diese Aussage unabhängig davon, welche Vektornorm $\|\cdot\| : \mathbb{R}^{n_x} \to \mathbb{R}$ man für $\|q(t)\|$ wählt. Wird in der vorliegenden Arbeit eine nicht näher spezifizierte Vektornorm $\|\cdot\|$ im $\mathbb{R}^n$ verwendet, so setze man z. B. $\|\cdot\| = \|\cdot\|_2$.

Die für das elementare Beispiel (1.2) beobachteten Besonderheiten des Lösungsverhaltens sind charakteristisch für lineare DA–Systeme (1.1) mit konstanten Koeffizienten ([69, Kapitel XII], [84, Satz VII.1.1]). Dagegen ist die analytische Untersuchung von nichtlinearen DA–Systemen (1.1) i. allg. sehr viel komplizierter als im linearen Fall.

Aus der Literatur sind zahlreiche Ansätze zur Klassifikation von nichtlinearen DA–Systemen bekannt (vgl. [39, Kapitel 2 und Abschnitt 7.2]). Die auf Campbell, Gear und andere zurückgehende Idee, (1.1) durch wiederholte Differentiation der Gleichungen des DA–Systems in ein System gewöhnlicher Differentialgleichungen zu transformieren, zählt ebenso zu den klassischen Ansätzen wie der erstmals von Rheinboldt untersuchte Zusammenhang zwischen DA–Systemen und Vektorfeldern auf differenzierbaren Mannigfaltigkeiten. In beiden Fällen ist es möglich, bekannte Ergebnisse aus der Theorie der gewöhnlichen Differentialgleichungen auf nichtlineare DA–Systeme (1.1) zu übertragen.

Beispiel 2 Zunächst wird der Ansatz von Gear betrachtet. Wollte man in Beispiel 1 ein zu (1.2) äquivalentes System gewöhnlicher Differentialgleichungen erhalten, so „fehlt“ zunächst eine Differentialgleichung für die Komponente x_1. Differenziert man die erste Gleichung des DA–Systems (1.2)

$$x_1(t) + x'_2(t) = q_1(t) \, ,$$

bezüglich t, so ergibt sich

$$x'_1(t) + x''_2(t) = q'_1(t) \, .$$

$x''_2(t)$ bestimmt man nun aus der zweiten Ableitung der zweiten Gleichung in (1.2) usw. usf. . Am Ende ist die m-te Gleichung von (1.2) m-fach zu differenzieren und man erhält das zu (1.2) äquivalente System gewöhnlicher Differentialgleichungen

$$\begin{aligned} x'_1(t) &= \sum_{k=1}^{m} (-1)^{k+1} \frac{d^k}{dt^k} q_k(t) \, , \\ x'_2(t) &= q_1(t) - x_1(t) \, , \\ &\vdots \\ x'_{m-1}(t) &= q_{m-2}(t) - x_{m-2}(t) \, , \\ x'_m(t) &= q_{m-1}(t) - x_{m-1}(t) \, . \end{aligned}$$

Definition 2 ([73], [74], [75]) Sei die Funktion F in (1.1) hinreichend oft stetig differenzierbar. Zu einem gegebenen $m \in \mathbb{N}$ betrachtet man das DA–System

$$F(t, x(t), x'(t)) = 0 \, , \quad \frac{d}{dt} F(t, x(t), x'(t)) = 0 \, , \quad \ldots \, , \quad \frac{d^m}{dt^m} F(t, x(t), x'(t)) = 0 \, . \tag{1.4}$$

Wenn man aus (1.4) Gleichungen auswählen kann, die durch algebraische Umformungen auf die Gestalt $x'(t) = \varphi(t, x(t))$ gebracht werden können, und wenn m die kleinste natürliche Zahl mit dieser Eigenschaft ist, dann heißt m *Differentiationsindex* des DA–Systems (1.1).

Bemerkung 1 a) Für das lineare DA–System (1.2) fallen nach Beispiel 2 der Differentiationsindex und der Nilpotenzindex m zusammen.

b) Unter geeigneten Glattheitsvoraussetzungen erfüllt jede Lösung $x(t)$ von (1.1) wegen $F(t, x(t), x'(t)) = 0$ die Gleichungen (1.4).

c) Das in Definition 2 konstruierte System $x'(t) = \varphi(t, x(t))$ wird als ein dem DA–System (1.1) *zugrundeliegendes System gewöhnlicher Differentialgleichungen* bezeichnet.

d) Verwenden wir in der vorliegenden Arbeit den Begriff „Index“ ohne einen Zusatz, so ist damit stets der Differentiationsindex eines DA–Systems gemeint.

e) Zur Vereinfachung der Schreibweise beschränken wir uns hier auf DA–Systeme 1. Ordnung. Systeme höherer Ordnung können wie bei gewöhnlichen Differentialgleichungen durch Einführung von Hilfsvariablen $x^{[1]} := x$, $x^{[2]} := x' = \frac{d}{dt}x^{[1]}$, ... auf DA–Systeme 1. Ordnung zurückgeführt werden (vgl. [158, §11.I]). Als Differentiationsindex des DA–Systems höherer Ordnung wird der Differentiationsindex des zugehörigen Systems 1. Ordnung bezeichnet.

f) In der vorliegenden Arbeit werden ausschließlich DA–Systeme (1.1) betrachtet, für die gilt $F : \Omega \to \mathbb{R}^{n_x}$ mit einer geeigneten Menge $\Omega \subset [0,T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. Definition 2 kann aber ebenso auf die in der Literatur untersuchten über- bzw. unterbestimmten DA–Systeme übertragen werden.

Differentiell-algebraische Systeme in Hessenbergform

Beispiel 3 a) Läßt sich die Lösung x von (1.1) in Komponenten η und ζ aufspalten und hängt F nur von η', aber nicht von ζ' ab, so heißt (1.1) *semi-explizites* DA–System, wenn (1.1) die Ableitung $\eta'(t)$ explizit als Funktion von t, $\eta(t)$ und $\zeta(t)$ definiert. Hierzu betrachten wir als Beispiel das semi-explizite DA–System ($\eta \to y$, $\zeta \to z$)

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)) \\ 0 &= g(t, y(t), z(t)) \end{aligned} \tag{1.5}$$

mit Funktionen $f : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_y}$ und $g : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_z}$.
Bei der Bestimmung des dem DA–System (1.5) zugrundeliegenden Systems gewöhnlicher Differentialgleichungen ist es wegen der semi-expliziten Struktur ausreichend, statt der Ableitung $\frac{d}{dt}F(t,x,x')$ in (1.4) nur die Ableitung des algebraischen Teils zu betrachten. Differenziert man die Zwangsbedingungen $0 = g(t, y(t), z(t))$ bezüglich t, so folgt

$$0 = g_t(t,y,z) + g_y(t,y,z)y'(t) + g_z(t,y,z)z'(t) = g_t(t,y,z) + [g_y f](t,y,z) + g_z(t,y,z)z'(t) \; .$$

Wenn in einer Umgebung der Lösung von (1.5) die *Index-1-Bedingung*

$$\text{„ } g_z(t,y,z) \text{ ist regulär “} \tag{1.6}$$

erfüllt ist, dann erhält man hieraus

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)) \; , \\ z'(t) &= -[g_z^{-1} g_t](t, y(t), z(t)) - [g_z^{-1} g_y f](t, y(t), z(t)) \; , \end{aligned} \tag{1.7}$$

also hat (1.5) den (Differentiations-)Index 1. Hierbei bezeichnet $g_z = \frac{\partial}{\partial z} g(t,y,z)$ die Jacobimatrix von g bezüglich z und

$$[g_y f](t,y,z) := g_y(t,y,z) \cdot f(t,y,z) \; , \quad [g_z^{-1} g_t](t,y,z) := g_z^{-1}(t,y,z) \cdot g_t(t,y,z) \; , \; \ldots \; .$$

b) Aus dem semi-expliziten DA–System

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)) \; , \\ 0 &= g(t, y(t)) \end{aligned} \tag{1.8}$$

mit $f : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_y}$ und $g : [0,T] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_z}$ folgt durch Differentiation der Zwangsbedingungen $0 = g(t,y)$

$$0 = g_t(t,y) + g_y(t,y)y'(t) = g_t(t,y) + [g_y f](t,y,z) \; . \tag{1.9}$$

Ersetzt man in (1.8) die Zwangsbedingungen $0 = g(t,y)$ durch ihre Ableitung (1.9), so ergibt sich ein DA–System der Form (1.5), das wie oben durch eine weitere Differentiation der Zwangsbedingungen in ein System gewöhnlicher Differentialgleichungen überführt werden kann, wenn die *Index-2-Bedingung*

$$\text{„ } [g_y f_z](t,y,z) \text{ ist regulär “} \tag{1.10}$$

in einer Umgebung der Lösung erfüllt ist. Deshalb hat ein DA–System (1.8) mit (1.10) den (Differentiations-)Index 2.

c) Ist zu dem semi-expliziten DA–System

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)) \; , \\ z'(t) &= k(t, y(t), z(t), u(t)) \; , \\ 0 &= g(t, y(t)) \end{aligned} \tag{1.11}$$

mit $f : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_y}$, $k : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_z}$ und $g : [0,T] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_u}$ die *Index-3-Bedingung*

$$\text{„ } [g_y f_z k_u](t,y,z,u) \text{ ist regulär “} \tag{1.12}$$

in einer Umgebung der Lösung erfüllt, so zeigt man durch dreimalige Differentiation der Zwangsbedingungen, daß der (Differentiations-)Index 3 beträgt.

Da DA–Systeme der Struktur (1.8) und (1.11) in zahlreichen praktischen Anwendungen auftreten, werden sie unter einem einheitlichen Begriff zusammengefaßt ([48]):

Definition 3 ([39, Definition 2.5.3]) Das System (1.1) heißt DA–System in *Hessenbergform* der Größe m ($m \geq 2$), wenn es die Blockstruktur

$$\begin{aligned} \frac{d}{dt}x^{[1]}(t) &= F^{[1]}(t, x^{[1]}, x^{[2]}, \ldots, x^{[m]}) \; , \\ \frac{d}{dt}x^{[k]}(t) &= F^{[k]}(t, x^{[k-1]}, x^{[k]}, \ldots, x^{[m-1]}) \; , \quad (2 \leq k \leq m-1) \\ 0 &= F^{[m]}(t, x^{[m-1]}) \end{aligned}$$

hat, die Funktion F stetig differenzierbar ist und

$$\Big(\frac{\partial F^{[m]}}{\partial x^{[m-1]}}\Big)\cdot\Big(\frac{\partial F^{[m-1]}}{\partial x^{[m-2]}}\Big)\cdots\Big(\frac{\partial F^{[2]}}{\partial x^{[1]}}\Big)\cdot\Big(\frac{\partial F^{[1]}}{\partial x^{[m]}}\Big)$$

in einer Umgebung der Lösung regulär ist. (Die Bezeichnung „Hessenbergform“ wurde in Anlehnung an die Struktur der Jacobimatrix $\partial F/\partial x$ gewählt. Hat F die angegebene Form, so ist $\partial F/\partial x$ eine Block–Hessenbergmatrix.)

Beispiel 4 Index-2-Systeme in Hessenbergform haben die Gestalt (1.8) mit $x^{[1]} = y$, $x^{[2]} = z$, $F^{[1]} = f$ und $F^{[2]} = g$, sie erfüllen die Bedingung (1.10). DA–Systeme (1.11) mit (1.12) sind Index-3-Systeme in Hessenbergform.

Für DA–Systeme in Hessenbergform ist der Nachweis der eindeutigen Lösbarkeit von Anfangswertproblemen sehr viel einfacher als für nichtlineare DA–Systeme (1.1) mit beliebiger Struktur (vgl. z. B. [154, Abschnitt 9.4]). Wie im Satz von Picard–Lindelöf ist hierfür die wesentliche Voraussetzung, daß die rechte Seite des zugrundeliegenden Systems gewöhnlicher Differentialgleichungen $x'(t) = \varphi(t, x(t))$ aus Definition 2 Lipschitz-stetig bezüglich x ist. Zur Vereinfachung der Schreibweise fordern wir jedoch in Satz 1, daß die Funktionen f, g und k aus (1.5), (1.8) und (1.11) so oft stetig differenzierbar sind, daß die Lipschitz–Stetigkeit von φ garantiert ist.

Satz 1 *Gegeben sei ein DA–System (1.1) der Form (1.5), (1.8) oder (1.11), für das die Funktion F für alle $t \in [0,T]$ in einer Umgebung des Anfangswerts x_0 hinreichend oft stetig differenzierbar ist.*

a) Ist in einer Umgebung des Anfangswerts $x_0 = (y_0^T, z_0^T)^T$ die Index-1-Bedingung (1.6) erfüllt und gilt

$$0 = g(0, y_0, z_0) \ ,$$

so ist für (1.5) das Anfangswertproblem $y(0) = y_0$, $z(0) = z_0$ lokal eindeutig lösbar, d. h., es gibt ein $T_0 \in (0,T]$, so daß das Anfangswertproblem auf dem Zeitintervall $[0,T_0]$ eindeutig lösbar ist.

b) Gegeben sei ein semi-explizites DA–System (1.8), für das die Index-2-Bedingung (1.10) in einer Umgebung des Anfangswerts $x_0 = (y_0^T, z_0^T)^T$ erfüllt ist. Gilt für y_0 und z_0

$$\begin{aligned} 0 &= g(t,y(t))\Big|_{(t,y)=(0,y_0)} = g(0,y_0) \ , \\ 0 &= \frac{d}{dt}g(t,y(t))\Big|_{(t,y,z)=(0,y_0,z_0)} = g_t(0,y_0) + g_y(0,y_0)\cdot y'(t)\Big|_{(t,y,z)=(0,y_0,z_0)} \qquad (1.13) \\ &= g_t(0,y_0) + [g_y f](0,y_0,z_0) \ , \end{aligned}$$

so ist das Anfangswertproblem $y(0) = y_0$, $z(0) = z_0$ für (1.8) lokal eindeutig lösbar.

c) Gegeben sei ein semi-explizites DA–System (1.11), für das die Index-3-Bedingung (1.12) in einer Umgebung des Anfangswerts $x_0 = (y_0^T, z_0^T, u_0^T)^T$ erfüllt ist. Gilt für y_0, z_0 und u_0

$$\begin{aligned} 0 &= g(0,y_0) \ , \\ 0 &= \frac{d^k}{dt^k}g(t,y(t))\Big|_{(t,y,z,u)=(0,y_0,z_0,u_0)} \ , \quad (k=1,2) \ , \end{aligned}$$

so ist das Anfangswertproblem $y(0) = y_0$, $z(0) = z_0$, $u(0) = u_0$ für (1.11) lokal eindeutig lösbar.

Beweis a) Wie in Beispiel 3a erhält man das dem Index-1-System (1.5) zugrundeliegende System gewöhnlicher Differentialgleichungen (1.7). Sind die Funktionen f, $[g_z^{-1}g_t]$ und $[g_z^{-1}g_y f]$ Lipschitz-stetig bezüglich y und z, so folgt aus der Theorie der gewöhnlichen Differentialgleichungen, daß das Anfangswertproblem $y(0) = y_0$, $z(0) = z_0$ für (1.7) lokal eindeutig lösbar ist ([158, §6.III und §10]). Für diese Lösung $y(t)$, $z(t)$ gilt

$$g_t(t,y(t),z(t)) + g_y(t,y(t),z(t))y'(t) + g_z(t,y(t),z(t))z'(t) = 0 \ , \quad (t \in [0,T_0]) \ ,$$

also

$$g(t,y(t),z(t)) = g(0,y_0,z_0) + \int_0^t \frac{d}{d\tau}g(\tau,y(\tau),z(\tau))\,d\tau = g(0,y_0,z_0) = 0 \ , \quad (t \in [0,T_0]) \ .$$

Damit ist $y(t)$, $z(t)$ auch Lösung des DA–Systems (1.5).

b) und c) führt man durch einmalige bzw. durch zweimalige Differentiation der Zwangsbedingungen $0 = g(t,y(t))$ auf a) zurück ([154, Satz 9.4.3]). ∎

Bemerkung 2 a) Unter geeigneten Voraussetzungen an f, g und k kann die Lösung des Anfangswertproblems wie bei Systemen gewöhnlicher Differentialgleichungen über das gesamte Intervall $[0,T]$ fortgesetzt werden (vgl. [158, Satz §6.VII]).

b) Wie schon im einführenden Beispiel 1 sind auch in Satz 1 die Anfangswertprobleme nur unter zusätzlichen Bedingungen an die Anfangswerte lösbar. Dabei müssen die Anfangswerte nicht nur die im DA–System explizit enthaltenen Zwangsbedingungen, sondern u. U. weitere (*versteckte*) Zwangsbedingungen erfüllen (z. B. $0 = g_t(0,y_0) + [g_y f](0,y_0,z_0)$ für (1.8)). Generell ist die Bestimmung konsistenter Anfangswerte nicht trivial und für DA–Systeme der Form (1.1) bisher nicht allgemein gelöst (vgl. die Lösungsansätze in [125], [101], [41]). Sehr viel einfacher ist es, in semi-expliziten DA–Systemen konsistente Anfangswerte zu ermitteln. So kann man z. B. in (1.8) in einem ersten Schritt einen Vektor $y_0 \in \mathbb{R}^{n_y}$ mit $0 = g(0,y_0)$ bestimmen. Anschließend wird das nichtlineare Gleichungssystem

$$0 = g_t(0,y_0) + [g_y f](0,y_0,\zeta)$$

nach $\zeta \in \mathbb{R}^{n_z}$ aufgelöst und $z_0 := \zeta$ gesetzt. Schließlich erhält man $y'(0) = f(0,y_0,z_0)$ aus (1.8). Für viele Anwendungsgebiete lassen sich solche angepaßten Algorithmen zur Bestimmung konsistenter Anfangswerte angeben (vgl. z. B. [109], [112], [120], [148] und auch (3.86) und (3.87) in Abschnitt 3.3.2).

Deskriptorform und Zustandsform

Die Konsistenzbedingungen an die Anfangswerte widerspiegeln die Tatsache, daß die n_x Variablen des DA–Systems (1.1) i. allg. nicht voneinander unabhängig sind. Durch die Zwangsbedingungen werden die möglichen Zustandsänderungen im DA–System auf eine Teilmenge des $\mathbb{R}^{n_x}$ eingeschränkt.

Beispiel 5 (vgl. [84, S. 457f und S. 474f]) Gegeben sei ein autonomes DA–System (1.8), für das die Index-2-Bedingung (1.10) erfüllt ist, so daß insbesondere g_y Vollrang hat. Die Zwangsbedingungen $g(y) = 0$ mit $g : \mathbb{R}^{n_y} \to \mathbb{R}^{n_z}$ definieren unter geeigneten Glattheitsvoraussetzungen eine $(n_y - n_z)$–dimensionale Mannigfaltigkeit

$$\mathfrak{M} := \{\, y \in \mathbb{R}^{n_y} \,:\, g(y) = 0 \,\} ,$$

für die in einer Umgebung $V \subset \mathfrak{M}$ eines gegebenen Vektors $y_0 \in \mathfrak{M}$ eine Parameterdarstellung $\omega : U \to V$ angegeben werden kann. Die bijektive Funktion ω bildet eine offene Menge $U \subset \mathbb{R}^{n_y - n_z}$ nach V ab, dabei seien ω und ω^{-1} stetig differenzierbare Funktionen, für die die Jacobimatrix $\omega_\eta(\eta)$ für alle $\eta \in U$ Vollrang hat. Aus $\omega(\eta) \in V \subset \mathfrak{M}$ folgt $g(\omega(\eta)) \equiv 0$, also $\frac{\partial}{\partial \eta} g(\omega(\eta)) = g_y(\omega(\eta)) \cdot \omega_\eta(\eta) = 0$. Da ω_η Vollrang hat, bilden die Spaltenvektoren von $\omega_\eta(\eta)$ deshalb eine Basis des Tangentialraums

$$T_y\mathfrak{M} := \{\, v \in \mathbb{R}^{n_y} \,:\, g_y(y)v = 0 \,\}$$

an die Mannigfaltigkeit $\mathfrak{M}$ im Punkt $y := \omega(\eta)$, und für beliebige Vektoren $v \in T_y\mathfrak{M}$, $\eta \in U$ und $\tilde{\eta} \in \mathbb{R}^{n_y - n_z}$ gilt

$$\omega_\eta(\eta)\tilde{\eta} = v \quad \Leftrightarrow \quad \tilde{\eta} = \omega_\eta(\eta)^+ v \tag{1.14}$$

mit der Pseudoinversen $\omega_\eta(\eta)^+ := (\omega_\eta(\eta)^T\omega_\eta(\eta))^{-1}\omega_\eta(\eta)^T$ (vgl. [111, Satz 3.4.1]).
Unter Verwendung einer solchen Parameterdarstellung ω kann — wiederum unter geeigneten Glattheitsvoraussetzungen — ein zum DA–System (1.8) äquivalentes System von $n_y - n_z$ gewöhnlichen Differentialgleichungen angegeben werden: Hierzu sei ein Vektor $z_0 \in \mathbb{R}^{n_z}$ mit $0 = [g_y f](y_0, z_0)$ gegeben (vgl. (1.13)), und die Umgebung V von y_0 sei hinreichend klein gewählt. Nach dem Satz über die implizite Funktion sind die in (1.8) enthaltenen versteckten Zwangsbedingungen $0 = \frac{d}{dt} g(y(t)) = [g_y f](y, z)$ wegen (1.10) für $y \in V$ lokal eindeutig nach z auflösbar, d. h., es gibt eine Funktion $h : V \to \mathbb{R}^{n_z}$, so daß für $y \in V$ und z in einer geeigneten Umgebung von z_0 gilt

$$0 = [g_y f](y, z) \quad \Leftrightarrow \quad z = h(y)$$

und $z_0 = h(y_0)$. Dann ist $f(y, h(y)) \in T_y\mathfrak{M}$ und das DA–System (1.8) ist äquivalent zu

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

Ersetzt man hier $y(t)$ durch $\omega(\eta(t))$, so folgt aus $y'(t) = \frac{d}{dt}\omega(\eta(t)) = \omega_\eta(\eta(t))\eta'(t)$ zunächst

$$\omega_\eta(\eta(t))\eta'(t) = f(\omega(\eta(t)), h(\omega(\eta(t))))$$

und schließlich mit (1.14)

$$\eta'(t) = \psi(\eta(t)) \quad \text{mit} \quad \psi(\eta) := \omega_\eta(\eta)^+ f(\omega(\eta), h(\omega(\eta))) . \tag{1.15}$$

In einer Umgebung von (y_0, z_0) ist $(y(t), z(t))$ genau dann eine Lösung des DA–Systems (1.8), wenn es eine Lösung $\eta(t)$ des Systems (1.15) von $n_y - n_z$ gewöhnlichen Differentialgleichungen gibt, für die gilt $y(t) = \omega(\eta(t))$ und $z(t) = h(\omega(\eta(t)))$.
Das System (1.15) ist eine *lokale Zustandsform* des DA–Systems (1.8). Während ein Anfangswertproblem für ein System (1.15) mit Lipschitz-stetiger rechter Seite ψ für beliebige Anfangsbedingungen $\eta(0) = \eta_0$ lokal eindeutig lösbar ist, ergeben sich die Bedingungen (1.13) aus Satz 1b unmittelbar aus $y_0 = \omega(\eta_0)$, $z_0 = h(y_0)$.

Die in Beispiel 5 skizzierte Identifizierung von DA–Systemen mit Differentialgleichungen, die auf differenzierbaren Mannigfaltigkeiten definiert sind, wurde erstmals von Rheinboldt ([139]) untersucht. Sie ist Grundlage des von verschiedenen Autoren betrachteten geometrischen Zugangs zur analytischen Untersuchung von DA–Systemen (1.1) ([135], vgl. auch die Einführung in [84, S. 457f und S. 474ff]). Durch den Übergang zu einer Parametrisierung der Zwangsmannigfaltigkeit $\mathfrak{M}$ läßt sich mit den Mitteln der Differentialgeometrie u. a. ein eleganter Beweis für Satz 1, d. h. für die Existenz und Eindeutigkeit der Lösung von Anfangswertproblemen für (1.5), (1.8) und (1.11) angeben ([140]).

Solche Parameterdarstellungen der Zwangsmannigfaltigkeit lassen sich i. allg. nur lokal konstruieren. In technischen Anwendungen entspricht der Übergang von den Variablen x in (1.1) zu einer Parameterdarstellung der Zwangsmannigfaltigkeit dem Übergang von der Modellierung in Deskriptorform zur Modellierung in Minimalkoordinaten, d. h. zur Modellierung in Zustandsform. Dabei ist ein Differentialgleichungsmodell in *Zustandsform* gegeben, wenn es ein System gewöhnlicher Differentialgleichungen bildet, so daß die Anfangswerte unabhängig voneinander vorgegeben werden können (Minimalkoordinaten). Läßt man im Differentialgleichungsmodell jedoch zusätzliche (redundante) Variablen zu, so ergeben sich Zwangsbedingungen an die Variablen bzw. an ihre zeitlichen Ableitungen. Die Modellgleichungen haben dann die allgemeine Gestalt (1.1) und werden als Modellgleichungen in *Deskriptorform* bezeichnet (vgl. auch [50, S. 38])[3]. In Beispiel 5 liegt (1.8) in Deskriptorform und (1.15) in Zustandsform vor.

Ebenso wie man z. B. in der klassischen Mechanik versucht, durch *analytische* Transformationen zu Modellgleichungen in Zustandsform zu gelangen, lassen sich Parameterdarstellungen einer differenzierbaren Mannigfaltigkeit auch *numerisch* stabil und effizient bestimmen. Damit ermöglicht der differentialgeometrische Zugang lokal den Übergang von einem DA–System zu einem äquivalenten System gewöhnlicher Differentialgleichungen. Für die Integration eines Anfangswertproblems über das gesamte Zeitintervall $[0, T]$ ist in der Regel der Wechsel zwischen verschiedenen Parameterdarstellungen der Mannigfaltigkeit erforderlich. Auf der Grundlage dieser lokalen Zustandsformen wurden verschiedene Integrationsverfahren für die dynamische Simulation von mechanischen Mehrkörpersystemen entwickelt und implementiert ([133], [134], [159], [165]).

Diskretisierungsverfahren für differentiell-algebraische Systeme

Neben diesem Lösungsansatz, ein DA–System analytisch oder numerisch auf Zustandsform zu transformieren und auf diese Zustandsform eines der bekannten Verfahren für gewöhnliche Differentialgleichungen anzuwenden, wird etwa seit Beginn der achtziger Jahre intensiv die direkte Anwendung von Diskretisierungsverfahren auf DA–Systeme (1.1) untersucht. Der klassische Ansatz hierzu geht auf Gear ([70]) zurück und ist im einfachsten Fall (implizites Eulerverfahren) gegeben durch

$$F(t_{n+1}, x_{n+1}, \frac{x_{n+1} - x_n}{h}) = 0 \tag{1.16}$$

mit der Integrationsschrittweite h und $t_n = nh$, $x_n \approx x(t_n)$, $(n \geq 0)$. Im Vergleich zur Theorie der Diskretisierungsverfahren für Systeme gewöhnlicher Differentialgleichungen

[3] Selbstverständlich gibt es weder eine eindeutig bestimmte Zustandsform noch eine eindeutig bestimmte Deskriptorform der Modellgleichungen.

ergeben sich für die numerische Lösung von DA–Systemen sowohl bei der Konstruktion der Verfahren als auch bei ihrer theoretischen Untersuchung und bei ihrer praktischen Umsetzung zahlreiche neue Fragen und Probleme.

Es gibt in der Literatur verschiedene Konzepte zur theoretischen Untersuchung von Diskretisierungsverfahren für DA–Systeme. Wir folgen in der vorliegenden Arbeit dem von Hairer, Lubich und anderen beschrittenen Weg, zunächst die Anwendung der numerischen Verfahren auf nichtlineare DA–Systeme mit möglichst einfacher Struktur (z. B. Hessenbergform) im Detail zu untersuchen ([81], [84, Kapitel VI und VII] u. v. a.). Durch diese Einschränkung vermeidet man zahllose beweistechnische Details beim Nachweis der Konvergenz der Verfahren, so daß die Bestimmung optimaler Verfahrensparameter sehr erleichtert wird.

Viele in praktischen Anwendungen auftretende DA–Systeme der allgemeinen Form (1.1) können durch (lineare oder nichtlineare) Koordinatentransformationen und durch Umformungen der Gleichungen (1.1) auf Hessenbergform transformiert werden. Ist ein Diskretisierungsverfahren gegenüber dieser Transformation invariant, so lassen sich die für Systeme in Hessenbergform gezeigten Ergebnisse direkt auf solche DA–Systeme der allgemeinen Form (1.1) übertragen (vgl. hierzu [81, S. 3ff], [84, S. 456] und für die in der vorliegenden Arbeit untersuchten Verfahren die Bemerkungen 29 und 34). Dabei setzt man in der Regel voraus, daß in (1.1) die Funktion $F(t, x, x')$ und die Lösung $x(t)$ hinreichend oft stetig differenzierbar sind. Unter Verwendung des von März, Griepentrog und anderen entwickelten Projektorenkalküls ([76], [114]) wurde in [12] und [15] nachgewiesen, daß einige der für Index-1- und Index-2-Systeme in Hessenbergform entwickelten Beweisgedanken auch unter sehr viel schwächeren Glattheitsvoraussetzungen auf DA–Systeme der allgemeinen Form (1.1) übertragen werden können.

In den nachfolgenden Kapiteln 2 und 3 werden für Index-2-Systeme und Index-3-Systeme in Hessenbergform neben numerischen Lösungsverfahren auch einige Eigenschaften der analytischen Lösung untersucht. Schon das lineare DA–System (1.2) aus Beispiel 1 zeigt, daß die direkte Anwendung von Diskretisierungsverfahren auf DA–Systeme vom Index > 1 problematisch ist, denn die Lösung hängt (bezüglich $\|\cdot\|_{C^0}$) nicht stetig von der rechten Seite $q(t)$ ab (vgl. (1.3)). Unabhängig vom konkreten Diskretisierungsverfahren kann deshalb die numerische Lösung durch Rundungsfehler, die bei der Computer–Rechnung in Gleitpunktarithmetik unvermeidbar sind, stark verfälscht werden. Aus diesem Grund wurde im Zusammenhang mit Konvergenzbeweisen für BDF und implizite Runge–Kutta–Verfahren auch der Einfluß von Rundungsfehlern und anderen kleinen Störungen auf die numerische Lösung untersucht (z. B. in [104], [40], [81], [114, Abschnitt 3]).

Auf der Grundlage dieser aus der Literatur bekannten Fehlerabschätzungen wird in Kapitel 2 in Anlehnung an die Definition des Störungsindexes ([81, S. 1], vgl. auch Definition 4) eine Störungstheorie für nichtlineare Index-2- und Index-3-Systeme in Hessenbergform entwickelt. Diese Störungstheorie beschreibt detailliert die Empfindlichkeit der analytischen und der numerischen Lösung gegenüber kleinen Störungen im DA–System. Das zentrale Ergebnis ist, daß die differentiellen Lösungskomponenten (y in (1.8) bzw. y und z in (1.11)) sehr viel robuster gegenüber Störungen sind als die algebraischen Komponenten (z in (1.8) bzw. u in (1.11)). Der Einfluß von Störungen auf die differentiellen

Lösungskomponenten hängt entscheidend davon ab, ob das DA–System bezüglich der algebraischen Komponenten linear oder nichtlinear ist.

Sowohl für die analytische Lösung als auch für die numerische Lösung werden scharfe Fehlerschranken für die einzelnen Lösungskomponenten bewiesen. Dabei wird auch der enge Zusammenhang zwischen dem Einfluß von Störungen auf die analytische Lösung und dem Einfluß von Störungen auf die numerische Lösung gezeigt. Die Ergebnisse der Störungstheorie unterstreichen, daß die direkte Anwendung von geeigneten Diskretisierungsverfahren auf Index-2-Systeme in Hessenbergform zu numerisch stabilen Integrationsverfahren führt, wenn man die stärkere Empfindlichkeit der algebraischen Lösungskomponenten z gegenüber Rundungsfehlern und anderen Störungen bei der Implementierung der Verfahren berücksichtigt (z. B. bei der Schrittweitensteuerung [81, Kapitel 8]).

Gleichzeitig erklärt die Störungstheorie, warum diese Integrationsverfahren prinzipiell versagen, wenn die Integrationsschrittweite h zu klein wird (vgl. Beispiel 8). Aus Sicht der Störungstheorie sind nichtlineare Index-2-Systeme (1.8) besonders gutartig, wenn sie linear bezüglich der algebraischen Komponenten z sind ($f(y, z) = f_0(y) + f_z(y)z$), denn in diesem Fall sind diejenigen Terme, die in den Fehlerschranken für die differentiellen Lösungskomponenten y nicht stetig von den Störungen abhängen, aus Sicht der praktischen Rechnung vernachlässigbar klein (vgl. Beispiel 10). Index-2-Systeme dieser speziellen Form treten in verschiedenen praktischen Anwendungen auf, als Beispiel sei die Gear–Gupta–Leimkuhler–Formulierung (3.6) der Modellgleichungen für mechanische Mehrkörpersysteme genannt.

Gegenstand von Kapitel 3 ist die Konstruktion und Implementierung von numerischen Verfahren für DA–Systeme, deren Index größer als 1 ist (*DA–Systeme von höherem Index*). Fehlerabschätzungen und aus der Literatur bekannte Konvergenzbeweise belegen, daß Diskretisierungsverfahren, die aus der Theorie der gewöhnlichen Differentialgleichungen bekannt sind (z. B. BDF, implizite Runge–Kutta–Verfahren), auf Index-2- und Index-3-Systeme in Hessenbergform übertragen werden können. Weite Verbreitung haben u. a. der BDF–Code DASSL von Petzold ([129], [132], [39, Kapitel 5 und 7.5]) und das auf einem impliziten Runge–Kutta–Verfahren 5. Ordnung basierende Programm RADAU5 von Hairer und Wanner ([83], [84, Anhang]) gefunden. Ebenso wie das implizite Eulerverfahren (1.16) basieren DASSL und RADAU5 auf impliziten Verfahren und erfordern in jedem Integrationsschritt die Lösung von nichtlinearen Gleichungssystemen der Dimension $\geq n_x$ zur Bestimmung von x_{n+1}.

Für semi-explizite DA–Systeme kann der in einem einzelnen Integrationsschritt erforderliche Rechenaufwand stark reduziert werden, wenn man die Struktur des DA–Systems bei der Konstruktion der Verfahren berücksichtigt. So erfordert z. B. das halb-explizite Eulerverfahren

$$\begin{aligned} \frac{y_{n+1} - y_n}{h} &= f(t_n, y_n, z_n) \\ 0 &= g(t_{n+1}, y_{n+1}) \end{aligned} \tag{1.17}$$

für Index-2-Systeme (1.8) nur die Lösung eines Gleichungssystems der Dimension n_z zur Berechnung von z_n. Die differentiellen Komponenten y_{n+1} können dagegen explizit berechnet werden ($n_z < n_x = n_y + n_z$). Das Verfahren (1.17) gehört zur Klasse der *partitionierten* Diskretisierungsverfahren für semi-explizite DA–Systeme, die seit Ende der achtziger Jahre intensiv untersucht werden (vgl. den aktuellen Überblick in [84, Kapitel VII.6]).

Bemerkung 3 Das klassische Anwendungsgebiet von BDF und impliziten Runge–Kutta-Verfahren sind steife Systeme gewöhnlicher Differentialgleichungen. Dagegen ist das halb-explizite Eulerverfahren (1.17) bei der Integration von steifen Systemen ebenso wie das explizite Eulerverfahren für gewöhnliche Differentialgleichungen nur für sehr kleine Integrationsschrittweiten h numerisch stabil, seine praktische Anwendung kommt deshalb vor allem für nicht-steife Systeme in Betracht. Für eine ausführliche Diskussion des Phänomens (und des Begriffs) „Steifheit" sei auf [50, Abschnitt 4.1.3], [154, Abschnitt 5.2.1] und [84, Kapitel IV.1] verwiesen. Gegenüber dieser klassischen Theorie für gewöhnliche Differentialgleichungen ergeben sich bei der Integration von steifen Index-2-Systemen (1.8) neue (und bisher ungelöste) Probleme, weil man steife DA–Systeme (1.8) angeben kann, für die sowohl die BDF als auch A-stabile (implizite) Runge–Kutta–Verfahren nur für sehr kleine Schrittweiten h numerisch stabil sind ([32], [87], [163]).

Im Mittelpunkt von Kapitel 3 stehen partitionierte Verfahren für nicht-steife Index-2-Systeme in Hessenbergform. Hierzu werden 2 neue Verfahrensklassen betrachtet: halb-explizite Runge–Kutta–Verfahren mit expliziter Stufe und partitionierte lineare Mehrschrittverfahren vom Adams–Typ. In beiden Fällen sind die zugehörigen Verfahren für gewöhnliche Differentialgleichungen (explizite Runge–Kutta–Verfahren bzw. implizite Adams–Verfahren) sehr gut zur Integration nicht-steifer Systeme geeignet.

Als Verallgemeinerung von (1.17) auf mehrstufige (explizite) Runge–Kutta–Verfahren haben Hairer, Lubich und Roche in [81, S. 20f] halb-explizite Runge–Kutta–Verfahren für Index-2-Systeme (1.8) eingeführt. Die Konstruktion von halb-expliziten Verfahren höherer Ordnung vereinfacht sich erheblich, wenn man die erste (halb-explizite) Stufe eines solchen Verfahrens durch eine *explizite Stufe* ersetzt. In den Abschnitten 3.2.1, 3.3.1 und 3.3.2 werden diese halb-expliziten Runge–Kutta–Verfahren mit expliziter Stufe detailliert untersucht. Ein besonders effizientes sechsstufiges Verfahren der Ordnung 5 wurde als Integrator HEDOP5 implementiert und bewies in zahlreichen Testrechnungen die Vorteile des neuen Verfahrensansatzes.

Die aus der Theorie der gewöhnlichen Differentialgleichungen bekannten impliziten Adams–Verfahren (auch: Adams–Moulton–Verfahren) sind bei Anwendung auf Index-2-Systeme (1.8) instabil und konvergieren nicht ([84, Satz VII.3.6]). Als Alternative werden in den Abschnitten 3.2.2 und 3.3.3 *partitionierte* Mehrschrittverfahren untersucht, die die differentiellen Komponenten y mit einem Adams–Moulton–Verfahren und die algebraischen Komponenten z mit BDF bestimmen. Kombiniert man dabei zwei k-Schritt-Verfahren ($k \geq 1$), so konvergiert dieses partitionierte lineare Mehrschrittverfahren vom Adams–Typ mit der Ordnung $k+1$ in y und mit der Ordnung k in z.

Die in Kapitel 3 vorgestellten halb-expliziten und partitionierten Verfahren zeigen, daß man bekannte Verfahren für nicht-steife gewöhnliche Differentialgleichungen so verallgemeinern kann, daß sie auch für Index-2-Systeme in Hessenbergform eine hohe Konvergenzordnung erreichen. Bei Anwendung auf nicht-steife DA–Systeme (1.8) sind diese neu entwickelten Verfahren den traditionell verwendeten impliziten Verfahren überlegen.

Differentiell-algebraische Systeme in praktischen Anwendungen

Die seit einigen Jahren verfügbare Software zur Integration von DA–Systemen ermöglicht es in vielen Anwendungsgebieten, zur Modellierung in Deskriptorform überzugehen. Die

Modellgleichungen haben dabei nicht immer die „klassische" Form (1.1), sondern können z. B. auch partielle Differentialgleichungen, retardierte Differentialgleichungen, Integralgleichungen und einseitige Beschränkungen enthalten ([39, Kapitel 7.1]).

Obwohl diese Erweiterungen zahlreiche neue Fragen aufwerfen, können Konzepte, die für DA–Systeme der Form (1.1) entwickelt wurden, z. T. auf Systeme mit komplizierterer Struktur übertragen werden. Hierzu wenden wir in Abschnitt 2.4 die Störungstheorie auf DA–Systeme an, die bei Anwendung der Linienmethode als Semidiskretisierungen eines Systems partieller Differentialgleichungen entstehen. Zur Charakterisierung der Empfindlichkeit der Lösung gegenüber kleinen Störungen wird der gleichmäßige Störungsindex für Klassen von DA–Systemen eingeführt. Der gleichmäßige Störungsindex gibt für die semidiskreten DA–Systeme Fehlerschranken an, die gleichmäßig bezüglich der (Orts–)Diskretisierungsschrittweite Δ_x sind.

Die computergestützte Generierung von Differentialgleichungsmodellen birgt die Gefahr, daß durch die automatisierte Kopplung von verschiedenen Modellkomponenten Modellgleichungen aufgestellt werden, in denen Singularitäten auftreten können. Eine solche Singularität in den Bewegungsgleichungen für mechanische Mehrkörpersysteme (MKS) gab die Motivation für die in Kapitel 4 zusammengefaßten Arbeiten zur Simulation des dynamischen Verhaltens von MKS mit Kontaktbedingungen.

Die in Abschnitt 4.1 im einzelnen besprochenen MKS–Modellgleichungen

$$\begin{aligned} M(q)\ddot{q} &= f(q,\dot{q},\lambda,t) - G^T(q,t)\lambda \\ 0 &= g(q,t) \end{aligned} \tag{1.18}$$

beschreiben in Abhängigkeit von den im System wirkenden Kräften die zeitliche Änderung der Lagekoordinaten q eines MKS[4]. Sie bilden unter den Voraussetzungen aus Abschnitt 4.1 ein DA–System vom Index 3.

Da sich Starrkörper nicht gegenseitig durchdringen können, sind ihre Lagekoordinaten im Mehrkörpersystem nicht unabhängig voneinander. Bei Berücksichtigung der geometrischen Abmessungen der Körper eines MKS ergeben sich deshalb einseitige Beschränkungen

$$\tilde{g}(q,t) \geq 0$$

an die Lagekoordinaten q des MKS. Betrachtet man den Spezialfall, daß zwei Starrkörper eines MKS permanent in Kontakt sind, so können diese einseitigen Beschränkungen durch eine (skalare) Zwangsbedingung

$$\gamma(q,t) = 0 \tag{1.19}$$

ersetzt werden, die in (1.18) Teil der Zwangsbedingungen $g(q,t) = 0$ ist.

Schwerpunkt von Kapitel 4 ist die dynamische Simulation von Rad–Schiene–Systemen. Hierfür ist aus der Literatur bekannt, daß der geometrische Kontakt zwischen einer Schiene und einem kegelförmigen Rad sehr gut durch ein derartiges Starrkörperkontaktmodell beschrieben werden kann (vgl. z. B. [148] und auch Abschnitt 4.1). In praxi werden jedoch

[4] Synonym zu $x'(t) = \frac{d}{dt}x(t)$ in (1.1) verwenden wir in (1.18) und in Kapitel 4 die in der Mechanik übliche Bezeichnung $\dot{x} = \frac{d}{dt}x(t)$ für die zeitliche Ableitung einer Funktion.

Räder mit sog. Verschleißprofil verwendet, diese Räder sind im Unterschied zu kegelförmigen Rädern nicht konvex. Wenn man das Starrkörperkontaktmodell formal auf den Kontakt zwischen einer Schiene und einem solchen Rad mit Verschleißprofil überträgt, so ist die Kontaktbedingung (1.19) nur stückweise differenzierbar, denn die Lage des Punkts, in dem sich die Schiene und das nicht-konvexe Rad berühren, kann sich unstetig ändern.

Die Differenzierbarkeit der Zwangsbedingungen in (1.18) ist jedoch eine entscheidende Voraussetzung für die Existenz einer klassischen Lösung $q(t)$, $\lambda(t)$ (vgl. Beispiel 5). Erreicht eine Trajektorie einen Punkt $P^* = (t^*, q^*, \dot{q}^*, \lambda^*)$, in dem die Zwangsbedingungen nicht differenzierbar sind, so kann die Lösung von (1.18) i. allg. nicht stetig über diesen Punkt hinaus fortgesetzt werden (P^* ist „unpassierbar“, engl.: „Impasse point“ [39, S. 214]). Bei Verwendung des Starrkörperkontaktmodells haben die Bewegungsgleichungen also i. allg. eine nur stückweise stetige Lösung.

Um diese Unstetigkeiten zu überwinden, wird eine Regularisierung der Kontaktbedingung (1.19) eingeführt, die an die Besonderheiten des Rad–Schiene–Kontakts angepaßt ist und u. a. die einseitigen Beschränkungen verwendet, die (1.19) zugrunde liegen. Aus Sicht der praktischen Anwendung ist entscheidend, daß die regularisierten Kontaktbedingungen effizient numerisch ausgewertet werden können. Hierzu werden verschiedene Implementierungen angegeben. Das regularisierte Kontaktmodell wird im MKS–Simulationspaket SIMPACK ([142]) zur dynamischen Simulation von Rad–Schiene–Systemen genutzt und steht in diesem Rahmen für Simulationsaufgaben im Schienenfahrzeugbau zur Verfügung.

Für die in Kapitel 4 betrachtete Verallgemeinerung des Starrkörperkontaktmodells auf Rad–Schiene–Systeme, deren Räder Verschleißprofil haben, treten die Details der Zeitintegration von DA–Systemen, die den Schwerpunkt der Kapitel 2 und 3 bilden, in den Hintergrund. Wesentlich ist hier statt dessen der enge Zusammenhang zwischen der Modellierung, der analytischen Untersuchung der Bewegungsgleichungen, der numerischen Lösung der Bewegungsgleichungen und der Implementierung im Simulationspaket.

Bezeichnungen, Verweise auf das Internet usw.

Die umfangreiche Literatur zu DA–Systemen enthält neben analytischen und numerischen Untersuchungen auch zahlreiche gut dokumentierte Integrationsverfahren und Benchmark–Probleme. In den nachfolgenden Kapiteln werden wir vielfach auf solche aus der Literatur bekannten Ergebnisse verweisen, ohne sie hier im einzelnen wiederzugeben. Die Untersuchung der numerischen Verfahren wird zum großen Teil in Anlehnung an die Monographien von Hairer, Nørsett und Wanner ([82]) und von Hairer und Wanner ([84]) dargestellt.

Den gewachsenen Möglichkeiten der elektronischen Kommunikation wollen wir im Text durch Angabe einiger Verweise auf FORTRAN–Quelltexte u. ä. im Internet Rechnung tragen. Die Verfügbarkeit solcher Dateien ist leider zeitlich befristet (z. B. wegen der Änderung von Internet–Adressen), die angegebenen Adressen entsprechen dem Stand von November 1997.

Kapitel 2

Eine Störungstheorie für differentiell-algebraische Systeme von höherem Index

Während der numerischen Lösung von Anfangswertproblemen entsteht der Diskretisierungsfehler des numerischen Verfahrens, bei der Implementierung des Verfahrens in Gleitpunktarithmetik kommen Rundungsfehler und ggf. Fehler durch Abbruch der iterativen Lösung nichtlinearer Gleichungen hinzu. Sowohl die Konvergenzuntersuchungen von Diskretisierungsverfahren für gewöhnliche Differentialgleichungen als auch ihre praktische Umsetzung im Computerprogramm beruhen u. a. auf der Tatsache, daß die Lösung eines Anfangswertproblems für Systeme gewöhnlicher Differentialgleichungen mit Lipschitzstetiger rechter Seite stetig von Störungen des Anfangswerts und der rechten Seite abhängt ([158, Satz §12.VI]).

Bei der Entwicklung und Untersuchung von Verfahren für DA–Systeme zeigte sich jedoch in (praktisch wichtigen) Spezialfällen, daß man bei einer geeigneten Implementierung der Verfahren Anfangswertprobleme auch dann zuverlässig, effizient und genau numerisch lösen kann, wenn die analytische Lösung *nicht* stetig von kleinen Störungen abhängt. Eine wesentliche Voraussetzung hierfür ist die detaillierte Untersuchung der Sensitivität der analytischen und der numerischen Lösung gegenüber kleinen Störungen. Für nichtlineare DA–Systeme (1.1) wurde dieses Problem erstmals systematisch im Zusammenhang mit der Definition des Störungsindexes ([81, S. 1ff]) betrachtet.

In diesem Kapitel entwickeln wir eine solche Störungstheorie u. a. für Index-2- und Index-3-Systeme in Hessenbergform. Zunächst zeigen wir in Abschnitt 2.1 das Konzept der Störungstheorie und illustrieren das untersuchte Phänomen an einem Testbeispiel. Den Schwerpunkt bildet Abschnitt 2.2, in dem Index-2-Systeme in Hessenbergform betrachtet werden. Hier analysieren wir insbesondere auch den Zusammenhang zwischen Fehlerschranken für die analytische und die numerische Lösung. In Abschnitt 2.3 werden diese Ergebnisse auf Index-3-Systeme erweitert.

Wir konzentrieren unsere Untersuchungen auf diejenigen Fehlerterme, die nicht stetig von Störungen im DA–System abhängen. Für diese Fehlerterme werden Schranken angegeben, die die Struktur des DA–Systems berücksichtigen. Ein weiteres zentrales Ergebnis dieses Kapitels ist der Nachweis, daß diese Fehlerschranken optimal sind und i. allg. nicht

verbessert werden können.

Während die klassische Störungstheorie Fehlerschranken für ein *einzelnes* DA–System bestimmt, benötigt man u. a. bei der Untersuchung von Semidiskretisierungen partieller Differentialgleichungen Fehlerschranken für *Klassen* von DA–Systemen. Hierzu wird im abschließenden Abschnitt 2.4 der Begriff „Störungsindex“ in natürlicher Weise zum gleichmäßigen Störungsindex einer Klasse differentiell-algebraischer Systeme erweitert.

In den nachfolgenden Kapiteln ziehen wir unmittelbar Nutzen aus den Resultaten der Störungstheorie für Index-2- und Index-3-Systeme: Die Ergebnisse von Abschnitt 2.2 sind Grundlage der Konvergenzbeweise für Runge–Kutta–Verfahren und lineare Mehrschrittverfahren (Abschnitt 3.2). Sie motivieren darüberhinaus die Auswahl der Gear–Gupta–Leimkuhler–Formulierung der Modellgleichungen für die dynamische Simulation von mechanischen Mehrkörpersystemen (vgl. Bemerkung 15 in Abschnitt 3.1).

2.1 Motivation

Im Unterschied zur Integration von Systemen gewöhnlicher Differentialgleichungen oder von Index-1-Systemen erweist sich die direkte Übertragung von Integrationsverfahren aus der Theorie der gewöhnlichen Differentialgleichungen auf Anfangswertprobleme für DA–Systeme von höherem Index als kompliziert ([130]). Ursachen hierfür sind einerseits eine mögliche Ordnungsreduktion der Verfahren und algorithmische Details bei der Ordnungs- und Schrittweitensteuerung und bei der Lösung nichtlinearer Gleichungssysteme. Andererseits setzen die Eigenschaften der analytischen Lösung eines DA–Systems von höherem Index *prinzipielle* Grenzen für die Anwendung der aus der Theorie der gewöhnlichen Differentialgleichungen bekannten Verfahren, weil kleine Störungen (Rundungsfehler usw.) während der Integration drastisch verstärkt werden können, so daß die numerischen Ergebnisse unbrauchbar werden.

Ist der (Störungs-)Index des DA–Systems größer als 3, so erfordert die zuverlässige Integration eine analytische Transformation des DA–Systems vor der Diskretisierung (Indexreduktion, vgl. Abschnitt 3.1). I. allg. ist bereits die numerische Lösung von Index-3-Systemen kompliziert.

Beispiel 6 Die Modellgleichungen des starren Radsatzes mit Kegelprofil in Beispiel 27 bilden ein DA–System vom Index 3. Die mit dem Integrator RADAU5 ([84, Anhang]) berechnete numerische Lösung hat *unabhängig* von der vorgegebenen Toleranz TOL einen Fehler der Größe $10^{-3} \dots 10^{-4}$ in *allen* Lösungskomponenten, wenn für die Parameter des Integrators die Standardvorgaben verwendet werden. Ursache dieser großen Fehler im Endergebnis ist die Verstärkung von Fehlern, die in jedem Integrationsschritt durch Abbruch der Newtoniteration bei der Lösung der Runge–Kutta–Gleichungen entstehen. RADAU5 verwendet als Abbruchschranke im vereinfachten Newton–Verfahren $\kappa \cdot \mathrm{TOL}$, Standardvorgabe ist $\kappa = 0.03$ ([84, S. 120f]). Während die durch das Runge–Kutta–Verfahren definierte Lösung q_n die Zwangsbedingungen $g(q_n) = 0$ exakt erfüllt, berechnet RADAU5 nur eine numerische Lösung $\hat{q}_n$ mit $\|g(\hat{q}_n)\| \leq \kappa \cdot \mathrm{TOL}$.
Verringert man daher, wie in der Dokumentation von RADAU5 vorgeschlagen ([81, Kapitel 9]), die Abbruchschranke für das vereinfachte Newtonverfahren, so wird die vorgegebene Genauigkeitsschranke TOL wesentlich besser eingehalten. Abb. 2.1 zeigt links für

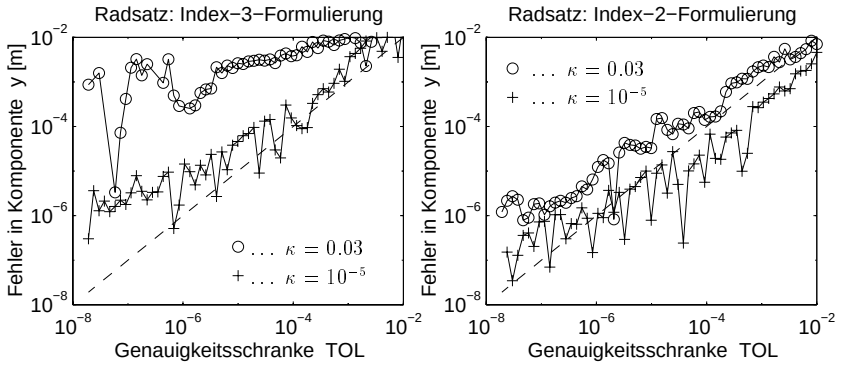


Abbildung 2.1: Fehler der numerischen Lösung bei der dynamischen Simulation eines starren Radsatzes (RADAU5, vgl. Beispiel 6).

verschiedene Genauigkeitsforderungen TOL den Fehler in einer der Lagekoordinaten q (zum Zeitpunkt $T = 5\,\mathrm{s}$) bei $\kappa = 0.03$ („○“, Standardvorgabe) bzw. $\kappa = 10^{-5}$ („+“). Dieser Test zeigt einen Effekt, der für die Integration eines DA–Systems von höherem Index typisch ist. Wendet man nämlich RADAU5 nicht direkt auf die Index-3-Formulierung der Bewegungsgleichungen sondern auf die hierzu analytisch äquivalente Index-2-Formulierung (3.4) an (vgl. Abschnitt 3.1), so wird schon mit der Standardvorgabe $\kappa = 0.03$ die geforderte Genauigkeit erreicht, die Abbruchfehler des Newtonverfahrens sind (wie auch bei der Anwendung von RADAU5 auf gewöhnliche Differentialgleichungen üblich) gegenüber dem Diskretisierungsfehler vernachlässigbar.

Beispiel 6 unterstreicht die Notwendigkeit, für DA–Systeme von höherem Index nicht nur den Diskretisierungsfehler, sondern auch die bei der Implementierung von Diskretisierungsverfahren in Gleitpunktarithmetik auftretenden Fehler zu untersuchen. Diese *Störungstheorie* basiert auf der klassischen Abschätzung für Störungen im Anfangswertproblem

$$y'(t) = f(t, y(t)) \,, \quad (\, t \in [0, T] \,) \,, \quad y(0) = y_0 \tag{2.1}$$

für Systeme gewöhnlicher Differentialgleichungen (vgl. z. B. [158, §12]). Wir setzen voraus, daß (2.1) eine stetig differenzierbare Lösung $y(t)$ hat und daß $f : [0, T] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ in einer Umgebung $\mathcal{U}$ dieser Lösung stetig und bezüglich y Lipschitz-stetig ist, d. h., es gibt eine Konstante L mit

$$\|f(t, y_2) - f(t, y_1)\| \leq L \|y_2 - y_1\| \,, \quad (\, t \in [0, T] \,,\ y_1, y_2 \in \mathcal{U} \,) \,. \tag{2.2}$$

Satz 2 *Sei* $\hat{y} \in C^1([0, T])$ *und*

$$\hat{y}'(t) = f(t, \hat{y}(t)) + \delta(t) \,, \quad (\, t \in [0, T] \,) \,, \quad \hat{y}(0) = \hat{y}_0 \,.$$

Dann gilt für alle $t \in [0,T]$

$$\|\hat{y}(t) - y(t)\| \le e^{Lt} \left(\|\hat{y}_0 - y_0\| + \max_{\tau\in[0,t]} \| \int_0^\tau \delta(w)\, dw \| \right) , \tag{2.3}$$

sofern der Ausdruck auf der rechten Seite von (2.3) hinreichend klein ist.

Beweisidee Man beschränkt sich zunächst auf Funktionen $\hat{y}(t)$, die für $t \in [0,T]$ in $\mathcal{U}$ verbleiben, und verwendet (2.2), um eine Abschätzung für $v(t) := \hat{y}(t) - y(t)$ zu erhalten:

$$\begin{aligned} v(t) &= v(0) + \int_0^t v'(\tau)\, d\tau = v(0) + \int_0^t (f(\tau,\hat{y}(\tau)) - f(\tau,y(\tau)))\, d\tau + \int_0^t \delta(\tau)\, d\tau , \\ \|v(t)\| &\le \|v(0)\| + L\int_0^t \|v(\tau)\|\, d\tau + \| \int_0^t \delta(\tau)\, d\tau \| . \end{aligned}$$

Hieraus folgt mit dem Lemma von Gronwall (s. u.) die Fehlerschranke (2.3). Wie im Beweis von Satz 3 wird anschließend unter Verwendung von (2.3) bewiesen, daß $\hat{y}(t)$ tatsächlich stets in $\mathcal{U}$ verbleibt, wenn die rechte Seite von (2.3) hinreichend klein ist. ∎

Zentrales Beweishilfsmittel ist das *Lemma von Gronwall* (1919), das hier in der in [50] verwendeten Darstellung wiedergegeben wird:

Lemma 1 *Gegeben seien stetige nichtnegative Funktionen* $\psi, \chi : [0,T] \to \mathbb{R}$ *und ein* $\varrho \ge 0$ *mit*

$$\psi(t) \le \varrho + \int_0^t \chi(w)\psi(w)\, dw , \quad (t \in [0,T]) .$$

Dann gilt für alle $t \in [0,T]$

$$\psi(t) \le \varrho \exp(\int_0^t \chi(w)\, dw) .$$

Beweis z. B. [82, Übung I.10.2], [50, Lemma 3.9]. ∎

Satz 2 zeigt, daß für Systeme gewöhnlicher Differentialgleichungen mit Lipschitz-stetiger rechter Seite die Lösung eines Anfangswertproblems stetig von kleinen Störungen abhängt. Für die in der Theorie der gewöhnlichen Differentialgleichungen üblichen Diskretisierungsverfahren kann man entsprechende Fehlerschranken auch für die numerische Lösung beweisen.

Bei der quantitativen Untersuchung des in Beispiel 6 beobachteten Phänomens betrachtet man deshalb die Empfindlichkeit der analytischen Lösung von DA–Systemen (1.1) gegenüber kleinen Störungen $\delta : [0,T] \to \mathbb{R}^{n_x}$, d. h., man sucht für Funktionen $\hat{x} : [0,T] \to \mathbb{R}^{n_x}$ mit

$$F(t,\hat{x}(t),\hat{x}'(t)) = \delta(t) , \quad (t \in [0,T]) , \quad \hat{x}(0) = \hat{x}_0 \tag{2.4}$$

Schranken der Form (2.3) für $\|\hat{x}(t) - x(t)\|$, $(t \in [0,T])$. Die Lösungsdarstellung (1.3) eines linearen Index-m-Systems mit konstanten Koeffizienten zeigt bereits, daß eine solche Schranke neben $\int_0^\tau \delta(w)\, dw$ auch Ableitungen

$$\frac{d^j}{d\tau^j} \int_0^\tau \delta(w)\, dw = \delta^{(j-1)}(\tau) , \quad (j = 1,\ldots,m)$$

dieses Terms enthalten muß.

Hierbei verwenden wir für Funktionen $\delta \in C^r([0,T],\mathbb{R}^n)$ die Bezeichnung

$$\|\delta\|_{C^r([0,t])} := \max_{\tau\in[0,t]} \|\delta(\tau)\| + \max_{\tau\in[0,t]} \|\delta'(\tau)\| + \ldots + \max_{\tau\in[0,t]} \|\delta^{(r)}(\tau)\| , \tag{2.5}$$

ist das Zeitintervall $[0,t]$ aus dem Zusammenhang ersichtlich, so schreiben wir auch kurz $\|\delta\|_{C^r}$ für $\|\delta\|_{C^r([0,t])}$.

Definition 4 Zu einem gegebenen DA–System (1.1) mit Lösung $x : [0,T] \to \mathbb{R}^{n_x}$ betrachte man beliebige Funktionen $\hat{x} : [0,T] \to \mathbb{R}^{n_x}$, die (2.4) erfüllen. Gibt es ein $r \in \mathbb{N}$ und eine Konstante $C > 0$, so daß für $t \in [0,T]$ stets

$$\|\hat{x}(t) - x(t)\| \le C \left(\|\hat{x}_0 - x_0\| + \|\delta\|_{C^r([0,t])} \right) \tag{2.6}$$

gilt, wenn die rechte Seite in (2.6) hinreichend klein ist, und ist r die kleinste natürliche Zahl mit dieser Eigenschaft, so heißt $m = r+1$ *Störungsindex* des DA–Systems (1.1) entlang der Lösung $x(t)$.

Bemerkung 4 a) Dieses an die Störungstheorie angelehnte Indexkonzept wurde in [81] im Zusammenhang mit Konvergenzuntersuchungen für Runge–Kutta–Verfahren eingeführt. Für lineare Systeme mit konstanten Koeffizienten und für DA–Systeme in Hessenbergform fällt der Störungsindex mit dem differentiellen Index zusammen ([73]), i. allg. kann der Störungsindex jedoch beliebig viel größer als der differentielle Index sein ([44]).

b) Wegen der Abschätzung (2.3) wird festgelegt, daß Systeme gewöhnlicher Differentialgleichungen mit Lipschitz-stetiger rechter Seite den Störungsindex 0 haben.

c) Der Störungsindex beschreibt die Sensitivität der analytischen Lösung von DA–Systemen gegenüber kleinen Störungen. I. allg. erwartet man bei Verwendung geeigneter Diskretisierungsverfahren, daß für die numerische Lösung entsprechende Fehlerschranken bewiesen werden können, wobei $\|\delta^{(r)}(\tau)\|$ durch das diskrete Analogon $\frac{1}{h^r}\|\delta_n\|$ zu ersetzen ist ([113], [81]). D. h., bei Diskretisierung eines DA–Systems vom Störungsindex m werden kleine Störungen δ_n (Rundungsfehler, Abbruchfehler) i. allg. mit dem Faktor $\frac{1}{h^{m-1}}$ verstärkt (h bezeichnet die Integrationsschrittweite).

Einer der Schwerpunkte dieses Kapitels ist es, für DA–Systeme von höherem Index diesen Zusammenhang zwischen Fehlerschranken für die analytische und die numerische Lösung zu untersuchen. Wir beschränken uns dabei auf DA–Systeme, von denen aus der Literatur bekannt ist, daß sie durch die direkte Übertragung von Verfahren aus der Theorie der gewöhnlichen Differentialgleichungen zufriedenstellend gelöst werden können. Die Störungstheorie profitiert in vielen Fällen von Ideen, die von Hairer, Lubich und Roche ([81]) in Konvergenzbeweisen für Runge–Kutta–Verfahren entwickelt wurden.

Insbesondere für die numerische Lösung benötigt man möglichst detaillierte Kenntnisse über die Fehlerfortpflanzung und -verstärkung während der Integration, der Begriff „Störungsindex m“ ist hier zu grob. Fehlerabschätzungen

- sollen berücksichtigen, daß man sich in verschiedenen Anwendungen auf Störungen δ von *spezieller Struktur* beschränken kann[1],

[1] Deuflhard und Bornemann [50, Abschnitt 3.1.3] führen hierfür den Störungsindex *bezüglich einer Familie von Störungen* ein.

- sollen für *jede* der Lösungskomponenten möglichst scharfe Schranken angeben und
- sind praktisch nur brauchbar, wenn die Konstante C in (2.6) *nicht zu groß* ist.

Beispiel 7 a) Gegenüber dem Abbruchfehler bei der iterativen Lösung nichtlinearer Gleichungen sind Rundungsfehler beim Auflösen linearer Gleichungssysteme häufig vernachlässigbar. So treten z. B. in der Schaltkreissimulation bei ladungsorientierter Modellierung DA–Systeme auf, für die ein Teil der Zwangsbedingungen explizit nach differentiellen Variablen auflösbar ist (im Beispiel: nach den Ladungen Y [77]). Für $\hat{m} = 1$ und $\hat{m} = 2$ zeigt Günther ([77, Abschnitt 2.4]), daß sich solche Systeme numerisch wie DA–Systeme vom Störungsindex $\hat{m} - 1$ verhalten können, obwohl ihr nach Definition 4 bestimmter Störungsindex $m = \hat{m}$ beträgt.

b) Für Systeme in Hessenbergform ist die Schranke (2.6) i. allg. scharf für die algebraischen Lösungskomponenten. Konvergenzbeweise für numerische Verfahren zeigen, daß der Einfluß kleiner Störungen auf die differentiellen Lösungskomponenten sehr viel geringer ist (vgl. z. B. [81]).

c) Wendet man die Linienmethode zur Lösung partieller Differentialgleichungen mit algebraischen Nebenbedingungen an, so entstehen durch Semidiskretisierung differentiell-algebraische Systeme (1.1) , die von der Ortsdiskretisierung abhängen. I. allg. hängt dann auch die Konstante C in (2.6) von der Ortsdiskretisierung ab, die Fehlerschranke ist hier nur aussagekräftig, wenn C für feiner werdende Ortsdiskretisierung beschränkt bleibt.

Verbesserte Fehlerschranken für einzelne Lösungskomponenten setzen eine spezielle Struktur des DA–Systems voraus. Zur Vereinfachung der Notation beschränken wir uns hier auf Index-2- und Index-3-Systeme in Hessenbergform; ähnliche Abschätzungen können jedoch auch für Systeme mit komplizierterer Struktur bewiesen werden ([14], [15]).

Das Kernstück der Störungstheorie sind die Fehlerabschätzungen für Index-2-Systeme ([28], vgl. auch [10]). Neben den Sätzen 3 und 5, die zeigen, daß die Empfindlichkeit der differentiellen Komponenten gegenüber Störungen entscheidend von der Nichtlinearität des Systems abhängt, illustrieren in Abschnitt 2.2 verschiedene Beispiele den Zusammenhang der Abschätzungen für die analytische und die numerische Lösung.

Auch für Index-3-Systeme bestimmt die nichtlineare Kopplung der Lösungskomponenten die Größe der Fehlerschranken für die differentiellen Komponenten, Abschnitt 2.3 faßt hierzu die Ergebnisse von [13], [17] und [20] zusammen. An 2 Fallstudien wird schließlich in Abschnitt 2.4 die Bedeutung der Größe von C in (2.6) diskutiert ([19]).

2.2 Fehlerschranken für differentiell-algebraische Systeme vom Index 2 in Hessenbergform

In diesem Abschnitt wird die Empfindlichkeit der Lösung von Index-2-Systemen

$$\begin{aligned} y'(t) &= f(y(t), z(t)) \,, \quad (t \in [0,T]) \,, \\ 0 &= g(y(t)) \,, \quad y(0) = y_0 \,, \quad z(0) = z_0 \end{aligned} \tag{2.7}$$

gegenüber kleinen Störungen untersucht und an verschiedenen Beispielen illustriert. Wir setzen dabei stets voraus, daß (2.7) eine hinreichend oft stetig differenzierbare Lösung $y : [0,T] \to \mathbb{R}^{n_y}$, $z : [0,T] \to \mathbb{R}^{n_z}$ hat und daß in einer Umgebung

$$\mathcal{U} := \{\, (\eta, \zeta) \,:\, \|\eta - y(t)\| \le \Delta_y \,,\ \|\zeta - z(t)\| \le \Delta_z \ \text{ für ein } t \in [0,T] \,\}$$

dieser Lösung die Funktionen f und g definiert und hinreichend oft stetig differenzierbar sind ($\Delta_y, \Delta_z > 0$).

Bemerkung 5 a) Im Unterschied zur Theorie gewöhnlicher Differentialgleichungen ist es i. allg. nicht sinnvoll, sich auf *kleine* Umgebungen $\mathcal{U}$ der Lösung zu beschränken: Ist f linear bez. z, so sind $f(\eta,\zeta)$, $g(\eta)$ auch dann auf ganz $\mathcal{U}$ definiert, wenn Δ_z beliebig groß ist. In diesem Fall sind $f(\eta,\zeta)$, $f_y(\eta,\zeta)$, ... nicht gleichmäßig beschränkt, wenn man beliebige $(\eta,\zeta) \in \mathcal{U}$ betrachtet. Wir setzen deshalb nicht für alle $(\eta,\zeta) \in \mathcal{U}$, sondern nur für alle Punkte $(\eta, z(t)) \in \mathcal{U}$ voraus, daß f, g und die Ableitungen dieser Funktionen durch eine Konstante $L = \mathcal{O}(1)$ beschränkt bleiben und daß die *Index-2-Bedingung*

$$[g_y f_z](\eta, z(t)) \quad \text{ist regulär} \,, \quad \|([g_y f_z](\eta, z(t)))^{-1}\| \le L < \infty \tag{2.8}$$

erfüllt ist. Schließlich sollen die Ableitungen von f_z für *alle* $(\eta,\zeta) \in \mathcal{U}$ beschränkt sein durch

$$\|f_{yz}(\eta,\zeta)\| \le L \,, \quad \|f_{zz}(\eta,\zeta)\| \le \mu \,, \quad ((\eta,\zeta) \in \mathcal{U}) \tag{2.9}$$

mit einer Konstanten $\mu \le L$.

b) Ist f linear bez. z, so ist $\mu = 0$. Ein typisches Beispiel für Systeme (2.7) mit $0 < \mu \ll 1$ ist die in Abschnitt 3.1 betrachtete Gear–Gupta–Leimkuhler–Formulierung (3.6) von Modellgleichungen für mechanische Mehrkörpersysteme, in denen kleine Reibungskräfte $\mu_0 \|F_N\|_2^2 \cdot \psi(q,q')$ wirken. Hier bezeichnet $F_N := -G^T(q)\lambda$ die Zwangskräfte und ψ eine von den Lage- und Geschwindigkeitskoordinaten abhängige Funktion; der Parameter μ_0 ist der Reibungskoeffizient. In der Notation (2.7) hat f für diese Modellgleichungen die Form $f(y,z) = f_0(y) + \mu_0 z^T f_1(y) z \cdot f_2(y) + f_3(y) z$, die Lagrangeschen Multiplikatoren λ sind Teil der algebraischen Variablen z. Es gilt also $\mu = \mathcal{O}(\mu_0)$ in (2.9) und die Voraussetzungen von a) sind mit $\Delta_z = \mathcal{O}(\frac{1}{\mu})$ erfüllt. Für $\mu \to 0$ konvergieren die Fehlerschranken der Sätze 3 und 5 gegen die entsprechenden Abschätzungen für Systeme, die bez. z linear sind (und für die deshalb $\mu = 0$ gilt).

Bemerkung 6 In der Literatur werden *verbesserte* Fehlerschranken für die differentiellen Komponenten y in (2.7) im Zusammenhang mit dem Konvergenzbeweis für implizite Runge–Kutta–Verfahren angegeben. Hierzu wird (2.7) entlang der analytischen Lösung $(y(t), z(t))$ linearisiert ([81, Satz 4.4]). Durch die zusätzliche Berücksichtigung von Termen 2. Ordnung (f_{zz}) gelingt es in den Sätzen 3 und 5, *optimale* Fehlerschranken für y zu beweisen. Die Beweise der Sätze 3 und 5 sind nicht nur wegen der Terme höherer Ordnung technisch aufwendig, sondern vor allem auch wegen der schwachen Voraussetzungen an die Größe der Störungen im DA–System.

Die Beschränkung auf autonome Systeme (2.7) dient der Vereinfachung der Schreibweise. Die Ergebnisse lassen sich unmittelbar auf nichtautonome Systeme übertragen (durch die übliche Erweiterung des Systems um die Differentialgleichung $t' = 1$).

2.2.1 Die Sensitivität der analytischen Lösung gegenüber kleinen Störungen

Zum DA–System (2.7) seien nun Funktionen $(\hat{y}(t), \hat{z}(t))$ gegeben, für die

$$\begin{aligned} \hat{y}'(t) &= f(\hat{y}(t), \hat{z}(t)) + \delta(t) \,, \quad (t \in [0,T]) \,, \\ \theta(t) &= g(\hat{y}(t)) \,, \quad \hat{y}(0) = \hat{y}_0 \,, \quad \hat{z}(0) = \hat{z}_0 \end{aligned} \tag{2.10}$$

mit Residuen $\delta \in C^0([0,T])$ und $\theta \in C^1([0,T])$ gilt. Aus der Literatur ([81, S. 4], [84, S. 459f]) ist bekannt, daß (2.7) den Störungsindex 2 hat, es gilt

$$\|\hat{y}(t) - y(t)\| + \|\hat{z}(t) - z(t)\| \le C \Big(\|\hat{y}_0 - y_0\| + \|\delta\|_{C^0([0,t])} + \|\theta\|_{C^1([0,t])} \Big) \tag{2.11}$$

mit einer von den Störungen unabhängigen Konstanten C, sofern die rechte Seite in (2.11) hinreichend klein ist.

Für das Beispiel linearer Systeme (2.7) mit konstanten Koeffizienten können die Differenzen $\hat{y}(t) - y(t)$ und $\hat{z}(t) - z(t)$ explizit angegeben werden ([28, Beispiel 1]). Man erkennt, daß die Abschätzung (2.11) für die algebraischen Komponenten z scharf ist. Insbesondere kann $\hat{z}(t) - z(t)$ auch dann beliebig groß werden, wenn $\|\delta\|_{C^0([0,T])}$ und $\|\theta\|_{C^0([0,T])}$ sehr klein sind, z hängt (bezüglich $\|\cdot\|_{C^0}$) nicht stetig von Störungen θ der algebraischen Gleichungen ab.

In die differentiellen Komponenten y geht dagegen der kritische Fehlerterm θ' nicht direkt, sondern nur über die Kopplung mit z ein. Der tatsächliche Einfluß von θ' auf y ist deshalb sehr viel geringer als in (2.11) angegeben und hängt insbesondere von der Struktur von f ab. Satz 3 gibt scharfe Fehlerschranken für die differentiellen Komponenten, die die Struktur von f berücksichtigen. Hierbei wird das Residuum $\delta(t)$ aufgespaltet in $\delta(t) = P(t)\delta(t) + (I - P(t))\delta(t)$ mit der Projektorfunktion

$$P(t) := I - [f_z(g_y f_z)^{-1} g_y](y(t), z(t)) \,. \tag{2.12}$$

Die Funktion $P(t)$ projiziert Vektoren $\eta \in \mathbb{R}^{n_y}$ auf den Tangentialraum der Mannigfaltigkeit $\{\, \eta : g(\eta) = 0 \,\}$ im Punkt $y(t)$ ([81, S. 35]), d. h., $P(t)\delta(t)$ ist die Projektion von $\delta(t)$ in diesen Tangentialraum.

Um $(\hat{y}(t), \hat{z}(t)) \in \mathcal{U}$, $(t \in [0,T])$ zu garantieren, wird in Satz 3 vorausgesetzt, daß $\overline{\mu} D(t)$ hinreichend klein ist. Dabei ist

$$D(t) := \|\delta\|_{C^0([0,t])} + \|\theta\|_{C^1([0,t])} \tag{2.13}$$

und $\overline{\mu} := \mu + \frac{1}{\Delta_z}$, d. h., in praxi ist i. allg. $\overline{\mu} = \mathcal{O}(\mu)$ (vgl. Bemerkung 5b); für Systeme, die linear in z sind, wird $\overline{\mu} := \mu = 0$ gesetzt.

Satz 3 *a) Es gibt eine Konstante C, so daß für alle Funktionen $(\hat{y}, \hat{z})$ mit (2.10), $\delta \in C^0([0,T])$, $\theta \in C^1([0,T])$ und $\|\delta\|_{C^0([0,T])} \le C_\delta$ die Abschätzung*

$$\|\hat{y}(t) - y(t)\| \le C \Big(\|\hat{y}_0 - y_0\| + \int P\delta + \|\theta\|_{C^0([0,t])} + \|\theta\|_{C^0([0,t])} \cdot D(t) + \mu D(t)^2 \Big) \tag{2.14}$$

mit

$$\int P\delta := \max_{\tau \in [0,t]} \| \int_0^\tau P(w)\delta(w)\, dw \|$$

für alle $t \in [0,T]$ erfüllt ist, wenn $\overline{\mu} D(t)$ und die rechte Seite in (2.14) hinreichend klein sind (C ist unabhängig von $\hat{y}$, $\hat{z}$, μ und den Störungen δ und θ und wird i. allg. durch L, T und C_δ bestimmt).

b) Ist $f(\eta, \zeta) = f_0(\eta) + f_z \cdot \zeta$ und $f_z \equiv const$, so gilt (2.14) mit $D(t) := 0$.

Beweis a) Zum Beweis von Teil a) wird das Lemma von Gronwall auf

$$v(t) := \Phi(\hat{y}(t)) - \Phi(y(t)) \quad \text{mit} \quad \Phi(\eta) := \eta - [f_z(g_y f_z)^{-1} g](\eta, z(t)) \tag{2.15}$$

angewendet. Hierbei setzen wir zu Beginn zusätzlich voraus, daß für alle $\tau \in [0,t] \subset [0,T]$ gilt

$$\|\hat{y}(\tau) - y(\tau)\| + \mu \|\hat{z}(\tau) - z(\tau)\| \le \gamma \,, \quad \|\hat{z}(\tau) - z(\tau)\| \le \Delta_z \tag{2.16}$$

mit einer hinreichend kleinen positiven Konstanten γ. Den Nachweis, daß (2.16) stets erfüllt ist, wenn $\overline{\mu} D(t)$ und die rechte Seite in (2.14) hinreichend klein sind, führen wir am Ende von Beweisteil a).

Es gilt

$$v(t) = v(0) + \int_0^t \frac{d}{d\tau} v(\tau)\, d\tau \tag{2.17}$$

und

$$\begin{aligned} \frac{d}{d\tau} v(\tau) &= \Phi_\eta(\hat{y}(\tau))\hat{y}'(\tau) - \frac{\partial}{\partial z}[(f_z(g_y f_z)^{-1})(\hat{y}, z)](g(\hat{y}), z'(\tau)) - \Phi_\eta(y(\tau))y'(\tau) \\ &= (I - [f_z(g_y f_z)^{-1} g_y](\hat{y}, z))\hat{y}'(\tau) - \frac{\partial}{\partial y}[(f_z(g_y f_z)^{-1})(\hat{y}, z)](g(\hat{y}), \hat{y}'(\tau)) + \\ &\quad + \mathcal{O}(\|\theta(\tau)\|) - P(\tau)y'(\tau) \end{aligned} \tag{2.18}$$

(denn $g(y(\tau)) = 0$ und $g(\hat{y}(\tau)) = \theta(\tau)$). Seien $\phi_{ik} = \phi_{ik}(y,z)$ die Elemente der Matrix $[f_z(g_y f_z)^{-1}](y,z)$, dann ist der in (2.18) verwendete Tensor definiert durch die Darstellung

$$\frac{\partial}{\partial y}[(f_z(g_y f_z)^{-1})(y,z)](\zeta, \eta) := \Big(\sum_{j=1}^{n_y} \sum_{k=1}^{n_z} \frac{\partial \phi_{ik}}{\partial y_j} \zeta_k \eta_j \Big)_{i=1}^{n_y} \,, \tag{2.19}$$

die für beliebige Vektoren $\zeta = (\zeta_k)_k \in \mathbb{R}^{n_z}$ und $\eta = (\eta_j)_j \in \mathbb{R}^{n_y}$ gelten soll.
Da wir im Unterschied zu den klassischen Abschätzungen in [81], [84] nicht voraussetzen, daß $\hat{y}'(\tau) = f(\hat{y}, \hat{z}) + \delta(\tau)$ durch eine Konstante der Größe $\mathcal{O}(1)$ beschränkt ist, erfordert dieser Term besondere Aufmerksamkeit:

$$\begin{aligned} \hat{y}'(\tau) &= f(\hat{y}, z) + f_z(\hat{y}, z)(\hat{z} - z) + f(\hat{y}, \hat{z}) - f(\hat{y}, z) - f_z(\hat{y}, z)(\hat{z} - z) + \delta(\tau) \\ &= f(\hat{y}, z) + f_z(\hat{y}, z)(\hat{z} - z) + \mathcal{O}(\mu)\|\hat{z} - z\|^2 + \delta(\tau) \end{aligned} \tag{2.20}$$

(dies folgt aus $\|\Psi(1) - \Psi(0) - \Psi'(0)\| = \| \int_0^1 (\Psi'(\vartheta) - \Psi'(0))\, d\vartheta \| \le \frac{1}{2} \max_{\vartheta \in (0,1)} \|\Psi''(\vartheta)\|$ für die Funktion $\Psi(\vartheta) := f(\hat{y}, z + \vartheta(\hat{z} - z))$). Wegen (2.20) gilt

$$\hat{y}'(\tau) = \mathcal{O}(1) + \mathcal{O}(1)\|\hat{z}(\tau) - z(\tau)\| + \mathcal{O}(\|\delta(\tau)\|) \,, \tag{2.21}$$

$$\begin{aligned} &(I - [f_z(g_y f_z)^{-1} g_y](\hat{y}, z))\hat{y}'(\tau) = \\ &\quad = (I - [f_z(g_y f_z)^{-1} g_y](\hat{y}, z))(f(\hat{y}, z) + \delta(\tau)) + \mathcal{O}(\mu)\|\hat{z} - z\|^2 \,, \\ &\quad = P(\tau)f(y, z) + P(\tau)\delta(\tau) + \mathcal{O}(1 + \|\delta(\tau)\|)\|\hat{y} - y\| + \mathcal{O}(\mu)\|\hat{z} - z\|^2 \end{aligned} \tag{2.22}$$

und

$$\begin{aligned}\theta'(\tau) &= \frac{d}{d\tau}g(\hat{y}(\tau)) = g_y(\hat{y})\hat{y}'(\tau)\\ &= [g_y f](\hat{y},z) + [g_y f_z](\hat{y},z)(\hat{z}-z) + \mathcal{O}(\mu)\|\hat{z}-z\|^2 + g_y(\hat{y})\delta(\tau)\ .\end{aligned}$$

Mit (2.16) ist $\mu\|\hat{z}-z\|^2 \ll \|\hat{z}-z\|$, so daß diese Gleichungen wegen der Index-2-Bedingung (2.8) nach $\hat{z}-z$ aufgelöst werden können:

$$\hat{z}(\tau)-z(\tau) = \mathcal{O}(1)(\|[g_y f](\hat{y},z)\| + \|\delta(\tau)\| + \|\theta'(\tau)\|) = \mathcal{O}(1)(\|\hat{y}(\tau)-y(\tau)\| + D(\tau))\ , \quad (2.23)$$

denn $[g_y f](\hat{y},z) = [g_y f](y,z) + \mathcal{O}(1)\|\hat{y}-y\|$ und $[g_y f](y,z) = g_y(y)y' = \frac{d}{d\tau}g(y(\tau)) = 0$. Deshalb ist in (2.22) der Term $\mu\|\hat{z}-z\|^2$ beschränkt durch

$$\mu\|\hat{z}-z\|^2 = \mathcal{O}(\mu)\|\hat{y}-y\|^2 + \mathcal{O}(\mu D(\tau))\|\hat{y}-y\| + \mathcal{O}(\mu D(\tau)^2) = \mathcal{O}(1)\|\hat{y}-y\| + \mathcal{O}(\mu D(\tau)^2) \quad (2.24)$$

denn $\|\hat{y}-y\| \le \gamma < 1$ und $\overline{\mu}D(\tau) \ll 1$.
Setzt man (2.23) und (2.24) in (2.21) bzw. (2.22) ein, so folgt in (2.18)

$$\frac{d}{d\tau}v(\tau) = P(\tau)\delta(\tau) + \mathcal{O}(1)(\|\hat{y}-y\| + \|\theta(\tau)\| + \|\theta(\tau)\|\cdot D(\tau) + \mu D(\tau)^2)\ .$$

Dieser Ausdruck und

$$\hat{y}(\tau) - y(\tau) = v(\tau) + \mathcal{O}(1)\|g(\hat{y}(\tau))\| = v(\tau) + \mathcal{O}(1)\|\theta(\tau)\|$$

werden in (2.17) verwendet, um mit dem Lemma von Gronwall die Abschätzung (2.14) zu zeigen, wobei die Konstante C unabhängig von der Konstanten γ aus (2.16) ist.
Schließlich wird mittels vollständiger Induktion bewiesen, daß unter den Voraussetzungen des Satzes stets die Abschätzung (2.16) gilt: Hierzu bezeichnen wir mit err_y die rechte Seite von (2.14) für $t = T$. Sei nun die Abschätzung (2.16) erfüllt für $\tau \in [0,t] \subset [0,T]$. Dann gilt wegen (2.14) und (2.23)

$$\begin{aligned}\|\hat{y}(t)-y(t)\| + \mu\|\hat{z}(t)-z(t)\| &\le C^*(err_y + \mu D(t))\\ \|\hat{z}(t)-z(t)\| &\le \Delta_z \cdot C^*(\frac{err_y}{\Delta_z} + \frac{1}{\Delta_z}D(t))\ .\end{aligned}$$

mit einer Konstanten C^*. Sind err_y und $\overline{\mu}D(t)$ so klein, daß $C^*(err_y + \mu D(T)) \le \gamma/2$ und $C^*(\frac{err_y}{\Delta_z} + \frac{1}{\Delta_z}D(T)) \le 1/2$ ist, dann gibt es also wegen der gleichmäßigen Stetigkeit von y, z, $\hat{y}$, $\hat{z}$ eine (von t unabhängige) Konstante $\Delta_t > 0$, so daß (2.16) auch für alle $\tau \in [0, t+\Delta_t] \cap [0,T]$ erfüllt ist. Hieraus folgt (2.16) sukzessive für alle $\tau \in [0,T]$.

b) Zum Beweis von b) wird die Fehlerfortpflanzung in der Projektion $\Phi(t,\eta) := P(t)\cdot\eta$ untersucht. Wegen

$$\int_0^1 g_y(y+\vartheta(\eta-y))(\eta-y)\,d\vartheta = g(y+1\cdot(\eta-y)) - g(y+0\cdot(\eta-y)) = g(\eta)$$

gilt in einer Umgebung der exakten Lösung

$$\begin{aligned}P(t)(\eta-y(t)) = \eta - y(t) - [f_z(g_y f_z)^{-1}](y(t))\cdot\\ \cdot\Big(\int_0^1 (g_y(y) - g_y(y+\vartheta(\eta-y)))d\vartheta\cdot(\eta-y(t)) + g(\eta)\Big)\end{aligned} \quad (2.25)$$

und man erhält für $\Phi(t,\eta) - \Phi(t,y(t)) = P(t)(\eta - y(t))$ die Abschätzung

$$\|\Phi(t,\eta) - \Phi(t,y(t))\| = \|\eta - y(t)\| + \mathcal{O}(\|\eta - y(t)\|^2) + \mathcal{O}(\|g(\eta)\|)\ . \quad (2.26)$$

Außerdem folgt aus $P(t)f_z = (I - [f_z(g_y f_z)^{-1}g_y](y(t)))f_z \equiv 0$ die Identität

$$[\Phi_\eta f_z](t,\eta) \equiv 0\ , \quad (2.27)$$

so daß die totalen Ableitungen von $\Phi(t,y(t))$ und $\Phi(t,\hat{y}(t))$ bezüglich t nicht von den algebraischen Komponenten z abhängen. Verwendet man

$$\begin{aligned}&\frac{d}{d\tau}(\Phi(\tau,\hat{y}(\tau)) - \Phi(\tau,y(\tau))) =\\ &\quad= \Phi_t(\tau,\hat{y}(\tau)) - \Phi_t(\tau,y(\tau)) + \Phi_\eta(\tau,\hat{y}(\tau))\hat{y}'(\tau) - \Phi_\eta(\tau,y(\tau))y'(\tau)\\ &\quad= \Phi_t(\tau,\hat{y}) - \Phi_t(\tau,y) + [\Phi_\eta f_0](\hat{y}) - [\Phi_\eta f_0](y) + \Phi_\eta(\tau,\hat{y})\delta(\tau)\end{aligned}$$

und $\Phi_\eta(\tau,\hat{y})\delta(\tau) = P(\tau)\delta(\tau)$, so folgt die Behauptung analog zum Beweisteil a). ∎

Bemerkung 7 Die gegenüber (2.11) verbesserten Fehlerschranken für y sind wesentlich, weil während der Integration von $t=0$ zu $t=T$ nur in diesen Komponenten Fehler übertragen werden. Die Komponenten z sind durch $0 = \frac{d}{dt}g(y(t)) = [g_y f](y,z)$ bestimmt. Wie im Konvergenzbeweis für implizite Runge–Kutta–Verfahren in Hairer et al. ([81]) wird die Fehlerfortpflanzung nicht nur separat für differentielle und algebraische Komponenten betrachtet, sondern außerdem innerhalb der differentiellen Komponenten getrennt nach Anteilen *tangential* zur Mannigfaltigkeit $\{\eta : g(\eta) = 0\}$ (also in der Projektion $P(t)(\hat{y}(t) - y(t))$) und *orthogonal* zur Mannigfaltigkeit (diese Komponente ist durch $g(\hat{y}(t)) = \theta(t)$ bestimmt). Während jedoch in [81] $f(\hat{y},\hat{z})$ entlang (y,z) linearisiert wird, erlaubt die (beweistechnisch aufwendigere) Linearisierung von $f(\hat{y},\hat{z})$ entlang $(\hat{y},z)$ im Beweis von Satz 3 den Nachweis von (scharfen) Fehlerschranken für y nicht nur für Systeme (2.7) mit $\mu = \mathcal{O}(1)$, sondern auch im Fall $0 \le \mu \ll 1$.

Obwohl die Fehlerschranke in Satz 3 deutlich kleiner ist als (2.11), werden die Komponenten y i. allg. trotzdem von Ableitungen der Störungen θ beeinflußt (vgl. Abschnitt 2.2.3). Nur für Systeme (2.7) mit sehr spezieller Struktur hängen die Komponenten y stetig von Störungen im System ab. Neben Satz 3b zeigen wir ein solches Ergebnis für Systeme (2.7), die linear bez. z sind und für die die Zwangsbedingung eine Hyperfläche $\{\eta : g(\eta) = 0\}$ definiert. Im Unterschied zu Satz 3 hat dieses sehr spezielle Ergebnis *kein* Gegenstück für die numerische Lösung (vgl. Beispiel 10).

