Lie-Gruppen-Zeitintegration für mechanische Systeme mit Bewegungsgleichungen vom Differentiationsindex 3

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

der
Naturwissenschaftlichen Fakultät II
Chemie, Physik und Mathematik

der
Martin-Luther-Universität Halle-Wittenberg,
Institut für Mathematik

vorgelegt von
Frau Victoria Wieloch
geb. am 11. Februar 1992
in Halle (Saale)

Gutachter: Prof. Dr. Martin Arnold
Prof. Dr. Andreas Müller

Tag der Verteidigung: 28. Januar 2022

Inhaltsverzeichnis

Kapitel 1

Einleitung

Im Denken der Menschheit ist schon immer der Wille zum Verstehen der Welt um sie herum verankert. Eine Vielzahl von Phänomenen sowohl in der Technik als auch in der Natur können durch Differentialgleichungen beschrieben werden. Durch das Lösen dieser Gleichungen können die physikalischen Naturgesetze besser verstanden werden, jedoch führt deren Komplexität häufig dazu, dass keine geschlossene (analytische) Lösung gefunden werden kann. Aus diesem Grund wird in dieser Arbeit versucht mithilfe von numerischen Methoden eine hinreichend genaue Annäherung für die analytische Lösung von speziellen technischen Anwendungen (den Mehrkörpersystemen) zu finden und damit das tatsächliche Verhalten möglichst genau zu approximieren.

1.1 Motivation

Die numerische Simulation von Mehrkörpersystemen, also von mechanischen Systemen mit endlich vielen starren oder elastischen Einzelkörpern, die durch Kopplungselemente verknüpft sind und auf die Kräfte einwirken, spielt in vielen technischen Anwendungen wie beispielsweise der Robotik, der Fahrzeugtechnik oder der Partikelsimulation eine bedeutende Rolle [46]. Die resultierenden Bewegungsgleichungen solch eines Mehrkörpersystems setzen sich häufig aus gewöhnlichen Differentialgleichungen und algebraischen Nebenbedingungen zusammen. Diese Art von Gleichungen werden differential-algebraische Gleichungen (DAEs) genannt und sie können nicht ohne weitere Überlegungen durch numerische Verfahren zur Lösung von gewöhnlichen Differentialgleichungen gelöst werden.
Zudem können Schwierigkeiten auftreten, wenn Systeme mit großen Rotationen untersucht werden, welche beispielsweise in den speziellen Anwendungen der Bewegungen von Rotorenflügeln von Windrädern, Kabelverbindungen von Roboterarmen oder Solarsegeln in der Raumfahrt auftreten. Die Modellierung von solchen schwer zu lösenden Systemen mit großen Rotationen kann zu Singularitäten führen, welche durch die Verwendung von sogenannten Lie-Gruppen vermieden werden können. Eine Lie-Gruppe G ist eine differenzierbare Mannigfaltigkeit, für die ein Produkt oder eine Hintereinanderausführung und die Umkehrabbildung glatte Abbildungen sind. Wird eine Differentialgleichung auf einer Lie-Gruppe definiert, so bleibt die Lösung dieser Differentialgleichung stets auf G. Auch im Fall von Mehrkörpersystemen mit großen Rotationen werden durch die Verwendung von Lie-Gruppen die angesprochenen Singularitäten vermieden. Numerische Verfahren zur Lösung solcher Differentialgleichungen auf Lie-Gruppen sind somit in der Literatur von großem Interesse [4, 16, 48, 42].

Daher untersucht die vorliegende Arbeit Lösungen von Bewegungsgleichungen in Lie-Gruppen-Formulierung für Mehrkörpersysteme, die differential-algebraische Gleichungen darstellen, mit geeigneten numerischen Verfahren. Der Fokus liegt dabei auf zwei Arten von Verfahren: den BDF-Verfahren und dem Generalized-α-Verfahren.

1.2 Stand der Forschung

Die Lösung von Anfangswertproblemen ist seit vielen Jahren nicht mehr aus wissenschaftlichen Arbeiten wegzudenken. Dabei wird die Lösung einer Differentialgleichung in Form einer Funktion gesucht, die durch einen fest vorgegebenen Punkt (den Anfangswert) verläuft. Ein prominentes Beispiel für eine Person, die sich mit solchen Aufgabenklassen auseinandersetzte, ist Newton. Er beschäftigte sich bereits im 17. Jahrhundert mit Differentialgleichungen und ihren Lösungen mit Hilfe von Reihenentwicklungen [45]. Je komplexer die zu lösenden Gleichungen werden, umso schwieriger ist es geschlossene analytische Lösungen zu erhalten. Praktische Anwendungen sind zudem häufig von komplexer Natur, weshalb die numerische Lösung von Differentialgleichungen in der Literatur eine entscheidende Rolle spielt. Dabei werden numerische Verfahren entwickelt, um Lösungen von unterschiedlichen Typen von Differentialgleichungen möglichst genau zu approximieren. Im Fall von gewöhnlichen Differentialgleichungen waren in der Vergangenheit besonders zwei Arten solcher Verfahren von Bedeutung: die Einschrittverfahren und die linearen Mehrschrittverfahren. Eine spezielle Art von linearen Mehrschrittverfahren sind BDF-Verfahren, welche in dieser Arbeit von großem Interesse sind, da sie in industriellen Simulationspaketen für Mehrkörpersysteme beliebte Methoden mit Schrittweitensteuerung und Ordnungskontrolle sind [3]. Die BDF-Verfahren wurden zuerst als implizite k-Schrittverfahren der Ordnung $p = k$ für gewöhnliche Differentialgleichungen durch Curtiss und Hirschfelder [17] entwickelt. Sie wurden bekannt für die Lösung von steifen Differentialgleichungen durch die Arbeit von Gear [23].
In der Strukturdynamik haben sich Integratoren vom Newmark-Typ durchgesetzt. Das Generalized-α-Verfahren, das auf Chung und Hulbert [15] zurückgeht, ist solch ein Integrator zweiter Ordnung. Géradin und Cardona [24] sowie Arnold und Brüls [1] verwendeten das Verfahren zur Lösung von dynamischen Problemen mit Zwangsbedingungen. Weitere Untersuchungen erfolgten von Lunk und Simeon [37], Jay und Negrut [34] sowie Arnold u.a. [4] zur Lösung von differential-algebraischen Gleichungen (DAEs) mithilfe einer Indexreduktion.
Ein weit verbreitetes Forschungsthema ist zudem die Lösung von Differentialgleichungen auf Mannigfaltigkeiten [27]. Numerische Methoden zur Lösung solcher Differentialgleichungen sollen sicherstellen, dass die numerische Lösung auf diesen Mannigfaltigkeiten bleibt. Solche Klassen von Methoden sind zum Beispiel Projektionsmethoden oder Methoden, die auf lokalen Koordinaten basieren. Mannigfaltigkeiten mit einer Gruppenstruktur besitzen weitere nützliche Eigenschaften. Lie-Gruppen sind solche Mannigfaltigkeiten, die auch eine Gruppe darstellen. In den letzten drei Jahrzehnten waren numerische Lösungsverfahren für Differentialgleichungen auf Lie-Gruppen in der Forschung von großem Interesse. Klassische Ansätze dazu sind die von Crouch und Grossman [16] und Munthe-Kaas [42, 43].
Die numerische Integration von gewöhnlichen Differentialgleichungen auf Mannigfaltigkeiten durch explizite Formeln für Mehrschrittverfahren und Runge-Kutta-Verfahren dritter Ordnung durch Approximationen direkt in der Lie-Gruppe wurde von Crouch und Grossman [16] untersucht. In [16] wurden mehrere Exponentialabbildun-

gen kombiniert. Deren Auswertung erfordert jedoch einen hohen Aufwand an Rechenleistung. Approximationen in einem speziellen Tangentialraum der Lie-Gruppe (Lie-Algebra) können verwendet werden, um die Anzahl dieser Exponentialabbildungen zu reduzieren. Munthe-Kaas [42, 43] verwendete diese Strategie für Runge-Kutta-Verfahren auf Mannigfaltigkeiten. Die Bewegungsgleichungen können dabei durch eine gewöhnliche Differentialgleichung gelöst werden, die auf der korrespondierenden Lie-Algebra agiert. Hierbei sind Lie-Klammern involviert, um die hohe Ordnung dieser Lie-Gruppen-Verfahren zu erhalten. Celledoni, Marthinsen und Owren [13] erreichten eine kleinere Anzahl an Exponentialabbildungen als Crouch und Grossman [16] durch die Verwendung von Linearkombinationen innerhalb dieser Abbildungen und vermieden zudem die Berechnung von Lie-Klammern für Runge-Kutta-Verfahren.
Mehrschrittverfahren für Differentialgleichungen auf Mannigfaltigkeiten wurden von Faltinsen, Marthinsen und Munthe-Kaas [20] eingeführt. In diesen Methoden wurden vor jedem Zeitschritt lokale Koordinaten verwendet, um die gegebenen Nichtlinearitäten der Lie-Gruppen-Formulierungen zu berechnen.
Weitere Referenzen zu Mehrkörpersystemen in Verbindung mit Lie-Gruppen und deren Verwendung sind unter anderem [14], [41] und [48].

1.3 Beitrag der vorliegenden Arbeit und Gliederung

Die effiziente Implementierung der oben genannten BDF-Verfahren für Mehrkörpersysteme mit Lie-Gruppen-Struktur ist eine bisher unbeantwortete Frage, derer sich die vorliegende Arbeit widmet. Dabei wird zum einen auf die BDF-Variante der Verfahren aus [20] eingegangen und zum anderen ein neues BDF-Lie-Gruppen-Verfahren, das BLieDF-Verfahren, eingeführt, in dem eine weniger zeitaufwendige Auswertung der vorhandenen Lie-Klammern erfolgt. Beide Varianten werden auf differential-algebraische Gleichungen in der Index-3-Formulierung angewendet und die Konvergenz der Ordnung $p = k$ bewiesen.
Zur möglichst effizienten Rechnung haben sich Verfahren mit variabler Schrittweite und Schrittweitensteuerung bewährt. Daher werden im Verlauf der Arbeit sowohl das Generalized-α-Verfahren als auch das BLieDF-Verfahren für variable Schrittweiten erweitert und die Konvergenz der Verfahren zur Lösung von beschränkten mechanischen Mehrkörpersystemen bewiesen werden. Dabei ist das Risiko einer Ordnungsreduktion zu beachten, die bei einer naiven Erweiterung aufgrund der Index-3-DAE-Struktur zu beobachten wäre.
Die vorliegende Arbeit soll somit einen Beitrag zur effizienten Lösung von mechanischen Mehrkörpersystemen leisten, indem bereits etablierte Verfahren so erweitert werden, dass mit ihnen Differentialgleichungen auf Konfigurationsräumen mit Lie-Gruppen-Struktur unter Verwendung von variablen Schrittweiten gelöst werden können.
Die Arbeit ist wie folgt gegliedert. Das nachfolgende Kapitel führt gewöhnliche Differentialgleichungen zweiter Ordnung ein und stellt lineare Mehrschrittverfahren (BDF- und Adams-Moulton-Verfahren) sowie das Generalized-α-Verfahren zur Lösung dieser vor. Außerdem wird ein Einblick in die Thematik der Lie-Gruppen und entsprechend in die Lösung von Differentialgleichungen für Konfigurationsräume mit Lie-Gruppen-Struktur gegeben.

Im dritten Kapitel erfolgt eine Einführung in die Zeitintegration solcher speziellen Differentialgleichungen durch die vorgestellten linearen Mehrschrittverfahren und das Generalized-α-Verfahren. Dabei werden nicht nur die vorhandenen Verfahren von [8] und [20] vorgestellt, sondern zusätzlich die Ideen aus [13] und [16] verwendet, um entsprechende Mehrschrittverfahren zur Lösung von Differentialgleichungen zweiter Ordnung zu definieren, die später als Vergleich für die BDF-Verfahren dienen sollen. Im Anschluss wird die Zeitintegration von mechanischen Mehrkörpersystemen ohne Zwangsbedingungen und von beschränkten Mehrkörpersystemen untersucht. Dazu werden die in vorgestellten Verfahren auf unbeschränkte mechanische Systeme angewendet und das Benchmarkproblem „schwerer Kreisel“ vorgestellt. Außerdem wird das neue BLieDF-Verfahren eingeführt. Anschließend werden das Generalized-α-Verfahren, das Munthe-Kaas-BDF-Verfahren (basierend auf [20]) und das BLieDF-Verfahren auf differential-algebraische Gleichungen vom Index 3 angewendet.
In Kapitel 4 wird der Konvergenzbeweis des Generalized-α-Verfahrens für differential-algebraische Gleichungen aus [2] vorgestellt und verwendet, um anschließend die Konvergenz des Verfahrens auch für variable Schrittweiten zu beweisen. Dazu muss das Verfahren zunächst so umgeschrieben werden, dass die gewünschte Ordnung auch bei wechselnder Schrittweite erhalten werden kann. Zusätzlich wird untersucht, für welche Schrittweitenverhältnisse das Verfahren in jedem Fall stabil ist.
Das darauffolgende Kapitel verwendet ähnliche Techniken, um die Konvergenz der Ordnung $p = k$ für das Munthe-Kaas-BDF-Verfahren und das BLieDF-Verfahren für $2 \leq k \leq 6$, konstante Schrittweiten und beschränkte mechanische Systeme zu beweisen. Eine Übertragung der BLieDF-Verfahren auf variable Schrittweiten und der Beweis der Konvergenz für $2 \leq k \leq 3$ wird im Anschluss thematisiert.
Kapitel 6 führt anhand des vorgestellten Benchmarkproblems numerische Tests für die verwendeten Verfahren sowohl für konstante, als auch gegebenenfalls für variable Schrittweiten aus. Die Arbeit schließt mit einer Zusammenfassung der Ergebnisse.

Kapitel 2

Vorbetrachtungen

2.1 Gewöhnliche Differentialgleichungen und ausgewählte Zeitintegrationsverfahren

In diesem Abschnitt werden gewöhnliche Differentialgleichungen und einige Zeitintegrationsverfahren zur Lösung von Anfangswertproblemen vorgestellt. Da im Verlauf der Arbeit vor allem Differentialgleichungen 2. Ordnung bzw. deren Äquivalent als System von Differentialgleichungen 1. Ordnung von Interesse sind, werden allgemeine Differentialgleichungen 1. Ordnung lediglich im Anhang A eingeführt. Ebenso werden dort grundlegende Begriffe wie Konsistenz, Konvergenz und Nullstabilität für lineare Mehrschrittverfahren für solche Gleichungen 1. Ordnung erläutert. Die Begriffe lassen sich problemlos auch auf Differentialgleichungen 2. Ordnung übertragen.

Definition 1 (Gewöhnliche Differentialgleichung 2. Ordnung (ODE) [28])
Ein *System gewöhnlicher Differentialgleichungen (engl.: ordinary differential equations, ODEs) 2. Ordnung* ist ein Gleichungssystem der Form

$$\mathbf{x}'' = \mathbf{f}(t, \mathbf{x}, \mathbf{x}') \tag{2.1}$$

mit einer gegebenen Funktion $\mathbf{f} : \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$. Eine Funktion $\mathbf{x}(t)$ wird *Lösung* der Differentialgleichung (2.1) auf dem Zeitintervall $[t_0, t_{\text{end}}]$ genannt, falls für alle $t \in [t_0, t_{\text{end}}]$ der Zusammenhang

$$\mathbf{x}''(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{x}'(t)) \tag{2.2a}$$

gilt. Ist die Funktion $\mathbf{f}$ lipschitzstetig, so ist die Lösung $\mathbf{x}(t)$ eindeutig durch zwei *Anfangswerte*

$$\mathbf{x}(t_0) = \mathbf{x}_0 \quad \text{und} \quad \mathbf{x}'(t_0) = \mathbf{x}'_0 \tag{2.2b}$$

festgelegt. Ein *Anfangswertproblem* ist die Suche nach einer Lösung $\mathbf{x}(t)$ von (2.2a) mit den Anfangswerten (2.2b).