Folgerung 1 *Die differentiellen Komponenten y hängen stetig von kleinen Störungen des Systems (2.7) ab, wenn f bez. z linear ist (d. h. $f(\eta,\zeta) = f_0(\eta) + f_z(\eta)\cdot\zeta$) und der algebraische Teil von (2.7) aus einer skalaren Gleichung $g = 0$ besteht (d. h. $n_z = 1$).*

Beweis Zum Beweis betrachten wir eine von Wensch ([161]) in anderem Zusammenhang eingeführte Funktion $\Phi : \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$, die definiert ist durch $\Phi(\eta) := \tilde{\eta}(1;\eta)$, wobei $\tilde{\eta}(\vartheta;\eta)$ die Lösung des Anfangswertproblems

$$\frac{d}{d\vartheta}\tilde{\eta}(\vartheta;\eta) = -[f_z(g_y f_z)^{-1}](\tilde{\eta}(\vartheta;\eta))\cdot g(\eta)\ , \quad (\vartheta \in [0,1])\ , \quad \tilde{\eta}(0;\eta) = \eta$$

bezeichnet. Man kann zeigen, daß Φ in einer Umgebung der analytischen Lösung von (2.7) wohldefiniert ist und die Bedingungen (2.26) und (2.27) erfüllt, so daß die Behauptung wie im Teil b) des Beweises von Satz 3 folgt.
Der Nachweis von (2.26) und (2.27) wird in [28, Folgerung 1] geführt, er ist technisch aufwendig und nutzt insbesondere die Voraussetzung $n_z = 1$ explizit aus. ∎

2.2.2 Die Sensitivität der numerischen Lösung gegenüber kleinen Störungen

Die Fehlerschranken (2.11) und (2.14) für Index-2-Systeme enthalten $\|\theta\|_{C^1}$ und unterscheiden sich deshalb *prinzipiell* von der klassischen Störungstheorie für gewöhnliche Differentialgleichungen (Satz 2). Als Gegenstück zu den Termen $\|\theta\|_{C^0}\|\theta\|_{C^1} + \mu\|\theta\|_{C^1}^2$ bzw. $\|\theta\|_{C^1}$ in (2.14) und (2.11) erwarten wir daher für die numerische Lösung neben dem Diskretisierungsfehler und Fehlern der Größenordnung von $\delta := \max_n \|\delta_n\|$ und $\theta := \max_n \|\theta_n\|$ einen (gegenüber der Theorie der gewöhnlichen Differentialgleichungen) *zusätzlichen Fehlerterm* $\mathcal{O}(\theta \cdot \frac{1}{h}\theta + \mu(\frac{1}{h}\theta)^2)$ in den differentiellen Komponenten y und $\mathcal{O}(\frac{1}{h}\theta)$ in den algebraischen Lösungskomponenten z (unabhängig vom konkreten Diskretisierungsverfahren). Hier bezeichnet h die Integrationsschrittweite und δ_n und θ_n die im n-ten Integrationsschritt auftretenden Störungen im differentiellen und im algebraischen Teil (Rundungsfehler, Abbruchfehler bei der iterativen Lösung nichtlinearer Gleichungen usw.).

Beispiel 8 Um den quantitativen Einfluß von Störungen zu verdeutlichen, werden BDF mit konstanter Schrittweite h auf das nichtautonome Index-2-System

$$\begin{aligned} y_1' &= 4y_1^2 - y_2^2 + z + \frac{1}{2}(1-\mu_0)(\sqrt{1-y_1^2} - z) \\ y_2' &= (y_1 y_2 + 2(1-\mu_0)(1-y_1^2) + 2\mu_0 z^2)z \\ 0 &= 4y_1 + y_2 - 6\sin t \quad , \qquad t \in [0.5, 1] \end{aligned}$$

angewendet, die Anfangswerte entnimmt man der exakten Lösung $y_1(t) = \sin t$, $y_2(t) = 2\sin t$, $z(t) = \cos t$ (der Parameter μ_0 bestimmt die Größe von μ in (2.9), $\mu = \mathcal{O}(\mu_0)$). Während alle anderen Berechnungen in `IEEE extended`–Arithmetik (10 Byte, Maschinengenauigkeit $< 10^{-19}$) ausgeführt werden, wird in den algebraischen Gleichungen künstlich eine Störung eingeführt, indem man $\sin t$ nur mit einfacher Genauigkeit (4 Byte) berechnet (damit sind die Störungen in den algebraischen Gleichungen von der Größenordnung $\theta = 6 \cdot \epsilon_{\text{single}} \approx 1.8_{10}\text{-}7$).
Abb. 2.2 zeigt links für $\mu_0 = 1$ den maximalen Fehler des impliziten Eulerverfahrens (d. h. der BDF mit $k = 1$) in den Komponenten y_1 („+“) und z („∘“), jeweils aufgetragen gegenüber h. Je nach Größe von h dominieren die Diskretisierungsfehler $\mathcal{O}(h)$ oder die zusätzlichen Fehlerterme $\mathcal{O}(\frac{1}{h^2}\theta^2)$, $\mathcal{O}(\frac{1}{h}\theta)$, wegen der logarithmischen Skaleneinteilung haben die Graphen deshalb die Anstiege $+1$, -2 bzw. -1. Wiederholt man beginnend mit $h_0 = 0.1$ die Integration für immer kleineres h, so reduziert sich zunächst der Gesamtfehler sowohl in y als auch in z (Bereich A), dann nur noch in y (Bereich B) und schließlich wächst der Gesamtfehler sowohl in y als auch in z rasch an. Dabei wächst der Fehler in y

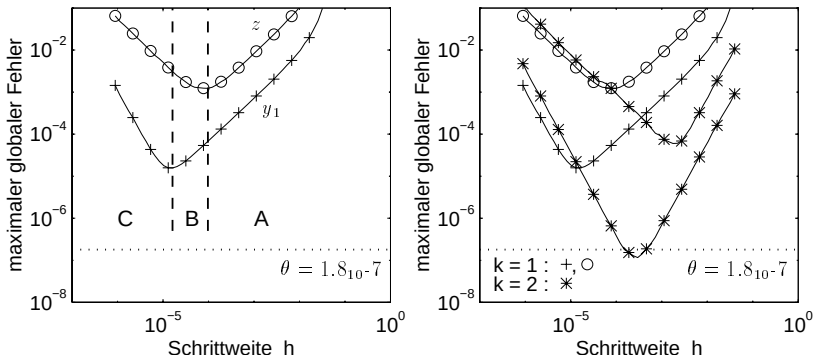


Abbildung 2.2: Gesamtfehler bei Anwendung der BDF auf ein Index-2-System (vgl. Beispiel 8).

schneller, bleibt jedoch trotzdem deutlich kleiner als der Fehler in z (Bereich C).
Qualitativ gleiche Ergebnisse erhält man für Verfahren höherer Ordnung: Abb. 2.2 zeigt rechts neben den Ergebnissen für das Eulerverfahren („+“ und „∘“ wie oben) die entsprechenden Graphen für BDF 2. Ordnung („∗“). Der Diskretisierungsfehler hat hier die Größenordnung $\mathcal{O}(h^2)$ (entspricht in Abb. 2.2 Geraden mit Anstieg $+2$), er ist sehr viel kleiner als für das Eulerverfahren. Die zusätzlichen Fehlerterme sind jedoch für beide Verfahren nahezu identisch.
Durch Variation von μ_0 wird schließlich die Bedeutung der Struktur von f für die Größe des Fehlers in y deutlich: Abb. 2.3 auf S. 30 zeigt für verschiedene μ_0 die Fehler des impliziten Eulerverfahrens in den Komponenten y_1 und z. Für große Schrittweiten h überwiegt der Diskretisierungsfehler, die Größe des Diskretisierungsfehlers variiert geringfügig mit der Größe von μ_0. Für kleine Schrittweiten h ist der Fehlerterm in den algebraischen Komponenten z, der in der Abschätzung (2.11) dominiert, weitgehend unabhängig von μ_0. Dagegen ist der zusätzliche Fehlerterm in y für kleines μ_0 sehr viel kleiner als für $\mu_0 = 1$. Ist das System linear bezüglich z (d. h. $\mu_0 = 0$), so überschreitet der Fehler in y selbst für sehr kleine h nicht die Größe der Störungen θ (vgl. jedoch Beispiel 10 für ein Index-2-System (2.7) mit $\mu = 0$, in dem der sehr kleine zusätzliche Fehlerterm $\mathcal{O}(\theta \cdot \frac{1}{h}\theta)$ auch numerisch nachgewiesen werden kann).

Der Einfluß von Störungen auf die numerische Lösung von Index-2-Systemen wird in der Literatur im Zusammenhang mit der Untersuchung der Konvergenz von Diskretisierungsverfahren betrachtet (vgl. z. B. [39], [81] und Abschnitt 3.2 der vorliegenden Arbeit). Dabei ist es üblich, Störungen der Größenordnung $\delta = \mathcal{O}(h)$ und $\theta = \mathcal{O}(h^2)$ zu betrachten. In Abb. 2.2 ist diese Voraussetzung nur im Bereich A erfüllt.

Dagegen ist der unten angegebene Satz 5 das direkte Analogon zu Satz 3 und erklärt *vollständig* das Verhalten des Gesamtfehlers in den Abb. 2.2 und 2.3. Dabei ist der Nachweis dieser Fehlerschranken wegen der Details der Diskretisierung beweistechnisch

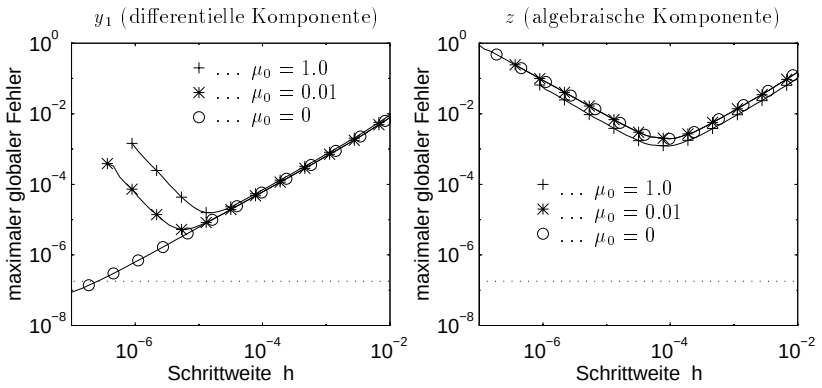


Abbildung 2.3: Gesamtfehler des impliziten Eulerverfahrens in Abhängigkeit von der Konstanten μ aus (2.9) (vgl. Beispiel 8).

noch aufwendiger als in Satz 3, weshalb wir uns in diesem Abschnitt auf das implizite Eulerverfahren beschränken. Die Übertragung der Ergebnisse auf mehrstufige implizite Runge–Kutta–Verfahren ([10]) und auf BDF ([15]) ist möglich.

Das implizite Eulerverfahren

$$\begin{aligned} \frac{y_{n+1}-y_n}{h} &= f(y_{n+1},z_{n+1}) \\ 0 &= g(y_{n+1}) \end{aligned} \tag{2.28}$$

wurde erstmals von Gear ([70]) auf DA–Systeme angewendet, es hat — wie bei der Integration von gewöhnlichen Differentialgleichungen — die Konvergenzordnung 1 ([104]). In jedem Integrationsschritt ist ein nichtlineares Gleichungssystem zu lösen, das unter der Voraussetzung $\|g(y_n)\| + h\|[g_y f](y_n,z_n)\| = \mathcal{O}(h^2)$ lokal eindeutig die Vektoren y_{n+1} und z_{n+1} bestimmt ([81, Satz 4.1]). Mit den Voraussetzungen aus Bemerkung 5 kann ein ähnliches Ergebnis auch unter schwächeren Bedingungen an (y_n,z_n) bewiesen werden:

Satz 4 *Gegeben seien (y_n,z_n) mit*

$$\|y_n - y(t_n)\| + \mu\|z_n - z(t_n)\| \le \gamma \quad , \quad \|z_n - z(t_n)\| \le \frac{1}{2}\Delta_z \quad , \quad \Delta + \overline{\mu}\frac{\Delta}{h} \le \Delta_0$$

und $\Delta := \|g(y_n)\| + h\|[g_y f](y_n,z_n)\|$. Sind die (positiven) Konstanten h_0, Δ_0 und γ hinreichend klein, so ist (2.28) für alle $h \in (0,h_0]$ (lokal) eindeutig nach (y_{n+1},z_{n+1}) auflösbar, und es gilt

$$\|y_{n+1} - y_n\| \le C_{h\Delta}(h+\Delta) \; , \quad \|z_{n+1} - z_n\| \le C_{h\Delta}(h + \frac{\Delta}{h}) \tag{2.29}$$

mit einer von h, Δ, μ und (y_n,z_n) unabhängigen Konstanten $C_{h\Delta}$.

Beweis Zum nichtlinearen Gleichungssystem (2.28) induziert das vereinfachte Newtonverfahren mit der Matrix

$$J = \begin{pmatrix} \frac{1}{h}I & -f_z(y_n,z_n) \\ g_y(y_n) & 0 \end{pmatrix}$$

eine Fixpunktabbildung

$$\begin{pmatrix} y \\ z \end{pmatrix} \mapsto \Phi(y,z) := \begin{pmatrix} y \\ z \end{pmatrix} - J^{-1}\begin{pmatrix} \frac{1}{h}(y-y_n) - f(y,z) \\ g(y) \end{pmatrix} \; .$$

In Teil a) des Beweises wird gezeigt, daß für den Startwert $y^{(0)} := y_n$, $z^{(0)} := z_n$ das Inkrement des ersten Newtonschritts durch

$$\|\Phi(y_n,z_n) - \begin{pmatrix} y_n \\ z_n \end{pmatrix}\|_\sigma \le \frac{1}{2}C_1(h+\Delta) \tag{2.30}$$

beschränkt ist. Hier bezeichnet C_1 eine von h und Δ unabhängige Konstante und $\|\cdot\|_\sigma$ eine geeignet skalierte Norm. Bezüglich dieser Norm ist für alle

$$(y,z) \in \mathcal{U}_h := \{\,(y,z) \;:\; \|y-y_n\| + \sigma\|z-z_n\| \le C_1(h+\Delta)\,\} \tag{2.31}$$

die Kontraktivitätsbedingung $\|\Phi'(y,z)\|_\sigma \le \frac{1}{2}$ erfüllt (Teil b) des Beweises), so daß die Folge der Newtoniterierten wegen (2.30) in $\mathcal{U}_h$ verbleibt und gegen eine Lösung von (2.28) konvergiert, die in $\mathcal{U}_h$ eindeutig ist (Banachscher Fixpunktsatz). (Außerdem wird nachgewiesen, daß die Abschätzungen (2.29) für alle $(y,z) \in \mathcal{U}_h$ erfüllt sind.)

a) Zum Nachweis der Regularität von J wird zunächst die Regularität von $[g_y f_z](y_n,z_n)$ gezeigt: wie im Beweis von Satz 3 (vgl. (2.20) und (2.23)) folgt

$$\begin{aligned} f(y_n,z_n) &= f(y_n,z(t_n)) + f_z(y_n,z(t_n))(z_n - z(t_n)) + \mathcal{O}(\mu)\|z_n - z(t_n)\|^2 \, , \\ \|z_n - z(t_n)\| &= \mathcal{O}(1)\|y_n - y(t_n)\| + \mathcal{O}(\|[g_y f](y_n,z_n)\|) = \mathcal{O}(1)\|y_n - y(t_n)\| + \mathcal{O}(\frac{1}{h}\Delta) \, , \\ f(y_n,z_n) &= f(y_n,z(t_n)) + \mathcal{O}(1)\|z_n - z(t_n)\| = \mathcal{O}(1) + \mathcal{O}(\frac{1}{h}\Delta) \end{aligned}$$

und

$$[g_y f_z](y_n,z_n) = [g_y f_z](y_n,z(t_n)) + \mathcal{O}(\mu)\|z_n - z(t_n)\| = [g_y f_z](y(t_n),z(t_n)) + \mathcal{O}(\gamma) + \mathcal{O}(\mu\frac{1}{h}\Delta),$$

d. h., $[g_y f_z](y_n,z_n)$ ist regulär, falls γ und Δ_0 hinreichend klein sind. Das Inkrement $\Phi(y_n,z_n) - (y_n^T,z_n^T)^T$ des ersten Newtonschritts läßt sich deshalb schreiben als

$$-J^{-1}\begin{pmatrix} -f(y_n,z_n) \\ g(y_n) \end{pmatrix} = \begin{pmatrix} h(I - [f_z(g_y f_z)^{-1} g_y](y_n,z_n)) & [f_z(g_y f_z)^{-1}](y_n,z_n) \\ -[(g_y f_z)^{-1} g_y](y_n,z_n) & \frac{1}{h}[(g_y f_z)^{-1}](y_n,z_n) \end{pmatrix}\begin{pmatrix} f(y_n,z_n) \\ -g(y_n) \end{pmatrix}$$

und kann in der skalierten Norm

$$\|\begin{pmatrix} y \\ z \end{pmatrix}\|_\sigma := \|y\|_2 + \sigma\|z\|_2 \quad \text{mit} \quad \sigma := h + \frac{1}{M + \Delta/h^2}$$

durch

$$\|\Phi(y_n,z_n)-\begin{pmatrix} y_n \\ z_n\end{pmatrix}\|_\sigma = \|\begin{pmatrix} \mathcal{O}(h)+\mathcal{O}(\Delta)+\mathcal{O}(1)\|g(y_n)\| \\ \sigma\cdot\mathcal{O}(1)(\frac{1}{h}\|g(y_n)\|+\|[g_yf](y_n,z_n)\|)\end{pmatrix}\|_2 \le \frac{1}{2}C_1(h+\Delta)$$

abgeschätzt werden, denn

$$\sigma\cdot\frac{\Delta}{h} = (1+\frac{h}{h^2M+\Delta})\Delta \le \Delta+h \ . \qquad (2.32)$$

(In der Definition von σ bezeichnet M eine (große) positive Konstante, die von h, Δ und μ unabhängig ist und am Ende des Beweises geeignet gewählt wird.)

b) In $\mathcal{U}_h$ ist die Abschätzung (2.29) mit $C_{h\Delta}=C_1(2+M)$ erfüllt, denn dort gilt $\sigma\|z-z_n\|_2 \le C_1(h+\Delta)$ (vgl. (2.31)) und

$$\frac{h+\Delta}{\sigma} = \frac{h}{\sigma}+\frac{\Delta}{\sigma} \le h(M+\frac{\Delta}{h^2})+\frac{\Delta}{h} = 2\frac{\Delta}{h}+Mh \ .$$

Insbesondere ist für hinreichend kleine γ, h_0 und Δ_0 stets $\mathcal{U}_h \subset \mathcal{U}$, denn

$$\|z-z_n\| \le C_{h\Delta}(h+\frac{\Delta}{h}) \le C_{h\Delta}\cdot h_0+\Delta_z\cdot C_{h\Delta}\frac{1}{\Delta_z}\frac{\Delta}{h} \le C_{h\Delta}\cdot h_0+\Delta_z\cdot C_{h\Delta}\Delta_0 \le \frac{1}{2}\Delta_z$$

und damit $\|z-z(t_n)\| \le \Delta_z$. D. h., in $\mathcal{U}_h$ sind f, g und die Ableitungen dieser Funktionen wohldefiniert. Es gilt

$$\begin{aligned}\Phi'(y,z) &= J^{-1}(J-\begin{pmatrix}\frac{1}{h}I-f_y(y,z) & -f_z(y,z)\\ g_y(y) & 0\end{pmatrix})\\ &= \begin{pmatrix}\mathcal{O}(h) & \mathcal{O}(1)\\ \mathcal{O}(1) & \mathcal{O}(\frac{1}{h})\end{pmatrix}\begin{pmatrix} f_y(y,z) & f_z(y,z)-f_z(y_n,z_n)\\ g_y(y_n)-g_y(y) & 0\end{pmatrix}\end{aligned}$$

und mit den abkürzenden Bezeichnungen $\delta_y := \|y-y_n\|$, $\delta_z := \|z-z_n\|$ erhält man

$$\begin{aligned}\|\Phi'(y,z)\|_\sigma &= \|\begin{pmatrix} I & \\ & \sigma I\end{pmatrix}\cdot\Phi'(y,z)\cdot\begin{pmatrix} I & \\ & \frac{1}{\sigma}I\end{pmatrix}\|_2\\ &= \|\begin{pmatrix}\mathcal{O}(h)+\mathcal{O}(\Delta)+\mathcal{O}(1)(\delta_y+h\delta_z) & \mathcal{O}(\frac{h}{\sigma})(\delta_y+\mu\delta_z)\\ \mathcal{O}(\sigma)+\mathcal{O}(\sigma\cdot\frac{\Delta}{h})+\mathcal{O}(\frac{\sigma}{h})(\delta_y+h\delta_z) & \mathcal{O}(1)(\delta_y+\mu\delta_z)\end{pmatrix}\|_2 \ ,\end{aligned}$$

denn

$$f_y(y,z) = f_y(y,z(t_n))+\mathcal{O}(1)(\|z-z_n\|+\|z_n-z(t_n)\|) = \mathcal{O}(1)+\mathcal{O}(\frac{1}{h}\Delta)+\mathcal{O}(1)\|z-z_n\| \ .$$

Die Konstanten in den $\mathcal{O}(.)$–Termen hängen dabei von der Konstante L aus (2.9) ab, sind aber von h, Δ, M und μ unabhängig.

Wegen $\frac{h}{\sigma} \le 1$, $(y,z)\in\mathcal{U}_h$ und (2.32) ist $\|\Phi'\|_\sigma$ beschränkt durch

$$\begin{aligned}\|\Phi'\|_\sigma &= \mathcal{O}(\sigma)+\mathcal{O}(h)+\mathcal{O}(\Delta)+\mathcal{O}(\sigma+h+\mu)\|z-z_n\|\\ &= \mathcal{O}(\sigma)+\mathcal{O}(h)+\mathcal{O}(\Delta)+\mathcal{O}(M(h+\mu)(h+\tfrac{\Delta}{h})) \ .\end{aligned}$$

Wählt man nun eine hinreichend große Konstante M, so ist für kleine $h>0$ der durch $\mathcal{O}(\sigma)$ abgeschätzte Term kleiner als 1/4. Anschließend können h_0 und Δ_0 so bestimmt werden, daß $\|\Phi'\|_\sigma \le \frac{1}{2}$ ist. Damit ist Satz 4 bewiesen. ∎

Wie in Beispiel 6 erfüllt die auf dem Computer berechnete Lösung $(\hat y_n,\hat z_n)$ die Gleichungen (2.28) nicht exakt, sondern mit (kleinen) Residuen δ_{n+1}, θ_{n+1}:

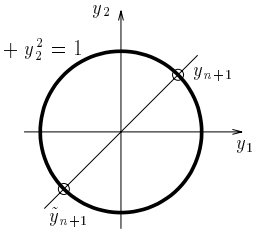
$$\begin{aligned}\frac{\hat y_{n+1}-\hat y_n}{h} &= f(\hat y_{n+1},\hat z_{n+1})+\delta_{n+1} \ ,\\ \theta_{n+1} &= g(\hat y_{n+1}) \ .\end{aligned} \qquad (2.33)$$

Bemerkung 8 Fehlerabschätzungen für $\|\hat y_n-y_n\|$ können nur dann angegeben werden, wenn (y_n,z_n) und $(\hat y_n,\hat z_n)$ in der Umgebung von ein und derselben analytischen Lösung von (2.7) verbleiben.

So haben die Gleichungen (2.28) für das Beispiel

$$y_1'(t) = -2y_1z \ , \quad y_2'(t) = -2y_2z \ , \quad y_1^2+y_2^2 = 1$$

neben der Lösung $y_{n+1}=y_n$, $z_{n+1}=0$ eine zweite Lösung $\tilde y_{n+1}=-y_n$, $\tilde z_{n+1}=\frac{1}{h}$ (falls $y_{n,1}^2+y_{n,2}^2=1$, vgl. Abbildung). Schranken für $\|\hat y_{n+1}-y_{n+1}\|$ können deshalb für $\hat y_{n+1}$ in einer Umgebung von y_{n+1}, aber nicht für $\hat y_{n+1}$ in einer Umgebung von $\tilde y_{n+1}$ gefunden werden.



Als Gegenstück zur Stetigkeit von $(\hat y(t),\hat z(t))$ in Satz 3 wird die Störungstheorie deshalb auf Folgen $(\hat y_n,\hat z_n)$ beschränkt, für die es unabhängig von h, δ, θ und μ eine Konstante $C^*_{h\Delta}$ gibt, so daß

$$\begin{aligned}\|\hat y_{n+1}-\hat y_n\| &\le C^*_{h\Delta}(h+\|g(\hat y_n)\|+h\|[g_yf](\hat y_n,\hat z_n)\|+h\delta+\theta) \ ,\\ \|\hat z_{n+1}-\hat z_n\| &\le C^*_{h\Delta}(h+\tfrac{1}{h}\|g(\hat y_n)\|+\|[g_yf](\hat y_n,\hat z_n)\|+\delta+\tfrac{1}{h}\theta)\end{aligned} \qquad (2.34)$$

mit

$$\delta := \max_m\|\delta_m\|+\|[g_yf](\hat y_0,\hat z_0)\| \ , \quad \theta := \max_m\|\theta_m\|+\|g(\hat y_0)\|$$

gilt. Wie in Satz 4 kann man zeigen, daß es stets mindestens ein $(\hat y_{n+1},\hat z_{n+1})$ mit (2.33) und (2.34) gibt, wenn $h\le h_0$, $\|\hat y_n-y(t_n)\|+\mu\|\hat z_n-z(t_n)\| \le \gamma$, $\|\hat z_n-z(t_n)\| \le \frac{1}{2}\Delta_z$ und $\Delta^*+\overline{\mu}\frac{\Delta^*}{h} \le \Delta_0$ gilt und die positiven Konstanten h_0, γ und Δ_0 hinreichend klein sind. ($\Delta^* := \|g(\hat y_n)\|+h\|[g_yf](\hat y_n,\hat z_n)\|+\|\theta_{n+1}\|+h\|\delta_{n+1}\|$)

Bedingung (2.34) beschreibt den Einfluß von Störungen in einem einzelnen Integrationsschritt. Nun untersuchen wir die Fortpflanzung und Verstärkung von Fehlern während der Integration:

Satz 5 *a) Es gibt eine Konstante C_0, so daß für alle Folgen $(\hat y_n,\hat z_n)$ mit (2.33), (2.34) und $\delta\le C_\delta$ die Abschätzung*

$$\|\hat y_n-y_n\| \le C_0\Big(\|\hat y_0-y_0\|+\int_h P\delta+\theta+\theta D_h+\mu D_h^2\Big) \qquad (2.35)$$

für alle $n \geq 0$, mit $nh \leq T$ erfüllt ist, sofern die Schrittweite h, die rechte Seite von (2.35) und $\overline{\mu} D_h$ hinreichend klein sind. Hier werden die Bezeichnungen

$$\int_h P\delta := h \sum_{m=0}^{n-1} \|P_m \delta_{m+1}\| + h\delta \quad , \quad D_h := \delta + \frac{1}{h}\theta$$

verwendet, dabei ist $P_m := P(t_m)$ der Projektor aus (2.12). Die Konstante C_0 ist unabhängig von h, δ, θ und μ, hängt jedoch von L, T und C_δ ab.
b) Ist $f(\eta, \zeta) = f_0(\eta) + f_z \cdot \zeta$ und $f_z \equiv const$, dann gilt (2.35) mit $D_h := 0$.
Darüberhinaus gilt in beiden Fällen für die algebraischen Komponenten

$$\|\hat{z}_n - z_n\| \leq C_0 \Big(\|\hat{y}_0 - y_0\| + \delta + \frac{1}{h}\theta \Big) .$$

Beweis Für die Untersuchung der Fehlerfortpflanzung wird wie für die analytische Lösung eine Projektion von $\hat{y}_m - y_m$ definiert:

$$v_m := \Phi_h(\hat{y}_m) - \Phi_h(y_m) \quad \text{mit} \quad \Phi_h(\eta) := P(y_m, z_m)(\eta - [f_z(g_y f_z)^{-1} g](\eta, z_m)) . \tag{2.36}$$

Dabei wird der Projektor $P(y_m, z_m) := I - [f_z(g_y f_z)^{-1} g_y](y_m, z_m)$ entlang der numerischen Lösung (y_m, z_m) entwickelt.
Ähnlich wie in Bedingung (2.16) setzen wir zunächst zusätzlich voraus, daß für alle $m \leq n$

$$\begin{aligned} \|y_m - y(t_m)\| + \|z_m - z(t_m)\| \leq \gamma \ , \ \ \|[g_y f](y_m, z_m)\| \leq C_h \cdot h \ , \\ \|\hat{y}_m - y(t_m)\| + \mu \|\hat{z}_m - z(t_m)\| \leq \gamma \ , \ \ \|[g_y f](\hat{y}_m, \hat{z}_m)\| \leq C_h(\gamma + D_h) \ , \\ \|\hat{z}_m - z(t_m)\| \leq \tfrac{1}{2}\Delta_z \end{aligned} \tag{2.37}$$

gilt, und führen am Ende des Beweisteils a) den Nachweis von (2.37). Hier sollen γ und C_h nicht von h, D_h und μ abhängen, γ sei hinreichend klein. Die Konstante C_h ist von der Größenordnung $\mathcal{O}(1)$ und wird später geeignet gewählt (s. u.).
Unter diesen Voraussetzungen ist das Gleichungssystem (2.28) lokal eindeutig lösbar, es gilt $\|y_{m+1} - y_m\| = \mathcal{O}(h + C_h h^2)$ und $\|z_{m+1} - z_m\| = \mathcal{O}(h + C_h h)$ (vgl. Satz 4), wobei die Konstanten der $\mathcal{O}(.)$–Terme von C_h unabhängig sind. Ebenso folgt aus (2.34)

$$\begin{aligned} \|\hat{y}_{m+1} - \hat{y}_m\| &= \mathcal{O}(h) + \mathcal{O}(hD_h) + \mathcal{O}(h)\|[g_y f](\hat{y}_m, \hat{z}_m)\| \\ &= \mathcal{O}(h + C_h h\gamma + (1 + C_h) h D_h) = \mathcal{O}((1 + C_h)\sqrt{h}) \ , \end{aligned} \tag{2.38}$$

$$\begin{aligned} \|\hat{z}_{m+1} - \hat{z}_m\| &= \mathcal{O}(h) + \mathcal{O}(C_h\gamma) + \mathcal{O}((1 + C_h) D_h) \ , \\ \theta \cdot \|\hat{y}_{m+1} - \hat{y}_m\| &= \mathcal{O}(h)(1 + C_h)(\theta + \theta D_h) \ , \end{aligned}$$

denn $hD_h = h\delta + \theta \leq C_\delta h + \sqrt{h} \cdot \sqrt{\theta \cdot D_h} = \mathcal{O}(\sqrt{h})$.
Das diskrete Gegenstück zu der im Beweis von Satz 3 verwendeten Gleichung $g_y(y)y' = 0$ ist

$$\begin{aligned} g_y(y_{m+1})(y_{m+1} - y_m) = g(y_{m+1}) - g(y_m) + \\ + \int_0^1 \Big(g_y(y_{m+1}) - g_y(y_m + \vartheta(y_{m+1} - y_m)) \Big) \, d\vartheta \cdot (y_{m+1} - y_m) \ , \end{aligned} \tag{2.39}$$

denn für $\Psi(\vartheta) := g(y_m + \vartheta(y_{m+1} - y_m))$ gilt $\Psi(1) - \Psi(0) = \int_0^1 \Psi'(\vartheta) \, d\vartheta$.
Aus (2.39) folgt

$$[g_y f](y_{m+1}, z_{m+1}) = \frac{1}{h} g_y(y_{m+1})(y_{m+1} - y_m) = \mathcal{O}(\frac{1}{h})\|y_{m+1} - y_m\|^2 = \mathcal{O}(h + C_h h^2) \leq C_h h \tag{2.40}$$

und (ersetze $y_m \to \hat{y}_m$, $y_{m+1} \to \hat{y}_{m+1}$)

$$\begin{aligned} [g_y f](\hat{y}_{m+1}, \hat{z}_{m+1}) &= \mathcal{O}(\|\delta_{m+1}\|) + \mathcal{O}(\frac{1}{h}(\|g(\hat{y}_{m+1})\| + \|g(\hat{y}_m)\|)) + \mathcal{O}(\frac{1}{h})\|\hat{y}_{m+1} - \hat{y}_m\|^2 \\ &= \mathcal{O}(D_h) + \mathcal{O}(\,(1 + C_h)^2(1 + C_\delta + \frac{1}{h}\theta)(h + hD_h)\,) \\ &\leq C_h(\gamma + D_h) \ , \end{aligned} \tag{2.41}$$

sofern man in (2.37) die Konstante C_h ausreichend groß wählt. Dabei ist zu berücksichtigen, daß $h \leq h_0$ und $\theta + \theta D_h \leq \Delta_0$ mit (von γ und C_h abhängigen) kleinen positiven Konstanten h_0 und Δ_0 gilt.
Ersetzt man in (2.39) $y_m \to \hat{y}_m$, $y_{m+1} \to y_m$, so folgt mit $g(\hat{y}_m) = \theta_m = \mathcal{O}(\theta)$

$$\begin{aligned} \|(I - P(y_m, z_m))(\hat{y}_m - y_m)\| &\leq \mathcal{O}(1)\|\hat{y}_m - y_m\|^2 + \mathcal{O}(\theta) \leq \tfrac{1}{2}\|\hat{y}_m - y_m\| + \mathcal{O}(\theta) \\ &\leq \tfrac{1}{2}\|P(y_m, z_m)(\hat{y}_m - y_m)\| + \tfrac{1}{2}\|(I - P(y_m, z_m))(\hat{y}_m - y_m)\| + \mathcal{O}(\theta) \end{aligned}$$

und damit $\|(I - P(y_m, z_m))(\hat{y}_m - y_m)\| \leq \|v_m\| + \mathcal{O}(\theta)$ und

$$\|\hat{y}_m - y_m\| \leq 2\|v_m\| + \mathcal{O}(\theta) \tag{2.42}$$

(vgl. (2.36)). Der Fehler in den differentiellen Komponenten ist also durch einen Ausdruck in θ und $\|v_m\|$ beschränkt.
Für die gestörte Folge erhält man wie zuvor in (2.39)

$$\begin{aligned} g_y(\hat{y}_{m+1})(\hat{y}_{m+1} - \hat{y}_m) &= \mathcal{O}(1)\|\hat{y}_{m+1} - \hat{y}_m\|^2 + g(\hat{y}_{m+1}) - g(\hat{y}_m) \\ &= g(\hat{y}_{m+1}) - g(\hat{y}_m) + \mathcal{O}(h) \end{aligned} \tag{2.43}$$

(vgl. auch (2.38)). Um eine Abschätzung für die algebraischen Komponenten zu erhalten, betrachten wir schließlich die Differenz zwischen (2.39) und der entsprechenden Beziehung mit $y_m \to \hat{y}_m$, $y_{m+1} \to \hat{y}_{m+1}$:

$$\begin{aligned} g_y(\hat{y}_{m+1})(\hat{y}_{m+1} - y_{m+1}) - (g(\hat{y}_{m+1}) - g(\hat{y}_m)) - g_y(y_{m+1})(y_{m+1} - y_m) = \\ = \mathcal{O}(1)(\|\hat{y}_{m+1} - \hat{y}_m\| + \|y_{m+1} - y_m\|)(\|\hat{y}_{m+1} - y_{m+1}\| + \|\hat{y}_m - y_m\|) \\ = \mathcal{O}(h + hD_h)(\|\hat{y}_{m+1} - y_{m+1}\| + \|\hat{y}_m - y_m\|) \ . \end{aligned} \tag{2.44}$$

Wie in (2.20) gilt

$$f(\hat{y}_{m+1}, \hat{z}_{m+1}) = f(\hat{y}_{m+1}, z_{m+1}) + f_z(\hat{y}_{m+1}, z_{m+1})(\hat{z}_{m+1} - z_{m+1}) + \mathcal{O}(\mu)\|\hat{z}_{m+1} - z_{m+1}\|^2 \tag{2.45}$$

und damit auch

$$[g_y f](\hat y_{m+1},\hat z_{m+1}) = [g_y f](\hat y_{m+1}, z_{m+1}) + [g_y f_z](\hat y_{m+1}, z_{m+1})(\hat z_{m+1} - z_{m+1}) + \\ + \mathcal{O}(\mu)\|\hat z_{m+1} - z_{m+1}\|^2 .$$

Wegen $\|z_{m+1} - z(t_{m+1})\| = \mathcal{O}(\gamma) + \mathcal{O}(h)$ ist die Matrix $[g_y f_z](\hat y_{m+1}, z_{m+1})$ regulär, so daß diese Gleichungen nach $\hat z_{m+1} - z_{m+1}$ aufgelöst werden können, denn im letzten Summanden auf der rechten Seite gilt $\mu\|\hat z_{m+1} - z_{m+1}\| = \mathcal{O}(\gamma) + \mathcal{O}(h) + \mathcal{O}(\mu D_h) \ll 1$. Man erhält die Abschätzung

$$\|\hat z_{m+1} - z_{m+1}\| = \mathcal{O}(1)\|[g_y f](\hat y_{m+1}, \hat z_{m+1}) - [g_y f](\hat y_{m+1}, z_{m+1})\| .$$

Setzt man hier $f(\hat y_{m+1}, \hat z_{m+1}) = \frac{1}{h}(\hat y_{m+1} - \hat y_m) - \delta_{m+1}$, $f(y_{m+1}, z_{m+1}) = \frac{1}{h}(y_{m+1} - y_m)$, $[g_y f](\hat y_{m+1}, z_{m+1}) = [g_y f](y_{m+1}, z_{m+1}) + \mathcal{O}(1)\|\hat y_{m+1} - y_{m+1}\|$ und (2.44) ein, so ergibt sich

$$\|\hat z_{m+1} - z_{m+1}\| = \mathcal{O}(1 + D_h)(\|\hat y_{m+1} - y_{m+1}\| + \|\hat y_m - y_m\|) + \mathcal{O}(D_h) \tag{2.46}$$

und

$$\mu\|\hat z_{m+1} - z_{m+1}\|^2 = \mathcal{O}(1)(\|\hat y_{m+1} - y_{m+1}\| + \|\hat y_m - y_m\|) + \mathcal{O}(\mu D_h^2) , \tag{2.47}$$

denn $\|\hat y_m - y_m\| = \mathcal{O}(\gamma)$, $\|\hat y_{m+1} - y_{m+1}\| = \mathcal{O}(\gamma) + \mathcal{O}(\sqrt{h})$ und die Terme μD_h und μD_h^2 sind nach Voraussetzung klein.

Die Abschätzungen (2.39) und (2.42)–(2.47) werden zu einer Schranke für $v_{m+1} - v_m$ zusammengefügt:

$$\begin{aligned} v_{m+1} - v_m = \\ &= P(y_{m+1}, z_{m+1})\Big(\hat y_{m+1} - \hat y_m - [f_z(g_y f_z)^{-1}](\hat y_{m+1}, z_{m+1})(g(\hat y_{m+1}) - g(\hat y_m))\Big) - \\ &\quad - P(y_{m+1}, z_{m+1})(y_{m+1} - y_m) + \mathcal{O}(h)\|\hat y_m - y_m\| + \mathcal{O}(h\theta) + \\ &\quad + P(y_{m+1}, z_{m+1})\Big([f_z(g_y f_z)^{-1}](\hat y_{m+1}, z_{m+1}) - [f_z(g_y f_z)^{-1}](\hat y_m, z_m)\Big) g(\hat y_m) , \\ &= P(y_{m+1}, z_{m+1})\Big((I - [f_z(g_y f_z)^{-1} g_y](\hat y_{m+1}, z_{m+1}))(\hat y_{m+1} - \hat y_m) + \\ &\qquad + [f_z(g_y f_z)^{-1}](\hat y_{m+1}, z_{m+1}) \cdot (g_y(\hat y_{m+1})(\hat y_{m+1} - \hat y_m) - (g(\hat y_{m+1}) - g(\hat y_m)))\Big) - \\ &\quad - P(y_{m+1}, z_{m+1})(y_{m+1} - y_m) + \mathcal{O}(h)\|\hat y_m - y_m\| + (\mathcal{O}(1)\|\hat y_{m+1} - \hat y_m\| + \mathcal{O}(h))\,\theta . \end{aligned}$$

Verwendet man hier die Identität

$$P(y_{m+1}, z_{m+1}) f_z(\hat y_{m+1}, z_{m+1}) = P(y_{m+1}, z_{m+1})(f_z(\hat y_{m+1}, z_{m+1}) - f_z(y_{m+1}, z_{m+1}))$$

und die Bezeichnung $\Psi(\eta) := P(y_{m+1}, z_{m+1})(I - [f_z(g_y f_z)^{-1} g_y](\eta, z_{m+1})) f(\eta, z_{m+1})$, so ergibt sich

$$\begin{aligned} v_{m+1} - v_m &= h(\Psi(\hat y_{m+1}) - \Psi(y_{m+1})) + hP(y_{m+1}, z_{m+1})\delta_{m+1} + \\ &\quad + (\mathcal{O}(h\|\delta_{m+1}\|) + \mathcal{O}(h))\|\hat y_{m+1} - y_{m+1}\| + \\ &\quad + \mathcal{O}(h)\|\hat y_m - y_m\| + \mathcal{O}(h\mu)\|\hat z_{m+1} - z_{m+1}\|^2 + \mathcal{O}(h\theta) + \mathcal{O}(h\theta D_h) , \end{aligned}$$

d. h.

$$\begin{aligned} \|v_{m+1}\| \;\le\; &\|v_m\| + \mathcal{O}(h)\|\hat y_m - y_m\| + \mathcal{O}(h)\|\hat y_{m+1} - y_{m+1}\| + \mathcal{O}(h\mu)\|\hat z_{m+1} - z_{m+1}\|^2 + \\ &+ h\|P_m\delta_{m+1}\| + \mathcal{O}(h^2\delta) + \mathcal{O}(h\theta) + \mathcal{O}(h\theta \cdot D_h) + \mathcal{O}(h) d_h\|\delta_{m+1}\| \end{aligned}$$

mit $d_h := \|y_{m+1} - y(t_{m+1})\| + \|z_{m+1} - z(t_{m+1})\|$. In diese Ungleichung setzt man schließlich die Abschätzungen für $\|\hat y_{m+1} - y_{m+1}\|$ und $\|\hat z_{m+1} - z_{m+1}\|$ ein und erhält

$$\|v_{m+1}\| \le (1 + \mathcal{O}(h))\|v_m\| + h\|P_m\delta_{m+1}\| + \mathcal{O}(h)(d_h\|\delta_{m+1}\| + h\delta + \theta + \theta D_h + \mu D_h^2) . \tag{2.48}$$

Mit Ungleichung (2.48) wird nun zunächst die Konvergenz des Eulerverfahrens nachgewiesen. Hierzu betrachtet man die *spezielle* Folge $(\hat y_n, \hat z_n) := (\, y(t_n), z(t_n)\,)$, so daß δ in (2.33) der lokale Diskretisierungsfehler ist, $\delta = \mathcal{O}(h)$ und $\theta = \max_n \|g(y(t_n))\| = 0$, also $D_h = \mathcal{O}(h)$. Damit vereinfacht sich (2.48) zu

$$\|v_{m+1}\| = (1 + \mathcal{O}(h))\|v_m\| + \mathcal{O}(h\delta) ,$$

so daß rekursiv $\|v_n\| = \mathcal{O}(1)\|v_0\| + \mathcal{O}(\delta) = \mathcal{O}(h)$ folgt. Mit (2.42) und (2.46) erhält man $y_n = y(t_n) + \mathcal{O}(h)$, $z_n = z(t_n) + \mathcal{O}(h)$, d. h. Konvergenz mit Ordnung 1. Ist h_0 hinreichend klein, so verbleibt die numerische Lösung (y_n, z_n) also für beliebiges $h \in (0, h_0]$ in der γ–Umgebung der analytischen Lösung, so daß die ersten beiden Bedingungen in (2.37) stets erfüllt sind ($\|[g_y f](y_n, z_n)\| \le C_h h$ folgt aus (2.40)). Insbesondere gilt $d_h = \mathcal{O}(h)$ in (2.48).

Kehren wir nun zu *beliebigen* Folgen $(\hat y_n, \hat z_n)$ mit (2.33) und (2.34) zurück. Setzt man $d_h = \mathcal{O}(h)$ ein, so folgt durch rekursive Anwendung von (2.48)

$$\|v_n\| = \mathcal{O}(1)\Big(\|v_0\| + \int_h P\delta + \theta + \theta \cdot D_h + \mu D_h^2\Big)$$

und damit auch (2.35) und die entsprechende Schranke für z. Wie für die analytische Lösung können auch hier diese Abschätzungen verwendet werden, um durch vollständige Induktion zu beweisen, daß die zusätzliche Voraussetzung (2.37) für alle $m \ge 0$ mit $mh \le T$ erfüllt ist, wenn $\overline{\mu} D_h$ und die Ausdrücke auf der rechten Seite von (2.35) hinreichend klein sind. Dabei schätzt man den Abstand zwischen $(\hat y_m, \hat z_m)$ und der analytischen Lösung unter Verwendung der Dreiecksungleichung ab ($\hat y_m - y(t_m) = \hat y_m - y_m + \mathcal{O}(h)$, $\hat z_m - z(t_m) = \hat z_m - z_m + \mathcal{O}(h)$) und gebraucht (2.41) zum Nachweis von $\|[g_y f](\hat y_m, \hat z_m)\| \le C_h(\gamma + D_h)$. Damit ist Teil a) des Satzes vollständig bewiesen.

b) Ist $f_z = const$, dann gilt $\mu = 0$ und $P(y_{m+1}, z_{m+1}) f_z \equiv 0$. Damit vereinfacht sich Definition (2.36) zu $v_m = P(y_m, z_m)(\hat y_m - y_m)$, so daß v_m nicht von $\hat z_m$ beeinflußt wird. Deshalb hängt die Fehlerschranke für y in diesem Fall stetig von den Störungen ab. ∎

Bemerkung 9 Der Beweis von Satz 5 lehnt sich an den Konvergenzbeweis für implizite Runge–Kutta–Verfahren von Hairer et al. an ([81, Satz 4.4]). Insbesondere können auch hier die Ergebnisse unmittelbar auf den Fall variabler Schrittweiten übertragen werden, dabei ist $\frac{1}{h}\theta$ durch $\max_m \frac{1}{2h_m}(\|\theta_{m+1}\| + \|\theta_m\|)$ zu ersetzen.

Während in [81] der Schwerpunkt auf Konvergenzuntersuchungen liegt und daneben Abschätzungen der Form (2.35) für Störungen $\delta = \mathcal{O}(h)$, $\theta = \mathcal{O}(h^2)$ bewiesen werden (hier vereinfacht sich (2.35) zu $C_0(\|\hat y_0 - y_0\| + \int_h P\delta + \theta)$), konzentrieren wir uns in

Tabelle 2.1: Schranken für den Einfluß von kleinen Störungen auf die Lösung von gewöhnlichen Differentialgleichungen und von Index-2-Systemen (2.7).

	analytische Lösung $\|\hat{y}(t)-y(t)\|,\ \|\hat{z}(t)-z(t)\|$	numerische Lösung $\|\hat{y}_n-y(t_n)\|,\ \|\hat{z}_n-z(t_n)\|$
gew. Dgl. (2.1)	$C\int\delta$	$C(h^p+\int_h\delta)$
DA–Systeme (2.7)		
Komponenten y	$C(\int P\delta+\|\theta\|+\|\theta\|D+\mu D^2)$	$C(h^q+\int_h P\delta+\theta+\theta D_h+\mu D_h^2)$
Komponenten z	$C(\|\delta\|_{C^0}+\|\theta\|_{C^1})$	$C(h^r+\delta+\frac{1}{h}\theta)$

Satz 5 auf möglichst scharfe Fehlerabschätzungen (2.35). Neben wesentlich schwächeren Voraussetzungen an die Größe der Störungen ($\theta\cdot D_h=o(1)$, $\mu D_h^2=o(1)$, $\overline{\mu}D_h=o(1)$) wird dabei insbesondere auch die Struktur von f berücksichtigt (sowohl für $\mu=\mathcal{O}(1)$ als auch für $0\le\mu\ll 1$). Für den Konvergenzbeweis in [81] reicht es aus, längs der analytischen Lösung $(y(t),z(t))$ zu linearisieren. Dagegen erfordert der Nachweis der verbesserten Fehlerschranke (2.35) sowohl die Linearisierung entlang (y_n,z_n) als auch entlang $(\hat{y}_n,z_n)$.

2.2.3 Zusammenfassung und Beispiele

Die Sätze 3 und 5 sind die zentralen Ergebnisse der Störungstheorie für Index-2-Systeme (2.7). In diesem Abschnitt werden die Abschätzungen für die analytische und die numerische Lösung einander gegenübergestellt. Beispiele zeigen, daß die Schranken (2.14) und (2.35) i. allg. nicht verbessert werden können.

Die Sensitivität der Lösung von Index-2-Systemen (2.7) bezüglich kleiner Störungen wird durch den Begriff „Störungsindex 2“ nur teilweise beschrieben, weil die differentiellen Lösungskomponenten y sehr viel robuster gegenüber Störungen sind, als die obere Schranke in (2.11) angibt, und die Fehlerfortpflanzung in y darüberhinaus von der Struktur von f abhängt.

Tab. 2.1 faßt die Sätze 3 und 5 zusammen. Die Analogie der Abschätzungen zeigt, daß es wie in der Theorie der gewöhnlichen Differentialgleichungen sinnvoll ist, die Störungstheorie zunächst für die analytische Lösung zu entwickeln und dann auf die numerische Lösung zu übertragen. Hier werden die Bezeichnungen $\int\delta=\max\|\int\delta(w)dw\|$, $\int P\delta=\max\|\int P(w)\delta(w)dw\|$, $\|\theta\|=\|\theta\|_{C^0}$, $D=D(t)=\|\delta\|_{C^0}+\|\theta\|_{C^1}$ und $\int_h\delta=h\sum_m\|\delta_{m+1}\|$, $\int_h P\delta=h\sum_m\|P_m\delta_{m+1}\|+h\delta$, $D_h=\delta+\frac{1}{h}\theta$ verwendet und die Fehler in den Anfangswerten vernachlässigt. Die Terme h^p, h^q und h^r entsprechen dem (globalen) Diskretisierungsfehler des numerischen Verfahrens, d. h. $p=q=r=1$ für das implizite Eulerverfahren.
Die Größe des zusätzlichen Fehlerterms $\|\theta\|_{C^0}\cdot D+\mu D^2$ bzw. $\theta D_h+\mu D_h^2$ in den differentiellen Komponenten hängt entscheidend von μ, d. h. von $\|f_{zz}\|$ ab. In dem Spezialfall $f_z=const$ kann in Tab. 2.1 $D=D_h=0$ gesetzt werden.

Bemerkung 10 Die genaue Gestalt des zusätzlichen Fehlerterms in y ist $\nu\|\theta\|_{C^0}D+\mu D^2$ bzw. $\nu\theta D_h+\mu D_h^2$ mit $\nu\le L$ und $\|f_{yz}(\eta,\zeta)\|\le\nu$, ($(\eta,\zeta)\in\mathcal{U}$) in (2.9). Der Nachweis dieser verbesserten Fehlerschranke verkompliziert die Beweise der Sätze 3 und 5 noch einmal erheblich (statt (2.15) ist z. B. $\Phi(\eta):=P(t)(\eta-[f_z(g_yf_z)^{-1}g](\eta,z(t)))$ zu verwenden). Da $\theta D_h\ll D_h^2$ gilt, beschränken wir uns deshalb auf die beiden praktisch relevanten Fälle $\nu=\mathcal{O}(L)$ (Sätze 3a, 5a) und $\nu=\mu=0$ (Sätze 3b, 5b).

Die Ergebnisse der Störungstheorie unterstreichen, daß Index-2-Systeme (2.7) durch direkte Anwendung von Verfahren aus der Theorie der gewöhnlichen Differentialgleichungen genau und zuverlässig gelöst werden können, wenn der zusätzliche Fehlerterm in y klein gehalten wird. Neben der Rechnung in doppelter Genauigkeit sind hierfür die Abbruchschranken für die Lösung nichtlinearer Gleichungen (sehr) klein zu wählen.

Da jede Schrittweitensteuerung voraussetzt, daß der Fehler bei Reduktion der Schrittweite kleiner wird, müssen Ordnungs- und Schrittweitensteuerung geeignet modifiziert werden, damit ein Integrator nicht nur mit Schrittweiten im Bereich A, sondern auch im Bereich B von Abb. 2.2 auf S. 29 arbeiten kann. Hierzu schließt man den Fehler in z aus der Schrittweitensteuerung aus (z. B. in DASSL [132]) oder skaliert ihn wie in RADAU5 mit der Schrittweite h ([81, S. 102]). Nur für sehr kleine Schrittweiten h kann die Fehlerverstärkung eine Integration unmöglich machen, Bereich C von Abb. 2.2 zeigt, warum hier jede automatische Schrittweitensteuerung versagt.

In diesem Fall kann man z. B. das DA–System vor Beginn der Integration in ein äquivalentes System transformieren, das linear in den algebraischen Komponenten ist ([71], vgl. auch Beispiel 18). Hierzu nutzt man aus, daß die versteckten Zwangsbedingungen $0=\frac{d}{dt}g(y(t))=g_y(y)y'(t)=[g_yf](y,z)$ in einer Umgebung der Lösung von (2.7) nach z auflösbar sind, denn nach dem Satz über die implizite Funktion gibt es wegen (2.8) eine Funktion $G:\Omega_y\to\mathbb{R}^{n_z}$ mit $\Omega_y\subset\mathbb{R}^{n_y}$, $0=[g_yf](y,G(y))$ und $z=G(y)$. Durch Einführung von Hilfsvariablen $\zeta\in\mathbb{R}^{n_z}$ erhält man mit $\tilde{f}(y)=f(y,G(y))$ das zu (2.7) analytisch äquivalente DA–System

$$\begin{aligned} y' &= \tilde{f}(y)-g_y^T(y)\zeta \\ 0 &= g(y)\,, \end{aligned} \tag{2.49}$$

das den Index 2 hat und linear in den (neuen) algebraischen Variablen ζ ist. Wegen $[g_y\tilde{f}](y)=0$ folgt aus $0=\frac{d}{dt}g(y(t))=g_y(y)y'(t)=[g_y\tilde{f}](y)-[g_yg_y^T](y)\cdot\zeta$, daß ζ für die analytische Lösung von (2.49) identisch verschwindet.

Mit den Bezeichnungen von (2.7) führt diese Transformation also auf Systeme mit $f(\eta,\zeta)=f_0(\eta)-g_y^T(\eta)\zeta$, die einen Schwerpunkt der nachfolgenden Beispiele bilden. Zunächst beginnen wir jedoch mit einem DA–System, das nichtlinear in z ist und in ähnlicher Form schon von Lubich ([105]) betrachtet wurde:

Beispiel 9 Zu dem Anfangswertproblem $y(0)=(0,0,0)^T$, $z(0)=(0,0)^T$ für das System

$$\begin{aligned} y_1' &= z_1 &&, \quad 0 = y_1\,, \\ y_2' &= z_2 &&, \quad 0 = y_2\,, \\ y_3' &= y_2z_1-y_1z_2+\mu(z_1^2+z_2^2) &&, \quad (t\in[0,1])\,, \end{aligned} \tag{2.50}$$

mit der analytischen Lösung $y(t)=(0,0,0)^T$, $z(t)=(0,0)^T$ seien Funktionen

$$\hat{y}(t)=\left(\Theta\sin\frac{t}{\varepsilon},\,\Theta\cos\frac{t}{\varepsilon},\,(\Theta D+\mu D^2)\,t\right)^T\,,\quad \hat{z}(t)=\left(D\cos\frac{t}{\varepsilon},\,-D\sin\frac{t}{\varepsilon}\right)^T$$

gegeben (hier bezeichnen Δ, Θ und ε kleine positive Konstanten, $D := \Delta + \frac{\Theta}{\varepsilon}$). Die Residuen in (2.10) betragen

$$\delta(t) = (\, -\Delta \cos\frac{t}{\varepsilon}\,,\, \Delta \sin\frac{t}{\varepsilon}\,,\, 0\,)^T\,,\quad \theta(t) = (\, \Theta \sin\frac{t}{\varepsilon}\,,\, \Theta\cos\frac{t}{\varepsilon}\,)^T\,,$$

es gilt also $\|\hat y_0 - y_0\| = \mathcal{O}(\Theta)$, $\|\delta\|_{C^0} = \mathcal{O}(\Delta)$, $\|\theta\|_{C^0} = \mathcal{O}(\Theta)$, $\|\theta\|_{C^1} = \mathcal{O}(\frac{\Theta}{\varepsilon})$ und $|\hat y_3(t) - y_3(t)| = \mathcal{O}(\Theta D + \mu D^2)$, $\|\hat z(t) - z(t)\| = \mathcal{O}(D)$.
Das implizite Eulerverfahren löst das Anfangswertproblem (2.50) exakt: $y_n = y(t_n)$, $z_n = z(t_n)$. Für die Folge $(\hat y_n, \hat z_n)$ mit

$$\begin{aligned}\hat y_n &= \Big(\,\Theta \sin n\frac{\pi}{2}\,,\, \Theta\cos n\frac{\pi}{2}\,,\, nh(\Theta D_h + 2\mu D_h^2)\Big)^T\,,\\ \hat z_n &= \sqrt{2}D_h\cdot\Big(\cos(2n-1)\frac{\pi}{4}\,,\, -\sin(2n-1)\frac{\pi}{4}\Big)^T\end{aligned}$$

und $D_h := \Delta + \frac{1}{h}\Theta$ betragen die Residuen in (2.33) $\delta = \mathcal{O}(\Delta)$, $\theta = \mathcal{O}(\Theta)$. Es gilt $|\hat y_{n,3} - y_{n,3}| = \mathcal{O}(\theta D_h + \mu D_h^2)$ und $\|\hat z_n - z_n\| = \mathcal{O}(D_h)$, d. h., als Analogon zu $\|\delta(t)\|_{C^0}$ und $\|\theta(t)\|_{C^1}$ in der Schranke für die analytische Lösung steht hier δ bzw. $\frac{1}{h}\theta$. Man erhält also für (2.50) sowohl für die analytische als auch für die numerische Lösung Fehler in der Größenordnung der Fehlerschranken aus den Sätzen 3a und 5a.

Hat man — wie für mechanische Mehrkörpersysteme — verschiedene analytisch äquivalente differentiell-algebraische Formulierungen von Modellgleichungen (vgl. Abschnitt 3.1), so ist es im Fall von Index-2-Systemen aus Sicht der Störungstheorie günstig, ein DA–System mit möglichst kleinem $\mu = \|f_{zz}\|$ auszuwählen. Jedoch selbst in dem Spezialfall (2.49) mit $\mu = 0$, $f(\eta,\zeta) = f_0(\eta) - g_y^T(y)\zeta$ hängen die differentiellen Komponenten nicht stetig von Störungen ab, sobald g nichtlinear ist (vgl. jedoch Satz 3b und Satz 5b für den Fall, daß g linear, d. h. g_y konstant ist).

Beispiel 10 Das Anfangswertproblem

$$\begin{aligned} y' &= -g_y^T(y)z\,, \qquad (\,t\in[0,1]\,)\,,\\ 0 &= g(y)\,, \qquad\quad\; g(y) = g(y_1,y_2) = y_1 + y_2^2 - 1\end{aligned} \tag{2.51}$$

mit $y(0) = (\,0\,,\,1\,)^T$ und $z(0) = 0$ hat die analytische Lösung $y(t)\equiv y(0)$, $z(t)\equiv z(0)$. Das implizite Eulerverfahren löst (2.51) exakt: $y_n = y(t_n)$, $z_n = z(t_n)$. Trotz der sehr einfachen Struktur von (2.51) kann man für jeden (kleinen) positiven Parameter Θ eine Folge $(\hat y_n, \hat z_n)$ konstruieren, die in (2.33) Residuen $\delta = 0$, $\theta = \mathcal{O}(\Theta)$ hervorruft und in keiner Lösungskomponente stetig von den Störungen θ_n abhängt:
Gleichung (2.33) mit $\delta_{n+1} = 0$ und $\theta_{n+1} = (-1)^{n+1}\Theta$ ist für (2.51) äquivalent zu

$$\hat y_{n+1,1} = 1 + \theta_{n+1} - \hat y_{n+1,2}^2\,,\quad \hat z_{n+1} = -(\hat y_{n+1,1} - \hat y_{n,1})/h\,,\quad \varphi(\hat y_{n+1,2}, \hat y_{n,2}) = 2(-1)^{n+1}\Theta$$

mit der Funktion

$$\varphi(\eta,\eta^*) := (\eta - \eta^*)(\eta^* + \eta + \frac{1}{2\eta})\,. \tag{2.52}$$

Sei $nh = 1$ und $\hat y_{n,2} := \eta_n$. Im nachfolgenden Lemma 2 wird eine Folge (η_n) konstruiert, für die

$$\hat y_{n,2} \ge \hat y_{n-2,2} + \kappa\Theta^2 \ge \ldots \ge \hat y_{0,2} + \frac{n}{2}\kappa\Theta^2 = y_{n,2} + \mathcal{O}(\frac{1}{h}\theta^2)$$

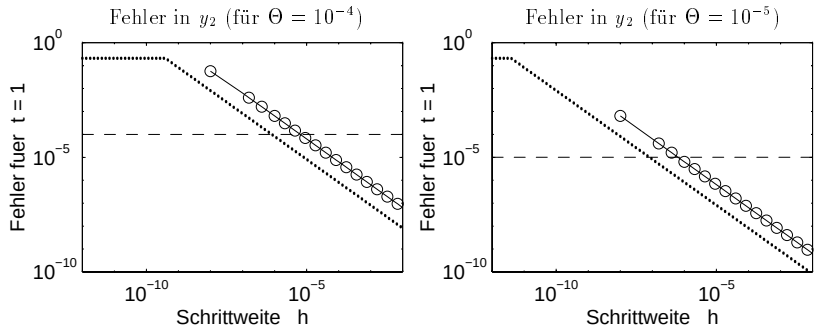


Abbildung 2.4: Globaler Fehler („o“) und untere Fehlerschranke („...“) für das implizite Eulerverfahren und ein Index-2-System mit $f(\eta,\zeta) = f_0(\eta) - g_y^T(y)\zeta$.

gilt oder aber $\hat y_{n,2} \ge Y$. Abb. 2.4 zeigt für $\Theta = 10^{-4}$ (links) und $\Theta = 10^{-5}$ (rechts) in Abhängigkeit von h den Fehler $|\hat y_{n,2} - y_{n,2}|$ zum Zeitpunkt $nh = 1$. Die gepunktete Linie deutet die (in Lemma 2 analytisch bewiesene) untere Schranke für den Fehler an: $\min\{\,Y - 1\,,\, n\kappa\Theta^2/2\,\}$, wobei die (frei wählbare) Konstante Y auf 1.2 gesetzt wurde.
Ebenso wie die Fehlerschranke aus Satz 5a hat der zusätzliche Fehlerterm in y_2 die Größenordnung $\mathcal{O}(\frac{1}{h}\theta^2)$. Unabhängig davon, wie klein der positive Parameter Θ gewählt wird, kann für kleines h der Fehler in jeder der Lösungskomponenten beliebig groß werden. Gleichzeitig verdeutlicht dieses Beispiel, daß der Fehlerterm $\mathcal{O}(\frac{1}{h}\theta^2)$ praktisch kaum relevant ist, weil er nur für außerordentlich kleines h die Größenordnung von θ übersteigt ($h < 10^{-6}$ für $\Theta = 10^{-5}$ in Abb. 2.4).

Lemma 2 *Ist ein $Y > 1$ gegeben, so gibt es stets positive Konstanten $\kappa = \kappa(Y)$, $\Theta_0 = \Theta_0(Y)$, so daß für beliebiges $\Theta \in (0,\Theta_0]$ eine Folge (η_n) mit*

$$\eta_0 = 1\,,\quad \eta_{n+1} = \eta_n + \mathcal{O}(\Theta)\,,\quad \varphi(\eta_{n+1},\eta_n) = 2(-1)^{n+1}\Theta\,,\ (n\ge 0)$$

definiert werden kann (vgl. (2.52)), für die gilt

$$\eta_{n+2} \ge \min(\,Y, \eta_n + \kappa\Theta^2\,)\,,\quad (\,n\ge 0\,,\ n\ \textit{gerade})\,.$$

Beweis Die Folge (η_n) wird rekursiv als Folge von Lösungen der nichtlinearen Gleichungen $\varphi(\eta,\eta_n) = 2(-1)^{n+1}\Theta$ definiert. Hierzu beweist man für jedes gerade $n \ge 0$ mit $\eta_n \le Y$ mittels Taylorentwicklung, daß $\varphi(\eta,\eta_n) - 2(-1)^{n+1}\Theta$ in

$$I_{n+1} := [\eta(-1),\eta(1)] \quad\text{mit}\quad \eta(\tau) := \eta_n + \beta(\eta_n)(-1)^{n+1}\Theta + (\alpha(\eta_n) + 2\kappa\tau)\Theta^2$$

eine Nullstelle hat, wenn $\Theta \in (0,\Theta_0]$ mit einem hinreichend kleinen $\Theta_0 > 0$ gilt. Dabei ist $\beta(\eta) := 4\eta/(1+4\eta^2)$, $\alpha(\eta) := -0.5\beta^3(\eta)(1 - 1/(2\eta^2))$ und $\kappa := \beta^3(Y)/(16Y^2)$. Für diese Nullstelle η_{n+1} zeigt man anschließend, daß $\varphi(\eta,\eta_{n+1}) - 2(-1)^{n+2}\Theta$ eine Nullstelle η_{n+2} in $I_{n+2} := [\eta_n + \kappa\Theta^2, \eta_n + (4+4\kappa)\Theta^2]$ hat usw. . Da die Restglieder der Taylorentwicklungen für $\eta_n > Y$ nicht gleichmäßig beschränkt sind, muß dieser Fall gesondert betrachtet werden.
Sämtliche Details dieses rein technischen Beweises enthält [28, Lemma 1]. ∎

Beispiel 10 zeigt ein weiteres interessantes Detail der Störungstheorie: Obwohl häufig die Fehlerschranken für die analytische und die numerische Lösung einander entsprechen, ist dies keineswegs immer der Fall. Die differentiellen Komponenten y der analytischen Lösung von (2.51) hängen nach Folgerung 1 stetig von Störungen ab. Das Fehlerwachstum in der numerischen Lösung entsteht für dieses Beispiel also erst durch die Diskretisierung.

Neben $f(\eta,\zeta) = f_0(\eta) + f_z(\eta)\zeta$ ist $n_z = 1$ eine entscheidende Voraussetzung von Folgerung 1. Sobald $n_z > 1$ ist, tritt das in Beispiel 10 beobachtete Fehlerwachstum i. allg. auch für die analytische Lösung ein:

Beispiel 11 Das Anfangswertproblem

$$\begin{array}{rcl} y' &=& -g_y^T(y)z \\ 0 &=& g(y) \end{array} \quad , \quad g(y) = g(y_1,y_2,y_3) = \begin{pmatrix} y_1^2 - y_2 \\ y_1 - y_3 \end{pmatrix} \quad , \quad (\, t \in [0,1]\,) \tag{2.53}$$

mit $y(0) = (\,0\,,\,0\,,\,0\,)^T$ und $z(0) = (\,0\,,\,0\,)^T$ hat die konstante Lösung $y(t) \equiv y(0)$, $z(t) \equiv z(0)$.
Seien positive Parameter Θ und ε gegeben, für die $\varepsilon \le \varepsilon_0$ und $\Theta \le \varepsilon_0 \varepsilon^{0.5}$ mit einer hinreichend kleinen Konstanten $\varepsilon_0 > 0$ gilt. Die Funktion $\theta(t) := (\,\Theta \sin \frac{t}{\varepsilon}\,,\, \Theta\cos\frac{t}{\varepsilon}\,)^T$ erfüllt $\|\theta\|_{C^0} = \mathcal{O}(\Theta)$, $\|\theta\|_{C^1} = \mathcal{O}(\frac{\Theta}{\varepsilon})$ und damit $\|\theta\|_{C^0} \ll 1$, $\|\theta\|_{C^0}\|\theta\|_{C^1} \ll 1$.
Wie beim Beweis der Existenz der Lösung von Anfangswertproblemen für gewöhnliche Differentialgleichungen (vgl. z. B. [158, Satz §12.VI]) folgt unter Verwendung von (2.11) und (2.14), daß es Funktionen $\hat y(t)$, $\hat z(t)$ mit $\hat y(0) = (\,0\,,\,0\,,\,-\Theta\,)^T$, $\hat z(0) = (\,-\frac{\Theta}{\varepsilon}\,,\,0\,)^T$ und

$$\begin{array}{rcl} \hat y'(t) &=& -g_y^T(\hat y)\hat z \\ \theta(t) &=& g(\hat y) \end{array}$$

gibt. Hier gilt $\theta'(t) = g_y(\hat y)\hat y'(t) = -[g_y g_y^T](\hat y)\hat z$ und $\hat y'(t) = [g_y^T(g_y g_y^T)^{-1}](\hat y)\cdot\theta'(t)$.
Zum Beweis von Satz 3 wird die Fehlerfortpflanzung in einem Term untersucht, der linear in $g(\eta)$ ist (vgl. (2.15)). Zum Nachweis, daß die Fehlerschranke (2.14) scharf ist, ist zusätzlich ein Term zu berücksichtigen, der in $g(\eta)$ quadratisch ist. Mit der in (2.19) eingeführten Tensorschreibweise gilt

$$\hat y(t) - y(t) = \Phi(\hat y(t)) - \Phi(y(t)) + \mathcal{O}(\Theta) \tag{2.54}$$

für

$$\Phi(\eta) := \eta - [g_y^T(g_y g_y^T)^{-1} g](\eta) + \frac{1}{2}[(\frac{\partial}{\partial y} g_y^T(g_y g_y^T)^{-1})(g, g_y^T(g_y g_y^T)^{-1} g)](\eta)\,,$$

denn $g(y(t)) = 0$ und $g(\hat y(t)) = \theta(t)$. Diese Funktion Φ ist bezüglich η stetig differenzierbar, man erhält $\frac{d}{dt}\Phi(y(t)) = 0$ (wegen $y'(t) = 0$) und

$$\frac{d}{dt}\Phi(\hat y(t)) = \Phi_\eta(\hat y)\hat y'(t) = \Phi_\eta(\hat y)[g_y^T(g_y g_y^T)^{-1}](\hat y)\cdot\theta'(t) = \Psi(\hat y(t))\theta'(t) + \mathcal{O}(\|g(\hat y)\|^2\|\theta'(t)\|)$$

mit der Matrix

$$\Psi(\eta) = \frac{1}{2}\Big([(\frac{\partial}{\partial y} g_y^T(g_y g_y^T)^{-1})(I, g_y^T(g_y g_y^T)^{-1} g)](\eta) - [(\frac{\partial}{\partial y} g_y^T(g_y g_y^T)^{-1})(g, g_y^T(g_y g_y^T)^{-1})](\eta)\Big)\,,$$

deren Elemente ψ_{kl} durch die Elemente ϕ_{kl} der Matrix $g_y^T(g_y g_y^T)^{-1}$ bestimmt sind:

$$\psi_{kl}(\eta) = \frac{1}{2}\sum_{j=1}^{n_z}\sum_{i=1}^{n_y}(\frac{\partial \phi_{kl}}{\partial y_i}\phi_{ij} - \frac{\partial \phi_{kj}}{\partial y_i}\phi_{il})g_j\,.$$

Für (2.53) ist $\Psi(\eta) = \lambda(\eta)\,(\,1\,,\,2\eta_1\,,\,1\,)^T\cdot(\,g_2(\eta)\,,\,-g_1(\eta)\,)$ mit $\lambda(\eta) := 1/(4(1+2\eta_1^2)^2)$, also

$$\frac{d}{dt}\Big(\Phi(\hat y(t)) - \Phi(y(t))\Big) = \lambda(\hat y(t))\begin{pmatrix} 1 \\ 2\hat y_1(t) \\ 1 \end{pmatrix}\cdot(\Theta\cos\frac{t}{\varepsilon}\,,\,-\Theta\sin\frac{t}{\varepsilon})\begin{pmatrix} \frac{\Theta}{\varepsilon}\cos\frac{t}{\varepsilon} \\ -\frac{\Theta}{\varepsilon}\sin\frac{t}{\varepsilon}\end{pmatrix} + \mathcal{O}(\frac{\Theta^3}{\varepsilon})$$

und die in (2.54) betrachtete Differenz $\hat y(t) - y(t)$ erfüllt deshalb

$$\begin{aligned} \hat y_1(1) - y_1(1) &= \hat y_1(0) - y_1(0) + \int_0^1 \frac{d}{dt}\Big(\Phi_1(\hat y(t)) - \Phi_1(y(t))\Big)\,dt + \mathcal{O}(\Theta)\,, \\ &= \int_0^1 \lambda(\hat y(t))\,dt\cdot\frac{\Theta^2}{\varepsilon} + \mathcal{O}(\Theta) + \mathcal{O}(\frac{\Theta^3}{\varepsilon})\,. \end{aligned}$$

Für $|\eta_1| \le 1$ ist $\frac{1}{36} \le \lambda(\eta) \le \frac{1}{4}$, so daß für $\varepsilon \le \varepsilon_0$, $\Theta \le \varepsilon_0\sqrt{\varepsilon}$ und hinreichend kleines $\varepsilon_0 > 0$ stets $\frac{1}{72}\frac{\Theta^2}{\varepsilon} \le \hat y_1(1) - y_1(1) \le \frac{1}{2}\frac{\Theta^2}{\varepsilon} \le \frac{1}{2}\varepsilon_0^2$ gilt. D. h., der Fehler in der differentiellen Komponente y_1 hängt (bezüglich $\|\cdot\|_{C^0}$) nicht stetig von Störungen $\theta(t)$ des algebraischen Teils ab. Wegen $\hat y_2 = \hat y_1^2 + \mathcal{O}(\Theta)$, $\hat y_3 = \hat y_1 + \mathcal{O}(\Theta)$ und $\hat z = \mathcal{O}(\frac{\Theta}{\varepsilon})$ gilt dies ebenso für alle anderen Lösungskomponenten.

2.3 Fehlerschranken für differentiell-algebraische Systeme vom Index 3 in Hessenbergform

Im Zusammenhang mit Konvergenzbeweisen für Runge–Kutta–Verfahren und BDF wurde auch für Index-3-Systeme in Hessenbergform die Fortpflanzung von Störungen während der Integration untersucht (vgl. die detaillierten Darstellungen in [81, Kapitel 6] und [93]). Ähnlich wie zuvor für Index-2-Systeme zeigen wir hier für die differentiellen Lösungskomponenten wesentlich schärfere Fehlerschranken als die zum Störungsindex $m = 3$ gehörende Schranke (2.6). Der Schwerpunkt liegt auf den in Kapitel 3 und 4 ausführlich betrachteten Modellgleichungen für mechanische Mehrkörpersysteme (MKS):

$$\begin{array}{rcl} M(q)q'' &=& f(q,q',\lambda) - G^T(q)\lambda\,, \quad q(0) = q_0\,, \quad q'(0) = v_0\,, \quad \lambda(0) = \lambda_0\,, \\ 0 &=& g(q)\,, \quad (\, t \in [0,T]\,)\,. \end{array} \tag{2.55}$$

Wie oben sei vorausgesetzt, daß das Anfangswertproblem eine hinreichend glatte Lösung $q(t)$, $v(t)$, $\lambda(t)$ mit $v(t) := q'(t)$ hat. In einer Umgebung $\mathcal{U}$ dieser Lösung seien f, g, $G(q) := g_q(q)$ und die symmetrische Matrixfunktion $M(q)$ (*Massenmatrix*) hinreichend oft stetig differenzierbar. Als Vereinfachung gegenüber Abschnitt 3.3.2 sei $M(q)$ in $\mathcal{U}$ regulär, ebenso die Matrix

$$[GM^{-1}\Gamma](q,v,\lambda) \quad \text{mit} \quad \Gamma(q,v,\lambda) := f_\lambda(q,v,\lambda) - G^T(q)\,. \tag{2.56}$$

Multipliziert man in (2.55) den differentiellen Teil mit $M^{-1}(q)$ (von links) und ersetzt $q' \to v$, $q'' \to v'$, so ergibt sich zusammen mit den kinematischen Gleichungen $q'(t) = v(t)$ ein zu (2.55) äquivalentes Index-3-System (1.11) in Hessenbergform, das auch den Störungsindex 3 hat ([81, S. 7f]).

In Abschnitt 2.3.1 beweisen wir Fehlerschranken für die analytische Lösung von (2.55), die wiederum stark von der Struktur von f abhängen. In dem praktisch wichtigen Spezialfall, daß f von λ unabhängig ist (mechanische Mehrkörpersysteme, in denen keine Reibungskräfte wirken), sind die differentiellen Komponenten q und v Index-2-Variable im Sinne des Störungsindexes. Hierfür zeigen wir in Abschnitt 2.3.2 exemplarisch entsprechende Fehlerschranken für die numerische Lösung, ein Testbeispiel illustriert darüberhinaus die Auswirkung von Reibungskräften im MKS.

2.3.1 Die Sensitivität der analytischen Lösung gegenüber kleinen Störungen

Gegeben seien Funktionen $\hat{q}$, $\hat{v} := \hat{q}'$ und $\hat{\lambda}$, die die Gleichungen (2.55) nicht exakt, sondern nur näherungsweise erfüllen:

$$\begin{aligned} M(\hat{q})\hat{q}'' &= f(\hat{q},\hat{q}',\hat{\lambda}) - G^T(\hat{q})\hat{\lambda} + \delta(t)\,, \quad \hat{q}(0) = \hat{q}_0\,,\ \ \hat{q}'(0) = \hat{v}_0\,,\ \ \hat{\lambda}(0) = \hat{\lambda}_0\,, \\ \theta(t) &= g(\hat{q})\ , \quad (\,t \in [0,T]\,)\,. \end{aligned} \tag{2.57}$$

Die Störungstheorie folgt den in Abschnitt 2.2 entwickelten Ideen, zur Vereinfachung der (ohnehin aufwendigen) Schreibweise gehen wir hierbei jedoch weit weniger ins Detail. So fordern wir hier für *sämtliche* Argumente $(q,v,\lambda) \in \mathcal{U}$, daß die Ableitungen von f, g und M durch eine Konstante $L = \mathcal{O}(1)$ gleichmäßig beschränkt sind, wiederum mit Ausnahme der Ableitungen von f bezüglich der algebraischen Variablen: hier wird nur

$$\|f_{v\lambda}(q,v,\lambda)\| \le \nu\,,\ \ \|f_{\lambda\lambda}(q,v,\lambda)\| \le \mu\,,\ (\,(q,v,\lambda) \in \mathcal{U}\,) \tag{2.58}$$

mit Konstanten $\nu, \mu \le L$ vorausgesetzt. (Ähnlich wie in Bemerkung 5 kann man auch für Index-3-Systeme (2.55) die Voraussetzungen an f, g und M abschwächen.) Wie in (2.12) wird ein Projektor auf den Tangentialraum von $\{\,\xi : g(\xi) = 0\,\}$ verwendet:

$$S(t,\xi) := I - [M^{-1}\Gamma(GM^{-1}\Gamma)^{-1}G](\xi, v(t), \lambda(t))\,. \tag{2.59}$$

Im Zentrum des Beweises für den nachfolgenden Satz 6 steht der Spezialfall $f = f(q,v)$; hier gilt $\mu = \nu = 0$ und der Projektor S hängt nicht explizit von t ab (vgl. auch Bemerkung 11c).

Satz 6 *Es gibt eine Konstante C, so daß für alle Funktionen $\hat{q}$, $\hat{v}$, $\hat{\lambda}$ mit (2.57), $\delta \in C^0([0,T])$, $\theta \in C^2([0,T])$ und $\|\delta\|_{C^0([0,T])} \le C_\delta$, $\|\theta\|_{C^2([0,T])} \le C_\theta$ die Abschätzungen*

$$\begin{aligned} \|\hat{q}(t) - q(t)\| + \|S(t,\hat{q}(t))\hat{v}(t) - S(t,q(t))v(t)\| \le C\Big(\Delta_0 + \int SM^{-1}\delta + \\ + \|\theta\|_{C^0} + D^2(t) + \nu D(t)(\|\delta\|_{C^0} + \|\theta\|_{C^2}) + \mu(\|\delta\|_{C^0} + \|\theta\|_{C^2})^2\Big)\,, \end{aligned} \tag{2.60}$$

$$\|[G(\hat{q})\hat{v}](t) - [G(q)v](t)\| \le C \cdot \|\theta\|_{C^1}\,, \tag{2.61}$$

$$\|\hat{\lambda}(t) - \lambda(t)\| \le C(\Delta_0 + \|\delta\|_{C^0} + \|\theta\|_{C^2}) \tag{2.62}$$

mit

$$\Delta_0 := \|\hat{q}_0 - q_0\| + \|S(0,\hat{q}_0)\hat{v}_0 - S(0,q_0)v_0\|\,, \quad D(t) := \int SM^{-1}\delta + \|\theta\|_{C^1([0,t])}\,,$$

$$\int SM^{-1}\delta := \max_{\tau\in[0,t]} \|\int_0^\tau [SM^{-1}](w,q(w))\delta(w)\,dw\|$$

für alle $t \in [0,T]$ erfüllt sind, wenn die rechte Seite von (2.60) hinreichend klein ist. (C ist unabhängig von $\hat{q}$, $\hat{v}$, $\hat{\lambda}$, μ, ν und den Störungen δ, θ und wird wie in Satz 3 durch L, T, C_δ und C_θ bestimmt. Im Spezialfall $\mu = \nu = 0$ ist C darüberhinaus auch unabhängig von C_θ.)

Beweis Teile des Beweises können in direkter Analogie zum Beweis von Satz 3 geführt werden, so daß wir uns hier auf die neu hinzukommenden Beweisgedanken beschränken. Insbesondere werden wie in (2.16) nur $(\hat{q},\hat{v},\hat{\lambda})$ in einer hinreichend kleinen Umgebung der analytischen Lösung betrachtet. Die Fehlerfortpflanzung untersucht man in den beiden Funktionen

$$\begin{aligned} \Delta^q(t) &:= S(t,q(t))(\hat{q}(t) - q(t)) \\ \Delta^v(t) &:= S(t,\hat{q}(t))\hat{v}(t) - S(t,q(t))v(t) - (\Phi(t,\hat{q}(t)) - \Phi(t,q(t))) \end{aligned} \tag{2.63}$$

mit

$$\begin{aligned} \Phi(t,\eta) := \frac{\partial S}{\partial t}(t,q(t))\eta + [SM^{-1}f_v](q(t),v(t),\lambda(t))\eta + \\ + \frac{\partial S}{\partial q}(t,q(t))(v(t),\eta) + \frac{\partial S}{\partial q}(t,q(t))(\eta,v(t)) \end{aligned} \tag{2.64}$$

(für die Tensornotation vgl. (2.19)). Im weiteren verzichten wir soweit wie möglich auf die explizite Angabe des Arguments t.
Die analytische Lösung von (2.55) erfüllt neben den Zwangsbedingungen $g(q) = 0$ auch

$$0 = \frac{d}{dt}g(q(t)) = g_q(q)q'(t) = G(q(t))v(t)\,, \tag{2.65}$$

$$0 = \frac{d^2}{dt^2}g(q(t)) = g_{qq}(q)(v,v) + [GM^{-1}](q)(f(q,v,\lambda) - G^T(q)\lambda)\,. \tag{2.66}$$

Aus (2.65), (2.66) und den entsprechenden Gleichungen für $(\hat{q},\hat{v},\hat{\lambda})$, in denen auf der linken Seite der Gleichungen die Ableitungen $\theta'(t)$ bzw. $\theta''(t)$ stehen, folgt einerseits $[G(\hat{q})\hat{v}](t) = \theta'(t)$, also (2.61), und andererseits wegen (2.56) auch

$$\|\hat{\lambda}(t) - \lambda(t)\| \le \mathcal{O}(1)(\|\hat{q}(t) - q(t)\| + \|\hat{v}(t) - v(t)\| + \|\delta\|_{C^0([0,t])} + \|\theta\|_{C^2([0,t])})\,. \tag{2.67}$$

Wie in (2.26) erhält man außerdem

$$\|\hat{q}(t) - q(t)\| \le \mathcal{O}(1)(\|\Delta^q(t)\| + \|\theta\|_{C^0})\,, \tag{2.68}$$

$$\begin{aligned} \|\hat{v}(t) - v(t)\| &\le \|\Delta^v(t)\| + \|\Phi(t,\hat{q}) - \Phi(t,q)\| + \mathcal{O}(1)\|[G(\hat{q})\hat{v}](t)\| \\ &\le \|\Delta^v(t)\| + \mathcal{O}(1)\|\hat{q}(t) - q(t)\| + \mathcal{O}(\|\theta\|_{C^1})\,. \end{aligned} \tag{2.69}$$

Nach (2.17) sind der entscheidende Beweisschritt Abschätzungen für $\frac{d}{dt}\Delta^q(t)$ und $\frac{d}{dt}\Delta^v(t)$:

$$\begin{aligned}\frac{d}{dt}\Delta^q(t) &= S(t,\hat q)\hat q' - S(t,q)q' + (S(t,q)-S(t,\hat q))\hat q' + \\ &\quad + \frac{\partial S}{\partial t}(t,q)(\hat q - q) + \frac{\partial S}{\partial q}(t,q)(\hat q - q, q') \\ &= \Delta^v(t) + \mathcal{O}(1)\|\hat q(t) - q(t)\| \,, \end{aligned} \tag{2.70}$$

denn $S(t,\eta)$ und $\Phi(t,\eta)$ sind Lipschitz-stetig bezüglich η. In

$$\begin{aligned}\frac{d}{dt}\Delta^v(t) &= S(t,\hat q)\hat v' - S(t,q)v' + \frac{\partial S}{\partial t}(t,\hat q)\hat v - \frac{\partial S}{\partial t}(t,q)v + \frac{\partial S}{\partial q}(t,\hat q)(\hat v,\hat q') - \frac{\partial S}{\partial q}(t,q)(v,q') \\ &\quad - \Big(\frac{\partial \Phi}{\partial t}(t,\hat q) - \frac{\partial \Phi}{\partial t}(t,q) + \frac{\partial \Phi}{\partial q}(t,\hat q)\hat q' - \frac{\partial \Phi}{\partial q}(t,q)q'\Big)\end{aligned}$$

setzt man

$$\begin{aligned}\frac{\partial \Phi}{\partial q}(t,\hat q)\hat q' - \frac{\partial \Phi}{\partial q}(t,q)q' &= \frac{\partial S}{\partial t}(t,q)(\hat v - v) + [SM^{-1}f_v](q,v,\lambda)(\hat v - v) + \\ &\quad + \frac{\partial S}{\partial q}(t,q)(v,\hat v - v) + \frac{\partial S}{\partial q}(t,q)(\hat v - v, v)\end{aligned}$$

und

$$\begin{aligned}&S(t,\hat q)\hat v' - S(t,q)v' = \\ &= [SM^{-1}](t,\hat q)(f(\hat q,\hat v,\hat\lambda) - G^T(\hat q)\hat\lambda + \delta(t)) - [SM^{-1}](t,q)(f(q,v,\lambda) - G^T(q)\lambda) \\ &= [SM^{-1}](t,\hat q)(f(\hat q,v,\lambda) - G^T(\hat q)\lambda) - [SM^{-1}](t,q)(f(q,v,\lambda) - G^T(q)\lambda) + \\ &\quad + [SM^{-1}](t,q)\delta(t) + \mathcal{O}(\|\delta(t)\|)\|\hat q - q\| + \\ &\quad + [SM^{-1}](t,\hat q)\Big((f(\hat q,\hat v,\hat\lambda) - G^T(\hat q)\hat\lambda) - (f(\hat q,v,\lambda) - G^T(\hat q)\lambda)\Big)\end{aligned}$$

ein und erhält unter Verwendung von $[SM^{-1}](t,\hat q)\Gamma(\hat q,v,\lambda) \equiv 0$

$$\begin{aligned}\frac{d}{dt}\Delta^v(t) &= [SM^{-1}](t,q)\delta(t) + \mathcal{O}(1)\|\hat q - q\| + \frac{\partial S}{\partial q}(t,q)(\hat v - v, \hat v - v) + \\ &\quad + [SM^{-1}](t,\hat q)\Big((f(\hat q,\hat v,\hat\lambda) - G^T(\hat q)\hat\lambda) - (f(\hat q,v,\lambda) - G^T(\hat q)\lambda) - \\ &\qquad - f_v(\hat q,v,\lambda)(\hat v - v) - \Gamma(\hat q,v,\lambda)(\hat\lambda - \lambda)\Big) \,, \\ &= [SM^{-1}](t,q(t))\delta(t) + \\ &\quad + \mathcal{O}(1)(\|\hat q - q\| + \|\hat v - v\|^2 + \nu\|\hat v - v\|\|\hat\lambda - \lambda\| + \mu\|\hat\lambda - \lambda\|^2) \,.\end{aligned}$$

Schließlich folgt wegen (2.67), (2.68) und (2.69)

$$\begin{aligned}\frac{d}{dt}\Delta^v(t) &= [SM^{-1}](t,q(t))\delta(t) + \mathcal{O}(1)(\|\Delta^q(t)\| + \|\Delta^v(t)\|) + \\ &\quad + \mathcal{O}(\|\theta\|_{C^0} + D^2(t) + \nu D(t)(\|\delta\|_{C^0} + \|\theta\|_{C^2}) + \mu(\|\delta\|_{C^0} + \|\theta\|_{C^2})^2) \,.\end{aligned} \tag{2.71}$$

Unter Verwendung des Lemmas von Gronwall wird nun mit (2.70) und (2.71) sowie (2.67) und (2.68) die Behauptung bewiesen (analog zum Teil a) des Beweises von Satz 3). ∎

Bemerkung 11 a) Ebenso wie für Index-2-Systeme erhält man für die differentiellen Komponenten deutlich kleinere Fehlerschranken als in der Definition des Störungsindexes. Die gegenüber Störungen robustesten Lösungskomponenten sind q und $S(t,q)v$. I. allg. hängt jedoch keine Komponente stetig von Störungen ab. Besonders klein ist die Schranke (2.60), falls $\mu,\nu \ll 1$ (z. B. MKS mit sehr kleinen Reibungskräften). Ist f von λ unabhängig (z. B. MKS mit Potentialkräften), so erhält man wegen $\mu = \nu = 0$ für die Komponenten q und v sogar die Abschätzung (2.6) mit $r = 1$, d. h., für Index-3-Systeme (2.55) mit $f = f(q,v)$ sind q und v *Index-2-Variable* im Sinne des Störungsindexes ([17, Satz 3]). Jay ([93]) zeigt für Index-3-Systeme in Hessenbergform, daß der Mechanismus der Fehlerfortpflanzung auch für implizite Runge–Kutta–Verfahren davon beeinflußt wird, ob das DA–System linear oder nichtlinear von den algebraischen Variablen abhängt. Ist das System linear in den algebraischen Variablen, so ist der globale Diskretisierungsfehler in den differentiellen Komponenten i. allg. deutlich kleiner als im nichtlinearen Fall.

b) Satz 6 nutzt explizit die Struktur von (2.55) aus. Die Aussagen lassen sich übertragen auf MKS–Modellgleichungen mit $q' = T(t,q)v$ (vgl. Abschnitt 3.3.2), der Beweis wird dabei technisch noch aufwendiger. Für allgemeine Index-3-Systeme in Hessenbergform gibt Satz 10 aus [17] entsprechende Fehlerabschätzungen an, die in der Notation von (2.58) den beiden Fällen $\mu = \mathcal{O}(1)$, $\nu = \mathcal{O}(1)$ und $\mu = 0$, $\nu = \mathcal{O}(1)$ entsprechen. Für die differentiellen Komponenten haben diese Schranken die Form (2.60) bzw. (2.61), dabei ist jedoch $\|\theta\|_{C^0}$ durch $\|\theta\|_{C^1}$ zu ersetzen. Für die algebraischen Komponenten erhält man wie in Satz 6 die Schranke aus Definition 4 mit $m = 3$.

c) Im Spezialfall $\mu = \nu = 0$ ist $S = S(q)$ und deshalb hängt $S(\hat q)\hat v'$ nicht von $\hat\lambda$ ab. Dann ist (2.55) äquivalent zu (2.66) und dem Index-2-System in Hessenbergform

$$\begin{aligned}q' &= v \\ (S(q)v)' &= [SM^{-1}f](q,v) + \frac{\partial S}{\partial q}(q)(v,v) \\ 0 &= g(q)\end{aligned} \tag{2.72}$$

mit den differentiellen Komponenten $y := (\,q^T\,,\,(S(q)v)^T\,)^T$ und den algebraischen Komponenten $z := G(q)v$ ([17, Abschnitt 2]). Hier ist Satz 6 eine unmittelbare Folgerung aus Satz 3.

Beispiel 12 Wie in Abschnitt 2.2.3 kann man am Beispiel nachweisen, daß die Fehlerschranken aus Satz 6 scharf sind. Als Erweiterung von Beispiel 9 betrachten wir das Anfangswertproblem $q(0) = q'(0) = 0$, $\lambda(0) = 0$ für (2.55) mit $M(q) = I$,

$$f(q,v,\lambda) = (\,0\,,\,0\,,\,v_1^2 + v_2^2 + \nu(v_1\lambda_2 - v_2\lambda_1) + \mu(\lambda_1^2 + \lambda_2^2)\,)^T, \quad g(q) = g(q_1,q_2,q_3) = (\,q_1\,,\,q_2\,)^T,$$

dessen Lösung identisch verschwindet. Für die Funktionen

$$\hat q(t) = \Big(\,\Theta\sin\frac{t}{\varepsilon}\,,\,\Theta\cos\frac{t}{\varepsilon}\,,\,\frac{1}{2}\Big(\frac{\Theta^2}{\varepsilon^2} + \nu\frac{\Theta^2}{\varepsilon^3} + \mu\frac{\Theta^2}{\varepsilon^4}\Big)t^2\Big)^T, \quad \hat\lambda(t) = \Big(\,\frac{\Theta}{\varepsilon^2}\sin\frac{t}{\varepsilon}\,,\,\frac{\Theta}{\varepsilon^2}\cos\frac{t}{\varepsilon}\,\Big)^T$$

gilt in (2.57) $\delta(t) \equiv 0$, $\theta(t) = (\,\Theta\sin\frac{t}{\varepsilon}\,,\,\Theta\cos\frac{t}{\varepsilon}\,)^T$, also $\|\theta\|_{C^0} = \mathcal{O}(\Theta)$, $\|\theta\|_{C^1} = \mathcal{O}(\frac{\Theta}{\varepsilon})$, $\|\theta\|_{C^2} = \mathcal{O}(\frac{\Theta}{\varepsilon^2})$. Wenn man $\Theta > 0$ fixiert und $\varepsilon > 0$ hinreichend klein wählt, dann zeigen die Fehler in q_3, v_3 (entspricht $S(t,q)v$), in v_1 (entspricht $G(q)v$) und in λ, daß die

Fehlerschranken (2.60), (2.61) und (2.62) i. allg. nicht verbessert werden können. Eine Ausnahme bilden die in praxi häufig betrachteten linearen DA–Systeme (2.55) mit konstanten Koeffizienten: Hier hängen q und Sv stetig von kleinen Störungen ab (vgl. Satz 3b und (2.72)) und sind deshalb sogar *Index-1-Variable* im Sinne von Definition 4.

2.3.2 Die Sensitivität der numerischen Lösung gegenüber kleinen Störungen: eine Fallstudie

Aus der Literatur bekannte Fehlerabschätzungen für die numerische Lösung von Index-3-Systemen wurden im Zusammenhang mit Konvergenzuntersuchungen für BDF (z. B. [104] für (2.55) mit $f = f(q,v)$) und für implizite Runge–Kutta–Verfahren ([81, Kapitel 6], [93]) bewiesen, eine typische Voraussetzung ist dabei $\theta = \mathcal{O}(h^3)$. Wir zeigen hier — wiederum unter möglichst schwachen Voraussetzungen an die Größe der Störungen — das direkte Gegenstück zu Satz 6 für das implizite Eulerverfahren

$$\begin{aligned} \frac{q_{n+1}-q_n}{h} &= v_{n+1} \\ M(q_{n+1})\frac{v_{n+1}-v_n}{h} &= f(q_{n+1},v_{n+1}) - G^T(q_{n+1})\lambda_{n+1} \\ 0 &= g(q_{n+1}) \end{aligned} \tag{2.73}$$

im Spezialfall $f = f(q,v)$ (d. h. $\mu = \nu = 0$ in (2.58)). Mit dieser Einschränkung ist (2.73) wohldefiniert und hat die Konvergenzordnung 1 ([104], [40], vgl. auch den Beweis von Satz 8). Hängt dagegen f auch von λ ab, so ist nicht garantiert, daß das nichtlineare Gleichungssystem (2.73) lösbar ist ([81, S. 75]).

Sei also $f = f(q,v)$. Wie in Bemerkung 11c angedeutet, kann hier die Äquivalenz von (2.55) und (2.72) ausgenutzt werden, wobei (2.73) jedoch nicht äquivalent zur Anwendung des impliziten Eulerverfahrens auf das Index-2-System (2.72) ist. Statt dessen folgt für $\Psi(\vartheta) := S(q_n + \vartheta h v_{n+1})v_n$ aus $\Psi(1) - \Psi(0) = \int_0^1 \Psi'(\vartheta)\,d\vartheta$

$$(S(q_{n+1}) - S(q_n))v_n = h\int_0^1 \frac{\partial S}{\partial q}(q_n + \vartheta h v_{n+1})(v_n, v_{n+1})\,d\vartheta$$

und damit ist (2.73) äquivalent zu $\frac{1}{h}(q_{n+1}-q_n) = v_{n+1}$, $g(q_{n+1}) = 0$,

$$\frac{S(q_{n+1})v_{n+1} - S(q_n)v_n}{h} = [SM^{-1}f](q_{n+1},v_{n+1}) + \int_0^1 \frac{\partial S}{\partial q}(q_n+\vartheta h v_{n+1})(v_n,v_{n+1})\,d\vartheta \tag{2.74}$$

und

$$\lambda_{n+1} = [(GM^{-1}G^T)^{-1}](q_{n+1})\Big([GM^{-1}f](q_{n+1},v_{n+1}) - G(q_{n+1})\frac{v_{n+1}-v_n}{h}\Big)\,. \tag{2.75}$$

Die rechte Seite von (2.74) enthält neben q_{n+1}, v_{n+1} auch q_n, v_n, die Verfahrensvorschrift (2.73) ergibt also eine Diskretisierung von (2.72), die teilweise implizit und teilweise explizit ist. Hiervon unberührt kann die Lösbarkeit des nichtlinearen Gleichungssystems (2.73) mit nur geringfügigen Modifikationen ebenso wie für Index-2-Systeme (Satz 4) bewiesen werden:

Satz 7 *Gegeben seien (q_n, v_n) mit*

$$\|q_n - q(t_n)\| + \|v_n - v(t_n)\| \le \gamma \quad \text{und} \quad \Delta := \|g(q_n)\| + h\|G(q_n)v_n\| \le \Delta_0 h\,.$$

Sind die (positiven) Konstanten h_0, Δ_0 und γ hinreichend klein, so ist (2.73) für alle $h \in (0,h_0]$ (lokal) eindeutig nach $(q_{n+1}, v_{n+1}, \lambda_{n+1})$ auflösbar und es gilt

$$\begin{aligned} \|q_{n+1} - q_n\| + \|S(q_{n+1})v_{n+1} - S(q_n)v_n\| &\le C_{h\Delta}(h+\Delta)\,, \\ \|G(q_{n+1})v_{n+1}\| &\le C_{h\Delta}(h + \frac{\Delta}{h}) \end{aligned}$$

mit einer von h, Δ und (q_n, v_n) unabhängigen Konstanten $C_{h\Delta}$.

Wie in Abschnitt 2.2.2 kann dieser Satz erweitert werden, um zu zeigen, daß es für hinreichend kleine Schrittweiten $h > 0$ unter allen Folgen $(\hat q_n, \hat v_n, \hat\lambda_n)$ mit

$$\begin{aligned} \frac{\hat q_{n+1} - \hat q_n}{h} &= \hat v_{n+1} \\ M(\hat q_{n+1})\frac{\hat v_{n+1} - \hat v_n}{h} &= f(\hat q_{n+1}, \hat v_{n+1}) - G^T(\hat q_{n+1})\hat\lambda_{n+1} + \delta_{n+1} \\ \theta_{n+1} &= g(\hat q_{n+1}) \end{aligned} \tag{2.76}$$

stets solche gibt, die die Abschätzungen

$$\begin{aligned} \|\hat q_{n+1} - \hat q_n\| + \|S(\hat q_{n+1})\hat v_{n+1} - S(\hat q_n)\hat v_n\| &\le C^*_{h\Delta}(h + \Delta^*)\,, \\ \|G(\hat q_{n+1})\hat v_{n+1}\| &\le C^*_{h\Delta}(h + \frac{\Delta^*}{h}) \end{aligned} \tag{2.77}$$

mit

$$\Delta^* := \|g(\hat q_n)\| + h\|G(\hat q_n)\hat v_n\| + h\delta + \theta\,,$$
$$\delta := \max_m \|\delta_m\| + \|G(\hat q_0)\hat v_0\|\,, \quad \theta := \max_m \|\theta_m\| + \|g(\hat q_0)\|$$

und einer von h, δ und θ unabhängigen Konstanten $C^*_{h\Delta}$ erfüllen, sofern $\Delta^* \le \Delta_0 h$ mit hinreichend kleinem $\Delta_0 > 0$ gilt.

Da die kinematischen Gleichungen $q' = v$ linear sind, können sie während der Integration sehr genau gelöst werden. Ebenso wie in Satz 6 vernachlässigen wir deshalb zur Vereinfachung der Darstellung die Fehler in $\frac{1}{h}(\hat q_{m+1} - \hat q_m) - \hat v_{m+1} = 0$ (andernfalls wäre in Satz 8 jeweils θ durch $\theta + h\max_m \|\frac{1}{h}(\hat q_{m+1} - \hat q_m) - \hat v_{m+1}\|$ zu ersetzen).

Für $(\hat q_n, \hat v_n, \hat\lambda_n)$ kann man die Fehlerschranken aus Satz 6 (mit $\mu = \nu = 0$) direkt auf die numerische Lösung übertragen:

Satz 8 *Es gibt eine Konstante C_0, so daß für alle Folgen $(\hat q_n, \hat v_n, \hat\lambda_n)$ mit (2.76) und (2.77) und $\delta \le C_\delta$ die Abschätzungen*

$$\begin{aligned} \|\hat q_n - q_n\| + \|S(\hat q_n)\hat v_n - S(q_n)v_n\| &\le C_0\Big(\Delta_0 + \int_h SM^{-1}\delta + \theta + (\delta + \frac{1}{h}\theta)^2\Big)\,, \\ \|G(\hat q_n)\hat v_n - G(q_n)v_n\| &\le C_0(\Delta_0 + \delta + \frac{1}{h}\theta)\,, \\ \|\hat\lambda_n - \lambda_n\| &\le C_0(\Delta_0 + \delta + \frac{1}{h^2}\theta) \end{aligned} \tag{2.78}$$

für alle $n \geq 0$ mit $nh \leq T$ erfüllt sind, sofern die Schrittweite h und die rechte Seite von (2.78) hinreichend klein sind. Hier bezeichnet $\Delta_0 := \|\hat{q}_0 - q_0\| + \|S(\hat{q}_0)\hat{v}_0 - S(q_0)v_0\|$ und

$$\int_h SM^{-1}\delta := h\sum_{m=0}^{n-1} \|[SM^{-1}](q(t_m))\delta_{m+1}\| + h\delta \ .$$

Die Konstante C_0 ist von h, δ und θ unabhängig und wird i. allg. durch L, T und C_δ bestimmt.

Beweisskizze Der vollständige Beweis ist wegen der schwachen Voraussetzungen an die Größe der Störungen außerordentlich umfangreich. Wir zeigen hier deshalb nur die grundlegenden Beweisgedanken, zahlreiche Details können analog zu den Beweisen der Sätze 5 und 6 nachgewiesen werden.
Für (q_n, v_n), $(\hat{q}_n, \hat{v}_n)$ in einer hinreichend kleinen Umgebung der analytischen Lösung wird die Fehlerfortpflanzung in den Folgen

$$\begin{aligned} \Delta_m^q &:= S(q_m)(\hat{q}_m - q_m) \\ \Delta_m^v &:= S(\hat{q}_m)\hat{v}_m - S(q_m)v_m - (\Phi_m(\hat{q}_m, \hat{v}_m) - \Phi_m(q_m, v_m)) \end{aligned} \tag{2.79}$$

mit

$$\begin{aligned} \Phi_m(\eta,\xi) := [SM^{-1}f_v](q(t_m), v(t_m))\eta + \frac{\partial S}{\partial q}(q(t_m))\Big(v(t_m), T(q(t_m))g(\eta)\Big) + \\ + \frac{\partial S}{\partial q}(q(t_m))\Big(T(q(t_m))(g(\eta) - hG(\eta)\xi), v(t_m)\Big) \end{aligned} \tag{2.80}$$

und $T(q) := [M^{-1}G^T(GM^{-1}G^T)^{-1}](q)$ untersucht (vgl. (2.63)). Wegen $\mu = \nu = 0$ ist $\partial S/\partial t \equiv 0$ und es gilt $\Gamma(q, v, \lambda) = -G^T(q)$ und $T(q)G(q) = I - S(q)$.
Wie für die analytische Lösung sind die Komponenten $G(q)v$ und λ durch q und $S(q)v$ bestimmt: Wendet man $\Psi(1) - \Psi(0) = \int_0^1 \Psi'(\vartheta)\, d\vartheta$ auf $\Psi(\vartheta_1) := g(q_{m+1} + \vartheta_1(q_m - q_{m+1}))$ und auf $\tilde{\Psi}(\vartheta_2) := G(q_{m+1} - h\vartheta_2\vartheta_1 v_{m+1})v_{m+1}$ an, so folgt

$$\begin{aligned} G(q_{m+1})v_{m+1} &= \int_0^1 \Big(G(q_{m+1}) - G(q_{m+1} - h\vartheta_1 v_{m+1})\Big)v_{m+1}\, d\vartheta_1 + \\ &\qquad + \int_0^1 G(q_{m+1} + \vartheta_1(q_m - q_{m+1}))\frac{q_{m+1} - q_m}{h}\, d\vartheta_1 \\ &= -h\int_0^1\int_0^1 g_{qq}(q_{m+1} - h\vartheta_2\vartheta_1 v_{m+1})(v_{m+1}, \vartheta_1 v_{m+1})\, d\vartheta_2 d\vartheta_1 + \frac{g(q_{m+1}) - g(q_m)}{h} \end{aligned}$$

und die entsprechenden Gleichungen für $G(\hat{q}_{m+1})\hat{v}_{m+1}$, also

$$G(\hat{q}_{m+1})\hat{v}_{m+1} - G(q_{m+1})v_{m+1} = \frac{\theta_{m+1} - \theta_m}{h} + \mathcal{O}(h)(\|\hat{q}_{m+1} - q_{m+1}\| + \|\hat{v}_{m+1} - v_{m+1}\|) \ . \tag{2.81}$$

Auf gleiche Weise erhält man aus (2.75)

$$\begin{aligned} \hat{\lambda}_{m+1} - \lambda_{m+1} = -[(GM^{-1}G^T)^{-1}](\hat{q}_{m+1})\frac{\theta_{m+1} - 2\theta_m + \theta_{m-1}}{h^2} + \\ + \mathcal{O}(1)(\|\hat{q}_{m+1} - q_{m+1}\| + \|\hat{q}_m - q_m\| + \|\hat{v}_{m+1} - v_{m+1}\| + \|\hat{v}_m - v_m\| + \delta) \ . \end{aligned} \tag{2.82}$$

Wegen (2.81) können die Fehler in q und v durch $\|\Delta_m^q\|$ und $\|\Delta_m^v\|$ ausgedrückt werden (vgl. (2.68), (2.69)):

$$\begin{aligned} \|\hat{q}_m - q_m\| &\leq \mathcal{O}(1)(\|\Delta_m^q\| + \theta) \ , \\ \|S(\hat{q}_m)\hat{v}_m - S(q_m)v_m\| &\leq \mathcal{O}(1)(\|\Delta_m^q\| + \|\Delta_m^v\|) + \mathcal{O}(\theta) \ , \\ \|\hat{v}_m - v_m\| &\leq \mathcal{O}(1)(\|\Delta_m^q\| + \|\Delta_m^v\|) + \mathcal{O}(\frac{1}{h}\theta) \ . \end{aligned} \tag{2.83}$$

Für die diskreten Analoga zu $\frac{d}{dt}\Delta^q(t)$ und $\frac{d}{dt}\Delta^v(t)$ (vgl. (2.70), (2.71)) beweist man die Abschätzungen

$$\begin{aligned} \frac{1}{h}(\Delta_{m+1}^q - \Delta_m^q) &\leq \mathcal{O}(1)(\|\Delta_{m+1}^q\| + \|\Delta_m^q\| + \|\Delta_{m+1}^v\| + \theta) \ , \\ \frac{1}{h}(\Delta_{m+1}^v - \Delta_m^v) &= [SM^{-1}](q(t_m))\delta_{m+1} + \mathcal{O}(1)(\theta + (h + d_m)(\delta + \tfrac{1}{h}\theta)) + \\ &\qquad + \mathcal{O}(1)(\|\Delta_{m+1}^q\| + \|\Delta_m^q\| + \|\Delta_{m+1}^v\| + \|\Delta_m^v\|) \end{aligned} \tag{2.84}$$

mit

$$d_m := \epsilon_{m+1}^q + \hat{\epsilon}_{m+1}^q + \epsilon_m^q + \hat{\epsilon}_m^q + \epsilon_{m+1}^v + \hat{\epsilon}_{m+1}^v + \epsilon_m^v + \hat{\epsilon}_m^v \tag{2.85}$$

und $\epsilon_m^q := \|q_m - q(t_m)\|$, $\hat{\epsilon}_m^q := \|\hat{q}_m - q(t_m)\|$, $\epsilon_m^v := \|v_m - v(t_m)\|$ usw. .
Während dabei $\frac{1}{h}(\Delta_{m+1}^q - \Delta_m^q)$ und die meisten Terme in $\frac{1}{h}(\Delta_{m+1}^v - \Delta_m^v)$ analog zum Beweis von Satz 6 abgeschätzt werden können, erfordern die Ausdrücke, die $\partial S/\partial q$ enthalten, eine gesonderte Betrachtung. Für $\frac{\partial S}{\partial q}(q)(v, T(q)g(\eta))$ zeigt (2.81)

$$\begin{aligned} &\frac{1}{h}T(q(t_{m+1}))(g(\hat{q}_{m+1}) - g(\hat{q}_m)) = \\ &\quad = (T(q(t_{m+1})) - T(q_{m+1}) + T(q_{m+1}))(G(\hat{q}_{m+1})\hat{v}_{m+1} - G(q_{m+1})v_{m+1}) + \\ &\qquad + \mathcal{O}(h)(\|\hat{q}_{m+1} - q_{m+1}\| + \|\hat{v}_{m+1} - v_{m+1}\|) \ , \\ &\quad = (I - S(\hat{q}_{m+1}))\hat{v}_{m+1} - (I - S(q_{m+1}))v_{m+1} + \mathcal{O}(1)\|\hat{q}_{m+1} - q_{m+1}\| + \\ &\qquad + \mathcal{O}(h)\|\hat{v}_{m+1} - v_{m+1}\| + \mathcal{O}(d_m)\|G(\hat{q}_{m+1})\hat{v}_{m+1} - G(q_{m+1})v_{m+1}\| \ , \\ &\quad = \hat{v}_{m+1} - v_{m+1} + \mathcal{O}(1)(\|\Delta_{m+1}^q\| + \|\Delta_{m+1}^v\| + \theta) + \\ &\qquad + \mathcal{O}(h + d_m)\|G(\hat{q}_{m+1})\hat{v}_{m+1} - G(q_{m+1})v_{m+1}\| \ , \end{aligned}$$

deshalb entspricht der zweite Summand auf der rechten Seite von (2.80) dem Term $\frac{\partial S}{\partial q}(t, q)(v, \eta)$ in (2.64).
Ebenso entspricht der dritte Summand in (2.80) dem Term $\frac{\partial S}{\partial q}(t, q)(\eta, v)$ in (2.64), wobei zusätzlich beachtet werden muß, daß v hier auf der Zeitschicht t_m zu betrachten ist ($\frac{\partial S}{\partial q}(q)(v_m, v_{m+1})$ in (2.74)). Deshalb enthält der zweite Summand auf der rechten Seite von (2.80) den Vektor $g(\eta)$, der dritte jedoch den Vektor $g(\eta) - hG(\eta)\xi$, denn

$$g(\hat{q}_m) = g(\hat{q}_{m+1}) - h(G(\hat{q}_{m+1})\hat{v}_{m+1} - G(q_{m+1})v_{m+1}) + \mathcal{O}(h)(\|\hat{q}_{m+1} - q_{m+1}\| + \|\hat{v}_{m+1} - v_{m+1}\|)$$

(vgl. (2.81)).

Zum Abschluß des Beweises zeigt man unter Verwendung von (2.84) wie am Ende des Beweises von Satz 5a die Konvergenz des Eulerverfahrens mit der Ordnung 1, denn mit der speziellen Folge $\hat{q}_n = q(t_n)$, $\hat{v}_n = v(t_n)$, $\hat{\lambda}_n = \lambda(t_n)$ ist $\delta = \mathcal{O}(h)$ und $\theta = 0$ in (2.76). (Wegen $\frac{1}{h}(q(t_{m+1}) - q(t_m)) = v(t_{m+1}) + \mathcal{O}(h)$ gilt hier nicht $\frac{1}{h}(\hat{q}_{m+1} - \hat{q}_m) = \hat{v}_{m+1}$, so daß in den rechten Seiten von (2.84) ein zusätzlicher Term $\mathcal{O}(h)$ zu berücksichtigen ist.) Für die globalen Diskretisierungsfehler in (2.85) gilt $e_m^q = \mathcal{O}(h)$, ... , also folgt aus der Dreiecksungleichung und aus (2.83)

$$d_m = \mathcal{O}(h) + \mathcal{O}(1)(\|\Delta_{m+1}^q\| + \|\Delta_m^q\| + \|\Delta_{m+1}^v\| + \|\Delta_m^v\|) + \mathcal{O}(\delta) + \mathcal{O}(\theta) + \mathcal{O}(\frac{1}{h}\theta) .$$

Durch rekursive Anwendung von (2.84) beweist man die Behauptung analog zu Satz 5. Schließlich erhält man aus (2.82) die Fehlerschranke für λ. ∎

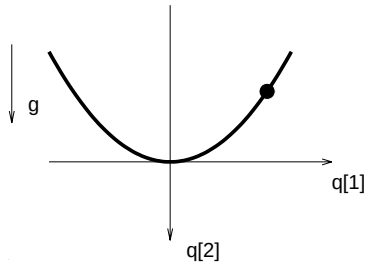
Bemerkung 12 a) Der Beweis von Satz 8 orientiert sich an der Anwendung von Satz 5 auf Index-2-Systeme der speziellen Form (2.72). Der entscheidende Unterschied liegt in dem expliziten Anteil $\int_0^1 \frac{\partial S}{\partial q}(q_\vartheta)(v_n, v_{n+1})\, d\vartheta$ der Diskretisierung (2.74). Ebenso wie Satz 5 kann auch Satz 8 auf variable Schrittweiten h verallgemeinert werden. Wegen des Diskretisierungsfehlers in den kinematischen Gleichungen $q' = v$ geht dabei aber für (2.73) die Konvergenz in λ verloren ([40]).

b) Setzt man $\mu = \nu = 0$ in Satz 6, so zeigt Satz 8 wieder die Analogie der Fehlerschranken für die analytische und die numerische Lösung. Der Einfluß kleiner Störungen auf q_n und v_n im Index-3-System (2.55) mit $f = f(q, v)$ ist vergleichbar mit dem Einfluß kleiner Störungen auf die Lösung eines Index-2-Systems (2.7), das nichtlinear in z ist. Der zusätzliche Fehlerterm ist am größten in λ und am kleinsten in q und in $S(q)v$. Ähnlich wie in Beispiel 9 kann man mit Beispiel 12 zeigen, daß die Fehlerschranken aus Satz 8 scharf sind. Statt der bei der Untersuchung von Index-3-Systemen üblichen Voraussetzung $\theta = \mathcal{O}(h^3)$ verwendet Satz 8 dabei nur $\theta = o(h)$.

Als Ergänzung zu Beispiel 8 zeigen wir auch für die MKS–Modellgleichungen (2.55) die quantitativen Auswirkungen des zusätzlichen Fehlerterms auf den Gesamtfehler der Integration. Dabei wird deutlich, daß der Fehler in q und v sehr viel größer ist, wenn (2.55) Reibungskräfte enthält.

Beispiel 13

Eine Punktmasse m möge sich unter dem Einfluß der parallel zur q_2–Achse wirkenden Schwerkraft auf der Parabel $q_2 + q_1^2 = 0$ bewegen (vgl. Abbildung). Die dynamischen Gleichungen in (2.55) sind dann gegeben durch



$$mq'' = \begin{pmatrix} 0 \\ mg \end{pmatrix} + F_R(q, q', \lambda) - \lambda \begin{pmatrix} 2q_1 \\ 1 \end{pmatrix}$$

mit den Reibungskräften $F_R(q, q', \lambda)$ und den Zwangskräften $F_N := -\lambda(\, 2q_1 \,,\, 1\,)^T$.

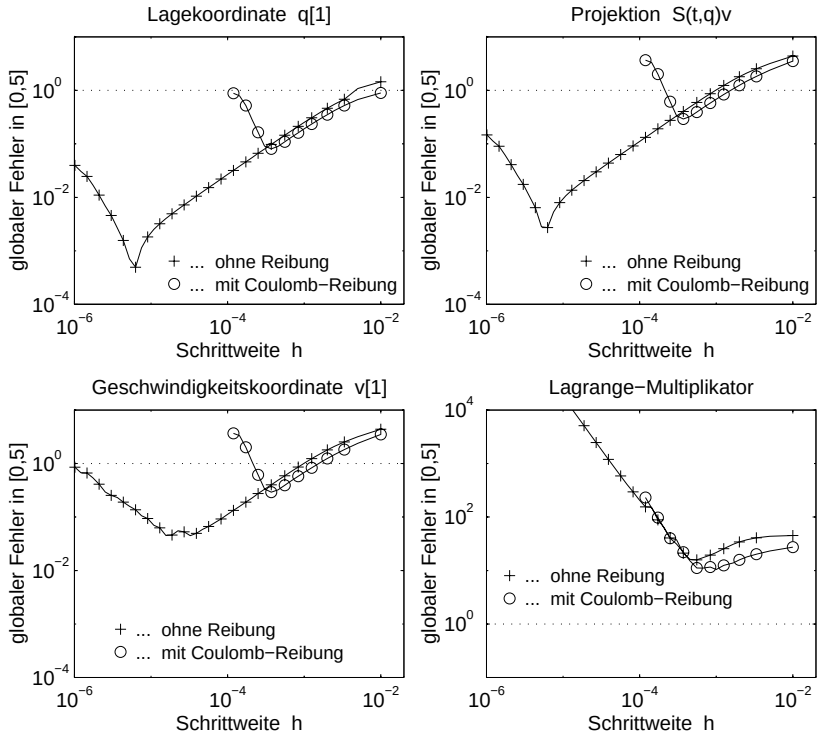


Abbildung 2.5: Globaler Fehler bei der Simulation der Bewegung einer Punktmasse auf einer Parabel.

Im Fall von Coulombscher Reibung wirkt F_R entgegen dem Geschwindigkeitsvektor q' und es gilt $|F_R| = \mu_0 |F_N|$ mit dem Reibungskoeffizienten μ_0 (für den Kontakt Stahl–Stahl ist $\mu_0 = 0.03 \ldots 0.09$ [42, Kapitel I.11]):

$$F_R(q, q', \lambda) := -\mu_0 \frac{|\lambda|\sqrt{1 + 4q_1^2}}{\sqrt{(q_1')^2 + (q_2')^2}} \begin{pmatrix} q_1' \\ q_2' \end{pmatrix} .$$

Sei $m = 1\,\mathrm{kg}$. Das Anfangswertproblem $q(0) = (\, 1 \,,\, -1\,)^T$, $q'(0) = (\, 0 \,,\, 0\,)^T$ wird wie in Beispiel 8 mit dem impliziten Eulerverfahren und fester Schrittweite h integriert, wobei die Zwangsbedingung künstlich gestört wird: statt $g(q_n) = 0$ wird $g(q_n) = 30 \cdot \theta(t_n)$ gesetzt. Hier bezeichnet $\theta(t)$ die Differenz zwischen $\sin t$ in `extended`- und $\sin t$ in `single`-Arithmetik, $|30 \cdot \theta(t_n)| \approx 10^{-6}$.
Obwohl in diesem Beispiel Satz 8 formal nur für $\mu_0 = 0$ anwendbar ist (denn $\mu = \mathcal{O}(\mu_0)$

und F_R ist für $\mu_0 \neq 0$ nicht differenzierbar in $\lambda = 0$), bestätigen die Simulationsergebnisse exakt die Abschätzungen von Satz 6: Abb. 2.5 zeigt für verschiedene h den globalen Fehler in q und in der Projektion $S(t,q)v$ sowie in v und in λ. Hierbei markiert „+“ die Simulationsergebnisse für den Fall $\mu_0 = 0$ (keine Reibungskräfte) und „∘“ die entsprechenden Ergebnisse für Coulombsche Reibung mit $\mu_0 = 0.03$. Gezeigt wird der globale Fehler der gestörten numerischen Lösung auf einem festen Zeitgitter $t_i = 0.01 \cdot i$, $(i = 1, \ldots, 500)$, zum Vergleich wird eine sehr genau berechnete numerische Lösung verwendet.
Für größere Schrittweiten dominiert in q und v der Diskretisierungsfehler, der zusätzliche Fehlerterm ist vernachlässigbar. Wegen der logarithmischen Skaleneinteilung haben die Graphen den Anstieg $+1$ (Konvergenzordnung 1). In λ ist der Fehler sehr viel größer und wächst mit $1/h^2$ (Anstieg -2, in dieser Graphik reicht die Ordinatenachse nicht wie in den anderen Diagrammen von 10^{-4} bis 10^{1}, sondern von 10^{-1} bis 10^{4}). Bei Reduktion von h verringert sich der Fehler in q und v nicht weiter, der zusätzliche Fehlerterm dominiert. Für $\mu_0 = 0$ erwarten wir Fehlerterme $\mathcal{O}(\frac{1}{h^2}\theta^2)$ in q und $S(q)v$ und Fehlerterme $\mathcal{O}(\frac{1}{h}\theta)$ in $G(q)v$, für kleine Schrittweiten h haben die Graphen in Abb. 2.5 die Anstiege -2 in q_1 und $S(q)v$ und den Anstieg -1 in v_1.
Ein völlig anderes Bild ergibt sich für das System mit Reibung: während die Fehler in λ etwa so groß sind wie zuvor, ist der Einfluß der Störungen auf q und v stark gewachsen, darüberhinaus sind die Fehler in q_1, $S(t,q)v$ und v_1 jetzt nahezu gleich groß.

Die Simulationsergebnisse von Beispiel 13 unterstreichen, daß die Fehlerschranken der Sätze 6 und 8 die Größe der tatsächlich auftretenden Fehler qualitativ richtig wiedergeben. Abschließend sei bemerkt, daß man in praktischen Rechnungen zu sehr viel kleineren zusätzlichen Fehlertermen als in Abb. 2.5 kommen kann, wenn die Störungen während der Integration reduziert werden (vgl. hierzu Beispiel 6).

2.4 Der gleichmäßige Störungsindex

Der Störungsindex eines DA–Systems dient nicht nur wie z. B. der differentielle Index der Klassifikation von differentiell-algebraischen Systemen, sondern gibt darüberhinaus eine (grobe) Vorstellung von der Empfindlichkeit der numerischen Lösung gegenüber kleinen Störungen. Bei der quantitativen Beschreibung des zusätzlichen Fehlerterms ist die Fehlerschranke (2.6) aus Definition 4 jedoch nur dann nützlich, wenn die Konstante C nicht zu groß ist.

Singulär gestörte Systeme gewöhnlicher Differentialgleichungen und Semidiskretisierungen von gewissen Systemen partieller Differentialgleichungen führen auf *Klassen* von DA–Systemen (1.1), in denen man für jedes individuelle DA–System eine Abschätzung (2.3) bzw. (2.6) zeigen kann. Betrachtet man aber die Klasse *aller* dieser Systeme, so kann die Konstante C beliebig groß werden. Die Arbeiten [79], [80] von Hairer, Lubich und Roche und [107] von Lubich zeigen, daß der bei der numerischen Lösung entstehende Fehler sehr viel realistischer durch Fehlerschranken beschrieben wird, die *gleichmäßig* für alle Systeme einer Aufgabenklasse gelten. Das entsprechende Gegenstück zu Definition 4 bezeichnen wir hier als gleichmäßigen Störungsindex einer Klasse von DA–Systemen (Definition 5).

In diesem Abschnitt wird das Konzept gleichmäßiger Fehlerabschätzungen für die analytische Lösung am Beispiel der nach Baumgarte stabilisierten Indexreduktion von DA–Systemen im Detail vorgestellt. Im kritischen Fall großer Baumgarte–Parameter („DAEs of nearly higher index“ [164]) beschreibt der gleichmäßige Störungsindex die Empfindlichkeit der Lösung gegenüber Störungen sehr viel besser als der (klassische) Störungsindex (Beispiel 15b).

In jüngster Zeit ist das Interesse an der Untersuchung partieller Differentialgleichungen mit algebraischen Nebenbedingungen gewachsen (engl.: „partial differential-algebraic equations“ (PDAEs), vgl. [46] und die dort angegebene Literatur). Campbell und Marszalek ([45]) untersuchen u. a. die Anwendung der Linienmethode auf ein spezielles System aus 2 linearen partiellen Differentialgleichungen. Beispiel 16 zeigt, daß jedes der dabei entstehenden semidiskreten DA–Systeme den Störungsindex 1 hat. In Abhängigkeit von den Koeffizienten der partiellen Differentialgleichungen hat die Klasse all dieser Semidiskretisierungen jedoch entweder

- den gleichmäßigen Störungsindex 2 (dies entspricht in diesem Beispiel der Empfindlichkeit der Lösung des Systems partieller Differentialgleichungen gegenüber kleinen Störungen) oder
- überhaupt keinen gleichmäßigen Störungsindex.

Beispiel I: Baumgarte–Stabilisierung

Die gleichmäßigen Fehlerschranken werden am Beispiel der Index-1-Formulierung von Index-2-Systemen in Hessenbergform illustriert. Hierzu ersetzt man zunächst in (2.7) die Zwangsbedingung $0 = g(y(t))$ durch $0 = \frac{d}{dt}g(y(t)) = g_y(y)y' = [g_y f](y,z)$, anschließend fügt man analog zur Baumgarte–Stabilisierung (3.5) von MKS–Modellgleichungen (vgl. Abschnitt 3.1) einen stabilisierenden Term $\alpha g(y)$ mit einem Parameter $\alpha > 0$ hinzu, um ein mögliches Abdriften der Lösung von $\{\eta : g(\eta) = 0\}$ zu verhindern. Wir ersetzen also (2.7) durch

$$\begin{aligned} y'(t) &= f(y(t), z(t)) \,, \\ 0 &= \frac{1}{\alpha}\frac{d}{dt}g(y(t)) + g(y(t)) \,. \end{aligned} \tag{2.86}$$

Durch diese Substitution wird die analytische Lösung nicht geändert, denn

$$g(y(t)) = 0 \,,\ (t \in [0,T]) \quad \Leftrightarrow \quad \left(0 = \frac{1}{\alpha}\frac{d}{dt}g(y(t)) + g(y(t)) \quad \text{und} \quad g(y(0)) = 0 \right) .$$

Wie in Abschnitt 2.2 wird die Auswirkung von Störungen auf die Lösung an Hand von Funktionen $(\hat{y}_\alpha(t), \hat{z}_\alpha(t))$ mit

$$\begin{aligned} \hat{y}'_\alpha(t) &= f(\hat{y}_\alpha(t), \hat{z}_\alpha(t)) + \delta(t) \quad , \quad (t \in [0,T]) \,, \\ \theta(t) &= \frac{1}{\alpha}\frac{d}{dt}g(\hat{y}_\alpha(t)) + g(\hat{y}_\alpha(t)) \,, \quad \hat{y}_\alpha(0) = \hat{y}_{\alpha,0} \,, \quad \hat{z}_\alpha(0) = \hat{z}_{\alpha,0} \end{aligned} \tag{2.87}$$

untersucht. Das wesentliche Hilfsmittel sind dabei folgende Abschätzungen:

Lemma 3 *a) Gegeben sei ein $\alpha > 0$ und Funktionen $\tilde\delta \in C[0,T]$, $\tilde\theta \in C^1[0,T]$. Dann gelten für die Lösungen der linearen Differentialgleichung*

$$w'(t) + \alpha w(t) = \tilde\delta(t) + \alpha\tilde\theta(t) \tag{2.88}$$

die Abschätzungen

$$\begin{aligned} |w(t) - \tilde\theta(t)| &\le |w(0) - \tilde\theta(0)|e^{-\alpha t} + \frac{1}{\alpha}\cdot(\,\|\tilde\delta\|_{C^0([0,t])} + \|\tilde\theta\|_{C^1([0,t])}\,)\,, \\ |w'(t)| &\le |w'(0)|e^{-\alpha t} + \mathcal{O}(1)\|\tilde\delta\|_{C^0([0,t])} + \|\tilde\theta\|_{C^1([0,t])}\,. \end{aligned}$$

b) Ist $\tilde\delta(t) \equiv 0$, $\tilde\theta(t) = \Theta\cos\frac{t}{\varepsilon}$ und $w(0) = \varepsilon^2\alpha^2\Theta/(1+\varepsilon^2\alpha^2)$ mit (kleinen) positiven Parametern Θ und ε, so ist die Lösung von (2.88) gegeben durch

$$w_\alpha(t) = (1 - \frac{1}{1+\varepsilon^2\alpha^2})\,\Theta\cos\frac{t}{\varepsilon} + \frac{\varepsilon\alpha}{1+\varepsilon^2\alpha^2}\,\Theta\sin\frac{t}{\varepsilon}\,. \tag{2.89}$$

Beweis In der Lösung

$$w(t) = w(0)e^{-\alpha t} + \int_0^t e^{-\alpha(t-\tau)}(\tilde\delta(\tau) + \alpha\tilde\theta(\tau))\,d\tau \tag{2.90}$$

von (2.88) erhält man durch partielle Integration

$$\alpha\int_0^t e^{-\alpha(t-\tau)}\tilde\theta(\tau)\,d\tau = \Big[e^{-\alpha(t-\tau)}\tilde\theta(\tau)\Big]_0^t - \int_0^t e^{-\alpha(t-\tau)}\tilde\theta'(\tau)\,d\tau$$

und damit wegen $\int_0^t e^{-\alpha(t-\tau)}\,d\tau = \frac{1}{\alpha}(1-e^{-\alpha t}) < \frac{1}{\alpha}$ auch

$$\Big|\int_0^t e^{-\alpha(t-\tau)}(\tilde\delta(\tau) - \tilde\theta'(\tau))\,d\tau\,\Big| < \frac{1}{\alpha}\cdot\max_{\tau\in[0,t]}(|\tilde\delta(\tau)| + |\tilde\theta'(\tau)|)\,.$$

Aus $w'(t) = \tilde\delta(t) - \alpha(w(t) - \tilde\theta(t))$ folgt dann auch die Abschätzung für $w'(t)$. Schließlich werden zum Beweis von Teil b) des Lemmas die gegebenen Funktionen $\tilde\delta$ und $\tilde\theta$ in (2.90) eingesetzt. ∎

Lemma 3 wird nun auf (2.87) angewendet: Wegen

$$\frac{1}{\alpha}\frac{d}{dt}g(y(t)) + g(y(t)) = \frac{1}{\alpha}g_y(y(t))y'(t) + g(y(t)) = \frac{1}{\alpha}[g_y f](y(t),z(t)) + g(y(t))$$

kann der algebraische Teil von (2.86) nach den algebraischen Komponenten z aufgelöst werden (vgl. (2.8)), d. h., das stabilisierte System (2.86) hat den Störungsindex 1 ([84, S. 459f]). Deshalb ist

$$\begin{aligned} &\|\hat y_\alpha(t) - y(t)\| + \|\hat z_\alpha(t) - z(t)\| \le \\ &\qquad \le C_\alpha(\|\hat y_\alpha(0) - y(0)\| + \|\hat z_\alpha(0) - z(0)\| + \|\delta\|_{C^0([0,t])} + \|\theta\|_{C^0([0,t])})\,. \end{aligned} \tag{2.91}$$

Damit diese Abschätzung auch für große α gilt, muß jedoch i. allg. $\lim_{\alpha\to\infty} C_\alpha = \infty$ sein. Dies ist plausibel, denn (2.86) ist für große α nahezu identisch mit dem Ausgangsproblem (2.7) und eine Schranke der Form (2.91) kann es für DA–Systeme vom Störungsindex 2 per definitionem nicht geben.

Beispiel 14 Das Anfangswertproblem $y_1(0) = y_{1,0}$ zu dem Index-2-System $y_1' = y_2' = z$, $0 = y_1 + y_2$ hat die konstante Lösung $y_1(t) \equiv y_{1,0}$, $y_2(t) \equiv -y_{1,0}$, $z(t) \equiv 0$. Mit $w_\alpha(t)$ aus (2.89) werden nun Funktionen $(\hat y_\alpha, \hat z_\alpha)$ definiert:

$$\hat y_{\alpha,1}(t) = y_{1,0} + \frac{1}{2}(w_\alpha(t) - w_\alpha(0))\,,\quad \hat y_{\alpha,2}(t) = w_\alpha(t) - \hat y_{\alpha,1}(t)\,,\quad \hat z_\alpha(t) = \frac{1}{2}w_\alpha'(t)\,.$$

Für diese Funktionen ergeben sich in (2.87) die Residuen $\delta(t) \equiv 0$ und $\theta(t) = \Theta\cos\frac{t}{\varepsilon}$, denn $g(\hat y_\alpha(t)) = \hat y_{\alpha,1}(t) + \hat y_{\alpha,2}(t) = w_\alpha(t)$ (vgl. Lemma 3).
Wie in Beispiel 9 gilt $\|\delta\|_{C^0} = 0$, $\|\theta\|_{C^0} = \mathcal{O}(\Theta)$, $\|\theta\|_{C^1} = \mathcal{O}(\frac{\Theta}{\varepsilon})$. Andererseits ist

$$|\hat z_\alpha(t) - z(t)| = \frac{1}{2}|w_\alpha'(t)| = \frac{1}{2}|-\frac{\varepsilon\alpha^2}{1+\varepsilon^2\alpha^2}\sin\frac{t}{\varepsilon} + \frac{\alpha}{1+\varepsilon^2\alpha^2}\cos\frac{t}{\varepsilon}|\cdot\Theta$$

und man erhält für $\alpha \ge 1$ und den speziellen Wert $\varepsilon = 1/\sqrt{\alpha}$ in (2.91) $\hat y_{\alpha,1}(0) = y_1(0)$, $|\hat y_{\alpha,2}(0) - y_2(0)| = |w_\alpha(0)| = \frac{\alpha}{1+\alpha}\Theta \le \Theta$, $|\hat z_\alpha(0) - z(0)| = \frac{1}{2}\frac{\alpha}{1+\alpha}\Theta \le \Theta$, $\|\theta\|_{C^0} \le \Theta$ und $|\hat z_\alpha(\frac{\pi}{2}\varepsilon) - z(\frac{\pi}{2}\varepsilon)| = \frac{1}{2}\frac{\alpha^{3/2}}{1+\alpha}\Theta \ge \frac{1}{4}\sqrt{\alpha}\Theta$, d. h., für die Konstante C_α muß $C_\alpha \ge \frac{1}{12}\sqrt{\alpha}$ gelten.

Für $\alpha\to\infty$ wächst also die klassische Fehlerschranke (2.91) für das stabilisierte System sehr schnell, Beispiel 15b wird zeigen, daß (2.91) den Fehler in $(y(t), z(t))$ stark überschätzt, wenn $\|\theta\|_{C^1}$ klein und $\alpha \gg 1$ ist.

Es ist ein bekannter Nachteil der Baumgarte–Stabilisierung, daß für (sehr) große Baumgarte–Parameter α die Lösung des indexreduzierten Systems sogar aufwendiger als die Lösung des Ausgangsproblems sein kann (vgl. z. B. [29]). Weder der differentielle noch der (klassische) Störungsindex sind geeignet, um für große α die Schwierigkeiten bei der numerischen Lösung des indexreduzierten Systems zu beschreiben: Wegen der Terme $e^{-\alpha t}$ in Lemma 3 enthält die Lösung (sehr) schnell abklingende Anteile, darüberhinaus ist die Lösung weit weniger robust gegenüber Störungen als es der (niedrige) klassische Störungsindex suggeriert.

Der gleichmäßige Störungsindex: Definition

Als Alternative zu Fehlerschranken (2.6) mit sehr großen Konstanten C erweitern wir die Störungstheorie von der Betrachtung einzelner DA–Systeme auf die Untersuchung von Klassen differentiell-algebraischer Systeme, die von einem Parameter $\alpha \in M_\alpha$ abhängen:

$$F_\alpha(t, x(t), x'(t)) = 0\,,\ (\,t \in [0,T]\,)\,,\quad x(0) = x_{0,\alpha}\,. \tag{2.92}$$

In der vorliegenden Arbeit beschränken wir uns auf skalare Parameter α, i. allg. kann dabei die Dimension von x mit α variieren (vgl. Beispiel 16).

Definition 5 Gegeben sei eine Klasse von DA–Systemen (2.92) mit Lösungen $x_\alpha(t)$, d. h. $F_\alpha(t, x_\alpha(t), x_\alpha'(t)) = 0$, $(t \in [0,T])$. Gibt es ein $r \in \mathbb{N}$ und eine von α unabhängige Konstante C, so daß für alle $\alpha \in M_\alpha$, alle Funktionen $\hat x_\alpha(t)$ mit $F_\alpha(t, \hat x_\alpha(t), \hat x_\alpha'(t)) = \delta(t)$, $(t \in [0,T])$ und für jedes $t \in [0,T]$ stets

$$\|\hat x_\alpha(t) - x_\alpha(t)\| \le C\,(\,\|\hat x_\alpha(0) - x_\alpha(0)\| + \|\delta\|_{C^r([0,t])}\,) \tag{2.93}$$

gilt, sofern die rechte Seite in (2.93) hinreichend klein ist, und ist darüberhinaus r die kleinste natürliche Zahl mit dieser Eigenschaft, so heißt $m = r + 1$ *gleichmäßiger Störungsindex* der Klasse von DA–Systemen (2.92) bezüglich der Lösungen $\hat x_\alpha(t)$.

Bemerkung 13 a) Auch für Klassen von DA–Systemen kann man i. allg. wie in den Abschnitten 2.2 und 2.3 kleinere Fehlerschranken für die differentiellen Komponenten nachweisen. Wir beschränken uns in diesem Abschnitt jedoch auf das Analogon (2.93) zu Definition 4.

b) Mattheij ([115]) und Wijckmans ([164, Kapitel 4 und 5]) untersuchen lineare DA–Systeme (1.1) mit zeitabhängigen Koeffizienten, die einen kleinen Parameter ε enthalten. Dabei ist der Index der betrachteten Systeme für $\varepsilon = 0$ größer als im Fall $\varepsilon \neq 0$. Für den Nachweis von (bez. ε) gleichmäßigen Fehlerschranken ist die partielle Integration das wesentliche Beweishilfsmittel (wie im Beweis von Lemma 3). Ein System mit $0 < |\varepsilon| \ll 1$ wird verbal als „System, das fast einen höheren Index hat" charakterisiert (engl.: „almost higher index system"). In Analogie zur B–Konvergenz–Theorie für die Anwendung impliziter Runge–Kutta–Verfahren auf steife gewöhnliche Differentialgleichungen ([61]) bevorzugen wir die gleichzeitige Betrachtung sämtlicher DA–Systeme einer gegebenen Klasse. Der gleichmäßige Störungsindex einer Klasse (2.92) ist dabei stets mindestens so groß wie das Maximum der klassischen Störungsindizes jedes einzelnen DA–Systems dieser Klasse.

Beispiel 15 a) Die von Hairer et al. ([79], [80]) betrachtete Klasse singulär gestörter Systeme gewöhnlicher Differentialgleichungen hat den gleichmäßigen Störungsindex 1, jedes einzelne dieser Systeme dagegen den (klassischen) Störungsindex 0 (vgl. Bemerkung 4b). Die von Lubich ([107]) im Zusammenhang mit der Modellierung steifer Mehrkörpersysteme betrachtete Klasse singulärer singulär gestörter Systeme gewöhnlicher Differentialgleichungen hat den gleichmäßigen Störungsindex 3, wiederum hat jedes einzelne dieser Systeme den Störungsindex 0.

b) Die Klasse der nach Baumgarte stabilisierten Systeme (2.86) mit Baumgarte–Parameter $\alpha \geq \alpha_0 > 0$ hat den gleichmäßigen Störungsindex 2. Zum Nachweis wendet man auf jede der n_z Komponenten von $g(\hat{y}_\alpha)$ Lemma 3 mit $\frac{d}{dt}g(\hat{y}_\alpha(t)) + \alpha g(\hat{y}_\alpha(t)) = \alpha\theta(t)$ an und erhält $g(\hat{y}_\alpha(t)) = \hat{\theta}(t)$ mit

$$\begin{aligned} \|\hat{\theta}(t) - \theta(t)\| &\leq \|g(\hat{y}_\alpha(0)) - \theta(0)\| \cdot e^{-\alpha t} + \frac{1}{\alpha} \cdot \|\theta\|_{C^1([0,t])} , \\ \|\hat{\theta}'(t)\| &\leq \|\frac{d}{dt} g(\hat{y}_\alpha(t))\Big|_{t=0}\| \cdot e^{-\alpha t} + \|\theta\|_{C^1([0,t])} . \end{aligned}$$

Wegen

$$\frac{d}{dt} g(\hat{y}_\alpha(t)) = g_y(\hat{y}_\alpha(t))\hat{y}'_\alpha(t) = [g_y f](\hat{y}_\alpha(t), \hat{z}_\alpha(t)) + g_y(\hat{y}_\alpha(t))\delta(t)$$

und $g(y(0)) = [g_y f](y(0), z(0)) = 0$ folgt hieraus

$$\begin{aligned} \hat{y}'_\alpha(t) &= f(\hat{y}_\alpha, \hat{z}_\alpha) + \delta(t) \\ \hat{\theta}(t) &= g(\hat{y}_\alpha(t)) \end{aligned}$$

mit

$$\begin{aligned} \|\hat{\theta}(t)\| &\leq \|\theta(t)\| + \frac{1}{\alpha_0}\|\theta\|_{C^1([0,t])} + \mathcal{O}(1)(\|\hat{y}_\alpha(0) - y(0)\| + \|\theta(0)\|) , \\ \|\hat{\theta}'(t)\| &\leq \|\theta\|_{C^1([0,t])} + \mathcal{O}(1)(\|\hat{y}_\alpha(0) - y(0)\| + \|\hat{z}_\alpha(0) - z(0)\| + \|\theta'(0)\| + \|\delta(0)\|) , \end{aligned}$$

wobei die Konstanten in den $\mathcal{O}(\cdot)$–Termen von α unabhängig sind. Aus Satz 3 folgt damit die Abschätzung (2.93) mit $r = 1$, d. h. $m = 2$.

Der gleichmäßige Störungsindex der Klasse von DA–Systemen (2.86) mit $\alpha \geq \alpha_0 > 0$ bleibt unverändert, wenn die Klasse um das Index-2-System (2.7), d. h. um den Grenzfall $\alpha \to \infty$, erweitert wird.

c) Die Klasse der nach Baumgarte stabilisierten Systeme (2.86) mit Baumgarte–Parameter $\alpha \leq \overline{\alpha} < \infty$ hat den gleichmäßigen Störungsindex 1, in (2.93) ist dann $C = C(\overline{\alpha})$.

Bemerkung 14 a) Das diskrete Gegenstück zu der gleichmäßigen Fehlerschranke in Beispiel 15b ist eine Fehlerschranke $C^*_{0,\infty} \cdot \frac{1}{h}\Delta$ mit einer von α unabhängigen Konstanten $C^*_{0,\infty}$ (vgl. Satz 5 mit $\Delta := \delta + \theta$). Wendet man geeignete Diskretisierungsverfahren auf (2.86) an, so läßt sich nachweisen, daß die Verstärkung kleiner Fehler Δ durch $\min(C^*_{0,\alpha}\Delta, C^*_{0,\infty} \cdot \frac{1}{h}\Delta)$ beschränkt ist, wobei $\lim_{\alpha\to\infty} C^*_{0,\alpha} = \infty$ und $C^*_{0,\infty} = \mathcal{O}(1)$ gilt. Ist die Schrittweite h nicht übermäßig klein, so ist für große α die gleichmäßige Fehlerschranke $C^*_{0,\infty} \cdot \frac{1}{h}\Delta$ sehr viel kleiner als die auf herkömmliche Weise (z. B. mit Satz 5) nachgewiesene Fehlerschranke $C^*_{0,\alpha}\Delta$.

b) Modifiziert man die Beispiele 14 und 15b geeignet, so lassen sich für beliebiges $m > 0$ Klassen differentiell-algebraischer Systeme vom gleichmäßigen Störungsindex m angeben, für die jedes einzelne DA–System den klassischen Störungsindex 1 hat. Die in Abschnitt 3.1 betrachtete Baumgarte–Stabilisierung (3.5) für MKS–Modellgleichungen führt auf eine Klasse (2.92) vom gleichmäßigen Störungsindex 3.

Beispiel II: Semidiskretisierung eines Systems partieller Differentialgleichungen mit der Linienmethode

In vielfältigen praktischen Anwendungen entstehen DA–Systeme (1.1) als Ergebnis der (räumlichen) Semidiskretisierung von Systemen partieller Differentialgleichungen mit der Linienmethode ([46]). Campbell und Marszalek ([45], [46]) versuchen, die Empfindlichkeit der Lösung des Systems partieller Differentialgleichungen gegenüber Störungen analog zum Störungsindex für (gewöhnliche) differentiell-algebraische Systeme zu beschreiben, um damit eine (grobe) Vorstellung zu erhalten, mit welchen numerischen Schwierigkeiten man bei der Zeitintegration des semidiskreten Systems rechnen muß.

In dem verbleibenden Teil dieses Abschnitts zeigen wir an Beispiel 1 aus [45], daß sich im Rahmen dieses Konzepts zwangsläufig die Notwendigkeit gleichmäßiger, d. h. vom Ortsdiskretisierungsparameter unabhängiger Fehlerschranken für das semidiskrete Problem ergibt. In der Störungstheorie für differentiell-algebraische Systeme ist dies das Analogon zu den gleichmäßigen Schranken für den Diskretisierungsfehler bei der Zeitintegration von semidiskretisierten Anfangs–Randwertproblemen für parabolische Differentialgleichungen (z. B. [156], [110], vgl. auch die ausführliche Darstellung in [153, Kapitel 7]).

Beispiel 16 ([45, Beispiel 1]) Gegeben sei das System von zwei linearen partiellen Differentialgleichungen

$$\left.\begin{aligned} u_t - \frac{1}{4}v_{xx} + \varrho v &= f^u(x,t) \\ -\frac{1}{4}u_{xx} + \frac{1}{4}v_{xx} + v &= f^v(x,t) \end{aligned}\right\} \quad , \quad (\, x \in [0,L],\ t \in [0,T]\,) \tag{2.94}$$

mit Anfangsbedingungen

$$u(x,0) = g^u(x) \ , \quad v(x,0) = g^v(x) \ , \quad (\, x \in [0,L]\,)$$

und homogenen Randbedingungen. In (2.94) bezeichnet $\varrho \in \mathbb{R}$ einen fixierten Parameter, für den wir speziell die beiden Fälle $\varrho = 0$ und $\varrho = -2$ betrachten werden. Die Funktionen $f := (f^u, f^v)^T$ und $g := (g^u, g^v)^T$ mögen hinreichend oft differenzierbar sein und die Randbedingungen erfüllen. Außerdem sei vorausgesetzt, daß u, v, f und g dargestellt werden können als

$$u(x,t) = \sum_{n=1}^{\infty} \phi_n(x) u_n(t) \ , \quad v(x,t) = \sum_{n=1}^{\infty} \phi_n(x) v_n(t) \ , \quad \dots$$

mit $\phi_n(x) := \sin(\frac{n\pi x}{L})$. Unter geeigneten Glattheitsvoraussetzungen sind die Koeffizienten $u_n(t)$, $v_n(t)$ Lösungen der Anfangswertprobleme $u_n(0) = g_n^u$, $v_n(0) = g_n^v$ für die linearen DA–Systeme

$$\begin{aligned} u_n'(t) + (\varrho + \frac{1}{4}\lambda_n^2) v_n(t) &= f_n^u(t) \\ \frac{1}{4}\lambda_n^2 u_n(t) + (1 - \frac{1}{4}\lambda_n^2) v_n(t) &= f_n^v(t) \end{aligned} \tag{2.95}$$

mit $\lambda_n := n\pi/L$, $(n \geq 1)$.
Für die Lösung von (2.94) mit der Linienmethode wird auf $[0, L]$ das äquidistante Gitter $0 = x_0 < x_1 < \ldots < x_N < x_{N+1} = L$, $x_i := i\Delta_x$ mit $\Delta_x = L/(N+1)$ definiert, auf dem man die partiellen Ableitungen bez. x durch zentrale Differenzenquotienten 2. Ordnung approximiert. Sei $u_i^N(t) \approx u(x_i, t)$, $v_i^N(t) \approx v(x_i, t)$, $(i = 1, \ldots, N)$ und

$$U(t) = (u_1^N(t), \ldots, u_N^N(t))^T \ , \quad V(t) = (v_1^N(t), \ldots, v_N^N(t))^T \ .$$

Mit den Bezeichnungen

$$F^u(t) = (f^u(x_1,t), \ldots, f^u(x_N,t))^T \ , \quad F^v(t) = (f^v(x_1,t), \ldots, f^v(x_N,t))^T \ ,$$

$$G^u = (g^u(x_1), \ldots, g^u(x_N))^T \ , \quad G^v = (g^v(x_1), \ldots, g^v(x_N))^T$$

ist die Finite–Differenzen–Approximation $U(t)$, $V(t)$ Lösung des Anfangswertproblems $U(0) = G^u$, $V(0) = G^v$ für das DA–System

$$\begin{aligned} U'(t) - \frac{1}{4} A_\Delta \cdot V(t) + \varrho \cdot V(t) &= F^u(t) \\ -\frac{1}{4} A_\Delta \cdot U(t) + \frac{1}{4} A_\Delta \cdot V(t) + V(t) &= F^v(t) \end{aligned} \tag{2.96}$$

mit der symmetrischen Tridiagonalmatrix

$$A_\Delta = \frac{1}{\Delta_x^2} \begin{pmatrix} -2 & 1 & 0 & & & 0 \\ 1 & -2 & 1 & & & \\ & & & \ddots & & \\ & & & 1 & -2 & 1 \\ 0 & & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{N \times N} \ .$$

Die Eigenwerte von A_Δ sind $\mu_i = -\frac{4}{\Delta_x^2} \sin^2(\frac{i\pi}{2(N+1)})$, die Eigenvektoren

$$\Phi_i = (\ \sin(\frac{i\pi}{N+1}) \ , \ \sin(\frac{2i\pi}{N+1}) \ , \ \ldots \ , \ \sin(\frac{Ni\pi}{N+1}) \ )^T \ , \quad (i = 1, \ldots, N) \ .$$

Bez. der Basis $\{\Phi_1, \ldots, \Phi_N\}$ werden die Vektoren $U(t)$, $V(t)$, $F^u(t)$, $F^v(t)$, G^u, G^v dargestellt durch

$$U(t) = \sum_{i=1}^{N} \frac{1}{\|\Phi_i\|_2} \Phi_i \cdot U_i(t) \ , \quad V(t) = \sum_{i=1}^{N} \frac{1}{\|\Phi_i\|_2} \Phi_i \cdot V_i(t) \ , \quad \dots \ ,$$

so daß man aus (2.96) nach Multiplikation mit Φ_1^T, Φ_2^T, $\ldots$, Φ_N^T die Gleichungssysteme

$$\begin{aligned} U_i'(t) + (\varrho + \frac{1}{4}\Lambda_i^2) V_i(t) &= F_i^u(t) \\ \frac{1}{4}\Lambda_i^2 U_i(t) + (1 - \frac{1}{4}\Lambda_i^2) V_i(t) &= F_i^v(t) \end{aligned} \tag{2.97}$$

mit $\Lambda_i := \sqrt{-\mu_i} = 2\sin(\frac{i\pi}{2(N+1)})/\Delta_x$, $(i = 1, \ldots, N)$ erhält, denn $A_\Delta \Phi_i = \mu_i \Phi_i = -\Lambda_i^2 \Phi_i$ und $\Phi_j^T \Phi_i = \|\Phi_i\|_2^2 \cdot \delta_{ij}$, $(i, j = 1, \ldots, N)$ mit dem Kroneckerschen Delta δ_{ij}.

Die Transformationen von (2.94) zu (2.95) und von (2.96) zu (2.97) berühren nicht die Empfindlichkeit der Lösungen gegenüber Störungen. Statt des Systems (2.94) und der Semidiskretisierungen (2.96) betrachten wir daher die Systeme (2.95) und (2.97), die für jedes n bzw. jedes i aus je einer Differentialgleichung und einer algebraischen Gleichung bestehen. Im Grenzfall $N \to \infty$ geht (2.97) in (2.95) über, denn für fixiertes $i \leq N$ gilt $\lim_{N\to\infty} \Lambda_i = \lambda_i = i\pi/L$.

Da die Dimension von U und V in (2.96) mit N unbeschränkt wächst, machen gleichmäßige Fehlerschranken nur Sinn, wenn die in Definition 5 verwendeten Normen für verschiedene N zueinander passend gewählt werden. Bez. x wird in diesem Beispiel für die Untersuchung der partiellen Differentialgleichungen die L^2–Norm auf $[0, L]$ und für die Semidiskretisierungen das diskrete Analogon $\|U\|_{2,N} := (\frac{1}{N}\sum_{i=1}^{N} (U_i^N)^2)^{1/2}$ verwendet:

$$\|\delta(t)\|_{C^r([0,t]),N} = \max_{\tau\in[0,t]} \|\delta(\tau)\|_{2,N} + \max_{\tau\in[0,t]} \|\delta'(\tau)\|_{2,N} + \ldots + \max_{\tau\in[0,t]} \|\delta^{(r)}(\tau)\|_{2,N} \ .$$

Campbell und Marszalek ([45]) betrachten das System (2.94) für $\varrho = 0$ und geben dabei auch die Transformationen zu (2.95) bzw. (2.97) an. Sie untersuchen detailliert den Zusammenhang zwischen den Indizes von (2.95) und (2.97): Ist $\varrho \in \{0, -2\}$, so hat (2.95) für $\lambda_n^2 = 4$ den Störungsindex 2, in jedem anderen Fall dagegen den Störungsindex 1. Ist $\lambda_n^2 \neq 4$ für alle $n > 0$ und ist die Diskretisierung hinreichend fein, so haben nicht nur die Systeme (2.95), sondern auch ihre diskreten Gegenstücke (2.97) sämtlich den Störungsindex 1, denn $\lim_{N\to\infty} \Lambda_i = \lambda_i$.

Besondere Beachtung verdient der Fall, daß in (2.95) mit $\varrho \in \{0, -2\}$ für ein gewisses $n_0 \in \mathbb{N}$ gilt $\lambda_{n_0}^2 = 4$. Wegen $\Lambda_i \neq \lambda_i$ haben die DA–Systeme (2.97) für hinreichend feine Diskretisierung auch in diesem Fall den Störungsindex 1. Nach den Ergebnissen des Beispiels 14 ist jedoch nicht zu erwarten, daß die Klasse *aller* Semidiskretisierungen (2.96) den gleichmäßigen Störungsindex 1 hat, wenn $\lambda_{n_0}^2 = 4$ für ein $n_0 \in \mathbb{N}$ gilt.

Für eine detaillierte Untersuchung betrachten wir deshalb DA–Systeme (2.97) mit $\Lambda_i^2 \approx 4$ aber $\Lambda_i^2 \neq 4$ (dabei soll $\varrho = 0$ oder $\varrho = -2$ sein). Interpretiert man diese Index-1-Systeme als DA–Systeme, die durch eine kleine Störung aus dem Index-2-System (2.95) mit $\lambda_n^2 = 4$ hervorgehen, so zeigt Söderlind ([151]), daß das Stabilitätsverhalten der Lösungen der Index-1-Systeme (2.97) mit $0 < |\Lambda_i^2 - 4| \ll 1$ entscheidend vom Vorzeichen dieser Störung abhängt.

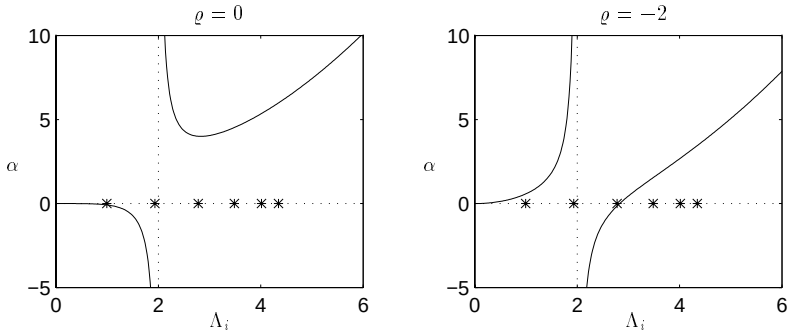


Abbildung 2.6: Abhängigkeit des Koeffizienten α aus (2.98) von Λ_i. Mit „$*$“ sind die Eigenwerte $\Lambda_1, \Lambda_2, \ldots, \Lambda_6$ der Semidiskretisierung (2.96) mit $N = 6$ und $L = \pi$ markiert.

Für ein solches Index-1-System erhält man das äquivalente DA–System

$$\begin{aligned} U_i'(t) + \alpha U_i(t) &= F_i^u(t) + \alpha \cdot \frac{1}{\frac{1}{4}\Lambda_i^2} F_i^v(t) \\ V_i(t) &= \frac{1}{\varrho + \frac{1}{4}\Lambda_i^2}(F_i^u(t) - U_i'(t)) \end{aligned} \tag{2.98}$$

mit

$$\alpha := -\frac{1}{4}\Lambda_i^2 \, \frac{\varrho + \frac{1}{4}\Lambda_i^2}{1 - \frac{1}{4}\Lambda_i^2} \quad ,$$

indem man die zweite Gleichung von (2.97) nach V_i auflöst und in die erste Gleichung einsetzt.

In (2.98) bestimmt der Parameter α das Stabilitätsverhalten der Lösung. Ist $\alpha > 0$, so haben diese Gleichungen die Struktur des Gleichungssystems (2.88) (ersetze $U_i(t)$ durch $w(t)$) und Lemma 3 kann angewendet werden. Ist dagegen $\alpha < 0$, so können Fehler in (2.98) mit dem Faktor $e^{-\alpha t}$ verstärkt werden. In dem kritischen Fall $\Lambda_i^2 \to 4$ gilt dabei $\alpha \to -\infty$. In Abb. 2.6 wird für $\varrho = 0$ (links) und $\varrho = -2$ (rechts) der Wert des Parameters α in Abhängigkeit von Λ_i dargestellt. In einer Umgebung von $\Lambda_i^2 = 4$ erhält man dabei je nach Wahl von ϱ qualitativ unterschiedliche Ergebnisse.

Lemma 4 *Gegeben sei eine positive Konstante $\Delta_\lambda \in (0, 1]$.*

a) Für $\varrho \in \{0, -2\}$ hat die Klasse der DA–Systeme (2.97) mit $|\Lambda_i^2 - 4| \geq \Delta_\lambda > 0$ den gleichmäßigen Störungsindex 1.

b) Für $\varrho = 0$ hat die Klasse der DA–Systeme (2.97) mit $\Lambda_i^2 \in (0, 4 - \Delta_\lambda] \cup [4, \infty)$ den gleichmäßigen Störungsindex 2.

c) Für $\varrho = -2$ hat die Klasse der DA–Systeme (2.97) mit $\Lambda_i^2 \in (0, 4] \cup [4 + \Delta_\lambda, \infty)$ den gleichmäßigen Störungsindex 2.

d) Weder für $\varrho = 0$ noch für $\varrho = -2$ hat die Klasse aller DA–Systeme (2.97) (mit beliebigem Λ_i^2) einen gleichmäßigen Störungsindex.

Beweis Ist $|\Lambda_i^2 - 4| \geq \Delta_\lambda$, dann gibt es eine von Λ_i unabhängige Konstante $\alpha_0 > 0$ mit $\alpha \geq -\alpha_0$. Unter Verwendung des Lemmas von Gronwall kann man deshalb nachweisen, daß die Abschätzung (2.93) mit $r = 0$ und $C = \mathcal{O}(e^{\alpha_0 t})$ erfüllt ist. (α_0 hängt jedoch von Δ_λ ab, es gilt $\lim_{\Delta_\lambda \to 0} C = \infty$.) In dem Streifen $\{\Lambda_i : |\Lambda_i^2 - 4| \leq \Delta_\lambda\}$ hängt das Stabilitätsverhalten der Lösung von (2.97) vom Vorzeichen von α und damit auch von ϱ ab (vgl. Abb. 2.6). Ist $\alpha > 0$, so kann man unter Verwendung von Lemma 3 wie in Beispiel 15b die gleichmäßige Fehlerschranke (2.93) mit $r = 1$, d. h. $m = 2$, nachweisen. Ist $\alpha < 0$, so gibt es wegen $C = \mathcal{O}(e^{-\alpha t})$ keine gleichmäßige Fehlerschranke der Form (2.93), denn bei Annäherung von Λ_i^2 an 4 kann $-\alpha$ beliebig groß werden. ∎

Sowohl für $\varrho = 0$ als auch für $\varrho = -2$ trennt also das Index-2-System (2.95) mit $\lambda_n^2 = 4$ eine Klasse von Index-1-Systemen (2.97), die sich nahezu wie Index-2-Systeme verhalten, von einer Klasse von Index-1-Systemen (2.97), in denen Fehler mit unbeschränkt großen Faktoren verstärkt werden können ([151]).

Als Folgerung aus Lemma 4 ergeben sich Aussagen zum gleichmäßigen Störungsindex der Klasse aller Semidiskretisierungen (2.96):

Folgerung 2 *Die Klasse aller hinreichend feinen Semidiskretisierungen (2.96) der partiellen Differentialgleichungen (2.94) hat*

a) den gleichmäßigen Störungsindex 1, falls $\varrho \in \{0, -2\}$ und $\lambda_n^2 \neq 4$ für alle $n > 0$,

b) den gleichmäßigen Störungsindex 2, falls $\varrho = -2$ und $\lambda_{n_0}^2 = 4$ für ein $n_0 \in \mathbb{N}$ und

c) keinen gleichmäßigen Störungsindex, falls $\varrho = 0$ und $\lambda_{n_0}^2 = 4$ für ein $n_0 \in \mathbb{N}$.

Beweis a) und b) folgen aus Lemma 4, $\lim_{N\to\infty} \Lambda_i = \lambda_i$ und $\Lambda_i < \lambda_i$, $(i = 1, \ldots, N)$ (vgl. die Markierungen „$*$“ in Abb. 2.6).
In c) gilt für $\lambda_{n_0}^2 = 4$ die Abschätzung $\Lambda_{n_0}^2 = (\frac{2}{\Delta_x} \sin \Delta_x)^2 = 4(1 - \frac{1}{3}\Delta_x^2) + \mathcal{O}(\Delta_x^4)$, also $\alpha = -3((N+1)/L)^2 + \mathcal{O}(N^4)$, so daß Fehler in (2.98) mit $\exp(3(\frac{N+1}{L})^2 t)$ verstärkt werden können. Da dieser Faktor für $N \to \infty$ unbeschränkt wächst, können keine gleichmäßigen Fehlerschranken angegeben werden. ∎

Ergebnis Für jedes der beiden hier betrachteten Systeme partieller Differentialgleichungen mit algebraischen Nebenbedingungen führt die Finite–Differenzen–Diskretisierung auf semidiskrete DA–Systeme, die bei fixiertem L und hinreichend kleinem Δ_x stets den Störungsindex 1 haben. Für kleine Δ_x wird die Empfindlichkeit der Lösung gegenüber Störungen nur durch Fehlerschranken, die gleichmäßig bezüglich Δ_x sind, korrekt beschrieben. Ist L ein ganzzahliges Vielfaches von $\pi/2$, so hat eines der Systeme (2.95) den Index 2. In Abhängigkeit von den Koeffizienten der partiellen Differentialgleichungen ist es hier möglich, daß die Klasse der Semidiskretisierungen den gleichmäßigen Störungsindex 2 hat, i. allg. können kleine Fehler jedoch um unbeschränkt große Faktoren verstärkt werden. Es ist in diesem Beispiel *nicht* möglich, allein aus Kenntnissen über die Empfindlichkeit der Lösung der partiellen Differentialgleichungen (2.94) gegenüber Störungen Rückschlüsse zu ziehen auf den Einfluß kleiner Störungen in den semidiskreten Problemen (2.96); hierzu sind zusätzlich Informationen über Details der Ortsdiskretisierung erforderlich.

2.5 Zusammenfassung

Der Störungsindex und die im Zusammenhang mit seiner Definition entwickelte Störungstheorie sind wesentliche und bewährte Hilfsmittel bei der theoretischen Untersuchung und bei der praktischen Realisierung von Diskretisierungsverfahren für differentiell-algebraische Systeme. In diesem Abschnitt haben wir für verschiedene DA–Systeme von höherem Index, die häufig in praktischen Anwendungen auftreten, die Störungstheorie für die analytische und die numerische Lösung bis ins Detail entwickelt.

Ein Schwerpunkt lag dabei auf dem Nachweis der (plausiblen) Analogie der Fehlerschranken für die analytische und die numerische Lösung. In der Literatur nutzt man in Konvergenzbeweisen für Diskretisierungsverfahren aus, daß die differentiellen Lösungskomponenten robust gegenüber kleinen Störungen sind. Für Index-2- und Index-3-Systeme in Hessenbergform wurde gezeigt, daß diese Robustheit ihre Ursache in den Eigenschaften der analytischen Lösung hat. Im Unterschied zu linearen Systemen gibt es jedoch im Fall nichtlinearer Kopplung zwischen dem differentiellen und dem algebraischen Teil i. allg. keine Lösungskomponenten, die stetig von kleinen Störungen abhängen. Die Fehlerschranken beweisen, daß der Fehlerterm in den differentiellen Komponenten entscheidend davon beeinflußt wird, ob und wie stark das System nichtlinear in den algebraischen Variablen ist. Mit zahlreichen Beispielen wurde belegt, daß die angegebenen Fehlerschranken scharf sind und i. allg. nicht verbessert werden können.

Für nichtlineare Systeme führt der höhere Index bei der numerischen Lösung in jeder Lösungskomponente zu einem (gegenüber der Theorie der gewöhnlichen Differentialgleichungen) zusätzlichen Fehlerterm, der für größere Schrittweiten gegenüber dem Diskretisierungsfehler vernachlässigbar ist, für kleinere Schrittweiten jedoch rasch unbeschränkt wächst. Der zusätzliche Fehlerterm wurde in Testrechnungen nachgewiesen, er entspricht in der Regel denjenigen Termen in den Fehlerschranken für die analytische Lösung, die nicht stetig von Störungen im DA–System abhängen. An einem Beispiel wurde jedoch gezeigt, daß der zusätzliche Fehlerterm in Einzelfällen auch allein durch die Diskretisierung des DA–Systems entstehen kann.

In der Störungstheorie haben wir den Einfluß kleiner Störungen auf die Lösung von DA–Systemen bewußt nicht nur qualitativ, sondern vor allem quantitativ charakterisiert. So werden wir in den folgenden Abschnitten Diskretisierungsverfahren für Index-2-Systeme analysieren und die MKS–Modellgleichungen in Index-2-Formulierung integrieren, obwohl die Lösung von Index-2-Systemen *nicht* stetig von kleinen Störungen abhängt — der zusätzliche Fehlerterm ist hier bei geeigneter Implementierung in praxi gegenüber dem Diskretisierungsfehler vernachlässigbar. Umgekehrt zeigen die Abschätzungen für Systeme mit großen Baumgarte–Koeffizienten und für semidiskretisierte Systeme partieller Differentialgleichungen, daß Störungen auch in Systemen mit niedrigem (klassischem) Störungsindex erheblich verstärkt werden können.

Während die Störungstheorie für nichtsteife (gewöhnliche) differentiell-algebraische Systeme vom Index ≤ 3 in Hessenbergform im wesentlichen abgeschlossen ist, sind für die kommenden Jahre Erweiterungen auf Systeme partieller Differentialgleichungen mit algebraischen Nebenbedingungen zu erwarten. Es ist jedoch fraglich, ob hierzu ein ähnlich universeller Begriff wie der Störungsindex für gewöhnliche DA–Systeme gefunden werden kann.

Kapitel 3

Zur numerischen Lösung von Anfangwertproblemen für differentiell-algebraische Systeme

Häufig führt die Modellierung naturwissenschaftlicher oder technischer Prozesse auf differentiell-algebraische Systeme. Obwohl diese DA–Systeme in der Regel zumindest lokal in äquivalente Systeme gewöhnlicher Differentialgleichungen transformiert werden könnten, ist es für die numerische Lösung meist günstiger, direkt das DA–System zu diskretisieren. Die Frage nach *geeigneten* Diskretisierungsverfahren erweist sich dabei als außerordentlich komplex, denn einerseits müssen (wie in der Theorie gewöhnlicher Differentialgleichungen) verschiedene Verfahren für eine bestimmte Klasse differentiell-algebraischer Systeme miteinander verglichen werden, andererseits können die Modellgleichungen auf verschiedene Art formuliert werden. Diese *analytisch äquivalenten* Formulierungen der Modellgleichungen haben stets dieselbe analytische Lösung, dagegen hängen oft Effizienz und Robustheit der numerischen Verfahren sehr stark von der Auswahl einer geeigneten Formulierung ab.

Generell können DA–Systeme von höherem Index nur dann direkt, d. h. ohne vorherige analytische Indexreduktion, numerisch gelöst werden, wenn man bei der Implementierung die Struktur des DA–Systems berücksichtigt (z. B. bei der Schrittweitensteuerung). Bei der Implementierung der Verfahren kann man sich also nicht auf die in den theoretischen Untersuchungen betrachteten DA–Systeme mit vergleichsweise einfacher Struktur (z. B. Hessenbergform) beschränken, die problemabhängigen Schnittstellen der Integrationssoftware müssen statt dessen einen möglichst breiten Anwendungsbereich abdecken. Neben der theoretischen Untersuchung von Diskretisierungsverfahren für DA–Systeme von höherem Index hat deshalb vor allem die Umsetzung im Computerprogramm besondere praktische Bedeutung.

In verschiedenen Monographien wird eine Übersicht der umfangreichen Literatur zur numerischen Lösung von DA–Systemen gegeben, wir verweisen hier insbesondere auf die überarbeiteten 2. Auflagen der Bücher von Brenan, Campbell und Petzold ([39]) und Hairer und Wanner ([84, vor allem Kapitel VI und VII]), daneben auf die klassische Monographie von Griepentrog und März ([76]), die detaillierte Untersuchung von Runge–Kutta–Verfahren in [81] und die Einführung in [154, Kapitel 9 und 10]). Zur dynamischen Simulation von mechanischen Mehrkörpersystemen (MKS) mit Zwangsbedingungen ist

eine Monographie von Eich und Führer ([57]) in Vorbereitung.

Im vorliegenden Kapitel analysieren wir einige Teilaspekte der numerischen Lösung von DA–Systemen mit glatten Zwangsbedingungen. Der Schwerpunkt liegt dabei auf den MKS–Modellgleichungen und auf Index-2-Systemen in Hessenbergform, z. T. können die Ergebnisse auch auf andere Anwendungsgebiete übertragen werden. Zunächst konzentrieren wir uns in Abschnitt 3.1 auf analytische Transformationen von DA–Systemen und auf die Auswahl einer für die numerische Integration geeigneten Formulierung der MKS–Modellgleichungen. Dabei erweisen sich die Index-2-Formulierung und die Gear–Gupta–Leimkuhler–Formulierung (GGL–Formulierung) als besonders vorteilhaft.

In Anlehnung an die Störungstheorie aus Kapitel 2 wird anschließend die Konvergenz partitionierter Ein- und Mehrschrittverfahren für Index-2-Systeme in Hessenbergform untersucht (Abschnitt 3.2). Partitionierte Verfahren sind besonders zur Integration nichtsteifer DA–Systeme geeignet, als Beispiele betrachten wir in Abschnitt 3.3 halb-explizite Runge–Kutta–Verfahren und partitionierte Mehrschrittverfahren vom Adams–Typ. Durch neuartige Verfahrensansätze (explizite Stufen, Kombination von Adams–Verfahren mit BDF) erreicht man wesentlich effizientere Verfahren höherer Ordnung als sie bisher aus der Literatur bekannt sind. Für die dynamische Simulation von mechanischen Mehrkörpersystemen wird schließlich der Integrator HEDOP5 vorgestellt, der auf einem halbexpliziten Runge–Kutta–Verfahren 5. Ordnung aufbaut.

Leistungsfähige Integratoren für DA–Systeme sind heute Bestandteil von Simulationspaketen in vielen technischen Anwendungsgebieten. Am Beispiel der dynamischen Simulation von Rad–Schiene–Systemen mit dem Simulationspaket SIMPACK kommen wir im nachfolgenden Kapitel 4 auf diese praktische Anwendung der numerischen Lösungsverfahren zurück.

3.1 Indexreduktion und numerische Lösungsverfahren für differentiell-algebraische Systeme von höherem Index

Wegen der erheblichen Schwierigkeiten bei der numerischen Lösung von DA–Systemen vom Index > 1 ist es bewährte Praxis, bereits vor der Anwendung numerischer Verfahren das DA–System (1.1) analytisch zu transformieren. Für Systeme von höherem Index (Index 2, Index 3) vereinfacht dies in der Regel die anschließende numerische Integration (vgl. z. B. Beispiel 6), für nichtlineare Systeme von hohem Index (Index > 3) ist die analytische Indexreduktion sogar eine zwingende Voraussetzung für eine zuverlässige numerische Lösung.

Diese analytischen Transformationen beruhen auf der Tatsache, daß (unter geeigneten Glattheitsvoraussetzungen) jede Lösung $x(t)$ von (1.1) auch

$$\frac{d^r}{dt^r} F(t, x(t), x'(t)) = 0 \ , \quad (r > 0)$$

erfüllt. Enthält das DA–System (1.1) algebraische Zwangsbedingungen, so genügt jede Lösung $x(t)$ auch den in $\frac{d^r}{dt^r} F(t, x(t), x'(t)) = 0$ auftretenden *versteckten* Zwangsbedingungen. Ersetzt man in (1.1) Zwangsbedingungen durch ihre zeitlichen Ableitungen, d. h. durch versteckte Zwangsbedingungen, so bleibt die analytische Lösung des Problems unverändert, das neu entstandene *analytisch äquivalente* DA–System hat jedoch i. allg. einen niedrigeren Index ([72]).

In diesem Abschnitt geben wir einen Überblick über numerische Verfahren, die das indexreduzierte DA–System lösen und durch geeignete Projektionen oder stabilisierende Terme garantieren, daß die numerische Lösung des indexreduzierten DA–Systems auch die ursprünglich in (1.1) enthaltenen Zwangsbedingungen möglichst gut erfüllt. Insbesondere erlaubt die in Kapitel 2 entwickelte Störungstheorie den detaillierten Vergleich verschiedener indexreduzierter Formulierungen der Modellgleichungen für mechanische Mehrkörpersysteme ([17]). Für die Gear–Gupta–Leimkuhler–Formulierung wird eine Erweiterung auf MKS mit Kontaktbedingungen angegeben ([20], [23]).

Indexreduktion und Drift-off-Effekt

Das Prinzip der Indexreduktion kann für DA–Systeme (1.1) allgemeiner Struktur formuliert werden, dabei sind neben $F(t, x(t), x'(t)) = 0$ die Gleichungen des Ableitungsfeldes (engl.: derivative array) $\frac{d^r}{dt^r} F(t, x(t), x'(t)) = 0$, $(r = 1, \ldots, m)$ für ein gewisses $m > 0$ zu betrachten ([72]). In den letzten Jahren wurden verschiedene Konzepte entwickelt und in Programmpakete umgesetzt, um die Indexreduktion zu automatisieren und die indexreduzierten Gleichungen zu integrieren ([116], [47], [100]). Die partiellen Ableitungen von F können dabei z. B. mittels algebraischer Formelmanipulationsprogramme (Maple, Mathematica) oder mittels automatischer Differentiation ([34]) berechnet werden. Weitgehend unabhängig von der Struktur von (1.1) ist die automatisierte Indexreduktion zur numerischen Lösung von DA–Systemen von hohem Index geeignet, wenn die Dimension des Systems nicht zu groß ist. Der numerische Aufwand zur Berechnung der Lösung ist jedoch erheblich.

Für Anwendungen, die auf Modellgleichungen (1.1) einer bestimmten Struktur (z. B. Hessenbergform) führen, gelangt man dagegen bei Ausnutzung dieser Struktur zu Integrationsverfahren, die um mehrere Größenordnungen schneller als bei automatisierter Indexreduktion sind. Wir beschränken uns hier auf die Modellgleichungen für mechanische Mehrkörpersysteme, für die nicht nur die schon in Abschnitt 2.3 betrachteten Modellgleichungen

$$\begin{aligned} M(q)q''(t) &= f(q, q', \lambda) - G^T(q)\lambda \\ 0 &= g(q) \ , \end{aligned} \tag{3.1}$$

sondern auch alle im Ableitungsfeld benötigten Ableitungen der Zwangsbedingungen mit Hilfe von Mehrkörperformalismen automatisch generiert werden können (vgl. z. B. [141]). In (3.1) werden die Bezeichnungen von Abschnitt 2.3 verwendet, dabei sind $v(t) = q'(t)$ wiederum die Geschwindigkeitskoordinaten des MKS. Neben den (holonomen) Zwangsbedingungen $r^{(\mathrm{L})}(q) = g(q) = 0$ auf Ebene der **L**agekoordinaten erfüllt die Lösung von (3.1) die versteckten Zwangsbedingungen auf Ebene der **G**eschwindigkeits- bzw. **B**eschleunigungskoordinaten:

$$0 = \frac{d}{dt} g(q(t)) = G(q(t))q'(t) = [G(q)v](t) =: r^{(\mathrm{G})}(q(t), v(t)) \ , \tag{3.2}$$

$$0 = \frac{d^2}{dt^2} g(q(t)) = [GM^{-1}](q)(f(q, v, \lambda) - G^T(q)\lambda) + g_{qq}(q)(v, v) =: r^{(\mathrm{B})}(q, v, \lambda) \ . \tag{3.3}$$

Lemma 5 *Wenn gegebene Funktionen $q(t)$, $v(t)$ und $\lambda(t)$ mit $v(t) = q'(t)$ die Gleichungen*

$$M(q(t))v'(t) = f(q(t), v(t), \lambda(t)) - G^T(q(t))\lambda(t)$$

in (3.1) erfüllen, dann gilt

$$\begin{aligned} r^{(\mathrm{L})}(q(t)) = g(q(t)) = 0\ , \quad (t \in [0,T]) \quad &\Leftrightarrow \quad r^{(\mathrm{G})}(q(t), v(t)) = 0\ , \quad (t \in [0,T]) \\ &\Leftrightarrow \quad r^{(\mathrm{B})}(q(t), v(t), \lambda(t)) = 0\ , \quad (t \in [0,T])\ , \end{aligned}$$

sofern $r^{(\mathrm{L})}(q(0)) = 0$ und $r^{(\mathrm{G})}(q(0), v(0)) = 0$ ist.

Beweis Aus $r^{(\mathrm{G})}(q(\tau), v(\tau)) = \frac{d}{d\tau} r^{(\mathrm{L})}(q(\tau))$, $(\tau \in [0,T])$ folgt

$$g(q(t)) = r^{(\mathrm{L})}(q(t)) = r^{(\mathrm{L})}(q(0)) + \int_0^t \frac{d}{d\tau} r^{(\mathrm{L})}(q(\tau))\, d\tau = \int_0^t r^{(\mathrm{G})}(q(\tau), v(\tau))\, d\tau\ ,$$

also $r^{(\mathrm{L})}(q(t)) = 0$, falls $r^{(\mathrm{G})}(q(\tau), v(\tau)) = 0$, $(\tau \in [0,t])$ und $r^{(\mathrm{L})}(q(0)) = 0$. Analog beweist man die anderen Behauptungen des Lemmas. ■

Ersetzt man in (3.1) die Zwangsbedingung $r^{(\mathrm{L})}(q) = g(q) = 0$ durch $r^{(\mathrm{G})}(q, v) = 0$, so bleibt also für konsistente Anfangswerte mit $g(q(0)) = 0$ die Lösung des Anfangswertproblems unverändert. Das entstehende DA–System

$$\begin{aligned} q' &= v \\ M(q)v' &= f(q, v, \lambda) - G^T(q)\lambda \\ 0 &= G(q)v \end{aligned} \tag{3.4}$$

hat den Index 2 und heißt *Index-2-Formulierung* der Modellgleichungen. Ebenso erhält man ein zu (3.1) analytisch äquivalentes Index-1-System, die *Index-1-Formulierung* der Modellgleichungen, indem $g(q) = 0$ durch $r^{(\mathrm{B})}(q, v, \lambda) = 0$ ersetzt wird ([154, S. 404f]). Integriert man die Index-1- oder die Index-2-Formulierung, so wird die numerische Lösung die versteckten Zwangsbedingungen (3.3) bzw. (3.2) mit hoher Genauigkeit erfüllen. Dagegen verletzt die numerische Lösung i. allg. die holonomen Zwangsbedingungen $g(q) = 0$ in (3.1), die numerische Lösung entfernt sich mit zunehmender Zeit immer weiter von der Mannigfaltigkeit $\{\eta : g(\eta) = 0\}$ (Drift-off-Effekt).

Beispiel 17 Integriert man die Bewegungsgleichungen eines starren Radsatzes für das in Beispiel 27 beschriebene Manöver in Index-1- oder Index-2-Formulierung, so ergibt sich ein quadratisches bzw. lineares Fehlerwachstum in den Zwangsbedingungen. Abb. 3.1 auf S. 69 zeigt den Abstand des rechten Rades zur Schiene für die Integratoren DASSL ([39], Toleranz $\mathrm{RTOL} = \mathrm{ATOL} = 10^{-5}$) und RADAU5 ([84], Toleranz $\mathrm{RTOL} = \mathrm{ATOL} = 10^{-4}$), (links: Index-2-Formulierung, rechts: Index-1-Formulierung, man beachte die unterschiedliche Skaleneinteilung auf der Ordinatenachse). Wegen der Unterschiede in der Schrittweitensteuerung ist der Drift für die beiden Integratoren verschieden groß, jedoch heben für beide Integratoren die Räder als Folge des Drift-off-Effekts von den Schienen ab („der fliegende Radsatz“ [55]), für praktische Anwendungen sind diese Simulationsergebnisse vollständig unbrauchbar.

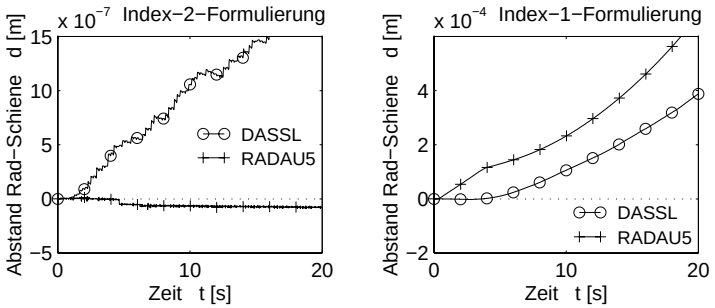


Abbildung 3.1: Drift-off-Effekt bei der dynamischen Simulation eines starren Radsatzes.

Alishenas ([1], [2]) untersucht detailliert die Auswirkungen des Drift-off-Effekts und weist insbesondere nach, daß dieser Effekt einen (gegenüber der Integration der zu (3.1) gehörenden Zustandsform) zusätzlichen Fehlerterm in der numerischen Lösung zur Folge hat (dieser durch Indexreduktion entstehende Fehlerterm ist nicht zu verwechseln mit dem in Kapitel 2 nachgewiesenen Fehlerterm $\frac{1}{h}\theta$, $\frac{1}{h^2}\theta$, ..., der für DA–Systeme von höherem Index charakteristisch ist). Alishenas weist nach, daß der Drift für die Index-2-Formulierung deutlich kleiner ist als für die Index-1-Formulierung. Für die praktische Anwendung ist es zwingend erforderlich, die numerische Lösung zu projizieren, so daß die Zwangsbedingungen eingehalten werden (vgl. Abschnitt 3.3.2).