In dieser Arbeit stellt t stets die Zeit dar, weshalb die Suche nach einer Lösung des Anfangswertproblems (2.2) auch *Zeitintegration* genannt wird und die Ableitungen nach t, wie in der Literatur üblich, durch einen Punkt dargestellt werden

$$\dot{\mathbf{x}}(t) \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t).$$

Außerdem soll die Bewegung von Körpern im Raum beschrieben werden. Dabei wird von dem Modell des Starrkörpers ausgegangen und dessen Drehung im Inertial- bzw. körperfesten System untersucht.

Definition 2 (Starrkörper, Inertialsystem, körperfestes System [19, 21])
Ein *Starrkörper* besteht aus vielen einzelnen Massenpunkten, deren Lage zueinander unverändert bleibt, unabhängig davon, welchen Einflüssen und Kräften der Körper unterliegt.
Ein Bezugssystem heißt *Inertialsystem*, wenn jeder Körper, auf den keine Kräfte wirken, relativ zu diesem Bezugssystem in Ruhe verharrt oder sich gleichförmig geradlinig bewegt. Bei einem *körperfesten Bezugssystem* sind ein beliebig festgelegter Punkt des Starrkörpers als Ursprung und beliebige, relativ zum Körper ruhende Achsen als Koordinatenachsen gewählt.

Im Folgenden bezeichnet $\mathbf{x}(t)$ die Konfiguration eines oder mehrerer Starrkörper zur Zeit t, deren Geschwindigkeit durch $\mathbf{v}(t) = \dot{\mathbf{x}}(t)$ beschrieben wird. Somit kann (2.2) zu dem System 1. Ordnung

$$\dot{\mathbf{x}}(t) = \mathbf{v}(t), \tag{2.3a}$$
$$\dot{\mathbf{v}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{v}(t)), \tag{2.3b}$$
$$\mathbf{x}(t_0) = \mathbf{x}_0, \tag{2.3c}$$
$$\mathbf{v}(t_0) = \mathbf{v}_0 \tag{2.3d}$$

mit $\mathbf{v}_0 \coloneqq \dot{\mathbf{x}}_0$ umformuliert werden [28].
Die Zeitintegration von (2.2) bzw. (2.3) erfolgt durch numerische Verfahren, die die Lösung auf einem *Punktgitter* [46]

$$I_h = \{t_0, t_1, ..., t_{N_{\text{end}}}\} \quad \text{mit} \quad t_0 < t_1 < ... < t_{N_{\text{end}}} = t_{\text{end}} \tag{2.4}$$

mit Schrittweiten $h_n = t_{n+1} - t_n$, $(n = 0, ..., N_{\text{end}} - 1)$, im Zeitintervall $t \in [t_0, t_{\text{end}}]$ approximieren. Ist $h = h_n$ für alle $n = 0, ..., N_{\text{end}} - 1$ konstant, so ist das Punktgitter *äquidistant* und ansonsten *variabel*. Zunächst soll die Schrittweite h konstant sein. Im Verlauf der Arbeit erfolgt die Untersuchung einiger numerischer Verfahren für variable Schrittweiten h_n (vgl. Kapitel 4 und Abschnitt 5.2).
Die zwei bekanntesten Arten von Zeitintegrationsverfahren sind Einschrittverfahren und lineare Mehrschrittverfahren. In dieser Arbeit sind die Mehrschrittverfahren von Interesse. Einschrittverfahren werden nicht weiter thematisiert. Im Folgenden sollen zwei Vertreter solcher linearer Mehrschrittverfahren zur Lösung von (2.3) eingeführt werden. Dies sind zum einen die BDF-Verfahren und zum anderen die Adams-Moulton-Verfahren. Als weiteres Verfahren wird das Generalized-α-Verfahren vorgestellt, das sich nicht eindeutig in eine der beiden Gruppen (Einschritt- bzw. Mehrschrittverfahren) einordnen lässt.

2.1.1 BDF-Verfahren

Definition 3 (BDF-Verfahren [28, für ODEs 1. Ordnung])
Die *BDF-Verfahren* zur Lösung von (2.3) sind k-Schritt-Verfahren der Form

$$\frac{1}{h} \sum_{i=0}^{k} \alpha_i \mathbf{x}_{n+1-i} = \mathbf{v}_{n+1}, \tag{2.5a}$$

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{v}_{n+1-i} = \mathbf{f}(t_{n+1}, \mathbf{x}_{n+1}, \mathbf{v}_{n+1}) \tag{2.5b}$$

für den Zeitschritt $t_n \to t_{n+1} := t_n + h$ mit konstanter Schrittweite h. Dabei werden zu vorgegebenen numerischen Lösungen $\mathbf{x}_{n+1-i}$ und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, die Approximationen $\mathbf{x}_{n+1} \approx \mathbf{x}(t_{n+1})$ und $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ berechnet. Die Parameter α_i erfüllen die Konsistenzbedingungen (vgl. Satz A.1)

$$\sum_{i=0}^{k}\alpha_i = 0, \tag{2.6a}$$

$$\sum_{i=0}^{k}\alpha_i \frac{(k-i)^\ell}{k^{\ell-1}} = \ell, \quad (\ell = 1, ..., k), \tag{2.6b}$$

und sind daher durch

$$k = 1: \quad \alpha_0 = 1,\ \alpha_1 = -1, \tag{2.7a}$$

$$k = 2: \quad \alpha_0 = \frac{3}{2},\ \alpha_1 = -2,\ \alpha_2 = \frac{1}{2}, \tag{2.7b}$$

$$k = 3: \quad \alpha_0 = \frac{11}{6},\ \alpha_1 = -3,\ \alpha_2 = \frac{3}{2},\ \alpha_3 = -\frac{1}{3}, \tag{2.7c}$$

$$k = 4: \quad \alpha_0 = \frac{25}{12},\ \alpha_1 = -4,\ \alpha_2 = 3,\ \alpha_3 = -\frac{4}{3},\ \alpha_4 = \frac{1}{4}, \tag{2.7d}$$

$$k = 5: \quad \alpha_0 = \frac{137}{60},\ \alpha_1 = -5,\ \alpha_2 = 5,\ \alpha_3 = -\frac{10}{3},\ \alpha_4 = \frac{5}{4},\ \alpha_5 = -\frac{1}{5}, \tag{2.7e}$$

$$k = 6: \quad \alpha_0 = \frac{49}{20},\ \alpha_1 = -6,\ \alpha_2 = \frac{15}{2},\ \alpha_3 = -\frac{20}{3},\ \alpha_4 = \frac{15}{4},\ \alpha_5 = -\frac{6}{5},\ \alpha_6 = \frac{1}{6} \tag{2.7f}$$

gegeben.

Diese impliziten Mehrschrittverfahren wurden durch Curtiss und Hirschfelder [17] eingeführt und erlangten durch die Arbeit von Gear [23] große Bekanntheit zur Lösung von steifen Differentialgleichungen. Die Verfahren sind für $1 \leq k \leq 6$ nullstabil und haben die Konvergenzordnung $p = k$. Für $k \leq 2$ sind sie A-stabil, jedoch nur $A(\alpha)$-stabil für $3 \leq k \leq 6$. Ab $k \geq 7$ werden die Verfahren instabil und wurden deshalb in den numerische Rechnungen nicht näher untersucht, vgl. [29].
Um die Zeitintegration mit einem Mehrschrittverfahren wie (2.5) zu starten, reichen die Anfangswerte (2.3c) und (2.3d) nicht aus. Es werden Approximationen $\mathbf{x}_i \approx \mathbf{x}(t_i)$ und $\mathbf{v}_i \approx \mathbf{v}(t_i)$, $(i = 0, ..., k-1)$, benötigt. Diese können zum Beispiel durch ein Einschrittverfahren mit hinreichender Ordnung (vgl. Gleichung (A.5)) berechnet werden, vgl. [28].
Für einige theoretische Untersuchungen ist es von Vorteil, die Konsistenzbedingungen (2.6) in einer anderen Art und Weise zu formulieren. Dazu kann das nachfolgende Lemma bewiesen werden.

Lemma 1
Die Konsistenzbedingungen (2.6) sind äquivalent zu

$$\sum_{i=0}^{k}\alpha_i i^\ell = 0, \quad (\ell = 0, 2, ..., k),$$

$$\sum_{i=0}^{k} \alpha_i i = -1.$$

Beweis:
Jedes Polynom $\pi \in \Pi_k$ stimmt mit seinem k-ten Taylorpolynom, entwickelt in $t = t_{n-(k-1)}$, überein. Daher gelten die Zusammenhänge

$$\pi(t_{n+1-i}) = \pi\big(t_{n-(k-1)} + (k-i)h\big) = \sum_{\ell=0}^{k} \pi^{(\ell)}\big(t_{n-(k-1)}\big)\frac{(k-i)^\ell h^\ell}{\ell!} \tag{2.8}$$

und

$$\begin{aligned} h\dot{\pi}(t_{n+1}) = h\dot{\pi}\big(t_{n-(k-1)} + kh\big) &= \sum_{\ell'=0}^{k-1} \pi^{(1+\ell')}\big(t_{n-(k-1)}\big)\frac{k^{\ell'} h^{\ell'+1}}{\ell'} \\ &= \sum_{\ell=1}^{k} \pi^{(\ell)}\big(t_{n-(k-1)}\big)\frac{k^{\ell-1} h^\ell}{(\ell-1)!}. \end{aligned} \tag{2.9}$$

Durch die Konsistenzbedingungen (2.6) folgt

$$\sum_{i=0}^{k} \alpha_i \pi(t_{n+1-i}) = h\dot{\pi}(t_{n+1}) \tag{2.10}$$

und die Taylorentwicklung dieses Ausdrucks in $t = t_{n-(k-1)}$ beweist die Behauptung.

■

Für die späteren Betrachtungen der BDF-Verfahren für Konfigurationsräume mit Lie-Gruppen-Struktur ist eine Umformulierung der Gleichung (2.5a) zielführend. Dazu wird das Inkrement definiert durch

$$\boldsymbol{\Delta}\mathbf{x}_n \coloneqq \mathbf{x}_{n+1} - \mathbf{x}_n \tag{2.11}$$

und die nachfolgenden Lemmata können bewiesen werden.

Lemma 2
Für Parameter $\alpha_i \in \mathbb{R}$, $(i = 0, ..., k)$, die die Konsistenzbedingungen (2.6) erfüllen und

$$\gamma_i = \sum_{j=0}^{i-1} \alpha_j, \quad (i = 1, ..., k), \tag{2.12}$$

sowie Inkrementen (2.11) gilt

$$\sum_{i=0}^{k} \alpha_i \mathbf{x}_{n+1-i} = \sum_{i=1}^{k} \gamma_i (\mathbf{x}_{n+2-i} - \mathbf{x}_{n+1-i}) = \sum_{i=1}^{k} \gamma_i \boldsymbol{\Delta}\mathbf{x}_{n+1-i}.$$

Beweis:
Es gilt

$$\begin{aligned} \sum_{i=0}^{k} \alpha_i \mathbf{x}_{n+1-i} &= \alpha_0 \mathbf{x}_{n+1} + \alpha_1 \mathbf{x}_n + ... + \alpha_k \mathbf{x}_{n+1-k} \\ &= \alpha_0(\mathbf{x}_{n+1} - \mathbf{x}_n) + (\alpha_0 + \alpha_1)(\mathbf{x}_n - \mathbf{x}_{n-1}) + ...+ \\ &\quad + (\alpha_0 + ... + \alpha_{k-1})(\mathbf{x}_{n+2-k} - \mathbf{x}_{n+1-k}) + (\alpha_0 + ... + \alpha_k)\mathbf{x}_{n+1-k} \\ &= \sum_{i=1}^{k} \gamma_i (\mathbf{x}_{n+2-i} - \mathbf{x}_{n+1-i}) \end{aligned}$$

$$= \sum_{i=1}^{k} \gamma_i \boldsymbol{\Delta}\mathbf{x}_{n+1-i}$$

mit (2.6a) und (2.12).

■

Lemma 3
Gleichung (2.5a) ist äquivalent zu

$$\frac{1}{h} \sum_{i=1}^{k} \gamma_i \boldsymbol{\Delta}\mathbf{x}_{n+1-i} = \mathbf{v}_{n+1} \tag{2.13}$$

mit (2.11) und Parametern γ_i, die (2.12) erfüllen.

Beweis:
Die Behauptung folgt direkt mit Lemma 2.

■

Die Darstellung

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \boldsymbol{\Delta}\mathbf{x}_n, \tag{2.14a}$$

$$\frac{1}{h} \sum_{i=1}^{k} \gamma_i \boldsymbol{\Delta}\mathbf{x}_{n+1-i} = \mathbf{v}_{n+1}, \tag{2.14b}$$

$$\frac{1}{h} \sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i} = \mathbf{f}(t_{n+1}, \mathbf{x}_{n+1}, \mathbf{v}_{n+1}) \tag{2.14c}$$

wird auch *BDF-Verfahren in Inkrementform* genannt. Die Parameter γ_i, $(i = 1, ..., k)$, definiert durch (2.12), müssen ebenso wie α_i in (2.6) Konsistenzbedingungen genügen, damit auch die BDF-Verfahren in Inkrementform (2.14) mit der Konvergenzordnung $p = k$ konvergieren.

Lemma 4
Die BDF-Verfahren in Inkrementform (2.14) konvergieren nur dann mit Ordnung $p = k$, wenn die Parameter γ_i aus (2.12) die Konsistenzbedingungen

$$\sum_{i=1}^{k} \gamma_i \frac{(k+1-i)^\ell - (k-i)^\ell}{k^{\ell-1}} = \ell, \quad (\ell = 1, ..., k), \tag{2.15}$$

erfüllen.

Beweis:
Die BDF-Verfahren (2.5) konvergieren nur dann mit der Ordnung $p = k$, wenn die Parameter α_i, $(i = 0, ..., k)$, die Bedingungen (2.6) erfüllen, vgl. [28] bzw. Satz A.1. Nach Lemma 3 sind die Gleichungen (2.13) und (2.5a) äquivalent für γ_i, $(i = 1, ..., k)$, aus (2.12). Um die Behauptung des Lemmas zu beweisen, muss somit gezeigt werden, dass (2.15) aus (2.6) folgt. Dazu wird wie im Beweis von Lemma 1 begonnen. Mit Gleichung (2.10) und Lemma 2 gilt

$$\sum_{i=1}^{k} \gamma_i \big(\pi(t_{n+2-i}) - \pi(t_{n+1-i})\big) = h\dot{\pi}(t_{n+1})$$

für jedes Polynom $\pi \in \Pi_k$. Unter Verwendung der Gleichungen (2.8) und (2.9) folgt die Behauptung.

■

Aufgrund von Gleichung (2.12) und Lemma 4 sind die Parameter γ_i für $k = 1, ..., 6$ gegeben durch

$$k = 1: \quad \gamma_1 = 1, \tag{2.16a}$$

$$k = 2: \quad \gamma_1 = \frac{3}{2},\ \gamma_2 = -\frac{1}{2}, \tag{2.16b}$$

$$k = 3: \quad \gamma_1 = \frac{11}{6},\ \gamma_2 = -\frac{7}{6},\ \gamma_3 = \frac{1}{3}, \tag{2.16c}$$

$$k = 4: \quad \gamma_1 = \frac{25}{12},\ \gamma_2 = -\frac{23}{12},\ \gamma_3 = \frac{13}{12},\ \gamma_4 = -\frac{1}{4}, \tag{2.16d}$$

$$k = 5: \quad \gamma_1 = \frac{137}{60},\ \gamma_2 = -\frac{163}{60},\ \gamma_3 = \frac{137}{60},\ \gamma_4 = -\frac{21}{20},\ \gamma_5 = \frac{1}{5}, \tag{2.16e}$$

$$k = 6: \quad \gamma_1 = \frac{49}{20},\ \gamma_2 = -\frac{71}{20},\ \gamma_3 = \frac{79}{20},\ \gamma_4 = -\frac{163}{60},\ \gamma_5 = \frac{31}{30},\ \gamma_6 = -\frac{1}{6}. \tag{2.16f}$$

2.1.2 Adams-Moulton-Verfahren

Definition 4 (Adams-Moulton-Verfahren [28, für ODEs 1. Ordnung])
Die *Adams-Moulton-Verfahren* zur Lösung von (2.3) sind k-Schritt-Verfahren der Form

$$\frac{1}{h}(\mathbf{x}_{n+1} - \mathbf{x}_n) = \sum_{i=0}^{k} \beta_i \mathbf{v}_{n+1-i}, \tag{2.17a}$$

$$\frac{1}{h}(\mathbf{v}_{n+1} - \mathbf{v}_n) = \sum_{i=0}^{k} \beta_i \mathbf{f}(t_{n+1-i}, \mathbf{x}_{n+1-i}, \mathbf{v}_{n+1-i}) \tag{2.17b}$$

für den Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit konstanter Schrittweite h. Dabei werden für gegebene numerische Lösungen $\mathbf{x}_{n+1-i}$ und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, die Approximationen $\mathbf{x}_{n+1} \approx \mathbf{x}(t_{n+1})$ und $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ berechnet. Die Parameter β_i sind gegeben durch

$$\begin{aligned}
k = 0: &\quad \beta_0 = 1, \\
k = 1: &\quad \beta_0 = \frac{1}{2},\ \beta_1 = \frac{1}{2}, \\
k = 2: &\quad \beta_0 = \frac{5}{12},\ \beta_1 = \frac{2}{3},\ \beta_2 = -\frac{1}{12}, \\
k = 3: &\quad \beta_0 = \frac{9}{24},\ \beta_1 = \frac{19}{24},\ \beta_2 = -\frac{5}{24},\ \beta_3 = \frac{1}{24}.
\end{aligned}$$

Das Verfahren für $k = 0$ ist das implizite Euler-Verfahren und theoretisch ein Einschrittverfahren. Die Trapezregel ergibt sich für $k = 1$. Adams-Moulton-Verfahren sind implizite Verfahren mit der Konvergenzordnung $p = k+1$, die unter Verwendung von Quadraturformeln konstruiert werden. Außerdem sind sie optimal nullstabil [46].

2.1.3 Generalized-α-Verfahren

Das Generalized-α-Verfahren ist ein Verfahren vom Newmark-Typ, das zurückgeht auf Arbeiten von Chung und Hulbert [15], die die Zeitintegration für lineare Systeme in linearen Räumen betrachtet haben.

Definition 5 (Generalized-α-Verfahren [15, für Matrixgl. der lin. Strukturdynamik])
Bei dem *Generalized-α-Verfahren* zur Lösung von (2.3) werden die numerischen Lösungen $\mathbf{x}_n$, $\mathbf{v}_n$, $\dot{\mathbf{v}}_n$ und $\mathbf{a}_n$ im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit konstanter Schrittweite h verwendet und es ist gegeben durch

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\mathbf{v}_n + (0.5 - \beta)h^2\mathbf{a}_n + \beta h^2\mathbf{a}_{n+1}, \tag{2.18a}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + (1 - \gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1}, \tag{2.18b}$$

$$(1 - \alpha_m)\mathbf{a}_{n+1} + \alpha_m\mathbf{a}_n = (1 - \alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n, \tag{2.18c}$$

$$\dot{\mathbf{v}}_{n+1} = \mathbf{f}(t_{n+1}, \mathbf{x}_{n+1}, \mathbf{v}_{n+1}) \tag{2.18d}$$

mit den Parametern aus (2.20). In jedem Zeitschritt werden Variablen $\mathbf{x}_{n+1} \approx \mathbf{x}(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$, $\dot{\mathbf{v}}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1})$ und die Beschleunigung

$$\mathbf{a}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h)$$

an einer Zwischenstelle mit

$$\Delta_\alpha = \alpha_m - \alpha_f \tag{2.19}$$

approximiert.

Die beschleunigungsähnliche Variable $\mathbf{a}_n$ ist eine Hilfsvariable. In [15] wurden außerdem für die Verfahrensparameter α_f, α_m, β und γ optimale Werte angegeben in Bezug auf Stabilitäts- und Genauigkeitseigenschaften in Abhängigkeit von einem Dämpfungparameter $\rho_\infty \in [0, 1)$. Dabei wurde in [15] die lineare Testgleichung

$$\ddot{x} + \omega^2 x = 0, \quad \omega \in \mathbb{R},$$

untersucht, deren Anwendung auf das Generalized-α-Verfahren (2.18) in einer von $h\omega$ abhängigen linearen Abbildung $(q_n, v_n, a_n) \mapsto (q_{n+1}, v_{n+1}, a_{n+1})$ resultiert. Der Spektralradius für $h\omega \to \infty$ der Matrix, die diese lineare Abbildung definiert, ist genau dieser Dämpfungsparameter ρ_∞. Die Verfahrensparameter ergeben sich zu

$$\alpha_m = \frac{2\rho_\infty - 1}{\rho_\infty + 1}, \quad \alpha_f = \frac{\rho_\infty}{\rho_\infty + 1}, \quad \beta = \frac{1}{4}\left(\gamma + \frac{1}{2}\right)^2, \quad \gamma = \frac{1}{2} + \alpha_f - \alpha_m. \tag{2.20}$$

Mit diesen Parametern ist das Verfahren (2.18) ein Zeitintegrationsverfahren zweiter Ordnung [15].

2.2 Lie-Gruppen

In diesem Abschnitt werden Lie-Gruppen – im Speziellen Matrix-Lie-Gruppen – eingeführt sowie einige Beispiele vorgestellt.

Definition 6 (Lie-Gruppen [27, 5])
Eine *Lie-Gruppe* G ist eine differenzierbare Mannigfaltigkeit (vgl. [27]), auf der eine Gruppenoperation $\circ : G \times G \to G$ und eine inverse Abbildung $inv : G \to G$ definiert sind, wobei sowohl $\circ$ als auch inv differenzierbare Abbildungen sind.

Da jede endliche Lie-Gruppe $(G, \circ)$ mit der Gruppenoperation $\circ$ isomorph zu einer Matrix-Lie-Gruppe ist, werden nur diese näher betrachtet und im folgenden Abschnitt erläutert.

2.2.1 Matrix-Lie-Gruppen

Definition 7 (Matrix-Lie-Gruppen [27])
Eine *Matrix-Lie-Gruppe* G ist eine Lie-Gruppe, die eine Untergruppe der allgemeinen linearen Gruppe

$$\mathrm{GL}(n_G) \coloneqq \{\mathbf{A} \in \mathbb{R}^{n_G \times n_G} : \ \det \mathbf{A} \neq 0\}$$

darstellt.

Ein Element $q \in G$ kann somit stets als Matrix aufgefasst werden, wobei die Gruppenoperation $\circ$ die Matrixmultiplikation darstellt. In praktischen Implementierungen ist es jedoch in Hinblick auf die Rechenzeit oft ungünstig, q als Matrix zu charakterisieren, weshalb weiterhin die Gruppenoperation $\circ$ verwendet wird.
Die Konfiguration eines Körpers (vgl. Definition 2) zur Zeit t wird durch die Funktion

$$q(t) : \ \mathbb{R} \to G \tag{2.21}$$

beschrieben. Die stetig differenzierbare Funktion $q(t)$ mit $q(t_0) \in G$ bleibt genau dann in der Lie-Gruppe G, wenn ihre Zeitableitung $\dot{q}(t)$ im *Tangentialraum* T_qG bzgl. des Punktes $q = q(t)$ liegt [27].

Definition 8 (Lie-Algebra, Lie-Klammer und Matrix-Kommutator [5])
Der Tangentialraum in der Identität e einer N-dimensionalen Lie-Gruppe G wird als *Lie-Algebra* $\mathfrak{g} = T_eG$ bezeichnet. Die *Lie-Klammer* $[\mathbf{A}, \mathbf{B}] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ ist eine bilineare, schiefsymmetrische Operation, für die die Jacobi-Identität

$$[\mathbf{A}, [\mathbf{B}, \mathbf{C}]] + [\mathbf{C}, [\mathbf{A}, \mathbf{B}]] + [\mathbf{B}, [\mathbf{C}, \mathbf{A}]] = \mathbf{0} \tag{2.22}$$

erfüllt ist. Ist G eine Matrix-Lie-Gruppe, so ist die Lie-Klammer isomorph zu dem *(Matrix)-Kommutator* gegeben durch $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$.

Da alle Tangentialräume linear sind, ist auch die Lie-Algebra $\mathfrak{g}$ ein linearer Raum. Durch die invertierbare lineare Abbildung

$$\widetilde{(\bullet)} : \mathbb{R}^N \to \mathfrak{g} \tag{2.23}$$

wird ein Isomorphismus zwischen dem euklidischen Raum $\mathbb{R}^N$ und der Lie-Algebra $\mathfrak{g}$ beschrieben [8]. Im Folgenden wird ein Element der Lie-Algebra stets durch die Abbildung (2.23) kenntlich gemacht.
Durch die Gruppenstruktur von G kann jedes Element von T_qG in jedem Punkt $q \in G$ durch Elemente $\widetilde{\mathbf{v}}$ der Lie-Algebra repräsentiert werden. Dafür definiert die Linkstranslation

$$L_q : \ G \to G, \ y \mapsto L_q(y) = q \circ y \tag{2.24}$$

eine Bijektion in G, und ihre Ableitung $DL_q(e)$ in der Identität e ist eine Bijektion zwischen den Tangentialräumen $\mathfrak{g} \coloneqq T_eG$ und T_qG, das heißt [4]

$$T_qG = \{DL_q(e) \cdot \widetilde{\mathbf{v}} : \ \widetilde{\mathbf{v}} \in \mathfrak{g}\} = \{DL_q(e) \cdot \widetilde{\mathbf{v}} : \ \mathbf{v} \in \mathbb{R}^N\}.$$

Somit kann die Ableitung von (2.21) als

$$\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\mathbf{v}}(t) \tag{2.25}$$

beschrieben werden [4]. Ein Großteil der Rechnungen soll in der Lie-Algebra bzw. im euklidischen Raum $\mathbb{R}^N$ vonstattengehen. Anschließend erfolgt eine Transformation in die Lie-Gruppe G. Diese kann lokal mit Hilfe der Exponentialabbildung

$$\exp : \mathfrak{g} \to G \quad \text{mit} \quad \exp(\widetilde{\mathbf{w}}) = \sum_{i=0}^{\infty} \frac{1}{i!} \widetilde{\mathbf{w}}^i \tag{2.26}$$

in einer Umgebung von $\mathbf{w} = \mathbf{0}$ erfolgen. Die Hintereinanderausführung von mehreren Exponentialabbildungen

$$\exp\left(h\widetilde{\mathbf{w}}_1\right) \circ \exp\left(h\widetilde{\mathbf{w}}_2\right) = \exp\left(\mathrm{BCH}\left(h\widetilde{\mathbf{w}}_1, h\widetilde{\mathbf{w}}_2\right)\right) \tag{2.27}$$

wird mit Hilfe der Baker-Campbell-Hausdorff-Formel beschrieben, dabei ist

$$\mathrm{BCH}\left(h\widetilde{\mathbf{w}}_1, h\widetilde{\mathbf{w}}_2\right) = h\widetilde{\mathbf{w}}_1 + h\widetilde{\mathbf{w}}_2 + \frac{1}{2}\left[h\widetilde{\mathbf{w}}_1, h\widetilde{\mathbf{w}}_2\right] + \mathcal{O}(h)\left\|\left[h\widetilde{\mathbf{w}}_1, h\widetilde{\mathbf{w}}_2\right]\right\| \tag{2.28}$$

für $h \to 0$ mit $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^N$ [49]. Wenn $\widetilde{\mathbf{w}}_1$ und $\widetilde{\mathbf{w}}_2$ kommutieren, dann werden die Kommutatoren gleich null, und in diesem Fall folgt $\exp\left(h\widetilde{\mathbf{w}}_1\right) \circ \exp\left(h\widetilde{\mathbf{w}}_2\right) = \exp\left(h\widetilde{\mathbf{w}}_1 + h\widetilde{\mathbf{w}}_2\right)$.
Im Verlauf der Arbeit werden numerische Lösungen von (2.25) gesucht. Dazu schlug Magnus [38] vor, eine Lösung der Form

$$q(t) = q(t_m) \circ \exp(\widetilde{\boldsymbol{\nu}}_m(t)) \tag{2.29}$$

mit einer Matrixfunktion $\widetilde{\boldsymbol{\nu}}_m(t) \in \mathfrak{g}$ zu finden. Diese genügt für t in einer Umgebung von t_m der Differentialgleichung

$$\dot{\widetilde{\boldsymbol{\nu}}}_m(t) = \mathrm{dexp}^{-1}_{-\widetilde{\boldsymbol{\nu}}_m(t)} \widetilde{\mathbf{v}}(t) \quad \text{mit} \quad \boldsymbol{\nu}_m(t_m) = \mathbf{0} \tag{2.30}$$

(Anhang B), siehe auch [33].
Hierbei beschreibt dexp^{-1} die Inverse der linksseitig trivialisierten Ableitung der Exponentialabbildung (2.26) mit der Reihenentwicklung

$$\mathrm{dexp}^{-1}_{\widetilde{\mathbf{w}}} = \sum_{i=0}^{\infty} \frac{B_i}{i!} \mathrm{ad}^i_{\widetilde{\mathbf{w}}}, \tag{2.31}$$

wobei $\mathrm{ad}^i_{\widetilde{\mathbf{w}}}$ der adjungierte Operator mit $\mathrm{ad}^0_{\widetilde{\mathbf{w}}_1} \widetilde{\mathbf{w}}_2 = \widetilde{\mathbf{w}}_2$ und $\mathrm{ad}^i_{\widetilde{\mathbf{w}}_1} \widetilde{\mathbf{w}}_2 \coloneqq \left[\widetilde{\mathbf{w}}_1, \mathrm{ad}^{i-1}_{\widetilde{\mathbf{w}}_1} \widetilde{\mathbf{w}}_2\right]$ die i-te Anwendung des Kommutators und B_i die Bernoullizahlen [30]

$$B_0 = 1, \quad B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = 0, \quad B_4 = -\frac{1}{30}, \quad B_5 = 0, \quad B_6 = \frac{1}{42}, \ldots \tag{2.32}$$

sind. Durch Integration von (2.30) und geeignetes Abschneiden der Reihe (2.31) kann eine Lösung von (2.29) erhalten werden, die Magnus-Entwicklung. Im Verlauf dieser Arbeit wird ein Resultat der Magnus-Entwicklung in einer speziellen Form von Müller [40] verwendet. Dort wurde eine Berechnungsvorschrift für die einzelnen Glieder der Reihe angegeben. Somit ist $\widetilde{\boldsymbol{\nu}}_m(t)$ in (2.29) gegeben durch

$$\begin{aligned}\widetilde{\boldsymbol{\nu}}_m(t) = {} & h\widetilde{\mathbf{v}}(t_m) + \frac{h^2}{2}\dot{\widetilde{\mathbf{v}}}(t_m) + \frac{h^3}{6}\ddot{\widetilde{\mathbf{v}}}(t_m) + \frac{h^3}{12}\left[\widetilde{\mathbf{v}}(t_m), \dot{\widetilde{\mathbf{v}}}(t_m)\right] + \frac{h^4}{24}\dddot{\widetilde{\mathbf{v}}}(t_m) \\ & + \frac{h^4}{24}\left[\widetilde{\mathbf{v}}(t_m), \ddot{\widetilde{\mathbf{v}}}(t_m)\right] + \frac{h^5}{5!}\widetilde{\boldsymbol{\mu}}_5(t_m) + \frac{h^6}{6!}\widetilde{\boldsymbol{\mu}}_6(t_m) + \frac{h^7}{7!}\widetilde{\boldsymbol{\mu}}_7(t_m) + \mathcal{O}(h^8)\end{aligned} \tag{2.33}$$