Die Kombination von Diskretisierungsverfahren für die Index-1- bzw. Index-2-Formulierung mit Projektionsschritten wird u. a. in [2], [55], [106] und [144] untersucht, bei geeigneter Implementierung benötigen die Projektionsschritte nur geringfügigen zusätzlichen numerischen Aufwand. In Abschnitt 3.3 werden wir partitionierte Verfahren für die Index-2-Formulierung konstruieren, die bei der Implementierung mit zusätzlichen Projektionsschritten gekoppelt werden.

Gear–Gupta–Leimkuhler–Formulierung

Ein Nachteil der Koordinatenprojektionen ist, daß sie Eingriffe in vorhandene Integrationssoftware erforderlich machen. Ändert sich die Struktur der Modellgleichungen (z. B. zusätzliche algebraische Gleichungen wie in (3.14)), so müssen nicht nur die vom Nutzer bereitzustellenden Teile der Simulationssoftware, sondern außerdem der Integrator selbst modifiziert werden. Dies kann vermieden werden, wenn die Zwangsbedingungen auf Lageebene nicht in einem separaten Projektionsschritt, sondern bereits bei der Formulierung der Bewegungsgleichungen berücksichtigt werden (vgl. [58] für einen Vergleich von zahlreichen aus der Literatur bekannten Lösungsansätzen).

Das klassische Verfahren hierzu geht auf Baumgarte ([33]) zurück: Die Zwangsbedingung $g(q) = 0$ wird weder durch (3.2) noch durch (3.3) ersetzt sondern durch

$$0 = \frac{d^2}{dt^2} g(q(t)) + 2\alpha \frac{d}{dt} g(q(t)) + \beta g(q(t)) = r^{(\mathrm{B})}(q, v, \lambda) + 2\alpha r^{(\mathrm{G})}(q, v) + \beta r^{(\mathrm{L})}(q)\ . \tag{3.5}$$

Die Parameter α und β werden so gewählt, daß die Lösung der linearen Differentialgleichung $0 = w''(t) + 2\alpha w'(t) + \beta w(t)$ asymptotisch stabil ist, meist setzt man $\alpha = \beta$ (aperiodischer Grenzfall). Das entscheidende (und ungelöste) Problem ist hier die Wahl der Baumgarte–Koeffizienten α, β. Sind α und β klein, so ist (3.5) nahezu identisch mit (3.3) und die numerische Lösung kann sich wiederum weit von der Mannigfaltigkeit $\{\,\eta : g(\eta) = 0\,\}$ entfernen. Sind α und β sehr groß, so hat das nach Baumgarte stabilisierte System zwar den (klassischen) Index 1, seine numerische Lösung ist aber (mindestens) ebenso kompliziert wie die des Index-3-Systems (3.1) (vgl. Abschnitt 2.4 und [29]).

Statt der Linearkombination in (3.5) erweist es sich als wesentlich günstiger, sowohl die ursprünglichen als auch die versteckten Zwangsbedingungen *direkt* in der Formulierung der Modellgleichungen zu berücksichtigen. Der dadurch wachsenden Zahl von Gleichungen im DA–System trägt man durch Einführung von (künstlichen) Hilfsvariablen Rechnung. Ein derartiger Ansatz wurde erstmals von Gear, Gupta und Leimkuhler ([74]) verwendet, die gleichzeitig die Zwangsbedingungen auf Ebene der Lage- und der Geschwindigkeitskoordinaten betrachten:

$$\begin{aligned} q' &= v - G^T(q)\eta \,, \\ M(q)v' &= f(q,v,\lambda) - G^T(q)\lambda \,, \\ 0 &= G(q)v \quad (= r^{(\mathrm{G})}(q,v)\,) \,, \\ 0 &= g(q) \qquad (= r^{(\mathrm{L})}(q)\,) \,. \end{aligned} \tag{3.6}$$

Satz 9 ([74]) *Die in (3.6) definierte* Gear–Gupta–Leimkuhler–Formulierung *(GGL–Formulierung) der MKS–Modellgleichungen bildet ein DA–System vom Index 2, das zu den ursprünglichen Modellgleichungen (3.1) analytisch äquivalent ist.*

Beweis Multipliziert man in (3.6) die dynamischen Gleichungen „$M(q)v' = \ldots$“ von links mit $M^{-1}(q)$, so erhält man ein DA–System der Form (1.8) mit $y = (q^T, v^T)^T$, $z = (\eta^T, \lambda^T)^T$. Durch elementares Nachrechnen zeigt man, daß für dieses System wegen (2.56) die Index-2-Bedingung (1.10) erfüllt ist. Da der Differentiationsindex invariant gegenüber der Multiplikation mit $M^{-1}(q)$ ist, hat (3.6) nach Beispiel 3b den Differentiationsindex 2. In (3.6) folgt für $\eta : [0,T] \to \mathbb{R}^{n_\lambda}$ aus $g(q) = 0$

$$0 = \frac{d}{dt} g(q(t)) = G(q)q'(t) = G(q)v - [GG^T](q)\eta = -[GG^T](q)\eta \,,$$

also verschwinden (für die analytische Lösung) die Hilfsvariablen η identisch, denn $G(q)$ hat Vollrang (vgl. (2.56)). ∎

Bemerkung 15 a) Durch die Berücksichtigung der Zwangsbedingungen auf Ebene der Geschwindigkeitskoordinaten wird der Index der MKS–Modellgleichungen auf 2 reduziert, wobei für die GGL–Formulierung darüber hinaus ein Abdriften der Lösung von der Zwangsmannigfaltigkeit unmöglich ist. Es verringert sich jedoch nicht nur der Index des DA–Systems und damit die Empfindlichkeit der algebraischen Komponenten λ gegenüber kleinen Störungen. Besonders wichtig für die praktische Umsetzung (z. B. Schrittweitensteuerung) ist die Tatsache, daß auch die differentiellen Komponenten q und v sehr viel

Tabelle 3.1: Fehlerschranken für verschiedene Formulierungen der MKS–Modellgleichungen (3.1) mit $f = f(q,v)$, (wie in Satz 6 ist $D(t) = \int SM^{-1}\delta + \|\theta\|_{C^1}$).

	Index-3-Formulierung	GGL–Formulierung	Index-2-Formulierung
$\|\hat{q} - q\|$	$\mathcal{O}(\int SM^{-1}\delta + \|\theta\|_{C^0} + D^2(t))$	$\mathcal{O}(\int SM^{-1}\delta + \|\theta\|_{C^0} + \|\theta\|_{C^0} \cdot D(t))$	$\mathcal{O}(\int SM^{-1}\delta + \|\theta\|_{C^0})$
$\|S(\hat{q})\hat{v} - S(q)v\|$			
$\|G(\hat{q})\hat{v} - G(q)v\|$	$\mathcal{O}(\|\delta\|_{C^0} + \|\theta\|_{C^1})$		
$\|\hat{\lambda} - \lambda\|$	$\mathcal{O}(\|\delta\|_{C^0} + \|\theta\|_{C^2})$	$\mathcal{O}(\|\delta\|_{C^0} + \|\theta\|_{C^1})$	$\mathcal{O}(\|\delta\|_{C^0} + \|\theta\|_{C^1})$

robuster gegenüber Störungen sind als im Index-3-System (3.1). Dies folgt aus den Ergebnissen der in Kapitel 2 entwickelten Störungstheorie.
Als Beispiel werden in Tab. 3.1 die Fehlerschranken für Index-3-, GGL- und Index-2-Formulierung in dem aus praktischer Sicht wichtigen Spezialfall $f = f(q,v)$ zusammengefaßt. Wir verwenden hier die Bezeichnungen aus Satz 6, zur Vereinfachung werden die Fehler in den Anfangswerten vernachlässigt. Das Residuum im differentiellen Teil wird mit δ bezeichnet, θ ist das Residuum in den n_λ bzw. $2n_\lambda$ algebraischen Gleichungen. Während die Abschätzungen für Index-3- und GGL–Formulierung wegen $f_\lambda \equiv 0$ aus Satz 6 (mit $\mu = \nu = 0$) bzw. Satz 3a (mit $\mu = 0$) folgen, kann man die Fehlerschranke für die Index-2-Formulierung unter Verwendung von $\frac{d}{dt}\hat{q} = \hat{v}$ analog zu Satz 3b beweisen. Bis auf den in praxi kaum relevanten Fehlerterm $\|\theta\|_{C^0} \cdot D(t)$ (vgl. Beispiel 10) sind die Fehlerschranken für die GGL–Formulierung identisch zu denen der Index-2-Formulierung. Alle Fehlerterme, die Ableitungen von θ enthalten, sind dabei deutlich kleiner als für die Index-3-Formulierung: $\|\theta\|_{C^0} \cdot \|\theta\|_{C^1}$ statt $\|\theta\|^2_{C^1}$ (bzw. $\|\theta\|_{C^1}$) in den differentiellen Lösungskomponenten und $\|\theta\|_{C^1}$ statt $\|\theta\|_{C^2}$ in den algebraischen Lösungskomponenten. Für die numerische Lösung erhält man entsprechend für die GGL–Formulierung zusätzliche Fehlerterme der Größenordnung $\frac{1}{h}\theta^2$ (in q, v) und $\frac{1}{h}\theta$ (in λ) statt der Fehlerterme $\frac{1}{h^2}\theta^2$ (in q und $S(q)v$), $\frac{1}{h}\theta$ (in $G(q)v$) und $\frac{1}{h^2}\theta$ (in λ) für die Index-3-Formulierung. Obwohl Lage- und Geschwindigkeitskoordinaten sowohl für die Index-3- als auch für die GGL–Formulierung Index-2-Variable (im Sinne des Störungsindexes) sind, wird trotzdem der zusätzliche Fehlerterm in *allen* Lösungskomponenten um den (kleinen) Faktor h reduziert, wenn man die versteckte Zwangsbedingung $G(q)v = 0$ explizit in die Modellgleichungen aufnimmt.

b) Ausgehend von der Index-2-Formulierung der Modellgleichungen kann man die GGL–Formulierung (3.6) auch als Hinzunahme der ursprünglichen Zwangsbedingungen $g(q) = 0$ zur Vermeidung des Abdriftens (und damit zur Stabilisierung der Integration) interpretieren. Dies motiviert die synonyme Bezeichnung *stabilisierte Index-2-Formulierung*. Führer und Leimkuhler ([66], [68]) verallgemeinern diese Idee und untersuchen die *stabilisierte*

Index-1-Formulierung der Modellgleichungen:

$$\begin{aligned} q' &= v - (\tfrac{\partial}{\partial q} r^{(\mathrm{L})}(q))^T \eta^{(1)} - (\tfrac{\partial}{\partial q} r^{(\mathrm{G})}(q,v))^T \eta^{(2)} \\ M(q)v' &= f(q,v,\lambda) - G^T(q)\lambda - (\tfrac{\partial}{\partial v} r^{(\mathrm{G})}(q,v))^T \eta^{(2)} \\ 0 &= r^{(\mathrm{B})}(q,v,\lambda) \\ 0 &= r^{(\mathrm{L})}(q) \\ 0 &= r^{(\mathrm{G})}(q,v) \end{aligned} \tag{3.7}$$

mit Funktionen $\eta^{(1)}, \eta^{(2)} : [0,T] \to \mathbb{R}^{n_\lambda}$. (In (3.7) ist $\frac{\partial}{\partial q} r^{(\mathrm{L})}(q) = \frac{\partial}{\partial v} r^{(\mathrm{G})}(q,v) = G(q)$). Ähnlich wie in (3.6) werden dabei den Gleichungen der Index-1-Formulierung die Zwangsbedingungen auf Ebene der Lage- und Geschwindigkeitskoordinaten hinzugefügt, die man durch Hilfsvariablen $\eta = ((\eta^{(1)})^T, (\eta^{(2)})^T)^T \in \mathbb{R}^{2n_\lambda}$ an den differentiellen Teil der Modellgleichungen ankoppelt. Mit den Bezeichnungen von Tab. 3.1 führt dies zu Fehlerschranken $\mathcal{O}(\int SM^{-1}\delta + \|\theta\|_{C^0} + \|\theta\|_{C^0} \cdot D(t))$ für *alle* Lösungskomponenten q, v und λ. Wie in Abschnitt 2.2.3 läßt sich zeigen, daß diese Fehlerschranken ebenso wie die Schranken aus Tab. 3.1 scharf sind. Verglichen mit der GGL–Formulierung verringert die Hinzunahme der n_λ Zwangsbedingungen (3.3) auf Ebene der Beschleunigungskoordinaten zwar den zusätzlichen Fehlerterm in λ, jedoch *nicht* denjenigen in q und v. Die stabilisierte Index-1-Formulierung ist deshalb der GGL–Formulierung nur vorzuziehen, wenn $n_\lambda \ll n_q$ ist und die Lagrangeschen Multiplikatoren bzw. die Zwangskräfte mit hoher Genauigkeit berechnet werden sollen.

Bemerkung 15b zeigt, daß die Hinzunahme versteckter Zwangsbedingungen keineswegs in jedem Fall den für DA–Systeme von höherem Index charakteristischen zusätzlichen Fehlerterm reduziert. Dies hat seine Ursache in dem komplizierten Mechanismus der Fehlerfortpflanzung in DA–Systemen, der nicht allein durch den (Störungs-)Index, sondern vor allem auch durch die Struktur des Systems bestimmt wird (vgl. Kapitel 2). Der aussagefähige Vergleich verschiedener Formulierungen von MKS–Modellgleichungen gab ursprünglich die Motivation für die aufwendigen Untersuchungen des Kapitels 2.

Als Warnung, daß durch Hinzunahme versteckter Zwangsbedingungen die Empfindlichkeit von Lösungskomponenten gegenüber kleinen Störungen sogar *verstärkt* werden kann, diene das nachfolgende (akademische) Beispiel:

Beispiel 18 Das Anfangswertproblem $y_1(0) = 0$, $y_2(0) = 1$, $z(0) = 0$ für das Index-2-System

$$\begin{aligned} y_1' &= z \\ y_2' &= z \\ 0 &= y_1 + y_2^2 - 1 \end{aligned}$$

hat eine konstante Lösung $y_1(t) \equiv 0$, $y_2(t) \equiv 1$, $z(t) \equiv 0$, die vom impliziten Eulerverfahren exakt bestimmt wird. Wegen $f_z = (1,1)^T = \text{const}$ hängen die differentiellen Lösungskomponenten y sowohl für die analytische als auch für die numerische Lösung stetig von Störungen ab (Sätze 3b und 5b). Koppelt man jedoch analog zu (3.7) die versteckte Zwangsbedingung $\frac{d}{dt}(y_1 + y_2^2 - 1) = 0$ an dieses System an, so erhält man das

analytisch äquivalente System

$$\begin{aligned} y_1' &= z - \eta \\ y_2' &= z - 2y_2\eta \\ 0 &= z + 2y_2 z \\ 0 &= y_1 + y_2^2 - 1 \,, \end{aligned} \tag{3.8}$$

d. h. das System (2.51) aus Beispiel 10 (löse $0 = z + 2y_2 z$ nach z auf und setze dies in die Differentialgleichungen ein; die in (2.51) mit z bezeichnete algebraische Variable entspricht der Variablen η in (3.8)). In Beispiel 10 wurde gezeigt, daß für die mit dem impliziten Eulerverfahren berechnete numerische Lösung von (3.8) die Komponenten y nicht stetig von Störungen der Zwangsbedingung $y_1 + y_2^2 - 1 = 0$ abhängen, d. h., beim Hinzufügen der versteckten Zwangsbedingung entsteht in y der in Abb. 2.4 gezeigte zusätzliche Fehlerterm $\mathcal{O}(\frac{1}{h}\theta^2)$.

Für die Erweiterung der GGL–Formulierung und der stabilisierten Index-1-Formulierung auf DA–Systeme komplizierterer Struktur führen wir für (3.6) und (3.7) eine einheitliche Bezeichnung ein. Sei

$$x := \begin{pmatrix} q \\ v \\ \lambda \end{pmatrix} \,, \quad \tilde{F}(t, x(t), x'(t)) = \begin{pmatrix} q' - v \\ M(q)v' - f(q,v,\lambda) + G^T(q)\lambda \\ \gamma(q,v,\lambda) \end{pmatrix} \,. \tag{3.9}$$

Dann erfüllt die Lösung der MKS–Modellgleichungen (3.1) die Gleichungen

$$\tilde{F}(t, x(t), x'(t)) = 0 \quad \text{und} \quad \tilde{\gamma}(t, x(t)) = 0 \,, \tag{3.10}$$

wobei für die GGL–Formulierung

$$\gamma(q,v,\lambda) := r^{(\mathrm{G})}(q,v) \quad \text{und} \quad \tilde{\gamma}(t,x) := r^{(\mathrm{L})}(q) \tag{3.11}$$

gewählt wird und für die stabilisierte Index-1-Formulierung

$$\gamma(q,v,\lambda) := r^{(\mathrm{B})}(q,v,\lambda) \quad \text{und} \quad \tilde{\gamma}(t,x) := \begin{pmatrix} r^{(\mathrm{L})}(q) \\ r^{(\mathrm{G})}(q,v) \end{pmatrix} \,. \tag{3.12}$$

In (3.10) ist $x \in \mathbb{R}^{n_x}$ mit $n_x = 2n_q + n_\lambda$ und $\tilde{F} : \Omega_1 \to \mathbb{R}^{n_x}$, $\tilde{\gamma} : \Omega_2 \to \mathbb{R}^{n_\eta}$ mit geeigneten Mengen $\Omega_1 \subset [0,T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$ und $\Omega_2 \subset [0,T] \times \mathbb{R}^{n_x}$. Mit diesen Bezeichnungen können (3.6) und (3.7) zusammengefaßt werden zu

$$\begin{aligned} \tilde{F}(t, x(t), x'(t)) + \tilde{\gamma}_x^T(t, x(t))\eta &= 0 \\ \tilde{\gamma}(t, x(t)) &= 0 \end{aligned} \tag{3.13}$$

mit Hilfsvariablen $\eta : [0,T] \to \mathbb{R}^{n_\eta}$, die für die analytische Lösung von (3.1) identisch verschwinden.

Stabilisierte Integration mit dem Integrator ODASSL ([66])

Aus der Literatur sind verschiedene Ansätze zur numerischen Lösung von DA–Systemen bekannt, die ähnlich wie die GGL–Formulierung sowohl den Index des Systems reduzieren als auch garantieren, daß die ursprünglichen Zwangsbedingungen eingehalten werden (vgl. z. B. [84, Kapitel VII.2]). Ebenso wie für (ursprüngliche und versteckte) Zwangsbedingungen kann man dabei erreichen, daß die numerische Lösung auch andere Invarianten der analytischen Lösung — wenn es solche gibt — erfüllt. Je nach dem Anwendungsgebiet ergeben sich solche Invarianten z. B. aus der Masse- oder der Energieerhaltung (vgl. z. B. [71], [1], [143], [59]). Eine Minimalanforderung an all diese Lösungsansätze ist, daß sie auf eindeutig lösbare Anfangswertprobleme führen müssen, die zu dem ursprünglich gegebenen Anfangswertproblem analytisch äquivalent sind.

Von Führer ([66]) wurde mit dem Integrator ODASSL eine Modifikation des BDF–Codes DASSL ([129]) entwickelt, die es — zunächst unabhängig von der konkreten Anwendung — prinzipiell ermöglicht, eine numerische Lösung für ein Anfangswertproblem eines DA–Systems $\tilde{F}(t, x(t), x'(t)) = 0$ zu bestimmen, die gewissen vorgegebenen Gleichungen $\tilde{\gamma}(t, x) = 0$ genügt ($\tilde{\gamma} = 0$ sind z. B. Zwangsbedingungen oder Invarianten der analytischen Lösung). Während die analytische Lösung $x(t)$ wie in (3.10) sowohl $\tilde{F}(t, x, x') = 0$ als auch $\tilde{\gamma}(t, x) = 0$ erfüllen soll, führt die Diskretisierung in jedem Integrationsschritt auf überbestimmte nichtlineare Gleichungssysteme, die i. allg. keine klassische, sondern nur eine verallgemeinerte Lösung haben (ODASSL = DASSL für überbestimmte (**O**verdetermined) DA–Systeme). ODASSL berechnet eine (klassische) BDF–Diskretisierung des erweiterten DA–Systems (3.13) und kann deshalb als Verallgemeinerung der GGL–Formulierung angesehen werden ([68]). ODASSL arbeitet besonders effizient, weil in (3.13) die Hilfsvariablen η bei der Schrittweitensteuerung unberücksichtigt bleiben, außerdem wird statt des Terms $\tilde{\gamma}_x^T(t, x)\eta$ ein Term $\hat{G}^T\hat{\eta}$ verwendet mit einer Matrix $\hat{G} \approx \tilde{\gamma}_x(t, x)$, die man jeweils über mehrere Integrationsschritte konstant hält.

ODASSL wird (u. a. als Bestandteil des MKS–Simulationspakets SIMPACK, vgl. Abschnitt 4.4) erfolgreich zur Integration der GGL–Formulierung und der stabilisierten Index-1-Formulierung der MKS–Modellgleichungen (3.1) eingesetzt. So ist es naheliegend, ODASSL auch zur Integration von MKS–Modellgleichungen komplizierterer Struktur, die z. B. bei der Modellierung von MKS mit Kontaktbedingungen auftreten, zu verwenden (vgl. Abschnitt 4.1). Eine direkte Übertragung des Ansatzes (3.13) auf DA–Systeme beliebiger Struktur führt jedoch nicht zum Erfolg ([23]):

Beispiel 19 Gegeben sei das Anfangswertproblem $y_1(0) = y_2(0) = s(0) = z(0) = 0$ für das Index-2-System $F(x, x') = 0$ mit $x = (y_1, y_2, s, z)^T$ und

$$F(x, x') = (\, y_1' - z\,,\; y_2' - z\,,\; -s\,,\; y_1 + y_2 + s^2\,)^T\,.$$

Durch Differentiation der Zwangsbedingung $y_1 + y_2 + s^2 = 0$ erhalten wir wegen $y_1' = z$, $y_2' = z$ und $s' = 0$ das äquivalente Index-1-System $\tilde{F}(x, x') = 0$ mit

$$\tilde{F}(x, x') = (\, y_1' - z\,,\; y_2' - z\,,\; -s\,,\; 2z\,)^T\,,$$

dessen (identisch verschwindende) Lösung die ursprüngliche Zwangsbedingung $\tilde{\gamma}(x) = 0$ mit $\tilde{\gamma}(x) := y_1 + y_2 + s^2$ erfüllt. Mit diesen Funktionen $\tilde{F}$, $\tilde{\gamma}$ hat aber das erweiterte

DA–System (3.13) neben der ursprünglichen Lösung $y_1(t) \equiv y_2(t) \equiv s(t) \equiv z(t) \equiv 0$ eine zweite Lösung

$$y_1(t) = -t/2\,,\quad y_2(t) = -t/2\,,\quad s(t) = \sqrt{t}\,,\quad z(t) = 0\,,\quad \eta(t) = -1/2\,,$$

d. h., das DA–System (3.13) ist *nicht* analytisch äquivalent zu dem ursprünglichen DA–System $F(x, x') = 0$.

Bemerkung 16 Ähnlich wie in Beispiel 19 führt der in Abschnitt 4.1 beschriebene Zugang zur Modellierung von MKS mit Kontaktbedingungen auf Modellgleichungen der Form

$$\begin{aligned} M(q)q'' &= \tilde{f}(q, s, q', \lambda) - \tilde{G}^T(q, s)\lambda \\ 0 &= h(q, s) \\ 0 &= \tilde{g}(q, s) \end{aligned} \tag{3.14}$$

mit $\tilde{G}(q, s) := [\tilde{g}_q - \tilde{g}_s h_s^{-1} h_q](q, s)$, in denen neben den Zwangsbedingungen $\tilde{g}(q, s) = 0$ weitere algebraische Gleichungen auftreten. Diese Gleichungen $h(q, s) = 0$ bestimmen in Abhängigkeit von den Lagekoordinaten q eindeutig die Kontaktpunktkoordinaten s.

Lemma 6 *Ist die Jacobimatrix h_s in einer Umgebung einer Lösung von (3.14) regulär, so ist das Gleichungssystem $h(q, s) = 0$ lokal eindeutig nach s auflösbar und definiert eine Funktion $\sigma : \mathbb{R}^{n_q} \to \mathbb{R}^{n_s}$ mit $h(q, \sigma(q)) \equiv 0$. Mit dieser Funktion σ ist (3.14) äquivalent zu MKS–Modellgleichungen in der Standardform (3.1) mit*

$$f(q, v, \lambda) := \tilde{f}(q, \sigma(q), v, \lambda)\,,\quad g(q) := \tilde{g}(q, \sigma(q)) \quad \textit{und} \quad G(q) = \frac{\partial}{\partial q} g(q)\,.$$

Beweis Die Existenz von σ folgt aus dem Satz über die implizite Funktion. Wegen $h(q, \sigma(q)) = 0$ ergibt sich $0 = D_q h(q, \sigma(q)) = h_q + h_s \cdot \sigma_q$ und

$$g_q(q) = D_q \tilde{g}(q, \sigma(q)) = \tilde{g}_q + \tilde{g}_s \sigma_q = \tilde{g}_q - \tilde{g}_s h_s^{-1} h_q = \tilde{G}(q, \sigma(q))\,.$$

Ersetzt man in (3.14) jeweils s durch $\sigma(q)$, so ergibt sich also (3.1) mit $G(q) = g_q(q)$. ∎

Im weiteren betrachten wir MKS–Modellgleichungen (3.14), die auf dem in Lemma 6 angegebenen Weg in ein äquivalentes DA–System (3.1) transformiert werden können, das die Voraussetzungen aus Abschnitt 2.3 erfüllt. Die Transformation (3.14) → (3.1) erleichtert die analytischen Untersuchungen, numerisch ist es jedoch günstiger, die Index-1-Gleichungen $h(q, s) = 0$ nicht explizit aufzulösen, sondern $s = \sigma(q)$ als Teil der Korrektoriteration im Integrator berechnen zu lassen ([148]).

Durch Differentiation der Zwangsbedingung $\tilde{r}^{(\mathrm{L})}(q, s) := \tilde{g}(q, s) = 0$ in (3.14) ergeben sich wie in (3.2) und (3.3) die Zwangsbedingungen auf Ebene der **G**eschwindigkeits- und der **B**eschleunigungskoordinaten, wobei die zeitlichen Ableitungen $s'(t)$ und $s''(t)$ durch implizite Differentiation aus $h(q(t), s(t)) = 0$ bestimmt werden (z. B. $h_q q' + h_s s' = 0$):

$$0 = \frac{d}{dt}\tilde{g}(q, s) = \tilde{g}_q \cdot q'(t) + \tilde{g}_s \cdot s'(t) = [\tilde{g}_q - \tilde{g}_s h_s^{-1} h_q](q, s) \cdot q'(t) = \tilde{G}(q, s)v =: \tilde{r}^{(\mathrm{G})}(q, s, v),$$

$$0 = \frac{d^2}{dt^2}\tilde{g}(q,s) =: \tilde{r}^{(\mathrm{B})}(q,s,v,\lambda)\ .$$

Analog zu (3.9) fassen wir die Funktionen in (3.14) zusammen zu

$$\Phi(t,\xi,s,\xi') = \begin{pmatrix} q'-v \\ M(q)v' - \tilde{f}(q,s,v,\lambda) + \tilde{G}^T(q,s)\lambda \\ \gamma(q,s,v,\lambda) \end{pmatrix},\ \tilde{F}(t,x,x') = \begin{pmatrix} \Phi(t,\xi,s,\xi') \\ h(q,s) \end{pmatrix} \quad (3.15)$$

mit $x := (\xi^T, s^T)^T$ und $\xi := (q^T, v^T, \lambda^T)^T$. Es gilt $x \in \mathbb{R}^{n_x}$ mit $n_x = 2n_q + n_s + n_\lambda$ und wie in (3.9) bildet $\tilde{F}$ eine Teilmenge des $\mathbb{R}^{2n_x+1}$ nach $\mathbb{R}^{n_x}$ ab.

Mit diesen Bezeichnungen werden als Verallgemeinerung der GGL–Formulierung und der stabilisierten Index-1-Formulierung auf Modellgleichungen der Form (3.14) die beiden DA–Systeme (3.13) und (3.18) betrachtet. Dabei nennen wir die erweiterten DA–Systeme (3.13) und (3.18) wie in (3.11) eine *GGL–Formulierung von (3.14)*, falls

$$\gamma(t,x) := \tilde{r}^{(\mathrm{G})}(q,s,v) \quad \text{und} \quad \tilde{\gamma}(t,x) := \tilde{r}^{(\mathrm{L})}(q,s) \qquad (3.16)$$

und eine *stabilisierte Index-1-Formulierung von (3.14)*, falls

$$\gamma(t,x) := \tilde{r}^{(\mathrm{B})}(q,s,v,\lambda) \quad \text{und} \quad \tilde{\gamma}(t,x) := \begin{pmatrix} \tilde{r}^{(\mathrm{L})}(q,s) \\ \tilde{r}^{(\mathrm{G})}(q,s,v) \end{pmatrix}\ . \qquad (3.17)$$

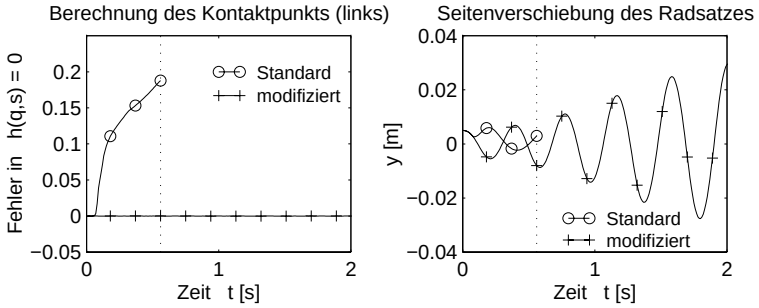


Abbildung 3.2: Vergleich verschiedener Korrektoriterationen in ODASSL bei Anwendung auf Modellgleichungen (3.14).

Die möglichen praktischen Auswirkungen des Effekts, den wir oben an dem einfachen Beispiel 19 theoretisch untersucht hatten, zeigt Abb. 3.2. Hier wurden die Bewegungsgleichungen eines starren Radsatzes in stabilisierter Index-1-Formulierung integriert (das Fahrmanöver wird in Beispiel 27 beschrieben; die analytische Lösung der Bewegungsgleichungen strebt gegen einen Grenzzyklus, vgl. Abb. 4.4). Die Funktion $h(q,s)$, die die Lage des Kontaktpunkts definiert, ist Bestandteil von $\tilde{F}$. Für die Standardimplementierung (3.13) von ODASSL (in Abb. 3.2 mit „o“ markiert) erfüllt die BDF–Lösung deshalb *nicht* die Gleichungen $h(q,s) = 0$, sondern nur $h(q_n,s_n) + \tilde{\gamma}_s^T(x_n)\eta_n = 0$ (Abb. 3.2 links,

„o“). Der große Fehler in s verfälscht die Simulationsergebnisse vollkommen, bei $t \approx 0.56\,\mathrm{s}$ bricht die Integration ab.

Wir konstruieren deshalb eine zu (3.14) analytisch äquivalente Formulierung, für die die numerische Lösung nicht nur die Zwangsbedingungen $\tilde{g}(q,s) = 0$, sondern auch die Index-1-Gleichungen $h(q,s) = 0$ im Rahmen der vorgegebenen Genauigkeit einhält. Hierzu wird statt (3.13) eine BDF–Diskretisierung von

$$\begin{aligned} \Phi(t,\xi(t),s(t),\xi'(t)) + ([\tilde{\gamma}_\xi - \tilde{\gamma}_s h_s^{-1} h_\xi](t,\xi(t),s(t)))^T\eta &= 0 \\ h(\xi(t),s(t)) &= 0 \\ \tilde{\gamma}(t,\xi(t),s(t)) &= 0 \end{aligned} \qquad (3.18)$$

berechnet, d. h., der Term

$$\tilde{\gamma}_x^T(t,x)\eta = \begin{pmatrix} \tilde{\gamma}_\xi^T(t,\xi,s) \\ \tilde{\gamma}_s^T(t,\xi,s) \end{pmatrix}\eta$$

in (3.13) wird hier durch

$$\begin{pmatrix} ([\tilde{\gamma}_\xi - \tilde{\gamma}_s h_s^{-1} h_\xi](t,\xi,s))^T \\ 0 \end{pmatrix}\eta$$

ersetzt. Man beachte, daß sich (3.13) und (3.18) nur für MKS–Modellgleichungen der erweiterten Form (3.14) unterscheiden; für Gleichungen in Standardform (3.1) fallen beide Ansätze zusammen.

Satz 10 *Ist in dem erweiterten DA–System (3.18) die Funktion Φ durch (3.15) bestimmt und werden die Funktionen γ und $\tilde{\gamma}$ entweder nach (3.16) oder nach (3.17) gewählt, so bildet (3.18) ein Index-2-System, das zu den Modellgleichungen (3.14) analytisch äquivalent ist.*

Beweis Wie in Lemma 6 kann in (3.18) das Gleichungssystem $h(\xi,s) = h(q,s) = 0$ nach s aufgelöst werden: $s = \sigma(q)$. Setzt man in (3.18) für s diese Funktion $\sigma(q)$ ein und ersetzt anschließend Φ, γ und $\tilde{\gamma}$ durch die Funktionen aus (3.15) und (3.16), so ergibt sich die klassische GGL–Formulierung (3.6) für MKS–Modellgleichungen der Standardform (3.1) mit $f(q,v,\lambda) = \tilde{f}(q,\sigma(q),v,\lambda)$ und $g(q) = \tilde{g}(q,\sigma(q)) = 0$. Nach Lemma 6 ist dieses DA–System (3.1) äquivalent zu den Modellgleichungen (3.14). Damit folgt die erste Behauptung aus Satz 9. Analog kann man nachweisen, daß (3.18) auch für die stabilisierte Index-1-Formulierung, d. h. für die in (3.17) definierten Funktionen γ und $\tilde{\gamma}$ äquivalent zu (3.14) ist. ∎

Verwendet man den modifizierten Ansatz (3.18), so führen die GGL–Formulierung und die stabilisierte Index-1-Formulierung also auch für MKS–Modellgleichungen der Form (3.14) auf DA–Systeme vom Index 2, die zu dem gegebenen Index-3-System (3.14) analytisch äquivalent sind. Aus der Konvergenztheorie für Mehrschrittverfahren folgt, daß die BDF bei Anwendung auf diese Index-2-Systeme unter geeigneten Voraussetzungen gegen die analytische Lösung von (3.14) konvergieren ([74], [84, Satz VII.3.5], vgl. auch den nachfolgenden Abschnitt 3.2.2).

Die vorgeschlagene Modifikation (3.18) erfordert innerhalb des Integrators ODASSL ausschließlich Änderungen der Korrektoriteration; für die Details des Algorithmus und verschiedene Simulationsergebnisse verweisen wir auf [23]. Mit dieser Modifikation des Korrektors steht der volle Leistungsumfang von ODASSL auch für Modellgleichungen der Form (3.14) zur Verfügung. In Abb. 3.2 markiert „+“ die mit dem nach (3.18) modifizierten Integrator berechneten Simulationsergebnisse, das rechte Diagramm zeigt eine der Lagekoordinaten q (vgl. Abb. 4.4). Die numerische Lösung ist korrekt und erfüllt im Rahmen der vorgegebenen Genauigkeit die Gleichungen $h(q,s)=0$ (linkes Diagramm).

Zusammenfassung

Analytische Transformationen zur Indexreduktion sind das entscheidende Hilfsmittel bei der numerischen Lösung von DA–Systemen von höherem und hohem Index. Unter Berücksichtigung der Struktur des DA–Systems können solche indexreduzierten Systeme numerisch effizient gelöst werden. Speziell für MKS–Modellgleichungen ist die Kopplung von Indexreduktion und Projektion vorteilhaft, daneben erweist sich auch die GGL–Formulierung der Modellgleichungen als besonders geeignet. Sowohl die Projektionsschritte als auch die GGL–Formulierung lassen sich von den oft ausschließlich betrachteten Systemen der (Standard-)Form (3.1) auf DA–Systeme allgemeinerer Struktur (z. B. (3.14)) übertragen. Die Anpassung vorhandener Integrationssoftware an Modellgleichungen mit komplizierter Struktur erfordert dabei besondere Sorgfalt.

3.2 Konvergenz von numerischen Verfahren für differentiell-algebraische Systeme vom Index 2 in Hessenbergform

Bei der Untersuchung numerischer Verfahren für DA–Systeme wurde zunächst vorwiegend die direkte Übertragung von numerischen Verfahren für gewöhnliche Differentialgleichungen auf DA–Systeme betrachtet (beginnend mit den Arbeiten von Gear [70] für BDF und Petzold [131] für Runge–Kutta–Verfahren). Inzwischen gibt es jedoch eine Vielzahl von Diskretisierungsverfahren, die der speziellen Struktur von DA–Systemen angepaßt sind. Einen besonderen Schwerpunkt bilden hierbei Einschrittverfahren für Index-1-Systeme in Hessenbergform (vgl. hierzu z. B. den aktuellen Überblick in [138, Abschnitt 3.2] und die detaillierte Untersuchung von linear-impliziten Verfahren in [9]).

Für Index-1-Systeme in Hessenbergform sind die Zwangsbedingungen direkt nach den algebraischen Variablen auflösbar, bei der Konstruktion numerischer Verfahren versucht man daher im wesentlichen, die Auflösung der Zwangsbedingungen möglichst effizient mit der zeitlichen Integration des differentiellen Teils zu koppeln. Dagegen ergeben sich für Index-2-Systeme qualitativ neue Probleme, weil hier die Lösung des DA–Systems nicht mehr stetig von kleinen Störungen (wie sie bei Rechnungen in Gleitpunktarithmetik unvermeidbar sind) abhängt. Es gibt deshalb in der Literatur verschiedentlich Ansätze, Index-2-Systeme für die numerische Lösung durch Regularisierung in Index-1-Systeme zu überführen (z. B. [30], [60], [85], [121]). Wegen der Schwierigkeiten, die hierbei entstehenden steifen Index-1-Systeme numerisch zu lösen, ist es jedoch günstiger, Index-2-Systeme

direkt zu lösen und den in Abschnitt 2.2 untersuchten zusätzlichen Fehlerterm (der für $h \to 0$ unbeschränkt wächst) bei der Implementierung der Verfahren zu berücksichtigen.

Inzwischen gibt es sowohl für allgemeine Index-2-Systeme in Hessenbergform als auch speziell angepaßt an nicht-steife Systeme (vgl. Abschnitt 3.3) effektive numerische Verfahren einschließlich der zugehörigen Integrationssoftware. Die Entwicklung von Verfahren konzentriert sich dabei auf Index-2-Systeme, weil sie in verschiedenen Anwendungsgebieten direkt als Modellgleichungen oder nach der in Abschnitt 3.1 beschriebenen Indexreduktion auftreten. Die direkte Diskretisierung von DA–Systemen vom Index > 2 ist zwar (zumindest für Index-3-Systeme in Hessenbergform) prinzipiell möglich (vgl. z. B. [104], [81, Kapitel 6], [123], [124], [93], [94], [162]), hat aber wegen der enormen Probleme bei der Implementierung (vgl. Beispiel 6) in praxi kaum Bedeutung.

Bei der Untersuchung numerischer Verfahren betrachten wir Index-2-Systeme in Hessenbergform

$$\begin{aligned} y'(t) &= f(y(t),z(t))\,, \quad (t\in[0,T])\,, \\ 0 &= g(y(t))\,, \quad y(0)=y_0\,, \quad z(0)=z_0\,, \end{aligned} \tag{3.19}$$

die — wie in Abschnitt 2.2 — eine hinreichend oft stetig differenzierbare Lösung $y : [0,T] \to \mathbb{R}^{n_y}$, $z : [0,T] \to \mathbb{R}^{n_z}$ haben; in einer Umgebung dieser Lösung sollen auch f und g hinreichend oft stetig differenzierbar sein. Als Vereinfachung gegenüber Bedingung (2.8) in Abschnitt 2.2 fordern wir hier, daß die *Index-2-Bedingung*

$$[g_y f_z](\eta,\zeta) \quad \text{regulär} \tag{3.20}$$

für *alle* Vektoren (η,ζ) in dieser Umgebung erfüllt ist.

Einen systematischen Weg, ausgehend von Verfahren für gewöhnliche Differentialgleichungen Diskretisierungsverfahren für DA–Systeme zu konstruieren, beschreibt der *direkte Zugang* ([81], vgl. auch [70], [131]): Man ersetzt in (3.19) die algebraische Gleichung $0=g(y(t))$ durch $\varepsilon z'(t)=g(y(t))$ mit $0<\varepsilon\ll 1$ und betrachtet in der Verfahrensvorschrift eines für die Lösung solcher singulär gestörter Systeme gewöhnlicher Differentialgleichungen geeigneten Verfahrens den Grenzübergang $\varepsilon \to 0$. Die auf diese Art konstruierten Verfahren sind nicht an die spezielle semi-explizite Struktur von (3.19) gebunden (vgl. z. B. [70], [131], [81] und insbesondere auch die Arbeiten von März und anderen [76], [114], [86], [155], [15]). Wegen des engen Zusammenhangs von singulär gestörten Systemen gewöhnlicher Differentialgleichungen zu DA–Systemen erklärt die theoretische Untersuchung der mit dem direkten Zugang konstruierten Verfahren außerdem z. T. das Verhalten von Diskretisierungsverfahren für Probleme mit singulären Störungen ([79], [80], [107]).

Die in der vorliegenden Arbeit näher untersuchten Verfahren folgen dagegen dem *indirekten Zugang*, d. h. ähnlich wie bei partitionierten Verfahren für Systeme gewöhnlicher Differentialgleichungen (vgl. z. B. [160]) wird bereits während der Konstruktion der Verfahren die spezielle Struktur des DA–Systems (3.19) ausgenutzt. Dies kann schon für steife DA–Systeme nützlich sein (z. B. die Kopplung von impliziten Runge–Kutta–Verfahren vom Gauß–Typ mit der Projektion der numerischen Lösung auf die Mannigfaltigkeit $\{\eta : g(\eta)=0\}$ [31]), vor allem wird so aber die effiziente Lösung nicht-steifer DA–Systeme ermöglicht. In den nachfolgenden Abschnitten 3.2.1 und 3.2.2 beweisen wir die Konvergenz von halb-expliziten Runge–Kutta–Verfahren (HERK) bzw. von partitionierten linearen Mehrschrittverfahren (PLMSV). In beiden Fällen sind (nichtlineare) Glei-

chungssysteme nicht — wie für den direkten Zugang — zur Berechnung aller Komponenten y und z, sondern nur zur Berechnung der algebraischen Komponenten z zu lösen.

Für Index-2-Systeme gibt es in der Literatur eine Reihe verschiedener Ansätze zum Beweis der Konvergenz numerischer Verfahren (z. B. [40], [74], [81], [104], [114]). Wir werden in diesem Abschnitt weitgehend der einheitlichen Darstellung der Konvergenzbeweise für Ein- und Mehrschrittverfahren in [84, Kapitel VII] folgen, die auf die Arbeiten von Hairer, Lubich und anderen zurückgeht. Der Konvergenzbeweis gliedert sich hier in die Untersuchung des lokalen Fehlers (in Abschnitt 3.3) und die Untersuchung der Fehlerfortpflanzung während der Integration (im vorliegenden Abschnitt), letztere lehnt sich an die bekannten Mechanismen der Fehlerfortpflanzung in der analytischen Lösung an (vgl. Abschnitt 2.2.1).

Wie in Abschnitt 2.2.2 für das implizite Eulerverfahren gezeigt (und in der Theorie der numerischen Verfahren für gewöhnliche Differentialgleichungen üblich), ergibt sich prinzipiell die Konvergenz der Verfahren als Spezialfall von Schranken für den Einfluß beliebiger kleiner Fehler auf die numerische Lösung. Zur Vereinfachung der Darstellung konzentrieren wir uns in diesem Abschnitt jedoch ausschließlich auf Konvergenzuntersuchungen (d. h. auf die Fortpflanzung der Diskretisierungsfehler) und verweisen für den Nachweis allgemeinerer Fehlerschranken, die auch Störungen in den algebraischen Gleichungen berücksichtigen, auf Abschnitt 2.2.2 und auf [10], [15].

3.2.1 Halb-explizite Runge–Kutta–Verfahren für differentiell–algebraische Systeme vom Index 2 in Hessenbergform: Definition und Konvergenz

Unter den Einschrittverfahren für gewöhnliche Differentialgleichungen sind insbesondere explizite Runge–Kutta–Verfahren zur Integration nicht-steifer Anfangswertprobleme geeignet. Als Verallgemeinerung auf Index-2-Systeme (3.19) wurden 1989 von Hairer et al. ([81]) *halb-explizite Runge–Kutta–Verfahren* (HERK–Verfahren) eingeführt, die wir hier in der (gegenüber der Darstellung in [81] verallgemeinerten) Form

$$\left.\begin{aligned} Y_{ni} &= y_n + h\sum_{j=1}^{i-1} a_{ij} f(Y_{nj}, Z_{nj}) \\ \eta_{ni} &= y_n + h\sum_{j=1}^{i} \gamma_{ij} f(Y_{nj}, Z_{nj}) \ , \quad g(\eta_{ni}) = 0 \end{aligned}\right\} \ , \ (i = 1, \ldots, \hat{s}) \ ,$$
$$y_{n+1} = y_n + h\sum_{j=1}^{s} b_j f(Y_{nj}, Z_{nj}) \ , \quad z_{n+1} = \sum_{j=1}^{\hat{s}} d_j Z_{nj} \qquad (3.21)$$

betrachten. Neben den Koeffizienten b_j, a_{ij}, $(i, j = 1, \ldots, s)$ des zugrundeliegenden s-stufigen expliziten Runge–Kutta–Verfahrens treten in (3.21) zusätzliche Parameter γ_{ij}, d_j auf. Hinsichtlich der Stufenzahl unterscheiden wir die Fälle $\hat{s} = s$ (also gleiche Stufenzahl im expliziten Runge–Kutta–Verfahren und im HERK–Verfahren) und $\hat{s} = s + 1$ (eine zusätzliche Stufe wird angefügt, vgl. Bemerkung 25). Insbesondere sei $b_j = a_{s+1,j}$, $(j = 1, \ldots, s)$ und außerdem $b_j := 0$, $(j > s)$, $a_{ij} := 0$, $(j \geq i)$ sowie $\gamma_{ij} := 0$, $(j > i)$.

Die Parameter a_{ij} bestimmen $c_i := \sum_j a_{ij}$, wobei hier wie im folgenden stets von 1 bis $\hat{s}$ zu summieren ist, wenn nicht explizit andere Summationsgrenzen vorgegeben werden. Es ist $c_1 = 0$, außerdem soll gelten $c_2 \neq 0$, d. h. $a_{21} \neq 0$. Für die Koeffizienten d_j der Linearkombination $z_{n+1} = \sum_j d_j Z_{nj}$ wird vorausgesetzt

$$\sum_{j=1}^{\hat{s}} d_j = 1 \ .$$

Bemerkung 17 a) In der i-ten Stufe des Verfahrens liegen neben y_n bereits die Stufenvektoren $Y_{n1}, \ldots, Y_{ni}, Z_{n1}, \ldots, Z_{n,i-1}$ vor. Hieraus wird (falls $\gamma_{ii} \neq 0$) zunächst Z_{ni} als Lösung ζ des nichtlinearen Gleichungssystems

$$0 = \Phi(\zeta) := g\Big(y_n + h\sum_{j=1}^{i-1} \gamma_{ij} f(Y_{nj}, Z_{nj}) + h\gamma_{ii} f(Y_{ni}, \zeta)\Big) \qquad (3.22)$$

bestimmt, so daß schließlich $f(Y_{ni}, Z_{ni})$ und damit (falls $i < \hat{s}$) auch $Y_{n,i+1}$ berechnet werden kann. In jeder Stufe von (3.21) mit $\gamma_{ii} \neq 0$ ist ein n_z-dimensionales Gleichungssystem zu lösen.

b) Die in [81] eingeführten HERK–Verfahren wurden (unabhängig voneinander) von Higueras Sanz ([90]) und von Brasey und Hairer ([36], [37], [38]) im Detail untersucht. Sie sind in (3.21) durch $\hat{s} = s$, $\gamma_{ij} = a_{i+1,j}$, $(i = 1, \ldots, s-1,\ j = 1, \ldots, i)$, $\gamma_{sj} = b_j$, $(j = 1, \ldots, s)$ charakterisiert. Für die Konstruktion von Verfahren höherer Ordnung ist es günstiger, statt dieser Parameter HERK–Verfahren mit einer *expliziten* ersten Stufe zu betrachten ([118], [21], vgl. Bemerkung 21). Im folgenden beschränken wir uns daher auf Verfahren mit

$$\gamma_{1j} = 0 \ , \quad (j = 1, \ldots, \hat{s})$$

und setzen (da hier in der ersten Stufe kein Gleichungssystem zu lösen ist)

$$Z_{n1} := z_n \ . \qquad (3.23)$$

c) Um die Lösbarkeit der Gleichungssysteme (3.22) zu garantieren, wird vorausgesetzt $\gamma_{ii} \neq 0$, $(i = 2, \ldots, \hat{s})$. Faßt man nun die Koeffizienten γ_{ij} zu einer (unteren Dreiecks-) Matrix zusammen und ersetzt γ_{11} durch die Zahl 1, so ergibt sich eine reguläre Matrix. Für die Elemente w_{ij} der inversen Matrix

$$W = \begin{pmatrix} 1 & & & \\ \gamma_{21} & \gamma_{22} & & \\ \vdots & \vdots & \ddots & \\ \gamma_{\hat{s}1} & \gamma_{\hat{s}2} & \cdots & \gamma_{\hat{s}\hat{s}} \end{pmatrix}^{-1}$$

gilt $w_{i1}\delta_{1k} + \sum_{j=2}^{\hat{s}} w_{ij}\gamma_{jk} = \delta_{ik}$, $(i, k = 1, \ldots, \hat{s})$ mit dem Kroneckerschen Delta δ_{ik}.

Lemma 7 (vgl. [38]) *Gegeben seien Vektoren* (y_n, z_n) *, die für ein* $\mu \in (0, 1]$ *die Bedingungen*

$$\|y_n - y(t_n)\| + \|z_n - z(t_n)\| = \mathcal{O}(h^\mu) \ , \quad \|g(y_n)\| = \mathcal{O}(h^{1+\mu})$$

erfüllen. Unter der Voraussetzung $\gamma_{11} = 0$, $\gamma_{ii} \neq 0$, $(i = 2, \ldots, \hat{s})$ *ist das HERK–Verfahren (3.21) mit (3.23) wohldefiniert, d. h., für hinreichend kleine Schrittweiten* h *sind die Gleichungssysteme (3.22) lokal eindeutig lösbar.*

Beweis Mittels vollständiger Induktion wird für $i = 2, \ldots, \hat{s}$ bewiesen, daß in der i-ten Stufe das Gleichungssystem (3.22) eine Lösung $\zeta = z_n + \mathcal{O}(h^\mu)$ hat, so daß $Y_{n,i+1} = y_n + \mathcal{O}(h)$ und $Z_{ni} = z_n + \mathcal{O}(h^\mu)$ gilt.

Ist die Induktionsvoraussetzung erfüllt, so gibt es Konstanten C_0, C_1, C_2, die von y_n, z_n, h und μ unabhängig sind und für die gilt

$$\|(\Phi_\zeta(z_n))^{-1}\| \leq \frac{C_1}{h} \ , \quad \|\Phi_{\zeta\zeta}(z_n)\| \leq C_2 h \ , \quad \|(\Phi_\zeta(z_n))^{-1}\Phi(z_n)\| \leq C_0 h^\mu \ ,$$

denn $g(y_n) = \mathcal{O}(h^{1+\mu})$ und $[g_y f](y_n, z_n) = [g_y f](y(t_n), z(t_n)) + \mathcal{O}(h^\mu) = \mathcal{O}(h^\mu)$. Deshalb folgt die lokal eindeutige Auflösbarkeit von (3.22) (und damit auch die Behauptung des Lemmas) aus dem Satz von Newton–Kantorovich ([122]). ∎

Wegen der expliziten Stufe pflanzen sich während der Integration Fehler in den hier betrachteten HERK–Verfahren nicht nur (wie bei impliziten Runge–Kutta–Verfahren und bei den HERK–Verfahren aus [81]) in den differentiellen Komponenten y, sondern auch in den algebraischen Komponenten z fort. Zum Konvergenzbeweis wird zunächst (analog zu [81]) die Fehlerfortpflanzung in einem einzelnen Integrationsschritt untersucht. Hierzu seien Vektoren $\hat{Y}_{ni}$, $\hat{Z}_{ni}$ gegeben mit $\hat{Z}_{n1} := \hat{z}_n$ und

$$\left.\begin{aligned} \hat{Y}_{ni} &= \hat{y}_n + h\sum_{j=1}^{i-1} a_{ij} f(\hat{Y}_{nj}, \hat{Z}_{nj}) + h\delta_i \\ \theta_i &= g\Big(\hat{y}_n + h\sum_{j=1}^{i} \gamma_{ij} f(\hat{Y}_{nj}, \hat{Z}_{nj})\Big) \end{aligned}\right\} \ , \quad (i = 1, \ldots, \hat{s}) \ . \tag{3.24}$$

Satz 11 (vgl. [81, Satz 4.2]) *Gegeben sei ein HERK–Verfahren (3.21) mit (3.23) und* $\gamma_{11} = 0$, $\gamma_{ii} \neq 0$, $(i = 2, \ldots, \hat{s})$. *Dann gibt es für beliebige Vektoren* (y_n, z_n), $(\hat{y}_n, \hat{z}_n)$, *die*

$$\|y_n - y(t_n)\| + \|z_n - z(t_n)\| = \mathcal{O}(h^\mu) \ , \quad \|\hat{y}_n - y(t_n)\| + \|\hat{z}_n - z(t_n)\| = \mathcal{O}(h^\mu)$$

und

$$\|g(y_n)\| + \|g(\hat{y}_n)\| = \mathcal{O}(h^{1+\mu}) \ , \quad \|\delta\| := \max_i \|\delta_i\| = \mathcal{O}(h^\mu) \ , \quad \|\theta\| := \max_i \|\theta_i\| = \mathcal{O}(h^{1+\mu})$$

mit einem $\mu \in (0, 1]$ *erfüllen, Vektoren* (Y_{ni}, Z_{ni}), $(\hat{Y}_{ni}, \hat{Z}_{ni})$, $(i = 1, \ldots, \hat{s})$ *mit (3.21) bzw. (3.24) und*

$$\|Y_{ni} - y(t_n)\| + \|Z_{ni} - z(t_n)\| = \mathcal{O}(h^\mu) \ , \quad \|\hat{Y}_{ni} - y(t_n)\| + \|\hat{Z}_{ni} - z(t_n)\| = \mathcal{O}(h^\mu) \ ,$$

sofern die Schrittweite $h > 0$ *hinreichend klein ist.*

Es gibt eine von h, δ *und* θ *unabhängige Konstante* C, *so daß für diese Vektoren gilt*

$$\|\hat{Y}_{ni} - Y_{ni}\| \leq C\big(\|\hat{y}_n - y_n\| + h\|\hat{z}_n - z_n\| + h\|\delta\| + \|\theta\|\big) \ , \tag{3.25}$$

$$\begin{aligned} \|\hat{Z}_{ni} - Z_{ni} - w_{i1}(\hat{z}_n - z_n)\| \leq\ & C\Big(\frac{1}{h}\|g_y(y(t_n))(\hat{y}_n - y_n)\| + h^{\mu-1}\|\hat{y}_n - y_n\| + \\ & + h^\mu\|\hat{z}_n - z_n\| + \|\delta\| + \frac{1}{h}\|\theta\|\Big) \ . \end{aligned} \tag{3.26}$$

Beweis Die Existenz der Vektoren (Y_{ni}, Z_{ni}), $(\hat{Y}_{ni}, \hat{Z}_{ni})$ folgt wie in Lemma 7. Für den Beweis von (3.25), (3.26) werden die Bezeichnungen $g_y^n := g_y(y(t_n))$, $f_z^n := f_z(y(t_n), z(t_n))$,

$$Y := (\, Y_{n1}^T, \ldots, Y_{n\hat{s}}^T\,)^T \ , \quad Z := (\, Z_{n1}^T, \ldots, Z_{n\hat{s}}^T\,)^T \ , \quad \ldots \ ,$$

die Hilfsvektoren $\mathbf{1} := (1, \ldots, 1)^T \in \mathbb{R}^{\hat{s}}$ und $\boldsymbol{e_1} := (1, 0, \ldots, 0)^T \in \mathbb{R}^{\hat{s}}$ sowie die Kroneckerprodukt–Schreibweise verwendet. Dann folgt aus (3.21) und (3.24)

$$\hat{Y} - Y = \mathbf{1} \otimes (\hat{y}_n - y_n) + \mathcal{O}(h)\|\hat{Y} - Y\| + \mathcal{O}(h)\|\hat{Z} - Z\| + \mathcal{O}(h\|\delta\|) \ , \tag{3.27}$$

weil f Lipschitz-stetig ist.
Wendet man $\Psi(1) - \Psi(0) = \int_0^1 \Psi'(\vartheta)\, d\vartheta$ auf $\Psi(\vartheta) := g(\eta_{ni} + \vartheta(\hat{\eta}_{ni} - \eta_{ni}))$ mit $\hat{\eta}_{ni} := \hat{y}_n + h\sum_j \gamma_{ij} f(\hat{Y}_{nj}, \hat{Z}_{nj})$ an, so erhält man aus der trivialen Identität

$$h\,[g_y^n f_z^n]\,(\hat{Z}_{n1} - Z_{n1} - (\hat{z}_n - z_n)) = 0$$

und aus dem algebraischen Teil

$$g(\hat{y}_n + h\sum\nolimits_j \gamma_{ij} f(\hat{Y}_{nj}, \hat{Z}_{nj})) - g(y_n + h\sum\nolimits_j \gamma_{ij} f(Y_{nj}, Z_{nj})) = \theta_i \ , \quad (i = 2, \ldots, \hat{s})$$

von (3.21) und (3.24) den Ausdruck

$$\begin{aligned} &(\mathbf{1} - \boldsymbol{e_1}) \otimes (g_y^n \cdot (\hat{y}_n - y_n)) + \mathcal{O}(h^\mu)\|\hat{y}_n - y_n\| + \mathcal{O}(h)\|\hat{Y} - Y\| + \\ &+ h(W^{-1} \otimes I)((I \otimes [g_y^n f_z^n]) + \mathcal{O}(h^\mu))(\hat{Z} - Z) - h(I \otimes [g_y^n f_z^n])(\boldsymbol{e_1} \otimes (\hat{z}_n - z_n)) = \mathcal{O}(\|\theta\|) \end{aligned}$$

und damit

$$\begin{aligned} \hat{Z} - Z - W\boldsymbol{e_1} \otimes (\hat{z}_n - z_n) = \mathcal{O}(\frac{1}{h})(\|g_y^n \cdot (\hat{y}_n - y_n)\| + h^\mu\|\hat{y}_n - y_n\|) + \\ + \mathcal{O}(1)\|\hat{Y} - Y\| + \mathcal{O}(h^\mu)\|\hat{z}_n - z_n\| + \mathcal{O}(\frac{1}{h}\|\theta\|) \ . \end{aligned} \tag{3.28}$$

Hierbei wurde ausgenutzt, daß W^{-1} mit Ausnahme der ersten Zeile die Koeffizienten γ_{ij} enthält (Bemerkung 17c), daß $W \otimes I$ und $I \otimes [g_y^n f_z^n]$ vertauschbar sind und daß $[g_y^n f_z^n]$ wegen der Index-2-Bedingung (3.20) regulär ist. Setzt man (3.28) in (3.27) ein, so folgt

$$\hat{Y} - Y = \mathcal{O}(1)\|\hat{y}_n - y_n\| + \mathcal{O}(h)\|\hat{Y} - Y\| + \mathcal{O}(h)\|\hat{z}_n - z_n\| + \mathcal{O}(h\|\delta\|) + \mathcal{O}(\|\theta\|)$$

und damit Behauptung (3.25) sowie mit (3.28) auch Behauptung (3.26). ∎

Für den Konvergenzbeweis für HERK–Verfahren gibt Satz 11 an, wie sich die Diskretisierungsfehler in einem einzelnen Integrationsschritt fortpflanzen.

Definition 6 In (3.24) sei $\delta_i = 0$, $\theta_i = 0$, $(i = 1, \ldots, \hat{s})$ und $\hat{y}_n$, $\hat{z}_n$ durch die analytische Lösung von (3.19) bei t_n gegeben: $\hat{y}_n = y(t_n)$, $\hat{z}_n = z(t_n)$.
Mit $\hat{y}_{n+1} := \hat{y}_n + h\sum_j b_j f(\hat{Y}_{nj}, \hat{Z}_{nj})$ und $\hat{z}_{n+1} := \sum_j d_j \hat{Z}_{nj}$ heißen

$$\delta y_h(t_n) := \hat{y}_{n+1} - y(t_n + h) \quad \text{und} \quad \delta z_h(t_n) := \hat{z}_{n+1} - z(t_n + h)$$

lokale Diskretisierungsfehler des HERK–Verfahrens (3.21) im differentiellen bzw. im algebraischen Teil.

Satz 12 *Gegeben sei ein HERK–Verfahren (3.21) mit (3.23) und*

$$\gamma_{11} = 0\ ,\quad \gamma_{ii} \neq 0\ ,\quad (i = 2, \ldots, \hat{s}) \quad \textit{und} \quad \gamma_{sj} = b_j\ ,\ (j = 1, \ldots, s)\ , \tag{3.29}$$

das die Kontraktivitätsbedingung

$$\kappa := |\sum_{j=1}^{\hat{s}} d_j w_{j1}| < 1 \tag{3.30}$$

erfüllt und lokale Diskretisierungsfehler der Ordnung

$$\delta y_h(t_n) = \mathcal{O}(h^{q_y})\ ,\quad P(t_n)\delta y_h(t_n) = \mathcal{O}(h^{q_y+1})\ ,\quad \delta z_h(t_n) = \mathcal{O}(h^{q_z+1}) \tag{3.31}$$

hat (vgl. die Definition (2.12) des Projektors $P(t)$). Ist $q := \min(q_y, q_z + 2) \geq 2$, so konvergiert das HERK–Verfahren mit der Ordnung q in y und mit der Ordnung $q-1$ in z, d. h., für hinreichend kleine Schrittweiten $h > 0$ gilt

$$\|y_m - y(t_m)\| + h\|z_m - z(t_m)\| = \mathcal{O}(h^q)\ ,\quad (t_m = mh\ ,\ t_m \in [0, T])\ .$$

Beweis Zum Beweis von Satz 12 wird das nachfolgende Lemma 8 erst mit $\mu = 1/2$ und anschließend mit $\mu = 1$ angewendet.
a) Sei also zunächst ein $C_1^* > 0$ so gewählt, daß für $n = 0$ die Voraussetzung (3.33) von Lemma 8 mit $\mu = 1/2$ und $C_1 = C_1^*$ erfüllt ist. Zu $\mu = 1/2$ und $C_1 = C_1^*$ bezeichne $C_2^* := C_2(C_1^*, \frac{1}{2})$ die Konstante C_2 aus Lemma 8. Dann bestimmt man $h_0 > 0$ so, daß für alle Schrittweiten $h \in (0, h_0]$ gilt: $C_2^* h^{q-\frac{1}{2}} \leq \frac{1}{2} C_1^* h^{1/2}$ und $C_2^* h^{q-1} \leq \frac{1}{2} C_1^* h^{1/2}$. Ist für ein gewisses $m > 0$ die Voraussetzung (3.33) für alle $n < m$ erfüllt, so gilt nach Lemma 8

$$\|y_m - y(t_m)\| + \|z_m - z(t_m)\| \leq C_2^* h^{q-\frac{1}{2}} + C_2^* h^{q-1} \leq \frac{1}{2} C_1^* h^{1/2} + \frac{1}{2} C_1^* h^{1/2} \leq C_1^* h^{1/2}\ ,$$

also ist (3.33) mit $\mu = 1/2$ und $C_1 = C_1^*$ auch für alle $n < m + 1$ erfüllt. Mittels vollständiger Induktion folgt deshalb aus Lemma 8 für alle m mit $mh \leq T$

$$\|y_m - y(t_m)\| \leq C_2^* h^{q-\frac{1}{2}} \quad \text{und} \quad \|z_m - z(t_m)\| \leq C_2^* h^{q-1}\ . \tag{3.32}$$

b) Wegen (3.32) und $q \geq 2$ ist in Lemma 8 für hinreichend kleine Schrittweiten $h > 0$ stets die Bedingung (3.33) mit $\mu := 1$ und $C_1 := 2C_2^*$ erfüllt; hieraus folgt die Behauptung des Satzes. ∎

Lemma 8 *Zusätzlich zu den Voraussetzungen von Satz 12 sei für ein fixiertes $\mu \in [\frac{1}{2}, 1]$ und für alle $n < m$ die Bedingung*

$$\|y_n - y(t_n)\| + \|z_n - z(t_n)\| \leq C_1 h^\mu \tag{3.33}$$

erfüllt. Dann gilt für hinreichend kleine Schrittweiten $h > 0$

$$\|y_m - y(t_m)\| + h^\mu \|z_m - z(t_m)\| \leq C_2 h^{q-(1-\mu)}\ ,\quad (t_m = mh\ ,\ t_m \in [0, T])$$

mit einer Konstanten C_2, die von C_1 und μ abhängen kann, aber von m und h unabhängig ist.

Beweis a) Wählt man die Koeffizienten des HERK–Verfahrens wie in (3.29), so ist (für $n \geq 1$) $y_n = \eta_{n-1,s}$, also $g(y_n) = 0$. Für $\hat{y}_n = y(t_n)$ gilt deshalb $g(y_n) = g(\hat{y}_n) = 0$ und man erhält in (3.26) wegen (3.33)

$$g_y(y(t_n))(\hat{y}_n - y_n) = \int_0^1 \Big(g_y(y(t_n)) - g_y(y_n + \vartheta(\hat{y}_n - y_n)) \Big)\, d\vartheta \cdot (\hat{y}_n - y_n) = \mathcal{O}(h^\mu)\|\hat{y}_n - y_n\| \tag{3.34}$$

(vgl. (2.25)). Damit folgt auch wie in (2.26) (ersetze $\eta \to y_m$)

$$y_m - y(t_m) = P(t_m)(y_m - y(t_m)) + \mathcal{O}(h^\mu)\|y_m - y(t_m)\|$$

und

$$\|y_m - y(t_m)\| \leq (1 + \mathcal{O}(h^\mu))\|P(t_m)(y_m - y(t_m))\|\ , \tag{3.35}$$

so daß die Konvergenzordnung $q - (1 - \mu)$ für die Komponenten y schon bewiesen ist, wenn nur $\|P(t_m)(y_m - y(t_m))\| = \mathcal{O}(h^{q-(1-\mu)})$ gilt.
b) Mit (3.33) und den Vektoren $\hat{y}_{n+1}$, $\hat{z}_{n+1}$ aus Definition 6 sind die Voraussetzungen von Satz 11 erfüllt und es gilt

$$P(t_{n+1})(\hat{y}_{n+1} - y_{n+1}) = P(t_{n+1})(\hat{y}_n - y_n) + \mathcal{O}(h) \sum\nolimits_j \|\hat{Y}_{nj} - Y_{nj}\| + \mathcal{O}(h^{1+\mu}) \sum\nolimits_j \|\hat{Z}_{nj} - Z_{nj}\|\ , \tag{3.36}$$

denn $P(t_{n+1}) f_z(y(t_{n+1}), z(t_{n+1})) \equiv 0$ und

$$f(\hat{Y}_{nj}, \hat{Z}_{nj}) - f(Y_{nj}, Z_{nj}) = \mathcal{O}(1)\|\hat{Y}_{nj} - Y_{nj}\| + f_z(y(t_{n+1}), z(t_{n+1}))(\hat{Z}_{nj} - Z_{nj}) + \\ + \mathcal{O}(h^\mu) \sum\nolimits_j \|\hat{Z}_{nj} - Z_{nj}\|\ .$$

Wegen (3.34) gilt außerdem in (3.26)

$$\|\hat{Z}_{nj} - Z_{nj} - w_{j1}(\hat{z}_n - z_n)\| \leq C(h^{\mu-1}\|\hat{y}_n - y_n\| + h^\mu \|\hat{z}_n - z_n\|) \tag{3.37}$$

und

$$\sum\nolimits_j d_j(\hat{Z}_{nj} - Z_{nj}) = \kappa(\hat{z}_n - z_n) + \mathcal{O}(h^{\mu-1}\|\hat{y}_n - y_n\| + h^\mu \|\hat{z}_n - z_n\|)\ . \tag{3.38}$$

c) Seien die Vektoren $\hat{y}_{n+1}$, $\hat{z}_{n+1}$ wie in Definition 6 bestimmt. Man erhält

$$\begin{aligned} P(t_{n+1})(y_{n+1} - y(t_{n+1})) &= P(t_{n+1})(y_{n+1} - \hat{y}_{n+1}) + P(t_{n+1})\delta y_h(t_n)\ , \\ h^\mu(z_{n+1} - z(t_{n+1})) &= h^\mu(z_{n+1} - \hat{z}_{n+1}) + h^\mu\, \delta z_h(t_n) \end{aligned} \tag{3.39}$$

und $P(t_{n+1})\delta y_h(t_n) = P(t_n)\delta y_h(t_n) + \mathcal{O}(h)\delta y_h(t_n) = \mathcal{O}(h^{q+1})$, $h^\mu\, \delta z_h(t_n) = \mathcal{O}(h^{q-(1-\mu)})$. Für hinreichend kleine Schrittweiten $h > 0$ folgt nun aus $P(t_{n+1}) = P(t_n) + \mathcal{O}(h)$, aus den Abschätzungen (3.35)–(3.38) und aus $\mu \geq 1/2$

$$\begin{pmatrix} \|P(t_{n+1})(\hat{y}_{n+1} - y_{n+1})\| \\ h^\mu \|\hat{z}_{n+1} - z_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h^r) \\ \mathcal{O}(1) & \overline{\kappa} \end{pmatrix} \begin{pmatrix} \|P(t_n)(\hat{y}_n - y_n)\| \\ h^\mu \|\hat{z}_n - z_n\| \end{pmatrix} \tag{3.40}$$

mit $r := 1$ und einer beliebig vorgebbaren Zahl $\overline{\kappa} \in (\kappa, 1)$. (Ungleichung (3.40) ist komponentenweise zu lesen.) Wendet man das nachfolgende Lemma 9 mit

$$u_n = \|P(t_n)(\hat{y}_n - y_n)\|\,, \quad v_n = h^{\mu}\|\hat{z}_n - z_n\|\,, \quad M_u = \mathcal{O}(h^{q_y})\,, \quad M_v = \mathcal{O}(h^{q_z+2-(1-\mu)})$$

auf (3.39) und (3.40) an, so ergibt sich

$$\|P(t_m)(y_m - y(t_m))\| \le \tilde{C}_2 h^{q-(1-\mu)}\,, \quad \|z_m - z(t_m)\| \le \tilde{C}_2 h^{q-1}$$

mit einer von m und h unabhängigen Konstanten $\tilde{C}_2$. Hieraus folgt mit (3.35) die Behauptung des Lemmas. ∎

Zur Untersuchung der Fehlerfortpflanzung in den Komponenten $\|P(t_n)(\hat{y}_n - y_n)\|$ und $\|\hat{z}_n - z_n\|$ transformiert man die 2×2–Matrix in (3.40) auf Diagonalgestalt ([51]):

Lemma 9 (vgl. [84, Lemma VI.3.9]) *Gegeben seien zwei Folgen (u_n), (v_n) von nichtnegativen Zahlen, für die (komponentenweise) die Ungleichung*

$$\begin{pmatrix} u_{n+1} \\ v_{n+1} \end{pmatrix} \le \begin{pmatrix} 1+\mathcal{O}(h) & \mathcal{O}(h^r) \\ \mathcal{O}(1) & \overline{\kappa} \end{pmatrix} \begin{pmatrix} u_n \\ v_n \end{pmatrix} + \begin{pmatrix} hM_u \\ M_v \end{pmatrix}$$

mit $r \ge 1$, $0 \le \overline{\kappa} < 1$ und $M_u \ge 0$, $M_v \ge 0$ für alle $n \ge 0$ erfüllt ist. Dann gibt es eine von h und m unabhängige Konstante C_0, so daß für hinreichend kleines $h > 0$ und für alle m mit $mh \le T$ die folgenden Abschätzungen gelten:

$$\begin{aligned} u_m &\le C_0(u_0 + h^r v_0 + M_u + h^{r-1} M_v)\,, \\ v_m &\le C_0(u_0 + (\overline{\kappa}^m + h^r)v_0 + M_u + M_v)\,. \end{aligned}$$

Bemerkung 18 Satz 12 kann ebenso wie die Konvergenzbeweise für Runge–Kutta–Verfahren in [81] und [84] direkt auf den Fall variabler Schrittweiten übertragen werden. Entscheidend für die Konvergenz ist die Kontraktivitätsbedingung $\kappa < 1$. HERK–Verfahren mit $\kappa > 1$ divergieren, im Fall $\kappa = 1$ kann Ordnungsreduktion auftreten. Sind die Anfangswerte von (3.21) nicht exakt, so bleibt die Aussage von Satz 12 gültig, falls $y_0 = y(0) + \mathcal{O}(h^{q_y+1})$, $z_0 = z(0) + \mathcal{O}(h^{q_z+1})$.

Wie erstmals von Jay ([92]) für implizite Runge–Kutta–Verfahren vom Lobatto–Typ untersucht, ergibt sich für die differentiellen Lösungskomponenten y eine höhere Konvergenzordnung, wenn $q_y > q_z + 2$ ist und in (3.40) $r \ge 2$ gilt. Murua ([117]) konstruiert ein Baummodell, um Bedingungen an die Koeffizienten a_{ij}, b_j, γ_{ij} angeben zu können, die ein möglichst großes r in (3.40) garantieren. Bei der Konstruktion von Verfahren höherer Ordnung finden diese Ergebnisse bisher Anwendung für $r = 2$ (z. B. bei der Konstruktion des Verfahrens HEDOP5 in Beispiel 22d). Für diesen Spezialfall wird in Folgerung 3 ein elementarer Beweis der höheren Konvergenzordnung in y gegeben:

Folgerung 3 *Gegeben sei ein HERK–Verfahren (3.21), das zusätzlich zu den Voraussetzungen von Satz 12 die Bedingungen*

$$\gamma_{ij} = a_{i+1,j}\,,\ (i = 2, \dots, \hat{s}-1,\ j = 1, \dots, i)\,, \quad b_2 = 0\,, \quad \sum\nolimits_j b_j c_j w_{j1} = 0 \tag{3.41}$$

erfüllt. Gilt mit $\hat{y}_n := y(t_n)$, $\hat{z}_n := z(t_n)$ und $\delta_i = 0$, $\theta_i = 0$, $(i = 1, \dots, \hat{s})$ für die Stufenvektoren in (3.24)

$$\hat{Y}_{ni} = y(t_n + c_i h) + \mathcal{O}(h^2)\,, \quad \hat{Z}_{ni} = z(t_n + c_i h) + \mathcal{O}(h^2)\,, \quad (i = 3, \dots, \hat{s}) \tag{3.42}$$

und ist $q := \min(q_y, q_z + 3) \ge 4$, so konvergiert das HERK–Verfahren mit der Ordnung q in y und mit der Ordnung $q-2$ in z und es gilt

$$\|y_m - y(t_m)\| = \mathcal{O}(h^{q_y}) + \mathcal{O}(h^2) \max_{n \le m} \|\delta z_h(t_n)\|\,, \quad (\,t_m = mh\,,\ t_m \in [0, T]\,)\,. \tag{3.43}$$

Beweis a) Wegen (3.41) ist in (3.21) $Y_{ni} = \eta_{n,i-1}$, $(i = 3, \dots, \hat{s})$, also $g(Y_{ni}) = 0$ und wie in (3.35) und (3.36) erhält man die gegenüber (3.25) verbesserte Abschätzung

$$\|\hat{Y}_{ni} - Y_{ni}\| \le C(\|\hat{y}_n - y_n\| + h^2\|\hat{z}_n - z_n\| + h\|\delta\| + \|\theta\|)\,, \quad (i = 3, \dots, \hat{s})\,. \tag{3.44}$$

b) Nach Satz 12 konvergiert das Verfahren (mindestens) mit der Ordnung $q-1$ in y und mit der Ordnung $q-2$ in z. Wegen (3.42) gilt also auch

$$Y_{ni} = y(t_n + c_i h) + \mathcal{O}(h^2)\,, \quad Z_{ni} = z(t_n + c_i h) + \mathcal{O}(h^2)\,, \quad (i = 3, \dots, \hat{s})$$

und damit

$$f(\hat{Y}_{nj}, \hat{Z}_{nj}) - f(Y_{nj}, Z_{nj}) = \mathcal{O}(1)\|\hat{Y}_{nj} - Y_{nj}\| + f_z(y(t_n + c_j h), z(t_n + c_j h))(\hat{Z}_{nj} - Z_{nj}) + \\ + \mathcal{O}(h^2)\|\hat{Z}_{nj} - Z_{nj}\|$$

für $j = 3, \dots, \hat{s}$. Mit den Bezeichnungen $f_z^n := f_z(y(t_n), z(t_n))$, ... und der Tensorschreibweise (2.19) ergibt sich durch Taylorentwicklung

$$\begin{aligned} &\sum\nolimits_j b_j(f(\hat{Y}_{nj}, \hat{Z}_{nj}) - f(Y_{nj}, Z_{nj})) = \\ &\quad = \sum\nolimits_{j \ne 2}(\mathcal{O}(1)\|\hat{Y}_{nj} - Y_{nj}\| + \mathcal{O}(h^2)\|\hat{Z}_{nj} - Z_{nj}\|) + \sum\nolimits_j b_j f_z^n \cdot (\hat{Z}_{nj} - Z_{nj}) + \\ &\qquad + \sum\nolimits_j b_j f_{yz}^n(\hat{Z}_{nj} - Z_{nj}, c_j h y'(t_n)) + \sum\nolimits_j b_j f_{zz}^n(\hat{Z}_{nj} - Z_{nj}, c_j h z'(t_n))\,. \end{aligned}$$

Erfüllen die Koeffizienten des HERK–Verfahrens die Bedingung (3.41), so können die Ausdrücke $\sum_j b_j f_{yz}^n \dots$ und $\sum_j b_j f_{zz}^n \dots$ weiter umgeformt werden, z. B.

$$\sum\nolimits_j b_j f_{yz}^n(\hat{Z}_{nj} - Z_{nj}, c_j h y'(t_n)) = h \sum\nolimits_j b_j c_j f_{yz}^n(\hat{Z}_{nj} - Z_{nj} - w_{j1}(\hat{z}_n - z_n), y'(t_n))\,.$$

Auf diese Weise ergibt sich die Abschätzung

$$P(t_n) \sum\nolimits_j b_j(f(\hat{Y}_{nj}, \hat{Z}_{nj}) - f(Y_{nj}, Z_{nj})) = \mathcal{O}(1) \sum\nolimits_{j \ne 2} \|\hat{Y}_{nj} - Y_{nj}\| + \\ + \mathcal{O}(h) \sum\nolimits_{j \ne 2} \|\hat{Z}_{nj} - Z_{nj} - w_{j1}(\hat{z}_n - z_n)\| + \mathcal{O}(h^2) \sum\nolimits_{j \ne 2} \|\hat{Z}_{nj} - Z_{nj}\|\,.$$

Unter Verwendung von (3.26) und (3.44) kann man deshalb analog zum Beweisteil b) von Lemma 8 zeigen, daß unter den Voraussetzungen von Folgerung 3 die Ungleichung (3.40) mit $\mu = 1$ und $r = 2$ erfüllt ist, so daß die Behauptung aus Lemma 9 folgt. ∎

Bevor wir in Abschnitt 3.3 zu HERK–Verfahren zurückkehren, wird zunächst die Konvergenz von Mehrschrittverfahren untersucht.

3.2.2 Konvergenz von partitionierten linearen Mehrschrittverfahren für differentiell-algebraische Systeme vom Index 2 in Hessenbergform

In diesem Abschnitt wird auf der Grundlage von Kapitel VII.3 aus [84] die Konvergenz von *partitionierten linearen Mehrschrittverfahren* (PLMSV)

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j y_{n+j} &= h \sum_{j=0}^{k-1} \beta_j f(y_{n+j}, z_{n+j}) + h\beta_k f(y_{n+k}, z_{n+k} + \zeta_{n+k}) \\ 0 &= g(y_{n+k}) \\ h \sum_{j=0}^{k} \hat{\beta}_j z_{n+j} &= \Psi(y_n, \ldots, y_{n+k}, z_n, \ldots, z_{n+k}; f, g, h) \end{aligned} \tag{3.45}$$

untersucht. In (3.45) bezeichnet ζ_{n+k} künstliche Hilfsvariablen und Ψ eine zunächst weitgehend beliebige Verfahrensfunktion, die mindestens ebenso oft stetig differenzierbar sei wie f und g. Diese Verfahrensfunktion Ψ soll für $t_{n+j} := t_n + jh$ und beliebige Vektoren y_{n+j}, z_{n+j} und $\hat{y}_{n+j}$, $\hat{z}_{n+j}$ mit

$$\|y_{n+j} - y(t_{n+j})\| + \|z_{n+j} - z(t_{n+j})\| = \mathcal{O}(h) \ , \quad \|g(y_{n+j})\| = \mathcal{O}(h^2) \ ,$$

$$\|\hat{y}_{n+j} - y(t_{n+j})\| + \|\hat{z}_{n+j} - z(t_{n+j})\| = \mathcal{O}(h) \ , \quad \|g(\hat{y}_{n+j})\| = \mathcal{O}(h^2) \ , \quad (j = 0, 1, \ldots, k)$$

die Bedingung

$$\begin{aligned} &\|\Psi(\hat{y}_n, \ldots, \hat{y}_{n+k}, \hat{z}_n, \ldots, \hat{z}_{n+k}; f, g, h) - \Psi(y_n, \ldots, y_{n+k}, z_n, \ldots, z_{n+k}; f, g, h)\| \leq \\ &\leq \ \mathcal{O}(1) \sum_{j=0}^{k} (\|g_y(y(t_{n+j}))(\hat{y}_{n+j} - y_{n+j})\| + h\|\hat{y}_{n+j} - y_{n+j}\| + h^2\|\hat{z}_{n+j} - z_{n+j}\|) \end{aligned} \tag{3.46}$$

erfüllen, außerdem sei

$$h \sum_{j=0}^{k} \hat{\beta}_j z(t_{n+j}) - \Psi(y(t_n), \ldots, y(t_{n+k}), z(t_n), \ldots, z(t_{n+k}); f, g, h) = \mathcal{O}(h^2) \ . \tag{3.47}$$

In allen hier betrachteten PLMSV ist z_{n+k} implizit als Lösung eines nichtlinearen Gleichungssystems definiert, (3.45) wurde nur für die theoretischen Untersuchungen explizit bezüglich $\sum_j \hat{\beta}_j z_{n+j}$ formuliert. Definition (3.45) umfaßt die in [84, Kapitel VII.3] untersuchten (klassischen) linearen Mehrschrittverfahren (vgl. auch die dort zusammengestellten Literaturangaben) und außerdem partitionierte Verfahren, die speziell für nicht-steife Index-2-Systeme entwickelt wurden (insbesondere β-geblockte Verfahren [7] und die in Abschnitt 3.3.3 eingeführten PLMSV vom Adams–Typ).

Ebenso wie in der Definition (2.28) des impliziten Eulerverfahrens ist für (3.45) ein nichtlineares Gleichungssystem zu lösen (vgl. Satz 4). Die lokal eindeutige Lösbarkeit von (3.45) ergibt sich unmittelbar aus dem entsprechenden Satz für (klassische) Mehrschrittverfahren:

Satz 13 (vgl. [84, Satz VII.3.1]) *Gegeben sei ein PLMSV (3.45) mit (3.46) und (3.47), für das* $\sum_{j=0}^{k} \alpha_j = 0$, $\alpha_k \neq 0$, $\beta_k \neq 0$ *und* $\hat{\beta}_k \neq 0$ *gilt. Erfüllen die Anfangswerte*

$$\|y_{n+i} - y(t_{n+i})\| + \|z_{n+i} - z(t_{n+i})\| = \mathcal{O}(h) \ , \quad \|g(y_{n+i})\| = \mathcal{O}(h^2) \ , \quad (i = 0, 1, \ldots, k-1) \ ,$$

so hat das nichtlineare Gleichungssystem (3.45) für hinreichend kleine Schrittweiten $h > 0$ *eine (lokal) eindeutig bestimmte Lösung, für die gilt*

$$\|y_{n+k} - y(t_{n+k})\| + \|z_{n+k} - z(t_{n+k})\| = \mathcal{O}(h) \ , \quad \|\zeta_{n+k}\| = \mathcal{O}(h) \ .$$

Beweis Nach Satz VII.3.1 aus [84] sind die Vektoren y_{n+k} und $\zeta := z_{n+k} + \zeta_{n+k}$ für hinreichend kleines $h > 0$ durch (3.45) lokal eindeutig bestimmt und es gilt $\|y_{n+k} - y(t_{n+k})\| = \mathcal{O}(h)$ und $\|\zeta - z(t_{n+k})\| = \mathcal{O}(h)$; darüberhinaus ist $\|g_y(y(t_{n+i}))(y_{n+i} - y(t_{n+i}))\| = \|g(y_{n+i})\| + \mathcal{O}(h)\|y_{n+i} - y(t_{n+i})\| = \mathcal{O}(h^2)$ für $i = 0, 1, \ldots, k$ (vgl. (3.34)).

Nach dem Satz über die implizite Funktion ist dann wegen $\hat{\beta}_k \neq 0$ und $\dfrac{\partial \Psi}{\partial z_{n+k}} = \mathcal{O}(h^2)$ auch der Vektor z_{n+k} durch $h\sum_j \hat{\beta}_j z_{n+j} = \Psi$ (lokal eindeutig) bestimmt und es gilt wegen (3.47) $\|z_{n+k} - z(t_{n+k})\| = \mathcal{O}(h)$, also folgt die Behauptung. ∎

Wie für HERK–Verfahren untersucht man auch für Mehrschrittverfahren zunächst die Fehlerfortpflanzung in einem einzelnen Integrationsschritt:

Satz 14 (vgl. [84, Satz VII.3.2]) *Zu einem PLMSV, das den Voraussetzungen von Satz 13 genügt, werden Vektoren* $\hat{y}_n, \ldots, \hat{y}_{n+k}$, $\hat{z}_n, \ldots, \hat{z}_{n+k}$, $\hat{\zeta}_{n+k}$ *betrachtet, für die gilt*

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j \hat{y}_{n+j} &= h \sum_{j=0}^{k-1} \beta_j f(\hat{y}_{n+j}, \hat{z}_{n+j}) + h\beta_k f(\hat{y}_{n+k}, \hat{z}_{n+k} + \hat{\zeta}_{n+k}) + h\delta_y \\ \theta &= g(\hat{y}_{n+k}) \\ h \sum_{j=0}^{k} \hat{\beta}_j \hat{z}_{n+j} &= \Psi(\hat{y}_n, \ldots, \hat{y}_{n+k}, \hat{z}_n, \ldots, \hat{z}_{n+k}; f, g, h) + h\delta_z \end{aligned} \tag{3.48}$$

und

$$\begin{gathered} \|\hat{y}_{n+j} - y(t_{n+j})\| = \mathcal{O}(h^2) \ , \quad \|\hat{z}_{n+j} - z(t_{n+j})\| = \mathcal{O}(h) \ , \quad (j = 0, 1, \ldots, k) \ , \\ \|\hat{\zeta}_{n+k}\| = \mathcal{O}(h) \ , \quad \|\delta_y\| + \|\delta_z\| = \mathcal{O}(h) \ . \end{gathered} \tag{3.49}$$

Ist die Schrittweite $h > 0$ *hinreichend klein, so gilt*

$$\|\hat{y}_{n+k} - y_{n+k}\| \ \leq \ C\Big(\sum_{j=0}^{k-1} (\|\hat{y}_{n+j} - y_{n+j}\| + h\|\hat{z}_{n+j} - z_{n+j}\|) + h\|\delta_y\| + \|\theta\|\Big) \ , \tag{3.50}$$

$$\|\hat{z}_{n+k} - z_{n+k}\| \ \leq \ C\Big(\frac{1}{h} \sum_{j=0}^{k} \|\hat{y}_{n+j} - y_{n+j}\| + \sum_{j=0}^{k-1} \|\hat{z}_{n+j} - z_{n+j}\| + D_h\Big) \ , \tag{3.51}$$

$$\|\hat{\zeta}_{n+k} - \zeta_{n+k}\| \ \leq \ C\Big(\sum_{j=0}^{k} (\frac{1}{h}\|\hat{y}_{n+j} - y_{n+j}\| + \|\hat{z}_{n+j} - z_{n+j}\|) + D_h\Big) \tag{3.52}$$

mit $D_h := \|\delta_y\| + \|\delta_z\| + \frac{1}{h}\|\theta\|$ *und einer Konstanten* C*, die von* h*,* δ_y*,* δ_z*,* θ *und den Vektoren* $\hat{y}_{n+j}$*,* $\hat{z}_{n+j}$*,* $\hat{\zeta}_{n+k}$ *unabhängig ist.*

Beweis Mit $\hat{\zeta} := \hat{z}_{n+k} + \hat{\zeta}_{n+k}$ kann wie im Beweis von Satz 13 der entsprechende Satz VII.3.2 aus [84] direkt angewendet werden, insbesondere ist (3.50) direkte Folgerung aus [84, (VII.3.18)]. Aus (3.46) folgt dann die Abschätzung (3.51) und hieraus schließlich mit [84, (VII.3.18)] auch die Fehlerschranke (3.52), denn $\|\hat{\zeta}_{n+k} - \zeta_{n+k}\| \le \|\hat{z}_{n+k} - z_{n+k}\| + \|\hat{\zeta} - \zeta\|$. ∎

Wie für Mehrschrittverfahren üblich, führt man die *charakteristischen Polynome* des PLMSV (3.45) ein (engl.: generating polynomials):

$$\varrho(\xi) := \sum_{j=0}^{k} \alpha_j \xi^j \ , \quad \sigma(\xi) := \sum_{j=0}^{k} \beta_j \xi^j \ , \quad \hat{\varrho}(\xi) := \sum_{j=0}^{k} \hat{\alpha}_j \xi^j \ , \quad \hat{\sigma}(\xi) := \sum_{j=0}^{k} \hat{\beta}_j \xi^j \ ,$$

wobei $\hat{\varrho}$ hier schon in Vorbereitung von Abschnitt 3.3.3 definiert wird.

Bemerkung 19 Im folgenden wird ausgiebig auf bekannte Ergebnisse und Begriffe aus der Theorie linearer Mehrschrittverfahren für gewöhnliche Differentialgleichungen zurückgegriffen ohne diese hier im Detail einzuführen. Die Darstellung orientiert sich dabei an Kapitel III aus [82]. So wird z. B. in Satz 15 gefordert, daß ϱ die Wurzelbedingung und $\hat{\sigma}$ die strenge Wurzelbedingung erfüllt, d. h. für $\xi \in \mathbb{C}$ gilt ([82, Definition III.3.2])

$$\varrho(\xi) = 0 \ \Rightarrow \ |\xi| \le 1 \qquad \text{und} \qquad \varrho(\xi) = \varrho'(\xi) = 0 \ \Rightarrow \ |\xi| < 1 \ ,$$

$$\hat{\sigma}(\xi) = 0 \ \Rightarrow \ |\xi| < 1 \ .$$

Die *Ordnung* p eines durch (ϱ, σ) gegebenen Mehrschrittverfahrens mit $\sum_{j=0}^{k} \alpha_j = 0$ ist die größtmögliche Zahl $p \in \mathbb{N}$, für die

$$\sum_{j=0}^{k} \alpha_j j^q = q \sum_{j=0}^{k} \beta_j j^{q-1} \ , \quad (q = 1, \ldots, p)$$

gilt ([82, Satz III.2.4]).