mit $h = t - t_m$. In [40] werden nur die Glieder bis zur Ordnung fünf berechnet und die Rechtstranslation anstelle der Linkstranslation (2.24) verwendet. Die hier angegebenen Reihenglieder wurden unter Verwendung der Formel aus [40] in Anhang C berechnet. Dort sind außerdem konkrete Terme für $\widetilde{\boldsymbol{\mu}}_i$ für $i = 5, 6, 7$ angegeben.
Weiterhin erlaubt die Exponentialabbildung die Konstruktion einer lokalen Parametrisierung der Lie-Gruppe G um einen gegebenen Punkt $q_m \in G$. Die Beziehung

$$q = q_m \circ \exp(\widetilde{\mathbf{v}}) \tag{2.34}$$

beschreibt lokal in einer Umgebung von q_m einen Diffeomorphismus zwischen G und $\mathfrak{g}$ bzw. zwischen G und $\mathbb{R}^N$, der als Koordinatenabbildung $\mathbb{R}^N \to G : \ \mathbf{v} \mapsto q = q_m \circ \exp(\widetilde{\mathbf{v}})$ geschrieben werden kann [9].
Um die Rechnungen im euklidischen Raum $\mathbb{R}^N$ anstatt direkt in der Lie-Algebra $\mathfrak{g}$ auszuführen, ist es sinnvoll, eine Abbildung zu definieren, die die auftretenden Kommutatoren, z.B. in (2.28) und (2.33), im $\mathbb{R}^N$ repräsentiert. Dieser lineare Operator, der einem $N \times 1$ Vektor die $N \times N$ Matrix zuordnet, ist definiert durch [7, 4]

$$\widehat{(\bullet)} : \mathbb{R}^N \to \mathbb{R}^{N\times N} \quad \text{mit} \quad \widehat{\widetilde{\mathbf{w}}_1 \mathbf{w}_2} = \mathrm{ad}_{\widetilde{\mathbf{w}}_1}(\widetilde{\mathbf{w}}_2) = [\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2] \quad \text{für alle} \quad \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^N. \tag{2.35}$$

Mit Hilfe dieser Abbildung kann dexp aus (2.30) auch als lineare Beziehung von $\mathbb{R}^N$ nach $\mathbb{R}^N$ durch

$$\mathbf{v}(t) = \mathbf{T}\big(\boldsymbol{\nu}_m(t)\big)\dot{\boldsymbol{\nu}}_m(t)$$

repräsentiert werden, wobei $\mathbf{T}\big(\boldsymbol{\nu}_m(t)\big)$ der so genannte $N \times N$ Tangentialoperator der Exponentialabbildung ist [9] mit

$$\mathrm{dexp}_{-\widetilde{\boldsymbol{\nu}}_m(t)}(\widetilde{\mathbf{w}}) = \widetilde{\mathbf{T}\big(\boldsymbol{\nu}_m(t)\big)\mathbf{w}}$$

für $\mathbf{w} \in \mathbb{R}^N$. Dieser besitzt die Reihenentwicklung

$$\mathbf{T}\big(\boldsymbol{\nu}_m(t)\big) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(i+1)!} \widehat{\boldsymbol{\nu}}_m^i(t), \tag{2.36}$$

vgl. [9, 33] und Lemma B.4.

2.2.2 Beispiele

Im folgenden Abschnitt sollen die verwendeten Matrix-Lie-Gruppen-Formulierungen näher vorgestellt werden.

Beispiel 1: $G = SO(3)$
Die spezielle orthogonale Gruppe $SO(3)$ ist definiert durch [5, 27]

$$SO(3) = \{\mathbf{R} \in \mathbb{R}^{3\times 3} |\ \mathbf{R}^\top \mathbf{R} = \mathbf{I}_3,\ \det \mathbf{R} = 1\}.$$

Die zugehörige Lie-Algebra [27]

$$\mathfrak{so}(3) = \{\widetilde{\boldsymbol{\Omega}} \in \mathbb{R}^{3\times 3} |\ \widetilde{\boldsymbol{\Omega}}^\top + \widetilde{\boldsymbol{\Omega}} = \mathbf{0}\}$$

enthält alle schiefsymmetrischen Matrizen der Form

$$\widetilde{\boldsymbol{\Omega}} = \begin{bmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{bmatrix} \tag{2.37}$$

mit $\mathbf{\Omega} = [\Omega_1\ \Omega_2\ \Omega_3]^\top \in \mathbb{R}^{3\times 1}$, siehe (2.23). Die Abbildung $\widehat{\bullet}$ aus (2.35) ist für $G = SO(3)$ identisch mit $\widetilde{\bullet}$ ([51, Lemma A.1]):

$$\widetilde{\mathbf{w}} = \widehat{\mathbf{w}},\ \mathbf{w} \in \mathbb{R}^3. \tag{2.38}$$

Die Exponentialabbildung (2.26) in $SO(3) \setminus \{0\}$ ist gegeben durch die sogenannte Rodriguez-Formel [39]

$$\exp_{SO(3)}(\mathbf{w}) = \mathbf{I}_3 + \frac{\sin(\omega)}{\omega}\widetilde{\mathbf{w}} + \frac{(1-\cos(\omega))}{\omega^2}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}$$

mit $\omega = \|\mathbf{w}\|_2$, $\mathbf{w} \in \mathbb{R}^3$. Der Tangentialoperator (2.36) ist in $SO(3)$ anhand von [7]

$$\mathbf{T}_{SO(3)}(\mathbf{w}) = \mathbf{I}_3 + \frac{\cos(\omega)-1}{\omega^2}\widetilde{\mathbf{w}} + \left(1 - \frac{\sin(\omega)}{\omega}\right)\frac{\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}}{\omega^2}$$

bestimmt.

Beispiel 2: $G = \mathbb{R}^3 \times SO(3)$
Die Beschreibung des direkten Produktes $\mathbb{R}^3 \times SO(3)$ erfolgt analog zu [7]. Ein Element der 6-dimensionalen Lie-Gruppe $\mathbb{R}^3 \times SO(3)$ kann durch das Paar $q = (\mathbf{x}, \mathbf{R})$ mit dem Translationsvektor $\mathbf{x} \in \mathbb{R}^3$ und der Rotationsmatrix $\mathbf{R} \in SO(3)$ bzw. alternativ in der Matrixschreibweise

$$q = \begin{bmatrix} \mathbf{R} & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 3} & \mathbf{I}_3 & \mathbf{x} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$$

beschrieben werden. Die Gruppenoperation $\circ$ ist gegeben durch

$$(\mathbf{x}_1, \mathbf{R}_1) \circ (\mathbf{x}_2, \mathbf{R}_2) = (\mathbf{x}_1 + \mathbf{x}_2, \mathbf{R}_1\mathbf{R}_2).$$

Die Lie-Algebra ist die Menge der 7×7-Matrizen

$$\widetilde{\mathbf{v}} = \begin{bmatrix} \widetilde{\mathbf{\Omega}} & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 3} & \mathbf{u} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & 0 \end{bmatrix} \tag{2.39}$$

mit $\mathbf{u} \in \mathbb{R}^3$ und einer schiefsymmetrischen Matrix $\widetilde{\mathbf{\Omega}}$ aus (2.37). Die Lie-Algebra kann mit dem euklidischen Raum $\mathbb{R}^6$ identifiziert werden, da die Matrix (2.39) durch einen Vektor $\mathbf{v} = \left[\mathbf{u}^\top, \mathbf{\Omega}^\top\right]^\top$ repräsentiert wird. In diesem Fall kann Gleichung (2.25) zu

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{u}, \\ \dot{\mathbf{R}} &= \mathbf{R}\widetilde{\mathbf{\Omega}} \end{aligned}$$

umformuliert werden, so dass die Geschwindigkeit $\mathbf{v}$ unterteilt ist in die Translationsgeschwindigkeit $\mathbf{u}$ im Inertialsystem und die Winkelgeschwindigkeit $\mathbf{\Omega}$ im körperfesten System. Der Operator $\widehat{\bullet}$ ist gegeben durch

$$\widehat{\mathbf{v}} = \begin{bmatrix} \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 3} \\ \mathbf{0}_{3\times 3} & \widetilde{\mathbf{\Omega}} \end{bmatrix}.$$

Schließlich berechnet sich die Exponentialabbildung (2.26) für $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^3$ zu

$$\exp_{\mathbb{R}^3 \times SO(3)}(\mathbf{w}_1, \mathbf{w}_2) = \begin{bmatrix} \exp_{SO(3)}(\mathbf{w}_2) & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 3} & \mathbf{I}_3 & \mathbf{w}_1 \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$$

und der Tangentialoperator (2.36) zu

$$\mathbf{T}_{\mathbb{R}^3 \times SO(3)}(\mathbf{w}_1, \mathbf{w}_2) = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_{3\times 3} \\ \mathbf{0}_{3\times 3} & \mathbf{T}_{SO(3)}(\mathbf{w}_2) \end{bmatrix}.$$

Beispiel 3: $G = SE(3)$
Die Beschreibung der speziellen euklidischen Gruppe $SE(3) = \mathbb{R}^3 \ltimes SO(3)$ erfolgt analog zu [7]. Ein Element der 6-dimensionalen Lie-Gruppe $SE(3)$ kann durch das Paar $q = (\mathbf{x}, \mathbf{R})$ mit dem Translationsvektor $\mathbf{x} \in \mathbb{R}^3$ und der Rotationsmatrix $\mathbf{R} \in SO(3)$ bzw. alternativ in der Matrixschreibweise

$$q = \begin{bmatrix} \mathbf{R} & \mathbf{x} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$$

beschrieben werden. Die Gruppenoperation $\circ$ ist gegeben durch

$$(\mathbf{x}_1, \mathbf{R}_1) \circ (\mathbf{x}_2, \mathbf{R}_2) = (\mathbf{x}_1 + \mathbf{R}_1\mathbf{x}_2, \mathbf{R}_1\mathbf{R}_2).$$

Die Lie-Algebra ist die Menge der 4×4-Matrizen

$$\widetilde{\mathbf{v}} = \begin{bmatrix} \widetilde{\mathbf{\Omega}} & \mathbf{U} \\ \mathbf{0}_{1\times 3} & 0 \end{bmatrix} \tag{2.40}$$

mit $\mathbf{U} \in \mathbb{R}^3$ und der schiefsymmetrischen Matrix $\widetilde{\mathbf{\Omega}}$ aus (2.37). Die Lie-Algebra kann mit $\mathbb{R}^6$ identifiziert werden, da die Matrix (2.40) durch einen Vektor $\mathbf{v} = \begin{bmatrix} \mathbf{U}^\top & \mathbf{\Omega}^\top \end{bmatrix}^\top$ repräsentiert wird. Gleichung (2.25) kann in $SE(3)$ umformuliert werden zu

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{R}\mathbf{U}, \\ \dot{\mathbf{R}} &= \mathbf{R}\widetilde{\mathbf{\Omega}}, \end{aligned}$$

so dass die Geschwindigkeit $\mathbf{v}$ unterteilt ist in die Translationsgeschwindigkeit $\mathbf{U}$ und die Winkelgeschwindigkeit $\mathbf{\Omega}$, beide im körperfesten System. Der Operator $\widehat{\bullet}$ wird zu

$$\widehat{\mathbf{v}} = \begin{bmatrix} \widetilde{\mathbf{\Omega}} & \widetilde{\mathbf{U}} \\ \mathbf{0}_{3\times 3} & \widetilde{\mathbf{\Omega}} \end{bmatrix}. \tag{2.41}$$

Schließlich berechnet sich die Exponentialabbildung (2.26) für $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^3$ zu

$$\exp_{SE(3)}(\mathbf{w}_1, \mathbf{w}_2) = \begin{bmatrix} \exp_{SO(3)}(\mathbf{w}_2) & \mathbf{A}(\mathbf{w}_2)\mathbf{w}_1 \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$$

mit

$$\mathbf{A}(\mathbf{w}_2) \coloneqq \frac{1}{\omega^2}\left(\omega^2\mathbf{I}_3 + (\mathbf{I}_3 - \exp_{SO(3)}(\mathbf{w}_2))\widetilde{\mathbf{w}}_2 + \widetilde{\mathbf{w}}_2\widetilde{\mathbf{w}}_2\right),$$

wobei $\omega = \|\mathbf{w}_2\|_2$ ist, und der Tangentialoperator (2.36) zu

$$\mathbf{T}_{SE(3)}(\mathbf{w}_1, \mathbf{w}_2) = \begin{bmatrix} \mathbf{T}_{SO(3)}(\mathbf{w}_2) & \mathbf{C}(\mathbf{w}_1, \mathbf{w}_2) - \frac{1}{2}\widetilde{\mathbf{w}}_1 \\ \mathbf{0}_{3\times 3} & \mathbf{T}_{SO(3)}(\mathbf{w}_2) \end{bmatrix}$$

mit

$$\begin{aligned}\mathbf{C}(\mathbf{w}_1, \mathbf{w}_2) = \frac{1-a}{2}\widetilde{\mathbf{w}}_1 + \frac{1-b}{\omega^2}(\widetilde{\mathbf{w}}_1\widetilde{\mathbf{w}}_2 + \widetilde{\mathbf{w}}_2\widetilde{\mathbf{w}}_1) - \frac{b-a}{\omega^2}\mathbf{w}_2^\top\mathbf{w}_1\widetilde{\mathbf{w}}_2 \\ + \frac{1}{\omega^2}\left(\frac{a}{2} - \frac{3(1-b)}{\omega^2}\right)\mathbf{w}_2^\top\mathbf{w}_1\widetilde{\mathbf{w}}_2\widetilde{\mathbf{w}}_2,\end{aligned}$$

wobei $a = 2(1-\cos(\omega))/\omega^2$ und $b = \sin(\omega)/\omega$ gilt.

Kapitel 3

Zeitintegrationsverfahren für mechanische Systeme auf Lie-Gruppen

Das nachfolgende Kapitel behandelt die Zeitintegration der Bewegungsgleichungen von mechanischen Systemen ohne und mit Zwangsbedingungen für Konfigurationsräume mit Lie-Gruppen-Struktur. Zunächst werden Zeitintegrationsverfahren für gewöhnliche Differentialgleichungen auf Lie-Gruppen thematisiert. Die vorgestellten Lie-Gruppen-Verfahren werden auf mechanischen Systeme ohne Zwangsbedingungen angewendet. Außerdem wird ein neues BDF-Lie-Gruppen-Verfahren vorgestellt, das BLieDF-Verfahren, das eine effizientere Auswertung der benötigten Lie-Klammern bzw. Matrix-Kommutatoren verwendet als das Munthe-Kaas-BDF-Lie-Gruppen-Verfahren.
Zudem werden mechanische Systeme mit Zwangsbedingungen, die durch differential-algebraische Gleichungen vom Index 3 beschrieben werden, thematisiert. Für diese Aufgabenklasse werden nur drei der vorgestellten Verfahren genauer untersucht, die beiden BDF-Verfahren und das Generalized-α-Verfahren.

3.1 Ausgewählte Zeitintegrationsverfahren für gewöhnliche Differentialgleichungen auf Lie-Gruppen

Bei der Anwendung eines numerischen Integrators auf Bewegungsgleichungen in Konfigurationsräumen mit Lie-Gruppen-Struktur muss die kinematische Gleichung (2.3a) durch (2.25) ersetzt werden. Somit wird eine Lösung des Anfangswertproblems

$$\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\mathbf{v}}(t), \tag{3.1a}$$
$$\dot{\mathbf{v}}(t) = \mathbf{f}(t, q(t), \mathbf{v}(t)), \tag{3.1b}$$
$$q(t_0) = q_0, \tag{3.1c}$$
$$\mathbf{v}(t_0) = \mathbf{v}_0 \tag{3.1d}$$

gesucht.
Im folgenden Abschnitt werden verschiedene aus der Literatur bekannte Verfahren zur Lösung von (3.1) vorgestellt. Crouch und Grossman [16] haben explizite Mehrschritt-

Lie-Gruppen-Verfahren eingeführt, die eine Vielzahl von Auswertungen der Exponentialabbildungen (2.26) erfordern, die die Elemente der Lie-Gruppe darstellen. In dieser Arbeit erfolgt eine Berechnung von Verfahrensparametern auch für implizite Mehrschritt-Lie-Gruppen-Verfahren.
Da die Berechnung der Exponentialfunktion rechentechnisch aufwendig sein kann, sind Verfahren in den Fokus gerückt, die den Großteil der Rechnungen in der Lie-Algebra ausführen. Dadurch wird nur eine Exponentialabbildung benötigt, im Gegenzug werden Kommutatoren ausgewertet. Mehrschrittverfahren dieser Art wurde von Faltinsen, Marthinsen und Munthe-Kaas [20] vorgestellt. Hier ist eine Umparametrisierung in der Lie-Gruppe nötig, um die gewünschte Konvergenzordnung des Verfahrens zu erhalten. Das Verfahren aus [20] wird in dieser Arbeit jedoch unter Verwendung der $\widehat{\bullet}$-Abbildung modifiziert, so dass die Rechnungen im euklidischen Raum $\mathbb{R}^N$ und nicht in der Lie-Algebra $\mathfrak{g}$ erfolgen.
Ein weiteres Verfahren geht auf Celledoni, Marthinsen und Owren [13] zurück. Es wird versucht, möglichst wenige Funktionswerte der Exponentialabbildung zu berechnen und gleichzeitig keine Matrix-Kommutatoren zu verwenden. In [13] wird diese Strategie auf Einschrittverfahren (Runge-Kutta-Verfahren) angewendet. In dieser Arbeit wird diese Idee aufgegriffen, und es werden implizite Mehrschrittverfahren mit entsprechenden Verfahrensparametern eingeführt.
Welche der genannten Varianten eingesetzt werden sollten, hängt stets von der untersuchten Anwendung ab und wird an dieser Stelle nicht vertieft werden.
Als letztes Verfahren wird das Generalized-α-Verfahren auf (3.1) angewendet, dessen Lie-Gruppen-Formulierung auf Brüls und Cardona [8] zurückgeht.

3.1.1 Crouch-Grossman-Verfahren

Crouch und Grossman führten in [16] sowohl Einschritt- als auch Mehrschrittverfahren zur Lösung von Differentialgleichungen auf Mannigfaltigkeiten ein. In der vorliegenden Arbeit werden vorrangig Mehrschrittverfahren untersucht, weshalb an dieser Stelle nur jene vorgestellt werden.
Weiterhin betrachteten Crouch und Grossman lediglich Mehrschrittverfahren der Form

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h \sum_{i=0}^{k} \beta_i \mathbf{f}(t_{n+1-i}, \mathbf{x}_{n+1-i}, \mathbf{v}_{n+1-i}) \tag{3.2}$$

mit Parametern $\beta_i \in \mathbb{R}$, $(i = 0, ..., k)$, vgl. Gleichung (A.3a). Es können also nur Verfahren verwendet werden, für die $\alpha_i = 0$, $(i \geq 2)$, gilt. BDF-Verfahren (2.5) sind somit in dieser Art von Verfahren nicht enthalten, Adams-Moulton-Verfahren (2.17) hingegen schon. Aus diesem Grund wird für die Crouch-Grossman-Verfahren nur die Adams-Moulton-Variante untersucht. Für die Anwendung auf (3.1) werden $\mathbf{x}_n$ und $\mathbf{x}_{n+1}$ aus $\mathbb{R}^N$ durch Elemente q_n und q_{n+1} aus der Lie-Gruppe G ersetzt. Die Summe (inklusive Summenzeichen) wird durch die Gruppenoperation substituiert, und die rechte Seite $\mathbf{f}$ soll die Elemente aus der Lie-Algebra $\mathfrak{g}$ darstellen. Hier zeigt sich auch die Notwendigkeit der Einschränkung an die Verfahrensparameter $\alpha_i = 0$, $(i \geq 2)$. Da durch die Umwandlung in die Lie-Gruppen-Elemente keine Skalierung dieser Lie-Gruppen-Elemente existiert, würde z.B. für $k = 2$ ein Term der Form $(\alpha_0 q_{n+1}) \circ (\alpha_1 q_n) \circ (\alpha_2 q_{n-1})$ nicht definiert werden können und deshalb können nur Verfahren der Form (3.2) untersucht werden.
In Konfigurationsräumen mit Lie-Gruppen-Struktur müssen zusätzliche Terme, die Matrix-Kommutatoren enthalten, in die Herleitung der Verfahren mit einbezogen

werden. Aus diesem Grund existieren zusätzliche Bedingungen, die das Verfahren erfüllen muss, um die selbe Konsistenz- und Konvergenzordnung wie (3.2) zu erhalten. Deshalb reichen die Verfahrensparameter β_i für Konfigurationsräume mit Lie-Gruppen-Struktur nicht aus und es werden Parameter β_i^j mit

$$\beta_i = \sum_{j=0}^{m-1} \beta_i^j, \quad (i = 0, ..., k,\ m \in \mathbb{N}), \tag{3.3}$$

eingeführt. Das Mehrschrittverfahren für die Geschwindigkeiten $\mathbf{v}$ kann die allgemeine Form (A.3a) beibehalten. Da für beide Variablen jedoch das selbe Verfahren verwendet werden soll, wird auch hier die Variante (3.2) eingesetzt. Zusammengefasst ergibt sich das folgende Verfahren.

Definition 9 (Crouch-Grossman-Verfahren [16, für ODEs 1. Ordnung])
Ein *Crouch-Grossman-Verfahren* für den Zeitschritt $t_n \to t_{n+1} = t_n + h$ zur Lösung von (3.1) ist gegeben durch

$$\begin{aligned} u_n^j &= u_n^{j+1} \circ \exp(h\beta_k^j \widetilde{\mathbf{v}}_{n+1-k}) \circ ... \circ \exp(h\beta_1^j \widetilde{\mathbf{v}}_n) \\ &\quad \circ \exp(h\beta_0^j \widetilde{\mathbf{v}}_{n+1}), \quad (j = 0, ..., m-1), && \text{(3.4a)} \\ u_n^m &= q_n, && \text{(3.4b)} \\ q_{n+1} &= u_n^0, && \text{(3.4c)} \\ \mathbf{v}_{n+1} &= \mathbf{v}_n + h \sum_{i=0}^{k} \beta_i \dot{\mathbf{v}}_{n+1-i}, && \text{(3.4d)} \\ \dot{\mathbf{v}}_{n+1-i} &= \mathbf{f}(t_{n+1-i}, q_{n+1-i}, \mathbf{v}_{n+1-i}), \quad (i = 0, ..., k). && \text{(3.4e)} \end{aligned}$$

Zu den vorgegebenen numerischen Lösungen q_{n+1-i} und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, werden dabei die Approximationen $q_{n+1} \approx q(t_{n+1})$ und $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ berechnet.

Crouch und Grossman haben Bedingungen für die Parameter β_i^j untersucht, um ein explizites Verfahren der Konvergenzordnung drei zu erhalten. In dieser Arbeit werden vor allem implizite Verfahren untersucht. Aus diesem Grund wurde die Vorgehensweise von [16] aufgegriffen und Parameter für implizite Crouch-Grossman-Verfahren bis zu einer Konvergenz der Ordnung vier bestimmt. Dabei wurden die Parameter stets so gewählt, dass durch (3.3) das Adams-Moulton-Verfahren (2.17) der entsprechenden Ordnung erhalten wird.
Zur Berechnung der Verfahrensparameter kann wie folgt vorgegangen werden. Ziel ist es, die Parameter in Definition 9 so zu bestimmen, dass

$$q(t_n) \circ \exp(h\beta_k^{m-1} \widetilde{\mathbf{v}}(t_{n+1-k})) \circ ... \circ \exp(h\beta_0^0 \widetilde{\mathbf{v}}(t_{n+1})) = q(t_n) \circ \exp(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})) + \mathcal{O}(h^{k+2}) \tag{3.5}$$

gilt (vgl. (2.29)). Dann besitzt der lokale Abbruchfehler (vgl. Definition A.3) die benötigte Ordnung, um ein Verfahren der Konvergenzordnung $p = k + 1$ zu erhalten. Die Hintereinanderausführungen der Exponentialabbildungen werden anhand der Baker-Campbell-Hausdorff-Formel (2.28) bestimmt. Das Argument dieser kombinierten Exponentialabbildungen wird anschließend mit $\widetilde{\boldsymbol{\nu}}_n(t_{n+1})$ aus (2.33) verglichen. Werden zudem alle Geschwindigkeiten $\mathbf{v}$ und deren Ableitungen durch Taylorentwicklung auf den Zeitpunkt t_n zurückgeführt, so entsteht ein Term der Form

$$\begin{aligned} &C_1 h\widetilde{\mathbf{v}}(t_n) + C_2 h^2 \widetilde{\mathbf{v}}^{(2)}(t_n) + C_3 h^3 \widetilde{\mathbf{v}}^{(3)}(t_n) + C_4 h^4 \widetilde{\mathbf{v}}^{(4)}(t_n) + C_5 h^3 [\widetilde{\mathbf{v}}(t_n), \dot{\widetilde{\mathbf{v}}}(t_n)] \\ &+ C_6 h^4 [\widetilde{\mathbf{v}}(t_n), \ddot{\widetilde{\mathbf{v}}}(t_n)] + C_7 h^4 [\widetilde{\mathbf{v}}(t_n), [\widetilde{\mathbf{v}}(t_n), \dot{\widetilde{\mathbf{v}}}(t_n)]] + \mathcal{O}(h^5), \end{aligned} \tag{3.6}$$

wobei die $C_\ell \in \mathbb{R}$, $(\ell = 1, ..., 7)$, von den Verfahrensparametern β_i^j abhängen. Wird nun für alle Terme, die nicht in $\mathcal{O}(h^{k+2})$ liegen, $C_\ell = 0$ gesetzt, lassen sich Konsistenzbedingungen angeben. Die Konsistenzbedingungen, die sich durch $C_\ell = 0$, $(\ell = 1, 2, 3, 4)$, herleiten lassen, stimmen mit denen aus (A.7) überein, die auch für Verfahren in linearen Räumen existieren. Die anderen sind zusätzliche Bedingungen, die nur für die Lie-Gruppen-Variante der Verfahren existieren. Für $k = 0, 1$ gibt es keine, für $k = 2$ eine und für $k = 3$ drei zusätzliche Bedingungen. Entsprechend werden insgesamt für $k = 0$ ein Verfahrensparameter, für $k = 1$ zwei, für $k = 2$ vier und für $k = 3$ sieben Verfahrensparameter benötigt. Aus Effizienzgründen beim später verwendeten Newton-Verfahren wurde darauf geachtet, dass stets nur ein Parameter β_0^j, $(j = 0, ..., m-1)$, vorhanden ist. Um schließlich die Verfahrensparameter zu bestimmen, werden die Konsistenzbedingungen unter Einbeziehung der Parameter

$$\begin{aligned}
k = 0: &\quad \beta_0^0, \quad (m = 1, k = 0),\\
k = 1: &\quad \beta_0^0,\ \beta_1^0, \quad (m = 1, k = 1),\\
k = 2: &\quad \beta_0^0,\ \beta_1^0,\ \beta_2^0,\ \beta_1^1, \quad (m = 2, k = 2),\\
k = 3: &\quad \beta_0^0,\ \beta_1^0,\ \beta_2^0,\ \beta_3^0,\ \beta_1^1,\ \beta_2^1,\ \beta_1^2, \quad (m = 3, k = 3),
\end{aligned}$$

gelöst. Im Fall $k = 3$ wurde β_1^2 an Stelle von β_3^1 verwendet, da die Lösung der Konsistenzbedingungen ansonsten komplexe Parameter ergeben hätte. Dadurch ergeben sich die zu betrachtenden Verfahrensparameter

$$\begin{aligned}
p = 1: &\quad \beta_0^0 = \beta_0 = 1,\\
p = 2: &\quad \beta_0^0 = \beta_0 = \frac{1}{2},\ \beta_1^0 = \beta_1 = \frac{1}{2},\\
p = 3: &\quad \beta_0^0 = \frac{5}{12},\ \beta_1^0 = \frac{7}{12},\ \beta_2^0 = -\frac{1}{12},\ \beta_1^1 = \frac{1}{12},\\
p = 4: &\quad \beta_0^0 = \frac{3}{8},\ \beta_1^0 = \frac{923 + \sqrt{42351}}{1176},\ \beta_2^0 = \frac{155 - \sqrt{42351}}{1176},\ \beta_3^0 = \frac{1}{24},\\
&\quad \beta_1^1 = \frac{-204 - \sqrt{42351}}{1176},\ \beta_2^1 = \frac{-400 + \sqrt{42351}}{1176},\ \beta_1^2 = \frac{53}{294}.
\end{aligned}$$

Theoretisch könnten durch gleiches Vorgehen auch die Parameter für höhere Ordnungen bestimmt werden, worauf in dieser Arbeit jedoch verzichtet wird. Die numerischen Tests aus Kapitel 6 bestätigen eine Konvergenzordnung von $p = k + 1$.

3.1.2 Kommutatorfreie Lie-Gruppen-Verfahren

Die in diesem Abschnitt vorgestellten Verfahren greifen die Ideen von Crouch und Grossman (Abschnitt 3.1.1) auf, doch sie benötigen aufgrund von Linearkombinationen innerhalb der Exponentialabbildungen weniger Auswertungen dieser Exponentialabbildungen. In [13] wurde diese Strategie für Runge-Kutta-Verfahren eingeführt und getestet. Da in der vorliegenden Arbeit vorrangig implizite Mehrschrittverfahren von Interesse sind, wird das Prinzip auf diese angewendet. Aus demselben Grund wie zuvor bei den Crouch-Grossman-Verfahren können nur Mehrschrittverfahren der Form (3.2) in ein kommutatorfreies Lie-Gruppen-Verfahren umgewandelt werden. Auch hier sind somit keine kommutatorfreien BDF-Lie-Gruppen-Mehrschrittverfahren möglich.

Definition 10 (Kommutatorfreie Lie-Gruppen-Mehrschrittverfahren)
Ein *kommutatorfreies Lie-Gruppen-Mehrschrittverfahren* für den Zeitschritt $t_n \to$

$t_{n+1} = t_n + h$ mit Schrittweite h zur Lösung von (3.1) ist gegeben durch

$$q_{n+1} = q_n \circ \exp\left(h \sum_{j=0}^{k} \beta_j^{m-1} \widetilde{\mathbf{v}}_{n+1-j}\right) \circ ... \circ \exp\left(h \sum_{j=0}^{k} \beta_j^0 \widetilde{\mathbf{v}}_{n+1-j}\right), \tag{3.7a}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + h \sum_{i=0}^{k} \beta_i \dot{\mathbf{v}}_{n+1-i}, \tag{3.7b}$$

$$\dot{\mathbf{v}}_{n+1-i} = \mathbf{f}(t_{n+1-i}, q_{n+1-i}, \mathbf{v}_{n+1-i}), \quad (i = 0, ..., k). \tag{3.7c}$$

Zu den vorgegebenen numerischen Lösungen q_{n+1-i} und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, werden dabei die Approximationen $q_{n+1} \approx q(t_{n+1})$ und $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ berechnet.