Analog zu Definition 6 werden lokale Fehler von PLMSV eingeführt:

Definition 7 In (3.48) sei $\delta_y = 0$, $\delta_z = 0$, $\theta = 0$ und $\hat{y}_{n+j}$, $\hat{z}_{n+j}$, $(j = 0, 1, \ldots, k-1)$ durch die analytische Lösung von (3.19) gegeben: $\hat{y}_{n+j} := y(t_{n+j})$, $\hat{z}_{n+j} := z(t_{n+j})$. Dann heißen

$$\delta y_h(t_n) := \hat{y}_{n+k} - y(t_{n+k}) \quad \text{und} \quad \delta\zeta_h(t_n) := \hat{z}_{n+k} + \hat{\zeta}_{n+k} - z(t_{n+k})$$

lokale Fehler im differentiellen Teil und

$$\delta z_h(t_n) := \hat{z}_{n+k} - z(t_{n+k})$$

lokaler Fehler im algebraischen Teil des PLMSV (3.45).

Lemma 10 ([84, Lemma VII.3.4]) *Enthält das PLMSV (3.45) ein durch (ϱ, σ) charakterisiertes lineares Mehrschrittverfahren der Ordnung p, so gilt für die lokalen Fehler*

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

Der entscheidende Vorteil partitionierter Verfahren gegenüber den klassischen Mehrschrittverfahren für Index-2-Systeme ist eine höhere Konvergenzordnung in y (vgl. Abschnitt 3.3.3). Der nachfolgende Konvergenzbeweis für PLMSV, der sich z. T. an den Beweisen der Sätze VII.3.5 und VII.3.6 aus [84] orientiert, ist dieser Situation angepaßt. (Es sei ausdrücklich darauf verwiesen, daß die in Satz 15 für die Komponenten z angegebenen Fehlerschranken im Spezialfall klassischer Mehrschrittverfahren nicht optimal sind, vgl. hierzu [84, Satz VII.3.6]).

Satz 15 *Gegeben sei ein PLMSV (3.45) mit (3.46) und (3.47), $\sum_{j=0}^{k} \alpha_j = 0$, $\alpha_k \neq 0$, $\beta_k \neq 0$ und $\hat{\beta}_k \neq 0$, dessen charakteristische Polynome ϱ und $\hat{\sigma}$ die Wurzelbedingung (ϱ) bzw. die strenge Wurzelbedingung ($\hat{\sigma}$) erfüllen. Hat das durch (ϱ, σ) charakterisierte lineare Mehrschrittverfahren die klassische Ordnung q_y und gilt für den lokalen Fehler des PLMSV $\|\delta z_h(t)\| = \mathcal{O}(h^{q_z})$ mit $q := \min(q_y, q_z + 1) \ge 3$, so konvergiert das PLMSV mit der Ordnung q in y und mit der Ordnung $q-1$ in z, d. h.*

$$\|y_m - y(t_m)\| + h\|z_m - z(t_m)\| = \mathcal{O}(h^q) \ , \quad (t_m = mh \, , \ t_m \in [0, T]) \ ,$$

sofern die Anfangswerte folgende Bedingungen erfüllen:

$$\|y_i - y(t_i)\| = \mathcal{O}(h^{q_y+1}) \ , \quad \|z_i - z(t_i)\| = \mathcal{O}(h^{q_z}) \ , \quad (i = 0, 1, \ldots, k-1) \ .$$

Beweis Ebenso wie im Konvergenzbeweis von HERK–Verfahren wird der Projektor $P(t)$ aus (2.12) verwendet. Für Vektoren (η, ζ) mit $\|\eta - y(t)\| + \|\zeta - z(t)\| = \mathcal{O}(h)$ gilt dabei (vgl. (3.35) und (3.36))

$$P(t)(f(\eta, \zeta) - f(y(t), z(t))) = \mathcal{O}(1)\|\eta - y(t)\| + \mathcal{O}(h)\|\zeta - z(t)\| \tag{3.53}$$

und

$$\|\eta - y(t)\| \le (1 + \mathcal{O}(h))\|P(t)(\eta - y(t))\| + \mathcal{O}(1)\|g(\eta)\| \ . \tag{3.54}$$

Die Fehlerfortpflanzung während der Integration wird in den Vektoren $P(t_n)\Delta y_n$ und Δz_n mit

$$\Delta y_n := y_n - y(t_n) \ , \quad \Delta z_n := z_n - z(t_n)$$

untersucht. Wie in Lemma 8 wird dabei zunächst zusätzlich vorausgesetzt, daß die numerische Lösung in einer kleinen Umgebung der analytischen Lösung verbleibt, d. h., es soll gelten

$$\|\Delta y_n\| \le C_1 h^2 \ , \quad \|\Delta z_n\| \le C_2 h \ , \quad \|\zeta_n\| \le C_3 h \ , \quad (n \ge 0) \tag{3.55}$$

mit Konstanten C_1, C_2 und C_3, die von h unabhängig sind.

a) Setzt man in (3.48) die analytische Lösung ein ($\hat{y}_{n+j} = y(t_{n+j})$, $\hat{z}_{n+j} = z(t_{n+j})$, $\hat{\zeta}_{n+k} = 0$), so folgt wie in (3.51) und (3.52) (vgl. auch [84, Satz VII.3.2]) die Abschätzung

$$\|z(t_{n+k}) - (z_{n+k} + \zeta_{n+k})\| = \mathcal{O}(1)\Big(\frac{1}{h}\sum_{j=0}^{k}\|\Delta y_{n+j}\| + \sum_{j=0}^{k-1}\|\Delta z_{n+j}\| + \tilde{D}_h\Big)$$

mit $\tilde{D}_h := \frac{1}{h}\sum_{j=0}^{k}\alpha_j y(t_{n+j}) - \sum_{j=0}^{k}\beta_j y'(t_{n+j}) = \mathcal{O}(h^{q_y})$ (vgl. [82, Satz III.2.4]). Unter Verwendung von (3.53) ergibt sich

$$\begin{aligned}\|P(t_{n+k})(f(y(t_{n+k}), z(t_{n+k})) - f(y_{n+k}, z_{n+k} + \zeta_{n+k}))\| = \\ = \mathcal{O}(1)\Big(\sum_{j=0}^{k}\|\Delta y_{n+j}\| + h\sum_{j=0}^{k-1}\|\Delta z_{n+j}\|\Big) + \mathcal{O}(h^{q_y+1}) \ .\end{aligned}$$

b) Seien Vektoren $\hat{y}_{n+k}$, $\hat{z}_{n+k}$ und $\hat{\zeta}_{n+k}$ wie in Definition 7 gegeben. Dann ist nach (3.48)

$$\begin{aligned} P(t_{n+k}) \sum_{j=0}^{k} \alpha_j y(t_{n+j}) = \\ &= h \sum_{j=0}^{k} \beta_j P(t_{n+k}) f(y(t_{n+j}), z(t_{n+j})) + \alpha_k P(t_{n+k})(y(t_{n+k}) - \hat{y}_{n+k}) + \\ &\quad + h\beta_k P(t_{n+k})(f(\hat{y}_{n+k}, \hat{z}_{n+k} + \hat{\zeta}_{n+k}) - f(y(t_{n+k}), z(t_{n+k}))) \\ &= h \sum_{j=0}^{k} \beta_j P(t_{n+k}) f(y(t_{n+j}), z(t_{n+j})) + \mathcal{O}(h^{q_y+1}) , \end{aligned} \tag{3.56}$$

denn

$$\|\hat{y}_{n+k} - y(t_{n+k})\| + h^2 \|\hat{z}_{n+k} + \hat{\zeta}_{n+k} - z(t_{n+k})\| = \mathcal{O}(h^{q_y+1})$$

(vgl. Lemma 10 und auch (3.53)). Ebenso folgt

$$h \sum_{j=0}^{k} \hat{\beta}_j z(t_{n+j}) = \Psi(y(t_n), \ldots, y(t_{n+k}), z(t_n), \ldots, z(t_{n+k}); f, g, h) + \mathcal{O}(h^{q_y+1}) + \mathcal{O}(h^{q_z+1}) . \tag{3.57}$$

Multipliziert man in (3.45) das erste Gleichungssystem mit $P(t_{n+k})$ und bildet dann die Differenzen zu (3.56) und (3.57), so folgt aus (3.53), aus Beweisteil a) und aus $P(t_{n+j}) = P(t_{n+k}) + \mathcal{O}(h)$

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j P(t_{n+j}) \Delta y_{n+j} &= \mathcal{O}(h) \sum_{j=0}^{k} \|\Delta y_{n+j}\| + \mathcal{O}(h^2) \sum_{j=0}^{k-1} \|\Delta z_{n+j}\| + \mathcal{O}(h^{q+1}) , \\ h \sum_{j=0}^{k} \hat{\beta}_j \Delta z_{n+j} &= \mathcal{O}(1) \sum_{j=0}^{k} \|\Delta y_{n+j}\| + \mathcal{O}(h^2) \sum_{j=0}^{k} \|\Delta z_{n+j}\| + \mathcal{O}(h^q) . \end{aligned} \tag{3.58}$$

Wegen (3.54) und

$$g(y_n) = 0 , \ (n \geq k) , \quad \|g(y_i)\| = \mathcal{O}(\|y_i - y(t_i)\|) = \mathcal{O}(h^{q+1}) , \ (i = 0, 1, \ldots, k-1)$$

kann man in (3.58) $\sum_j \|\Delta y_{n+j}\|$ durch $\sum_j \|P(t_{n+j})\Delta y_{n+j}\|$ ersetzen.

c) Mit den Bezeichnungen

$$U_n := (\, P(t_{n+k-1})\Delta y_{n+k-1} , \ \ldots , \ P(t_n)\Delta y_n \,)^T , \quad V_n := (\, h\Delta z_{n+k-1} , \ \ldots , \ h\Delta z_n \,)^T \tag{3.59}$$

läßt sich (3.58) daher schreiben als

$$\begin{aligned} U_{n+1} &= (A \otimes I) U_n + \mathcal{O}(h\|U_n\| + h\|V_n\|) + \mathcal{O}(h^{q+1}) \\ V_{n+1} &= (\hat{B} \otimes I) V_n + \mathcal{O}(\|U_n\| + h\|V_n\|) + \mathcal{O}(h^q) \end{aligned} \tag{3.60}$$

mit den Matrizen

$$A := \begin{pmatrix} -\alpha'_{k-1} & \cdots & -\alpha'_1 & -\alpha'_0 \\ 1 & \cdots & 0 & 0 \\ & \ddots & \vdots & \vdots \\ & & 1 & 0 \end{pmatrix} , \quad \hat{B} := \begin{pmatrix} -\hat{\beta}'_{k-1} & \cdots & -\hat{\beta}'_1 & -\hat{\beta}'_0 \\ 1 & \cdots & 0 & 0 \\ & \ddots & \vdots & \vdots \\ & & 1 & 0 \end{pmatrix}$$

und $\alpha'_j := \alpha_j/\alpha_k$, $\hat{\beta}'_j := \hat{\beta}_j/\hat{\beta}_k$. Hierbei können die beiden Normen $\|U\|$ und $\|V\|$ wie beim Beweis der Konvergenz linearer Mehrschrittverfahren für gewöhnliche Differentialgleichungen so gewählt werden, daß

$$\|A \otimes I\| \leq 1 \qquad \text{und} \qquad \|\hat{B} \otimes I\| \leq \kappa \ \text{ mit einem } \kappa \in (0,1)$$

ist, denn ϱ erfüllt die Wurzelbedingung und $\hat{\sigma}$ die strenge Wurzelbedingung ([82, Lemma III.4.4], [152, Satz 6.9.2]). Geht man in (3.60) zu diesen Normen über, so folgt

$$\begin{pmatrix} \|U_{n+1}\| \\ \|V_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \kappa + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|U_n\| \\ \|V_n\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{q+1}) \\ \mathcal{O}(h^q) \end{pmatrix}$$

und $\|U_0\| = \mathcal{O}(h^{q+1})$, $\|V_0\| = \mathcal{O}(h^q)$. Deshalb kann Lemma 9 angewendet werden: Es gilt $\|U_m\| + \|V_m\| = \mathcal{O}(h^q)$, hieraus folgt wegen (3.59) und (3.54) die Behauptung des Satzes.

d) Zum Abschluß des Beweises zeigt man auf dem im Beweis von Satz 12 ausführlich besprochenen Weg, daß wegen $q \geq 3$ die zusätzliche Voraussetzung (3.55) stets erfüllt ist, wenn die Schrittweite $h > 0$ hinreichend klein ist. ∎

Bemerkung 20 a) Ebenso wie der Konvergenzbeweis für klassische Mehrschrittverfahren aus [84] kann auch Satz 15 auf den Fall variabler Schrittweiten verallgemeinert werden (vgl. hierzu [82, Kapitel III.5]).

b) Wie für HERK–Verfahren kann die Voraussetzung $q \geq 3$ in Satz 15 zu $q \geq 2$ abgeschwächt werden, wenn unter geeigneten Voraussetzungen an $(\hat{y}_{n+j}, \hat{z}_{n+j})$ eine zu (3.26) analoge Abschätzung

$$\begin{aligned} \|\sum_{j=0}^{k} \hat{\beta}_j (\hat{z}_{n+j} - z_{n+j})\| \leq C \Big(\sum_{j=0}^{k} \frac{1}{h} \|g_y(y(t_{n+j}))(\hat{y}_{n+j} - y_{n+j})\| + \\ + \sum_{j=0}^{k} (h^{\mu-1} \|\hat{y}_{n+j} - y_{n+j}\| + h^{\mu} \|\hat{z}_{n+j} - z_{n+j}\|) + \|\delta_y\| + \|\delta_z\| + \frac{1}{h}\|\theta\| \Big) \end{aligned} \tag{3.61}$$

für $\mu \in [\frac{1}{2}, 1]$ nachgewiesen werden kann (vgl. Satz 11 und den Beweis von Satz 12). Eine solche Abschätzung (3.61) gilt z. B. für die in Abschnitt 3.3.3 betrachteten PLMSV.

3.3 Partitionierte Verfahren für nicht-steife differentiell-algebraische Systeme vom Index 2 in Hessenbergform

Partitionierte Verfahren sind der speziellen Struktur von DA–Systemen in Hessenbergform angepaßt und sollen insbesondere Anfangswertprobleme für nicht-steife DA–Systeme schneller als vorhandene Standardsoftware, die auf impliziten Verfahren basiert (z. B. DASSL, RADAU5), lösen ([84, Kapitel VII.6]). In diesem Abschnitt wird am Beispiel von

halb-expliziten Runge–Kutta–Verfahren (Abschnitt 3.3.1) und von partitionierten linearen Mehrschrittverfahren vom Adams–Typ (Abschnitt 3.3.3) die Übertragung von Verfahren, die zur Integration nicht-steifer Systeme gewöhnlicher Differentialgleichungen geeignet sind, auf Index-2-Systeme gezeigt. Praktisch finden solche partitionierten Verfahren derzeit vor allem Anwendung bei der dynamischen Simulation mechanischer Mehrkörpersysteme. Der auf einem halb-expliziten Runge–Kutta–Verfahren 5. Ordnung aufbauende Integrator HEDOP5, der in Abschnitt 3.3.2 vorgestellt wird, erweist sich hierfür als besonders effizient. Die neu entwickelten Verfahren werden an Hand von verschiedenen Benchmark–Problemen mit aus der Literatur bekannten Verfahren verglichen.

3.3.1 Konstruktion von halb-expliziten Runge–Kutta-Verfahren für differentiell-algebraische Systeme vom Index 2 in Hessenbergform

Für mehrstufige Einschrittverfahren ist neben dem Nachweis der Konvergenz (vgl. Abschnitt 3.2.1) die Untersuchung des lokalen Diskretisierungsfehlers ein wesentlicher Baustein zur Konstruktion effizienter Verfahren. In diesem Abschnitt wird zunächst motiviert, warum HERK–Verfahren (3.21) mit Koeffizienten

$$\gamma_{11} = 0\ ,\ \ \gamma_{ij} = a_{i+1,j}\ ,\ (i = 2,\ldots,\hat{s}-1,\ j = 1,\ldots,i)\ ,\ \ \gamma_{sj} = b_j\ ,\ (j = 1,\ldots,s) \quad (3.62)$$

besonders zur Konstruktion von effizienten Verfahren höherer Ordnung geeignet sind. Unter Verwendung der von Hairer et al. ([81, Kapitel 5]) entwickelten Verallgemeinerung des Baummodells von Butcher ([43]) auf DA–Systeme vom Index 2 in Hessenbergform werden anschließend die Konsistenzbedingungen für diese Verfahren formuliert und HERK–Verfahren der Ordnung $q = 2$, $q = 3$, $q = 4$ und $q = 5$ konstruiert.

Halb-explizite Runge–Kutta–Verfahren mit expliziter Stufe

Im Unterschied zu impliziten Runge–Kutta–Verfahren gibt es für explizite Runge–Kutta–Verfahren kein allgemein anwendbares Prinzip zur Konstruktion effizienter Verfahren höherer Ordnung. Die Konstruktion von Verfahren, die mit möglichst wenig Stufen eine vorgegebene Ordnung erreichen und dabei kleine Koeffizienten im führenden Fehlerterm (vgl. [82, S. 158]) und ein möglichst großes Stabilitätsgebiet haben, ist deshalb sehr aufwendig (vgl. z. B. den Überblick über solche Verfahren in [82, Kapitel II.5]). Die Konstruktion von HERK–Verfahren höherer Ordnung knüpft an diese Ergebnisse aus der Theorie für gewöhnliche Differentialgleichungen an.

Bemerkung 21 Es erscheint naheliegend, in (3.21) die Vektoren Y_{ni} und η_{ni} nur einmal zu generieren und $\eta_{ns} = y_{n+1}$, $\eta_{ni} = Y_{n,i+1}$, d. h. $g(y_{n+1}) = g(Y_{ni}) = 0$, $(i = 1,\ldots,\hat{s})$ zu fordern. Dieser Ansatz von Hairer et al. ([81, S. 20f]) läßt sich als (3.21) mit

$$\hat{s} = s\ ,\ \ \gamma_{ij} = a_{i+1,j}\ ,\ (i = 1,\ldots,s-1,\ j = 1,\ldots,i)\ ,\ \ \gamma_{sj} = b_j\ ,\ (j = 1,\ldots,s) \quad (3.63)$$

schreiben. Insbesondere das Verfahren 4. Ordnung HEM4 (Brasey und Hairer [38]) und das Verfahren 5. Ordnung HEM5 (Brasey [36]) finden Anwendung bei der dynamischen Simulation von MKS. Für die aus der Literatur bekannten expliziten Runge–Kutta–Verfahren

für gewöhnliche Differentialgleichungen führt die Parameterwahl (3.63) jedoch i. allg. nur auf HERK–Verfahren der Konvergenzordnung $q \le 2$ ([81, S. 68ff]). Als Grund für diese Ordnungsreduktion erkennt man, daß nur die Verfahren von Brasey und Hairer, nicht aber die aus der Literatur bekannten expliziten Runge–Kutta–Verfahren berücksichtigen, daß die Parameterwahl (3.63) i. allg. zu einer großen Differenz zwischen $f(Y_{n1}, Z_{n1}) =: Y'_{n1}$ und $f(y(t_n + c_1 h), z(t_n + c_1 h)) = y'(t_n)$ führt: Für $Y_{n1} = y_n := y(t_n)$ folgt aus

$$Y_{n2} = y_n + c_2 h f(Y_{n1}, Z_{n1})\ ,\ \ 0 = g(Y_{n2}) \quad (3.64)$$

durch Taylorentwicklung ([81, S. 68f])

$$\begin{aligned} Y'_{n1} = f(Y_{n1}, Z_{n1}) &= f(y_n, Z_{n1}) = f(y(t_n), z(t_n)) + f_z(y(t_n), z(t_n))(Z_{n1} - z(t_n)) + \mathcal{O}(h^2) \\ &= y'(t_n) + \frac{c_2 h}{2}[f_z(-g_y f_z)^{-1} g_{yy}(f,f)](y(t_n), z(t_n)) + \mathcal{O}(h^2)\ , \end{aligned}$$

d. h. $Y'_{n1} - y'(t_n) = \mathcal{O}(h)$, falls die Zwangsbedingungen nichtlinear sind ($g_y \neq \text{const}$). (Dagegen gilt bei Anwendung eines expliziten Runge–Kutta–Verfahrens auf gewöhnliche Differentialgleichungen $Y'_{n1} - y'(t_n) = 0$, falls $y_n = y(t_n)$.)

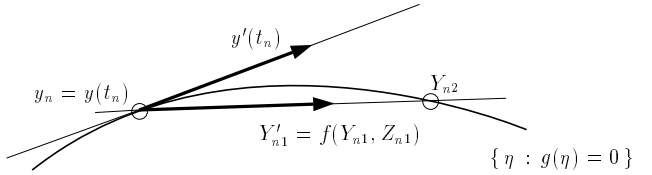


Abbildung 3.3: Differenz zwischen $f(Y_{n1}, Z_{n1})$ und $y'(t_n)$ in HERK–Verfahren (3.21) mit Parametern (3.63).

Abb. 3.3 zeigt, daß die große Differenz zwischen Y'_{n1} und $y'(t_n)$ durch die Bedingung $\gamma_{11} = a_{21}$, also $g(Y_{n2}) = 0$ in (3.64), verursacht wird: Ist $y_n = y(t_n)$, so gilt $f(Y_{n1}, Z_{n1}) = \frac{1}{c_2 h}(Y_{n2} - y(t_n))$, d. h., $Y'_{n1} = f(Y_{n1}, Z_{n1})$ ist parallel zu einem Vektor, der zwei Punkte der Mannigfaltigkeit $\{\,\eta\ :\ g(\eta) = 0\,\}$ miteinander verbindet (die Punkte $y(t_n)$ und Y_{n2}). Andererseits liegt aber $y'(t_n)$ wegen $0 = \frac{d}{dt} g(y(t))|_{t=t_n} = g_y(y(t_n)) y'(t_n)$ in der Tangentialebene an diese Mannigfaltigkeit im Punkt $y(t_n)$.
Die Verfahren von Brasey und Hairer halten den Einfluß von $f(Y_{n1}, Z_{n1})$ auf y_{n+1} klein, indem für HEM4 und HEM5 die Parameter a_{ij}, b_j so gewählt werden, daß $b_1 = 0$, $\sum_j b_j a_{j1} = 0$ gilt, und für HEM5 die Bedingungen $\sum_j b_j c_j a_{j1} = \sum_{j,k} b_j a_{jk} a_{k1} = 0$ erfüllt sind. Dies führt letztlich dazu, daß HEM4 und HEM5 mit 5 bzw. 8 Stufen mehr Stufen (und damit mehr Aufrufe von f) benötigen als die aus der Theorie der gewöhnlichen Differentialgleichungen bekannten expliziten Runge–Kutta–Verfahren, die mit $s = 4$ Stufen die Ordnung 4 und mit $s = 6$ Stufen die Ordnung 5 erreichen ([43]).

Als Alternative zu dem Konstruktionsprinzip von HEM4 und HEM5 wird in den zeitgleich, aber unabhängig voneinander entstandenen Arbeiten [118] und [21] vorgeschlagen, die halb-explizite erste Stufe (3.64) der Verfahren von Hairer et al. durch die explizite Stufe

$$Y_{n2} = y_n + c_2 h f(Y_{n1}, Z_{n1})\ ,\ \ Z_{n1} = z_n \quad (3.65)$$

zu ersetzen (vgl. (3.23) und Abschnitt 3.2.1).

Bemerkung 22 Die HERK–Verfahren mit expliziter Stufe (3.65) sind durch $\gamma_{11} = 0$ charakterisiert. Murua, auf den auch die Notation (3.21) für HERK–Verfahren zurückgeht ([118], dort als partitionierte HERK–Verfahren bezeichnet), betrachtet Verfahren mit

$$\gamma_{11} = 0\ ,\ \ \gamma_{ii} \neq 0\ ,\ (i = 2, \ldots, \hat{s})\ ,\ \ \gamma_{sj} = b_j\ ,\ (j = 1, \ldots, s)\ , \tag{3.66}$$

deren Konvergenz in Satz 12 untersucht wurde. Enthalten die Funktionen f und g in (3.19) wie für die Index-2-Formulierung der MKS–Modellgleichungen eine Funktion $G(y)$, die sowohl in f als auch in g auftritt (vgl. Abschnitt 3.3.2), so ist es darüberhinaus besonders günstig, wenn $\eta_{ni} = Y_{n,i+1}$ ist (bei der Auswertung von $f(Y_{n,i+1}, Z_{n,i+1})$ kann der bereits zuvor bei der Auswertung von $g(\eta_{ni})$ berechnete Wert $G(Y_{n,i+1}) = G(\eta_{ni})$ verwendet werden). In der vorliegenden Arbeit stehen wie in [21] diese durch (3.62) charakterisierten HERK–Verfahren im Vordergrund. Im Vergleich von HERK–Verfahren 5. Ordnung, deren Koeffizienten nach (3.63) (HEM5 [36]), nach (3.66) (PHEM56 [118]) bzw. nach (3.62) (HEDOP5) gewählt werden, sind PHEM56 und HEDOP5 dem Code HEM5 deutlich überlegen, wobei HEDOP5 meist geringfügig schneller als PHEM56 arbeitet (vgl. Abschnitt 3.3.2).

Konsistenzbedingungen

Wie im Abschnitt 3.2.1 werden im folgenden HERK–Verfahren mit expliziter erster Stufe (3.65) untersucht. In Satz 12 führen die Voraussetzungen (3.31) an den lokalen Fehler auf *Konsistenzbedingungen* an die Koeffizienten a_{ij}, b_j, γ_{ij}, die durch Vergleich der Taylorentwicklungen von $y(t_{n+1})$, $z(t_{n+1})$ und $\hat{y}_{n+1}$, $\hat{z}_{n+1}$ aus Definition 6 erhalten werden. Hierbei drückt man wie im klassischen Baummodell von Butcher ([43]) in

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hy'(t_n) + \frac{h^2}{2}y''(t_n) + \mathcal{O}(h^3) \\ z(t_{n+1}) &= z(t_n) + hz'(t_n) + \frac{h^2}{2}z''(t_n) + \mathcal{O}(h^3) \end{aligned}$$

die Ableitungen $y'(t_n)$, $z'(t_n)$, ... durch elementare Differentiale aus.
Zunächst gilt $y'(t) = f(y(t), z(t))$ und

$$y''(t) = f_y(y(t), z(t))y'(t) + f_z(y(t), z(t))z'(t)\ .$$

Differenziert man nun in (3.19) zweimal die Zwangsbedingung $g(y) = 0$, so folgt

$$0 = g_y(y(t))y'(t) = [g_y f](y(t), z(t))$$

und

$$\begin{aligned} 0 &= g_{yy}(y(t))(y'(t), y'(t)) + g_y(y(t))y''(t) \\ &= [g_{yy}(f, f)](y(t), z(t)) + [g_y f_y f](y(t), z(t)) + [g_y f_z](y(t), z(t))z'(t)\ , \end{aligned}$$

also wegen (3.20)

$$z'(t) = [(-g_y f_z)^{-1} g_{yy}(f, f)](y(t), z(t)) + [(-g_y f_z)^{-1} g_y f_y f](y(t), z(t))\ . \tag{3.67}$$

Entsprechende Darstellungen für höhere Ableitungen von y und z erhält man in gleicher Weise durch wiederholte Differentiation der Zwangsbedingungen.

Um die sehr schnell wachsende Zahl elementarer Differentiale handhaben zu können, identifiziert man sie mit Wurzelbäumen. Im weiteren werden für dieses Baummodell die Bezeichnungen von Hairer et al. ([81, Kapitel 5]) verwendet. Dabei wird auf die detaillierte Wiedergabe des Baummodells verzichtet, weil sie über den Rahmen der vorliegenden Arbeit hinausginge.

Bemerkung 23 Für Index-2-Systeme (3.19) werden Mengen T_y und T_z von speziellen endlichen gerichteten Graphen mit zweierlei Arten von Knoten (mageren „•“ und fetten „○“ Knoten) betrachtet, die ausgehend von dem Baum $\tau = \bullet$ rekursiv definiert werden:

i) $\tau \in T_y$.

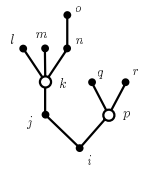
ii) Sind $t_1, \ldots, t_m \in T_y$ und $u_1, \ldots, u_n \in T_z$, so ist $[t_1, \ldots, t_m, u_1, \ldots, u_n]_y \in T_y$.

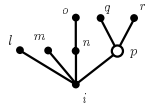
iii) Sind $t_1, \ldots, t_m \in T_y$ und $m \geq 2$, so ist $[t_1, \ldots, t_m]_z \in T_z$.

iv) Sind $t_1, \ldots, t_m \in T_y$ und $u_1, \ldots, u_n \in T_z$ und $m \geq 1$ oder $n \geq 2$, so ist $[\,[t_1, \ldots, t_m, u_1, \ldots, u_n]_y]_z \in T_z$.

Hierbei entsteht $t = [t_1, \ldots, t_m, u_1, \ldots, u_n]_y$, indem man die Wurzelknoten der Graphen $t_1, \ldots, t_m$, $u_1, \ldots, u_n$ durch $m + n$ Kanten mit einem (neu hinzukommenden) mageren Knoten verbindet. Dieser zusätzliche Knoten ist die *Wurzel* von t, er hat in t keinen Vorgänger, seine Nachfolger sind die Wurzeln der Teilgraphen $t_1, \ldots, t_m$, $u_1, \ldots, u_n$ von t. Entsprechend wird $u = [t_1, \ldots, t_m]_z$ mit einem fetten Wurzelknoten konstruiert ([81, Definition 5.1]). Als *Ordnung* $\varrho(t)$ eines Baumes $t \in T_y \cup T_z$ wird die Differenz zwischen der Anzahl seiner mageren Knoten und der Anzahl seiner fetten Knoten bezeichnet ([81, Definition 5.2]).

Beispiel 20

$t_1 =$ 

, $t_2 =$ 

Die Bäume $t_1 = [\,[\,[\tau, \tau, [\tau]_y]_z]_y, [\tau, \tau]_z]_y \in T_y$ und $t_2 = [\tau, \tau, [\tau]_y, [\tau, \tau]_z]_y \in T_y$ (die hier schon für die spätere Formulierung der Konsistenzbedingungen mit Indizes versehen wurden) haben die Ordnung 6 und stehen in der Taylorentwicklung von $y(t + h)$ in $y^{(VI)}(t)$ für die elementaren Differentiale

$$[f_{yz}(f_z(-g_y f_z)^{-1} g_{yyy}(f, f, f_y f), (-g_y f_z)^{-1} g_{yy}(f, f))](y(t), z(t))$$

bzw.

$$[f_{yyyz}(f, f, f_y f, (-g_y f_z)^{-1} g_{yy}(f, f))](y(t), z(t))\ .$$

Für die Taylorentwicklung der numerischen Lösung $\hat{y}_{n+1}$, $\hat{z}_{n+1}$ aus Definition 6 wird wie in Abschnitt 3.2.1 vorausgesetzt, daß

$$\sum_{j=1}^{\hat{s}} d_j = 1 \ , \quad \gamma_{11} = 0 \ , \quad \gamma_{ii} \neq 0 \ , \ (i = 2, \ldots, \hat{s}) \ , \quad Z_{n1} = z_n$$

gilt. Faßt man in (3.21) die Stufenvektoren Y_{ni}, η_{ni} und Z_{ni} als Funktionen von h auf und definiert $Y'_{ni} := f(Y_{ni}, Z_{ni})$, so folgt durch Differentiation nach h

$$\dot{Y}'_{ni} = f_y(Y_{ni}, Z_{ni})\dot{Y}_{ni} + f_z(Y_{ni}, Z_{ni})\dot{Z}_{ni} \ , \tag{3.68}$$

wobei $\dot{Y}$, $\dot{Z}$, ... die Ableitungen von Y, Z, ... nach h bezeichnen.

Auch hier erhält man eine Darstellung für $\dot{Z}_{ni}$ durch Differentiation der Zwangsbedingungen ($g(\eta_{ni}) = 0$ in (3.21), vgl. [81, S. 59f]): Es gilt $Y'_{ni}(0) = f(y_n, z_n)$ und damit

$$\dot{Y}_{ni}(0) = (\sum\nolimits_j a_{ij}) f(y_n, z_n) \ , \quad \dot{\eta}_{ni}(0) = (\sum\nolimits_j \gamma_{ij}) f(y_n, z_n) \tag{3.69}$$

und

$$\ddot{\eta}_{ni}(0) = 2 \sum\nolimits_j \gamma_{ij} \dot{Y}'_{nj}(0) \ . \tag{3.70}$$

Differenziert man $g(\eta_{ni}) = 0$ zweimal bezüglich h, so folgt

$$0 = g_y(\eta_{ni})\dot{\eta}_{ni}$$

und

$$0 = g_{yy}(\eta_{ni})(\dot{\eta}_{ni}, \dot{\eta}_{ni}) + g_y(\eta_{ni})\ddot{\eta}_{ni} \ .$$

Für $h = 0$ ergibt sich unter Verwendung von (3.68) und (3.70)

$$0 = \psi_i + 2 \sum\nolimits_j \gamma_{ij} [g_y f_z](y_n, z_n) \dot{Z}_{nj}(0)$$

mit $\Psi := (\psi_1, \ldots, \psi_{\hat{s}})^T$ und

$$\psi_i = g_{yy}(y_n)(\dot{\eta}_{ni}(0), \dot{\eta}_{ni}(0)) + 2 \sum\nolimits_j \gamma_{ij} [g_y f_y](y_n, z_n) \dot{Y}_{nj}(0) \ .$$

Faßt man diese Gleichungen für $i = 1, \ldots, \hat{s}$ zusammen, so erhält man mit der in Bemerkung 17c eingeführten Matrix W und mit $\dot{Z} := (\dot{Z}_{n1}, \ldots, \dot{Z}_{n\hat{s}})^T$ das Gleichungssystem

$$0 = \Psi + 2\,(W^{-1} \otimes [g_y f_z](y_n, z_n))\dot{Z}(0) \ , \tag{3.71}$$

denn mit Ausnahme der ersten Zeile enthält die Matrix W^{-1} die Koeffizienten γ_{ij} und wegen $Z_{n1} = z_n = z(t_n)$ ist $\gamma_{11}\dot{Z}_{n1}(0) = 0 = 1 \cdot \dot{Z}_{n1}(0)$. Löst man (3.71) nach $\dot{Z}(0)$ auf und setzt hier (3.69) ein, so folgt $[g_y f_z](y_n, z_n)\dot{Z}_{ni}(0) = -\frac{1}{2}\sum_k w_{ik}\psi_k$ für $i = 1, \ldots, \hat{s}$, also

$$\dot{Z}_{ni}(0) = \frac{1}{2} \sum\nolimits_k w_{ik} (\sum\nolimits_j \gamma_{kj})^2 [(-g_y f_z)^{-1} g_{yy}(f, f)](y_n, z_n) + \\ + \sum\nolimits_{k,j} w_{ik}\gamma_{kj} [(-g_y f_z)^{-1} g_y f_y f](y_n, z_n)$$

als Gegenstück zu (3.67). Durch wiederholte Differentiation von (3.68) und $g(\eta_{ni}) = 0$ erhält man auf diese Weise Darstellungen für die Koeffizienten $\dot{Y}_{ni}(0)$, $\dot{Z}_{ni}(0)$, ... der Taylorentwicklungen

$$\begin{aligned} Y_{ni}(h) &= Y_{ni}(0) + h\dot{Y}_{ni}(0) + \frac{h^2}{2}\ddot{Y}_{ni}(0) + \mathcal{O}(h^3) \ , \\ Z_{ni}(h) &= Z_{ni}(0) + h\dot{Z}_{ni}(0) + \frac{h^2}{2}\ddot{Z}_{ni}(0) + \mathcal{O}(h^3) \end{aligned}$$

und hieraus schließlich die Taylorentwicklungen von $\hat{y}_{n+1}$ und $\hat{z}_{n+1}$. Bis auf die Berücksichtigung der unterschiedlichen Koeffizienten für Y_{ni} und η_{ni} (a_{ij} bzw. γ_{ij}) und der Definition des Stufenvektors Z_{n1} geht man hierbei exakt genauso vor, wie Hairer et al. in [81, Kapitel 5] und [38].

Beim Vergleich der Taylorentwicklungen von analytischer und numerischer Lösung entspricht nun jedem elementaren Differential (und damit jedem der Bäume aus Bemerkung 23) eine Konsistenzbedingung, die nach folgendem Algorithmus aufgestellt werden kann (vgl. auch [81, S. 70], [38, S. 541]):

Algorithmus 1

1. Ordne den Knoten eines gegebenen Baumes der Menge $T_y \cup T_z$ je genau einen Summationsindex $i, j, \ldots$ zu. Ist der Wurzelknoten des Baumes fett (d. h. $u \in T_z$), so ordne dem Baum einen zusätzlichen Summationsindex k zu.

2. Bilde Produkte, die zu jedem Summationsindex genau einen Faktor enthalten. Diese Faktoren sind gegeben durch

   b_i, falls „i“ der Index der Wurzel des Baumes ist und der Wurzelknoten mager ist,

   $d_k w_{ki}$, falls der Baum einen fetten Wurzelknoten hat und „i“ der Index der Wurzel des Baumes sowie „k“ der dem Baum zugeordnete zusätzliche Summationsindex ist,

   a_{ij}, falls der magere Knoten „j“ den mageren Knoten „i“ zum Vorgänger hat,

   γ_{ij}, falls der magere Knoten „j“ den fetten Knoten „i“ zum Vorgänger hat,

   w_{ij}, falls der fette Knoten „j“ den mageren Knoten „i“ zum Vorgänger hat.

   Σ sei die über sämtliche Summationsindizes genommene Summe dieser Produkte.

3. Bilde das Produkt Π von rationalen Zahlen, das zu jedem Knoten des Baumes genau einen Faktor enthält: Dieser Faktor beträgt $1/r$ (für einen mageren Knoten) bzw. $r + 1$ (für einen fetten Knoten), wobei r die Ordnung desjenigen Teilgraphen des Baumes bezeichnet, für den der jeweilige Knoten die Wurzel ist.

4. Ist die Konsistenzbedingung $\Sigma = \Pi$ erfüllt, so stimmen in den Taylorentwicklungen der analytischen und der numerischen Lösung die Koeffizienten des dem Baum zugeordneten elementaren Differentials überein.

Das Baummodell eröffnet einen systematischen Weg, die Voraussetzungen (3.31) im Konvergenzsatz für HERK–Verfahren als Konsistenzbedingungen an die Verfahrensparameter zu formulieren:

Satz 16 *Gegeben sei ein HERK–Verfahren (3.21) mit*

$$\sum_j d_j = 1\ ,\quad Z_{n1} = z_n\ ,\quad \gamma_{11} = 0\ ,\quad \gamma_{ii} \neq 0\ ,\ (i = 2,\ldots,\hat{s})\ \text{ und }\ \gamma_{sj} = b_j\ ,\ (j = 1,\ldots,s)\ .$$

Für den lokalen Fehler des Verfahrens gilt

$$\delta y_h(t_n) = \mathcal{O}(h^{q_y})\ ,\quad P(t_n)\delta y_h(t_n) = \mathcal{O}(h^{q_y+1})\ ,\quad \delta z_h(t_n) = \mathcal{O}(h^{q_z+1})\ ,$$

wenn die Konsistenzbedingungen nach Algorithmus 1 erfüllt sind für alle Bäume $t \in T_y$ *der Ordnung* $\varrho(t) \leq q_y - 1$ *, für alle Bäume* $u \in T_z$ *mit* $\varrho(u) \leq q_z$ *sowie zusätzlich für alle Bäume* $t \in T_y$ *der Ordnung* $\varrho(t) = q_y$ *, die sich nicht als* $t = [u]_y$ *mit* $u \in T_z$ *schreiben lassen.*

Beweis Die Behauptung ergibt sich durch Vergleich der Taylorentwicklungen von analytischer und numerischer Lösung (vgl. [81, Satz 5.8]). Einzige Besonderheit sind die Konsistenzbedingungen zu Bäumen $t \in T_y$ der Form $t = [u]_y$ mit $u \in T_z$ und $\varrho(t) = q_y$: Sind $t_1,\ldots,t_m \in T_y$ und $u := [t_1,\ldots,t_m]_z \in T_z$, so hat die nach Algorithmus 1 aufgestellte Konsistenzbedingung zum Baum $[u]_y$ die Form $\sum_{i,j} b_i w_{ij}\phi_j = \Pi$ mit geeignet definierten ϕ_j. Wegen $\gamma_{sj} = b_j$, $(j = 1,\ldots,s)$ ist $\sum_i b_i w_{ij} = \delta_{sj}$ mit dem Kroneckerschen Delta δ_{sj}, also $\sum_{i,j} b_i w_{ij}\phi_j = \phi_s$. Man überprüft, daß die Konsistenzbedingung zu $[u]_y$ bereits dann erfüllt ist, wenn die Konsistenzbedingungen zu $t_1,\ldots,t_m$ erfüllt sind. Hieraus folgt die Behauptung, denn nach Bemerkung 23 gilt für $u := [t_1,\ldots,t_m]_z \in T_z$ entweder $m \geq 2$ (und damit $\varrho(t_i) < \varrho([u]_y) = q_y$, $i = 1,\ldots,m$) oder $u = [t']_z$ mit einem Baum $t' \in T_y$ der Ordnung $\varrho(t') = \varrho([u]_y) = q_y$, der sich nicht in der Form $t' = [u']_y$ mit einem $u' \in T_z$ darstellen läßt. ■

Bemerkung 24 a) Bäume mit ausschließlich mageren Knoten führen auf die aus der Theorie der gewöhnlichen Differentialgleichungen bekannten Konsistenzbedingungen für explizite Runge–Kutta–Verfahren.

b) Ähnlich wie im Beweis von Satz 16 zeigt man, daß beim Aufstellen der Konsistenzbedingungen weitere Bäume unbeachtet bleiben können, weil sie keine neuen Bedingungen ergeben. Dies trifft unter den Voraussetzungen des Satzes auf alle Bäume zu, die einen fetten Knoten enthalten, der nicht der Wurzelknoten des Baumes ist und der genau einen Nachfolger hat, denn für alle $k \geq 2$ gilt $\sum_j w_{ij}\gamma_{jk} = \delta_{ik}$.
Sind zusätzlich für $i \geq 2$ die Parameter γ_{ij} durch $\gamma_{ij} = a_{i+1,j}$, $(j = 1,\ldots,i)$ gegeben (vgl. (3.62) und [21]), so entfallen wegen $\sum_j a_{ij}w_{jk} = \delta_{i,k+1}$, $(k \geq 2)$ außerdem alle Bäume, die einen fetten Knoten enthalten, der

- einziger Nachfolger seines Vorgängers ist und
- Wurzel eines Teilgraphen $u = [t_1,\ldots,t_m]_z \in T_z$ dieses Baumes ist, in dem $t_{i_0} \neq \tau$ für mindestens ein $i_0 \in \{1,\ldots,m\}$ gilt.

(Betrachte als Beispiel den mit „k“ indizierten Knoten im Baum t_1 aus Beispiel 20; hier ist $u = [\tau,\tau,[\tau]_y]_z$ und $i_0 = m = 3$.)

Im nachfolgenden Beispiel werden einige Konsistenzbedingungen aufgeführt. Sofern nichts anderes angegeben ist, ist stets über alle auftretenden Indizes zu summieren. Neben

Tabelle 3.2: Konsistenzbedingungen für HERK–Verfahren mit expliziter Stufe und $\gamma_{sj} = b_j$, $(j = 1,\ldots,s)$.

Nr.	Baum	Ordnung	Konsistenzbedingung
1_y		1	$\sum b_i = 1$
2_y		2	$\sum b_i c_i = \frac{1}{2}$
3_y		3	$\sum b_i c_i^2 = \frac{1}{3}$
4_y		3	$\sum b_i a_{ij} c_j = \frac{1}{6}$
5_y		3	$\sum b_i c_i w_{ij} \hat{c}_{j+1}^2 = \frac{2}{3}$
6_y		3	$\sum b_i w_{ij} \hat{c}_{j+1}^2 w_{ik} \hat{c}_{k+1}^2 = \frac{4}{3}$
7_y		3	$\sum b_i a_{ij} w_{jk} \hat{c}_{k+1}^2 = \frac{1}{3}$
1_z		1	$\sum d_i w_{ij} \hat{c}_{j+1}^2 = 2$
2_z		1	$\sum d_i w_{ij} \gamma_{jk} c_k = 1 \quad \Leftrightarrow \quad \sum d_i c_i = 1$

der Bezeichnung $c_i = \sum_j a_{ij}$ wird $\hat{c}_{i+1} := \sum_{j=1}^{i} \gamma_{ij}$ verwendet, insbesondere gilt für die Verfahren, deren Koeffizienten die Bedingungen (3.62) und $\sum_j b_j = 1$ erfüllen,

$$\hat{c}_2 = 0\ ,\quad \hat{c}_i = c_i\ ,\ (i = 3,\ldots,\hat{s}) \quad \text{und} \quad \hat{c}_{s+1} = 1\ .$$

Beispiel 21 a) Als Konsistenzbedingung für den Baum t_1 aus Beispiel 20 erhält man

$$\sum b_i a_{ij} w_{jk} \gamma_{kl} \gamma_{km} \gamma_{kn} a_{no} w_{ip} \gamma_{pq} \gamma_{pr} = \frac{1}{2} \cdot 4 \cdot \frac{1}{4} \cdot 2 \cdot \frac{1}{6} = \frac{1}{6}\ .$$

Für Verfahren mit Koeffizienten nach (3.62) ergibt sich wegen

$$\sum_{j,k,l,m,n,o} a_{ij} w_{jk} \gamma_{kl} \gamma_{km} \gamma_{kn} a_{no} = \sum_{j,k,n} a_{ij} w_{jk} \hat{c}_{k+1}^2 \gamma_{kn} c_n = \sum_{k,n} \delta_{i,k+1} \hat{c}_{k+1}^2 a_{k+1,n} c_n = c_i^2 \sum_n a_{in} c_n$$

dieselbe Konsistenzbedingung $\sum b_i c_i^2 a_{in} c_n w_{ip} \hat{c}_{p+1}^2 = \frac{1}{6}$ wie für Baum t_2 aus Beispiel 20.
b) In Tab. 3.2 sind alle Konsistenzbedingungen zu Satz 16 mit $q_y = 3$, $q_z = 1$ zusammengefaßt.

Halb-explizite Runge–Kutta–Verfahren der Konvergenzordnung $q \leq 5$

Mit steigender Ordnung q_y, q_z wächst die Zahl der Konsistenzbedingungen sehr schnell an, für die Konstruktion von Verfahren werden deshalb *vereinfachende Bedingungen* an die Koeffizienten b_j, a_{ij}, γ_{ij} gestellt.

Bemerkung 25 Bei der Verfahrenskonstruktion geht man von einem expliziten Runge–Kutta–Verfahren aus, dem ggf. eine Stufe hinzugefügt wird, um die Konsistenzbedingungen für z zu erfüllen ([118], [21]). Es dient der Übersichtlichkeit, die Stufenzahl des expliziten Runge–Kutta–Verfahrens wie üblich mit „s“ zu bezeichnen. Ein HERK–Verfahren mit zusätzlicher Stufe hat dann also $\hat{s} = s+1$ Stufen, in (3.21) ist $a_{s+1,j} = b_j$, $(j = 1, \ldots, s)$, d. h. $Y_{n,s+1} = y_{n+1}$. Man beachte, daß für $\hat{s} > s$ der Funktionswert $f(Y_{n\hat{s}}, Z_{n\hat{s}})$ nicht in die Berechnung von y_{n+1} eingeht, denn $b_{s+1} = 0$. Wählt man die Parameter d_j in einem Verfahren mit zusätzlicher Stufe als $d_j = \delta_{j,s+1}$, $(j = 1, \ldots, s+1)$, so ist $y_{n+1} = Y_{n,s+1}$, $z_{n+1} = Z_{n,s+1}$ und der bei der Berechnung von $g(\eta_{n,s+1})$ auszuwertende Funktionswert $f(Y_{n,s+1}, Z_{n,s+1})$ kann als $f(y_{n+1}, z_{n+1})$ in der ersten Stufe des nachfolgenden Integrationsschritts noch einmal verwendet werden (FSAL: „first same as last“ [82, S. 167]).

Nun beschränken wir uns auf die in [21] betrachteten Verfahren mit (3.62), d. h.

$$\gamma_{11} = 0\,,\quad \gamma_{ii} \neq 0\,,\ (i = 2, \ldots, \hat{s})\,,\quad \gamma_{ij} = a_{i+1,j}\,,\ (i = 1, \ldots, \hat{s},\ j = 1, \ldots, i) \tag{3.72}$$

mit $\hat{s} = s$ bzw. (für Verfahren mit zusätzlicher Stufe) $\hat{s} = s+1$. Für die einheitliche Darstellung der vereinfachenden Bedingungen wird hier $a_{s+1,j} := b_j$, $(j = 1, \ldots, s+1)$ gesetzt; im Fall $\hat{s} = s+1 > s$ sind $a_{s+2,j} := \gamma_{\hat{s}j}$, $(j = 1, \ldots, s+1)$ geeignet zu bestimmende neue Parameter des Verfahrens. Ein solches Verfahren erfüllt die vereinfachenden Bedingungen C(r) und D(1), wenn gilt:

$$\mathrm{C}(r)\ :\quad \sum_{j=1}^{i-1} a_{ij} c_j^l = \frac{1}{l+1} c_i^{l+1}\quad,\quad (i = 3, \ldots, \hat{s}+1,\ l = 0, \ldots, r-1)\,,$$

$$\mathrm{D}(1)\ :\quad \sum_{j=i+1}^{s} b_j a_{ji} = b_i(1 - c_i)\quad,\quad (i = 1, \ldots, s)\,.$$

(Die Bedingung C(1) ist wegen $\sum_j a_{ij} = c_i$ stets erfüllt.)

Bemerkung 26 Ist (3.72) und die Bedingung C(r) erfüllt, so gilt $\sum_j \gamma_{ij} c_j^l = \hat{c}_{i+1}^{l+1}/(l+1)$, $(i = 2, \ldots, \hat{s},\ l = 0, 1, \ldots, r-1)$. Wegen $1 \cdot c_1 = 0 = \hat{c}_2^{l+1}/(l+1)$ können diese Gleichungen mit den Bezeichnungen $c^l := (c_1^l, \ldots, c_s^l)^T$, $\hat{c}_+^{l+1} := (\hat{c}_2^{l+1}, \ldots, \hat{c}_{s+1}^{l+1})^T$ zusammengefaßt werden zu $W^{-1} c^l = \hat{c}_+^{l+1}/(l+1)$, denn mit Ausnahme der ersten Zeile enthält die Matrix W^{-1} die Parameter γ_{ij}. Multipliziert man dieses Gleichungssystem mit W, so folgt aus C(r) die *reziproke* Bedingung ([36])

$$\mathrm{C}(r)\mathrm{R}\ :\quad \sum_{i=1}^{j} w_{ji} \hat{c}_{i+1}^{l+1} = (l+1) c_j^l\,,\quad (j = 1, \ldots, \hat{s},\ l = 0, 1, \ldots, r-1)\,.$$

Zu einem Baum $t \in T_y \cup T_z$ treten in den nach Algorithmus 1 aufgestellten Konsistenzbedingungen Ausdrücke der Form $\sum_i w_{ji} \hat{c}_{i+1}^2$ auf, wenn t den Baum $u_1 = [\tau, \tau]_z$ aus Abb. 3.4 als Teilgraphen enthält. Dabei führt der Baum t_3 aus Abb. 3.4 für jeden beliebigen Rumpf $t^* \in T_y \cup T_z$ auf dieselbe Konsistenzbedingung wie t_4, wenn das Verfahren die Bedingung C(2) (und damit auch C(2)R) erfüllt. In Tab. 3.2 entfallen für diese Verfahren die Bedingungen 5_y, 6_y und 7_y. Allgemein brauchen für ein Verfahren, das die

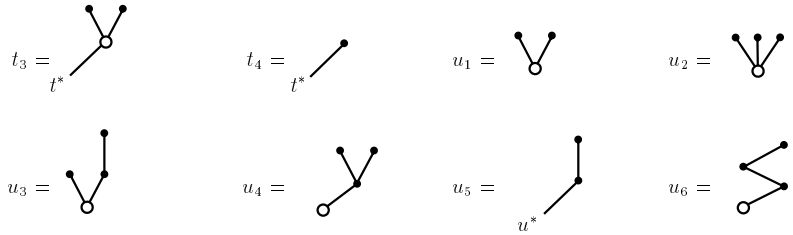
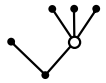


Abbildung 3.4: Bäume und vereinfachende Bedingungen (vgl. Bemerkung 26).

Bedingung C(r) erfüllt, alle diejenigen Bäume nicht betrachtet zu werden, die einen „buschigen“ Baum $[\tau, \tau, \ldots, \tau]_z \in T_z$ der Ordnung $\varrho(u) < r$ (echt) enthalten. Als Beispiele zeigt Abb. 3.4 die buschigen Bäume der Ordnung 1 und 2 (u_1, u_2).

Satz 17 *Mit den Parametern γ_{ij} nach (3.72) führt ein explizites Runge–Kutta–Verfahren der Ordnung p mit $2 \le p \le 5$ auf ein HERK–Verfahren mit lokalem Diskretisierungsfehler $\delta y_h(t) = \mathcal{O}(h^{p+1})$, wenn die vereinfachenden Bedingungen* C(2) *(für $p = 3$),* C(2) *und* D(1) *(für $p = 4$) bzw.* C(3) *und* D(1) *(für $p = 5$) erfüllt sind.*

Beweis Für $p = 3$ folgt die Behauptung aus Tab. 3.2 und Bemerkung 26.
Für $p = 4$ zeigt man für Verfahren, die die vereinfachende Bedingung C(2) erfüllen, durch vollständige Fallunterscheidung, daß neben den klassischen Konsistenzbedingungen an das explizite Runge–Kutta–Verfahren nur eine Konsistenzbedingung verbleibt:



$$\sum\nolimits_{i,j} b_i c_i w_{ij} \hat{c}_{j+1}^3 = \frac{3}{4}\,. \tag{3.73}$$

Für diese Verfahren gilt $0 = \frac{1}{3} - \frac{1}{3} = \sum_i b_i c_i^2 - 2\sum_{i,j} b_i a_{ij} c_j = \sum_i b_i (c_i^2 - 2\sum_j a_{ij} c_j) = b_2 c_2^2$, also $b_2 = 0$. Wie in Bemerkung 24b folgt deshalb unter Verwendung von D(1), daß auch die Konsistenzbedingung (3.73) erfüllt ist:

$$\begin{aligned}\sum\nolimits_{i,j} b_i c_i w_{ij} \hat{c}_{j+1}^3 &= \sum\nolimits_{i,j} b_i w_{ij} \hat{c}_{j+1}^3 - \sum\nolimits_{i,j,k} b_k a_{ki} w_{ij} \hat{c}_{j+1}^3 \\ &= \sum\nolimits_{j} \delta_{sj} \hat{c}_{j+1}^3 - \sum\nolimits_{j,k} b_k \delta_{k,j+1} \hat{c}_{j+1}^3 = \hat{c}_{s+1}^3 - \sum\nolimits_{k} b_k c_k^3 = 1 - \frac{1}{4} = \frac{3}{4}\,.\end{aligned}$$

$p = 5$: Ist C(3) erfüllt, so gilt $\hat{c}_{j+1} \sum_k a_{j+1,k} c_k = \frac{1}{2} \hat{c}_{j+1}^3$ für alle $j = 1, \ldots, \hat{s}$. Wie in Bemerkung 26 zeigt man, daß deshalb Bäume $t \in T_y \cup T_z$, die einen der Bäume u_1, u_2, u_3 aus Abb. 3.4 (echt) enthalten, keine neuen Konsistenzbedingungen ergeben. Nutzt man dies aus, so folgt die Behauptung für $p = 5$ ebenso, wie zuvor für $p = 4$ gezeigt. ∎

Die Bedingungen an den lokalen Fehler $\delta y_h(t)$ sind also automatisch erfüllt, wenn man von einem geeigneten expliziten Runge–Kutta–Verfahren ausgeht. Die neu hinzukommenden Parameter d_j und (gegebenenfalls) $a_{s+2,j}$ werden so gewählt, daß auch die Bedingungen an $\delta z_h(t)$ erfüllt sind.

Satz 18 *Mit den Parametern γ_{ij} nach (3.72) hat ein HERK–Verfahren, das die vereinfachende Bedingung* C(q_z) *mit einem* $q_z \in \{1,2,3\}$ *erfüllt, einen lokalen Diskretisierungsfehler* $\delta z_h(t) = \mathcal{O}(h^{q_z+1})$ *, wenn gilt:*

$$\sum_{i=1}^{\hat{s}} d_i c_i^l = 1 \ , \quad (l = 0,1,\ldots,q_z) \ , \tag{3.74}$$

$$d_2 = 0 \ \ (\textit{für} \ \ q_z \geq 2 \ ) \ \ \textit{und} \quad \sum_{i=1}^{\hat{s}} d_i a_{i2} = \sum_{i,j=1}^{\hat{s}} d_i w_{ij} c_{j+1} a_{j+1,2} = 0 \ \ (\textit{für} \ \ q_z \geq 3 \ ) \tag{3.75}$$

und

$$\frac{1}{q_z+1} \sum_{i,j=1}^{\hat{s}} d_i w_{ij} \hat{c}_{j+1}^{q_z+1} = 1 \ . \tag{3.76}$$

Beweis Die Konsistenzbedingungen (3.74) entsprechen den Bäumen $[[\tau,\ldots,\tau]_y]_z$, so ist z. B. für $u_4 = [[\tau,\tau]_y]_z$ aus Abb. 3.4

$$1 = \sum\nolimits_{i,j,k} d_i w_{ij} \gamma_{jk} c_k^2 = \sum\nolimits_{i,j,k} d_i w_{ij} a_{j+1,k} c_k^2 = \sum\nolimits_{i,k} d_i \delta_{ik} c_k^2 = \sum\nolimits_i d_i c_i^2 \ .$$

Die Bedingungen (3.75) garantieren, daß die Konsistenzbedingungen zu beliebigen Bäumen der Form u_5 aus Abb. 3.4 mit $\varrho(u_5) \leq q_z$ erfüllt sind, wenn C(q_z) und (3.74) gelten. Als Beispiel betrachte für $q_z \geq 2$ den Baum u_6 aus Abb. 3.4:

$$\sum\nolimits_{i,j} d_i a_{ij} c_j = \frac{1}{2} \sum\nolimits_i d_i c_i^2 + \sum\nolimits_{i \geq 3} d_i \Big(\sum\nolimits_j a_{ij} c_j - \frac{1}{2} c_i^2 \Big) - \frac{1}{2} d_2 c_2^2 = \frac{1}{2} \sum\nolimits_i d_i c_i^2 = \frac{1}{2} \ .$$

Die Konsistenzbedingungen für die buschigen Bäume $[\tau,\ldots,\tau]_z$ (z. B. u_1, u_2 aus Abb. 3.4) lauten: $\sum_{i,j} d_i w_{ij} \hat{c}_{j+1}^{l+1} = l+1$, $(l = 1,\ldots,q_z)$. Nach Bemerkung 26 sind diese Bedingungen für $l < q_z$ äquivalent zu (3.74), wenn C(q_z) erfüllt ist; im Fall $l = q_z$ ergibt sich die Bedingung (3.76).
Durch vollständige Fallunterscheidung (die dem Computer übertragen werden kann) überzeugt man sich, daß alle anderen Bäume $u \in T_z$ mit $\varrho(u) \leq q_z$ einen der Bäume u_1, u_2, u_3 aus Abb. 3.4 echt enthalten. Wie in Satz 17 folgt, daß keiner dieser Bäume auf eine neue Konsistenzbedingung führt. ∎

Bemerkung 27 a) Erfüllt ein HERK–Verfahren neben den Voraussetzungen von Satz 18 die vereinfachende Bedingung C(r) mit $r = q_z + 1$, so ist Bedingung (3.76) wegen der reziproken Bedingung C(r)R äquivalent zu (3.74) mit $l = q_z$ (vgl. Bemerkung 26).

b) Die Bedingungen in Satz 18 vereinfachen sich für HERK–Verfahren, die unter Verwendung der FSAL–Technik durch Anfügen einer zusätzlichen Stufe aus einem s-stufigen expliziten Runge–Kutta–Verfahren der Ordnung $p \geq 1$ hervorgehen (vgl. Bemerkung 25). Ist $\hat{s} = s+1 \geq 3$ und $d_j = \delta_{j,s+1}$ (d. h. $z_{n+1} = Z_{n,s+1}$ und $\sum_j d_j c_j^l = c_{s+1}^l = 1$), so gilt für den lokalen Fehler $\delta z_h(t) = \mathcal{O}(h^{q_z+1})$ mit einem $q_z \in \{1,2,3\}$, wenn die Bedingungen C(q_z) und $\sum_j w_{s+1,j} \hat{c}_{j+1}^{q_z+1} = q_z + 1$ und zusätzlich für $q_z = 3$ die Bedingungen $b_2 = 0$ und $\sum_j w_{s+1,j} c_{j+1} a_{j+1,2} = 0$ erfüllt sind.

Die Sätze 12, 17 und 18 geben Kriterien zur Konstruktion von HERK–Verfahren der Konvergenzordnung q in y und $q-1$ in z für $q \leq 5$ an. Neben den Konsistenzbedingungen ist dabei die Kontraktivitätsbedingung $|\sum_j d_j w_{j1}| < 1$ zu erfüllen. Bei Ausnutzung der FSAL–Technik benötigen die Verfahren pro Integrationsschritt im wesentlichen ebenso viele Aufrufe der rechten Seite f, um eine gegebene Ordnung q zu erreichen, wie explizite Runge–Kutta–Verfahren für gewöhnliche Differentialgleichungen. Sie sind damit deutlich effektiver als die nach dem klassischen Ansatz aus [81] konstruierten Verfahren (Tab. 3.3).

Tabelle 3.3: Vergleich des numerischen Aufwands (pro Integrationsschritt) für explizite Runge–Kutta–Verfahren, HERK–Verfahren nach [81] und HERK–Verfahren mit expliziter Stufe.

	$q=2$	$q=3$	$q=4$	$q=5$
explizite Runge–Kutta–Verfahren				
Aufrufe der rechten Seite	2	3	4	6
HERK–Verfahren mit (3.63)				
Aufrufe von f	2	3	5	8
Gleichungssysteme (3.22)	2	3	5	8
HERK–Verfahren mit (3.72)				
Aufrufe von f	2	3	4	6
Gleichungssysteme (3.22)	1	2	4	6

Beispiel 22 a) $q = 2$. Zu einem gegebenen expliziten Runge–Kutta–Verfahren der Ordnung $p = s = 2$ wählt man Parameter γ_{ij} nach (3.72) mit $\hat{s} := s = 2$ und setzt $d_j := b_j$, $(j = 1,2)$. Dieses HERK–Verfahren konvergiert mit der Ordnung $q = 2$ in y (und mit der Ordnung $q-1$ in z), denn $\sum_j d_j = \sum_j b_j = 1$ und $\sum_j d_j w_{j1} = \sum_j b_j w_{j1} = \delta_{21} = 0$.

b) $q = 3$. Ein explizites Runge–Kutta–Verfahren der Ordnung $p = s = 3$, das C(2) erfüllt, hat Parameter $b_2 = 0$ (denn $b_3 a_{32} c_2 = \frac{1}{6}$ und $\sum_i b_i c_i^2 = \frac{1}{3}$), $c_3 = \frac{2}{3}$, $b_1 = \frac{1}{4}$, $b_3 = \frac{3}{4}$ (denn $\sum_i b_i c_i^l = \frac{1}{l+1}$, $l = 0,1,2$) und $a_{32} = \frac{2}{9c_2}$ (denn $b_3 a_{32} c_2 = \frac{1}{6}$). Der Parameter $c_2 \neq 0$ bleibt frei wählbar ([82, Kapitel II.1]). Mit $\hat{s} := s = 3$, $q_y = 3$, $q_z = 1$ ergeben die Bedingungen (3.74) und $\sum_j d_j w_{j1} = 0$ (vgl. (3.30)) ein System aus 3 linearen Gleichungen in d_1, d_2, d_3, das die eindeutig bestimmte Lösung $d_1 = -2 + \frac{1}{c_2}$, $d_2 = -\frac{1}{c_2}$, $d_3 = 3$ hat. Mit diesen Parametern und mit (3.72) ergibt sich ein HERK–Verfahren der Ordnung $q = p = 3$ in y und der Ordnung $q-1$ in z (vgl. Satz 18 und Bemerkung 27a).

c) $q = 4$. Für ein explizites Runge–Kutta–Verfahren der Ordnung $p = s = 4$ ist D(1) erfüllt und damit auch $c_4 = 1$ (denn $b_4(1-c_4) = \sum b_j a_{j4} = 0$). Die Verfahren, die C(2) erfüllen, bilden eine Familie mit c_2 als freiem Parameter und $b_2 = \sum b_j a_{j2} = 0$, $c_3 = \frac{1}{2}$ ([82, S. 138]). Für $\hat{s} := s = 4$ führen C(2)R und die Konsistenzbedingungen (3.74) und (3.76) mit $q_z = 2$ auf

$$\beta_1 + \sum_{i=2}^{4} \beta_i \hat{c}_{i+1} = 1 \ , \quad \sum_{i=1}^{4} \beta_i \hat{c}_{i+1}^2 = 2 \ , \quad \sum_{i=1}^{4} \beta_i \hat{c}_{i+1}^3 = 3 \tag{3.77}$$

mit $\beta_i := \sum_j d_j w_{ji}$, $(i = 1, \ldots, 4)$, denn

$$w_{j1} + \sum_{i\geq 2} w_{ji}\hat{c}_{i+1} = w_{j1} \cdot (\sum\nolimits_k \delta_{1k}) + \sum_{i\geq 2} w_{ji}(\sum\nolimits_k \gamma_{ik}) = \sum\nolimits_k \delta_{jk} = 1$$

nach Definition von W. Die Lösungen von (3.77) sind $\beta_1 = 1$, $\beta_2 = -8$, $\beta_3 + \beta_4 = 4$ (beachte $c_4 = c_5 = 1$). Wegen $\beta_1 = \sum_j d_j w_{j1}$ kann deshalb ein HERK–Verfahren mit $\hat{s} = s = 4$, (3.72), $P(t)\delta y_h(t) = \mathcal{O}(h^5)$ und $\delta z_h(t) = \mathcal{O}(h^3)$ nicht die Kontraktivitätsbedingung $|\sum_j d_j w_{j1}| < 1$ erfüllen. Ebensowenig läßt sich die Kontraktivitätsbedingung erfüllen, wenn man eine zusätzliche Stufe mit $c_6 = \sum_j a_{6j} = 1$ einführt.
Unter Ausnutzung der FSAL–Strategie (vgl. Bemerkung 25) wird deshalb eine Stufe mit $c_6 \neq 1$ hinzugefügt. Neben $d_j = \delta_{5j}$ werden Parameter $a_{6j} = \gamma_{5j}$ bestimmt, für die gilt

$$\sum\nolimits_j a_{6j}c_j^l = \frac{1}{l+1}c_6^{l+1} \ , \ (l = 0, 1) \ , \quad \sum\nolimits_j w_{5j}\hat{c}_{j+1}^3 = 3 \ , \quad w_{51} = 0 \ . \tag{3.78}$$

Dann ist $\sum_j d_j w_{j1} = 0$ und nach Satz 12 und Bemerkung 27b konvergiert das Verfahren mit der Ordnung $q = p = 4$ in y und mit der Ordnung $q - 1$ in z.
Die Gleichungen (3.78) haben für jedes $c_6 \notin \{0, \frac{1}{2}, 1\}$ eine Lösung mit $a_{65} \neq 0$ (d. h. $\gamma_{55} \neq 0$), die Parameter a_{62} und c_6 sind dabei frei wählbar. Da die Jacobimatrix Φ_ζ in (3.22) für $i = 5$ die Form $ha_{65}[g_y f_z](y(t_n), z(t_n)) + \mathcal{O}(h)$ hat, wählen wir c_6 so, daß $|a_{65}|$ möglichst groß wird: $c_6 := \frac{1}{2} - \frac{1}{6}\sqrt{3}$. Setzt man zur Vereinfachung $a_{62} := b_2 = 0$, so folgt

$$a_{61} = \frac{1}{6} - \frac{1}{108}\sqrt{3} \ , \quad a_{63} = \frac{1}{3} - \frac{4}{27}\sqrt{3} \ , \quad a_{64} = -\frac{7}{108}\sqrt{3} \ , \quad a_{65} = \frac{1}{18}\sqrt{3} \ .$$

d) $q = 5$. Ein explizites Runge–Kutta–Verfahren der Ordnung $p = 5$ hat $s \geq 6$ Stufen. Wir beschränken uns hier auf das Verfahren von Dormand und Prince 5. Ordnung, das eines der effizientesten Integrationsverfahren für nicht-steife gewöhnliche Differentialgleichungen ist ([82, S. 178f]). Es hat $s = 6$ Stufen und erfüllt die Bedingungen C(3) und D(1). Neben der geringen Stufenzahl sind vor allem die sehr gute Schrittweitensteuerung durch ein eingebettetes Verfahren 4. Ordnung und die sehr kleinen Koeffizienten im führenden Fehlerterm hervorzuheben.
Wählt man zum Verfahren von Dormand und Prince die Parameter γ_{ij} nach (3.72), so kann man wie bei der Untersuchung des lokalen Fehlers durch Taylorentwicklung zeigen, daß die Stufenvektoren die Bedingungen (3.42) aus Folgerung 3 erfüllen. Das Verfahren erfüllt außerdem die Voraussetzungen (3.41) in Folgerung 3, denn wegen C(3) ist $b_2 = 0$ und für $k > 2$ ist wegen (3.72) $\sum_j a_{kj} w_{j1} = 0$, also folgt $\sum_{j,k} b_k a_{kj} w_{j1} = 0$ und deshalb in (3.41) auch $\sum_j b_j c_j w_{j1} = 0$, weil D(1) gilt.
Nach Folgerung 3 ist prinzipiell also $\delta z_h(t) = \mathcal{O}(h^{q_z+1})$ mit $q_z = 2$ ausreichend, um Konvergenz mit der Ordnung $q = 5$ für y zu erreichen. Die auf diese Weise konstruierten Verfahren sind aber in praxi nicht brauchbar, weil die Koeffizienten des führenden Fehlerterms in $\delta z_h(t)$ sehr viel größer als beim Verfahren von Dormand und Prince sind und $\|y_m - y(t_m)\|$ einen Fehlerterm $\mathcal{O}(h^2)\max_{n\leq m}\|\delta z_h(t_n)\|$ enthält (vgl. (3.43)).
Statt dessen wird für die Konstruktion des HERK–Verfahrens HEDOP5 dem Verfahren von Dormand und Prince unter Ausnutzung der FSAL–Technik eine 7. Stufe hinzugefügt, deren Koeffizienten $a_{8j} = \gamma_{7j}$, $(j = 1, \ldots, 7)$ nach Bemerkung 27b mit $q_z = 3$ bestimmt

Tabelle 3.4: Parameter a_{ij}, b_j des HERK–Verfahrens HEDOP5 (vgl. auch [82, Tab. II.5.2]), γ_{ij} ergibt sich aus (3.72).

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
$\frac{19}{20}$	$-\frac{18611506045861}{19738176307200}$	$\frac{59332529}{14479296}$	$-\frac{2509441598627}{893904224850}$	a_{84}	a_{85}	$\frac{46310205821}{287848404480}$	$-\frac{3280}{75413}$

$$a_{84} := \tfrac{2763523204159}{3289696051200} \ , \quad a_{85} := -\tfrac{41262869588913}{116235927142400}$$

werden ($d_j = \delta_{7j}$). Der Fehlerterm $\mathcal{O}(h^2)\max_{n\leq m}\|\delta z_h(t_n)\|$ in (3.43) ist dann gegenüber dem Fehlerterm 5. Ordnung vernachlässigbar. Zusammen mit C(3) und der Bedingung, die $|\sum_j d_j w_{j1}| < 1$ garantiert, ergibt sich das Gleichungssystem

$$\sum_j a_{8j}c_j^l = \frac{1}{l+1}c_8^{l+1} \ , \ (l = 0, 1, 2) \ , \ \sum_j w_{7j}c_{j+1}a_{j+1,2} = 0 \ , \ \sum_j w_{7j}\hat{c}_{j+1}^4 = 4 \ , \ w_{71} = 0 \ , \tag{3.79}$$

das durch Multiplikation der letzten 3 Gleichungen mit $a_{87} = 1/w_{77}$ in ein *lineares* Gleichungssystem in a_{81}, a_{82}, ..., a_{87} umgeformt werden kann, denn wegen $\gamma_{7j} = a_{8j}$, $(j = 1, \ldots, 7)$ gilt

$$a_{87}w_{7j} = \sum_{i=1}^{7} a_{8i}w_{ij} - \sum_{i=1}^{6} a_{8i}w_{ij} = -\sum_{i=1}^{6} a_{8i}w_{ij} \ , \quad (j = 1(1)6) \ .$$

und in W hängen die Elemente w_{ij} mit Zeilenindex $i \leq 6$ nicht von a_{8k}, $(k = 1, \ldots, 7)$ ab. Zu beliebig vorgegebenen a_{87}, c_8 mit $c_8 \neq 1$ ist (3.79) eindeutig lösbar. Bei der Festlegung von a_{87} und c_8 fordert man einerseits, daß die Koeffizienten des führenden Fehlerterms von $\delta z_h(t)$ (im quadratischen Mittel) möglichst klein sind, und andererseits, daß die Parameter a_{8j}, $(j = 1, \ldots, 7)$ und $1/a_{87}$ betragsmäßig nicht zu groß werden. Dabei erweist sich $c_8 = 19/20$, $a_{87} = -3280/75413$ als guter Kompromiß zwischen beiden Forderungen. Damit ergibt sich ein 7-stufiges HERK–Verfahren der Ordnung $q = p = 5$ in y und der Ordnung $q - 1$ in z, dessen Parameter in Tab. 3.4 im erweiterten Butcher–Schema dargestellt sind. Die hier vorgeschlagene Wahl der Verfahrensparameter d_j, a_{8j} erlaubt die Übertragung und Anpassung der Schrittweitensteuerung nach Dormand und Prince an das halbexplizite Verfahren (vgl. Abschnitt 3.3.2). Leider ist es nicht möglich, durch Hinzunahme einer einzigen Stufe sowohl eine eingebettete Lösung $\tilde{y}_{n+1} = y_n + h\sum_j \tilde{b}_j f(Y_{nj}, Z_{nj})$ von 4. Ordnung zur Schrittweitensteuerung als auch eine Näherung $z_{n+1} = \sum_j d_j Z_{nj}$ mit $\delta z_h(t) = \mathcal{O}(h^4)$ zu erhalten ([18, Beispiel 2e]).

Vergleichsrechnungen

HERK–Verfahren für Index-2-Systeme werden vor allem zur dynamischen Simulation von MKS eingesetzt. Vor der ausführlichen Diskussion dieser Anwendung in Abschnitt 3.3.2 sollen zunächst — ohne alle Details der Implementierung — die hier konstruierten HERK–Verfahren mit expliziter Stufe mit aus der Literatur bekannten Verfahren verglichen werden. Hierzu wird für zwei bekannte nicht-steife Benchmark–Probleme die Index-2-Formulierung der MKS–Bewegungsgleichungen mit fixierter Schrittweite h (und ohne Projektionsschritte) integriert. Die Abb. 3.5 und 3.6 zeigen für sehr viele verschiedene Schrittweiten h den numerischen Aufwand, der erforderlich ist, um eine Lösung einer gewissen Genauigkeit zu berechnen. Als Vergleichslösung wird eine mit hoher Genauigkeit berechnete numerische Lösung $y^{\mathrm{ref}}(t)$, $z^{\mathrm{ref}}(t)$ verwendet. Mit $\epsilon_0 = 10^{-8}$ werden die Fehler ϵ^y, ϵ^z gemessen in den Normen

$$\|\epsilon^y(t)\| := \Big(\frac{1}{n_y} \sum_{i=1}^{n_y} \big(\frac{\epsilon_i^y(t)}{\epsilon_0 + |y_i^{\mathrm{ref}}(t)|} \big)^2 \Big)^{1/2}, \quad \|\epsilon^z(t)\| := \Big(\frac{1}{n_z} \sum_{i=1}^{n_z} \big(\frac{\epsilon_i^z(t)}{\epsilon_0 + |z_i^{\mathrm{ref}}(t)|} \big)^2 \Big)^{1/2}. \tag{3.80}$$

In Tab. 3.5 werden die 4 verglichenen HERK–Verfahren angegeben. Nutzt man die FSAL–Technik aus, so benötigen die Verfahren (für $n \geq 1$) je Integrationsschritt 4 (HENEW4), 5 (HEM4), 6 (HEDOP5) bzw. 8 (HEM5) Aufrufe von f (vgl. Tab. 3.3).

Tabelle 3.5: Liste der in Abb. 3.5 und 3.6 verglichenen HERK–Verfahren.

Verfahren	Markierung	Ordnung in y	Ordnung in z
HEM4 ([38])	„$*$“	4	2
HEM5 ([36])	„$\times$“	5	3
HENEW4 (Beispiel 22c mit $c_2 := 2/5$)	„$\circ$“	4	3
HEDOP5 (Tab. 3.4)	„$+$“	5	4

Die für den Vergleich benutzten Benchmark–Probleme sind in der Literatur ausführlich dokumentiert und sollen hier nicht im Detail beschrieben werden. In beiden Fällen hängen die Modellgleichungen linear von den algebraischen Komponenten ab, so daß die Gleichungssysteme (3.22) linear in ζ sind.

Abb. 3.5 zeigt Ergebnisse für den z. B. in [84, Kapitel VII.7] beschriebenen 7–Körper–Mechanismus[1] („Andrews' squeezing mechanism“), der auf $n_y = 14$ Differentialgleichungen und $n_z = 6$ algebraische Gleichungen in (3.19) führt ($t \in [0, 0.03]$).

Als zweites Beispiel wird die von Führer ([56], vgl. auch [57]) geringfügig modifizierte Fassung des Lastwagenmodells von Simeon et al. ([149]) verwendet.[2] In Abb. 3.6 ist $n_y = 18$, $n_z = 2$ und $t \in [0, 1]$.

Mit wachsender Genauigkeit (d. h. kleiner werdendem Fehler) wächst der Aufwand der Verfahren an. Hinsichtlich des Fehlers in y sind für beide Benchmark–Probleme die beiden

[1] Unter `http://www.cwi.nl/cwi/projects/IVPtestset.shtml` sind am CWI Amsterdam FORTRAN–Quellen für dieses Benchmark–Problem im Internet verfügbar.

[2] Der Autor dankt Herrn Dr. C. Führer (Lund) für die Überlassung des FORTRAN–Quelltexts.

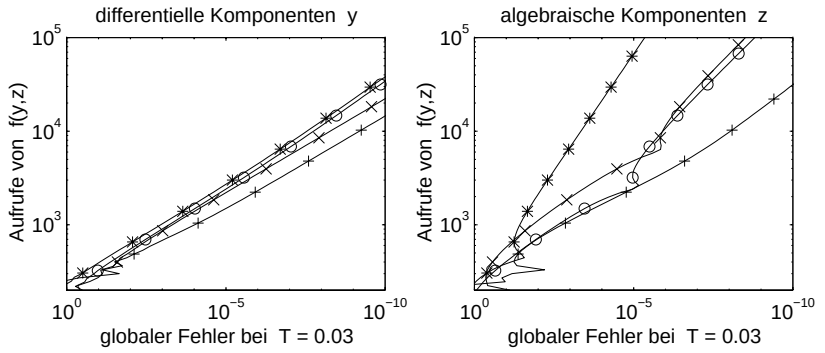


Abbildung 3.5: Aufwand und globaler Diskretisierungsfehler von HERK-Verfahren nach Tab. 3.5: Benchmark 7–Körper–Mechanismus ([84, Kapitel VII.7]).

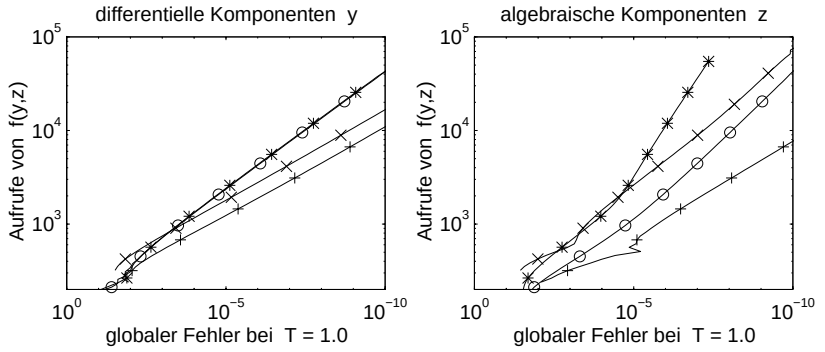


Abbildung 3.6: Aufwand und globaler Diskretisierungsfehler von HERK-Verfahren nach Tab. 3.5: Benchmark Lastwagenmodell ([56], [149]).

Verfahren 4. Ordnung HEM4 und HENEW4 („∗“, „o“) etwa gleichwertig. Für schärfere Genauigkeitsforderungen sind die Verfahren 5. Ordnung überlegen, wobei HEDOP5 („+“) deutlich effizienter als HEM5 („×“) ist. Wegen der höheren Konvergenzordnung in z sind die Unterschiede zwischen den Verfahren mit expliziter Stufe (HENEW4, HEDOP5) und den Verfahren HEM4 und HEM5 in den rechten Diagrammen der Abb. 3.5 und 3.6 noch wesentlich größer. (Mit der doppelt logarithmischen Skaleneinteilung ergeben sich für Fehler der Ordnung $\mathcal{O}(h^q)$ in der Regel lineare Kurvenverläufe mit Anstieg q in Abb. 3.5 und 3.6. Wechselt die Differenz zwischen der numerischen Lösung und der Vergleichslösung jedoch das Vorzeichen (z. B. im rechten Diagramm von Abb. 3.5), so kann sich lokal ein anderer Verlauf ergeben.)

Die Testergebnisse belegen, daß die Einführung einer expliziten ersten Stufe in HERK–Verfahren höherer Ordnung auf Verfahren führt, die den bisher aus der Literatur bekannten Verfahren HEM4 und HEM5 z. T. deutlich überlegen sind.

Zusammenfassung und Ausblick

Halb-explizite Runge–Kutta–Verfahren sind effektive Integrationsverfahren für nicht-steife DA–Systeme in Hessenbergform. Durch Einführung einer expliziten ersten Stufe wird es möglich, aus der Literatur bekannte explizite Runge–Kutta–Verfahren zu HERK–Verfahren gleicher Ordnung zu erweitern. Für Verfahren höherer Ordnung ist dabei eine zusätzliche Stufe erforderlich. Statt der aufwendigen Auflösung komplizierter Konsistenzbedingungen sind (bis zur Ordnung $q = 5$) ausschließlich lineare Bedingungsgleichungen zu erfüllen, wenn das gegebene explizite Runge–Kutta–Verfahren bestimmte vereinfachende Bedingungen erfüllt.

Bisher wurden Verfahren der Ordnung $q \le 5$ konstruiert, jedoch ist der Verfahrensansatz nicht hierauf beschränkt. Prinzipiell lassen sich so auch Verfahren höherer Ordnung konstruieren (auch mit mehr als einer expliziten Stufe), ebenso ist eine Übertragung auf die von Ostermann ([124]) untersuchten HERK–Verfahren für Index-3-Systeme in Hessenbergform denkbar.

3.3.2 HEDOP5 – Ein Integrator zur dynamischen Simulation von mechanischen Mehrkörpersystemen

Eines der wichtigsten Anwendungsgebiete für partitionierte Integratoren ist die dynamische Simulation von mechanischen Mehrkörpersystemen (MKS). Auf der Basis des im vorhergehenden Abschnitt konstruierten HERK–Verfahrens 5. Ordnung wurde der Integrator HEDOP5 entwickelt, der besonders für nicht-steife MKS geeignet ist. In diesem Abschnitt wird HEDOP5 (**H**alf–**E**xplicit integrator on the basis of the **5**th order method of **DO**rmand and **P**rince) beschrieben und mit anderen Integrationsverfahren verglichen.

Schrittweitensteuerung

Als Basis des Integrators wurde das explizite Runge–Kutta–Verfahren 5. Ordnung von Dormand und Prince gewählt, weil es für mittlere Genauigkeitsforderungen ($10^{-3} \dots 10^{-8}$)

als eines der effektivsten Einschrittverfahren gilt ([82, S. 249ff]). Mit einer expliziten ersten Stufe sind die Koeffizienten des HERK–Verfahrens durch Tab. 3.4 und $\gamma_{11} = 0$, $\gamma_{ij} = a_{i+1,j}$, $(i = 2, \dots, 7,\ j = 1, \dots, i)$ gegeben. Für die Schrittweitensteuerung wird — wie für gewöhnliche Differentialgleichungen — ein eingebettetes Verfahren 4. Ordnung verwendet.

Bemerkung 28 Das HERK–Verfahren hat im Unterschied zu dem expliziten Runge–Kutta–Verfahren bereits ohne Schrittweitensteuerung 7 Stufen (in der letzten Stufe wird $z_{n+1} = Z_{n7}$ bestimmt). Der Funktionswert $f(Y_{n7}, Z_{n7}) = f(y_{n+1}, z_{n+1})$ kann für die Konstruktion des eingebetteten Verfahrens verwendet werden. Sei

$$\tilde\eta := y_n + h \sum_{j=1}^{7} \tilde b_j f(Y_{nj}, Z_{nj}) \ , \quad \tilde y_{n+1} := \tilde\eta - [f_z(g_y f_z)^{-1}](y_n, z_n) g(\tilde\eta) \tag{3.81}$$

mit den Koeffizienten $\tilde b_j$ des eingebetteten Verfahrens 4. Ordnung ([82, Tab. II.5.2]):

$$\tilde b = (\ \frac{5179}{57600}\ ,\ 0\ ,\ \frac{7571}{16695}\ ,\ \frac{393}{640}\ ,\ -\frac{92097}{339200}\ ,\ \frac{187}{2100}\ ,\ \frac{1}{40}\ )^T \ .$$

Mit dem Baummodell zeigt man $\tilde\eta - y(t_{n+1}) = \mathcal{O}(h^4)$ und $P(t_n)(\tilde\eta - y(t_{n+1})) = \mathcal{O}(h^5)$ (falls $y_n = y(t_n)$, $z_n = z(t_n)$). Dann folgt aber aus $P(t_n)(\tilde y_{n+1} - \tilde\eta) = 0$ und

$$\begin{aligned} g_y(y_n)(\tilde y_{n+1} - y(t_{n+1})) &= g_y(y_n)(\tilde y_{n+1} - \tilde\eta) + g_y(y_n)(\tilde\eta - y(t_{n+1})) \\ &= -g(\tilde\eta) + (g(\tilde\eta) - g(y(t_{n+1})) + \mathcal{O}(h)\|\tilde\eta - y(t_{n+1})\|) = \mathcal{O}(h^5) \ , \end{aligned}$$

daß $\tilde y_{n+1} = y(t_{n+1}) + \mathcal{O}(h^5)$ ist ([31], [118]), d. h., $\tilde y_{n+1}$ kann als Näherung 4. Ordnung wie gewohnt in der Schrittweitensteuerung verwendet werden (gleiches gilt, wenn in (3.81) das Argument (y_n, z_n) von $[f_z(g_y f_z)^{-1}]$ durch (η, ζ) mit $\eta = y_n + \mathcal{O}(h)$, $\zeta = z_n + \mathcal{O}(h)$ ersetzt wird). Auf diese Weise erfordert die Berechnung der eingebetteten Lösung 4. Ordnung im HERK–Verfahren HEDOP5 nur 1 Aufruf von g und je 1 Vorwärts- und 1 Rückwärtssubstitution (mit der schon zuvor faktorisierten Matrix $[g_y f_z](\eta, \zeta)$).
Neben dem unter Verwendung von $\tilde y_{n+1} - y_{n+1}$ geschätzten lokalen Fehler $\delta y_h(t)$, der auf den globalen Fehler $\mathcal{O}(h^5)$ führt, werden die differentiellen Komponenten y auch von dem Term $\mathcal{O}(h^2) \max \|\delta z_h(t)\| = \mathcal{O}(h^6)$ beeinflußt (vgl. (3.43) und Beispiel 22d). Wegen der höheren Ordnung könnte dieser zweite Fehlerterm bei der Schrittweitensteuerung ignoriert werden. Ohne zusätzlichen Aufwand steht aber andererseits die Norm des Vektors $\tilde y_{n+1} - \tilde\eta$ als Näherung für $\|h\tilde b_7 \cdot f_z(y(t), z(t))\delta z_h(t)\|$ zur Verfügung. Man beweist nämlich für den durch $g(y_n + h\sum_{j=1}^{6} \tilde b_j f(Y_{nj}, Z_{nj}) + h\tilde b_7 f(y_{n+1}, \tilde\zeta)) = 0$ definierten Vektor $\tilde\zeta$ mit dem Baummodell $\tilde\zeta - z(t_{n+1}) = \mathcal{O}(h^3)$ (falls $y_n = y(t_n)$, $z_n = z(t_n)$). Deshalb folgt

$$\begin{aligned} g(\tilde\eta) &= g(\tilde\eta) - g\Big(y_n + h\sum_{j=1}^{6} \tilde b_j f(Y_{nj}, Z_{nj}) + h\tilde b_7 f(y_{n+1}, \tilde\zeta)\Big) \\ &= [g_y f_z](y_n, z_n) \cdot h\tilde b_7 (Z_{n7} - \tilde\zeta) + \mathcal{O}(h^2)\|Z_{n7} - \tilde\zeta\| \ , \\ \tilde y_{n+1} - \tilde\eta &= -[f_z(g_y f_z)^{-1}](y_n, z_n) g(\tilde\eta) = -h\tilde b_7 f_z(y_n, z_n)(Z_{n7} - \tilde\zeta) + \mathcal{O}(h^2)\|Z_{n7} - \tilde\zeta\| \end{aligned}$$

und $\|h\tilde b_7 f_z(y_n, z_n)(\tilde\zeta - z(t_{n+1}))\| = \|\tilde y_{n+1} - \tilde\eta\| + \mathcal{O}(h^5)$, denn $Z_{n7} = z(t_{n+1}) + \mathcal{O}(h^4)$.

Ähnlich wie im Code DOPRI5 von Hairer et al. ([82, Anhang]) wird unter Verwendung der eingebetteten Lösung $\tilde{y}_{n+1}$ in jedem Integrationsschritt die Größe

$$\texttt{err} := \|\tilde{y}_{n+1} - y_{n+1}\| := \Big(\frac{1}{n_y} \sum_{i=1}^{n_y} \big(\frac{\tilde{y}_{n+1,i} - y_{n+1,i}}{\mathrm{ATOL}_i + |y_{n+1,i}| \cdot \mathrm{RTOL}_i} \big)^2 \Big)^{1/2}$$

berechnet, wobei ATOL_i, RTOL_i die für die i-te differentielle Lösungskomponente vorgegebenen Fehlerschranken (absoluter und relativer Fehler) bezeichnen. Eine neue Schrittweite h_{neu} bestimmt man entweder mit der Standardstrategie

$$h_{\mathrm{neu}} = 0.9\, h \cdot (1/\texttt{err})^{0.2}$$

oder mit einer modifizierten Strategie in Anlehnung an die „PI–Steuerung“ nach Gustafsson et al. ([78], vgl. [84, S. 28ff] für die Anwendung auf DOPRI5). Ist $\texttt{err} \le 1$, so wird die Integration mit dem nächsten Integrationsschritt $(t_{n+1} \to t_{n+2})$ fortgesetzt, andernfalls muß der Integrationsschritt $t_n \to t_{n+1}$ mit der verkleinerten Schrittweite h_{neu} wiederholt werden.