Ziel ist es, die Parameter in Definition 10 so zu bestimmen, dass

$$\begin{aligned} &q(t_n) \circ \exp\left(h \sum_{j=0}^{k} \beta_j^{m-1} \widetilde{\mathbf{v}}(t_{n+1-j})\right) \circ ... \circ \exp\left(h \sum_{j=0}^{k} \beta_j^0 \widetilde{\mathbf{v}}(t_{n+1-j})\right) \\ &= q(t_n) \circ \exp\left(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\right) + \mathcal{O}(h^{k+2}) \end{aligned} \tag{3.8}$$

gilt (vgl. (2.29)). Dann besitzt der lokale Abbruchfehler (vgl. Definition A.3) die benötigte Konsistenzordnung, um ein Verfahren der Konvergenzordnung $p = k + 1$ zu erhalten. Die weitere Vorgehensweise ist analog zum Crouch-Grossman-Verfahren (Definition 9), lediglich die Baker-Campbell-Hausdorff-Formel wird seltener benötigt (vgl. (3.5) und (3.8)).
Durch die Hintereinanderausführungen der Exponentialabbildungen anhand der Baker-Campbell-Hausdorff-Formel (2.28) und den Vergleich mit $\widetilde{\boldsymbol{\nu}}_n(t_{n+1})$ aus (2.33) entsteht erneut ein Term der Form (3.6). Die Konsistenzbedingungen ergeben sich, indem für alle Terme, die nicht in $\mathcal{O}(h^{k+2})$ liegen, $C_\ell = 0$ gesetzt wird. Die Konsistenzbedingungen, die sich durch $C_\ell = 0$, $(\ell = 1, 2, 3, 4)$, herleiten lassen, stimmen mit denen aus (A.7) überein, die auch für Verfahren in linearen Räumen existieren. Die anderen sind zusätzliche Bedingungen, die nur in der Lie-Gruppen-Variante der Verfahren auftreten. Für $k = 0, 1$ gibt es keine, für $k = 2$ eine und für $k = 3$ drei zusätzliche Bedingungen. Jedoch zeigt sich, dass $C_7 = 0$ automatisch erfüllt ist, wenn $\beta_0^1 = 0$ und $C_\ell = 0$, $(i = 1, ..., 6)$, gelten. Daher werden für $k = 0$ ein Verfahrensparameter, für $k = 1$ zwei, für $k = 2$ vier und für $k = 3$ sechs Verfahrensparameter benötigt. Um schließlich diese Verfahrensparameter zu bestimmen, werden die Konsistenzbedingungen unter Einbeziehung der Parameter

$$\begin{aligned} k = 0: &\quad \beta_0^0, \quad (m = 1, k = 0), \\ k = 1: &\quad \beta_0^0,\ \beta_1^0, \quad (m = 1, k = 1), \\ k = 2: &\quad \beta_0^0,\ \beta_1^0,\ \beta_2^0,\ \beta_1^1, \quad (m = 2, k = 2), \\ k = 3: &\quad \beta_0^0,\ \beta_1^0,\ \beta_2^0,\ \beta_1^1,\ \beta_2^1,\ \beta_3^1, \quad (m = 2, k = 3), \end{aligned}$$

gelöst. Daher sind die zu verwendenden Verfahren von erster bis vierter Ordnung durch die Parameter

$$\begin{aligned} p = 1: &\quad \beta_0^0 = \beta_0 = 1, \quad (m = 1, k = 0), \\ p = 2: &\quad \beta_0^0 = \frac{1}{2},\ \beta_1^0 = \frac{1}{2}, \quad (m = 1, k = 1), \\ p = 3: &\quad \beta_0^0 = \frac{5}{12},\ \beta_1^0 = \frac{1}{3},\ \beta_2^0 = -\frac{1}{12},\ \beta_1^1 = \frac{1}{3}, \quad (m = 2, k = 2), \\ p = 4: &\quad \beta_0^0 = \frac{3}{8},\ \beta_1^0 = \frac{1}{6},\ \beta_2^0 = -\frac{1}{24},\ \beta_1^1 = \frac{5}{8},\ \beta_2^1 = -\frac{1}{6},\ \beta_3^1 = \frac{1}{24},\ (m = 2, k = 3), \end{aligned}$$

und (3.3) definiert. Alle restlichen Parameter sind null. Es fällt auf, dass das Crouch-Grossman-Verfahren (3.4) und das kommutatorfreie Lie-Gruppen-Mehrschrittverfahren von Ordnung eins identisch sind. Theoretisch könnten durch gleiches Vorgehen auch die Parameter für höhere Ordnungen bestimmt werden, dies soll in dieser Arbeit jedoch nicht vertieft werden. Die numerischen Tests aus Kapitel 6 bestätigen die Konvergenzordnung von $p = k + 1$.

3.1.3 Munthe-Kaas-Mehrschrittverfahren

Im Gegensatz zu den Crouch-Grossman-Mehrschrittverfahren in Abschnitt 3.1.1 berechnen die Munthe-Kaas-Mehrschrittverfahren (Name analog zu Runge-Kutta-Munthe-Kaas-Verfahren [42, 43]) die numerische Lösung in jedem Integrationsschritt zunächst in der Lie-Algebra $\mathfrak{g}$ und transformieren das Ergebnis durch die einmalige Anwendung der Exponentialabbildung in ein Element der Lie-Gruppe G. Dieses Verfahren wurde in [20] vorgestellt. Die Idee ist es, in jedem Zeitschritt die Differentialgleichung (2.30) im linearen Raum $\mathfrak{g}$ zu betrachten und sie numerisch durch klassische Ansätze zu lösen. Dabei repräsentiert $\operatorname{dexp}_{-\widetilde{\boldsymbol{\nu}}_m}(t)$ die Jacobi-Matrix der Koordinatenabbildung (2.34). Da $\operatorname{dexp}_{\mathbf{0}} = \mathbf{I}$ ist, ist die Abbildung in der Nähe von q_m gut konditioniert. Je mehr sich von dem Punkt q_m entfernt wird, um so schlechter konditioniert ist das Problem [20]. In einem Zeitschritt ist es daher sinnvoll, zur Berechnung von q_{n+1} durch (2.34) den Punkt $q_m := q_n$ als nächstbekannten Punkt aus der Lie-Gruppe G zu verwenden. Daher muss das Zentrum der Koordinatenabbildung in jedem Zeitschritt neu gewählt und die bekannten Informationen in das neue Koordinatensystem transformiert werden. Dazu seien $\widetilde{\boldsymbol{\omega}}_i^{(n)} \in \mathfrak{g}$ lokale Koordinaten von $q_{n+1-i} \in G$ in Bezug auf $q_n \in G$, so dass

$$q_{n+1-i} = q_n \circ \exp\left(\widetilde{\boldsymbol{\omega}}_i^{(n)}\right), \quad (i = 0, 1, ..., k), \tag{3.9}$$

ist. Das lineare k-Schrittverfahren (A.3a) für (3.1a) ist dann gegeben durch

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \widetilde{\boldsymbol{\omega}}_i^{(n)} = \sum_{i=0}^{k} \beta_i \operatorname{dexp}_k^{-1}\big(-\widetilde{\boldsymbol{\omega}}_i^{(n)}, \widetilde{\mathbf{v}}_{n+1-i}\big), \tag{3.10}$$

wobei $\operatorname{dexp}_k^{-1}\big(\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2\big)$ eine Approximation k-ter Ordnung der Abbildung $\operatorname{dexp}_{\widetilde{\mathbf{w}}_1}^{-1}\widetilde{\mathbf{w}}_2$ (vgl. (2.31)) ist, die lipschitzstetig mit einer Konstanten der Größe $\mathcal{O}(1)$ sein soll. Um diese Gleichung im euklidischen Raum darzustellen, kann der $\widehat{\bullet}$-Operator aus (2.35) verwendet werden. Dafür wird $\widehat{\operatorname{dexp}_k}^{-1}(\mathbf{w}_1, \mathbf{w}_2)$ definiert durch

$$\widehat{\operatorname{dexp}_1}^{-1}(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \mathbf{w}_2, \tag{3.11a}$$

$$\widehat{\operatorname{dexp}_2}^{-1}(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \mathbf{w}_2, \tag{3.11b}$$

$$\widehat{\operatorname{dexp}_3}^{-1}(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\operatorname{dexp}_2}^{-1}(\mathbf{w}_1, \mathbf{w}_2) - \frac{1}{2}\widehat{\mathbf{w}}_1\mathbf{w}_2, \tag{3.11c}$$

$$\widehat{\operatorname{dexp}_4}^{-1}(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\operatorname{dexp}_3}^{-1}(\mathbf{w}_1, \mathbf{w}_2) + \frac{1}{12}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2, \tag{3.11d}$$

$$\widehat{\operatorname{dexp}_5}^{-1}(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\operatorname{dexp}_4}^{-1}(\mathbf{w}_1, \mathbf{w}_2), \tag{3.11e}$$

$$\widehat{\operatorname{dexp}_6}^{-1}(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\operatorname{dexp}_5}^{-1}(\mathbf{w}_1, \mathbf{w}_2) - \frac{1}{720}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2 \tag{3.11f}$$

mit $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^N$ und es gilt $\operatorname{dexp}_k^{-1}(\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2) = \widetilde{\operatorname{dexp}_k^{-1}}(\mathbf{w}_1, \mathbf{w}_2)$. Daher ist Gleichung (3.10) äquivalent zu

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\omega}_i^{(n)} = \sum_{i=0}^{k} \beta_i \widehat{\operatorname{dexp}_k}^{-1}\big(-\boldsymbol{\omega}_i^{(n)}, \mathbf{v}_{n+1-i}\big).$$

Die Frage nach der Berechnung der aktuellen $\widetilde{\boldsymbol{\omega}}_i^{(n)}$ bleibt jedoch noch zu beantworten. Dafür zeigt der Vergleich von (3.9) mit (2.29) für $m = n$ und $t = t_{n+1-i}$, dass $\widetilde{\boldsymbol{\omega}}_i^{(n)} \approx \widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})$ gilt, da $q_n \approx q(t_n)$ und $q_{n+1-i} \approx q(t_{n+1-i})$ sind. Durch den Vergleich von (2.29) für $m = n$ mit $m = n-1$ für $t = t_{n+1-i}$ und $i = 1, ..., k$ folgt, dass

$$q(t_n) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})\big) = q(t_{n+1-i}) = q(t_{n-1}) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_{n-1}(t_{n+1-i})\big)$$

erfüllt ist und darum auch

$$\begin{aligned}\exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})\big) &= q(t_n)^{-1} \circ q(t_{n-1}) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_{n-1}(t_{n+1-i})\big) \\ &= \exp\big(-\widetilde{\boldsymbol{\nu}}_{n-1}(t_n)\big) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_{n-1}(t_{n+1-i})\big).\end{aligned}$$

Im Zeitschritt $t_n \to t_{n+1}$ sind Approximationen $-\widetilde{\boldsymbol{\omega}}_0^{(n-1)} \approx -\widetilde{\boldsymbol{\nu}}_{n-1}(t_n)$ und $\widetilde{\boldsymbol{\omega}}_{i-1}^{(n-1)} \approx \widetilde{\boldsymbol{\nu}}_{n-1}(t_{n+1-i})$ bekannt und somit können $\widetilde{\boldsymbol{\omega}}_i^{(n)} \approx \widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})$ berechnet werden durch

$$\exp\big(\widetilde{\boldsymbol{\omega}}_i^{(n)}\big) \approx \exp\big(-\widetilde{\boldsymbol{\omega}}_0^{(n-1)}\big) \circ \exp\big(\widetilde{\boldsymbol{\omega}}_{i-1}^{(n-1)}\big), \quad (i = 1, ..., k).$$

Die Kombination von zwei Exponentialabbildungen wird durch Approximationen der Baker-Campbell-Hausdorff-Formel BCH (2.28) berechnet. Praktisch wird eine Approximation k-ter Ordnung BCH_k verwendet, die lipschitzstetig mit einer Konstanten der Größe $\mathcal{O}(1)$ ist. Es folgt

$$\widetilde{\boldsymbol{\omega}}_i^{(n)} = \mathrm{BCH}_k\big(-\widetilde{\boldsymbol{\omega}}_0^{(n-1)}, \widetilde{\boldsymbol{\omega}}_{i-1}^{(n-1)}\big), \quad (i = 1, ..., k). \tag{3.12}$$

Zur Darstellung dieser Gleichung im euklidischen Raum wird erneut der $\widehat{\bullet}$-Operator verwendet und es können Approximationen (vgl. [44])

$$\widehat{\mathrm{BCH}}_2(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \mathbf{w}_1 + \mathbf{w}_2, \tag{3.13a}$$

$$\widehat{\mathrm{BCH}}_3(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\mathrm{BCH}}_2(\mathbf{w}_1, \mathbf{w}_2) + \frac{1}{2}\widehat{\mathbf{w}}_1\mathbf{w}_2, \tag{3.13b}$$

$$\widehat{\mathrm{BCH}}_4(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\mathrm{BCH}}_3(\mathbf{w}_1, \mathbf{w}_2) + \frac{1}{12}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2 - \frac{1}{12}\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_1\mathbf{w}_2, \tag{3.13c}$$

$$\widehat{\mathrm{BCH}}_5(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\mathrm{BCH}}_4(\mathbf{w}_1, \mathbf{w}_2) - \frac{1}{24}\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2, \tag{3.13d}$$

$$\begin{aligned}\widehat{\mathrm{BCH}}_6(\mathbf{w}_1, \mathbf{w}_2) \coloneqq \widehat{\mathrm{BCH}}_5(\mathbf{w}_1, \mathbf{w}_2) &- \frac{1}{720}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2 + \frac{1}{720}\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_1\mathbf{w}_2 \\ &- \frac{1}{180}\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2 + \frac{1}{180}\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_2\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2 - \frac{1}{120}\widehat{\widehat{\mathbf{w}}_1\mathbf{w}_2}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2\end{aligned} \tag{3.13e}$$

mit $\mathrm{BCH}_k(\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2) = \widetilde{\widehat{\mathrm{BCH}}_k}(\mathbf{w}_1, \mathbf{w}_2)$ definiert werden. Gleichung (3.12) ist damit äquivalent zu

$$\boldsymbol{\omega}_i^{(n)} = \widehat{\mathrm{BCH}}_k\big(-\boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_{i-1}^{(n-1)}\big), \quad (i = 1, ..., k).$$

Definition 11 (k-Schritt Munthe-Kaas-Verfahren [20, für ODEs 1. Ordnung])
Ein *k-Schritt Munthe-Kaas-Verfahren* zur Lösung von (3.1) für den Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit Schrittweite h ist gegeben durch

$$q_{n+1} = q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big), \tag{3.14a}$$

$$\boldsymbol{\omega}_i^{(n)} = \widehat{\mathrm{BCH}}_k\big(-\boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_{i-1}^{(n-1)}\big), \quad (i = 1, ..., k), \tag{3.14b}$$

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\omega}_i^{(n)} = \sum_{i=0}^{k} \beta_i \widehat{\mathrm{dexp}_k}^{-1}\big(-\boldsymbol{\omega}_i^{(n)}, \mathbf{v}_{n+1-i}\big), \tag{3.14c}$$

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i} = \sum_{i=0}^{k} \beta_i \mathbf{f}(t_{n+1-i}, q_{n+1-i}, \mathbf{v}_{n+1-i}). \tag{3.14d}$$

Dabei werden die vorgegebenen numerischen Lösungen q_{n+1-i}, $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, und $\boldsymbol{\omega}_i^{(n-1)}$, $(i = 0, ..., k)$ verwendet, um Approximationen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ und $\boldsymbol{\omega}_i^{(n)} \approx \boldsymbol{\nu}_n(t_{n+1-i})$, $(i = 0, ..., k-1)$, zu berechnen.

In [20] wurden jedoch statt (3.14b) und (3.14c) die hierzu äquivalenten Gleichungen (3.10) und (3.12) verwendet. Im Unterschied zu den Crouch-Grossman-Verfahren und den kommutatorfreien Lie-Gruppen-Verfahren können auf die k-Schritt Munthe-Kaas-Verfahren alle linearen Mehrschrittverfahren (A.3a) und nicht nur die der Form (3.2) angewendet werden. Dies ist ein deutlicher Vorteil der Munthe-Kaas-Verfahren. Da der Fokus dieser Arbeit auf BDF-Verfahren liegt, ist nachfolgend das Verfahren (3.14) mit Parametern α_i und β_i wie für die BDF-Verfahren (2.5) separat aufgeführt.

Definition 12 (k-Schritt Munthe-Kaas-BDF-Verfahren [20])
Ein *k-Schritt Munthe-Kaas-BDF-Verfahren* für $k \leq 6$ zur Lösung von (3.1) für den Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit Schrittweite h ist gegeben durch

$$\begin{aligned}
q_{n+1} &= q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big),\\
\boldsymbol{\omega}_i^{(n)} &= \widehat{\mathrm{BCH}}_k\big(-\boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_{i-1}^{(n-1)}\big), \quad (i = 1, ..., k),\\
\frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\omega}_i^{(n)} &= \widehat{\mathrm{dexp}_k}^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big),\\
\frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i} &= \mathbf{f}(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1})
\end{aligned}$$

mit Parametern α_i aus (2.7) und den Approximationen (3.11) und (3.13) der Abbildungen dexp^{-1} aus (2.31) und BCH aus (2.28). Dabei werden die numerischen Lösungen q_{n+1-i}, $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, und $\boldsymbol{\omega}_i^{(n-1)}$, $(i = 0, ..., k-1)$ verwendet, um die Approximationen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ und $\boldsymbol{\omega}_i^{(n)} \approx \boldsymbol{\nu}_n(t_{n+1-i})$, $(i = 0, ..., k)$, zu berechnen.

Die Konvergenzordnung ist wie bei den BDF-Verfahren (2.5) durch $p = k$ gegeben [20].

3.1.4 Generalized-α-Verfahren

Das Generalized-α-Verfahren (2.18) wurde in [8] auf Bewegungsgleichungen in Konfigurationsräumen mit Lie-Gruppen-Struktur angewendet. Hierzu wird (2.18a) umgeschrieben zu

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\boldsymbol{\Delta}\mathbf{x}_n, \tag{3.15a}$$
$$\boldsymbol{\Delta}\mathbf{x}_n = \mathbf{v}_n + (0.5-\beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1}. \tag{3.15b}$$

Nun kann in (3.15a) wie zuvor in Abschnitt 3.1 das Element $\mathbf{x}_n$ aus dem euklidischen Raum $\mathbb{R}^N$ durch ein Element q_n der Lie-Gruppe G ersetzt werden und die Gruppenoperation $\circ$ anstelle der Summe verwendet werden. Außerdem wird $\boldsymbol{\Delta}\mathbf{x}_n$ durch ein Element der Lie-Algebra $\widetilde{\boldsymbol{\Delta}\mathbf{q}_n}$ substituiert und es gilt

$$q_{n+1} = q_n \circ \exp\big(h\widetilde{\boldsymbol{\Delta}\mathbf{q}_n}\big),$$
$$\boldsymbol{\Delta}\mathbf{q}_n = \mathbf{v}_n + (0.5-\beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1}.$$

Die Variable $\boldsymbol{\Delta}\mathbf{q}_n$ kann als „eingefrorene“ bzw. Durchschnittsgeschwindigkeit auf dem Zeitintervall $[t_n, t_{n+1}]$ aufgefasst werden. Es folgt die Definition des Generalized-α-Verfahrens für Konfigurationsräume mit Lie-Gruppen-Struktur [8].

Definition 13 (Generalized-α-Verfahren zur Lösung von (3.1) [8])
Das *Generalized-α-Verfahren* verwendet die numerischen Lösungen q_n, $\mathbf{v}_n$, $\dot{\mathbf{v}}_n$ und $\mathbf{a}_n$ im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit konstanter Schrittweite h und ist gegeben durch

$$\begin{aligned}
q_{n+1} &= q_n \circ \exp\big(h\widetilde{\boldsymbol{\Delta}\mathbf{q}_n}\big),\\
\boldsymbol{\Delta}\mathbf{q}_n &= \mathbf{v}_n + (0.5-\beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1},\\
\mathbf{v}_{n+1} &= \mathbf{v}_n + (1-\gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1},\\
(1-\alpha_m)\mathbf{a}_{n+1} + \alpha_m\mathbf{a}_n &= (1-\alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n,\\
\dot{\mathbf{v}}_{n+1} &= \mathbf{f}(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}).
\end{aligned}$$

In jedem Zeitschritt werden Variablen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$, $\dot{\mathbf{v}}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1})$ und die Beschleunigung an einer Zwischenstelle

$$\mathbf{a}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h) \tag{3.16}$$

approximiert. Die Parameter sind durch (2.19) und (2.20) gegeben.

3.2 Zeitintegration für mechanische Systeme ohne Zwangsbedingungen

In diesem Abschnitt werden Gleichungen der Form (vgl. [8])

$$\dot{q}(t) = DL_{q(t)}(e)\cdot\widetilde{\mathbf{v}}(t), \tag{3.17a}$$
$$\mathbf{M}\big(q(t)\big)\dot{\mathbf{v}}(t) = -\mathbf{g}\big(t, q(t), \mathbf{v}(t)\big) \tag{3.17b}$$

betrachtet. Diese beschreiben mechanische Mehrkörpersysteme ohne Zwangsbedingungen mit Konfigurationsvariablen $q \in G$ und Geschwindigkeiten $\mathbf{v} \in \mathbb{R}^N$. Weiterhin bezeichnet $\mathbf{M}(q) \in \mathbb{R}^{N\times N}$ eine symmetrisch positiv definite Massematrix und

$-\mathbf{g}(t, q, \mathbf{v}) \in \mathbb{R}^N$ den Kraftvektor.
Werden numerische Verfahren zur Lösung von (3.17) verwendet, kann es vorkommen, dass nichtlineare Gleichungen gelöst werden müssen. Dafür soll das Newton-Raphson-Verfahren (vgl. Gleichung (3.60)) verwendet werden. Innerhalb dieses Verfahrens werden Iterationsmatrizen benötigt, die die Ableitungen der nichtlinearen Gleichungen enthalten. Zur allgemeinen Beschreibung werden dafür Bezeichnungen für die partiellen Ableitungen von (3.17b) nach q und $\mathbf{v}$ eingeführt. Die lokale Dämpfungsmatrix ist definiert durch

$$\mathbf{D}(t, q, \mathbf{v}) = \frac{\partial \mathbf{g}}{\partial \mathbf{v}}(t, q, \mathbf{v}) \in \mathbb{R}^{N \times N} \tag{3.18}$$

und die lokale Steifigkeitsmatrix $\mathbf{K}(t, q, \mathbf{v}, \dot{\mathbf{v}}) \in \mathbb{R}^{N \times N}$ durch

$$D_q\big(\mathbf{M}(q)\dot{\mathbf{v}} + \mathbf{g}(t, q, \mathbf{v})\big) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{K}(t, q, \mathbf{v}, \dot{\mathbf{v}})\mathbf{w} \tag{3.19}$$

für alle $\mathbf{w} \in \mathbb{R}^N$, vgl. [8].
Zunächst wird ein Beispielproblem für solche Gleichungen vorgestellt. Im Anschluss werden Lie-Gruppen-Verfahren zur Lösung von (3.17) eingeführt.

3.2.1 Benchmark: Schwerer Kreisel

Der schwere Kreisel (Abbildung 3.1; engl. Heavy Top; Übersetzung von Hackmann und Krähmer aus [39]) ist ein in der Literatur häufig verwendetes Benchmarkproblem (vgl. [7, 24]). Es handelt sich um einen rotierenden Kreisel, dessen Spitze im Nullpunkt durch ein Kugelgelenk befestigt ist. Die Orientierung des schweren Kreisels im Raum kann durch eine Rotationsmatrix $\mathbf{R} \in SO(3)$ beschrieben werden [8]. Die Bewegungsgleichungen (3.17) für einen solchen Körper in der Lie-Gruppen-Formulierung $SO(3)$ sind in der nachfolgenden Bemerkung angegeben.

Bemerkung 1 (Bewegungsgleichungen des schweren Kreisels in $SO(3)$ [8])
In $SO(3)$ sind die Bewegungsgleichungen (3.17) für den schweren Kreisel gegeben durch

$$\dot{\mathbf{R}} = \mathbf{R}\widetilde{\mathbf{\Omega}}, \tag{3.20a}$$

$$\mathbf{0}_{3\times 1} = \mathbf{J}\dot{\mathbf{\Omega}} + \widetilde{\mathbf{\Omega}}\mathbf{J}\mathbf{\Omega} - \widetilde{\mathbf{X}}\mathbf{R}^\top m\boldsymbol{\gamma}, \tag{3.20b}$$

wobei m die Masse des Körpers, $\mathbf{J}$ das Trägheitsmoment bezüglich des Fixpunktes, $\mathbf{X}$ der Ortsvektor des Körperschwerpunkts im körperfesten Bezugssystem und $\boldsymbol{\gamma}$ der

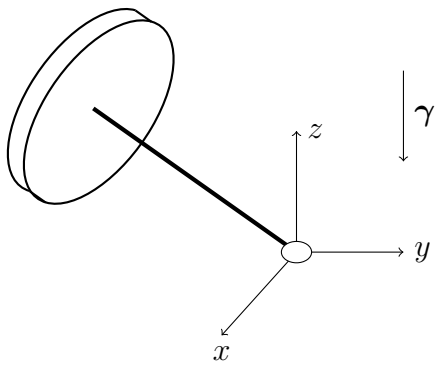

Abbildung 3.1: Schwerer Kreisel

Fallbeschleunigungsvektor ist. Speziell bedeutet dies in den Bezeichnungen aus (3.17), dass

$$\mathbf{M} = \mathbf{M}\big(\mathbf{R}\big) = \mathbf{J},\ \mathbf{g}\big(t, \mathbf{R}, \mathbf{\Omega}\big) = \widetilde{\mathbf{\Omega}}\mathbf{J}\mathbf{\Omega} - \widetilde{\mathbf{X}}\mathbf{R}^\top m\boldsymbol{\gamma}$$

die Massematrix und den Vektor der äußeren und inneren Kräfte darstellen. Für die Lösung der nichtlinearen Gleichungen innerhalb der Lie-Gruppen-Verfahren werden die Dämpfungsmatrix $\mathbf{D}$ aus (3.18) und die Steifigkeitsmatrix $\mathbf{K}$ aus (3.19) benötigt. Für (3.20) ist $\mathbf{D}$ gegeben durch [8]

$$\begin{aligned}\mathbf{D}(t, \mathbf{R}, \mathbf{\Omega}) &= \frac{\partial \mathbf{g}}{\partial \mathbf{\Omega}}(t, \mathbf{R}, \mathbf{\Omega}) = \frac{\partial(\widetilde{\mathbf{\Omega}}\mathbf{J}\mathbf{\Omega})}{\partial \mathbf{\Omega}} \\ &= \widetilde{\mathbf{\Omega}}\mathbf{J} - \widetilde{\mathbf{J}\mathbf{\Omega}},\end{aligned}$$

da $\widetilde{\mathbf{\Omega}}\mathbf{J}\mathbf{\Omega} = -\widetilde{\mathbf{J}\mathbf{\Omega}}\mathbf{\Omega}$ ist. Für alle $\mathbf{w} \in \mathbb{R}^N$ gilt aufgrund von $-\widetilde{\mathbf{w}}^T = \widetilde{\mathbf{w}}$ und $\widetilde{\mathbf{w}}\mathbf{R}^\top m\boldsymbol{\gamma} = -\widetilde{\mathbf{R}^\top m\boldsymbol{\gamma}}\mathbf{w}$ der Zusammenhang

$$\begin{aligned}D_{\mathbf{R}}\big(\mathbf{M}(\mathbf{R})\dot{\mathbf{\Omega}} + \mathbf{g}(t, \mathbf{R}, \mathbf{\Omega})\big) \cdot \big(DL_{\mathbf{R}}(e) \cdot \widetilde{\mathbf{w}}\big) &= -D_{\mathbf{R}}\big(\widetilde{\mathbf{X}}\mathbf{R}^\top m\boldsymbol{\gamma}\big) \cdot \big(DL_{\mathbf{R}}(e) \cdot \widetilde{\mathbf{w}}\big) \\ &= -\widetilde{\mathbf{X}}(\mathbf{R}\widetilde{\mathbf{w}})^\top m\boldsymbol{\gamma} = \widetilde{\mathbf{X}}\widetilde{\mathbf{w}}\mathbf{R}^\top m\boldsymbol{\gamma} \\ &= -\widetilde{\mathbf{X}}\widetilde{\mathbf{R}^\top m\boldsymbol{\gamma}}\mathbf{w}.\end{aligned}$$

Daher berechnet sich $\mathbf{K}$ zu

$$\mathbf{K}(t, \mathbf{R}, \mathbf{\Omega}, \dot{\mathbf{\Omega}}) = -\widetilde{\mathbf{X}}\widetilde{\mathbf{R}^\top m\boldsymbol{\gamma}}.$$

Bemerkung 2 (Anfangswerte und Modellparameter für den schweren Kreisel in $SO(3)$)
Um eine eindeutige Lösung von (2.2a) zu erhalten, werden Anfangswerte (2.2b) benötigt. Dies ist auch bei dem speziellen System (3.17) bzw. (3.20) der Fall. In dieser Arbeit werden die Anfangswerte des schweren Kreisels für die Lie-Gruppen-Formulierung $SO(3)$ wie in der Literatur üblich festgelegt (vgl. [8]).
Dies bedeutet, es werden $\mathbf{R}(0) = \mathbf{I}_3$ und $\mathbf{\Omega}(0) = [0\ 150\ -4.61538]^\top$ rad/s gewählt. Die Modellparameter werden als $m = 15$ kg, $\mathbf{J} = \mathrm{diag}(15.234375, 0.46875, 15.234375)$ kg $\cdot$ m^2, $\mathbf{X} = [0\ 1\ 0]^\top$ m und $\boldsymbol{\gamma} = [0\ 0\ -9.81]^\top$ m/s^2 festgesetzt.

3.2.2 Zeitintegrationsverfahren für die numerische Lösung von mechanischen Systemen ohne Zwangsbedingungen

Um die Lie-Gruppen-Verfahren aus Abschnitt 3.1 auf mechanische Systeme ohne Zwangsbedingungen (3.17) anzuwenden, wird in (3.1b)

$$\mathbf{f}(t, q, \mathbf{v}) \coloneqq -\mathbf{M}^{-1}\big(q(t)\big)\mathbf{g}\big(t, q(t), \mathbf{v}(t)\big)$$

gewählt. Es gelten die nachfolgenden Definitionen.

Definition 14 (Generalized-α-Verfahren zur Lösung von (3.17) [8])
Das *Generalized-α-Verfahren* verwendet im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit konstanter Schrittweite h die numerischen Lösungen q_n, $\mathbf{v}_n$, $\dot{\mathbf{v}}_n$ und $\mathbf{a}_n$ und ist gegeben durch

$$q_{n+1} = q_n \circ \exp\big(h\widetilde{\mathbf{\Delta}\mathbf{q}_n}\big), \tag{3.21a}$$

$$\mathbf{\Delta}\mathbf{q}_n = \mathbf{v}_n + (0.5 - \beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1}, \tag{3.21b}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + (1 - \gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1}, \tag{3.21c}$$

$$(1 - \alpha_m)\mathbf{a}_{n+1} + \alpha_m\mathbf{a}_n = (1 - \alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n, \tag{3.21d}$$

$$\mathbf{M}\big(q_{n+1}\big)\dot{\mathbf{v}}_{n+1} = -\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) \tag{3.21e}$$

mit Parametern (2.19) und (2.20). In jedem Zeitschritt werden Variablen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$, $\dot{\mathbf{v}}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1})$ und die Beschleunigung $\mathbf{a}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h)$ an einer Zwischenstelle berechnet, vgl. (3.16).