Prinzipiell werden für Index-2-Systeme in der Schrittweitensteuerung nur die Fehler in den differentiellen Komponenten betrachtet. Die globalen Fehler in z können um mehrere Größenordnungen größer als die in y sein. Hinsichtlich des in Bemerkung 28 betrachteten Fehlerterms $h^2 \|f_z(y(t_n), z(t_n)) \delta z_h(t_n)\|$ erwies sich in allen Testrechnungen, daß die Näherung $h\|\tilde{y}_{n+1} - \tilde{\eta}\|$ deutlich kleiner als $\|\tilde{y}_{n+1} - y_{n+1}\|$ war. In praxi führt deshalb die Definition $\texttt{err} := \max\{\|\tilde{y}_{n+1} - y_{n+1}\|, h\|f(y_{n+1}, z_{n+1})\| \|\tilde{y}_{n+1} - \tilde{\eta}\|\}$ auf dieselben Ergebnisse wie $\texttt{err} := \|\tilde{y}_{n+1} - y_{n+1}\|$.

Anwendung auf die Modellgleichungen für Mehrkörpersysteme

Für die Anwendung von HEDOP5 auf die MKS–Modellgleichungen (3.1) ist die Index-2-Formulierung als Index-2-System (3.19) in Hessenbergform zu schreiben. (Sowohl in den MKS–Modellgleichung als auch in den DA–Systemen vom Index 2 werden in der vorliegenden Arbeit die in der Literatur üblichen Bezeichner für die auftretenden Funktionen verwendet. Explizit sei deshalb darauf hingewiesen, daß f und g in (3.1) eine prinzipiell andere Bedeutung als in (3.19) haben.)

Für die Implementierung von HEDOP5 sollen möglichst schwache Voraussetzungen an die MKS–Modellgleichungen gestellt werden, um die breite Anwendbarkeit zu garantieren. Deshalb wird hier die — für theoretische Untersuchungen günstige — Beschränkung auf autonome Probleme mit regulärer Massenmatrix $M(q)$ aufgegeben. Mit den Bezeichnungen von Abschnitt 2.3 wird das HERK–Verfahren formal auf das Index-2-System

$$\begin{aligned} q' &= T(q,t)v \\ v' &= u \\ w' &= M(q)u - f(q,v,\lambda,t) + G^T(q,t)\lambda \\ 0 &= w \\ 0 &= G(q,t)v + g_t(q,t) \end{aligned} \tag{3.82}$$

mit $G(q,t) := (\frac{\partial}{\partial q} g(q,t)) \cdot T(q,t)$ angewendet ([37], [109]). In (3.82) sind die versteckten Zwangsbedingungen $0 = G(q,t)v + g_t(q,t) = \frac{d}{dt} g(q(t),t)$ enthalten. In einer Umgebung

der analytischen Lösung seien die Funktionen f, g und M hinreichend oft stetig differenzierbar, die Massenmatrix $M(q)$ positiv semidefinit und die Matrix

$$\begin{pmatrix} M(q) & \Gamma(q,v,\lambda,t) \\ G(q,t) & 0 \end{pmatrix} \quad \text{mit} \quad \Gamma(q,v,\lambda,t) := f_\lambda(q,v,\lambda,t) - G^T(q,t) \tag{3.83}$$

regulär. Aus praktischer Sicht bilden MKS, in denen keine Reibungskräfte wirken, einen wichtigen Spezialfall. Für diese Mehrkörpersysteme hängt f nicht von den Zwangskräften ab, und es gilt $f = f(q,v,t)$ und $\Gamma(q,v,\lambda,t) = -G^T(q,t)$ in (3.83).

Die hier getroffenen Voraussetzungen an (3.82) sind i. allg. von den mit Mehrkörperformalismen automatisch generierten MKS–Modellgleichungen, die nicht immer auf eine reguläre Massenmatrix $M(q)$ führen, erfüllt (vgl. z. B. [144, Abschnitt 2.1 und Anhang A2]). Sie garantieren darüberhinaus bei konsistenten Anfangswerten die eindeutige Lösbarkeit des Anfangswertproblems für die MKS–Modellgleichungen, denn das DA–System (3.82) bildet ein Index-2-System in Hessenbergform.

Wie üblich erweitert man die Definition der HERK–Verfahren auf nichtautonome Systeme durch formales Hinzufügen der trivialen Gleichung $t' = 1$ ([82, S. 143]). Betrachtet man nun wie in Bemerkung 25 ein $\hat{s}$-stufiges HERK–Verfahren mit $\hat{s} = s + 1$, das unter Verwendung der FSAL–Technik aus einem expliziten Runge–Kutta–Verfahren hervorgegangen ist, so folgt für (3.82) mit $y = (q,v,w)^T$ und $z = (u,\lambda)^T$ insbesondere

$$0 = w_n + h \sum_{j=1}^{i} \gamma_{ij} W'_{nj} , \quad (i = 1, \ldots, \hat{s})$$

mit

$$W'_{ni} = M(Q_{ni})U_{ni} - f(Q_{ni}, V_{ni}, \Lambda_{ni}, t_n + c_i h) + G^T(Q_{ni}, t_n + c_i h)\Lambda_{ni} , \quad (i = 1, \ldots, \hat{s})$$

und $W'_{n+1,1} = W'_{n\hat{s}}$. Ist $\gamma_{11} = 0$, $\gamma_{ii} \neq 0$, $(i = 2, \ldots, \hat{s})$, dann zeigt man mittels vollständiger Induktion, daß für konsistente Anfangswerte $w_m = W'_{mj} = 0$, $(j = 1, \ldots, \hat{s})$ für beliebiges $m \ge 0$ gilt.

Unter der Voraussetzung $\gamma_{ij} = a_{i+1,j}$, $(i = 2, \ldots, \hat{s},\ j = 1, \ldots, i)$ läßt sich ein solches HERK–Verfahren für (3.82) deshalb mit den Bezeichnungen $a_{s+1,j} = b_j$, $M_{nj} = M(Q_{nj})$, $T_{nj} = T(Q_{nj}, t_n + c_j h)$, ..., $(j = 1, \ldots, \hat{s})$ schreiben als

$$q_{n+1} = Q_{n,s+1} , \quad v_{n+1} = V_{n,s+1} , \quad u_{n+1} = U_{n,s+1} , \quad \lambda_{n+1} = \Lambda_{n,s+1}$$

mit Stufenvektoren Q_{ni}, V_{ni}, U_{ni}, Λ_{ni}, für die gilt

$$\begin{gathered} Q_{ni} = q_n + h \sum_{j=1}^{i-1} a_{ij} T_{nj} \cdot V_{nj} , \qquad V_{ni} = v_n + h \sum_{j=1}^{i-1} a_{ij} U_{nj} , \quad (i = 1, \ldots, \hat{s}+1) , \\ U_{n1} = u_n , \quad \Lambda_{n1} = \lambda_n , \\ \begin{pmatrix} M_{ni} & G_{ni}^T \\ G_{n,i+1} & 0 \end{pmatrix} \begin{pmatrix} U_{ni} \\ \Lambda_{ni} \end{pmatrix} = \begin{pmatrix} f(Q_{ni}, V_{ni}, \Lambda_{ni}, t_n + c_i h) \\ -\dfrac{1}{h a_{i+1,i}} r_i \end{pmatrix} , \quad (i = 2, \ldots, s+1) \end{gathered} \tag{3.84}$$

mit

$$r_i := G_{n,i+1} \cdot \left(v_n + h \sum_{j=1}^{i-1} a_{i+1,j} U_{nj} \right) + g_t(Q_{n,i+1}, t_n + c_{i+1} h) \; .$$

Bis auf die explizite erste Stufe ist diese Darstellung identisch zu den von Brasey und Hairer betrachteten Verfahren HEM4 und HEM5 ([36], [38]).

Bemerkung 29 Der Verfahrensansatz (3.84) kann unmittelbar auf Modellgleichungen (3.14) für MKS mit Kontaktbedingungen übertragen werden. Hierzu bestimmt man für $i = 1, \ldots, s$ Stufenvektoren S_{ni}, für die gilt $h(Q_{ni}, S_{ni}) = 0$. Diese S_{ni} sind dann bei der Berechnung der nachfolgenden Stufenvektoren in $\tilde{f}(Q_{ni}, S_{ni}, V_{ni}, \Lambda_{ni}, t_n + c_i h)$, ... einzusetzen (vgl. [81, S. 21]). Für hinreichend kleine Schrittweiten sind dabei zu gegebenen Q_{ni} und q_{n+1} die Gleichungssysteme $h(Q_{ni}, S_{ni}) = 0$ und $h(q_{n+1}, s_{n+1}) = 0$ lokal eindeutig auflösbar nach S_{ni} bzw. s_{n+1}, denn $\partial h / \partial s$ ist in einer Umgebung der analytischen Lösung regulär.

Implementierung

Bemerkung 30 In (3.84) werden $U_{ni}, \Lambda_{ni}, (i = 2, \ldots, \hat{s})$ als Lösung von Gleichungssystemen bestimmt. In einer Umgebung der analytischen Lösung sind diese Gleichungssysteme unter obigen Voraussetzungen lokal eindeutig lösbar, wenn die Schrittweite h klein ist. Hängt f von λ ab, so kann diese Lösung mit dem vereinfachten Newtonverfahren mit der Matrix

$$\begin{pmatrix} M(q_n) & -\Gamma(q_n, v_n, \lambda_n, t_n) \\ G(q_n, t_n) & 0 \end{pmatrix} \tag{3.85}$$

berechnet werden, als Startwerte wählt man $U_{ni}^{(0)} := u_n$, $\Lambda_{ni}^{(0)} := \lambda_n$ oder wie in [9] und [27] auch $U_{ni}^{(0)} := \sum_{j=1}^{i-1} \nu_{ij} U_{nj}$, $\Lambda_{ni}^{(0)} := \sum_{j=1}^{i-1} \nu_{ij} \Lambda_{nj}$ mit geeigneten Parametern ν_{ij} (vgl. auch [37, Abschnitt II.8]).
Für MKS ohne Reibungskräfte ist $f_\lambda \equiv 0$. Dann kann zur Verringerung des numerischen Aufwands ausgenutzt werden, daß $-\Gamma = G^T$ gilt und deshalb die Matrix in (3.85) symmetrisch ist. Testrechnungen haben aber gezeigt, daß es noch sehr viel günstiger ist, die in diesem Fall *linearen* Gleichungssysteme in (3.84) mit einem direkten Verfahren zu lösen (obwohl die s Koeffizientenmatrizen in (3.84) nicht symmetrisch sind).

Typische Anforderungen an einen Integrator für MKS–Modellgleichungen sind neben der reinen Zeitintegration

- die Berechnung konsistenter Anfangswerte,
- die Projektion der numerischen Lösung auf die Mannigfaltigkeit, die durch die Zwangsbedingungen auf Ebene der Lage- und der Geschwindigkeitskoordinaten bestimmt ist,
- eine stetige Lösungsdarstellung („dense output“, z. B. für graphische Darstellungen),
- die Arbeit mit Schaltfunktionen (z. B. zur Lokalisierung von Unstetigkeitsstellen),
- leistungsfähige Routinen zum Lösen der linearen Gleichungssysteme (ggf. unter Ausnutzung der Besetztheitsstruktur der Koeffizientenmatrix) und

- die Berücksichtigung von zusätzlichen dynamischen Variablen c, die in die Berechnung von f eingehen ($f = f(q, c, v, \lambda, t)$) und für die $c' = d(q, c, v, \lambda, t)$ gilt (z. B. in mechatronischen Systemen).

Für partitionierte Verfahren war der auf einem halb-expliziten Extrapolationsverfahren von Lubich ([106]) aufbauende Integrator MEXX ([108], [109]) das erste Programm, das all diesen Forderungen gerecht wurde. Die in MEXX umgesetzten Ideen finden seither auch für die Programmierung anderer halb-expliziter Verfahren Anwendung.

Sowohl für praktische Anwendungen als auch für numerische Vergleichsrechnungen sind dabei Programmbibliotheken (z. B. MBSPACK [145], MBSSIM [157]) besonders geeignet, weil sie verschiedene Integratoren zur Verfügung stellen, die mit einer einheitlichen Schnittstelle zum mechanischen Problem versehen sind.

HEDOP5 wurde auf der Basis anderer halb-expliziter Integratoren des Programmpakets MBSPACK in FORTRAN77 implementiert und in dieses Programmpaket integriert[3] (im Internet verfügbar: `ftp://ftp.mathematik.tu-darmstadt.de/pub/department/software/mbspack/`). Beim Aufruf des vom Nutzer bereitzustellenden Funktionsunterprogramms zur Auswertung von M, G, f, g, g_t, ... verwenden die halb-expliziten Integratoren in MBSPACK eine Schnittstelle, die nahezu identisch zu MEXX ist. Im Anhang wird mit dem Eingangskommentar zu `hedop5.f` ein Überblick über den Aufruf von HEDOP5 gegeben, eine ausführliche Beschreibung ist Bestandteil der Dokumentation von MBSPACK (Datei `hedop5.man`).

Liegen konsistente Anfangswerte vor, so erfüllt HEDOP5 die oben genannten Anforderungen mit Ausnahme der Schaltfunktionen. Zur Lösung der auftretenden linearen Gleichungssysteme werden in MBSPACK wahlweise LINPACK–, LAPACK– oder vom Nutzer bereitgestellte Routinen verwendet ([3], [52]). Die stetige Lösungsdarstellung berechnet man ebenso wie für das Verfahren von Dormand und Prince ([82, S. 191ff]).

Bei der Projektion von Vektoren q, v, die sowohl zur Vermeidung des Drift-off-Effekts als auch bei der Berechnung konsistenter Anfangswerte große Bedeutung hat, wird wie in MEXX ein der speziellen Struktur der MKS–Modellgleichungen angepaßtes Verfahren verwendet ([106], [144, Algorithmus 4.1]): Ausgehend von Vektoren $\tilde{q}$, $\tilde{v}$, die die Zwangsbedingungen zum Zeitpunkt t näherungsweise erfüllen, berechnet man $q = \tilde{q} + T(q,t)\Delta_q$ und η_q mit

$$\begin{aligned} 0 &= M(q)\Delta_q + G^T(q,t)\eta_q \; , \\ 0 &= g(q,t) \end{aligned} \tag{3.86}$$

und anschließend v, η_v mit

$$\begin{aligned} 0 &= M(q)(v - \tilde{v}) + G^T(q,t)\eta_v \; , \\ 0 &= G(q,t)v + g_t(q,t) \; . \end{aligned} \tag{3.87}$$

Im Projektionsschritt (3.87) ist nur ein lineares Gleichungssystem zu lösen. Für die Lösung des nichtlinearen Gleichungssystems (3.86) mit dem vereinfachten Newtonverfahren steht während der Integration bereits eine gute Approximation der Jacobimatrix zur Verfügung

[3] Der Autor dankt Herrn Dr. B. Simeon (Darmstadt) für die Überlassung der Quelltexte der Integratoren MDOP5 und MHERK5 aus MBSPACK und für die freundliche Unterstützung bei der Implementierung und bei den Tests von HEDOP5.

(vgl. (3.84)), außerdem ist für Startwerte $\Delta_q = 0$, $\eta_q = 0$, das Residuum in (3.86) klein, denn $q_n = q(t_n) + \mathcal{O}(h^5)$ und $g(q(t_n), t_n) = 0$. Deshalb erfordern die Projektionsschritte insgesamt nur einen geringfügigen zusätzlichen Aufwand während der Integration ([109]). Für eine ausführliche Beschreibung der Implementierung sei ebenso wie für die detaillierte Beschreibung der halb-expliziten Integratoren in MBSPACK auf [144] verwiesen.

Vergleichsrechnungen

Aus der Literatur sind zahlreiche numerische Tests bekannt, die für nicht-steife MKS die Überlegenheit halb-expliziter Integratoren gegenüber impliziten Verfahren belegen (vgl. z. B. [84, Kapitel VII.7], [36], [37], [144], [145]). Die nachfolgenden Testergebnisse zeigen, daß HEDOP5 den dabei schnellsten halb-expliziten Runge–Kutta–Verfahren gleichwertig und z. T. sogar überlegen ist. Aus verschiedenen Testrechnungen mit Integratoren aus MBSPACK wurden hierfür 2 Benchmark–Probleme ausgewählt. Im Unterschied zu den einfachen Testrechnungen aus Abschnitt 3.3.1, die vor allem die verschiedenen Diskretisierungsfehler der betrachteten Verfahren verdeutlichen sollen, orientieren sich die beiden hier verwendeten Benchmark–Probleme an typischen Anforderungen der industriellen Praxis (speziell des Straßenfahrzeugbaus).

Beispiel A Das bereits in Abschnitt 3.3.1 betrachtete Lastwagenmodell wird hier in der in [149] dokumentierten Form für $t \in [0, 20]$ integriert, es ist $n_q = n_v = 11$, $n_\lambda = 1$. Die Modellgleichungen enthalten eine (zeitabhängige) Anregungsfunktion, die Unebenheiten der Straße simuliert und aus einer Fourierreihenapproximation von Fahrbahnmeßdaten gewonnen wurde. (FORTRAN–Quelltext von Simeon verfügbar im Internet als `ftp://ftp.mathematik.tu-darmstadt.de/pub/department/software/mbspack/truck.f`).

Beispiel B Ein typisches Beispiel für MKS, bei denen es sehr viel aufwendiger ist, die Zwangsbedingungen nicht nur auf Ebene der Lage- und Geschwindigkeitskoordinaten, sondern außerdem auch noch auf Ebene der Beschleunigungskoordinaten auszuwerten (Gleichung (3.3)), ist die 5–Punkt–Hinterradaufhängung nach Hiller und Frik ([91]). Die Modellgleichungen für dieses Beispiel wurden von Leister mit dem Programmpaket NEWEUL ([99]) generiert[4], der FORTRAN–Code umfaßt ca. 7000 Zeilen Quelltext. Es ist $n_q = n_v = 14$, $n_\lambda = 12$ und $t \in [0, 0.6]$. Mit einer (glatten) Anregungsfunktion wird die Fahrt über eine Fahrbahnunebenheit simuliert ([147]).

Verglichen werden die folgenden halb-expliziten Verfahren:

- MDOP5: explizites Runge–Kutta–Verfahren 5. Ordnung von Dormand und Prince, als halb-explizites Runge–Kutta–Verfahren auf die Index-1-Formulierung angewendet ([81, S. 20f], in MBSPACK von Simeon implementiert),
- HEM5 ([36], hier in der Implementierung MHERK5 von Simeon verwendet),
- HEDOP5 aus MBSPACK und

[4] Der Autor dankt den Herren Dr. G. Leister (Stuttgart) und Dr. B. Simeon (Darmstadt) für die Überlassung des an die MBSPACK–Schnittstelle angepaßten FORTRAN–Quelltexts.

- PHEM56 ([118] mit den Verfahrensparametern des unter `ftp://ftp.unige.ch/pub/doc/math/mechanic/phem56.f` verfügbaren Codes): HERK–Verfahren auf der Basis des Verfahrens 5. Ordnung von Dormand und Prince. Wie in HEDOP5 ist $\gamma_{11} = 0$ (explizite Stufe), $\hat{s} = s + 1 = 7$, $\gamma_{ii} \neq 0$, $(i = 2, \ldots, 7)$, $\gamma_{6j} = b_j$, $(j = 1, \ldots, 6)$ und die FSAL–Technik wird ausgenutzt, es gilt jedoch nicht $\gamma_{ij} = a_{i+1,j}$, $(i \leq 5)$. (In den Tests wird für PHEM56 eine Modifikation von `hedop5.f` verwendet.)

In Tab. 3.6 wird für die Integratoren ein Teil des numerischen Aufwands zusammengefaßt, dabei bezeichnet DEC die Anzahl der Matrixzerlegungen und SOL die Anzahl der Vorwärts- und Rückwärtssubstitutionen (jeweils pro Integrationsschritt). Das Konstruktionsprinzip von PHEM56 läßt mit der Wahl von γ_{ij}, $(i = 2, \ldots, 5,\ j = 1, \ldots, i)$ mehr Freiheitsgrade als in HEDOP5. Damit erreicht man im Vergleich zu HEDOP5 etwas kleinere Koeffizienten im führenden Fehlerterm, muß dafür aber in (3.84) statt $G_{n,i+1} = G(Q_{n,i+1}, t_n + c_{i+1}h)$ zusätzlich auch $G(q_n + h\sum_{j=1}^{i} \gamma_{ij} T_{nj} V_{nj}, t_n + \hat{c}_{i+1}h)$ auswerten $(i = 2, 3, 4, 5)$.

Tabelle 3.6: Numerischer Aufwand von halb-expliziten Integratoren (Angaben pro Integrationsschritt).

Verfahren	Markierung	Aufrufe von f	Aufrufe von G	Aufrufe von g_{qq} aus (3.3)	DEC	SOL	Ordnung in y	Ordnung in z
MDOP5	„$\circ$“	6	6	6	6	6	5	5
HEM5	„$\times$“	8	8	0	8	8	5	3
HEDOP5	„$+$“	6	8	0	6	7	5	4
PHEM56	„$*$“	6	12	0	6	7	5	4

In Tab. 3.6 beziehen sich die Angaben für DEC und SOL auf den sowohl für Beispiel A als auch für Beispiel B vorliegenden Fall $f = f(q, v, t)$, $f_\lambda \equiv 0$ und auf direkte Verfahren zur Lösung der Gleichungssysteme in (3.84). Die hierbei auftretenden Koeffizientenmatrizen sind für MDOP5 symmetrisch, für die anderen Verfahren dagegen i. allg. nicht symmetrisch. Der geringfügige zusätzliche Aufwand für Projektionsschritte wurde in Tab. 3.6 nicht berücksichtigt. Für die auf der Index-2-Formulierung basierenden Verfahren HEM5, HEDOP5 und PHEM56 sind sehr viel weniger Projektionsschritte erforderlich als für MDOP5 (vgl. [1]).

In den Testrechnungen wurden jedem der Integratoren nacheinander verschiedene Fehlerschranken ATOL_i, RTOL_i, $(i = 1, \ldots, n_q + n_v)$ für die differentiellen Komponenten q und v vorgegeben (Toleranzen für absoluten und relativen Fehler, für alle Komponenten sind diese Toleranzen gleich: $\mathrm{ATOL}_i = \mathrm{ATOL}$, $\mathrm{RTOL}_i = \mathrm{RTOL}$, $(i = 1, \ldots, n_q + n_v)$, $\mathrm{ATOL} = 0.1\,\mathrm{RTOL}$). Abb. 3.7 zeigt für Beispiel A die Ergebnisse für $\mathrm{RTOL} = 10^{-3-j/8}$, $(j = 0, 1, \ldots, 56)$ und Abb. 3.8 für Beispiel B die Ergebnisse für $\mathrm{RTOL} = 10^{-4-j/8}$, wobei die Marker die Ergebnisse für $j/8 \in \mathbb{N}$ (also $\mathrm{RTOL} = 10^{-3}, 10^{-4}, \ldots$) kennzeichnen. Dargestellt ist in doppelt logarithmischer Skaleneinteilung die Rechenzeit (auf einer SUN Sparc5 Workstation) und die erreichte Genauigkeit (in der Norm aus (3.80)). Wie in Kapitel II.10 von [82] zeigen diese Diagramme, wieviel Rechenzeit erforderlich ist, um eine Lösung mit einer gewünschten Genauigkeit zu erreichen.

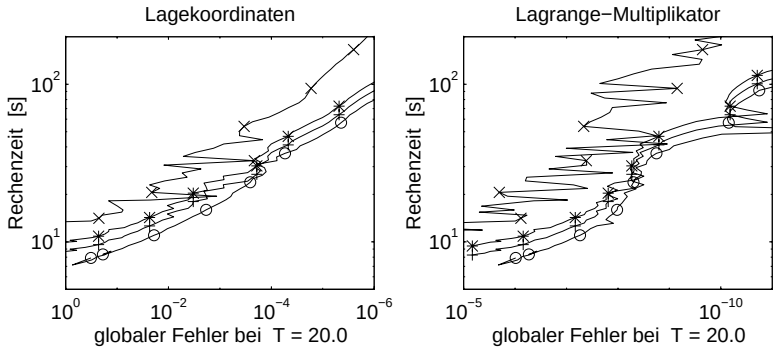


Abbildung 3.7: Vergleich der Integratoren MDOP5 („∘“), HEM5 („×“), HEDOP5 („+“) und PHEM56 („∗“) bei Anwendung auf Beispiel A.

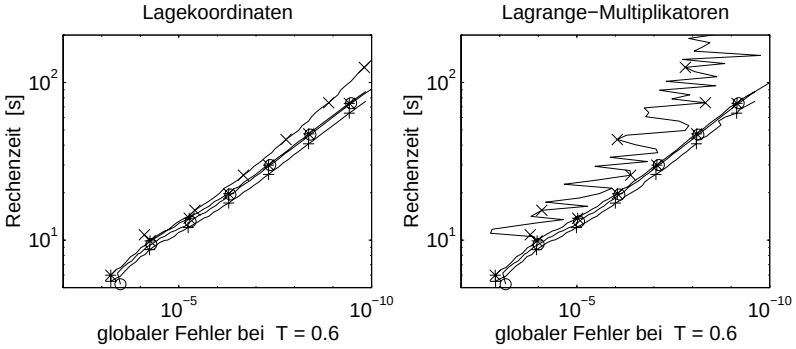


Abbildung 3.8: Vergleich der Integratoren MDOP5 („∘“), HEM5 („×“), HEDOP5 („+“) und PHEM56 („∗“) bei Anwendung auf Beispiel B.

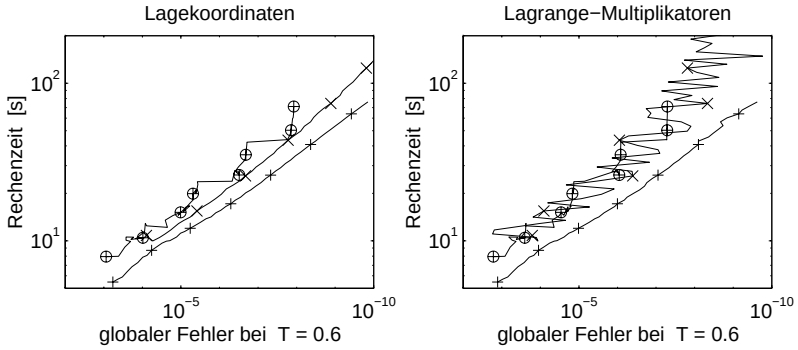


Abbildung 3.9: Vergleich der Integratoren HEM5 („×“), HEDOP5 („+“) und DASSL („⊕“, angewendet auf die GGL–Formulierung): Beispiel B.

Alle drei auf dem Verfahren von Dormand und Prince basierenden Verfahren erweisen sich gleichermaßen als sehr effektiv, sie sind dem Verfahren HEM5 deutlich überlegen. Insbesondere profitieren nicht nur MDOP5, sondern auch HEDOP5 und PHEM56 von der effizienten und sehr zuverlässigen Schrittweitensteuerung durch das eingebettete Verfahren 4. Ordnung (dies ist erkennbar an der gleichmäßigen Verringerung des Fehlers bei Verringerung der Genauigkeitsschranke).

Die Rechenzeitunterschiede zwischen MDOP5, HEDOP5 und PHEM56 liegen in beiden Beispielen zwischen 10% und 20% und sind problemabhängig. Ist die Auswertung von g_{qq} sehr aufwendig (Beispiel B), so ist HEDOP5 überlegen, dagegen ist MDOP5 der schnellste Integrator für Beispiel A. Bei der hier verwendeten Implementierung war HEDOP5 in allen Testrechnungen dem sehr ähnlichen Verfahren PHEM56 geringfügig überlegen; dies wird auch durch Testrechnungen von Murua ([118]) bestätigt. In den algebraischen Komponenten z führen die Integratoren HEDOP5 und PHEM56 trotz der gegenüber MDOP5 geringeren Konvergenzordnung zu sehr genauen Ergebnissen.

Schließlich werden zum Vergleich in Abb. 3.9 für Beispiel B die Ergebnisse eines impliziten Integrators aufgenommen („⊕“, DASSL angewendet auf die Gear–Gupta–Leimkuhler–Formulierung der Bewegungsgleichungen, vgl. Abschnitt 3.1), wobei die vergleichsweise hohen Rechenzeiten von DASSL noch einmal die Vorzüge der halb-expliziten Integratoren belegen.

Zusammenfassung

Die sehr guten Ergebnisse der Integratoren HEDOP5 und PHEM56 zeigen, daß durch die Einführung einer expliziten Stufe die Nachteile der bisher bekannten HERK–Verfahren für Index-2-Systeme überwunden werden konnten, wobei sich das Verfahren mit $\gamma_{ij} = a_{i+1,j}$, $(i = 2, \ldots, s+1, j = 1, \ldots, i)$ als besonders vorteilhaft erweist. Gegenüber Verfahren wie MDOP5, die auf der Index-1-Formulierung der Bewegungsgleichungen aufbauen, entfallen

die (je nach Beispiel) u. U. aufwendigen Auswertungen von g_{qq} (diese Funktion muß deshalb auch nicht als Quelltext zur Verfügung stehen).

Gleichzeitig unterstreichen die Testergebnisse, daß aus praktischer Sicht nicht nur die Wahl des Diskretisierungsverfahrens, sondern vor allem auch die Auswahl einer geeigneten Formulierung der Modellgleichungen wichtig ist, um effiziente Simulationssoftware zu entwickeln. So ist der Integrator MDOP5 vor allem deshalb so erfolgreich, weil es für die Index-1-Formulierung möglich ist, ein sehr effizientes Verfahren für gewöhnliche Differentialgleichungen direkt auf das DA–System zu übertragen. Die Entwicklung von PHEM56 und HEDOP5 zeigt, daß Verfahren, die auf der Index-2- oder sogar auf der Index-3-Formulierung aufbauen, nur dann konkurrenzfähig werden, wenn man auch hier an die effektivsten Verfahren für gewöhnliche Differentialgleichungen anknüpfen kann.

Für nicht-steife MKS–Modellgleichungen der Standardform (3.1) oder der erweiterten Form (3.14) sind halb-explizite Integratoren den auf impliziten Verfahren basierenden Standard–Integratoren (z. B. DASSL, RADAU5) überlegen. Darüberhinaus ist — insbesondere für MKS ohne Reibungskräfte (d. h. $f_\lambda \equiv 0$) — die Implementierung der HERK–Verfahren wesentlich einfacher. Halb-explizite Verfahren erreichen jedoch nicht den breiten Anwendungsbereich von impliziten Integratoren, denn sie nutzen die spezielle Struktur des DA–Systems bereits im Verfahrensansatz aus. In Simulationspaketen können sie deshalb die bewährten impliziten Integratoren zwar sinnvoll ergänzen, aber nie vollständig ersetzen.

3.3.3 Partitionierte lineare Mehrschrittverfahren vom Adams–Typ

Im Verhältnis zu der umfangreichen Literatur über Einschrittverfahren für DA–Systeme beschränken sich die Arbeiten zu Mehrschrittverfahren bisher auf relativ wenige Verfahrensklassen. Ein wesentlicher Grund hierfür ist die aufwendige Implementierung von Verfahren variabler Ordnung und Schrittweite. Neben dem am weitesten verbreiteten BDF–Code DASSL ([129]) gibt es verschiedene Implementierung der BDF für DA–Systeme. Für die Integration der MKS–Modellgleichungen — die hier im Vordergrund stehen wird — sind z. B. ODASSL ([66]) und MKS–DAESOL ([54]) zu nennen. Wegen der höheren Konvergenzordnung und der kleineren Fehlerkonstanten versucht man außerdem, für nicht-steife MKS (implizite) Adams–Verfahren zu verwenden. Der Integrator MBSABM ([4], [35], Bestandteil des Programmpakets MBSSIM [157]) verwendet hierzu die Index-1-Formulierung der MKS–Modellgleichungen. Dagegen sind implizite Adams–Verfahren höherer Ordnung bei Anwendung auf Index-2-Systeme instabil und konvergieren nicht. Verschiedene Ansätze, durch Verfahrensmodifikationen („β-blocking“) die Konvergenz der Verfahren auch bei Anwendung auf Index-2-Systeme (3.19) zu erreichen, wurden von Arévalo, Führer und Söderlind untersucht ([7], vgl. auch [6], [8]).

Der Konvergenzbeweis für partitionierte lineare Mehrschrittverfahren (PLMSV) aus Abschnitt 3.2.2 gibt den Konvergenzuntersuchungen für solche modifizierten Verfahren einen einheitlichen Rahmen auf der Grundlage des Konvergenzbeweises für klassische Mehrschrittverfahren in [84, Kapitel VII.3]. Neben den β-geblockten Verfahren erweisen sich die hier neu eingeführten PLMSV vom Adams–Typ, die ein Adams–Moulton–Verfahren für den differentiellen Teil eines Index-2-Systems (3.19) mit einer BDF–Diskretisierung für den algebraischen Teil kombinieren, als besonders geeignet. Ein solches k-Schritt-Verfahren erreicht für beliebiges $k \geq 1$ die Ordnung $q = k+1$ für y und die Ordnung $q-1 = k$ für z. Ergebnisse eines numerischen Tests belegen die Vorteile des neuen Verfahrensansatzes.

Im weiteren werden für das PLMSV (3.45) drei spezielle Verfahrensfunktionen betrachtet:

(I) Die algebraischen Komponenten z_{n+k} werden so bestimmt, daß y_{n+k} und z_{n+k} die versteckten Zwangsbedingungen $0 = \frac{d}{dt} g(y(t)) = [g_y f](y,z)$ erfüllen, d. h.

$$0 = [g_y f](y_{n+k}, z_{n+k}) \; .$$

(II) **β-blocking:** Sind zu einem linearen Mehrschrittverfahren mit charakteristischen Polynomen $\varrho(\xi) = \sum_{j=0}^{k} \alpha_j \xi^j$, $\sigma(\xi) = \sum_{j=0}^{k} \beta_j \xi^j$ und $\beta_k \neq 0$ zusätzliche Koeffizienten τ_j, $(j = 0, 1, \ldots, k)$ gegeben, so definiert ein β-geblocktes Mehrschrittverfahren nach Arévalo et al. die Vektoren y_{n+k} und z_{n+k} gemäß

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j y_{n+j} &= h \sum_{j=0}^{k-1} \beta_j f(y_{n+j}, z_{n+j}) + h\beta_k f(y_{n+k}, z_{n+k} - \sum_{j=0}^{k} \frac{\tau_j}{\beta_k} z_{n+j}) \; , \\ 0 &= g(y_{n+k}) \; . \end{aligned} \tag{3.88}$$

(III) Sei y_{n+k} wie in (3.45) definiert. Sind $\hat{\varrho}(\xi) = \sum_{j=0}^{k} \hat{\alpha}_j \xi^j$ und $\hat{\sigma}(\xi) = \sum_{j=0}^{k} \hat{\beta}_j \xi^j$ die charakteristischen Polynome eines zweiten Mehrschrittverfahrens mit $\hat{\alpha}_k \neq 0$ und $\hat{\beta}_k \neq 0$, so wird z_{n+k} bestimmt durch

$$\begin{aligned} \hat{\alpha}_k \tilde{y}_{n+k} + \sum_{j=0}^{k-1} \hat{\alpha}_j y_{n+j} &= h \sum_{j=0}^{k} \hat{\beta}_j f(y_{n+j}, z_{n+j}) \; , \\ 0 &= g(\tilde{y}_{n+k}) \; . \end{aligned} \tag{3.89}$$

(Gleichung (3.89) dient zur Definition von z_{n+k}, der Hilfsvektor $\tilde{y}_{n+k}$ muß nicht explizit berechnet werden.)

Lemma 11 *Für die Verfahren (I)–(III) ist die Voraussetzung (3.46) an die Verfahrensfunktion Ψ in (3.45) erfüllt. Die Koeffizienten $\hat{\beta}_j$, $(j = 0, 1, \ldots, k)$ in (3.45) sind dabei*

für (I) gegeben durch $\hat{\beta}_j = \delta_{kj}$,

für (II) gegeben durch $\hat{\beta}_j = \beta_j - \tau_j$ und

für (III) durch die Koeffizienten des charakteristischen Polynoms $\hat{\sigma}$ gegeben.

Beweis Für (I) trivial. Für (II) ist z_{n+k} und damit auch die Verfahrensfunktion Ψ implizit durch $R(y_n, \ldots, y_{n+k}, z_n, \ldots, z_{n+k}) = 0$ mit

$$R = g\Big(-\sum_{j=0}^{k-1} \frac{\alpha_j}{\alpha_k} y_{n+j} + h \sum_{j=0}^{k-1} \frac{\beta_j}{\alpha_k} f(y_{n+j}, z_{n+j}) + h \frac{\beta_k}{\alpha_k} f(y_{n+k}, z_{n+k} - \sum_{j=0}^{k} \frac{\tau_j}{\beta_k} z_{n+j})\Big)$$

definiert. Wegen der Regularität von $[g_y f_z](y(t_n), z(t_n))$ folgt die Behauptung mit

$r(\vartheta) := R(y_n + \vartheta(\hat{y}_n - y_n), \ldots, y_{n+k} + \vartheta(\hat{y}_{n+k} - y_{n+k}), z_n + \vartheta(\hat{z}_n - z_n), \ldots, z_{n+k} + \vartheta(\hat{z}_{n+k} - z_{n+k}))$

aus $r(0) = r(1) = 0$ und $r(1) - r(0) = \int_0^1 r'(\vartheta)\, d\vartheta$. Analog beweist man (III). ∎

Bemerkung 31 a) Für (I) erfüllt $\hat{\sigma}(\xi) = \sum_{j=0}^{k} \hat{\beta}_j \xi^j = \xi^k$ trivialerweise die strenge Wurzelbedingung, so daß das PLMSV mit (I) mit der Ordnung $q = p$ in y (und damit nach dem Satz über die implizite Funktion auch mit der Ordnung $q = p$ in z) konvergiert, wenn das durch (ϱ, σ) charakterisierte Mehrschrittverfahren die (klassische) Ordnung p hat und ϱ die Wurzelbedingung erfüllt.

b) β-geblockte Verfahren berechnen y_{n+k} und z_{n+k} simultan, durch Vergleich von (3.45) und (3.88) ergibt sich die Darstellung $\zeta_{n+k} = -(\sum_{j=0}^{k} \tau_j z_{n+j})/\beta_k$. Die Koeffizienten τ_j, $(j = 0, 1, \ldots, k)$ müssen nun einerseits so gewählt werden, daß das charakteristische Polynom $\hat{\sigma}(\xi) = \sum_{j=0}^{k} \hat{\beta}_j \xi^j = \sum_{j=0}^{k} (\beta_j - \tau_j)\xi^j$ die strenge Wurzelbedingung erfüllt, und andererseits so, daß der lokale Fehler eine möglichst hohe Ordnung hat. In [7] wird u. a. untersucht, unter welchen Voraussetzungen man Parameter τ_j finden kann, so daß ein k-Schritt-Verfahren der klassischen Ordnung $p = k + 1$ mit der Ordnung $q = k + 1$ in y und mit der Ordnung $q - 1 = k$ in z konvergiert. Für Adams–Moulton–Verfahren (ϱ, σ) ist dies möglich für $k \leq 3$ und unmöglich für $k > 3$. Sind $\varrho := \sum_j \alpha_j \xi^j$ und $\sigma := \sum_j \beta_j \xi^j$ durch DCBDF („difference corrected backward differentiation formulae“ [150]) gegeben, so erhält man ein entsprechendes β-geblocktes Verfahren für $k \leq 6$, d. h. immer dann, wenn das Verfahren für gewöhnliche Differentialgleichungen konvergiert.

c) Die Verfahren des Typs (III) berechnen (formal) y_{n+k} und z_{n+k} nacheinander, die charakteristischen Polynome (ϱ, σ) und $(\hat{\varrho}, \hat{\sigma})$ können unabhängig voneinander so gewählt werden, daß ϱ die Wurzelbedingung und $\hat{\sigma}$ die strenge Wurzelbedingung erfüllt und gleichzeitig die lokalen Fehler hinreichend klein sind. Für nicht-steife Index-2-Systeme (3.19) ist es besonders vorteilhaft, Adams–Moulton–Verfahren (ϱ, σ) mit BDF $(\hat{\varrho}, \hat{\sigma})$ zu koppeln, um die Vorzüge eines sehr effektiven Verfahrens für nicht-steife Differentialgleichungen (vgl. [82, Kapitel III]) zu verbinden mit $\hat{\sigma}(\xi) = \hat{\beta}_k \xi^k$.

Definition 8 Zu einem gegebenen $k \geq 1$ seien $\varrho = \sum_{j=0}^{k} \alpha_j \xi^j$ und $\sigma = \sum_{j=0}^{k} \beta_j \xi^j$ die charakteristischen Polynome des Adams–Moulton–Verfahrens und $\hat{\varrho} = \sum_{j=0}^{k} \hat{\alpha}_j \xi^j$ und $\hat{\sigma} = \sum_{j=0}^{k} \hat{\beta}_j \xi^j$ die charakteristischen Polynome der BDF. Dann heißt das PLMSV (3.45) mit der durch (3.89) definierten Verfahrensfunktion Ψ *partitioniertes lineares Mehrschrittverfahren vom Adams–Typ.*

Als Spezialfälle von Satz 15 ergeben sich Konvergenzaussagen für β-geblockte Verfahren und für PLMSV vom Adams–Typ.

Folgerung 4 (vgl. [84, Satz VII.6.5])
Gegeben sei ein k-Schritt-PLMSV (3.45) mit der durch (3.88) definierten Verfahrensfunktion Ψ und $k \geq 2$, $\alpha_k \neq 0$, $\beta_k \neq 0$, $\hat{\beta}_k = \beta_k - \tau_k \neq 0$. Hat das durch (ϱ, σ) charakterisierte Mehrschrittverfahren die klassische Ordnung $p = k + 1$, gilt

$$\sum_{j=0}^{k} \tau_j e^{jh} = \mathcal{O}(h^k) \,, \quad (h \to 0) \tag{3.90}$$

und erfüllt ϱ die Wurzelbedingung sowie $\hat{\sigma}(\xi) = \sum_{j=0}^{k} (\beta_j - \tau_j)\xi^j$ die strenge Wurzelbedingung, so konvergiert das β-geblockte Mehrschrittverfahren (3.88) mit der Ordnung

$q = p = k + 1$ in y und mit der Ordnung $q - 1 = k$ in z, wenn die Anfangswerte die Bedingungen

$$\|y_i - y(t_i)\| = \mathcal{O}(h^{k+2}) \,, \quad \|z_i - z(t_i)\| = \mathcal{O}(h^k) \,, \quad (i = 0, 1, \ldots, k-1)$$

erfüllen.

Beweis Setzt man in (3.88) $\zeta_{n+k} := -(\sum_{j=0}^{k} \tau_j z_{n+j})/\beta_k$, so ist $\|\delta y_h(t_n)\| = \mathcal{O}(h^{p+1})$, $\|\delta \zeta_h(t_n)\| = \mathcal{O}(h^p)$ nach Lemma 10. Für den Vektor $\hat{\zeta}_{n+k}$ aus Definition 7 gilt wegen $\hat{z}_{n+j} = z(t_{n+j})$, $(j = 0, 1, \ldots, k-1)$

$$\hat{\zeta}_{n+k} = -\sum_{j=0}^{k} \frac{\tau_j}{\beta_k} \hat{z}_{n+j} = -\frac{\tau_k}{\beta_k}(\hat{z}_{n+k} - z(t_{n+k})) - \frac{1}{\beta_k} \sum_{j=0}^{k} \tau_j z(t_{n+j}) \,,$$

also

$$\delta \zeta_h(t_n) = \hat{z}_{n+k} + \hat{\zeta}_{n+k} - z(t_{n+k}) = (1 - \frac{\tau_k}{\beta_k})(\hat{z}_{n+k} - z(t_{n+k})) + \mathcal{O}(h^k) \,,$$

denn aus (3.90) folgt $\sum_{j=0}^{k} \tau_j z(t_{n+j}) = \mathcal{O}(h^k)$ (vgl. [82, Satz III.2.4]). Damit ist aber wegen $0 \neq \hat{\beta}_k = \beta_k - \tau_k$

$$\|\delta z_h(t_n)\| = \|\hat{z}_{n+k} - z(t_{n+k})\| = \mathcal{O}(h^k)$$

und das PLMSV mit der durch (3.88) gegebenen Verfahrensfunktion Ψ erfüllt die Voraussetzungen von Satz 15 mit $q_y = p = k + 1$ und $q_z = k$. ∎

Bemerkung 32 a) Die Konvergenz von β-geblockten Mehrschrittverfahren für Index-2-Systeme (3.19) wurde erstmals von Arévalo, Führer und Söderlind nachgewiesen. Sie untersuchen β-geblockte Verfahren, die auf einem durch (ϱ, σ) charakterisierten linearen k-Schritt-Verfahren der (klassischen) Ordnung p mit $p \in \{k, k+1\}$ basieren. Unter ähnlichen Voraussetzungen wie in Folgerung 4 zeigt Satz 2.1 aus [7] für den Spezialfall von Index-2-Systemen (3.19) mit $f(y, z) = f_0(y) + f_z(y)z$, daß das β-geblockte Verfahren mit der Ordnung $q = p$ in y und mit der Ordnung k in z konvergiert. Später wurde diese Aussage auf beliebige Index-2-Systeme (3.19) erweitert ([6]). Satz 15 und Folgerung 4 der vorliegenden Arbeit beschränken sich auf den Fall $p = k + 1$, durch Kombination der Beweisidee von Folgerung 4 mit Satz VII.3.6 aus [84] kann jedoch auch für $p = k$ die Konvergenz nachgewiesen werden, dabei beträgt die Ordnung $q = k$ in y und z ([84, Satz VII.6.5]).

b) Die Voraussetzung $k \geq 2$ in Folgerung 4 kann wie in [7] abgeschwächt werden zu $k \geq 1$, denn für β-geblockte Mehrschrittverfahren läßt sich die Abschätzung (3.61) aus Bemerkung 20b zeigen.

Folgerung 5 *Gegeben sei ein PLMSV (3.45) mit der durch (3.89) definierten Verfahrensfunktion Ψ ($k \geq 2$, $\alpha_k \neq 0$, $\hat{\alpha}_k \neq 0$, $\beta_k \neq 0$, $\hat{\beta}_k \neq 0$). Hat das durch (ϱ, σ) charakterisierte Mehrschrittverfahren die klassische Ordnung $p = k + 1$ und das durch $(\hat{\varrho}, \hat{\sigma})$ charakterisierte Mehrschrittverfahren die klassische Ordnung k und erfüllt ϱ die Wurzelbedingung sowie $\hat{\sigma}$ die strenge Wurzelbedingung, so konvergiert das PLMSV mit*

der Ordnung $q = p = k+1$ *in* y *und mit der Ordnung* $q-1 = k$ *in* z, *wenn für die Anfangswerte*

$$\|y_i - y(t_i)\| = \mathcal{O}(h^{k+2}) \,, \quad \|z_i - z(t_i)\| = \mathcal{O}(h^k) \,, \quad (i = 0, 1, \ldots, k-1)$$

gilt.

Beweis Aus Lemma 10 folgt

$$\|\delta y_h(t)\| = \mathcal{O}(h^{k+2}) \,, \quad \|\delta \zeta_h(t)\| = \mathcal{O}(h^{k+1}) \,, \quad \|\delta z_h(t)\| = \mathcal{O}(h^k) \,,$$

so daß die Voraussetzungen von Satz 15 mit $q_y = k+1$ und $q_z = k$ erfüllt sind. ∎

Bemerkung 33 a) Für partitionierte k-Schritt-Verfahren vom Adams–Typ mit $k \geq 2$ ergibt sich also die Konvergenzordnung $q = k+1$ in y und $q-1 = k$ in z. Bemerkenswert ist, daß die Verfahren auch für $k > 6$ konvergieren, obwohl $(\hat{\varrho}, \hat{\sigma})$ durch BDF definiert werden (Folgerung 5 trifft keine Voraussetzungen bezüglich der Nullstellen von σ und $\hat{\varrho}$). Damit überwinden die PLMSV vom Adams–Typ die durch die für $k > 6$ fehlende Nullstabilität von BDF und DCBDF diktierte Ordnungsbarriere $q \leq 7$ der β-geblockten Mehrschrittverfahren aus [7].

b) Ebenso wie für β-geblockte Verfahren kann für PLMSV vom Adams–Typ die Voraussetzung $k \geq 2$ in Folgerung 5 abgeschwächt werden zu $k \geq 1$, denn nach [84, Satz VII.3.2] gilt die Abschätzung (3.61) aus Bemerkung 20b.

Bemerkung 34 Sind die charakteristischen Polynome $(\hat{\varrho}, \hat{\sigma})$ in (3.89) durch BDF definiert, so ist $\tilde{y}_{n+k}$ allein durch $y_n, \ldots, y_{n+k}$ bestimmt und hängt nicht von $z_n, \ldots, z_{n+k}$ ab (vgl. [84, Satz VII.3.5]). Es gilt nach Formel (VII.3.22) aus [84]

$$\|\tilde{y}_{n+k} - y(t_{n+k})\| \leq \mathcal{O}(1) \sum_{j=0}^{k} \|y_{n+j} - y(t_{n+j})\| + \mathcal{O}(h^{k+1}) \,.$$

Wie im Beweis von Satz 15 (vgl. (3.58)) kann man deshalb nachweisen, daß bei Anwendung eines partitionierten k-Schritt-Verfahrens vom Adams–Typ auf ein Index-2-System (3.19) mit $f(y,z) = \tilde{f}(y,z,d(y,z))$ die Aussage von Folgerung 5 unverändert gültig bleibt, wenn in (3.45) und (3.89) statt $f(y_{n+j}, z_{n+j}) = \tilde{f}(y_{n+j}, z_{n+j}, d(y_{n+j}, z_{n+j}))$ der Funktionswert $\tilde{f}(y_{n+j}, z_{n+j}, d(\tilde{y}_{n+j}, z_{n+j}))$ verwendet wird.
Index-2-Systeme (3.19), für die f diese spezielle Struktur hat, entstehen, wenn man formal ein semi-explizites Index-2-System

$$\begin{aligned} y' &= \varphi(y,w) \,, \\ 0 &= \gamma(y,w) \end{aligned} \tag{3.91}$$

transformiert in ein System (3.19) in Hessenbergform (hierbei sei vorausgesetzt, daß die Jacobimatrix γ_w in einer Umgebung der analytischen Lösung konstanten Rang hat). Unter Verwendung des Satzes über die implizite Funktion können die algebraischen Gleichungen in (3.91) nach einem Teil der algebraischen Variablen aufgelöst werden, d. h., es gibt eine Zerlegung des Vektors w in Komponenten s und z und hierzu Funktionen

$d : \Omega_y \times \Omega_z \to \mathbb{R}^{n_s}$, $g : \Omega_y \to \mathbb{R}^{n_z}$ mit $\Omega_y \subset \mathbb{R}^{n_y}$ und $\Omega_z \subset \mathbb{R}^{n_z}$, so daß $\gamma(y,w) = 0$ (lokal) äquivalent ist zu $s = d(y,z)$ und $g(y) = 0$ ([84, S. 456]). Setzt man z und $s = d(y,z)$ in $\varphi(y,w)$ ein, so erhält man ein zu (3.91) äquivalentes System (3.19) in Hessenbergform mit $f(y,z) = \tilde{f}(y,z,d(y,z))$.
Wird ein PLMSV vom Adams–Typ auf (3.91) angewendet, so führt diese formale Transformation wegen $0 = \gamma(\tilde{y}_{n+k}, w_{n+k})$ auf $s_{n+k} = d(\tilde{y}_{n+k}, z_{n+k})$ und damit auf Funktionswerte $\tilde{f}(y_{n+j}, z_{n+j}, d(\tilde{y}_{n+j}, z_{n+j}))$ in (3.45) und (3.89). Zusammenfassend ergibt sich, daß PLMSV vom Adams–Typ nicht nur für Index-2-Systeme in Hessenbergform, sondern allgemein für Index-2-Systeme der semi-expliziten Struktur (3.91) mit der Ordnung $q = k+1$ in den differentiellen Komponenten y und mit der Ordnung $q = k$ in den algebraischen Komponenten w konvergieren. Ein Spezialfall von (3.91) ist die Index-2-Formulierung der Modellgleichungen von MKS mit Kontaktbedingungen (vgl. (3.14) und Bemerkung 29). Hier ist $s = d(y)$ und man erhält deshalb die Konvergenzordnung $k+1$ nicht nur für y, sondern auch für s.

Bei der Implementierung von Mehrschrittverfahren für nicht-steife DA–Systeme werden in Analogie zu HERK–Verfahren nur die algebraischen Komponenten als Lösung von (je nach Anwendung linearen oder nichtlinearen) Gleichungssystemen berechnet, während man die differentiellen Komponenten y durch Funktionaliteration bestimmt (vgl. [82, S. 360], [154, Kapitel 4.6] und speziell für die Anwendung auf Index-1-Systeme auch [35]). Für Index-2-Systeme (3.19) in Hessenbergform wird ein $\mathrm{P(EC)^M E}$–Schema für PLMSV vom Adams–Typ durch folgenden Algorithmus realisiert:

Algorithmus 2

Schritt 0 Berechne Prädiktor $y_{n+k}^{(0)}$ mit explizitem k-Schritt–Adams–Verfahren.
Setze $l := 0$.

Schritt 1 Berechne $\zeta \in \mathbb{R}^{n_z}$ als Lösung des Gleichungssystems

$$0 = \Phi_{\mathrm{Adams}}(\zeta) := g\Big(r_{\mathrm{Adams}} + h\frac{\beta_k}{\alpha_k} f(y_{n+k}^{(l)}, \zeta)\Big) \,.$$

Schritt 2 Berechne $y_{n+k}^{(l+1)} := r_{\mathrm{Adams}} + h\frac{\beta_k}{\alpha_k} f(y_{n+k}^{(l)}, \zeta)$.

Schritt 3 Setze $l := l+1$. Ist $l < M$, so gehe zu Schritt 1, sonst gehe zu Schritt 4.

Schritt 4 Berechne $\zeta \in \mathbb{R}^{n_z}$ als Lösung des Gleichungssystems

$$0 = \Phi_{\mathrm{BDF}}(\zeta) := g\Big(r_{\mathrm{BDF}} + h\frac{\hat{\beta}_k}{\hat{\alpha}_k} f(y_{n+k}^{(M)}, \zeta)\Big) \,.$$

Schritt 5 Setze $y_{n+k} := y_{n+k}^{(M)}$, $z_{n+k} := \zeta$.

Hierbei ist

$$\begin{aligned} r_{\mathrm{Adams}} &:= -\sum_{j=0}^{k-1} \frac{\alpha_j}{\alpha_k} y_{n+j} + h \sum_{j=0}^{k-1} \frac{\beta_j}{\alpha_k} f(y_{n+j}, z_{n+j}) = y_{n+k-1} + h \sum_{j=0}^{k-1} \frac{\beta_j}{\alpha_k} f(y_{n+j}, z_{n+j}) \,, \\ r_{\mathrm{BDF}} &:= -\sum_{j=0}^{k-1} \frac{\hat{\alpha}_j}{\hat{\alpha}_k} y_{n+j} + h \sum_{j=0}^{k-1} \frac{\hat{\beta}_j}{\hat{\alpha}_k} f(y_{n+j}, z_{n+j}) = -\sum_{j=0}^{k-1} \frac{\hat{\alpha}_j}{\hat{\alpha}_k} y_{n+j} \,. \end{aligned}$$

Diese Implementierung unterscheidet sich ausschließlich in Schritt 4 von einem P(EC)^{M}E–Schema für die Anwendung des klassischen Adams–Moulton–Verfahrens auf Index-2-Systeme (3.19). Im Gegensatz zum klassischen Adams–Moulton–Verfahren ist jedoch das PLMSV vom Adams–Typ stabil. Wie für gewöhnliche Differentialgleichungen zeigt man, daß das nach Algorithmus 2 implementierte PLMSV für jedes $M \geq 1$ die Konvergenzordnung $q = k + 1$ in y und $q - 1 = k$ in z hat, d. h., im P(EC)^{M}E–Schema ist ein Korrektorschritt ausreichend, um die Konvergenzordnung des PLMSV zu erreichen (vgl. [154, Satz 4.6.1 und Bemerkung 4.6.3.(2)]).

Im Vergleich zu den BDF, die bei Anwendung auf das Index-2-System (3.19) für $k \leq 6$ mit der Ordnung k in y und z konvergieren ([84, Satz VII.3.5]), konvergieren die PLMSV vom Adams–Typ für beliebiges $k \geq 1$ und erreichen in y eine höhere Konvergenzordnung. Für die Implementierung mit variabler Ordnung und variabler Schrittweite erfordern die partitionierten Verfahren dagegen einen wesentlich höheren Organisationsaufwand („overhead“), da in Algorithmus 2 zusätzlich zu r_{BDF} auch r_{Adams} zu berechnen ist.

Abschließend werden die hier neu eingeführten PLMSV vom Adams–Typ unabhängig von diesen Details der Implementierung mit anderen Verfahren verglichen. Hierzu wird mit verschiedenen Verfahren das schon in Abschnitt 3.3.1 betrachtete Lastwagenmodell in der Variante von Führer ([56]) mit konstanter Schrittweite integriert. Dieses Benchmark–Problem ist besonders für Vergleiche von Mehrschrittverfahren geeignet, weil exakte Anfangswerte zur Verfügung stehen (für $t \leq 0$ liegt ein stationärer Zustand vor, außerdem wurde — im Unterschied zu Beispiel A aus Abschnitt 3.3.2 — eine unendlich oft stetig differenzierbare Anregungsfunktion $u(t)$ mit $u(t) = 0$, $(t \leq 0)$ verwendet).

In den Vergleichsrechnungen werden alle Mehrschrittverfahren in der P(EC)^{2}E–Technik implementiert, durch zusätzliche Korrektoriterationsschritte ändert sich der globale Fehler nur geringfügig. Damit erfordern die Verfahren je Integrationsschritt (unabhängig von k) die Lösung dreier Gleichungssysteme der Dimension n_z und 3 Aufrufe von f.

Ebenso wie in den Abb. 3.5 und 3.6 werden wieder für zahlreiche fixierte Schrittweiten h der globale Diskretisierungsfehler und die Anzahl der Funktionsaufrufe einander gegenübergestellt.

In Abb. 3.10 werden für verschiedene k die partitionierten k-Schritt-Verfahren vom Adams–Typ verglichen. Mit wachsender Ordnung verringert sich der Aufwand für die Berechnung einer Lösung von gewünschter Genauigkeit erheblich. Die Anstiege der Graphen entsprechen (für kleine Schrittweiten h) den theoretisch vorhergesagten Konvergenzordnungen $q = k + 1$ (in y) bzw. $q - 1 = k$ (in z).

Alternativen zu den PLMSV vom Adams–Typ sind z. B. für $k \leq 6$ die BDF und für $k \leq 3$ die β-geblockten Adams–Moulton–Verfahren. In Abb. 3.11 werden diese 3 Verfahrensklassen am Beispiel $k = 3$ verglichen. Bezüglich des Fehlers in y sind die beiden auf Adams–Verfahren basierenden partitionierten Verfahren vollkommen gleichwertig (Ordnung $q = k + 1 = 4$) und den BDF (Ordnung $q = k = 3$) deutlich überlegen. Wegen eines kleineren lokalen Fehlers $\delta z_h(t)$ und wegen der besseren Dämpfung von $\delta z_h(t)$ führt das PLMSV vom Adams–Typ auf kleinere Fehler in z als das β-geblockte Verfahren (im Beweis von Satz 15 ist $\kappa = 0$ für das PLMSV vom Adams–Typ und $\kappa > 0.71$ für das β-geblockte Verfahren).

Schließlich zeigt Abb. 3.12 den Vergleich dreier Verfahren, die für y die Konvergenzordnung $k = 5$ erreichen. Trotz des wesentlich höheren Aufwands je Integrationsschritt

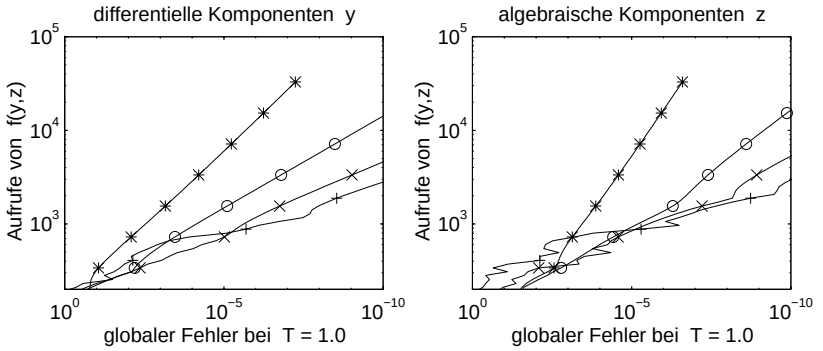


Abbildung 3.10: Aufwand und globaler Diskretisierungsfehler der PLMSV vom Adams–Typ mit $k = 2$ („$*$“), $k = 4$ („$\circ$“), $k = 6$ („$\times$“), $k = 8$ („$+$“): Benchmark Lastwagenmodell ([149], [56]).

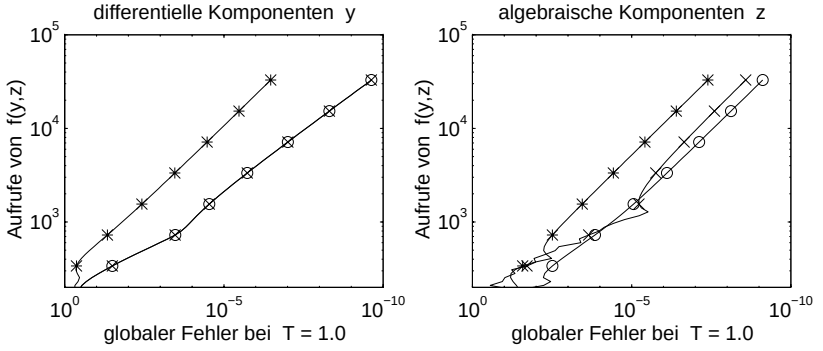


Abbildung 3.11: Aufwand und globaler Diskretisierungsfehler von k-Schritt-Verfahren mit $k = 3$: BDF („$*$“), PLMSV vom Adams–Typ („$\circ$“), β-geblocktes Adams–Moulton–Verfahren mit $\tau_0 = 0.1$, $\tau_1 = -0.3$, $\tau_2 = 0.3$, $\tau_3 = -0.1$ („$\times$“, [7]): Benchmark Lastwagenmodell ([149], [56]).

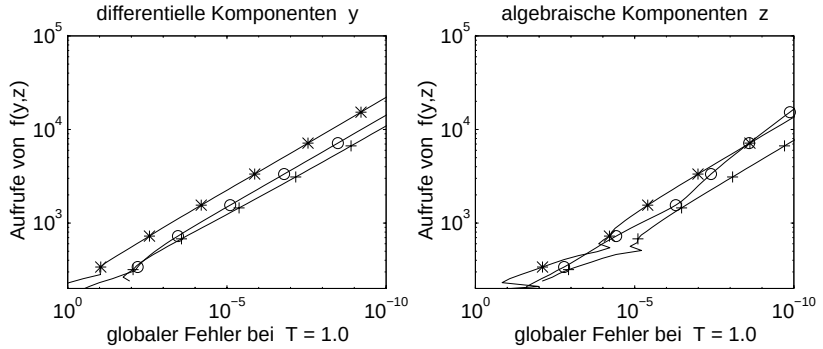


Abbildung 3.12: Aufwand und globaler Diskretisierungsfehler von Verfahren der Konvergenzordnung $q = 5$ in y: BDF mit $k = 5$ („∗“), PLMSV vom Adams–Typ mit $k = 4$ („o“), HEDOP5 („+“): Benchmark Lastwagenmodell ([149], [56]).

erweist sich hierbei das halb-explizite Runge–Kutta–Verfahren als sehr effizient. Von den beiden Mehrschrittverfahren hat die BDF wegen der größeren Fehlerkonstanten die größeren Fehler in y. Obwohl das PLMSV und das HERK–Verfahren nur die Konvergenzordnung $q - 1 = 4$ in z haben (BDF: Konvergenzordnung $q = p = k = 5$ in z), sind die BDF auch bezüglich des Fehlers in z keineswegs deutlich überlegen.

3.3.4 Zusammenfassung

In den Abschnitten 3.2 und 3.3 wurden Verfahren konstruiert, analysiert und in einem Fall als Integrator implementiert, die speziell zur Lösung nicht-steifer DA–Systeme vom Index 2 geeignet sind. Ausgehend von den aus der Literatur bekannten Konvergenzaussagen für implizite Runge–Kutta–Verfahren und für Mehrschrittverfahren kann man unter geeigneten Voraussetzungen sowohl für HERK–Verfahren als auch für PLMSV die Konvergenz nachweisen (Abschnitt 3.2); dabei ergibt sich für die differentiellen Komponenten y eine höhere Konvergenzordnung als für die algebraischen Komponenten z.

Neu eingeführte Verfahrensklassen (HERK–Verfahren mit expliziter Stufe, PLMSV vom Adams–Typ) erlauben es, effiziente Integrationsverfahren für nicht-steife gewöhnliche Differentialgleichungen (explizite Runge–Kutta–Verfahren höherer Ordnung, implizite Adams–Verfahren) mit entsprechenden Modifikationen auch für die Integration von Index-2-Systemen in Hessenbergform einzusetzen. Dabei sind nur zur Bestimmung der algebraischen Komponenten Gleichungssysteme zu lösen, die differentiellen Komponenten werden explizit bzw. mittels Funktionaliteration berechnet.

Für praxisnahe Berechnungen wurde auf der Grundlage des expliziten Runge–Kutta–Verfahrens 5. Ordnung von Dormand und Prince der leistungsfähige Integrator HEDOP5 zur dynamischen Simulation von mechanischen Mehrkörpersystemen in Deskriptorform implementiert, getestet und in die Programmbibliothek MBSPACK integriert.

Kapitel 4

Differentiell-algebraische Systeme und die dynamische Simulation von mechanischen Mehrkörpersystemen mit Kontaktbedingungen

In vielfältigen praktischen Anwendungen ist es sinnvoll, reale physikalische oder technische Systeme als mechanische Mehrkörpersysteme (MKS) zu modellieren. Ein solches MKS besteht aus endlich vielen Körpern und Verbindungselementen, wobei angenommen wird, daß die Masse des Systems ausschließlich in den Körpern des Systems konzentriert ist. Als Körper eines MKS betrachtet man sowohl Starrkörper als auch elastische Körper. Typische Verbindungselemente sind Gelenke, Federn, Dämpfer und mechatronische Elemente, die auch aktive Regelungskomponenten enthalten können.

Die dynamische Simulation von MKS ist eines derjenigen Anwendungsgebiete, die immer wieder den Anstoß zur Verbesserung der Diskretisierungsverfahren für DA–Systeme gegeben haben. In Industrie–Anwendungen (z. B. Robotik, Fahrzeugbau) werden dabei die Modellgleichungen in der Regel unter Verwendung von Mehrkörperformalismen automatisch generiert (vgl. z. B. [141]). Dabei wird dem MKS in natürlicher Weise ein Graph zugeordnet, dessen Knoten den Körpern des MKS und dessen Kanten den Verbindungselementen entsprechen. Während sich für MKS mit Baumstruktur Systeme gewöhnlicher Differentialgleichungen als Modellgleichungen ergeben, führen kinematisch geschlossene Schleifen im MKS und die Kopplung von Substrukturen (d. h. einzelner Baugruppen im MKS, deren Modellgleichungen unabhängig voneinander generiert wurden) zu Zwangsbedingungen und damit zu DA–Systemen.

Zahlreiche Integrationsverfahren wurden der speziellen Struktur dieser Modellgleichungen angepaßt (vgl. [57] für einen Überblick). Deshalb versucht man, auch kompliziertere mechanische und mechatronische Systeme auf ähnliche Weise zu modellieren. Im vorliegenden Kapitel konzentrieren wir uns auf eine solche Erweiterung des klassischen Konzepts der Euler–Lagrangeschen Bewegungsgleichungen: Berühren sich in einem MKS permanent zwei Starrkörper, so kann auch diese Kontaktbedingung als eine skalare Zwangsbedingung formuliert werden (vgl. z. B. [39, S. 5]).

Zunächst wird in Abschnitt 4.1 dieses Modell vorgestellt, die entsprechenden Modell-

gleichungen werden formuliert, und es wird gezeigt, wie sie im Spezialfall glatter Zwangsbedingungen mit einem Integrator für DA–Systeme effizient gelöst werden können. Die Erweiterung auf MKS mit Kontaktbedingungen, die auf nur stückweise differenzierbare Zwangsbedingungen führen, ist Gegenstand der Abschnitte 4.2 und 4.3. Hierzu werden 2 Strategien vorgeschlagen und am Beispiel der dynamischen Simulation von Rad–Schiene–Systemen miteinander verglichen: Dies ist einerseits die Integration der Modellgleichungen unter Berücksichtigung der unstetigen Zustandsänderungen (Abschnitt 4.2) und andererseits die Regularisierung der Kontaktbedingungen durch ein quasi-elastisches Kontaktmodell (Abschnitt 4.3). Aus physikalischen Gründen ist der zweite Ansatz für Rad–Schiene–Systeme, deren Räder ein sog. Verschleißprofil haben, zu bevorzugen.

Im abschließenden Abschnitt 4.4 werden verschiedene Aspekte der effizienten numerischen Umsetzung dieses quasi-elastischen Kontaktmodells beschrieben. Durch eine geeignete Implementierung konnte dabei im Rahmen des MKS–Simulationspakets SIMPACK die für die Simulation des Rad–Schiene–Kontakts erforderliche Rechenzeit so stark reduziert werden, daß nun auch komplizierte und große Simulationsaufgaben aus industriellen Anwendungen des Schienenfahrzeugbaus gelöst werden können.

4.1 Modellgleichungen für mechanische Mehrkörpersysteme mit Kontaktbedingungen

Die Bewegung mechanischer Mehrkörpersysteme wird durch Gleichungen beschrieben, die nach den Prinzipien der klassischen Mechanik aufgestellt werden. Im folgenden wird hierzu die kompakte Einführung von Hairer und Wanner ([84, S. 463f]) wiedergegeben, weil sie in aller Kürze die anschließend immer wieder verwendeten Ideen und Bezeichnungen zusammenfaßt (ausführliche Darstellungen sind in Mechanik–Lehrbüchern zu finden, z. B. [42, Teil II.A]). Wie in der Mechanik üblich gebrauchen wir in diesem Kapitel die Schreibweise $\frac{d}{dt}q(t) = \dot{q}(t)$.

Sind $q_1, \ldots, q_{n_q}$ die (verallgemeinerten) Koordinaten eines konservativen Systems mit der potentiellen Energie $U(q)$ und der kinetischen Energie $T(q,\dot{q})$, so ergeben sich die Bewegungsgleichungen nach Lagrange (Lagrangesche Gleichungen 2. Art) aus der Forderung, daß für $L(q,\dot{q}) := T(q,\dot{q}) - U(q)$ das Integral $\int_{t_0}^{t_1} L(q,\dot{q})\,dt$ minimal werden soll. Die Euler–Gleichungen zu diesem Variationsproblem lauten

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} - \frac{\partial L}{\partial q_k} = 0\ ,\quad (k = 1, \ldots, n_q)\ , \tag{4.1}$$

also

$$\sum_{l=1}^{n_q} L_{\dot{q}_k \dot{q}_l}\ddot{q}_l = L_{q_k} - \sum_{l=1}^{n_q} L_{\dot{q}_k q_l}\dot{q}_l\ ,\quad (k = 1, \ldots, n_q)\ . \tag{4.2}$$

Wird die Menge der zulässigen Zustände durch n_λ ideale Zwangsbedingungen $g_1(q) = 0$, $\ldots$, $g_{n_\lambda}(q) = 0$ beschränkt, so ist statt $L = T - U$ die Lagrange–Funktion

$$L(q,\dot{q},\lambda) = T(q,\dot{q}) - U(q) - \sum_{k=1}^{n_\lambda} \lambda_k g_k(q) \tag{4.3}$$

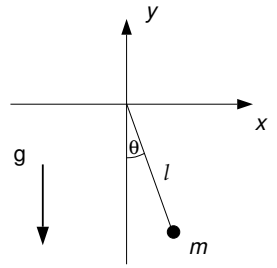
zu verwenden, in der die Zwangsbedingungen durch Lagrange–Multiplikatoren λ_k angekoppelt werden. Betrachtet man in (4.1) die Ableitung nach λ_k, so ergibt sich gerade die Zwangsbedingung $g_k(q) = 0$, denn L hängt nicht von $\dot{\lambda}_k$ ab.

Beispiel 23 Beschreibt man die Bewegung des in der Abbildung dargestellten mathematischen Pendels der Länge l durch den Auslenkungswinkel $\theta =: q_1$, so ist $T = ml^2\dot{\theta}^2/2$ und $U = -lmg\cos\theta$. Aus (4.2) folgt dann die bekannte Bewegungsgleichung $l\ddot{\theta} = -g\sin\theta$. Bei Modellierung in kartesischen Koordinaten ist dagegen die Zwangsbedingung $x^2 + y^2 = l^2$ zu erfüllen und mit der Lagrange–Funktion

$$L = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) - mgy - \lambda(x^2 + y^2 - l^2)$$

ergeben sich die Bewegungsgleichungen

$$\begin{aligned} m\ddot{x} &= -2x\lambda\ , \\ m\ddot{y} &= -mg - 2y\lambda\ , \\ 0 &= x^2 + y^2 - l^2\ . \end{aligned}$$



Mathematisches Pendel.

Auch für ein konservatives System mit Zwangsbedingungen folgen die Bewegungsgleichungen aus den Euler–Gleichungen zu dem Variationsproblem $\int L\,dt \to \min$. Wenn man zusätzlich Reibungskräfte und zeitabhängige Anregungsfunktionen berücksichtigt, so ergibt sich allgemein als *Deskriptorform* der MKS–Modellgleichungen

$$\begin{aligned} M(q)\ddot{q} &= f(q,\dot{q},\lambda,t) - G^T(q,t)\lambda\ , \\ 0 &= g(q,t)\ . \end{aligned} \tag{4.4}$$

Hier bezeichnet $M(q) = T_{\dot{q}\dot{q}}$ die Massenmatrix. M ist symmetrisch und positiv semidefinit, darüberhinaus gilt $\eta^T M(q)\eta > 0$ für alle $\eta \in \mathbb{R}^{n_q}$ mit $\eta \neq 0$ und $G(q,t)\eta = 0$ ($G(q,t) := \frac{\partial}{\partial q}g(q,t)$). In (4.4) werden die im MKS–Modell wirkenden Kräfte getrennt in $f(q,\dot{q},\lambda,t)$ und in die Zwangskräfte $-G^T(q,t)\lambda$, die garantieren, daß die Zwangsbedingungen $g(q,t) = 0$ eingehalten werden. Der Vektor f enthält u. a. die Kräfte, die von außen auf das MKS wirken (z. B. Schwerkraft, Corioliskraft), dabei gilt $f = f(q,\dot{q})$, wenn im MKS ausschließlich die Kräfte $L_{q_k} - \sum_l L_{\dot{q}_k q_l}\dot{q}_l$ aus (4.2) betrachtet werden. Berücksichtigt man jedoch auch die Reibung im MKS, so enthält f außerdem die Reibungskräfte, die i. allg. von den Zwangskräften $-G^T\lambda$ abhängen können.

Bis auf wenige Ausnahmen, die eine gesonderte Betrachtung erfordern (vgl. z. B. [30]), können die Zwangsbedingungen so formuliert werden, daß $G(q,t)$ in einer Umgebung der Lösung von (4.4) Vollrang hat. Prinzipiell beschränken wir uns hier wie schon in den Abschnitten 2.3 und 3.3.2 auf Probleme, für die

$$\begin{pmatrix} M(q) & \Gamma(q,v,\lambda,t) \\ G(q,t) & 0 \end{pmatrix} \quad \text{mit} \quad \Gamma(q,v,\lambda,t) := f_\lambda(q,v,\lambda,t) - G^T(q,t)$$

in einer Umgebung der Lösung (q,v,λ) regulär ist $(v(t) := \dot{q}(t))$. Dann hat (4.4) den Index 3. Im Unterschied zur Deskriptorform (4.4) wird eine zu (4.4) äquivalente Beschrei-

bung in Minimalkoordinaten, die auf ein System (4.2) von gewöhnlichen Differentialgleichungen führt, als *Zustandsform* der Bewegungsgleichungen bezeichnet (vgl. Einleitung).

Sind die Zwangsbedingungen $g(q,t) = 0$ hinreichend oft stetig differenzierbar, so ist — zumindest lokal — stets die Transformation auf eine solche Zustandsform möglich (vgl. z. B. [66], [134], [140]), damit können auch Aussagen über die Existenz und Eindeutigkeit der Lösung von Anfangswertproblemen für (4.4) direkt aus den entsprechenden Ergebnissen für Systeme gewöhnlicher Differentialgleichungen abgeleitet werden (vgl. z. B. [140]).

Typische Erweiterungen von (4.4), die in industrienahen Anwendungen zu betrachten sind, umfassen z. B. nicht-holonome Zwangsbedingungen (z. B. [81, S. 7]), die Kopplung mit partiellen Differentialgleichungen bei elastischen Körpern im MKS (z. B. [146]), die Berücksichtigung von mechatronischen Komponenten und (aktiven) Steuerungskomponenten im MKS und einseitige Beschränkungen an die MKS–Koordinaten q. Durch solche einseitigen Beschränkungen kann man insbesondere im Modell berücksichtigen, daß den idealisierten Körpern des MKS–Modells in der Realität dreidimensionale geometrische Körper entsprechen, die einander zwar berühren, aber nicht durchdringen können. Da hierbei die geometrische Form der Oberflächen dieser Körper zu berücksichtigen ist, sind die Ansätze zur Modellierung und numerischen Behandlung von geometrischen Beschränkungen sehr stark problemspezifisch.

Bemerkung 35 a) Im Mittelpunkt der nachfolgenden Untersuchungen stehen Modelle zur Beschreibung des Kontakts zwischen dem Rad eines Schienenfahrzeugs und der Schiene. Für dieses Problem ist charakteristisch,

- daß für Räder mit Verschleißprofil (s. u.) die Schienenoberfläche im Kontaktbereich nahezu parallel zur Lauffläche des Rades ist,
- daß bei der Simulation eines kompletten Schienenfahrzeugs sehr viele solche Rad–Schiene–Kontakte zu berücksichtigen sind (meist ≥ 8 Räder je Waggon) und
- daß die geometrische Kontaktbedingung mit der Berechnung der zwischen Rad und Schiene wirkenden Reibungskräfte gekoppelt ist.

Im Vergleich zu anderen Anwendungen (z. B. dynamische Simulation eines Industrieroboters mit komplizierter Geometrie, Andock–Manöver einer Raumkapsel an eine Raumstation) erweist es sich als günstig,

- daß man für die dynamische Simulation vereinfachend permanenten Kontakt zwischen Rad und Schiene voraussetzen kann (vgl. Bemerkung 37) und
- daß die Oberflächen der undeformierten Starrkörper Rad und Schiene verhältnismäßig einfach analytisch beschrieben werden können und durch Industriestandards vorgegeben sind (insbesondere ist das Rad rotationssymmetrisch und der Querschnitt des Gleiskopfs ist konvex).

Diese beiden Aspekte sind für eine effiziente numerische Lösung entscheidend.

b) Physikalisch betrachtet ist der Rad–Schiene–Kontakt kein Starrkörperkontakt, denn wegen der Achslast des Schienenfahrzeugs werden die Oberflächen von Rad und Schiene

im Kontaktbereich (elastisch) deformiert, es bildet sich eine Kontakt*fläche* aus ([97]). Vergleicht man jedoch für die Höhenauslenkung eines Rades über der Schiene das Starrkörperkontaktmodell mit elastischen Kontaktmodellen, so ergibt sich lediglich eine Differenz der Größenordnung $10 \ldots 100\,\mu$m, die bei den für die dynamische Simulation typischen Genauigkeitsforderungen (im Bereich von 1%) vernachlässigbar ist. Deshalb ist es bei der Formulierung der geometrischen Kontaktbedingung nicht erforderlich, die elastische Deformation quantitativ zu berücksichtigen. (Dagegen ist die elastische Deformation Kernpunkt der Berechnung der zwischen Rad und Schiene wirkenden Reibungskräfte, die in (4.4) Teil von $f(q,\dot{q},\lambda,t)$ sind.) Verzichtet man bei der geometrischen Modellierung auf die quantitative Berücksichtigung der elastischen Deformation, so vermeidet man in (4.4) Lösungsanteile, die mit hoher Frequenz und sehr kleiner Amplitude schwingen und deshalb zu großen Rechenzeiten führen (vgl. z. B. [107]). Formuliert man die Kontaktbedingungen als Teil der Zwangsbedingungen $g(q) = 0$, so ergibt sich eine erhebliche Rechenzeitersparnis gegenüber einem rein elastischen Modell (in Extremfällen bis zu 90%).

Zahlreiche anwendungsspezifische Details verdecken bei der Formulierung des Rad–Schiene–Kontaktmodells den Blick auf den mathematischen Kern des Problems. Zur Veranschaulichung des prinzipiellen Vorgehens betrachten wir deshalb zunächst in Anlehnung an ein Beispiel von Führer ([57]) ein einfaches Modellproblem:

Beispiel 24 In der Ebene sei eine masselose Kreisscheibe des Radius ϱ gegeben, in deren Mittelpunkt $q = (q_1, q_2)^T$ eine Punktmasse m befestigt wird. Diese Scheibe (einschließlich der Punktmasse) soll sich auf dem durch den Hyperbelast $\eta = \sqrt{\xi^2 + a^2}$ im (ξ,η)-Koordinatensystem definierten Untergrund $\{(\xi,\eta) : \eta \leq \sqrt{\xi^2 + a^2}\}$ so bewegen, daß Scheibe und Untergrund permanent in Kontakt sind, ohne einander zu durchdringen. Der Parameter a spezifiziert dabei genauer den Untergrund, für $a \to 0$ konvergiert

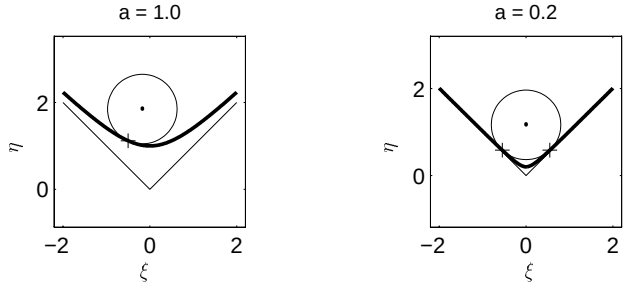


Abbildung 4.1: Geometrie des Modellproblems „Scheibe auf Hyperbelast“ ($\varrho = 0.8$).

der Hyperbelast punktweise gegen $\eta = |\xi|$ (vgl. Abb. 4.1). Der untere Rand der Scheibe $\{(\xi,\eta) : \eta = q_2 - \sqrt{\varrho^2 - (\xi - q_1)^2}\}$ wird durch $\xi \in (q_1 - \varrho, q_1 + \varrho)$ parametrisiert. Ordnet man mit der *Abstandsfunktion* $\Delta : (q_1 - \varrho, q_1 + \varrho) \times \mathbb{R}^2 \to \mathbb{R}$ jedem Punkt des unteren Scheibenrands seine (entlang der η-Achse bestimmte) vorzeichenbehaftete Entfernung zum Hyperbelast zu:

$$\Delta(\xi;q) := q_2 - \zeta(\xi,q_1) \quad \text{mit} \quad \zeta(\xi,q_1) := \sqrt{\varrho^2 - (\xi - q_1)^2} + \sqrt{\xi^2 + a^2}\,, \tag{4.5}$$

so kann die Kontaktbedingung analytisch beschrieben werden durch die (unendlich vielen) einseitigen Beschränkungen

$$\Delta(\xi;q) \geq 0\ , \quad (\ \xi \in (q_1-\varrho, q_1+\varrho)\ ) \tag{4.6}$$

und die Zusatzforderung, daß für (mindestens) ein $\xi_0 \in (q_1-\varrho, q_1+\varrho)$ gilt $\Delta(\xi_0;q)=0$ (in Abb. 4.1 durch „+“ markiert). Damit ergibt sich die Kontaktbedingung

$$0 = \gamma(q) := q_2 - \max_{|\xi-q_1|<\varrho} \zeta(\xi;q_1)\ , \tag{4.7}$$

die zu gegebenem q_1 die Höhenauslenkung q_2 der Scheibe eindeutig festlegt.

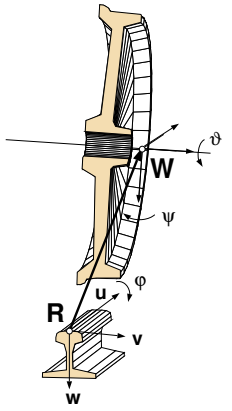
Beispiel 24 verdeutlicht, daß geometrische Beschränkungen in der Regel auf (unendlich viele) einseitige Bedingungen an die MKS–Lagekoordinaten q führen. Wählt man eine geeignete Abstandsfunktion Δ, so läßt sich die Bedingung, daß zwei Starrkörper des MKS permanent in Kontakt sind, als (zweiseitige) Zwangsbedingung $\gamma(q)=0$ formulieren, wobei γ das globale Minimum von Δ bezeichnet.

Bemerkung 36 Durch Übertragung von Beispiel 24 auf das räumliche Kontaktproblem Rad–Schiene stellen wir das Starrkörperkontaktmodell auf. Für Rad–Schiene–Systeme betrachtet man typischerweise eine Bewegung längs eines vorgegebenen Fahrwegs, so daß zu jeder längs des Fahrwegs zurückgelegten Wegstrecke x die lokale Geometrie des Fahrwegs (Gleisprofile, Gleisstörungen, Überhöhung von Gleisbögen, Weichen usw.) bekannt ist. In (4.4) ist $x = x(q)$ durch die MKS–Koordinaten q bestimmt, häufig ist x eine der Lagekoordinaten q.

In sehr guter Näherung kann man lokal die Gleiskrümmung und Änderungen des Gleisprofils vernachlässigen, deshalb wird im weiteren bei der Beschreibung der Kontaktbedingung das Gleis lokal durch einen prismatischen Körper ersetzt. Die nebenstehende Abbildung, die ebenso wie Abb. 4.3 freundlicherweise von Herrn Dipl.-Ing. H. Netter zur Verfügung gestellt wurde, zeigt im Schnitt ein einzelnes Rad über einem solchen Gleis. Zur geometrischen Beschreibung des Rades und des Gleises sind Profilfunktionen $F(s)$ und $G(v;x)$ vorgegeben. Die rotationssymmetrische Radoberfläche wird im Rad–Koordinatensystem $\boldsymbol{W}$ (engl.: **W**heel) beschrieben durch

$$\{\ (\ F(s)\sin\tau\ ,\ s\ ,\ F(s)\cos\tau\ )^T\ :\ \tau\in[0,2\pi),\ \underline{s}\leq s\leq\overline{s}\ \}\ .$$

Sie ist durch die Zylinderkoordinaten s und τ parametrisiert, dabei ist die s-Achse parallel zur Radachse.



Rad und Schiene.

Auf dem Gleis wird ein kartesisches Koordinatensystem $\boldsymbol{R}$ (engl.: **R**ail) mit Koordinaten (u,v,w) so eingeführt, daß die u-Achse längs und die v-Achse quer zur Laufrichtung des Gleises liegt sowie die w-Achse in Richtung des Gleisfußes orientiert ist (vgl. Abbildung). Der Koordinatenursprung von $\boldsymbol{W}$ soll in der (v,w)–Ebene liegen und bezüglich $\boldsymbol{R}$ die Koordinaten $(0,\xi_v,\xi_w)^T$ haben. Mit den angedeuteten Winkeln φ, ϑ und ψ, die Drehungen

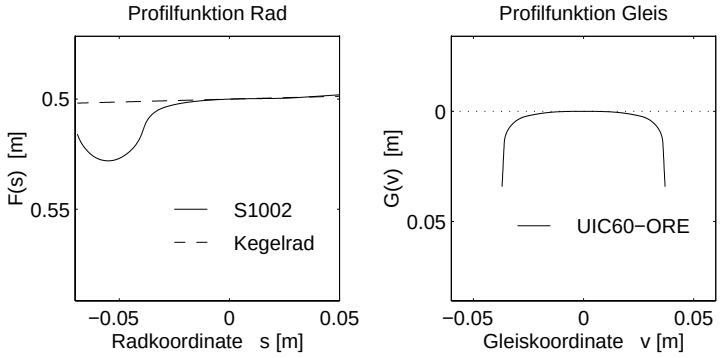


Abbildung 4.2: Profilfunktionen. Links: Rad mit nominellem Rollradius $r_0 = 0.5\,\mathrm{m}$, rechts: Gleis(kopf).

um die u-, v- bzw. w-Achse repräsentieren, ist die relative Lage des Rades zur Schiene vollständig durch $q_{\mathrm{rel}} = (\xi_v, \xi_w, \varphi, \vartheta, \psi)^T$ beschrieben. In (4.4) wird q_{rel} durch die MKS–Lagekoordinaten bestimmt: $q_{\mathrm{rel}} = q_{\mathrm{rel}}(q)$, wobei die konkrete Gestalt dieser Funktion wiederum vom Trassenverlauf, z. B. vom Anstellwinkel des Gleises und von Gleisstörungen, abhängt (vgl. [120] für typische Anwendungsbeispiele).

In praktischen Anwendungen ist es gleichermaßen wichtig, Systeme mit fahrwegunabhängigem Gleisprofil und Systeme, für die sich das Gleisprofil mit dem Fahrweg ändert, möglichst effizient simulieren zu können. Im ersten Fall ist G von x unabhängig (d. h. $G = G(v)$, z. B. bei Geradenfahrt und bei Fahrt im Gleisbogen), im zweiten Fall gilt dagegen $G = G(v;x)$ (z. B. bei Weichenfahrt). In praxi sind die Profilfunktionen durch Standards vorgegeben, Abb. 4.2 zeigt neben einem Kegelprofil die in den Testrechnungen verwendeten Profile S1002 (Rad) und UIC60–ORE (Gleis) (vgl. DIN 5573 [167]).

Ähnlich wie in Beispiel 24 wird jedem Punkt P_{W} der Radoberfläche längs der (nach unten (!) orientierten) w-Achse ein Punkt $P_{\mathrm{R}}(P_{\mathrm{W}})$ zugeordnet („**R**ail“, „**W**heel“). Wird der vorzeichenbehaftete Abstand $e_3^T \cdot (P_{\mathrm{R}}(P_{\mathrm{W}}) - P_{\mathrm{W}})$ in einem Punkt P_{W} der Radoberfläche minimal, so ist der Normalenvektor an die Radoberfläche in P_{W} parallel zum Normalenvektor an die Gleisoberfläche in $P_{\mathrm{R}}(P_{\mathrm{W}})$, denn andernfalls gäbe es in einer Umgebung von P_{W} einen Punkt mit kleinerem Abstand zur Gleisoberfläche. Weil $G(v;x)$ von u unabhängig ist und deshalb die Normalenvektoren an die Gleisoberfläche stets parallel zur (v,w)–Ebene sind, kann man die Suche nach Minimalstellen des Abstands also auf die Kurve $\mathfrak{C}$ derjenigen Punkte P_{W} der unteren Hälfte der Radoberfläche beschränken, in denen der Normalenvektor an die Radoberfläche parallel zur (v,w)–Ebene ist. Zur Veranschaulichung betrachte man den Spezialfall $\psi = 0$: hier ist $\mathfrak{C}$ die Menge derjenigen Punkte auf der Radoberfläche, die senkrecht unter der Radachse liegen. I. allg. ist s eine Parametrisierung von $\mathfrak{C}$, so daß die Abstandsfunktion Δ in Analogie zu (4.5) definiert

wird als

$$\Delta(s; q_{\rm rel}, x) := e_3^T \cdot (P_{\rm R}(P_{\rm W}(s)) - P_{\rm W}(s)) \tag{4.8}$$

(hier ist $P_{\rm W}(s) \in \mathfrak{C}$ und e_3 bezeichnet wie oben den dritten Einheitsvektor in $\boldsymbol{R}$). Zerlegt man Δ wie in (4.5) in $\Delta = -\xi_w - \zeta$, so ergibt sich die Kontaktbedingung

$$0 = \gamma(q) := -\xi_w - \max_{s\in[\underline{s},\overline{s}]} \{\, \zeta(s; \xi_v, \varphi, \psi, x) \,\} \tag{4.9}$$

mit der Funktion $\zeta(s; \xi_v, \varphi, \psi, x) := -\xi_w - \Delta(s; q_{\rm rel}, x)$, die weder von ξ_w noch von ξ_u und ϑ abhängt. Sind ξ_v, φ, ψ und x gegeben, so definiert (4.9) die Höhenauslenkung $-\xi_w$ des Rades.

In einem vollständigen Rad–Schiene–System mit M Rädern sind M Vektoren $q_{\rm rel}^{(i)}$ zu berechnen und man erhält als Zwangsbedingungen in (4.4)

$$0 = g(q) := \gamma(q) := (\, \gamma_1(q), \ldots, \gamma_M(q) \,)^T, \tag{4.10}$$

wobei $\gamma_i(q)$ die Kontaktbedingung (4.9) für das i-te Rad bezeichnet ($i = 1, \ldots, M$).

Bemerkung 37 Auf dem in Beispiel 24 angedeuteten Weg lassen sich für vielfältige Anwendungen Kontaktbedingungen im MKS zu Zwangsbedingungen umformen (vgl. z. B. [148], [5], [166]). Die bisher a priori getroffene Voraussetzung permanenten Kontakts läßt sich dabei leicht während der numerischen Integration überprüfen: In Beispiel 24 bleibt die einseitige Beschränkung $\min_{|\xi-q_1|<\varrho} \Delta(\xi; q) \geq 0$ aus (4.6) nur so lange aktiv, wie die durch (4.4) mit $g(q) := \min_{|\xi-q_1|<\varrho} \Delta(\xi; q)$ definierte Zwangskraft $-G^T(q)\lambda$ ins Innere des zulässigen Bereichs $\{\, q : \min_{|\xi-q_1|<\varrho} \Delta(\xi; q) \geq 0 \,\}$ gerichtet ist. Bezeichnet $\mathfrak{n}$ einen Normalenvektor an den Rand des zulässigen Bereichs in einem Punkt q mit $g(q) = 0$, so wirkt die Zwangskraft wegen $G(q) = \frac{\partial}{\partial q} g(q)$ stets längs des Vektors $\mathfrak{n}$. Ist die Kontaktbedingung also zunächst aktiv und wechselt λ später das Vorzeichen, so verläßt die Lösungstrajektorie den durch (4.7) definierten Rand des zulässigen Bereichs und die Scheibe verliert den Kontakt zum Untergrund.
Ausdrücklich sei darauf hingewiesen, daß derartige Systeme mit einer während der Integration veränderlichen Zahl der Freiheitsgrade *nicht* Gegenstand dieses Kapitels sind (vgl. hierzu z. B. [89], [103]). Bei der dynamischen Simulation von Schienenfahrzeugen überprüft man während der Integration unter Verwendung von Schaltfunktionen ([82, S. 196ff]) das Vorzeichen von λ. Ein Abheben des Rades signalisiert akute Entgleisungsgefahr, so daß die dynamische Simulation in der Regel abgebrochen wird.

In Abhängigkeit von der konkreten Gestalt von Δ lassen sich für Spezialfälle besonders effektive Lösungsverfahren angeben: Hat eine (hinreichend oft stetig differenzierbare) Abstandsfunktion Δ für fixiertes q nur einen lokalen Extremwert (bezüglich ξ) und ist dieses lokale Extremum gleichzeitig das globale Minimum von Δ, so reicht es aus, statt der unendlich vielen einseitigen Beschränkungen die Bedingung $\Delta \geq 0$ nur für einen einzigen Punkt, einen *Kontaktpunkt* zu fordern (vgl. z. B. [148]).

Beispiel 25 In Beispiel 24 sind die Kontaktpunkte durch ξ_0 mit

$$\Delta(\xi_0; q) = \min \{\, \Delta(\xi; q) \,:\, |\xi - q_1| < \varrho \,\} \tag{4.11}$$

charakterisiert und in Abb. 4.1 durch „+“ markiert.

Satz 19 *Gilt für das in Beispiel 24 beschriebene Modellproblem $a > \varrho$, so gibt es zu gegebenem $q \in \mathbb{R}^2$ ein eindeutig bestimmtes $\xi_0 \in \mathbb{R}$ mit (4.11), und die Kontaktbedingung (4.7) ist äquivalent zu*

$$\left.\begin{array}{rcl} 0 &=& \Delta_\xi(\xi; q) \\ 0 &=& \Delta(\xi; q) \end{array}\right\}. \tag{4.12}$$

Beweis In Beispiel 24 gilt

$$\lim_{\xi\to q_1-\varrho+0} \Delta_\xi(\xi; q) = -\infty, \quad \lim_{\xi\to q_1+\varrho-0} \Delta_\xi(\xi; q) = \infty$$

und

$$\Delta_{\xi\xi}(\xi; q) = \frac{\varrho^2}{\sqrt{\varrho^2 - (\xi - q_1)^2}^3} - \frac{a^2}{\sqrt{\xi^2 + a^2}^3} \geq \frac{1}{\varrho} - \frac{1}{a}, \quad (\, |\xi - q_1| < \varrho \,), \tag{4.13}$$

d. h., für $a > \varrho$ ist stets $\Delta_{\xi\xi} > 0$. Die Ableitung $\Delta_\xi(\xi; q)$ hat dann zu gegebenem q genau eine Nullstelle $\xi_0 \in (q_1 - \varrho, q_1 + \varrho)$ und für dieses ξ_0 nimmt $\Delta(\xi; q)$ sein globales Minimum an. Deshalb ist (4.11) für $a > \varrho$ äquivalent zu $\Delta_\xi(\xi_0; q) = 0$. Aus $\Delta(\xi_0; q) \leq \Delta(\xi; q)$, $(|\xi - q_1| < \varrho)$ folgt dann auch die Äquivalenz von (4.7) und (4.12). ∎

Beispiel 26 Wenn man für $a > \varrho$ in dem Modellproblem aus Beispiel 24 die Kontaktbedingung (4.7) durch (4.12) ersetzt, so müssen bei der Formulierung der Modellgleichungen für die Lagekoordinaten q und für $s := \xi$ nicht nur die Zwangsbedingung $0 = \tilde{g}(q, s) := \Delta(\xi; q)$, sondern auch die Gleichung $0 = h(q, s) := \Delta_\xi(\xi; q)$ berücksichtigt werden. Für diese Funktion h ist $h_s = \Delta_{\xi\xi}$ wegen (4.13) und $a > \varrho$ immer regulär.

Betrachtet man bei der Aufstellung der Bewegungsgleichungen allgemein Systeme, für die die n_λ Gleichungen $\tilde{g}_1(q, s) = 0, \ldots, \tilde{g}_{n_\lambda}(q, s) = 0$ und die n_s Gleichungen $h_1(q, s) = 0, \ldots, h_{n_s}(q, s) = 0$ erfüllt sein sollen, so ist mit den Vektoren $\tilde{q} := (q^T, s^T)^T$ und $\tilde{\lambda} = (\lambda^T, \nu^T)^T$ statt der Lagrange–Funktion (4.3) die Funktion

$$L(\tilde{q}, \dot{\tilde{q}}, \tilde{\lambda}) = T(q, \dot{q}) - U(q) - \sum_{k=1}^{n_\lambda} \lambda_k \tilde{g}_k(q, s) - \sum_{k=1}^{n_s} \nu_k h_k(q, s)$$

zu verwenden. Berücksichtigt man mögliche Reibungskräfte im MKS, so ergibt sich analog zu (4.4)

$$\begin{array}{rcl} \hat{M}(q) \begin{pmatrix} \ddot{q} \\ \ddot{s} \end{pmatrix} &=& \hat{f}(\tilde{q}, \dot{q}, \lambda, t) - \hat{G}^T(\tilde{q}, t) \begin{pmatrix} \lambda \\ \nu \end{pmatrix} \\ 0 &=& \hat{g}(\tilde{q}) = \begin{pmatrix} \tilde{g}(q, s) \\ h(q, s) \end{pmatrix} \end{array} \tag{4.14}$$

mit

$$\hat{M}(q) = \begin{pmatrix} M(q) & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{f}(\tilde{q}, \dot{q}, \lambda, t) = \begin{pmatrix} \tilde{f}(q, s, \dot{q}, \lambda, t) \\ 0 \end{pmatrix}, \quad \hat{G}(\tilde{q}, t) = \frac{\partial \hat{g}}{\partial \tilde{q}} = \begin{pmatrix} \tilde{g}_q & \tilde{g}_s \\ h_q & h_s \end{pmatrix}.$$

Ist h_s regulär, so können in (4.14) die Gleichungen

$$0 = -\tilde{g}_s^T(q, s)\lambda - h_s^T(q, s)\nu$$

nach ν aufgelöst werden: $\nu = -[(\tilde{g}_s h_s^{-1})^T](q, s)\lambda$. Setzt man diesen Ausdruck in (4.14) ein, so ergeben sich die schon in Abschnitt 3.1 betrachteten Modellgleichungen (3.14) für MKS mit Kontaktbedingungen.

Bemerkung 38 a) In Abschnitt 3.1 wurde gezeigt, daß man für (3.14) Äquivalente zur GGL- und zur stabilisierten Index-1-Formulierung definieren kann. Ist $\check{g}_s = \Delta_\xi = h$ wie in Beispiel 26, so vereinfacht sich sowohl das Aufstellen der Bewegungsgleichungen als auch die numerische Lösung, denn für die analytische Lösung gilt $\check{g}_s(q(t), s(t)) = 0$ und $\check{G}(q(t), s(t)) = \check{g}_q(q(t), s(t))$ wegen $h(q(t), s(t)) = 0$. Deshalb kann man auf die für die stabilisierte Index-1-Formulierung erforderlichen Modifikationen der Integrationssoftware (vgl. (3.18) und Abb. 3.2 auf S. 76) verzichten, solange man sich auf die GGL–Formulierung beschränkt ([23]).

b) Für einfache theoretische Untersuchungen wird oft das Profil der Räder eines Rad–Schiene–Systems linearisiert. Für den Kontakt zwischen einem solchen Kegelrad und einem Gleis mit konvexem Gleiskopf ist der Kontaktpunkt $P_{\mathrm{w}}^* = P_{\mathrm{w}}(s_*) \in \mathfrak{C}$ mit

$$\Delta(s_*; q_{\mathrm{rel}}, x) = \min \{ \Delta(s; q_{\mathrm{rel}}, x) : \underline{s} \le s \le \overline{s} \} \tag{4.15}$$

stets eindeutig bestimmt (vgl. (4.8) und (4.11)). Deshalb läßt sich das in Satz 19 am Modellproblem illustrierte Vorgehen auf den Kontakt eines kegelförmigen Rades mit der Schiene übertragen, und man erhält auch hier Modellgleichungen der Form (3.14). Dabei sind für ein System mit M Rädern die Kontaktpunktkoordinaten s_i, $(i = 1, \ldots, M)$ zu bestimmen, d. h., in (3.14) ist $s \in \mathbb{R}^{n_s}$ mit $n_s = M$. Für Rad–Schiene–Systeme wurde die direkte Anwendung eines Lösungsverfahrens für DA–Systeme auf (3.14) erstmals von Simeon, Führer und Rentrop untersucht ([148]), man vermeidet damit die im traditionellen Zugang erforderliche separate Auflösung der Gleichungen $h(q, s) = 0$ nach den Kontaktpunktkoordinaten s.

c) In [148] wird — im Unterschied zu der in Bemerkung 36 verwendeten Formulierung ([53], [120]) — die Suche nach dem Kontaktpunkt nicht auf $\mathfrak{C}$ eingeschränkt, so daß die Lage des Kontaktpunkts durch *zwei* Koordinaten beschrieben wird (s, τ). In praktischen Anwendungen ist die den Rad–Schiene–Systemen angepaßte Formulierung (4.9) vorzuziehen, denn sie führt auf DA–Systeme (3.14) kleinerer Dimension und ermöglicht außerdem die effiziente Bestimmung des Kontaktpunkts durch Suche nach dem globalen Maximum einer Funktion von *einer* reellen Variablen (z. B. bei der Berechnung konsistenter Anfangswerte).

Zum Test von neuen Modellen und Lösungsverfahren beschränkt man sich auf möglichst einfache Rad–Schiene–Systeme, die später bei der Simulation von vollständigen Schienenfahrzeugen als Substrukturen in einem modular aufgebauten Modell Verwendung finden. Das Spektrum der Möglichkeiten, wie sich ein Rad–Schiene–System dynamisch verhalten kann, ist schon für solche einfachen Modellprobleme weit gefächert und reicht von stabilen Trajektorien über Grenzzyklen bis hin zu chaotischen Bewegungen (vgl. z. B. [96]). Ein charakteristisches Testbeispiel, für das die Modellgleichungen noch sinnvoll von Hand aufgestellt werden können, ist die Bewegung eines starren Radsatzes längs eines geraden Gleises. Von Simeon, Führer und Rentrop wurde dieses Beispiel so aufbereitet, daß es auch für numerische Testrechnungen geeignet ist. In [148] werden hierzu die Modellgleichungen vollständig angegeben und Simulationsergebnisse für verschiedene Fahrmanöver gezeigt. Die in den Tests verwendeten FORTRAN–Quelltexte sind am CWI Amsterdam unter `http://www.cwi.nl/cwi/projects/IVPtestset.shtml` im Internet verfügbar.

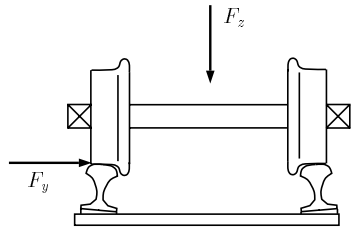


Abbildung 4.3: Substruktur „Starrer Radsatz“.

In [120] wurde dieses Beispiel aufgegriffen, wobei wir uns hier vor allem auf die Formulierung der Kontaktbedingungen konzentriert haben. Abb. 4.3 zeigt einen solchen starren Radsatz, wobei für die Räder das Verschleißprofil S1002 angedeutet ist. Der Radsatz wird als Starrkörper im MKS–Modell durch $n_q = 6$ Lagekoordinaten beschrieben. Hierzu wird in der Mitte des Gleisrosts in Analogie zu $\boldsymbol{R}$ ein kartesisches Koordinatensystem (x, y, z) eingeführt, so daß x die Position des Radsatzes längs des Fahrwegs, y die Seitenverschiebung und z die Höhenauslenkung des Radsatzes angibt. Zusammen mit den Drehwinkeln erhält man die Lagekoordinaten $q = (x, y, z, \alpha, \beta, \gamma)^T$, die in der nominellen Lage des Radsatzes die Zahlenwerte $y = 0$, $\alpha = 0$, $\gamma = 0$ haben ($z \approx -0.5\,\mathrm{m}$ für die in Abb. 4.2 gezeigten Radprofile). Betrachtet wird eine geführte Bewegung des Radsatzes mit konstanter Geschwindigkeit V_0, d. h. $x = V_0 \cdot t$. Die Abmessungen entsprechen denen eines in praxi verwendeten Radsatzes, für den Fahrweg werden zwei Gleise mit dem Profil UIC60–ORE im Winkel 1/40 angestellt (vgl. Abb. 4.3). Die vollständige Aufzählung der geometrischen und physikalischen Parameter und der aus ingenieurtechnischer Sicht interessanten Bezeichnungen enthält Tabelle 1 in [120]. Im Modell werden als zunächst frei zu wählende Parameter die Kräfte F_y und F_z eingeführt, die in y- bzw. in z-Richtung auf den Radsatz wirken (vgl. Abb. 4.3), hierbei entspricht F_z einer möglichen Achslast, F_y ist eine seitliche Führungskraft.

Beide Räder sollen permanent in Kontakt mit den Schienen sein, so daß q zwei Zwangsbedingungen erfüllen muß $(n_\lambda = 2)$. Im Starrkörperkontaktmodell mit den Modellgleichungen (3.14) treten zusätzlich $n_s = 2$ Kontaktpunktkoordinaten s auf (je 1 für das rechte und das linke Rad). Die für die Definition der Kontaktbedingungen (4.10) benötigte Transformation der MKS–Lagekoordinaten q in Relativkoordinaten $q_{\mathrm{rel}}^{(\mathrm{i})}(q)$ ergibt sich aus elementaren geometrischen Überlegungen ([148], [120]).

Die Berechnung der Reibungskräfte erfolgt nach der Kalkerschen Rollreibungstheorie mit dem Programm FASTSIM ([97]). Die Reibungskräfte F_R hängen nichtlinear von den Zwangskräften $F_N = -\check{G}^T(q, s)\lambda$, von Materialkonstanten, vom Schlupf zwischen Rad und Schiene und von anderen Einflußgrößen ab, dabei ist $\mu|F_N|$ eine obere Schranke für $|F_R|$. Der Reibungskoeffizient μ variiert in den verschiedenen Testrechnungen. Für den Stahl von Rad und Schiene werden die Materialkonstanten $\nu_{\mathrm{Stahl}} = 0.277$ (Querkontraktionszahl) und $E = 2.02275 \cdot 10^{11}\,\mathrm{Nm}^{-2}$ (Elastizitätsmodul) verwendet.

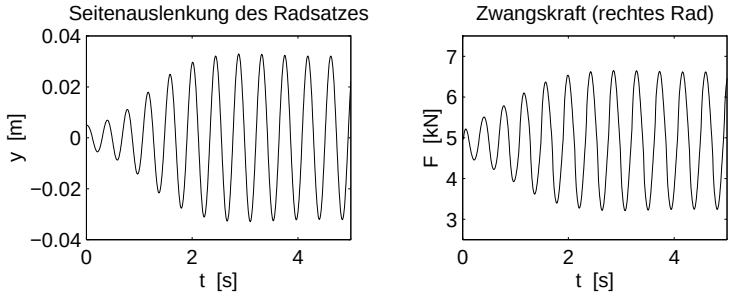


Abbildung 4.4: Bewegung eines starren Radsatzes mit Kegelrädern im Geradenlauf, $V_0 = 67.1\,\mathrm{ms}^{-1}$.

Beispiel 27 Ein seit langem bekanntes Phänomen der Laufdynamik von Schienenfahrzeugen ist die Instabilität der quasi-stationären Bewegung mit $y = 0$, wenn ein starrer Radsatz mit konstanter Geschwindigkeit $V_0 \geq V_0^*$ längs eines geraden Gleises geführt wird (V_0^* bezeichnet die sog. kritische Geschwindigkeit). Wird der Radsatz geringfügig seitlich ausgelenkt, so kehrt er nicht in die nominelle Lage ($y = 0$) zurück, für gewisse Geschwindigkeiten V_0 nähert sich die Bewegung sehr schnell einem Grenzzyklus an. Da sich diese rasch oszillierende Bewegung (engl.: „hunting motion“) ungünstig auf die Fahreigenschaften auswirkt, versucht man, durch konstruktive Veränderungen am Schienenfahrzeug ein möglichst großes V_0^* zu erreichen.
Als Beispiel zeigt Abb. 4.4 Simulationsergebnisse für einen starren Radsatz mit Kegelrädern ($F(s) = r_0 + c_{\mathrm{con}} \cdot s$ mit $r_0 = 0.5\,\mathrm{m}$ und $c_{\mathrm{con}} = -0.0262$) mit $F_y = F_z = 0\,\mathrm{N}$ und $\mu = 0.12$. Für $V_0 = 67.1\,\mathrm{ms}^{-1}$ ist die kritische Geschwindigkeit überschritten und die Schwingung in y klingt rasch auf. Zur Simulation wurde der BDF–Code ODASSL auf die nach (3.18) bestimmte Gear–Gupta–Leimkuhler–Formulierung der Bewegungsgleichungen (3.14) angewendet. Bei Toleranzen von 10^{-5} (für q und $v = \dot{q}$) bzw. 10^{-2} (für λ) betrug die Rechenzeit auf einer SUN Sparc5 Workstation 12.3 s.
Frequenz und Amplitude der quasi-periodischen Bewegung decken sich mit den Simulationsergebnissen von Simeon et al. (nach dem Modell aus [148]), sie sind in guter Übereinstimmung mit Daten, die auf einem Rollprüfstand experimentell bestimmt wurden.

4.2 Kontaktpunktsprünge und nicht differenzierbare Kontaktbedingungen

Kontaktbedingungen können auf Zwangsbedingungen führen, die nur stückweise differenzierbar sind. In diesem Abschnitt wird dieses Problem untersucht, und die Auswirkungen auf die dynamische Simulation von Rad–Schiene–Systemen werden verdeutlicht. Nur durch Einführung eines alternativen Kontaktmodells gelingt es, auch für Räder mit sog.

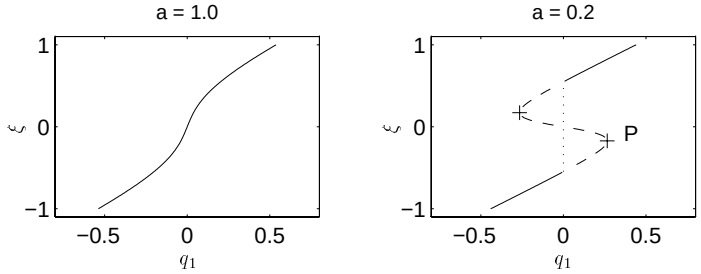


Abbildung 4.5: Lösungsmenge von $\Delta_\xi(\xi; q) = 0$, vgl. Beispiel 28 ($\varrho = 0.8$).

Verschleißprofil zufriedenstellende Simulationsergebnisse zu erhalten.

Der in Satz 19 skizzierte numerisch sehr günstige Zugang zur Formulierung von geometrischen Kontaktbedingungen setzt voraus, daß es stets einen eindeutig bestimmten Kontaktpunkt gibt, dessen Lage sich stetig mit q ändert. Schon für das einfache Modellproblem ist diese Voraussetzung verletzt:

Beispiel 28 Die Kontaktbedingung in Beispiel 24 ergibt sich aus dem globalen Minimum der Abstandsfunktion Δ. Abb. 4.5 zeigt für zwei verschiedene Werte des Parameters a die lokalen Extremstellen von $\Delta(\xi; q)$ in Abhängigkeit von q_1 (man beachte, daß Δ_ξ von q_2 unabhängig ist). Dabei entsprechen die globalen Minima von Δ den durchgezogenen Linien und die anderen lokalen Extrema von Δ der gestrichelten Linie. Insbesondere gibt die durchgezogene Linie an, wie sich in Abhängigkeit von q_1 die Lage des Kontaktpunkts ändert, d. h. sie zeigt das in (4.11) definierte $\xi_0(q_1)$.
Aus geometrischen Gründen gibt es für $q_1 = 0$ im Fall $a < \varrho$ zwei Kontaktpunkte, die symmetrisch bezüglich der η-Achse liegen (Abb. 4.1 rechts). Im rechten Diagramm von Abb. 4.5 entspricht das dem Kontaktpunkt*sprung* von $-\xi_*$ zu ξ_* im Punkt $q_1 = 0$ (aus $q_1 = 0$ und $\Delta_\xi(\xi_*; q) = \Delta_\xi(-\xi_*; q) = 0$ folgt $\xi_*^2 = (\varrho^2 - a^2)/2$).

Satz 20 *Gilt für das in Beispiel 24 beschriebene Modellproblem $a < \varrho$, so ergibt sich aus der Kontaktbedingung (4.7) eine Funktion $q_2 = q_2(q_1)$, die in $q_1 = 0$ nicht stetig differenzierbar ist. Die durch die Kontaktbedingung (4.7) definierte Mannigfaltigkeit $\{ q : \gamma(q) = 0 \}$ ist in Punkten $q \in \mathbb{R}^2$ mit $q_1 = 0$ nicht differenzierbar.*

Beweis Für $q_1 \neq 0$ definiert (4.11) eine Funktion $\xi_0 : (-\infty, 0) \cup (0, +\infty) \to \mathbb{R}$, für die insbesondere $\Delta_\xi(\xi_0(q_1); q) \equiv 0$ gilt. Deshalb läßt sich für $q_1 \neq 0$ die Kontaktbedingung (4.7) schreiben als

$$0 = \gamma(q) = q_2 - \zeta(\xi_0(q_1); q_1) = \Delta(\xi_0(q_1); q) \; .$$

Hieraus folgt für $q_1 \neq 0$ durch implizite Differentiation

$$\frac{dq_2}{dq_1}(q_1) = -\frac{\partial \Delta}{\partial q_1}(\xi_0(q_1); q) - \frac{\partial \Delta}{\partial \xi}(\xi_0(q_1); q) \cdot \frac{d\xi_0}{dq_1}(q_1) = -\frac{\partial \Delta}{\partial q_1}(\xi_0(q_1); q) \; , \tag{4.16}$$

denn $\Delta_\xi(\xi_0(q_1); q) = 0$.

Unter der Voraussetzung $a < \varrho$ gilt $\lim_{q_1 \to 0-} \xi_0(q_1) = -\xi_*$ und $\lim_{q_1 \to 0+} \xi_0(q_1) = \xi_*$. Setzt man diese Grenzwerte in (4.16) ein, so folgt durch elementare Umformungen

$$\lim_{q_1 \to 0-} \frac{dq_2}{dq_1}(q_1) = -\sqrt{\frac{\varrho^2 - a^2}{\varrho^2 + a^2}} \quad , \qquad \lim_{q_1 \to 0+} \frac{dq_2}{dq_1}(q_1) = \sqrt{\frac{\varrho^2 - a^2}{\varrho^2 + a^2}} \tag{4.17}$$

und damit auch die Behauptung des Satzes. ■

Kontaktbedingungen führen nicht nur in diesem einfachen Beispiel sondern auch in MKS–Modellen aus angewandten Problemen zu singulären Mannigfaltigkeiten. Einige Auswirkungen auf die dynamische Simulation von Rad–Schiene–Systemen werden in [65] diskutiert. Die Singularität hat weitreichende Konsequenzen: Sind in (4.4) die Zwangsbedingungen autonom ($g = g(q)$), so liegt der Geschwindigkeitsvektor $\dot{q}(t)$ stets im Tangentialraum an die Mannigfaltigkeit $\{\, \eta \,:\, g(\eta) = 0 \,\}$ im Punkt $q(t)$, denn $0 = \frac{d}{dt} g(q(t)) = G(q)\dot{q}(t)$. Ist g wie in Satz 20 nur stückweise differenzierbar, so führt dies in der Regel zu Unstetigkeiten in $\dot{q}(t)$ (und auch in $\lambda(t)$). Erreicht in Beispiel 28 mit $\varrho > a$ eine Trajektorie den Punkt $q_1 = 0$ mit einer Geschwindigkeit $v_1 = \dot{q}_1 \neq 0$, so kann sie wegen $v_2 = \dot{q}_2 = \frac{dq_2}{dq_1}(q_1) \cdot \dot{q}_1$ nicht stetig über $q_1 = 0$ hinaus fortgesetzt werden (vgl. (4.17)).

Bemerkung 39 a) Für stetige Trajektorien mit $\dot{q}_1 \neq 0$ ist also der Punkt $q_1 = 0$ ein *Impasse point* im ursprünglichen Wortsinn (dtsch.: „unpassierbarer Punkt"). Impasse points treten in verschiedenen praktischen Anwendungen auf und werden seit Ende der 80er Jahre intensiv untersucht. Im engeren Sinn konzentrieren sich diese Arbeiten in Anlehnung an die Theorie parameterabhängiger nichtlinearer Gleichungen auf DA–Systeme, für die gewisse nichtlineare Gleichungen $h(y, z) = 0$ mit Ausnahme einzelner Punkte (der „Impasse points") stets (lokal) eindeutig nach z auflösbar sind. Impasse points werden hierbei durch einen lokalen Rangabfall der Jacobimatrix h_z charakterisiert (vgl. z. B. die Arbeiten von Rabier und Rheinboldt [136], [137]). Dagegen entstehen die Singularitäten in den Modellgleichungen für MKS mit Kontaktbedingungen durch fehlende Differenzierbarkeit der Zwangsbedingungen.

b) Ebenso wie die von Rabier und Rheinboldt untersuchten Impasse points widerspiegeln auch die durch Kontaktbedingungen hervorgerufenen Singularitäten grundlegende Eigenschaften des mathematisch-physikalischen Modells, sie treten deshalb unabhängig von der konkreten Form der Modellgleichungen (Deskriptorform, Zustandsform) auf. Die für analytische Untersuchungen übliche Transformation der MKS–Modellgleichungen (4.4) auf Zustandsform setzt zweimalige stetige Differenzierbarkeit der Zwangsbedingungen voraus und ist für MKS mit Kontaktbedingungen nur außerhalb der Singularitäten von $\{\, \eta \,:\, g(\eta) = 0 \,\}$ möglich.

c) Ausdrücklich sei darauf verwiesen, daß für das Modellproblem auch im Fall $\varrho > a$ die Gleichung $\Delta_\xi(\xi; q) = 0$ in einer Umgebung der Lösung stets lokal eindeutig nach ξ auflösbar ist. Wählt man den Anfangswert für ξ auf dem passenden Lösungsast von $\Delta_\xi(\xi; q) = 0$ (d. h. $\xi(0) := \xi_0(q_1(0))$), so kann man wie in Satz 19 die Kontaktbedingung (4.7) durch (4.12) ersetzen, solange der kritische Punkt $q_1 = 0$ nicht erreicht wird. Eine Trajektorie, für die (4.7) und damit auch die unendlich vielen einseitigen Beschränkungen $\Delta(\xi; q) \geq 0$, $(|\xi - q_1| < \varrho)$ erfüllt sind, gelangt jedoch nie zu dem in Abb. 4.5 mit „P" markierten Umkehrpunkt der Kurve $\{\, (q_1, \xi) \,:\, \Delta_\xi(\xi; q) = 0 \,\}$. Dieser Umkehrpunkt ist ein klassischer Impasse point der Gleichungen (3.14) mit $s = \xi$, $h(q, s) = \Delta_\xi(\xi; q)$ und $\check{g}(q, s) = \Delta(\xi; q)$, denn in „P" gilt $h_s = \Delta_{\xi\xi} = 0$. Man beachte jedoch, daß (4.7) und (4.12) für das Modellproblem nur dann äquivalent sind, wenn ξ und q_1 gleiches Vorzeichen haben.

Die bisher im Detail am Modellproblem für $\varrho > a$ diskutierten Singularitäten sind auch charakteristisch für den Kontakt zwischen einem Rad mit *Verschleißprofil* und der Schiene. Verschleißprofile wie z. B. das Profil S1002 aus Abb. 4.2 werden von vielen europäischen Bahngesellschaften eingesetzt, sie wurden auf der Grundlage von Meßdaten eines nach langer Laufleistung verschlissenen Rades definiert. Deshalb ist die Profilfunktion des Rades über einen größeren Teil der *Lauffläche* des Rades nahezu parallel zur Profilfunktion des Gleises (vgl. Abb. 4.2 auf Seite 135, dort gehört der Bereich mit $s > -35\,\text{mm}$ zur Lauffläche und der Bereich mit $s < -35\,\text{mm}$ zur *Flanke* des Rades). Im Unterschied zu kegelförmigen Rädern sind Räder mit Verschleißprofil nicht konvex.

Als Gegenstück zu Abb. 4.5 zeigt Abb. 4.6 auf S. 144 wiederum den Kontaktpunkt („—"), die lokalen Extremstellen von Δ („- - -") und die Kontaktpunktsprünge („$\cdots$") in Abhängigkeit von den Lagekoordinaten, d. h. hier: in Abhängigkeit von q_{rel}. Dabei wurde von der nominellen Lage eines starren Radsatzes auf einem geraden Gleis mit Anstellwinkel 1/40 ausgegangen und anschließend ξ_v variiert (für die nominelle Lage ist $\xi_v \approx 9.50\,\text{mm}$, $\varphi = 0.025$ und $\psi = 0$). Das linke Diagramm entspricht dem einfachen Fall (Kegelprofil), das rechte dem komplizierten (Verschleißprofil). In dem Starrkörperkontaktmodell nach Bemerkung 36 treten Kontaktpunktsprünge nicht nur zwischen der Lauffläche und der Flanke des Rades ($s_* = -33\,\text{mm} \;\to\; s_* = -38\,\text{mm}$), sondern vor allem auch innerhalb der Lauffläche auf.

Der qualitative Unterschied zwischen linkem und rechtem Diagramm von Abb. 4.6 erscheint zunächst überraschend, da Kegel- und Verschleißprofil auf der Lauffläche nur wenig differieren (vgl. Abb. 4.2). Abb. 4.7 illustriert jedoch, wie weitreichend im Starrkörperkontaktmodell die Auswirkungen von sehr kleinen Profilunterschieden (im 1/10 mm–Bereich) sein können. Für beide Diagramme in Abb. 4.7 wurde ξ_w aus der Kontaktbedingung (4.9) bestimmt, die Winkel φ und ψ entsprechen der Konfiguration in Abb. 4.6. Wird das Rad von $\xi_v = 9.10\,\text{mm}$ (links) zu $\xi_v = 9.64\,\text{mm}$ (rechts) verschoben, so ändert sich die Abstandsfunktion für das Kegelprofil („-·-") nur geringfügig. Dagegen bildet sich für das Verschleißprofil („—") ein zweites lokales Minimum von Δ aus, schließlich springt (für $\xi_v \approx 9.64\,\text{mm}$) der Kontaktpunkt von $s_* \approx 2.4\,\text{mm}$ nach $s_* \approx -8.0\,\text{mm}$.

Kontaktpunktsprünge und die damit verbundenen Singularitäten der durch die Starrkörperkontaktbedingungen definierten Mannigfaltigkeit treten für Räder mit Verschleißprofil schon in unmittelbarer Nähe der nominellen Lage des Rades auf und werden deshalb zum entscheidenden Problem bei der dynamischen Simulation ([16]). Physikalisch entspricht der unstetigen Zustandsänderung beim Kontaktpunktsprung ein Stoß im MKS; wegen der zwischen Rad und Schiene wirkenden Reibungskräfte handelt es sich um einen Reibstoß. Während des Stoßes bleiben die Lagekoordinaten q unverändert, der Geschwindigkeitsvektor $\dot{q}$ und die Zwangskräfte $-G^T(q, t)\lambda$ ändern sich unstetig. Die Zustandsänderung des MKS wird nicht mehr allein durch die Bewegungsgleichungen (4.4) beschrieben, zusätzlich ist ein physikalisches Modell für die Zustandsänderung während des Reibstoßes

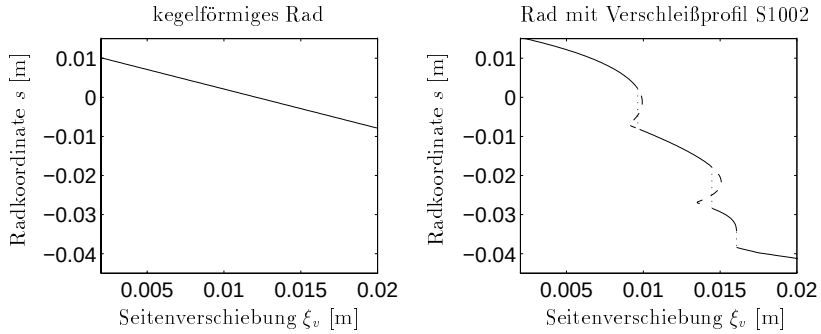


Abbildung 4.6: Kontaktpunkt und lokale Extremstellen der Abstandsfunktion Δ für Gleisprofil UIC60–ORE und kegelförmiges Rad (links) bzw. Radprofil S1002 (rechts). $\varphi = 0.025$, $\psi = 0$.

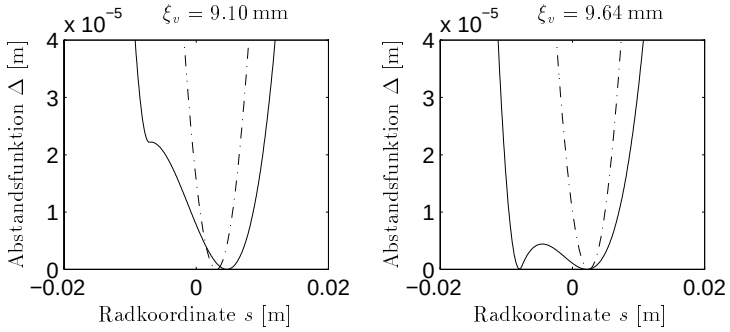


Abbildung 4.7: Abstandsfunktion $\Delta(s; q_{\mathrm{rel}})$ für Gleisprofil UIC60–ORE und kegelförmiges Rad („-·-“) bzw. Radprofil S1002 („—“). $\varphi = 0.025$, $\psi = 0$.

erforderlich.

In [120] wird das dynamische Verhalten von Rad–Schiene–Systemen, deren Räder Verschleißprofil haben, auf der Grundlage des Starrkörperkontaktmodells simuliert. Wegen der Kontaktpunktsprünge ist die Lösung der Bewegungsgleichungen mit den unendlich vielen einseitigen Beschränkungen

$$\Delta(s_i; q_{\mathrm{rel}}^{(i)}, x) \geq 0\ , \quad (\underline{s} \leq s_i \leq \overline{s},\ i = 1, \ldots, M)$$

aus Bemerkung 36 zu diskreten Zeitpunkten $t^{(j)}$ unstetig. Zwischen den Unstetigkeitsstellen, d. h. in den Teilintervallen $(t^{(j)}, t^{(j+1)})$, ändert sich dagegen die Lage der Kontaktpunkte stetig, so daß die in Abschnitt 4.1 für Rad–Schiene–Systeme mit Kegelrädern besprochenen Lösungsverfahren verwendet werden können.

Beginnend mit konsistenten Anfangswerten wendet man hierzu ein geeignetes Integrationsverfahren auf die Modellgleichungen (3.14) an (z. B. ODASSL für die Gear–Gupta–Leimkuhler–Formulierung). Nach jedem Integrationsschritt $t_n \to t_{n+1}$ wird überprüft, ob die vom Integrator stetig verfolgte Nullstelle s_i von $h_i(q, s) := \Delta_s(s_i; q_{\mathrm{rel}}^{(i)}, x)$ noch mit dem Kontaktpunkt zusammenfällt ($i = 1, \ldots, M$). Dazu ist für jedes Rad der Kontaktpunkt zu bestimmen, d. h., auf der Radoberfläche ist längs der Kurve $\mathfrak{C}$ derjenige Punkt zu suchen, für den die Abstandsfunktion minimal wird. In Abb. 4.6 rechts liegt der Kontaktpunkt stets auf einer der durchgezogenen Linien, dagegen folgt der Integrator der Kurve $\{\, s_i : h_i(q, s) = 0 \,\}$, die in der Abbildung durch die Vereinigung der durchgezogenen und der gestrichelten Linien dargestellt wird. Ist der Kontaktpunkt für jedes der Räder im Rahmen der vorgegebenen Genauigkeit mit dem vom Integrator berechneten Wert s_i identisch, so kann die Integration fortgesetzt werden.

Unterscheiden sich dagegen für eines der Räder die vom Integrator berechneten Koordinaten s von den Kontaktpunktkoordinaten, so hat während des Zeitschritts $t_n \to t_{n+1}$ ein Kontaktpunktsprung stattgefunden. Der genaue Zeitpunkt $t^{(j)} \in [t_n, t_{n+1}]$ wird unter Verwendung von Schaltfunktionen bestimmt, anschließend wird die Integration der Modellgleichungen in $t^{(j)}$ unterbrochen ([82, S. 196ff]). Das Reibstoßmodell gibt an, wie sich $\dot{q}$ und λ im Punkt $t^{(j)}$ ändern, und definiert damit die Anfangswerte $\dot{q}(t^{(j)})$ und $\lambda(t^{(j)})$ für die Integration der Modellgleichungen auf dem Zeitintervall $(t^{(j)}, t^{(j+1)})$. Mit diesen Anfangswerten (und mit den neuen Kontaktpunktkoordinaten s) kann nun die Integration fortgesetzt werden.

Beispiel 29 Ähnlich wie in Beispiel 27 zeigt Abb. 4.8 Simulationsergebnisse, die für die geführte Bewegung eines starren Radsatzes entlang eines geraden Gleises berechnet wurden ($V_0 = 30\,\mathrm{ms}^{-1}$, $F_y = 0\,\mathrm{N}$, $F_z = 10^5\,\mathrm{N}$, $\mu = 0.2$). Im Unterschied zu Beispiel 27 haben die Räder hier jedoch ein Verschleißprofil. Wegen des in Flankennähe rasch wachsenden Raddurchmessers ist die Amplitude der quasi-periodischen Bewegung sehr viel kleiner als bei Kegelrädern. In diesem Beispiel kommt für keines der beiden Räder die Flanke in Kontakt mit dem Gleis. Es finden jedoch zahlreiche Kontaktpunktsprünge innerhalb der Laufflächen der Räder statt, so daß sich die Zwangskräfte unstetig ändern.

Als Stoßmodell wird stark vereinfachend ein aus der Literatur bekanntes Modell für voll plastische, reibungsfreie Stöße verwendet ([89]): Der Geschwindigkeitsvektor $\dot{q}$ ändert

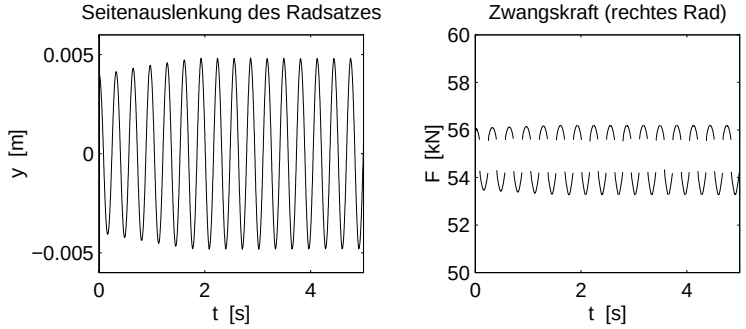


Abbildung 4.8: Bewegung eines starren Radsatzes mit Radprofil S1002 im Geradenlauf, $V_0 = 30.0\,\mathrm{ms}^{-1}$.

sich von $\dot{q}^-$ zu $\dot{q}^+$ mit

$$M(q)(\dot{q}^+ - \dot{q}^-) \in \text{range}\, \tilde{G}^T(q, s^+) \quad , \quad \tilde{G}(q, s^+)\dot{q}^+ = 0 \, , \tag{4.18}$$

wobei s^+ die Kontaktpunktkoordinaten nach dem Kontaktpunktsprung bezeichnet (vgl. die Definition von $\tilde{G}$ in (3.14)). Dadurch sind wegen $\frac{d^2}{dt^2}\tilde{g}(q,s) = 0$ auch die Zwangskräfte nach dem Stoß eindeutig bestimmt (vgl. (3.3)).

Diese Implementierung des Starrkörperkontaktmodells arbeitet zuverlässig und vergleichsweise effektiv. Gegenüber der Simulation von Rad–Schiene–Systemen mit Kegelrädern entsteht im wesentlichen der folgende zusätzliche Aufwand:

a) Auswertung der Schaltfunktion nach jedem Integrationsschritt (hierzu pro Rad die Bestimmung des globalen Minimums für eine von einer skalaren Größe abhängende skalare Funktion),

b) genaue Lokalisierung der Unstetigkeitsstelle $t^{(j)} \in [t_n, t_{n+1}]$ (hierauf kann wegen der meist kleinen Integrationsschrittweite und des ohnehin großen Modellfehlers verzichtet werden [120]),

c) Berechnung von s^+, Berechnung der unstetigen Zustandsänderung nach (4.18) und konsistente Initialisierung der Zwangskräfte nach dem Kontaktpunktsprung und

d) Neustart des Integrationsverfahrens.

Für die im Beispiel 29 gezeigte Simulation betrug die Rechenzeit auf einer SUN Sparc5 Workstation 202.1 s.

Obwohl der angegebene Algorithmus das Problem der numerischen Integration von Modellgleichungen mit Kontaktpunktsprüngen aus *numerischer* Sicht zufriedenstellend löst, ist er dennoch für praktische Rechnungen nur sehr eingeschränkt brauchbar. Ein ungelöstes Problem ist, welches Stoßmodell den Reibstoß zwischen Rad und Schiene ausreichend gut beschreibt und trotzdem effektiv numerisch umgesetzt werden kann. Außerdem führt in größeren Rad–Schiene–Systemen die wachsende Zahl von Rädern und damit die

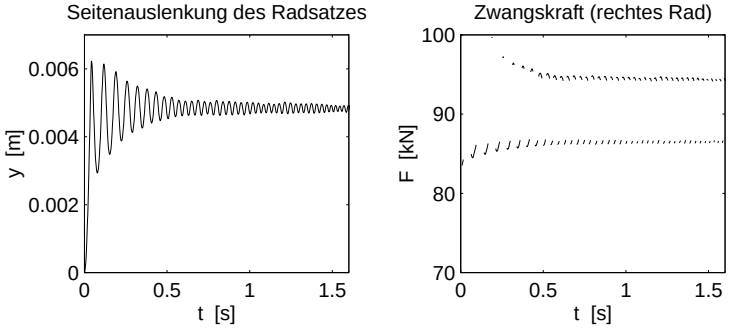


Abbildung 4.9: Starrer Radsatz im Geradenlauf. Anwendung des Starrkörperkontaktmodells auf ein Benchmark–Problem von Pascal ([126]).

wachsende Zahl von Kontaktbedingungen zu sehr viel häufigeren Unterbrechungen der Zeitintegration, so daß der nach jeder Unstetigkeitsstelle erforderliche große Aufwand zur Fortsetzung der Zeitintegration insgesamt stark anwächst.

Der entscheidende Mangel ist jedoch, daß der Modellfehler des Starrkörperkontaktmodells die Simulationsergebnisse stark verfälschen kann:

Beispiel 30 Der in den Beispielen 27 und 29 simulierte Grenzzyklus ist sowohl in der Realität als auch in der Simulation außerordentlich stabil und stellt sich selbst bei größeren Modellfehlern (z. B. im Reibstoßmodell) immer wieder ein. Von Pascal ([126]) wurde ein Benchmark–Problem formuliert, das die Auswirkungen verschiedener Modellierungen des geometrischen Kontakts einer Schiene mit einem Rad mit Verschleißprofil sehr viel deutlicher erkennen läßt. Mit den Bezeichnungen von Abb. 4.3 ist $V_0 = 30.0\,\mathrm{ms}^{-1}$, $F_y = 2 \cdot 10^4\,\mathrm{N}$, $F_z = 154715\,\mathrm{N}$, $\mu = 0.01$. Als Gleisprofil wird UIC60–ORE verwendet, die Räder haben das Verschleißprofil S1002, jedoch mit einem um 5 cm geringeren nominellen Rollradius als in Abb. 4.2 ($r_0 = 0.45\,\mathrm{m}$). Wegen der seitlichen Führungskraft F_y wird der Radsatz aus seiner anfänglichen nominellen Lage ausgelenkt und die Lösung erreicht eine quasi-stationäre Lage mit $\lim_{t\to\infty} y(t) \approx 5.0\,\mathrm{mm}$, d. h., die Radsatz–Lagekoordinaten y, z, α und γ nähern sich konstanten Werten an, während der Drehwinkel β und $x = V_0 \cdot t$ weiter wachsen. Für die quasi-stationäre Lage wird auf dem rechten Rad eine Konfiguration $q_{\mathrm{rel}}^{(i)}(q)$ erreicht, die in der Nähe einer Sprungstelle des Kontaktpunkts im Starrkörperkontaktmodell liegt.
Wendet man den oben beschriebenen Algorithmus zur dynamischen Simulation auf dieses Benchmark–Problem an, so wird die quasi-stationäre Lage nicht erreicht (vgl. Abb. 4.9). Wegen zahlreicher Kontaktpunktsprünge auf dem rechten Rad und wegen der damit verbundenen Unstetigkeiten in $\dot{q}$ und in den Zwangskräften oszilliert die Lösung statt dessen um die quasi-stationäre Lage.

4.3 Ein quasi-elastisches Modell für den Rad-Schiene-Kontakt

Prinzipiell wäre es denkbar — zum Beispiel bei plastischen Verformungen des Rades oder der Schiene während des Reibstoßes — daß die Simulationsergebnisse aus Beispiel 30 die Grenzen der Anwendbarkeit des MKS–Modells zeigen. Die wirkliche Ursache des inakzeptabel großen Fehlers der Simulationsergebnisse ist jedoch der Modellfehler des Starrkörperkontaktmodells. In diesem Abschnitt entwickeln wir als Alternative ein sog. quasi-elastisches Modell für den Rad–Schiene–Kontakt, das einerseits die elastische Deformation von Rad und Schiene qualitativ berücksichtigt und andererseits die Nachteile eines rein elastischen Kontaktmodells vermeidet.

Abb. 4.7 auf S. 144 zeigt, daß für Verschleißprofile die Aufgabe, zu gegebenem Profil und zu gegebener Lage q_{rel} den Kontaktpunkt zu bestimmen, außerordentlich schlecht konditioniert ist. Minimale Lageänderungen oder kleinste Profiländerungen (z. B. durch Verschleiß) können Kontaktpunktsprünge zur Folge haben. D. h., die zur Definition der Kontaktbedingung eingeführte Hilfsgröße „Kontaktpunkt“ ist im Fall von Verschleißprofilen ungeeignet. In der ingenieurwissenschaftlichen Literatur werden gelegentlich Ansätze untersucht, die bei der Formulierung der Kontaktbedingungen statt des Kontaktpunkts (also desjenigen Punkts, in dem Δ das *globale* Minimum annimmt) die Menge aller *lokalen* Minimalstellen von Δ zu verwenden („Mehrpunktkontakt“). Aber auch diese Menge ist ungeeignet, denn die Anzahl der Minimalstellen kann sich bei Lage- und / oder bei Profiländerungen sprunghaft verändern (vgl. [11] und Abb. 4.7).

Bemerkung 40 Ausgangspunkt des Starrkörperkontaktmodells sind zwei undeformierte Festkörper, die sich in (mindestens) einem Kontaktpunkt P_{W}^* berühren. Abb. 4.10 zeigt diese Konfiguration für ein Rad und eine Schiene, deren undeformierte Oberflächen schematisch durch die gestrichelten Linien angedeutet werden.
Sehr viel genauer als durch das Starrkörperkontaktmodell wird der Rad–Schiene–Kontakt durch ein elastisches Kontaktmodell beschrieben (vgl. z. B. [95] für eine allgemeine Einführung und [98] für eine spezifische Darstellung des Rad–Schiene–Kontakts). Ist $\mathcal{E}$ die Tangentialebene, die im Kontaktpunkt P_{W}^* an die Radoberfläche gelegt wird, so betrachtet man eine in Normalenrichtung zu $\mathcal{E}$ wirkende Kraft F_N. Zur analytischen Beschreibung des elastischen Kontakts definiert man ein kartesisches Koordinatensystem (ξ_x, ξ_y, ξ_z) mit dem Kontaktpunkt als Koordinatenursprung so, daß $\mathcal{E}$ mit der (ξ_x, ξ_y)–Ebene zusammenfällt und F_N längs der ξ_z-Achse wirkt (Abb. 4.12). Durch die elastische Deformation von Rad und Schiene bildet sich eine Kontaktfläche Ω aus, deren Lage und Größe zunächst unbekannt sind. Näherungsweise nimmt man an, daß Ω in $\mathcal{E}$ liegt. Die Andruckkraft F_N verteilt sich auf Ω, dabei ist $|F_N| = \iint_\Omega p(\xi'_x, \xi'_y)\, d\Omega$ mit der Normaldruckverteilung $p(\xi'_x, \xi'_y)$, für die gilt $p(\xi'_x, \xi'_y) \geq 0$ und

$$p(\xi'_x, \xi'_y) = 0\,, \quad \text{falls} \quad (\xi'_x, \xi'_y) \notin \Omega\,.$$

Nun betrachtet man die Wirkung dieser Normaldruckverteilung getrennt für die Festkörper Rad und Schiene. Die lineare Elastizitätstheorie beschreibt die durch p in diesen Körpern hervorgerufenen Deformationen, wobei es sinnvoll ist, Rad und Schiene jeweils

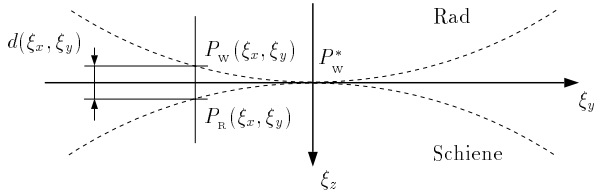


Abbildung 4.10: Starrkörperkontakt zweier geometrischer Körper.

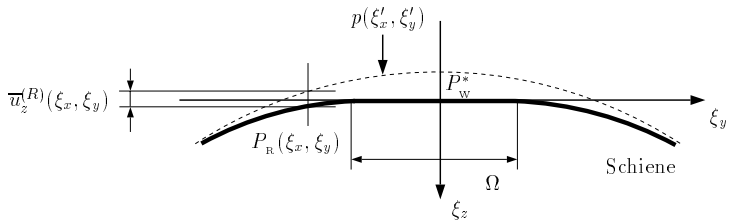


Abbildung 4.11: Elastische Deformation der Schiene (undeformierter „--“ und deformierter „—“ Zustand der Oberfläche).

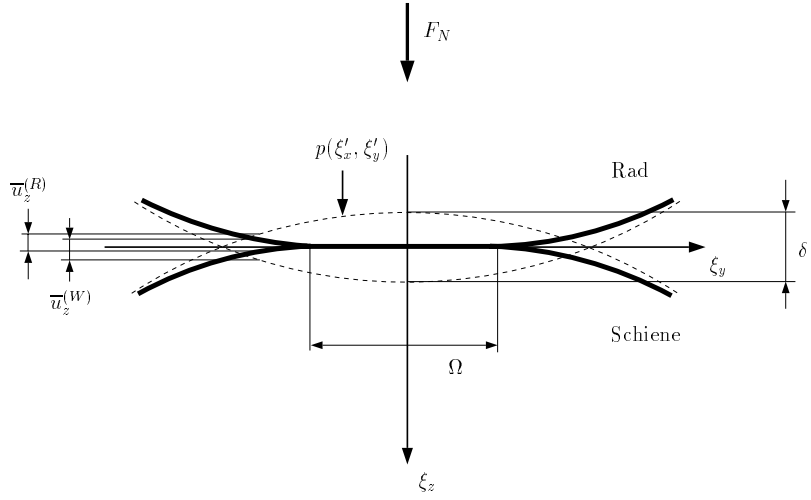


Abbildung 4.12: Elastische Deformation von Rad und Schiene (undeformierter „--“ und deformierter „—“ Zustand), nach [95, Abb. 4.2].

durch elastische Halbräume zu approximieren. Wenn man dabei den Einfluß der in Tangentialrichtung wirkenden Kräfte (z. B. Reibungskräfte) vernachlässigt, so ergeben sich an den Oberflächen von Rad und Schiene die Verschiebungen

$$\overline{u}_z^{(W)}(\xi_x,\xi_y) = \frac{1-\nu_{\mathrm{W}}^2}{\pi E^{(W)}} \iint_\Omega \frac{p(\xi'_x,\xi'_y)}{\sqrt{(\xi_x-\xi'_x)^2+(\xi_y-\xi'_y)^2}}\, d\Omega$$

$$\overline{u}_z^{(R)}(\xi_x,\xi_y) = \frac{1-\nu_{\mathrm{R}}^2}{\pi E^{(R)}} \iint_\Omega \frac{p(\xi'_x,\xi'_y)}{\sqrt{(\xi_x-\xi'_x)^2+(\xi_y-\xi'_y)^2}}\, d\Omega$$

in ξ_z-Richtung. Hierbei sind die Querkontraktionszahlen ν_{W}, ν_{R} und die Elastizitätsmoduln $E^{(W)}$, $E^{(R)}$ Materialkonstanten (für „**W**heel“ bzw. „**R**ail“, vgl. S. 139). Abb. 4.11 veranschaulicht die Deformation der Schiene. In Abb. 4.12 werden (wieder nur schematisch) neben den undeformierten Oberflächen („- - -“) auch die deformierten Oberflächen („———“) der beiden Körper sowie die Kontaktfläche Ω dargestellt. In Normalenrichtung zu $\mathcal{E}$ ergibt sich als Folge der elastischen Deformation eine veränderte relative Lage des Rades zur Schiene. Die undeformierten Oberflächen von Rad und Schiene würden sich in dieser Lage gegenseitig durchdringen (vgl. Abbildung).

Zu gegebenen Koordinaten ξ_x und ξ_y bezeichne $P_{\mathrm{W}}(\xi_x,\xi_y) = (\xi_x,\xi_y,\xi_z^{(W)})^T$ einen Punkt auf der Radoberfläche und $P_{\mathrm{R}}(\xi_x,\xi_y) = (\xi_x,\xi_y,\xi_z^{(R)})^T$ den zugehörigen Punkt auf der Gleisoberfläche. Berühren sich die *undeformierten* Körper Rad und Schiene im Kontaktpunkt $P_{\mathrm{W}}^* = P_{\mathrm{W}}(0,0) = P_{\mathrm{R}}(0,0)$, so haben $P_{\mathrm{W}}(\xi_x,\xi_y)$ und $P_{\mathrm{R}}(\xi_x,\xi_y)$ einen nichtnegativen Abstand $d(\xi_x,\xi_y)$, wobei $d(\xi_x,\xi_y)$ allein durch die Geometrie der Oberflächen von Rad und Schiene bestimmt ist (vgl. Abb. 4.10). Die der Normaldruckverteilung p entsprechenden Verschiebungen $\overline{u}_z^{(W)}$ und $\overline{u}_z^{(R)}$ bewirken in Abb. 4.12 eine Verschiebung von P_{W} nach „oben“ und eine Verschiebung von P_{R} nach „unten“. Damit ergibt sich im *deformierten* Zustand als Abstand zwischen P_{W} und P_{R}:

$$d(\xi_x,\xi_y) + \overline{u}_z^{(W)}(\xi_x,\xi_y) + \overline{u}_z^{(R)}(\xi_x,\xi_y) - \delta\ . \tag{4.19}$$

Hierbei bezeichnet

$$\delta := \overline{u}_z^{(W)}(0,0) + \overline{u}_z^{(R)}(0,0) \tag{4.20}$$

die *elastische Annäherung* der beiden Festkörper Rad und Schiene. Ω ist genau dann die Kontaktfläche, wenn gilt

$$\begin{aligned} d(\xi_x,\xi_y) + \overline{u}_z^{(W)}(\xi_x,\xi_y) + \overline{u}_z^{(R)}(\xi_x,\xi_y) &= \delta\ , \quad \text{falls } (\xi_x,\xi_y)\in\Omega\ , \\ d(\xi_x,\xi_y) + \overline{u}_z^{(W)}(\xi_x,\xi_y) + \overline{u}_z^{(R)}(\xi_x,\xi_y) &> \delta\ , \quad \text{falls } (\xi_x,\xi_y)\notin\Omega\ . \end{aligned}$$

Durch diese Bedingungen werden für den elastischen Rad–Schiene–Kontakt implizit die Größe und Lage der Kontaktfläche Ω sowie die Normaldruckverteilung $p(\xi'_x,\xi'_y)$ innerhalb der Kontaktfläche bestimmt. Hieraus erhält man schließlich $\overline{u}_z^{(W)}$ und $\overline{u}_z^{(R)}$ und insbesondere auch δ.

Die elastische Annäherung δ wird also neben Materialeigenschaften vor allem durch die Normalkraft F_N und durch die Geometrie der Oberflächen von Rad und Schiene in

der Kontaktfläche bestimmt. Dabei gilt z. B. im klassischen Hertzschen Kontaktmodell $\delta \sim |F_N|^{2/3}$. Umgekehrt ergibt sich für jede relative Lage q_{rel} des Rades zur Schiene, für die sich Rad und Schiene als Starrkörper durchdringen würden (d. h. $\delta > 0$ in Abb. 4.12), eine rücktreibende Kraft F_N (z. B. $|F_N| \sim \delta^{3/2}$), die in einem elastischen Kontaktmodell der Zwangskraft $-\tilde{G}'^T(q,s)\lambda$ des Starrkörperkontaktmodells (3.14) entspricht.

Wie schon in Bemerkung 35b erwähnt, führt die quantitative Berücksichtigung des Zusammenhangs von δ und F_N im elastischen Kontaktmodell auf einen Lösungsanteil, der mit hoher Frequenz oszilliert, dessen Amplitude jedoch weit unterhalb der für das MKS–Modell typischen Genauigkeitsforderungen liegt. Dieser in praxi nicht interessierende Lösungsanteil verkompliziert die numerische Lösung erheblich, so daß der Rechenaufwand sehr viel größer als im Starrkörperkontaktmodell ist.

Statt dessen soll im folgenden die elastische Deformation nur qualitativ berücksichtigt werden. Nach Johnson ([95, Abschnitt 4.1]) kann man die Lage und Größe der Kontaktfläche recht gut qualitativ (jedoch nicht quantitativ) beschreiben, indem für verschiedene Zahlenwerte Δ_0 die Niveaulinien des Abstands zwischen Rad und Schiene aufgetragen werden. Mit den Bezeichnungen aus Bemerkung 36 betrachten wir deshalb die Projektion der Kurven $\{\, P_{\mathrm{W}} : e_3^T \cdot (P_{\mathrm{R}}(P_{\mathrm{W}}) - P_{\mathrm{W}}) = \Delta_0 \,\}$ auf $\mathcal{E}$ (vgl. (4.8)). Abb. 4.13 auf S. 152 zeigt die projizierten Kurven für die beiden Konfigurationen aus Abb. 4.7 für $\Delta_0 = 20(20)100\,\mu$m, dabei sind die lokalen Minima von $\Delta(s;q_{\mathrm{rel}},x)$ durch „+“ markiert. Kleineren Werten von Δ_0 entsprechen kleinere Normalkräfte und damit in Abb. 4.13 die innen gelegenen Kurven, entsprechend veranschaulichen die Kurven für größere Werte von Δ_0 die Kontaktflächen für größere Normalkräfte.

Die Abb. 4.7 und 4.13 illustrieren, daß die Lage und die Größe der Kontaktfläche sehr viel weniger empfindlich gegenüber Änderungen von q_{rel} sind als die Lage des Kontaktpunkts. Statt der Starrkörperkontaktbedingung (4.9) fordern wir deshalb, daß ein *gewichteter Mittelwert* von $e_3^T \cdot (P_{\mathrm{R}}(P_{\mathrm{W}}) - P_{\mathrm{W}})$, d. h. ein gewichteter Mittelwert des Abstands des Rades zur Schiene, verschwindet. Der Mittelwert soll im wesentlichen durch die Funktionswerte derjenigen Punkte bestimmt sein, die im elastischen Kontaktmodell zur Kontaktfläche gehören. Der Einfluß der außerhalb der Kontaktfläche gelegenen Punkte auf die Kontaktbedingung soll dagegen vernachlässigbar klein sein. Wie im Starrkörperkontaktmodell reicht es aus, statt der zweidimensionalen Kontaktfläche nur die Kurve $\mathfrak{C}$ zu betrachten.

Verschiedene Ansätze dieser Art wurden implementiert und getestet ([65]). Als besonders vorteilhaft erweist sich die von Frischmuth ([22]) vorgeschlagene Funktion

$$\mathrm{smax}_s^{(\nu)}\zeta(s;q_{\mathrm{rel}},x) := \nu \ln\Big(\int_{\mathfrak{C}} \exp(\frac{1}{\nu}\zeta(s;q_{\mathrm{rel}},x))\, ds \;/\; \int_{\mathfrak{C}} ds \Big)\ , \tag{4.21}$$

in der ν einen kleinen positiven Parameter bezeichnet.

Lemma 12 *Gegeben sei eine stetige Funktion $\zeta(s;q_{\mathrm{rel}},x)$. Dann ist $\mathrm{smax}_s^{(\nu)}\zeta(s;q_{\mathrm{rel}},x)$ (mindestens) ebenso oft stetig differenzierbar bezüglich q_{rel} und x wie $\zeta(s;q_{\mathrm{rel}},x)$. Werden q_{rel} und x fixiert, so gilt für jedes $\nu > 0$*

a) $\mathrm{smax}_s^{(\nu)}\zeta(s;q_{\mathrm{rel}},x) \le \max_s \zeta(s;q_{\mathrm{rel}},x)$ *und*

b) $\lim\limits_{\nu\to 0} \mathrm{smax}_s^{(\nu)}\zeta(s;q_{\mathrm{rel}},x) = \max_s \zeta(s;q_{\mathrm{rel}},x)$.

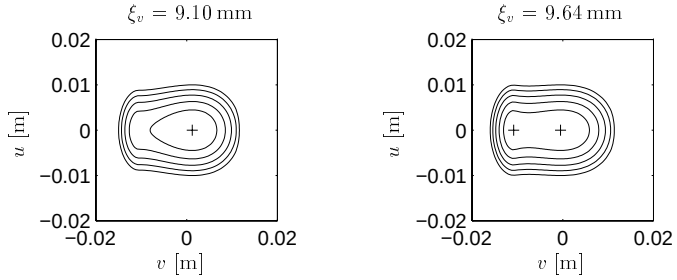


Abbildung 4.13: Konturlinien für den Abstand zwischen Rad (Profil S1002) und Schiene (Profil UIC60–ORE), $\varphi = 0.025$, $\psi = 0$, $\Delta_0 = 20(20)100\,\mu$m.

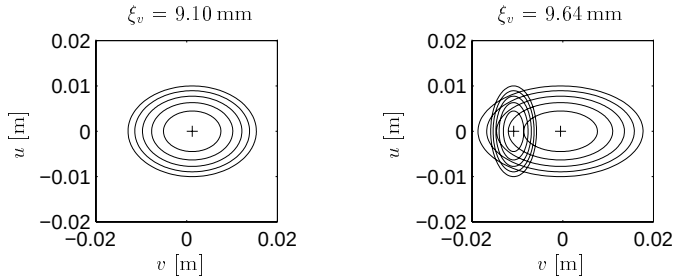


Abbildung 4.14: Konturlinien für den Abstand zwischen zwei Paraboloiden, die im klassischen Hertzschen Kontaktmodell die Oberflächen von Rad (Profil S1002) und Schiene (Profil UIC60–ORE) in der Umgebung eines Kontaktpunkts $P_{\mathrm{w}}(s_*)$ approximieren (aufgetragen für alle lokalen Minimalstellen s_* von Δ), $\varphi = 0.025$, $\psi = 0$, $\Delta_0 = 20(20)100\,\mu$m.

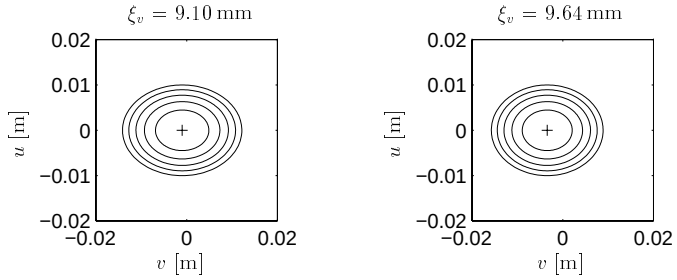


Abbildung 4.15: Konturlinien für den Abstand zwischen zwei Paraboloiden, die im verallgemeinerten Hertzschen Kontaktmodell die Oberflächen von Rad (Profil S1002) und Schiene (Profil UIC60–ORE) in der Umgebung von $P_{\mathrm{w}}(\check{s})$ approximieren, $\varphi = 0.025$, $\psi = 0$, $\Delta_0 = 20(20)100\,\mu$m.

Beweis a) folgt aus $\exp(\frac{1}{\nu}\zeta(s;q_{\mathrm{rel}},x)) \le \max_s \exp(\frac{1}{\nu}\zeta(s;q_{\mathrm{rel}},x))$.
b) Sei $\varepsilon > 0$ beliebig vorgegeben. Wegen der Stetigkeit von ζ existiert ein $\tilde{\varepsilon} > 0$, so daß es ein Teilstück $\mathfrak{K}$ von $\mathfrak{C}$ der Länge $\int_{\mathfrak{K}} ds \ge \tilde{\varepsilon} \cdot \int_{\mathfrak{C}} ds$ gibt, für das gilt

$$\zeta(s;q_{\mathrm{rel}},x) \ge \max_s \zeta(s;q_{\mathrm{rel}},x) - \varepsilon\ , \quad \text{falls } P_{\mathrm{w}}(s) \in \mathfrak{K}\ .$$

Damit folgt

$$\begin{aligned} \operatorname{smax}_s^{(\nu)}\zeta(s;q_{\mathrm{rel}},x) &\ge \nu \ln\Big(\exp(\frac{1}{\nu}(\max_s \zeta(s;q_{\mathrm{rel}},x)-\varepsilon)) \cdot \tilde{\varepsilon} \cdot \int_{\mathfrak{C}} ds \,/\, \int_{\mathfrak{C}} ds\Big) \\ &= \max_s \zeta(s;q_{\mathrm{rel}},x) - \varepsilon + \nu \ln \tilde{\varepsilon}\ , \end{aligned}$$

also $\lim_{\nu\to 0} \operatorname{smax}_s^{(\nu)}\zeta(s;q_{\mathrm{rel}},x) \ge \max_s \zeta(s;q_{\mathrm{rel}},x) - \varepsilon$. Da $\varepsilon > 0$ beliebig gewählt war, folgt die Behauptung. ■

Mit dieser Funktion $\operatorname{smax}_s^{(\nu)}$ wird die *quasi-elastische Kontaktbedingung*

$$0 = \gamma^{(\nu)}(q) := -\xi_w - \operatorname*{smax}_{s\in[\underline{s},\overline{s}]}^{(\nu)} \zeta(s;q_{\mathrm{rel}},x) \tag{4.22}$$

definiert, die die elastische Deformation von Rad und Schiene qualitativ, aber nicht quantitativ berücksichtigt. Nach Lemma 12 definiert (4.22) die Höhenauslenkung $-\xi_w$ so, daß sich Rad und Schiene als Starrkörper i. allg. geringfügig durchdringen würden:

$$\delta^{(\nu)} := \max_{s\in[\underline{s},\overline{s}]} \zeta(s;q_{\mathrm{rel}},x) - \operatorname*{smax}_{s\in[\underline{s},\overline{s}]}^{(\nu)} \zeta(s;q_{\mathrm{rel}},x) \ge 0$$

(vgl. auch (4.9)). So ist es naheliegend, den Parameter $\nu > 0$ so zu wählen, daß $\delta^{(\nu)}$ etwa der elastischen Annäherung δ aus (4.20) entspricht. Je nach Achslast liegen typische Werte für δ im Bereich $10\,\mu\mathrm{m} \ldots 100\,\mu\mathrm{m}$. In umfangreichen Testrechnungen erwiesen sich quasi-elastische Kontaktbedingungen (4.22) mit $\nu \in [10^{-5}, 5\cdot 10^{-5}]$ als besonders geeignet für die Profile S1002 (Rad) und UIC60–ORE (Schiene). Abb. 4.16 auf S. 154 zeigt die nach dem Starrkörperkontaktmodell (4.9) und die nach dem quasi-elastischen Kontaktmodell (4.22) definierte Höhenauslenkung $-\xi_w$ in Abhängigkeit von ξ_v. Die gepunkteten Linien deuten die Unstetigkeitsstellen der Ableitung der Starrkörperkontaktbedingung, d. h. die Sprungstellen des Kontaktpunkts an.

Verwendet man in den Modellgleichungen das quasi-elastische Kontaktmodell, so ergeben sich statt (4.10) die Kontaktbedingungen $0 = g^{(\nu)}(q) = (\,\gamma_1^{(\nu)}(q)\,,\,\ldots\,,\,\gamma_M^{(\nu)}(q)\,)^T$ mit $\lim_{\nu\to 0} g^{(\nu)}(q) = g(q)$. Da keine Kontaktpunktkoordinaten zu berücksichtigen sind, haben die Modellgleichungen die Standardform (4.4). Gleichzeitig werden die für das Starrkörperkontaktmodell charakteristischen Unstetigkeiten in $\dot{q}$ und λ vermieden, weil $g^{(\nu)}(q) = 0$ nach Lemma 12 eine hinreichend oft stetig differenzierbare Zwangsbedingung definiert. In diesem Sinne kann der Übergang vom Starrkörperkontaktmodell zum quasi-elastischen Kontaktmodell als *Regularisierung* der Bewegungsgleichungen interpretiert werden ([64]).

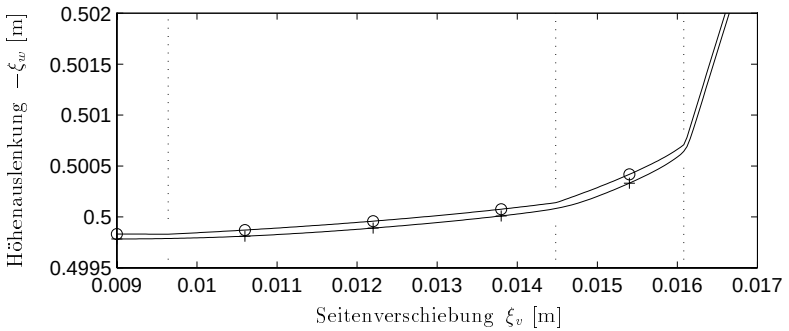


Abbildung 4.16: Höhenauslenkung des Rades nach der Starrkörperkontaktbedingung (4.9) („∘“) bzw. nach der quasi-elastischen Kontaktbedingung (4.22) mit $\nu = 2 \cdot 10^{-5}$ („+“). Gleisprofil UIC60–ORE, Radprofil S1002. $\varphi = 0.025$, $\psi = 0$.

Bemerkung 41 a) Die Starrkörperkontaktbedingung (4.9) für den Rad-Schiene-Kontakt ergab sich in Abschnitt 4.1 als Spezialfall von Kontaktbedingungen für beliebige Starrkörper (vgl. Beispiel 24). Noch allgemeiner ist (4.9) ein Beispiel für eine (einseitige) Beschränkung, die sich in einem DA–System ergibt, wenn aus einer Schar von (unendlich vielen) einseitigen Beschränkungen stets mindestens eine aktiv sein soll. Die Regularisierung der Starrkörperkontaktbedingung durch den Übergang von (4.9) zu (4.22) läßt sich direkt auf diese allgemeinere Situation übertragen ([22]). Dabei hängt es sehr stark von der jeweiligen Anwendung ab, ob der Übergang vom Grenzwert $\nu \to 0$ zu endlichen Werten von ν physikalisch gerechtfertigt oder sogar — wie im Rad–Schiene–Kontaktmodell — wünschenswert ist.

b) Im quasi-elastischen Kontaktmodell hängt die Lösung $q(t)$, $\lambda(t)$ der Modellgleichungen vom Parameter $\nu > 0$ ab. Für sehr einfache Modellprobleme (Bewegung einer Punktmasse auf einer Mannigfaltigkeit mit Knick) konnte nachgewiesen werden, daß diese Lösung für $\nu \to 0$ gegen einen wohl definierten Grenzwert $q^*(t)$, $\lambda^*(t)$ konvergiert, wobei $q^*(t)$ die (nur stückweise differenzierbaren) Zwangsbedingungen des Starrkörperkontaktmodells erfüllt ([62], [88]). Mit $q^*(t)$ und $\lambda^*(t)$ ist also für das Starrkörperkontaktmodell eine spezielle Lösung gegeben, deren Übergänge von $\dot{q}^*(t^{(j)} - 0)$ zu $\dot{q}^*(t^{(j)} + 0)$ und von $\lambda^*(t^{(j)} - 0)$ zu $\lambda^*(t^{(j)} + 0)$ für alle Zeitpunkte $t^{(j)} \in [0, T]$, für die sich $\dot{q}^*$ und λ^* unstetig ändern, durch den Grenzübergang $\nu \to 0$ eindeutig bestimmt sind. Der Grenzübergang $\nu \to 0$ im quasi-elastischen Kontaktmodell eröffnet damit einen systematischen Weg, Übergangsbedingungen für die unstetigen Zustandsänderungen im Starrkörperkontaktmodell zu formulieren.

Für die Modellierung des Rad–Schiene–Kontakts ist es jedoch wegen des Modellfehlers im Starrkörperkontaktmodell wenig sinnvoll, die quasi-elastische Kontaktbedingung (4.22) für $\nu \to 0$ zu betrachten, denn für Parameterwerte $\nu > 0$ mit $\delta^{(\nu)} \approx \delta$ erhält man eine aus *physikalischer* Sicht wesentlich bessere Kontaktbedingung für den elastischen Kontakt von Rad und Schiene.

Durch die qualitative Berücksichtigung der elastischen Deformation von Rad und Schiene im quasi-elastischen Kontaktmodell ist es möglich, auch bei der Berechnung der Reibungskräfte die konkrete Gestalt der Kontaktfläche sehr viel besser zu berücksichtigen als im Starrkörperkontaktmodell. Die Modellierung und Berechnung der Reibungskräfte zwischen Rad und Schiene ist kompliziert. Wir verzichten hier auf eine detaillierte Darstellung und verweisen statt dessen auf die Monographie [97] von Kalker, dessen Rollreibungstheorie als The state of the art gelten kann. Zur Berechnung der Reibungskräfte sind nicht nur wie in Bemerkung 40 Kräfte in normaler, sondern auch in tangentialer Richtung zu berücksichtigen. Wegen des außerordentlich hohen numerischen Aufwands zur vollständigen Lösung des normalen und des tangentialen Kontaktproblems sind in MKS–Modellen vereinfachende Annahmen zwingend erforderlich. Hierzu gibt es eine Vielzahl verschiedener Ansätze ([97]). Die Entscheidung, ob ein Ansatz geeignet ist oder nicht, basiert meist nicht auf Fehlerabschätzungen, sondern auf dem Vergleich von Simulationsergebnissen mit gemessenen Daten. Deshalb werden solche Einschätzungen in der Literatur oft kontrovers diskutiert, wobei sich die Entscheidungskriterien im Laufe der Zeit durchaus ändern können.

Ein typisches Beispiel ist das *klassische Hertzsche Kontaktmodell* zur Lösung des normalen Kontaktproblems aus Bemerkung 40. Hierzu werden die undeformierten Oberflächen von Rad und Schiene durch Paraboloide ersetzt, deren Scheitel im Kontaktpunkt P^*_{W} liegt (vgl. (4.15)) und deren Achsen parallel zur ξ_z-Achse sind. Die Paraboloide werden eindeutig durch die Bedingung definiert, daß im Kontaktpunkt die Krümmungen des ersten Paraboloids mit den Krümmungen der Radoberfläche und die Krümmungen des zweiten Paraboloids mit den Krümmungen der Gleisoberfläche übereinstimmen sollen. Der Abstand $d(\xi_x, \xi_y)$ der undeformierten Oberflächen von Rad und Schiene wird damit approximiert durch den Abstand der Paraboloide und in (4.19) gilt näherungsweise

$$d(\xi_x, \xi_y) = A\xi_x^2 + B\xi_y^2 \quad \text{mit gewissen Konstanten } A, B > 0\,. \tag{4.23}$$

Dabei wurden die ξ_x- und die ξ_y-Achse so festgelegt, daß der Koeffizient des Terms $\ldots \xi_x\xi_y$ verschwindet. Für Funktionen d der Gestalt (4.23) zeigte Hertz schon 1882, daß Ω elliptisch ist (*Kontaktellipse*). Sind die Halbachsen a und b der Kontaktellipse bekannt, so erhält man in Ω die Normaldruckverteilung $p(\xi'_x, \xi'_y) = p_0 \cdot (1 - \xi'^2_x/a^2 - \xi'^2_y/b^2)^{1/2}$ mit $p_0 := 3|F_N|/(2\pi ab)$, und die Koeffizienten A und B in (4.23) können unter Verwendung von elliptischen Integralen in geschlossener Form angegeben werden. Damit sind a, b und $p(\xi'_x, \xi'_y)$ implizit als Funktionen von A, B und F_N gegeben, und es gilt $\delta = c|F_N|^{2/3}$ mit einem Proportionalitätsfaktor c, der durch A und B und durch die Materialkonstanten ν_{W}, ν_{R}, $E^{(W)}$ und $E^{(R)}$ eindeutig bestimmt ist ([95, Kapitel 4]). Von Kalker wurde der Algorithmus FASTSIM entwickelt und implementiert, mit dem man auf dieser Grundlage auch das tangentiale Kontaktproblem sehr effizient numerisch lösen kann, um schließlich die Rollreibungskräfte zu berechnen ([97, Kapitel 3]).

Betrachtet man die Radoberfläche nur längs der Kurve $\mathfrak{C}$, so entspricht der Approximation der Oberflächen durch Paraboloide eine Approximation der in Abb. 4.7 auf S. 144 dargestellten Funktion Δ durch eine Parabel, deren Scheitelpunkt in s_* liegt und die in s_* dieselbe Krümmung wie Δ hat. Schon diese sehr einfache Darstellung macht deutlich, daß eine solche Approximation für kegelförmige Räder sinnvoll ist, aber für Räder mit Verschleißprofil weder ein quantitativ noch ein qualitativ richtiges Bild von Δ ergibt.

Als Pendant zu Abb. 4.13 zeigt Abb. 4.14 auf S. 152 einige Niveaulinien des Abstands zwischen Paraboloiden, die nach dem klassischen Hertzschen Kontaktmodell bestimmt wurden (jeweils für alle Kontaktpunkte). Für $\xi_v = 9.10\,\mathrm{mm}$ (linkes Diagramm) ergibt sich eine Approximation von akzeptabler Genauigkeit. Für $\xi_v \approx 9.64\,\mathrm{mm}$ (rechtes Diagramm) springt der Kontaktpunkt von $s_*^- \approx 2.4\,\mathrm{mm}$ zu $s_*^+ \approx -8.0\,\mathrm{mm}$ und die Approximation der Kontaktfläche springt von den Ellipsen um s_*^- (die sehr viel größer als die reale Kontaktfläche sind) zu den Ellipsen um s_*^+, die die Kontaktfläche völlig unzureichend beschreiben.

Wendet man das klassische Hertzsche Kontaktmodell auf den Kontakt einer Schiene mit einem Rad mit Verschleißprofil an, so ist der Modellfehler groß. Die Auswirkungen dieses Modellfehlers auf die Genauigkeit der berechneten Reibungskräfte und auf die Simulationsergebnisse insgesamt werden in der Literatur unterschiedlich bewertet. Knothe und Le The schreiben z. B. 1983 „[die Verwendung des klassischen Hertzschen Kontaktmodells] ... mag zulässig sein, solange man sich nur für das Bewegungsverhalten eines Radsatzes oder eines ganzen Schienenfahrzeugs auf dem Gleis interessiert“ ([98]). Dagegen sollte das 1990 von Pascal aufgestellt Benchmark–Problem aus Beispiel 30 belegen, daß die Laufdynamik von Schienenfahrzeugen falsch wiedergegeben werden kann, wenn man bei der dynamischen Simulation das klassische Hertzsche Kontaktmodell einsetzt ([126]).

Simulationspakete verwendeten bis Ende der achtziger Jahre vorwiegend das klassische Hertzsche Kontaktmodell (auch für Räder mit Verschleißprofil). Seitdem versucht man auf vielfältige Weise, die kompliziertere Kontaktgeometrie von Rad und Schiene zu berücksichtigen ohne den numerischen Aufwand zur Berechnung der Reibungskräfte zu stark anwachsen zu lassen (vgl. z. B. [98], [127], [128]). Einen besonders einfachen Zugang bietet das quasi-elastische Kontaktmodell: Wie im Starrkörperkontaktmodell nimmt man an, daß die Reibungskraft in einem einzelnen Punkt angreift. Statt des Kontaktpunkts $P_{\mathrm{W}}^* = P_{\mathrm{W}}(s_*)$ wird als Angriffspunkt von Zwangs- und Reibungskräften ein Punkt $P_{\mathrm{W}}(\tilde{s}) \in \mathfrak{C}$ verwendet, wobei $\tilde{s}$ in Anlehnung an (4.22) als gewichteter Mittelwert von s entlang der Kurve $\mathfrak{C}$ definiert ist:

$$\tilde{s} := \int_{\mathfrak{C}} s \, \exp(\frac{1}{\nu}\zeta(s; q_{\mathrm{rel}}, x))\, ds \;\; / \;\; \int_{\mathfrak{C}} \exp(\frac{1}{\nu}\zeta(s; q_{\mathrm{rel}}, x))\, ds \; . \tag{4.24}$$

Entsprechend wählt man zur Beschreibung der elastischen Deformation die Ebene $\mathcal{E}$ aus Bemerkung 40 so, daß die Zwangskraft $-G^{(\nu)}(q)^T\lambda$ mit $G^{(\nu)}(q) := \frac{\partial}{\partial q} g^{(\nu)}(q)$ in Normalenrichtung zu $\mathcal{E}$ wirkt und $P_{\mathrm{W}}(\tilde{s}) \in \mathcal{E}$ gilt. Wie in Bemerkung 40 wird ein kartesisches Koordinatensystem (ξ_x, ξ_y, ξ_z) so definiert, daß die ξ_x- und die ξ_y-Achse die Ebene $\mathcal{E}$ aufspannen, Koordinatenursprung ist $P_{\mathrm{W}}(\tilde{s})$. In natürlicher Weise ergibt sich aus (4.24) eine Verallgemeinerung des Hertzschen Kontaktmodells: Die ξ_x- und die ξ_y-Achse werden so vorgegeben, daß die Radachse und damit auch die s-Achse des in Bemerkung 36 eingeführten Koordinatensystems $\boldsymbol{W}$ parallel zur (ξ_y, ξ_z)–Ebene liegen. Nun betrachtet man Paraboloide der Form $\{\,(\xi_x, \xi_y, \xi_z)^T : \xi_z = \alpha\xi_x^2 + \beta\xi_y^2\,\}$ mit gewissen $\alpha, \beta \in \mathbb{R}$, d. h. Paraboloide, deren Scheitel in $P_{\mathrm{W}}(\tilde{s})$ liegt, deren Achse mit der ξ_z-Achse zusammenfällt und deren Hauptkrümmungsrichtungen durch die ξ_x- und die ξ_y-Achse gegeben sind. Diese Paraboloide sind durch Angabe ihrer Hauptkrümmungen im Scheitelpunkt, die mit $\kappa_x^{(\mathrm{W})}$, $\kappa_y^{(\mathrm{W})}$, $\kappa_x^{(\mathrm{R})}$ und $\kappa_y^{(\mathrm{R})}$ bezeichnet seien, eindeutig bestimmt (κ_x ist die Krümmung in

ξ_x-Richtung, κ_y diejenige in ξ_y-Richtung, jeweils für „**W**heel“ und „**R**ail“). Die Krümmungen $\kappa_x^{(\mathrm{W})}$ und $\kappa_y^{(\mathrm{W})}$ werden als gewichtete Mittelwerte der Krümmungen der Radoberfläche definiert:

$$\kappa_x^{(\mathrm{W})} := \int_{\mathfrak{C}} \kappa_x(P_{\mathrm{W}}(s))\, \exp(\frac{1}{\nu}\zeta(s; q_{\mathrm{rel}}, x))\, ds \;\; / \;\; \int_{\mathfrak{C}} \exp(\frac{1}{\nu}\zeta(s; q_{\mathrm{rel}}, x))\, ds \; , \;\; \ldots \; .$$

$\kappa_x(P_{\mathrm{W}}(s))$ bezeichnet dabei die im Punkt $P_{\mathrm{W}}(s) \in \mathfrak{C}$ bestimmte Krümmung der Radoberfläche in ξ_x-Richtung, entsprechend ist $\kappa_y(P_{\mathrm{W}}(s))$ die Krümmung in ξ_y-Richtung. Auch für $\kappa_x^{(\mathrm{R})}$ und $\kappa_y^{(\mathrm{R})}$ wird ein längs $\mathfrak{C}$ berechneter gewichteter Mittelwert verwendet. Für $\kappa_x(.)$ und $\kappa_y(.)$ sind dabei jedoch nicht die Krümmungen der Radoberfläche in $P_{\mathrm{W}}(s)$, sondern die Krümmungen der Gleisoberfläche in $P_{\mathrm{R}}(P_{\mathrm{W}}(s))$ einzusetzen (wiederum in ξ_x- und ξ_y-Richtung).

Durch diese lokale Approximation der undeformierten Oberflächen von Rad und Schiene durch Flächen 2. Ordnung wird der Abstand zwischen beiden Körpern wie in (4.23) durch eine in ξ_x und ξ_y quadratische Funktion angenähert. Die Kontaktfläche Ω sowie $p(\xi_x', \xi_y')$ und δ können analog zur klassischen Hertzschen Theorie bestimmt werden. Lösungsansätze für das tangentiale Kontaktproblem, die auf dem klassischen Hertzschen Kontaktmodell aufbauen, kann man direkt auf diese Verallgemeinerung übertragen. So ist es wie zuvor möglich, die Reibungskräfte mit dem Programm FASTSIM von Kalker schnell und zuverlässig zu berechnen.

Der entscheidende Vorteil des verallgemeinerten Hertzschen Kontaktmodells ist die wesentlich bessere Approximation der Oberflächen von Rad und Schiene. Hierzu zeigt Abb. 4.15 auf S. 152 wiederum die Niveaulinien des Abstands zwischen den Paraboloiden. Legt man die Genauigkeit eines MKS–Modells zu Grunde, so ergibt sich eine sehr gute Approximation der Daten aus Abb. 4.13. Durch die Bildung der Mittelwerte längs $\mathfrak{C}$ wird insbesondere auch erreicht, daß das verallgemeinerte Hertzsche Kontaktmodell in Fällen, für die schon das klassische Modell zu zufriedenstellenden Ergebnissen führt, in guter Näherung mit dem klassischen Hertzschen Kontaktmodell übereinstimmt (z. B. für kegelförmige Räder).

Ebenso wie andere Modelle für die Beschreibung des geometrischen Kontakts zwischen Rad und Schiene läßt sich auch das quasi-elastische Kontaktmodell (4.22) mit verschiedenen Modellen zur Berechnung der Reibungskräfte koppeln. In den Beispielen 31 und 32 werden die Reibungskräfte mit FASTSIM auf der Basis des verallgemeinerten Hertzschen Kontaktmodells bestimmt. Zahlreiche Testrechnungen belegen, daß das verallgemeinerte Hertzsche Kontaktmodell im Rahmen der Genauigkeit eines MKS–Modells gut als Grundlage der Berechnung der Reibungskräfte geeignet ist, wenn die Kontaktfläche zwischen Rad und Schiene einfach zusammenhängend ist. Bei speziellen Fahrmanövern (z. B. im Bogenlauf) ist es möglich, daß über einen längeren Zeitraum die Flanke und die Lauffläche eines Rades gleichzeitig in Kontakt mit dem Gleis kommen (sog. Mehrpunktkontakt). Für eine Erweiterung des Kontaktmodells, die dieser Situation angepaßt ist, liegen ebenfalls positive praktische Erfahrungen vor ([119]).

Bemerkung 42 Die Verwendung von Ellipsen zur Approximation der Kontaktfläche stößt auf prinzipielle Schwierigkeiten, wenn nicht nur die Laufdynamik des Rad–Schiene–Systems, sondern außerdem auch der Verschleiß von Rad und Schiene im Kontaktbereich

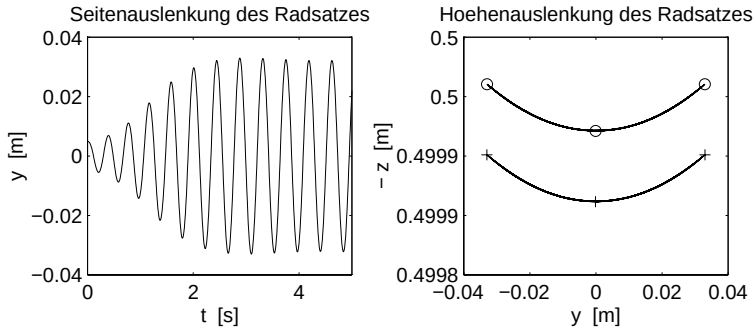


Abbildung 4.17: Bewegung eines starren Radsatzes mit Kegelrädern im Geradenlauf, $V_0 = 67.1\,\mathrm{ms}^{-1}$. Links: $y = y(t)$, rechts: Höhenauslenkung $-z$ für Starrkörperkontaktmodell (4.9) („∘“) und quasi-elastisches Kontaktmodell (4.22) mit $\nu = 2 \cdot 10^{-5}$ („+“).

Tabelle 4.1: Numerischer Aufwand in Beispiel 31 (NSTEP: Zahl der Integrationsschritte, NFUNC: Zahl der Funktionsaufrufe, NREP: Zahl der Schrittwiederholungen).

	NSTEP	NFUNC	NREP
Starrkörperkontaktmodell	796	1794	75
quasi-elastisches Modell	794	1811	81

simuliert werden soll. Hierzu ist die tatsächliche (nicht-elliptische) Gestalt der in Abb. 4.13 angedeuteten Kontaktfläche zu berücksichtigen, der Rechenaufwand wächst dadurch stark an (vgl. z. B. [97], [102]).

Zwei Simulationsbeispiele sollen am Ende dieses Abschnitts verdeutlichen, daß das quasi-elastische Kontaktmodell tatsächlich das für kegelförmige Räder bewährte Starrkörperkontaktmodell so verallgemeinert, daß es zur dynamischen Simulation von Rad–Schiene–Systemen auch dann geeignet ist, wenn die Räder ein Verschleißprofil haben.

Beispiel 31 Zunächst wird für den starren Radsatz mit Kegelrädern aus Beispiel 27 das Starrkörperkontaktmodell (4.9) (mit klassischem Hertzschen Kontaktmodell zur Berechnung der Reibungskräfte) verglichen mit dem quasi-elastischen Kontaktmodell (4.22) (Berechnung der Reibungskräfte nach dem verallgemeinerten Hertzschen Kontaktmodell mit $\nu = 2 \cdot 10^{-5}$).
Tab. 4.1 zeigt für das in Beispiel 27 beschriebene Fahrmanöver „Grenzzyklus“ den numerischen Aufwand für beide Modelle bei Integration der GGL–Formulierung der Modellgleichungen mit ODASSL (TOL = 10^{-5} für q, s, $\dot{q}$ und TOL = 10^{-2} für λ)[1]. Die Zahl

[1] Der Autor dankt Herrn Dipl.-Ing. H. Netter (Oberpfaffenhofen) für die Bereitstellung der FORTRAN–Quelltexte eines Radsatzmodells, auf deren Grundlage das quasi-elastische Kontaktmodell implementiert und getestet werden konnte.

der Integrationsschritte und der Funktionsaufrufe ist für beide Modelle nahezu gleich. Die geringe Zahl von Schrittwiederholungen unterstreicht, daß für kegelförmige Räder beide Modelle auf glatte Zwangsbedingungen führen. Die Rechenzeit für das quasi-elastische Modell wird im wesentlichen dadurch bestimmt, wie die Kurvenintegrale in (4.22) und (4.24) während der Integration ausgewertet werden (vgl. Abschnitt 4.4).
Bis auf die Höhenauslenkung $-z$ fallen die Simulationsergebnisse für beide Modelle im Rahmen der vorgegebenen Genauigkeit zusammen. Abb. 4.17 zeigt hierzu links die mit dem quasi-elastischen Kontaktmodell berechnete Seitenauslenkung y des Radsatzes (vgl. Abb. 4.4 links). Im rechten Diagramm wird für beide Modelle die Höhenauslenkung $-z$ des Radsatzes angegeben. Die mit dem quasi-elastischen Kontaktmodell berechneten Werte entsprechen dabei etwa einer elastischen Annäherung $\delta^{(\nu)} \approx 60.0\,\mu\mathrm{m}$.

Beispiel 32 Abb. 4.18 zeigt Simulationsergebnisse für die Anwendung des quasi-elastischen Kontaktmodells auf das in Beispiel 30 beschriebene Benchmark–Problem von Pascal. Zunächst wird der Radsatz durch die seitliche Führungskraft F_y weit nach rechts verschoben. Kurzzeitig kommt die Flanke des rechten Rades in Kontakt mit dem Gleis. Dabei bleibt die Kontaktfläche jedoch einfach zusammenhängend, das verallgemeinerte Hertzsche Kontaktmodell kann angewendet werden.
Im Unterschied zu den unbefriedigenden Simulationsergebnissen für das Starrkörperkontaktmodell (Abb. 4.9) wird die Annäherung der Lösung an die quasi-stationäre Lage bei Anwendung des quasi-elastischen Kontaktmodells qualitativ richtig berechnet. In [119] betrachtet Netter dieses Benchmark–Problem für verschiedene Reibungskoeffizienten μ. Die mit dem quasi-elastischen Kontaktmodell ermittelten Simulationsergebnisse sind dabei in guter Übereinstimmung mit der von Pascal angegebenen Referenzlösung ([126]).

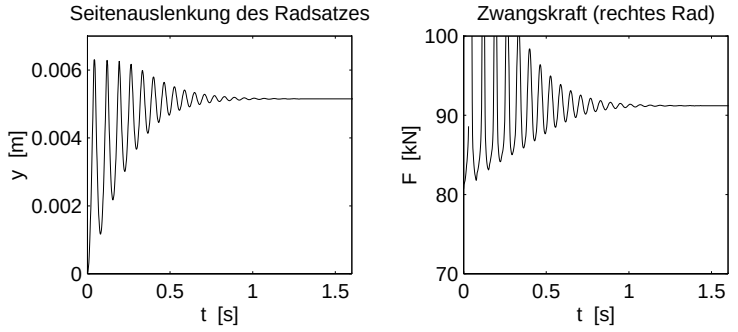


Abbildung 4.18: Starrer Radsatz im Geradenlauf: Anwendung des quasi-elastischen Kontaktmodells auf ein Benchmark–Problem nach Pascal ([126]).

4.4 Zur Implementierung des quasi-elastischen Kontaktmodells im Simulationspaket SIMPACK

Das quasi-elastische Kontaktmodell beschreibt den Kontakt zwischen Rad und Schiene mit einer Genauigkeit, die für einen großen Anwendungsbereich ausreichend ist. Für Räder mit Verschleißprofil vermeidet das quasi-elastische Kontaktmodell sowohl die für das Starrkörperkontaktmodell typischen Singularitäten im DA–System als auch die für rein elastische Kontaktmodelle charakteristischen Schwierigkeiten bei der numerischen Lösung der Modellgleichungen.

In diesem Abschnitt soll am Beispiel des MKS–Simulationspakets SIMPACK gezeigt werden, wie das quasi-elastische Kontaktmodell zur dynamischen Simulation von Schienenfahrzeugen eingesetzt werden kann. Dabei konzentrieren wir uns insbesondere auf die Auswertung der Funktionswerte in der Kontaktbedingung (4.22) und in (4.24).

Simulationspakete für mechanische Mehrkörpersysteme

Zur computergestützten Generierung der MKS–Modellgleichungen (4.4) wurden zahlreiche Mehrkörperformalismen entwickelt (vgl. z. B. [141]). Entsprechend groß ist die Zahl der verfügbaren MKS–Simulationspakete. Die Simulationspakete unterscheiden sich u. a. hinsichtlich ihres potentiellen Anwendungsbereichs (z. B. Robotik, Straßenfahrzeugbau, Schienenfahrzeuge), hinsichtlich der Benutzerführung (z. B. Pre- und Post-Processing, Erstellung des MKS–Modells) und insbesondere auch hinsichtlich der Qualität der verwendeten numerischen Verfahren. Im folgenden beschränken wir uns auf das MKS–Simulationspaket SIMPACK, das in den vergangenen Jahren von der SIMPACK–Arbeitsgruppe innerhalb der Abteilung Fahrzeugsystemdynamik des Instituts für Robotik und Systemdynamik der DLR (Oberpfaffenhofen) um eine sog. „vollständige Rad–Schiene–Funktionalität“ erweitert wurde ([142]). Damit wurde die Analyse der Laufdynamik von Rad–Schiene–Systemen zu einem der wesentlichen Anwendungsgebiete von SIMPACK.

Ein Vergleich von SIMPACK mit anderen Simulationspaketen, die ähnliche Anwendungsmöglichkeiten bieten, ist nicht Gegenstand der vorliegenden Arbeit. Hervorgehoben sei lediglich, daß SIMPACK moderne Integrationssoftware (u. a. DASSL, RADAU5) zur numerischen Lösung der Modellgleichungen in Deskriptorform (4.4) einsetzt (andere Simulationspakete verwenden z. T. noch heute Lösungsverfahren ohne Fehlerkontrolle und / oder ohne Schrittweitensteuerung). Darüberhinaus steht in SIMPACK mit dem in Abschnitt 3.1 besprochenen Integrator ODASSL von Führer ([66]) eine Modifikation von DASSL zur Verfügung, die speziell der Struktur der MKS–Modellgleichungen angepaßt ist. Effiziente numerische Lösungsverfahren sind ein wesentlicher Baustein des Simulationspakets. Mindestens ebenso wichtig sind

- Programmbausteine, die den Anwender beim schnellen und fehlerfreien Aufstellen des MKS–Modells unterstützen (Pre-Processing),
- Programmbausteine zur Analyse der Simulationsergebnisse (Post-Processing) und
- Schnittstellen zu anderen Programmpaketen (z. B. für Modellstrukturen zur computergestützten Konstruktion (CAD), zum computergestützten Reglerentwurf (CACE) und zur Verwendung von Finite–Elemente–Diskretisierungen (FEM) in der Festigkeitsberechnung).

Grundlage des Modellaufbaus in SIMPACK ist ein streng modulares Konzept, d. h., MKS–Modelle werden unter Verwendung von Modellbibliotheken aus einfachen Substrukturen aufgebaut. Für Schienenfahrzeuge ist eine solche Substruktur z. B. der Kontakt eines einzelnen Rades mit einer Schiene. Während der menü-geführten Eingabe und Definition der geometrischen und physikalischen Parameter des MKS–Modells werden im Pre-Processing zahlreiche Eingabedaten graphisch veranschaulicht („Modellkontrolle“). Allein schon die Vorgabe der Schienengeometrie kann (z. B. bei einer Weichenfahrt oder bei der Berücksichtigung von gemessenen Gleislagefehlern) sehr aufwendig sein und erfordert ggf. mehrere Megabyte Speicherplatz. Zur Analyse der Simulationsergebnisse bietet SIMPACK vielfältige Möglichkeiten für die automatisierte Variation von Systemparametern und für die statistische Auswertung der Simulationsergebnisse („Bewertungsfilter“). Damit wird es auch möglich, von der Analyse des MKS–Modells zur Systemauslegung, d. h. zur Optimierung von Systemparametern, überzugehen.

Aus diesen charakteristischen Merkmalen eines MKS–Simulationspakets ergeben sich Anforderungen an die Modellierung und an die Verfahren zur numerischen Lösung der MKS–Modellgleichungen, die weit über die in den Beispielen 27 und 32 betrachtete Simulation eines starren Radsatzes auf einem geraden Gleis hinausgehen. Das Modell für den Rad–Schiene–Kontakt muß sowohl für fahrwegabhängige Profile (z. B. Weichenfahrt) als auch für neuartige Konstruktionsansätze (z. B. Losradsätze, schlupfgeregelte Radsätze, gelenkte Einzelräder) uneingeschränkt einsetzbar sein. Dies gab die Motivation für die in Abschnitt 4.3 besprochene Entwicklung des quasi-elastischen Kontaktmodells.

Die Software zur Integration der Modellgleichungen muß zuverlässig funktionieren und robust sein gegenüber fehlerhaften Eingabedaten und gegenüber Modellkomponenten, die die in theoretischen Untersuchungen (vgl. z. B. Kapitel 3) üblichen Glattheitsvoraussetzungen nicht erfüllen. Wegen der großen Anzahl zu lösender Anfangswertprobleme für MKS–Modellgleichungen ist die effiziente numerische Lösung außerordentlich wichtig. Typische Vereinfachungen der Modellstruktur bei häufig wiederkehrenden Simulationsaufgaben müssen unbedingt zur Beschleunigung der numerischen Integration ausgenutzt werden. Beispiele sind

- die kleinere Zahl der Freiheitsgrade bei starrer Kopplung der Räder eines Radsatzes (Entwicklung des „schnellen Radsatzelements“ in SIMPACK [142]) und
- Gleise mit einem Profil, das längs des Fahrwegs unverändert bleibt (dann ist Δ in (4.8) von x unabhängig und die Kontaktbedingung (4.22) kann wie nachfolgend beschrieben durch eine sehr schnell auszuwertende Approximation ersetzt werden).

Im Zusammenhang mit Kapitel 3 sei erwähnt, daß SIMPACK bisher ebenso wie andere Simulationspakete die MKS–Modellgleichungen in einer Form generiert, in der die einzelnen Funktionen $M(q)$, $f(q,\dot{q},\lambda,t)$ und $G(q,t)$ in (4.4) nicht separat, sondern nur gemeinsam ausgewertet werden können (meist in Form des Residuums $M(q)\ddot{q} - f(q,\dot{q},\lambda,t) + G^T(q,t)\lambda$). Für Modellgleichungen in Deskriptorform schränkt das derzeit die Men-

ge der sinnvoll verwendbaren Integratoren auf implizite Verfahren wie DASSL, ODASSL oder RADAU5 ein. Die partitionierten Verfahren aus Kapitel 3, für deren effiziente Implementierung die Ausnutzung der Struktur der MKS–Modellgleichungen (4.4) entscheidend ist (vgl. Abschnitt 3.3.2), erfordern innerhalb des Mehrkörperformalismus Eingriffe in die Algorithmen, nach denen die Modellgleichungen automatisch generiert werden. Derartige Veränderungen am Mehrkörperformalismus sind nicht trivial, sie sind — nicht nur für SIMPACK — Gegenstand der aktuellen Forschung ([67]).

Implementierung des quasi-elastischen Kontaktmodells

Stellt man die Modellgleichungen wie in [148] und [120] von Hand auf, so ist es insbesondere aufwendig, anzugeben, wie die relative Lage $q_{\rm rel}$ des Rades zur Schiene aus den MKS–Koordinaten q bestimmt wird (zumal auch $\frac{\partial}{\partial q} q_{\rm rel}(q)$ benötigt wird). In SIMPACK sind diese Koordinatentransformationen Teil der computergestützten Generierung der Modellgleichungen, so daß zu gegebenem q die in der Kontaktbedingung (4.22) des quasi-elastischen Kontaktmodells verwendete Funktion $q_{\rm rel}(q)$ einschließlich ihrer partiellen Ableitungen unmittelbar zur Verfügung steht.

Bei der Implementierung von (4.22) konzentrieren wir uns im folgenden auf die Diskretisierung der Kurvenintegrale ([25]). Aus Gründen des Rechenaufwandes ist es nicht möglich, bei jeder Auswertung der Kontaktbedingung (4.22) die Kurvenintegrale so genau zu berechnen, daß der Fehler gegenüber der dem Integrator vorgegebenen Toleranz vernachlässigbar klein bleibt. Statt dessen ersetzen wir die Kontaktbedingung (4.22) durch eine Diskretisierung, die ebenso wie (4.22) die beiden wesentlichen Kriterien für ein quasi-elastisches Kontaktmodell erfüllt (Definition einer hinreichend oft stetig differenzierbaren Kontaktbedingung, wobei $-\xi_w$ für konsistente Lagekoordinaten $q_{\rm rel}$ nur wenig von der durch ein rein elastisches Kontaktmodell definierten Höhenauslenkung abweicht).

Hierzu sei $\{ s_i : i = 0, 1, \ldots, N \}$ ein Gitter mit $\underline{s} = s_0 < s_1 < \ldots < s_N = \overline{s}$ und $h_i := s_i - s_{i-1}$, $(i = 1, \ldots, N)$. In der Kontaktbedingung (4.22) wird $\operatorname{smax}^{(\nu)}$ ersetzt durch

$$\operatorname*{smax}_{s\in[\underline{s},\overline{s}]}{}^{(\nu,h)} \zeta(s; q_{\rm rel}, x) := \nu \ln\Big(\frac{1}{\overline{s}-\underline{s}} \sum_{i=1}^{N} h_i \exp(\frac{1}{\nu}\zeta(s_i; q_{\rm rel}, x)) \Big) \tag{4.25}$$

(vgl. (4.21)). Entscheidend ist, daß das Gitter $\{s_i\}$ während der gesamten Simulation fixiert bleibt, so daß der beim Übergang von (4.21) zu (4.25) entstehende Diskretisierungsfehler hinreichend oft stetig differenzierbar bezüglich $q_{\rm rel}$ und x ist.

Aus Sicht des Rechenaufwandes sollte N in (4.25) möglichst klein sein. Wählt man jedoch N zu klein, so kann die Kontaktbedingung (4.22) mit (4.25) u. U. die Höhenauslenkung $-\xi_w$ so definieren, daß die Abweichung zum rein elastischen Kontaktmodell inakzeptabel groß wird. Ursache ist, daß $\operatorname{smax}_s^{(\nu,h)}\zeta$ für $\nu \to 0$ nicht wie $\operatorname{smax}_s^{(\nu)}\zeta$ gegen $\max\{\zeta(s; q_{\rm rel}, x) : \underline{s} \le s \le \overline{s}\}$, sondern gegen $\max\{\zeta(s_i; q_{\rm rel}, x) : i = 1, \ldots, N\}$ konvergiert ([63]).

Beispiel 33 Nach Bemerkung 36 hängt ζ von s, ξ_v, φ, ψ und x ab (vgl. (4.9)). Als stark vereinfachtes Beispiel sei die Funktion $\zeta(s; q_{\rm rel}, x) := -(s - \xi_v)^2$ für $s \in [0, 4]$ gegeben. Ist $\xi_v \in [0, 4]$, so konvergiert $\operatorname{smax}_s^{(\nu)}\zeta$ für $\nu \to 0$ gegen $\max_s \zeta(s; q_{\rm rel}, x) = 0$. Der

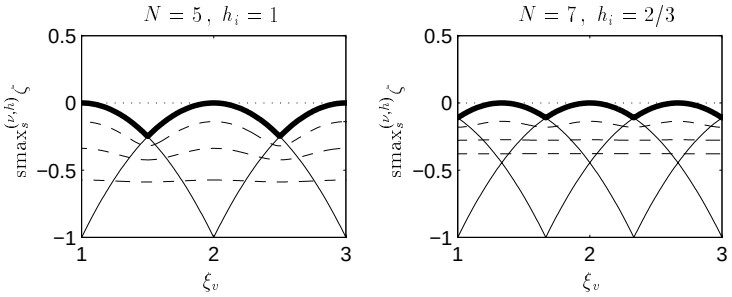


Abbildung 4.19: $\operatorname{smax}_s^{(\nu)}\zeta$ und $\operatorname{smax}_s^{(\nu,h)}\zeta$ für $\zeta(s; q_{\rm rel}, x) := -(s - \xi_v)^2$, vgl. Beispiel 33.

Grenzwert von $\operatorname{smax}_s^{(\nu,h)}\zeta$ für $\nu \to 0$ ist dagegen $\max_i \zeta(s_i; q_{\rm rel}, x) = -\min_i (s_i - \xi_v)^2$. Abb. 4.19 zeigt für äquidistante Gitter mit $N = 5$ (links) bzw. mit $N = 7$ (rechts) die Funktionen $\zeta(s_i; q_{\rm rel}, x)$, $(i = 0, 1, \ldots, N)$ als Funktionen von ξ_v (dünne durchgezogene Linien, dargestellt ist der Ausschnitt $\xi_v \in [1, 3]$). Die Grenzwerte von $\operatorname{smax}_s^{(\nu)}\zeta$ und $\operatorname{smax}_s^{(\nu,h)}\zeta$ für $\nu \to 0$ werden durch die gepunktete Linie bzw. durch die fette durchgezogene Linie angedeutet. Schließlich zeigen die gestrichelten Linien $\operatorname{smax}_s^{(\nu,h)}\zeta$ für $\nu = 0.1$, $\nu = 0.25$ und $\nu = 0.5$ (von oben nach unten).
Dieses einfache Beispiel zeigt anschaulich, daß für kleine Werte von ν eine zu grobe Diskretisierung in (4.25) zu künstlichen Oszillationen in der Kontaktbedingung führen kann. Zu gegebenem ν ergibt sich dagegen für hinreichend großes N eine Funktion $\operatorname{smax}_s^{(\nu,h)}\zeta$, die für die Verwendung in der Kontaktbedingung (4.22) sehr gut geeignet ist.

In praxi ist ν um mehrere Größenordnungen kleiner als in Abb. 4.19, und N wird in Abhängigkeit von ν bestimmt. Das Gitter $\{s_i\}$ wird auf der Lauffläche des Rades gröber und auf der Flanke feiner gewählt. Für die Profile UIC60–ORE (Gleis) und S1002 (Rad) und $\nu = 2 \cdot 10^{-5}$ hat sich ein Gitter mit $h_i \approx 0.5\,\text{mm}$ auf der Lauffläche und $h_i \approx 0.2\,\text{mm}$ auf der Flanke in zahlreichen Testrechnungen bewährt, hierzu sind $N \approx 200$ Auswertungen von ζ in (4.25) erforderlich.

Bei *jedem* Funktionsaufruf während der numerischen Integration der Modellgleichungen (4.4) sind also für *jedes* Rad etwa 200 Aufrufe der Abstandsfunktion erforderlich. Gegenüber dem Starrkörperkontaktmodell erscheint dieser Aufwand außerordentlich groß. Für Räder mit Verschleißprofil entsteht jedoch *unabhängig* von der verwendeten Kontaktbedingung ein wesentlich höherer numerischer Aufwand als für kegelförmige Räder, weil das klassische Hertzsche Kontaktmodell nicht anwendbar ist (vgl. Abschnitt 4.3). Bei der Berechnung der Reibungskräfte ist deshalb zur näherungsweisen Bestimmung der Kontaktfläche und der Normaldruckverteilung innerhalb der Kontaktfläche in jedem Fall die Auswertung der Abstandsfunktion $\Delta(s; q_{\rm rel}, x) = -\xi_w - \zeta(s; q_{\rm rel}, x)$ für zahlreiche Punkte der Radoberfläche erforderlich (vgl. z. B. [98], [102]). Im quasi-elastischen Kontaktmodell mit (4.25) werden diese Funktionswerte von ζ gleichzeitig zur Definition einer hinreichend oft differenzierbaren Zwangsbedingung verwendet.

Für das quasi-elastische Kontaktmodell basiert die Berechnung der Reibungskräfte auf

dem verallgemeinerten Hertzschen Kontaktmodell aus Abschnitt 4.3, wobei die Kurvenintegrale zur Berechnung der gewichteten Mittelwerte $\tilde{s}$, $\kappa_x^{(\mathrm{W})}$, $\kappa_y^{(\mathrm{W})}$, $\kappa_x^{(\mathrm{R})}$, $\kappa_y^{(\mathrm{R})}$ analog zu (4.25) diskretisiert werden.

Vergleicht man diese Implementierung des quasi-elastischen Kontaktmodells mit dem traditionell verwendeten Starrkörperkontaktmodell, so ergibt sich für Räder mit Verschleißprofil eine Regularisierung der Bewegungsgleichungen, die *während* der dynamischen Simulation vorgenommen wird (*on-line-Regularisierung*). Statt dessen ist es zur Verringerung des numerischen Aufwands für die dynamische Simulation vorteilhaft, während der dynamischen Simulation eine Approximation $\sigma(q_{\mathrm{rel}}, x)$ von $\mathrm{smax}_s^{(\nu,h)}\zeta(s; q_{\mathrm{rel}}, x)$ zu verwenden, die schneller ausgewertet werden kann. Bei der Auswahl und der Bestimmung von σ sind viele Details zu beachten, die die Besonderheiten der Geometrie der Oberflächen von Rad und Schiene berücksichtigen. In [26] werden die Einzelheiten der Lösungsansätze und der effizienten Berechnung einer Approximation des quasi-elastischen Kontaktmodells ausführlich diskutiert.

Aus Sicht der Zeitintegration — die hier im Vordergrund stehen soll — ist wesentlich,

- daß der Approximationsfehler $\sigma(q_{\mathrm{rel}}, x) - \mathrm{smax}_s^{(\nu,h)}\zeta(s; q_{\mathrm{rel}}, x)$ hinreichend klein ist,
- daß σ hinreichend oft stetig differenzierbar ist,
- daß σ zu gegebenen Profilen von Rad und Schiene zuverlässig und schnell berechnet werden kann und
- daß $\sigma(q_{\mathrm{rel}}, x)$ zu gegebenen q_{rel} und x (sehr) schnell ausgewertet werden kann.

Da ζ und damit auch $\mathrm{smax}_s^{(\nu,h)}\zeta$ von den drei relativen Lagekoordinaten ξ_v, φ und ψ abhängt (vgl. Bemerkung 36), ergibt sich im allgemeinen Fall ein vierdimensionales Approximationsproblem $\sigma = \sigma(\xi_v, \varphi, \psi, x)$. Praktisch brauchbare Approximationen sind deshalb nur in Spezialfällen möglich. Prinzipiell beschränkt man sich für die Approximation auf Rad–Schiene–Systeme, für die das Gleisprofil entlang des Fahrwegs unverändert bleibt. Dann ist die Profilfunktion $G(v; x)$ und damit auch ζ und $\mathrm{smax}_s^{(\nu,h)}\zeta$ von x unabhängig, d. h. $\sigma = \sigma(\xi_v, \varphi, \psi)$. Entscheidend ist die Abhängigkeit von ξ_v und φ, dagegen kann die Abhängigkeit von ψ in guter Näherung durch quadratische Interpolation beschrieben werden:

$$\sigma(\xi_v, \varphi, \psi) := \sigma(\xi_v, \varphi, 0) + (\psi/\psi_0)^2(\sigma(\xi_v, \varphi, \psi_0) - \sigma(\xi_v, \varphi, 0)) \tag{4.26}$$

(aus Symmetriegründen ist σ bezüglich ψ eine gerade Funktion). In (4.26) wird die Stützstelle ψ_0 geeignet gewählt, z. B. $\psi_0 = 3°$.

Für die Funktion $\sigma(\xi_v, \varphi, 0)$ verwenden wir einen auf einem Rechteckgitter definierten polynomialen 2D-Tensorproduktspline. Die Splinekoeffizienten werden so bestimmt, daß σ ein diskretes Analogon des Funktionals

$$\begin{aligned}&\int_{\underline{\xi}_v}^{\overline{\xi}_v}\int_{\underline{\varphi}}^{\overline{\varphi}} \omega(\xi_v, \varphi)\Big(\sigma(\xi_v, \varphi, 0) - \mathrm{smax}_s^{(\nu,h)}\zeta(\xi_v, \varphi, 0)\Big)^2 d\varphi d\xi_v + \\ &+ \int_{\underline{\xi}_v}^{\overline{\xi}_v}\int_{\underline{\varphi}}^{\overline{\varphi}} \Big(\mu_1^2(\frac{\partial^2\sigma}{\partial\xi_v^2}(\xi_v, \varphi, 0))^2 + 2\mu_1\mu_2(\frac{\partial^2\sigma}{\partial\xi_v\partial\varphi}(\xi_v, \varphi, 0))^2 + \mu_2^2(\frac{\partial^2\sigma}{\partial\varphi^2}(\xi_v, \varphi, 0))^2\Big)\, d\varphi d\xi_v\end{aligned} \tag{4.27}$$

minimiert. Dieses Funktional kombiniert eine gewichtete kleinste-Quadrate–Approximation von $\mathrm{smax}_s^{(\nu,h)}\zeta$ mit einer Verallgemeinerung des Thin–Plate–Funktionals zur Glättung des approximierenden Splines σ. Dabei sind $\omega(\xi_v, \varphi)$, μ_1 und μ_2 geeignete Gewichte. Wählt man eine B-Spline–Darstellung von $\sigma(\xi_v, \varphi, 0)$, so führt die Berechnung von σ auf ein sehr großes schwach besetztes überbestimmtes lineares Gleichungssystem. Hierfür wurden Lösungsalgorithmen, die die Besetztheitsstruktur der Koeffizientenmatrix ausnutzen, entwickelt und als FORTRAN–Programm WRSPL („**W**heel–**R**ail–**SPL**ine“) implementiert. Damit kann der approximierende Spline auch dann effizient und numerisch stabil berechnet werden, wenn der Splinegrad hoch ist und ein feines Gitter zur Definition des Tensorproduktsplines gewählt wurde. Durch Parallelisierung auf einem Workstation–Cluster läßt sich die Berechnung der Splinekoeffizienten zusätzlich beschleunigen. Für die Details des sequentiellen und des parallelisierten Lösungsalgorithmus sei auf [24] und [26] verwiesen.

Auf gleiche Weise bestimmt man $\sigma(\xi_v, \varphi, \psi_0)$ in (4.26). Analog zur Approximation der Kontaktbedingung werden auch die Eingabedaten des Reibgesetzes ($\tilde{s}$, $\kappa_x^{(\mathrm{W})}$, $\kappa_y^{(\mathrm{W})}$, $\kappa_x^{(\mathrm{R})}$, $\kappa_y^{(\mathrm{R})}$) durch derartige Linearkombinationen von approximierenden Tensorproduktsplines ersetzt. Die Berechnung der Splinekoeffizienten erfolgt vor Beginn der dynamischen Simulation (*off-line-Regularisierung*). Im MKS–Simulationspaket SIMPACK ist die Berechnung der Splinekoeffizienten Teil des Pre-Processings, die Koeffizienten werden in Tabellen abgespeichert und stehen während der Simulation zur Auswertung der Splinefunktionen zur Verfügung. Dabei ist die Berechnung dieser Tabellen nicht für jeden neuen Simulationslauf, sondern erst bei Änderung der Profilfunktionen $F(s)$ und $G(v)$ erforderlich, da σ allein durch das Rad- und das Gleisprofil bestimmt wird.

Beispiel 34 Approximiert man die in Abb. 4.16 auf S. 154 dargestellte Höhenauslenkung

$$-\xi_w = \mathrm{smax}_s^{(\nu,h)}\zeta(s; q_{\mathrm{rel}}, x)$$

eines Rades (Verschleißprofil S1002) über einem Gleis (Profil UIC60–ORE) durch einen Tensorproduktspline auf einem Gitter aus 20×10 Rechtecken, so fällt der approximierende Spline optisch mit der zu approximierenden Funktion, die in Abb. 4.16 durch „+“ markiert ist, zusammen.
Abb. 4.20 auf S. 166 zeigt die entsprechenden Ergebnisse für den Angriffspunkt der Zwangskraft, d. h. den Kontaktpunkt s_* des Starrkörperkontaktmodells („--“), den Angriffspunkt $\tilde{s}$ der Zwangskraft im quasi-elastischen Kontaktmodell („—“) und dessen Splineapproximation („···“). Mit den Markierungen „∘“ werden die Kontaktpunktsprünge im Starrkörperkontaktmodell angedeutet (vgl. auch Abb. 4.6). Im quasi-elastischen Kontaktmodell hängt sowohl der nach (4.24) bestimmte Punkt $\tilde{s}$ als auch der ihn approximierende Spline stetig von q_{rel} ab. Der relativ große Approximationsfehler bleibt noch in der Größenordnung des Modellfehlers des MKS–Modells.

Bemerkung 43 a) Die polynomiale Tensorproduktspline–Approximation erwies sich als besonders geeignet, weil das Approximationsproblem auf ein *lineares* Gleichungssystem zur Bestimmung der Splinekoeffizienten führt. Ein weiterer Vorteil gegenüber komplizierteren Ansatzfunktionen ist, daß im Simulationspaket nur die Tabellen der Splinekoeffizienten und keine komplizierteren oder übermäßig großen Datenmengen verwaltet werden müssen.

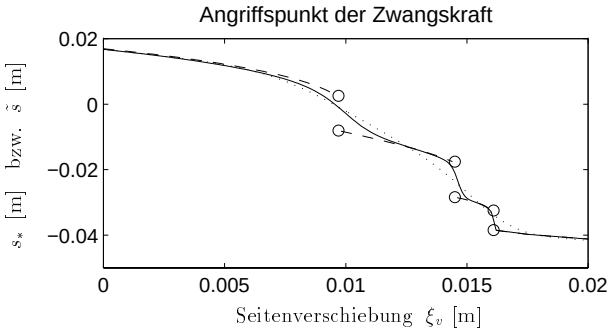


Abbildung 4.20: Kontaktpunkt s_* im Starrkörperkontaktmodell („--“) und Angriffspunkt $\tilde{s}$ der Zwangskraft im quasi-elastischen Kontaktmodell mit $\nu = 2 \cdot 10^{-5}$ („—“) sowie Splineapproximation („$\cdots$“). Gleisprofil UIC60-ORE, Radprofil S1002. $\varphi = 0.025$, $\psi = 0$.

b) Bekannte Nachteile von polynomialen Splines sind nicht nur die Unstetigkeiten in den höheren Ableitungen der Splinefunktion, sondern außerdem künstliche Oszillationen in den Splinefunktionen bei der Approximation von sehr stark veränderlichen oder unstetigen Funktionen. Zur Vermeidung solcher Oszillationen wurde in (4.27) der Glättungsterm mit positiven Gewichten μ_1 und μ_2 angefügt. Generell werden Splines hoher Ordnung verwendet (stückweise Polynome $(k+2)$-ten Grades für die Terme, die in (4.4) Teil der Zwangsbedingung $g(q) = 0$ sind, und stückweise Polynome k-ten Grades für die Eingabedaten des Reibgesetzes, die in (4.4) Teil der Kräfte $f(q, \dot{q}, \lambda, t)$ sind). Wie in vielen anderen technischen Anwendungen enthalten auch die Modellgleichungen für Rad–Schiene–Systeme Terme, die nur stetig, aber nicht differenzierbar sind. So wird z. B. in FASTSIM die Größe der Kontaktellipse durch lineare Interpolation zwischen tabellierten Daten bestimmt. Deshalb haben schon für $k \geq 2$ die Unstetigkeiten in den Ableitungen der Splinefunktionen nur geringen Einfluß auf die Schrittweiten- und Ordnungssteuerung des Integrators (vgl. hierzu auch die Tabellen 4.2 und 4.3).

c) Die Approximation von Kontaktbedingungen wird in verschiedenen Rad–Schiene–Simulationspaketen verwendet. Bei groben Genauigkeitsforderungen kann eine solche Approximation selbst dann zufriedenstellende Simulationsergebnisse liefern, wenn das Starrkörperkontaktmodell verwendet wird, denn bereits die Approximation der nur stückweise differenzierbaren Kontaktbedingung (4.9) durch eine glatte Funktion $\hat{\sigma}(q_{\mathrm{rel}}, x)$ wirkt als Regularisierung ([11]). Wegen des Modellfehlers des Starrkörperkontaktmodells bleibt jedoch der Gesamtfehler (= Modellfehler + Approximationsfehler) stets relativ groß, wenn die Räder ein Verschleißprofil haben. Außerdem bleibt diese Art der Regularisierung auf Anwendungen beschränkt, in denen ein solches $\hat{\sigma}$ angegeben werden kann. Vor allem der Fall fahrwegabhängiger Gleisprofile ist hierbei problematisch, weil $\hat{\sigma}$ dann von 4 Größen (ξ_v, φ, ψ, x) abhängt. Aus den genannten Gründen sind das quasi-elastische Kontaktmodell und seine Approximation diesem traditionellen Ansatz klar überlegen.

Die dynamische Simulation von Rad–Schiene–Systemen auf der Grundlage des quasi-elastischen Kontaktmodells

Sowohl für die einfachen Testbeispiele aus den Abschnitten 4.1, 4.2 und 4.3 als auch für die Anwendung von SIMPACK auf umfangreichere MKS–Modelle für Rad–Schiene–Systeme wurde die Approximation der Kontaktbedingungen und der Eingabeparameter des Reibgesetzes intensiv getestet. Ähnlich wie bei der Diskretisierung der Kurvenintegrale in (4.22) ist es auch bei der Approximation weder praktikabel noch sinnvoll, einen Approximationsfehler zu erreichen, der kleiner ist als die dem Integrator vorgegebene Fehlerschranke. Statt dessen werden Splinegrad und Splinegitter so gewählt, daß der Approximationsfehler klein bleibt gegenüber dem Modellfehler des MKS–Modells ([26]).

Beispiel 35 On-line- und off-line-Regularisierung sollen hier an Hand der Beispiele 31 und 32 verglichen werden. Dabei wird den Tensorproduktsplines ein Gitter mit 31 Knoten in ξ_v- und 21 Knoten in φ-Richtung vorgegeben. Die starken Änderungen der Daten in der Nähe der Singularitäten des Starrkörperkontaktmodells berücksichtigt man für Räder mit Verschleißprofil durch Vorgabe kleiner Gewichte $\omega(\xi_v, \varphi)$ und durch ein feineres Gitter in diesen Bereichen ([26]). Zur Approximation von $\mathrm{smax}_s^{(\nu,h)}$ werden bi-quartische Splines verwendet ($\mu_1 = \mu_2 = 10^{-8}$ in (4.27)), zur Approximation von $\tilde{s}$, $\kappa_x^{(\mathrm{W})}$, $\kappa_y^{(\mathrm{W})}$, $\kappa_x^{(\mathrm{R})}$ und $\kappa_y^{(\mathrm{R})}$ dagegen bi-quadratische Splines ($\mu_1 = \mu_2 = 10^{-6}$), im größten Teil des Definitionsbereichs von σ gilt dabei $\omega(\xi_v, \varphi) = 1$.
Mit diesen approximierenden Splines erhält man sowohl in Beispiel 31 als auch in Beispiel 32 nahezu identische Trajektorien für on-line- und off-line-Regularisierung. Dagegen ergaben sich drastische Unterschiede im numerischen Aufwand. Die Tabellen 4.2 und 4.3 zeigen die Ergebnisse für den Integrator ODASSL bei Anwendung auf die Gear–Gupta–Leimkuhler–Formulierung der Modellgleichungen. Dabei wurden für Beispiel 31, das wegen der Kegelräder schon mit dem Starrkörperkontaktmodell zufriedenstellend gelöst werden konnte, zum Vergleich auch die Resultate für die in Abschnitt 4.1 besprochene Implementierung des Starrkörperkontaktmodells angegeben. Sowohl die Zahl der Integrationsschritte (NSTEP) und der Funktionsaufrufe (NFUNC) als auch die Zahl der Schrittwiederholungen (NREP) bleibt beim Übergang zum quasi-elastischen Kontaktmodell nahezu unverändert, wobei die Unterschiede zwischen on-line- und off-line-Regularisierung vernachlässigbar sind. Die zur Simulation erforderliche Rechenzeit wird deshalb im wesentlichen vom numerischen Aufwand pro Funktionsaufruf bestimmt (Rechenzeitangaben in Tab. 4.2 und 4.3 für eine SUN Sparc5 Workstation). Für die off-line-Regularisierung (Splineapproximation) ergibt sich je nach Beispiel eine Rechenzeitersparnis zwischen 80% und 90%. Durch die Kombination des quasi-elastischen Kontaktmodells mit der Splineapproximation erreicht man nahezu die kleinen Rechenzeiten des Starrkörperkontaktmodells (Tab. 4.2).

Für MKS–Simulationspakete bietet das quasi-elastische Kontaktmodell daher eine sinnvolle Erweiterung des traditionell verwendeten Starrkörperkontaktmodells auf Rad–Schiene–Systeme, deren Räder Verschleißprofil haben. In SIMPACK wird dabei so weit wie möglich (d. h. für fahrwegunabhängige Gleisprofile) auf die off-line-Regularisierung, d. h. auf eine Splineapproximation der Kontaktbedingungen, zurückgegriffen. Wegen der enormen Rechenzeit bleibt die on-line-Regularisierung der Kontaktbedingungen auf Fahr-

Tabelle 4.2: Numerischer Aufwand in den Beispielen 27 und 31 (NSTEP: Zahl der Integrationsschritte, NFUNC: Zahl der Funktionsaufrufe, NREP: Zahl der Schrittwiederholungen). TOL = 10^{-5} für q, s, $\dot{q}$ und TOL = 10^{-2} für λ (vgl. auch Tab. 4.1).

	NSTEP	NFUNC	NREP	Rechenzeit
Starrkörperkontaktmodell	796	1794	75	12.3 s
on-line-Regularisierung	794	1811	81	141.0 s
off-line-Regularisierung	840	1864	69	24.7 s

Tabelle 4.3: Numerischer Aufwand in Beispiel 32 (NSTEP: Zahl der Integrationsschritte, NFUNC: Zahl der Funktionsaufrufe, NREP: Zahl der Schrittwiederholungen). TOL = 10^{-6} für q, $\dot{q}$ und TOL = 10^{-2} für λ.

	NSTEP	NFUNC	NREP	Rechenzeit
on-line-Regularisierung	2715	5780	71	621.3 s
off-line-Regularisierung	2671	5760	109	65.4 s

manöver beschränkt, für die sich das Gleisprofil mit dem Fahrweg ändert. In beiden Fällen steht zur Berechnung der Reibungskräfte nicht nur die Kombination des hier besprochenen verallgemeinerten Hertzschen Kontaktmodells mit dem Programm FASTSIM von Kalker zur Verfügung, sondern außerdem auch verschiedene andere Modelle, die teilweise aus der Literatur bekannt sind, z. T. aber auch neu entwickelt bzw. neu implementiert wurden ([97], [119], [142]).

Die Entwicklung und Implementierung des Modells für den Rad–Schiene–Kontakt ist ein wesentlicher Bestandteil der Entwicklung von Simulationssoftware für Rad–Schiene–Systeme. Daneben treten jedoch schon für einfache Testbeispiele (Radsätze, Drehgestelle) weitere Probleme auf, deren Modellierung und numerische Behandlung nicht trivial sind. Als Beispiele seien erwähnt:

1. Berechnung konsistenter Anfangswerte. Hierzu können die in Abschnitt 3.3.2 besprochenen Projektionen (3.86) und (3.87) verwendet werden (vgl. [144, Algorithmus 4.1] und speziell für Rad–Schiene–Systeme auch [120]). Simeon et al. entwickelten ein an die Struktur der MKS–Modellgleichungen (4.4) angepaßtes Homotopieverfahren, mit dem in (3.86) auch dann konsistente Lagekoordinaten q berechnet werden können, wenn kein ausreichend genauer Startwert $\tilde{q}$ zur Verfügung steht ([148, Abschnitte 5.2 und 6.2]).

2. Flankenanlauf. Kommt die Flanke eines Rades in Kontakt mit dem Gleis (z. B. bei der Einfahrt in einen Gleisbogen), so ändern sich $\dot{q}$ und die Zwangskräfte außerordentlich schnell, der numerische Aufwand zur Lösung der Modellgleichungen wächst sehr stark an (hierdurch entstehen z. B. die großen Rechenzeiten in Tab. 4.3).

3. Kurzzeitiges Abheben eines Rades. Hebt ein einzelnes Rad kurzzeitig vom Gleis ab, so sind die Auswirkungen auf die Laufdynamik des Rad–Schiene–Systems in praxi vernachlässigbar. Für die Simulation hat dieses Abheben dagegen weitreichende Konsequenzen, da die Zwangsbedingungen $g(q) = 0$ in (4.4) (kurzzeitig) nicht erfüllt sind (Änderung der Zahl der Freiheitsgrade, Unstetigkeiten in $\dot{q}$ und λ [103]).

Bei der Simulation realer Systeme sind zahlreiche weitere Probleme zu berücksichtigen, u. a. die Wechselwirkung des Schienenfahrzeugs mit dem bisher als starr angenommenen Fahrweg. Für die Details der Implementierung des quasi-elastischen Kontaktmodells in SIMPACK und für verschiedene erfolgreiche Anwendungen von SIMPACK zur dynamischen Simulation von Rad–Schiene–Systemen sei auf die Arbeiten von Netter, Rulka und anderen verwiesen ([142], [119]). Zu diesen Anwendungen zählt die Simulation verschiedenster Fahrmanöver (Geradenfahrt, Bogenlauf, Weichenfahrt, ...) für einzelne Waggons, für Wagen–Verbände und u. a. auch für Straßenbahnfahrzeuge. Besonders wichtig ist dabei die Verifikation der Simulationsergebnisse an Hand gemessener Daten.

Die Details der *industriellen* Anwendungen von SIMPACK im Schienenfahrzeugbau sind leider aus kommerziellen Gründen überwiegend nicht öffentlich zugänglich.

4.5 Zusammenfassung

Berücksichtigt man in einem mechanischen Mehrkörpersystem die geometrische Form der Körper, so treten in den Modellgleichungen zusätzlich zu den Zwangsbedingungen und Kräften, die klassischen Verbindungselementen (Gelenke, Federn, Dämpfer, ...) entsprechen, weitere Zwangsbedingungen, einseitige Beschränkungen und Kräfte auf. Im Starrkörperkontaktmodell entspricht dabei dem permanenten Kontakt zweier Körper eine skalare Zwangsbedingung an die Lagekoordinaten q. Sind die beiden Körper streng konvex, so berühren sie sich in genau einem Punkt, dem Kontaktpunkt. Die Lage des Kontaktpunkts ändert sich stetig mit der relativen Lage der Körper zueinander. Aus algorithmischen Gründen ist hier die differentiell-algebraische Formulierung (3.14) der Modellgleichungen besonders vorteilhaft, weil sie die effiziente Berechnung der Kontaktpunktkoordinaten s während der dynamischen Simulation ermöglicht.

Sind die sich berührenden Körper dagegen nicht konvex, so führt das Starrkörperkontaktmodell i. allg. auf nur stückweise differenzierbare Zwangsbedingungen, weil sich die Lage des Kontaktpunkts sprunghaft ändern kann. Den Kontaktpunktsprüngen entsprechen Singularitäten der Mannigfaltigkeit $\{\eta : g(\eta) = 0\}$, die im Fall autonomer Zwangsbedingungen $g(q) = 0$ durch den algebraischen Teil der Modellgleichungen definiert wird. Da die Lagekoordinaten q stets in der Zwangsmannigfaltigkeit verbleiben, liegt $v(t) = \dot{q}(t)$ im Tangentialraum an die Mannigfaltigkeit (wenn dieser existiert). Trajektorien $(q(t), v(t), \lambda(t))$, die einen Punkt erreichen, in dem die Zwangsmannigfaltigkeit nicht differenzierbar ist, können deshalb i. allg. nicht stetig über die Singularität der Mannigfaltigkeit hinaus fortgesetzt werden. Als Folge des Starrkörperkontaktmodells ergeben sich also unstetige Zustandsänderungen im Mehrkörpersystem, die sich in den Modellgleichungen als Singularitäten (bei Modellierung in Zustandsform) bzw. als Impasse points (bei Modellierung in Deskriptorform) widerspiegeln.

Hat die Lösung der Modellgleichungen endlich viele solche Unstetigkeitsstellen und sind Übergangsbedingungen für die Zustandsänderung in den Unstetigkeitsstellen gegeben, so kombiniert man bei der numerischen Lösung der Modellgleichungen ein vorhandenes Integrationsverfahren für differentiell-algebraische Systeme mit der numerischen Be-

rechnung der Zustandsänderung in den Unstetigkeitsstellen. Die Zeitpunkte, zu denen sich die Lage eines Kontaktpunkts sprunghaft ändert, d. h. die möglichen Unstetigkeitsstellen der Lösung, werden dabei unter Verwendung von geeigneten Schaltfunktionen automatisch während der Integration bestimmt. Die auf diese Weise berechneten Simulationsergebnisse zeigten für die dynamische Simulation von Rad–Schiene–Systemen, die in diesem Kapitel im Vordergrund stand, daß der Modellfehler des Starrkörperkontaktmodells inakzeptabel groß ist, wenn man den Kontakt zwischen der Lauffläche eines Rades mit Verschleißprofil und der Schiene betrachtet.

Als Alternative zum Starrkörperkontaktmodell, das den Abstand der beiden Körper nur in einem einzelnen Punkt (dem Kontaktpunkt) berücksichtigt, wurde deshalb ein Modell für den Rad–Schiene–Kontakt entwickelt, das ähnlich wie ein elastisches Kontaktmodell den Abstand zwischen Rad und Schiene über der gesamten Oberfläche der beiden Körper betrachtet. Wie zuvor im Starrkörperkontaktmodell wird auch in diesem sog. quasi-elastischen Kontaktmodell der permanente geometrische Kontakt zwischen zwei Körpern des Mehrkörpersystems durch eine skalare Zwangsbedingung beschrieben. Der Übergang vom Starrkörperkontaktmodell zum quasi-elastischen Modell verringert den Modellfehler des MKS–Modells erheblich und bewirkt eine Regularisierung der Bewegungsgleichungen, so daß für die dynamische Simulation des Rad–Schiene–Systems Standard–Integrationsverfahren verwendet werden können.

Es wurden 2 Algorithmen zur effizienten Auswertung der Kontaktbedingung des quasi-elastischen Kontaktmodells entwickelt und implementiert (on-line– und off-line–Regularisierung). Damit erreicht das quasi-elastische Modell einen sehr großen Einsatzbereich (u. a. Einzelräder, fahrwegabhängige Gleisprofile) und läßt sich trotzdem für klassische Simulationsaufgaben (z. B. fahrwegunabhängige Gleisprofile) fast ebenso schnell wie die Starrkörperkontaktbedingung auswerten. Einige Simulationsbeispiele illustrieren den Nutzen des vorgeschlagenen quasi-elastischen Kontaktmodells.

Das quasi-elastische Kontaktmodell und die beiden Varianten seiner Implementierung genügen typischen Anforderungen eines Simulationspakets hinsichtlich des Modellfehlers, der Breite des Anwendungsbereichs, der Robustheit der numerischen Algorithmen und vor allem auch hinsichtlich der zur Simulation erforderlichen Rechenzeit. Als Bestandteil des kommerziellen Simulationspakets SIMPACK wird das quasi-elastische Kontaktmodell bei der Simulation von Rad–Schiene–Systemen in umfangreichen und komplizierten industriellen Anwendungen eingesetzt.

So wie das Starrkörperkontaktmodell kann auch das quasi-elastische Modell für den Kontakt beliebiger Körper im Mehrkörpersystem formuliert werden. Es ist sehr stark von der konkreten Anwendung abhängig, ob die damit verbundene Regularisierung der Bewegungsgleichungen den Modellfehler verringert ohne den zur Simulation erforderlichen Rechenaufwand zu stark anwachsen zu lassen. Ebenso läßt sich nicht allgemein, sondern nur am konkreten mechanischen Beispiel entscheiden, ob der Modellfehler des quasi-elastischen Modells hinreichend klein ist und ob die quasi-elastische Kontaktbedingung effizient numerisch ausgewertet werden kann. Neben dem hier betrachteten Rad–Schiene–Kontakt zählt zum potentiellen Anwendungsbereich des quasi-elastischen Kontaktmodells der Kontakt von beliebigen Körpern, deren undeformierte Oberflächen nahezu parallel sind und auf die eine betragsmäßig große Kraft in Normalenrichtung zu ihrer Kontaktfläche wirkt.

Zusammenfassung

Die Modellierung von angewandten Problemen aus Naturwissenschaft und Technik führt häufig in natürlicher Weise auf differentiell-algebraische Systeme. Obwohl sich die analytischen Eigenschaften der differentiell-algebraischen Systeme von höherem Index grundlegend von den Eigenschaften gewöhnlicher Differentialgleichungen unterscheiden, können Anfangswertprobleme (und auch Randwertprobleme) für differentiell-algebraische Systeme mit numerischen Verfahren, die aus der Theorie der gewöhnlichen Differentialgleichungen bekannt sind, unter gewissen Voraussetzungen zufriedenstellend gelöst werden.

Hierzu wurde für nichtlineare Systeme vom Index 2 und 3 im Detail untersucht, welche Auswirkungen die Verstärkung kleiner Fehler (Rundungsfehler usw.) auf die numerische Integration hat. Die dabei nachgewiesenen zusätzlichen Fehlerterme sind für semi-explizite Index-2-Systeme so klein, daß die direkte Anwendung von Diskretisierungsverfahren auf diese Systeme zu robusten und effizienten numerischen Lösungsverfahren führt.

Unter Ausnutzung der speziellen Struktur von differentiell-algebraischen Systemen in Hessenbergform wurden partitionierte Verfahren für semi-explizite Index-2-Systeme konstruiert, die vor allem für nicht-steife Systeme geeignet sind. Kombiniert man ein explizites Runge–Kutta–Verfahren oder ein implizites Adams–Verfahren mit geeigneten Verfahren zur Bestimmung der algebraischen Lösungskomponenten, so erhält man stabile partitionierte Verfahren für semi-explizite Index-2-Systeme. Für die differentiellen Lösungskomponenten wird dabei die klassische Konvergenzordnung des Runge–Kutta– bzw. des Adams–Verfahrens erreicht.

Ein Starrkörperkontaktmodell für mechanische Mehrkörpersysteme führt auf differentiell-algebraische Systeme, deren algebraischer Teil aus den Kontaktbedingungen und aus nichtlinearen Gleichungen zur Bestimmung der Kontaktpunktkoordinaten besteht. Es wurde gezeigt, daß das Starrkörperkontaktmodell zu Singularitäten in den Modellgleichungen führen kann. Der Übergang zu einem quasi-elastischen Kontaktmodell bewirkt eine Regularisierung der Bewegungsgleichungen. Diese regularisierten Bewegungsgleichungen kann man mit Standardverfahren für differentiell-algebraische Systeme effizient numerisch lösen.

In allen Kapiteln der vorliegenden Arbeit werden die Modellgleichungen für mechanische Mehrkörpersysteme als Beispiel für differentiell-algebraische Systeme von höherem Index herangezogen. Detaillierte analytische Untersuchungen erklären einige der Unterschiede zwischen den verschiedenen analytisch äquivalenten Formulierungen der Modellgleichungen und sind gleichzeitig Grundlage der Konstruktion und Implementierung von Diskretisierungsverfahren. Die Verfahrensfunktion und die Koeffizienten der partitionierten Verfahren wurden anschließend so bestimmt, daß man mit möglichst geringem numerischem Aufwand eine hohe Konvergenzordnung erreicht.

Für Rad–Schiene–Systeme führt neben dem engen Zusammenspiel zwischen der Modellierung und den Zeitintegrationsverfahren erst die speziell angepaßte Diskretisierung von quasi-elastischen Kontaktbedingungen zu Software, die im Rahmen des Simulationspakets SIMPACK die effiziente dynamische Simulation in Anwendungen des Schienenfahrzeugbaus ermöglicht.

Anhang A

Aufruf des Integrators HEDOP5

```
C
C-----LIBRARY MBSPACK
C-----GROUP 1 STRUCTURAL INTERFACE
C-----CODE HEDOP5
C
C     HEDOP5 -- Multibody system integration with a
C     -------------------------------------------
C               Half Explicit RK method of 5th order
C               -----------------------------------
C  ** based on the implementation of HEM5
C  ** written by Bernd Simeon, Technical University of Munich, Feb.17/93 **
C  ** extended version for systems with friction               Feb.01/94 **
C  **
C  ** modified by Martin Arnold, University of Rostock
C  ** half-explicit Runge-Kutta method with explicit stage     Oct.16/95 **
C  ** extended version for systems with friction               Mar.08/96 **
C  ** extended version (with dense output)                     Apr.09/96 **
C
C     Subroutine HEDOP5 solves systems of differential-algebraic equations
C     (DAEs) arising in multibody system dynamics.
C
C     DOCUMENTATION:  hedop5.man
C
C     PARAMETERS - short description:
C
C       FMBS   This is a subroutine which you provide to define the
C              differential/algebraic system. FMBS must satisfy the
C              interface requirements defined in hedop5.man manual
C
C       NDIM(9) This array defines the dimensions and the structure of
C              the multibody system equations. NDIM(1) = NEQ is the
C              number of equations to be solved.
C
C       T      This is the current value of the independent variable.
C
C       X(*)   This array contains the solution components at T.
C
C       TOUT   This is a point at which a solution is desired.
C
C       INFO(20) The basic task of the code is to solve the system from T
C              to TOUT and return an answer at TOUT.  INFO is an integer
C              array which is used to communicate exactly how you want
C              this task to be carried out.
C              (See hedop5.man and below for details.)
```



```
C
C       RTOL,ATOL These quantities represent relative and absolute
C              error tolerances which you provide to indicate how
C              accurately you wish the solution to be computed.  You
C              may choose them to be both scalars or else both vectors.
C
C       DEX(20),IEX(30) These arrays provide conditional inputs
C              to perform the tasks specified by the INFO vector and report
C              also the performance of the code on return.
C              (See hedop5.man and below for details.)
C
C       IDID   This scalar quantity is an indicator reporting what the
C              code did.  You must monitor this integer variable to
C              decide  what action to take next.
C
C       RWORK  A real work array of length LRW which provides the
C              code with needed storage space.
C
C       LRW    The length of RWORK.
C              You must have
C                   LRW .GE. 11 + 27*NN + 8*NV*NV
C              where NN = NDIM(1), NV=NDIM(2)+NDIM(4).
C
C       IWORK  An integer work array of length LIW which provides the
C              code with needed storage space.
C
C       LIW    The length of IWORK.
C              You must have
C                    LIW .GE. 10+NN+NV+NDIM(7).
C
C       RPAR,IPAR These are real and integer parameter arrays which
C              you can use for communication between your calling
C              program and the FMBS subroutine.
C
C       *      This is the target label in case an error occurs such
C              that the solution of the problem cannot be continued
C              ( IDID .le. -11 ).
C
C     Quantities which may be altered by HEDOP5 are:
C
C     T, X(*), INFO(1), DEX(*), IEX(*),
C     IDID, RWORK(*), IWORK(*)
C
C     The integration process is controlled by the information provided in
C     INFO, DEX, IEX. The following table summarizes the features.
C
C     INFO   0   /  1                          to be supplied if INFO = 1:
C     ------------------------------------------------------------------
C     (1)  First / subsequent call.
C     (2)  Tolerances as scalars / arrays.      RTOL(*), ATOL(*)
C     (3)  Intervall / intermediate output.
C     (4)  No / yes dense output.               DEX(15)= dense interval
C          ( no interpolation! )                IEX(3) = output channel
C                                               IWORK(11:10+NN))
C                                                      = plot indices
C     (5)  No / yes select projection.          IEX(4) = 0: automatic
C                                                      = 1: pos./veloc.
C                                                      = 2,3: no proj.
C     (7)  No / yes supply max. stepsize.       DEX(2) = h_max
C     (8)  No / yes supply initial stepsize.    DEX(3) = h_0
```

```
C     (9)  No / yes select linear algebra meth. IEX(24)= 3: Linpack DGEFA (def.)
C                                                      = 4: Lapack DGETRF.
C     (11) No / yes compute consistent i.v.
C     (13) No / yes friction is part of forces. IEX(23)= 1: df/dlambda supplied
C                                                         in FL by call FMBS
C                                                      = 2: numerical differ.
C                                                         for df/dlambda
C     (14) No / yes define max. # steps.        IEX(25)= max. # steps
C     (18) No / yes control stepsize strategy.  DEX(11)= safety factor
C                                               DEX(12)= lower bound change
C                                               DEX(13)= upper bound change
C                                               DEX(14)= beta-stabilization
C
C     The performance of the code is reported via DEX, IEX:
C
C       DEX
C       (4)  The last used successful stepsize.
C       (5)  The new stepsize.
C       (9)  The last error estimation.
C       IEX
C       (7)  The number of FMBS evaluations so far.
C       (9)  The number of projection steps so far.
C       (10) The number of matrix decompositions so far.
C       (11) The number of steps taken so far.
C       (13) The number of error test failures so far.
C       (14) The number of convergence failures in case of projection.
```

Literaturverzeichnis

[1] T. Alishenas. *Zur numerischen Behandlung, Stabilisierung durch Projektion und Modellierung mechanischer Systeme mit Nebenbedingungen.* Dissertation, Königliche Technische Hochschule, Stockholm, 1992.

[2] T. Alishenas, Ö. Ólafsson. Modeling and velocity stabilization of constrained mechanical systems. *BIT*, 34:455–483, 1994.

[3] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, D. Sorensen. *LAPACK User's Guide.* SIAM, Philadelphia, 2. Auflage, 1992.

[4] T. Andrzejewski, H.G. Bock, E. Eich, R. von Schwerin. Recent advances in the numerical integration of multibody systems. In W. Schiehlen, editor, *Advanced Multibody System Dynamics – Simulation and Software Tools*, S. 127–151. Kluwer Academic Publishers, Dordrecht, NL, 1993.

[5] M. Anitescu, J.F. Cremer, F.A. Potra. Formulating 3D contact dynamics problems. *Mechanics of Structures and Machines*, 24:405–437, 1996.

[6] C. Arévalo. Optimized β-blocking for nonlinear semiexplicit index 2 DAEs. Vortrag zur Tagung „DAEs, Related Fields and Applications“ (gemeinsam mit C. Führer und G. Söderlind), Mathematisches Forschungsinstitut Oberwolfach, November 1995.

[7] C. Arévalo, C. Führer, G. Söderlind. Stabilized multistep methods for index 2 Euler–Lagrange DAEs. *BIT*, 36:1–13, 1996.

[8] C. Arévalo, G. Söderlind. Convergence of multistep discretizations of DAEs. *BIT*, 35:143–168, 1995.

[9] M. Arnold. *Numerische Behandlung von semi–expliziten Algebrodifferentialgleichungen vom Index 1 mit linear–impliziten Verfahren.* Dissertation, Martin–Luther–Universität Halle–Wittenberg, Sektion Mathematik, 1990.

[10] M. Arnold. Stability of numerical methods for differential–algebraic equations of higher index. *Applied Numerical Mathematics*, 13:5–14, 1993.

[11] M. Arnold. The geometry of wheel–rail contact. In K. Frischmuth, editor, *The dynamical simulation of wheel–rail systems, Proc. of the First Workshop on "Dynamics of Wheel–Rail–Systems", held at Rostock University (May, 5th and 6th, 1994)*, Universität Rostock, FB Mathematik, Preprint 94/21, November 1994.

[12] M. Arnold. *Implicit Runge–Kutta methods for transferable differential-algebraic equations*, volume 29 of *Banach Center Publications*, S. 267–274. Polish Academy of Sciences, 1994.

[13] M. Arnold. Index and stability of differential-algebraic systems. In U. Helmke, R. Mennicken, J. Saurer, editors, *Systems and Networks: Mathematical Theory and Applications II*, volume 79 of *Mathematical Research*, S. 41–44, Berlin, 1994. Akademie–Verlag.

[14] M. Arnold. Perturbation analysis for differential-algebraic equations of index 2. In W.F. Ames, editor, *Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics*, S. 20–23, Atlanta, 1994. Georgia Institute of Technology.

[15] M. Arnold. Applying BDF to quasilinear differential-algebraic equations of index 2: a perturbation analysis. Preprint 95/13, Universität Rostock, FB Mathematik, 1995.

[16] M. Arnold. Numerische Probleme in der dynamischen Simulation von Rad–Schiene–Systemen. *Z. Angew. Math. Mech.*, 75:677–678, 1995.

[17] M. Arnold. A perturbation analysis for the dynamical simulation of mechanical multibody systems. *Applied Numerical Mathematics*, 18:37–56, 1995.

[18] M. Arnold. Half-explicit Runge–Kutta methods with explicit stages for differential-algebraic systems of index 2. Preprint 95/19, Universität Rostock, FB Mathematik, 1995.

[19] M. Arnold. A note on the uniform perturbation index. Zur Veröffentlichung eingereicht bei *Rostocker Math. Kolloq.*, 1996.

[20] M. Arnold. Numerical problems in the dynamical simulation of wheel-rail systems. *Z. Angew. Math. Mech.*, Proceedings of ICIAM 95, Issue 3:151–154, 1996.

[21] M. Arnold. Half-explicit Runge–Kutta methods with explicit stages for differential-algebraic systems of index 2. Erscheint in: *BIT*, vol. 38, 1998.

[22] M. Arnold, K. Frischmuth. Solving problems with unilateral constraints by DAE methods. Zur Veröffentlichung eingereicht bei *Mathematics and Computers in Simulation*, 1997.

[23] M. Arnold, H. Netter. Ein modifizierter Korrektor für die stabilisierte Integration differential-algebraischer Systeme mit von Hessenbergform abweichender Struktur. Interner Bericht IB 515–93–03, DLR, D-5000 Köln 90, 1993.

[24] M. Arnold, H. Netter. The approximation of contact conditions in the dynamical simulation of wheel-rail systems. Interner Bericht IB 515–96–08, Institut für Robotik und Systemdynamik, DLR Oberpfaffenhofen, 1996.

[25] M. Arnold, H. Netter. Wear profiles and the dynamical simulation of wheel-rail systems. In M. Brøns, M.P. Bendsøe, M.P. Sørensen, editors, *Progress in Industrial Mathematics at ECMI 96*, S. 77–84. Teubner, Stuttgart, 1997.

[26] M. Arnold, H. Netter. Approximation of contact geometry in the dynamical simulation of wheel-rail systems. Erscheint in *Math. Modelling of Systems*, 1998.

[27] M. Arnold, K. Strehmel, R. Weiner. Half-explicit Runge–Kutta methods for semi-explicit differential-algebraic equations of index 1. *Numer. Math.*, 64:409–431, 1993.

[28] M. Arnold, K. Strehmel, R. Weiner. Errors in the numerical solution of nonlinear differential-algebraic systems of index 2. Reports on Computer Science and Scientific Computing 11(1995), Martin–Luther–Universität Halle–Wittenberg, FB Mathematik und Informatik, 1995.

[29] U.M. Ascher, H. Chin, S. Reich. Stabilization of DAEs and invariant manifolds. *Numer. Math.*, 67:131–149, 1994.

[30] U.M. Ascher, P. Lin. Sequential regularization methods for higher index DAEs with constraint singularities. The linear index-2 case. *SIAM J. Numer. Anal.*, 33(5):1921–1940, 1996.

[31] U.M. Ascher, L.R. Petzold. Projected implicit Runge–Kutta methods for differential–algebraic equations. *SIAM J. Numer. Anal.*, 28:1097–1120, 1991.

[32] U.M. Ascher, L.R. Petzold. Stability of computational methods for constrained dynamics systems. *SIAM J. Sci. Comput.*, 14:95–120, 1993.

[33] J. Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 1:1–16, 1972.

[34] C. Bischof, A. Carle, G. Corliss, A. Griewank, P. Hovland. ADIFOR – Generating derivative codes from Fortran programs. *Scientific Programming*, 1:11–29, 1992.

[35] H.G. Bock, R. von Schwerin. An inverse dynamics Adams method for constrained multibody systems. Preprint 93–27, IWR, Universität Heidelberg, 1993.

[36] V. Brasey. A half-explicit method of order 5 for solving constrained mechanical systems. *Computing*, 48:191–201, 1992.

[37] V. Brasey. *Half-Explicit Methods for Semi-Explicit Differential-Algebraic Equations of Index 2.* Dissertation, Université Genève, Départment de Mathématiques, 1994.

[38] V. Brasey, E. Hairer. Half-explicit Runge–Kutta methods for differential-algebraic systems of index 2. *SIAM J. Numer. Anal.*, 30:538–552, 1993.

[39] K.E. Brenan, S.L. Campbell, L.R. Petzold. *Numerical solution of initial–value problems in differential–algebraic equations*. SIAM, Philadelphia, 2. Auflage, 1996.

[40] K.E. Brenan, B.E. Engquist. Backward differentiation approximations of nonlinear differential/algebraic equations. *Math. Comp.*, vol. 51:659–676 und Supplement:S7–S16, 1988.

[41] P.N. Brown, A.C. Hindmarsh, L.R. Petzold. Consistent initial condition calculation for differential-algebraic systems. Technical report, Lawrence Livermore National Laboratory, August 1995, erscheint in *SIAM J. Sci. Comp.*.

[42] A. Budó. *Theoretische Mechanik*. Deutscher Verlag der Wissenschaften, Berlin, 9. Auflage, 1978.

[43] J.C. Butcher. *The numerical analysis of ordinary differential equations: Runge–Kutta and general linear methods*. Wiley&Sons, Chichester New York et al., 1987.

[44] S.L. Campbell, C.W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

[45] S.L. Campbell, W. Marszalek. ODE/DAE integrators and MOL problems. *Z. Angew. Math. Mech.*, Proceedings of ICIAM 95, Issue 1:251–254, 1996.

[46] S.L. Campbell, W. Marszalek. The index of an infinite dimensional implicit system. Technical Report 96/1996, North Carolina State University Raleigh, Department of Mathematics, Mai 1996, erscheint in *Mathematical Modelling of Systems*.

[47] S.L. Campbell, E. Moore. Constraint preserving integrators for general nonlinear higher index DAEs. *Numer. Math.*, 69:383–399, 1995.

[48] K.D. Clark. A structural form for higher index semistate equations I: Theory and applications to circuit and control. *Linear Algebra Appl.*, 98:169–197, 1988.

[49] J.E. Dennis, R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Englewood Cliffs, N.J., 1983.

[50] P. Deuflhard, F. Bornemann. *Numerische Mathematik II*. Walter de Gruyter, Berlin New York, 1994.

[51] P. Deuflhard, E. Hairer, J. Zugck. One–step and extrapolation methods for differential–algebraic systems. *Numer. Math.*, 51:501–516, 1987.

[52] J. Dongarra, J.R. Bunch, C.B. Moler, G.W. Stewart. *LINPACK Users Guide*. SIAM, Philadelphia, 1979.

[53] W. Duffek. Das räumliche Kontaktproblem bei starrem Radsatz und starrem Gleis. Interner Bericht IB 515–80–05, DLR, D-5000 Köln 90, 1980.

[54] E. Eich. *Projizierende Mehrschrittverfahren zur numerischen Lösung der Bewegungsgleichungen technischer Mehrkörpersysteme mit Zwangsbedingungen und Unstetigkeiten*. Fortschritt-Berichte VDI Reihe 18, Nr. 109. VDI–Verlag, Düsseldorf, 1992.

[55] E. Eich. Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints. *SIAM J. Numer. Anal.*, 30:1467–1482, 1993.

[56] E. Eich, C. Führer. Numerische Methoden in der Mehrkörperdynamik. In A. Bachem, M. Jünger, R. Schrader, editors, *Mathematik in der Praxis – Fallstudien aus Industrie, Wirtschaft, Naturwissenschaften und Medizin*, S. 41–60. Springer–Verlag, Berlin Heidelberg New York, 1995.

[57] E. Eich, C. Führer. *Numerical Methods in Multibody Dynamics*. Teubner – John Wiley, 1998, in Vorbereitung.

[58] E. Eich, C. Führer, B. Leimkuhler, S. Reich. Stabilization and projection methods for multibody dynamics. Technical Report A281, Helsinki University of Technology, Institute of Mathematics, 1990.

[59] E. Eich, C. Führer, J. Yen. On the error control for multistep methods applied to ODEs with invariants and DAEs in multibody dynamics. *Mechanics of Structures and Machines*, 23:159–180, 1995.

[60] E. Eich, M. Hanke. Regularization methods for constrained mechanical multibody systems. *Z. Angew. Math. Mech.*, 75:761–773, 1995.

[61] R. Frank, J. Schneid, C.W. Überhuber. The concept of B–convergence. *SIAM J. Numer. Anal.*, 18:753–780, 1981.

[62] K. Frischmuth. On the numerical solution of rail-wheel contact problems. *J. Theoretical and Applied Mechanics*, 34:1–15, 1996.

[63] K. Frischmuth. Regularisierungsmethoden für nichtstetige Funktionen. Vortrag TU Berlin, Institut für Luft- und Raumfahrt, Januar 1996.

[64] K. Frischmuth. Regularization methods for non-smooth dynamical problems. In R. Bogacz, G.-P. Ostermeyer, K. Popp, editors, *Dynamical Problems in Mechanical Systems, Proc. of the 4th German–Polish Workshop held in Berlin (July, 30th – August, 5th, 1995)*, Polish Academy of Sciences, Warsaw, 1996.

[65] K. Frischmuth, M. Arnold, M. Hänler, H. Netter. Differentialgleichungen und singuläre Mannigfaltigkeiten in der dynamischen Simulation von Rad–Schiene–Systemen. In K.-H. Hoffmann, W. Jäger, Th. Lohmann, H. Schunck, editors, *Mathematik – Schlüsseltechnologie für die Zukunft*, S. 331–342. Springer-Verlag, Berlin Heidelberg New York, 1997.

[66] C. Führer. Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen. Theorie, numerische Ansätze und Anwendungen. Dissertation, TU München, Mathematisches Institut und Institut für Informatik, 1988.

[67] C. Führer. Persönliche Mitteilung, 1996.

[68] C. Führer, B. Leimkuhler. Numerical solution of differential–algebraic equations for constrained mechanical motion. *Numer. Math.*, 59:55–69, 1991.

[69] F.R. Gantmacher. *Matrizentheorie II.* Deutscher Verlag der Wissenschaften, Berlin, 1959.

[70] C.W. Gear. The simultaneous numerical solution of differential-algebraic equations. *IEEE Trans. Circuit Theory*, CT-18:89–95, 1971.

[71] C.W. Gear. Maintaining solution invariants in the numerical solution of ODEs. *SIAM J. Sci. Stat. Comput.*, 7:734–743, 1986.

[72] C.W. Gear. Differential-algebraic equation index transformations. *SIAM J. Sci. Stat. Comput.*, 9:39–47, 1988.

[73] C.W. Gear. Differential-algebraic equations, indices, and integral algebraic equations. *SIAM J. Numer. Anal.*, 27:1527–1534, 1990.

[74] C.W. Gear, B. Leimkuhler, G.K. Gupta. Automatic integration of Euler–Lagrange equations with constraints. *J. Comp. Appl. Math.*, 12&13:77–90, 1985.

[75] C.W. Gear, L.R. Petzold. ODE methods for the solution of differential/algebraic systems. *SIAM J. Numer. Anal.*, 21:716–728, 1984.

[76] E. Griepentrog, R. März. *Differential–Algebraic Equations and Their Numerical Treatment.* Teubner Texte zur Mathematik. Teubner–Verlag, Leipzig, 1986.

[77] M. Günther. *Ladungsorientierte Rosenbrock–Wanner–Methoden zur numerischen Simulation digitaler Schaltungen.* Fortschritt-Berichte VDI Reihe 20, Nr. 168. VDI–Verlag, Düsseldorf, 1995.

[78] K. Gustafsson, M. Lundh, G. Söderlind. A PI stepsize control for the numerical solution of ordinary differential equations. *BIT*, 28:270–287, 1988.

[79] E. Hairer, Ch. Lubich, M. Roche. Error of Runge–Kutta methods for stiff problems studied via differential algebraic equations. *BIT*, 28:678–700, 1988.

[80] E. Hairer, Ch. Lubich, M. Roche. Error of Rosenbrock methods for stiff problems studied via differential algebraic equations. *BIT*, 29:77–90, 1989.

[81] E. Hairer, Ch. Lubich, M. Roche. *The numerical solution of differential-algebraic systems by Runge–Kutta methods.* Lecture Notes in Mathematics, 1409. Springer-Verlag, Berlin Heidelberg New York, 1989.

[82] E. Hairer, S.P. Nørsett, G. Wanner. *Solving Ordinary Differential Equations. I. Nonstiff Problems.* Springer-Verlag, Berlin Heidelberg New York, 2. Auflage, 1993.

[83] E. Hairer, G. Wanner. RADAU5 – an implicit Runge–Kutta code. Technical report, Université Genève, Département de Mathématiques, 1988.

[84] E. Hairer, G. Wanner. *Solving Ordinary Differential Equations. II. Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin Heidelberg New York, 2. Auflage, 1996.

[85] M. Hanke. Regularization methods for higher index differential-algebraic equations. In E. Griepentrog, M. Hanke, R. März, editors, *Toward a better understanding of differential-algebraic equations*, S. 105–141. Humboldt–Universität Berlin, FB Mathematik, Seminarbericht 92–1, 1992.

[86] M. Hanke, E.I. Macana. Implicit Runge–Kutta methods for general linear index 2 differential-algebraic equations with variable coefficients. Preprint 93–11, Humboldt–Universität Berlin, FB Mathematik, 1993.

[87] M. Hanke, E.I. Macana, R. März. On asymptotics in case of linear index-2-DAEs. Preprint 94–5, Humboldt–Universität Berlin, FB Mathematik, 1994.

[88] M. Hänler. A mass point moving on a non-smooth manifold in $\mathbb{R}^n$. *J. Theoretical and Applied Mechanics*, 34:17–29, 1996.

[89] E.J. Haug, S.C. Wu, S.M. Yang. Dynamics of mechanical systems with Coulomb friction, stiction, impact and constraint addition–deletion I. *Mechanism and Machine Theory*, 21:401–406, 1986.

[90] I. Higueras Sanz. *Metodos Runge–Kutta Explicitos para la Integracion Numerica de Ecuaciones Differenciales Algebraicas.* Dissertation, Universidad de Zaragoza, Departamento de Matematica Aplicada, 1991.

[91] M. Hiller, S. Frik. Road vehicle benchmark 2: five link suspension. In W. Kortüm, S. Sharp, A. de Pater, editors, *Application of Multibody Computer Codes to Vehicle System Dynamics. Progress Report to the 12th IAVSD Symposium, Lyon*, 1991.

[92] L.O. Jay. Convergence of a class of Runge–Kutta methods for differential-algebraic systems of index 2. *BIT*, 33:137–150, 1993.

[93] L.O. Jay. *Runge–Kutta Type Methods for Index Three Differential-Algebraic Equations with Applications to Hamiltonian Systems.* Dissertation, Université Genève, Départment de Mathématiques, 1994.

[94] L.O. Jay. Convergence of Runge–Kutta methods for differential-algebraic systems of index 3. *Applied Numerical Mathematics*, 17:97–118, 1995.

[95] K.L. Johnson. *Contact Mechanics.* Cambridge University Press, 1985.

[96] Ch. Kaas-Petersen. Chaos in a railway bogie. *Acta Mechanica*, 61:89–107, 1986.

[97] J.J. Kalker. *Three-Dimensional Elastic Bodies in Rolling Contact*. Kluwer Academic Publishers, Dordrecht Boston London, 1990.

[98] K. Knothe, H. Le The. Ermittlung der Normalspannungsverteilung beim Kontakt von Rad und Schiene. *Forsch. Ing.-Wes.*, 49:79–85, 1983.

[99] E. Kreuzer. *Symbolische Berechnung der Bewegungsgleichungen von Mehrkörpersystemen*. Fortschritt-Berichte VDI Reihe 11, Nr. 32. VDI–Verlag, Düsseldorf, 1979.

[100] P. Kunkel, V. Mehrmann, W. Rath, J. Weickert. GELDA: A software package for the solution of general linear differential algebraic equations. *SIAM J. Sci. Comp.*, 18:115–138, 1997.

[101] B.J. Leimkuhler, L.R. Petzold, C.W. Gear. Approximation methods for the consistent initialization of differential-algebraic systems of equations. *SIAM J. Numer. Anal.*, 28:205–226, 1991.

[102] Ch. Linder, H. Brauchli. Prediction of wheel wear. In E. Zobory, editor, *Proceedings of the 2nd Miniconference on Contact Mechanics and Wear of Rail/Wheel Systems. Budapest, July 1996*. TU Budapest, 1996.

[103] P. Lötstedt. Mechanical systems of rigid bodies subject to unilateral constraints. *SIAM J. Appl. Math.*, 42:281–296, 1982.

[104] P. Lötstedt, L.R. Petzold. Numerical solution of nonlinear differential equations with algebraic constraints I: convergence results for backward differentiation formulas. *Math. of Comp.*, 46:491–516, 1986.

[105] Ch. Lubich. Linearly implicit extrapolation methods for differential-algebraic systems. *Numer. Math.*, 55:197–211, 1989.

[106] Ch. Lubich. Extrapolation integrators for constrained mechanical systems. *Impact Comput. Sc. Eng.*, 3:213–234, 1991.

[107] Ch. Lubich. Integration of stiff mechanical systems by Runge–Kutta methods. *Z. Angew. Math. Phys.*, 44:1022–1053, 1993.

[108] Ch. Lubich, Ch. Engstler, U. Nowak, U. Pöhle. Numerical integration of constrained mechanical systems using MEXX. *Mechanics of Structures and Machines*, 23:473–495, 1995.

[109] Ch. Lubich, U. Nowak, U. Pöhle, Ch. Engstler. MEXX – Numerical software for the integration of constrained mechanical multibody systems. Preprint SC 92–12, ZIB Berlin, 1992.

[110] Ch. Lubich, A. Ostermann. Runge–Kutta approximation of quasi-linear parabolic equations. *Math. Comp.*, 64:601–627, 1995.

[111] G. Maeß. *Vorlesungen über Numerische Mathematik I*. Akademie–Verlag, Berlin, 1984.

[112] C. Majer, W. Marquardt, E.D. Gilles. Reinitialization of DAE's after discontinuities. *Computers and Chemical Engineering*, 19(Supplement):S507–S512, 1995.

[113] R. März. Analysis and numerical treatment of differential-algebraic systems. In R.E. Bank, R. Bulirsch, K. Merten, editors, *Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices. Proc. of a Conference held at the Math. Forschungsinstitut, Oberwolfach, Oct. 30 – Nov. 5, 1988*, S. 27–43, Basel, 1989. Birkhäuser-Verlag.

[114] R. März. Numerical methods for differential-algebraic equations. *Acta Numerica*, S. 141–198, 1992.

[115] R.M.M. Mattheij. On ill-conditioning for index-1 DAEs. Vortrag zur Tagung „DAEs, Related Fields and Applications", Mathematisches Forschungsinstitut Oberwolfach, November 1995.

[116] S.E. Mattson, G. Söderlind. Index reduction in differential-algebraic equations using dummy derivatives. *SIAM J. Sci. Comp.*, 14:677–692, 1993.

[117] A. Murua. Partitioned Runge–Kutta methods for semi-explicit differential-algebraic systems of index 2. Technical report, Konputazio Zientziak eta A. A., Informatika Fakultatea, Donostia/San Sebastián, 1996.

[118] A. Murua. Partitioned half-explicit Runge–Kutta methods for differential-algebraic systems of index 2. *Computing*, 59:43–61, 1997.

[119] H. Netter. *Rad–Schiene–Systeme in differential-algebraischer Darstellung*. Dissertation, TU München, Fakultät für Maschinenwesen, Lehrstuhl B für Mechanik, 1997.

[120] H. Netter, M. Arnold. Geometrie und Dynamik eines Rad-Schiene-Modells in Deskriptorform mit unstetigen Zustandsgrößen. Interner Bericht IB 515–93–02, DLR, D-5000 Köln 90, 1993.

[121] R.E. O'Malley jun., L.V. Kalachev. Regularization of nonlinear differential-algebraic equations. *SIAM J. Math. Anal.*, 25:615–629, 1994.

[122] J.M. Ortega, W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York London, 1970.

[123] A. Ostermann. A half-explicit extrapolation method for differential-algebraic systems of index 3. *IMA J. Numer. Anal.*, 10:171–180, 1990.

[124] A. Ostermann. A class of half-explicit Runge–Kutta methods for differential-algebraic systems of index 3. *Applied Numerical Mathematics*, 13:165–179, 1993.

[125] C.C. Pantelides. The consistent initialization of differential-algebraic systems. *SIAM J. Sci. Stat. Comput.*, 9:213–231, 1988.

[126] J.P. Pascal. Benchmark to test wheel/rail contact forces. Technical report, INRETS Paris, 1990.

[127] J.P. Pascal. About multi–Hertzian–contact hypothesis and equivalent conicity in the case of S1002 and UIC60 analytical wheel/rail profiles. *Vehicle System Dynamics*, 22:57–78, 1993.

[128] J.P. Pascal, G. Sauvage. New method for reducing the multicontact wheel/rail problem to one equivalent rigid contact patch. In *The Dynamics of Vehicles on Roads and on Railway Tracks, 12th IAVSD–Symposium Lyon*, S. 475–489. Swets & Zeitlinger, B.V. Lisse, 1991.

[129] L.R. Petzold. A description of DASSL: a differential/algebraic system solver. Technical Report SAND82–8637, Sandia National Laboratories Livermore, 1982.

[130] L.R. Petzold. Differential/algebraic equations are not ODEs. *SIAM J. Sci. Stat. Comput.*, 3:367–384, 1982.

[131] L.R. Petzold. Order results for implicit Runge–Kutta methods applied to differential/algebraic systems. *SIAM J. Numer. Anal.*, 23:837–852, 1986.

[132] L.R. Petzold, P. Lötstedt. Numerical solution of nonlinear differential equations with algebraic constraints II: practical implications. *SIAM J. Sci. Stat. Comput.*, 7:720–733, 1986.

[133] F.A. Potra, W.C. Rheinboldt. On the numerical solution of Euler–Lagrange equations. *Mechanics of Structures and Machines*, 19:1–18, 1991.

[134] F.A. Potra, J. Yen. Implicit numerical integration for Euler–Lagrange equations via tangent space parametrization. *Mechanics of Structures and Machines*, 19:77–98, 1991.

[135] P.J. Rabier, W.C. Rheinboldt. A geometric treatment of implicit differential-algebraic equations. *J. of Differential Equations*, 109:110–146, 1994.

[136] P.J. Rabier, W.C. Rheinboldt. On impasse points of quasi-linear differential-algebraic equations. *J. Math. Anal. Appl.*, 181:429–454, 1994.

[137] P.J. Rabier, W.C. Rheinboldt. On the computation of impasse points of quasi-linear differential-algebraic equations. *Math. Comp.*, 62:133–154, 1994.

[138] P. Rentrop, K. Strehmel, R. Weiner. Ein Überblick über Einschrittverfahren zur numerischen Integration in der technischen Simulation. *Mitt. Ges. Angew. Math. Mech.*, 19:9–43, 1996.

[139] W.C. Rheinboldt. Differential-algebraic systems as differential equations on manifolds. *Math. Comp.*, 43:473–482, 1984.

[140] W.C. Rheinboldt. On the existence and uniqueness of solutions of nonlinear semi-implicit differential-algebraic equations. *Nonlinear Anal., Theory Methods Appl.*, 16:647–661, 1991.

[141] R.E. Roberson, R. Schwertassek. *Dynamics of Multibody Systems*. Springer-Verlag, Berlin Heidelberg New York, 1988.

[142] W. Rulka, A. Haigermoser, L. Mauer, H. Netter. Anwendung moderner Auslegungsstrategien für Schienenfahrzeuge durch Einsatz des Simulationsprogramms SIMPACK. VDI Berichte Nr. 1219, Düsseldorf, 1995.

[143] V. Schulz, H.G. Bock, M.C. Steinbach. Exploiting invariants in the numerical solution of multipoint boundary value problems. Preprint 93–69, IWR, Universität Heidelberg, 1993, erscheint in *SIAM J. Sci. Comp.*.

[144] B. Simeon. *Numerische Integration mechanischer Mehrkörpersysteme: Projizierende Deskriptorformen, Algorithmen und Rechenprogramme*. Fortschritt-Berichte VDI Reihe 20, Nr. 130. VDI–Verlag, Düsseldorf, 1994.

[145] B. Simeon. MBSPACK – Numerical integration software for constrained mechanical motion. *Surveys on Mathematics for Industry*, 5:169–202, 1995.

[146] B. Simeon. Modelling a flexible slider crank mechanism by a mixed system of DAEs and PDEs. *Math. Modelling of Systems*, 2:1–18, 1996.

[147] B. Simeon. On the numerical solution of a wheel suspension benchmark problem. *Comp. Appl. Math.*, 66:443–456, 1996.

[148] B. Simeon, C. Führer, P. Rentrop. Differential-algebraic equations in vehicle system dynamics. *Surveys on Mathematics for Industry*, 1:1–37, 1991.

[149] B. Simeon, F. Grupp, C. Führer, P. Rentrop. A nonlinear truck model and its treatment as a multibody system. *J. Comput. Appl. Math.*, 50:523–532, 1994.

[150] G. Söderlind. A multi-purpose system for the numerical integration of ODEs. *Appl. Math. Comp.*, 31:346–360, 1989.

[151] G. Söderlind. Remarks on the stability of high-index DAEs with respect to parametric perturbations. *Computing*, 49:303–314, 1992.

[152] J. Stoer, R. Bulirsch. *Numerische Mathematik 2*. Springer–Verlag, Berlin Heidelberg New York, 3. Auflage, 1990.

[153] K. Strehmel, R. Weiner. *Linear-implizite Runge–Kutta–Methoden und ihre Anwendung*. Teubner, Stuttgart–Leipzig, 1992.

[154] K. Strehmel, R. Weiner. *Numerik gewöhnlicher Differentialgleichungen*. Teubner Studienbücherei Mathematik. Teubner, Stuttgart, 1995.

[155] C. Tischendorf. Feasibility and stability behaviour of the BDF applied to index-2 differential-algebraic equations. *Z. Angew. Math. Mech.*, 75:927–946, 1995.

[156] J.G. Verwer. Convergence and order reduction of diagonally implicit Runge–Kutta schemes in the method of lines. In *Numerical analysis, Proc. 11th Conference 1985, Dundee*, S. 220–237. Pitman, 1986.

[157] R. von Schwerin, M. Winckler. A guide to the integrator library MBSSIM – Version 1.00. Preprint 94–75, IWR, Universität Heidelberg, 1994.

[158] W. Walter. *Gewöhnliche Differentialgleichungen*. Springer-Verlag, Berlin Heidelberg New York, 4. Auflage, 1990.

[159] R.A. Wehage, E.J. Haug. Generalized coordinate partitioning for dimension reduction in analysis of constrained dynamic systems. *J. Mech. Design*, 104:247–255, 1982.

[160] R. Weiner, M. Arnold, P. Rentrop, K. Strehmel. Partitioning strategies in Runge–Kutta type methods. *IMA J. Numer. Anal.*, 13:303–319, 1993.

[161] J. Wensch. Stabilität von Runge–Kutta–Methoden für Algebrodifferentialgleichungen vom Index 2. Martin–Luther–Universität Halle–Wittenberg, FB Mathematik und Informatik, Diplomarbeit, 1992.

[162] J. Wensch, K. Strehmel, R. Weiner. A class of linearly-implicit Runge–Kutta methods for multibody systems. *Applied Numerical Mathematics*, 22:381–398, 1996.

[163] J. Wensch, R. Weiner, K. Strehmel. Stability investigations for index-2 systems. Reports on Computer Science and Scientific Computing 1(1994), Martin–Luther–Universität Halle–Wittenberg, FB Mathematik und Informatik, 1994.

[164] P.M.E.J. Wijckmans. *Conditioning of Differential Algebraic Equations and Numerical Solution of Multibody Dynamics*. PhD thesis, Technical University Eindhoven, 1996.

[165] J. Yen. Constrained equations of motion in multibody dynamics as ODEs on manifolds. *SIAM J. Numer. Anal.*, 30:553–568, 1993.

[166] J. Yen. Dynamic simulation of contact between geometric objects in three dimensional space. Technical Report TR 169, University of Iowa, Center for Computer Aided Design, 1995.

[167] Profile für Radreifen und Radkränze. DIN 5573, September 1983.

Selbständigkeitserklärung

Ich erkläre, daß ich die vorliegende Habilitationsschrift selbständig und ohne fremde Hilfe verfaßt, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Rostock, den 16.12.1996

Lebenslauf

1964 geboren in Halle/S.

1970-80 Besuch der Polytechnischen und der Erweiterten Oberschule in Ilmenau/Thür.

1980-82 Schüler der Spezialklassen für Mathematik und Physik der Martin–Luther–Universität Halle–Wittenberg, Abschluß mit dem Abitur (Juli 1982).

1982-88 Student der Fachrichtung „Mathematik Diplom“ an der Martin–Luther–Universität Halle–Wittenberg, Unterbrechungen:

November 1982 – April 1984 Militärdienst,

Oktober 1987 – Februar 1988 Teilstudium an der Lettischen Universität Riga.

Juni 1988 Hochschulabschluß als Diplom–Mathematiker (Prädikat: „Ausgezeichnet“). Thema der Diplomarbeit: Linear-implizite Runge–Kutta–Verfahren zur numerischen Lösung quasilinearer parabolischer Anfangs–Randwertprobleme mit der Linienmethode.

1987-90 Forschungsstudent am Wissenschaftsbereich Numerische Mathematik der Martin–Luther–Universität Halle–Wittenberg (Prof. Dr. K. Strehmel).

Dezember 1990 Promotion zum Dr. rer. nat. an der Fakultät für Naturwissenschaften der Martin–Luther–Universität Halle–Wittenberg (Prädikat: „summa cum laude“). Dissertation: Numerische Behandlung von semi-expliziten Algebrodifferentialgleichungen vom Index 1 mit linear-impliziten Verfahren.

September 1990 – März 1991 Aspirantur an der Sektion Mathematik der Universität Rostock.

April 1991 – heute Wissenschaftlicher Assistent am Lehrstuhl für Numerische Mathematik der Universität Rostock.

1992-93 halbjähriger Forschungsaufenthalt in der Abteilung Fahrzeugsystemdynamik (Prof. Dr. W. Kortüm) der Deutschen Forschungsanstalt für Luft- und Raumfahrt (DLR Oberpfaffenhofen).

verheiratet, 3 Kinder.

Wissenschaftliche Auszeichnung

1993 Leopoldina–Preis für Junge Wissenschaftler (verliehen auf der Jahrestagung der Deutschen Akademie der Naturforscher Leopoldina).