Für die Crouch-Grossman-Verfahren (3.4) und die kommutatorfreie Lie-Gruppen-Mehrschrittverfahren (3.7) werden zur Lösung von (3.17) die Gleichungen (3.4e) und (3.7c) durch

$$\mathbf{M}(q_{n+1-i})\dot{\mathbf{v}}_{n+1-i} = -\mathbf{g}\big(t_{n+1-i}, q_{n+1-i}, \mathbf{v}_{n+1-i}\big), \quad (i = 0, ..., k), \tag{3.22}$$

ersetzt.

Definition 15 (k-Schritt Munthe-Kaas-BDF-Verfahren zur Lösung von (3.17))
Ein *k-Schritt Munthe-Kaas-BDF-Verfahren (MKBDF)* zur Lösung von (3.17) verwendet im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit Schrittweite h die numerischen Lösungen q_n, $\boldsymbol{\omega}_{i-1}^{(n-1)}$ und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, zur Berechnung von q_{n+1}, $\boldsymbol{\omega}_i^{(n)}$, $(i = 0, ..., k)$, und $\mathbf{v}_{n+1}$ anhand von

$$q_{n+1} = q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big), \tag{3.23a}$$

$$\boldsymbol{\omega}_i^{(n)} = \widehat{\mathrm{BCH}}_k\big(-\boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_{i-1}^{(n-1)}\big), \ (i = 1, ..., k), \tag{3.23b}$$

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\omega}_i^{(n)} = \widehat{\mathrm{dexp}}_k^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big), \tag{3.23c}$$

$$\mathbf{M}\big(q_{n+1}\big)\frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i} = -\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) \tag{3.23d}$$

mit Parametern α_i aus (2.7) und den Approximationen (3.11) und (3.13) der Abbildungen dexp^{-1} aus (2.31) und BCH aus (2.28). In jedem Zeitschritt werden Variablen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$ und $\boldsymbol{\omega}_i^{(n)} \approx \boldsymbol{\nu}_n(t_{n+1-i})$, $(i = 0, ..., k)$, berechnet.

In den Verfahren wurde in den Gleichungen (3.21e), (3.22) und (3.23d) von links mit $\mathbf{M}\big(q_{n+1}\big)$ bzw. $\mathbf{M}(q_{n+1-i})$ multipliziert, da die Berechnung der Inversen in der computertechnischen Umsetzung ineffizient wäre.

3.2.3 BLieDF-Verfahren

Die BLieDF-Verfahren wurden in [54, 55] eingeführt. Im Unterschied zu den MKBDF-Verfahren (3.23) soll bei den BLieDF-Verfahren auf die aufwendige Berechnung der $\boldsymbol{\omega}_i^{(n)}$, $(i = 1, ..., k)$, in der Lie-Gruppe G anhand der Gleichung (3.23b) verzichtet werden. Deshalb werden in jedem Schritt nur die bereits berechneten $\boldsymbol{\omega}_0^{(n-i)}$, $(i = 1, ..., k-1)$, verwendet. Diese entsprechen wegen

$$q_{n+1} = q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big),$$

vgl. (3.9) für $i = 0$, den Inkrementen $\boldsymbol{\Delta}\mathbf{x}_{n-i}$, $(i = 1, ..., k-1)$, aus (2.14a) und (2.14b). Die Variablen $\mathbf{x}_n$ und $\mathbf{x}_{n+1}$ aus dem euklidischen Raum wurden durch die Elemente q_n und q_{n+1} der Lie-Gruppe G ersetzt. Das Inkrement $\boldsymbol{\Delta}\mathbf{x}_n$ wird durch ein Element der

Lie-Algebra $\widetilde{\boldsymbol{\omega}}_0^{(n)} \in \mathfrak{g}$ ersetzt und mit Hilfe der Exponentialfunktion in ein Element der Lie-Gruppe umgewandelt. Wie zuvor in Abschnitt 3.1.3 gilt $\boldsymbol{\omega}_0^{(n)} \approx \boldsymbol{\nu}_n(t_{n+1})$, da $q_n \approx q(t_n)$, $q_{n+1} \approx q(t_{n+1})$ und

$$q(t_{n+1}) = q(t_n) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\big),$$

vgl. (2.29). Analog zu Gleichung (2.14b) kann die gewichtete Summe $\sum_{i=1}^{k} \gamma_i \boldsymbol{\omega}_0^{(n+1-i)}$ betrachtet werden und das Update der Konfigurationsvariablen im Zeitschritt $t_n \to t_{n+1}$ wird durch

$$q_{n+1} = q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big), \tag{3.24a}$$

$$\frac{1}{h}\sum_{i=1}^{k} \gamma_i \boldsymbol{\omega}_0^{(n+1-i)} = \mathbf{v}_{n+1} + \mathbf{L}_{h,n}^{(k)} \tag{3.24b}$$

definiert. Der Verzicht auf die Umrechnung aus Gleichung (3.23b) im Vergleich zu den Munthe-Kaas-BDF-Verfahren (3.23) kann nur kompensiert werden, indem ein Korrekturterm

$$\mathbf{L}_{h,n}^{(k)} \coloneqq \mathbf{L}_h^{(k)}\big(\mathbf{v}_{n+1-k}, ..., \mathbf{v}_n, \mathbf{v}_{n+1}, \boldsymbol{\omega}_0^{(n+1-k)}, ..., \boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_0^{(n)}\big) \tag{3.25}$$

in (3.24b) eingeführt wird. Dieser soll für alle seine Argumente einer Lipschitzbedingung genügen mit einer Lipschitzkonstanten, die von der Größenordnung $\mathcal{O}(h)$ für $\mathbf{v}_{n+1-k}, ..., \mathbf{v}_n, \mathbf{v}_{n+1}$ und $\mathcal{O}(1)$ für $\boldsymbol{\omega}_0^{(n+1-k)}, ..., \boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_0^{(n)}$ ist. Außerdem wird dieser Korrekturterm so definiert, dass die lokalen Abbruchfehler (vgl. Definition A.3) in $\mathcal{O}(h^k)$ liegen. Um dies zu realisieren, wird die gewichtete Summe

$$\sum_{i=1}^{k} \gamma_i \boldsymbol{\nu}_{n+1-i}(t_{n+2-i})$$

näher untersucht. Eine klassische lokale Abbruchfehleranalyse in linearen Konfigurationsräumen würde zu

$$\frac{1}{h}\sum_{i=1}^{k} \gamma_i \boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) = \mathbf{v}(t_{n+1}) + \mathcal{O}(h^k) \tag{3.26}$$

mit $\boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) = h\mathbf{v}(t_{n+1-i})$ führen (vgl. [29]). In Konfigurationsräumen mit Lie-Gruppen-Struktur enthalten die $\boldsymbol{\nu}_{n+1-i}(t_{n+2-i})$ zusätzliche Terme mit Matrix-Kommutatoren, vgl. (2.33), die untersucht werden müssen und letztlich durch den Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ approximiert werden.

Lemma 5
Für $k+1$ mal stetig differenzierbares $\mathbf{v}(t) \in \mathbb{R}^N$ und $k \leq 6$ gilt

$$\begin{aligned}
&\frac{1}{h}\sum_{i=1}^{k} \gamma_i \boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) \\
&= \mathbf{v}(t_{n+1}) + \Bigg[\frac{h^2}{12}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{h^3}{12}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \frac{29}{720}h^4\widehat{\mathbf{v}}\dddot{\mathbf{v}} + \frac{1}{720}h^4\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} - \frac{1}{720}h^4\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{13}{360}h^4\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}} \\
&\quad - \frac{1}{240}h^4\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{1}{80}h^5\widehat{\mathbf{v}}\mathbf{v}^{(4)} + \frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dddot{\mathbf{v}} - \frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \frac{1}{864}h^5\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}} \\
&\quad - \frac{13}{8640}h^5\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{1}{48}h^5\dot{\widehat{\mathbf{v}}}\dddot{\mathbf{v}} - \frac{11}{4320}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} - \frac{11}{8640}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}} - \frac{19}{4320}h^5\ddot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}\Bigg](t_n) \\
&\quad + \mathcal{O}(h^k).
\end{aligned}$$

Beweis:
Gleichung (2.33) kann mit Hilfe des Operators $\widehat{\bullet}$ aus (2.35) umgeschrieben werden zu

$$\begin{aligned}\boldsymbol{\nu}_m(t) = \Bigg[& h\mathbf{v} + \frac{h^2}{2}\dot{\mathbf{v}} + \frac{h^3}{6}\ddot{\mathbf{v}} + \frac{h^3}{12}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{h^4}{24}\dddot{\mathbf{v}} + \frac{h^4}{24}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \frac{h^5}{120}\mathbf{v}^{(4)} + \frac{h^5}{120}\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}} - \frac{h^5}{240}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} \\ & + \frac{h^5}{80}\widehat{\mathbf{v}}\,\dddot{\mathbf{v}} + \frac{h^5}{720}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} - \frac{h^5}{720}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{h^6}{720}\mathbf{v}^{(5)} + \frac{h^6}{360}\widehat{\mathbf{v}}\mathbf{v}^{(4)} + \frac{h^6}{1440}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\,\dddot{\mathbf{v}} \\ & - \frac{h^6}{1440}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \frac{h^6}{360}\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}} - \frac{7h^6}{8640}\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{h^6}{288}\dot{\widehat{\mathbf{v}}}\,\dddot{\mathbf{v}} - \frac{h^6}{288}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} \\ & - \frac{h^6}{1728}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}}\Bigg](t_m) + \mathcal{O}(h^7) \end{aligned} \tag{3.27}$$

mit $h = t - t_m$. Wird $m = n + 1 - i$ gesetzt, so können die Taylorentwicklungen von $\mathbf{v}(t_{n+1-i})$ und deren Ableitungen berechnet werden und es gilt

$$\begin{aligned}\mathbf{v}(t_{n+1-i}) &= \mathbf{v}(t_n) + (1-i)h\dot{\mathbf{v}}(t_n) + \frac{(1-i)^2}{2}h^2\ddot{\mathbf{v}}(t_n) + \frac{(1-i)^3}{6}h^3\,\dddot{\mathbf{v}}(t_n) \\ &\quad + \frac{(1-i)^4}{24}h^4\mathbf{v}^{(4)}(t_n) + \frac{(1-i)^5}{120}h^5\mathbf{v}^{(5)}(t_n) + \mathcal{O}(h^6), \\ \dot{\mathbf{v}}(t_{n+1-i}) &= \dot{\mathbf{v}}(t_n) + (1-i)h\ddot{\mathbf{v}}(t_n) + \frac{(1-i)^2}{2}h^2\,\dddot{\mathbf{v}}(t_n) + \frac{(1-i)^3}{6}h^3\mathbf{v}^{(4)}(t_n) \\ &\quad + \frac{(1-i)^4}{24}h^4\mathbf{v}^{(5)}(t_n) + \mathcal{O}(h^5), \\ \ddot{\mathbf{v}}(t_{n+1-i}) &= \ddot{\mathbf{v}}(t_n) + (1-i)h\,\dddot{\mathbf{v}}(t_n) + \frac{(1-i)^2}{2}h^2\mathbf{v}^{(4)}(t_n) + \frac{(1-i)^3}{6}h^3\mathbf{v}^{(5)}(t_n) \\ &\quad + \mathcal{O}(h^4), \\ \dddot{\mathbf{v}}(t_{n+1-i}) &= \dddot{\mathbf{v}}(t_n) + (1-i)h\mathbf{v}^{(4)}(t_n) + \frac{(1-i)^2}{2}h^2\mathbf{v}^{(5)}(t_n) + \mathcal{O}(h^3), \\ \mathbf{v}^{(4)}(t_{n+1-i}) &= \mathbf{v}^{(4)}(t_n) + (1-i)h\mathbf{v}^{(5)}(t_n) + \mathcal{O}(h^2), \\ \mathbf{v}^{(5)}(t_{n+1-i}) &= \mathbf{v}^{(5)}(t_n) + \mathcal{O}(h). \end{aligned}$$

Das Einsetzen in (3.27) für $m = n+1-i$, $t = t_{n+2-i} = t_{n+1-i}+h$ und Zusammenfassen der Terme liefert

$$\begin{aligned}&\boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) \\ &= \Bigg[h\mathbf{v} + \left(\frac{3}{2} - i\right)h^2\dot{\mathbf{v}} + \left(\frac{7}{6} - \frac{3}{2}i + \frac{1}{2}i^2\right)h^3\ddot{\mathbf{v}} + \frac{h^3}{12}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \left(\frac{1}{8} - \frac{1}{12}i\right)h^4\widehat{\mathbf{v}}\ddot{\mathbf{v}} \\ &\quad + \left(\frac{5}{8} - \frac{7}{6}i + \frac{3}{4}i^2 - \frac{1}{6}i^3\right)h^4\,\dddot{\mathbf{v}} + \left(\frac{31}{120} - \frac{5}{8}i + \frac{7}{12}i^2 - \frac{1}{4}i^3 + \frac{1}{24}i^4\right)h^5\mathbf{v}^{(4)} \\ &\quad + \frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} - \frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}} + \left(\frac{23}{240} - \frac{1}{8}i + \frac{1}{24}i^2\right)h^5\widehat{\mathbf{v}}\,\dddot{\mathbf{v}} - \frac{1}{240}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}} \\ &\quad + \left(\frac{11}{120} - \frac{1}{8}i + \frac{1}{24}i^2\right)h^5\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}} + \left(\frac{7}{80} - \frac{31}{120}i + \frac{5}{16}i^2 - \frac{7}{36}i^3 + \frac{1}{16}i^4 - \frac{1}{120}i^5\right)h^6\mathbf{v}^{(5)} \\ &\quad + \left(\frac{1}{20} - \frac{23}{240}i + \frac{1}{16}i^2 - \frac{1}{72}i^3\right)h^6\widehat{\mathbf{v}}\mathbf{v}^{(4)} + \left(\frac{1}{480} - \frac{1}{720}i\right)h^6\widehat{\mathbf{v}}\widehat{\mathbf{v}}\,\dddot{\mathbf{v}} \\ &\quad + \left(-\frac{1}{480} + \frac{1}{720}i\right)h^6\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \left(\frac{1}{540} - \frac{1}{720}i\right)h^6\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}} + \left(-\frac{19}{8640} + \frac{1}{720}i\right)h^6\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} \\ &\quad + \left(\frac{3}{32} - \frac{3}{16}i + \frac{1}{8}i^2 - \frac{1}{36}i^3\right)h^6\dot{\widehat{\mathbf{v}}}\,\dddot{\mathbf{v}} + \left(-\frac{17}{4320} + \frac{1}{360}i\right)h^6\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\ddot{\mathbf{v}}\end{aligned}$$

$$+\left(-\frac{17}{8640}+\frac{1}{720}i\right)h^6\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}+\left(-\frac{7}{1080}+\frac{1}{240}i\right)h^6\ddot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}\Bigg](t_n)+\mathcal{O}(h^7). \tag{3.28}$$

Durch das Umschreiben der gewichteten Summe

$$\begin{aligned}\frac{1}{h}\sum_{i=1}^{k}\gamma_i\boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) &= \frac{1}{h}\sum_{i=0}^{k}\left(\sum_{j=0}^{i-1}\alpha_j\right)\boldsymbol{\nu}_{n+1-i}(t_{n+2-i})\\ &= \frac{1}{h}\sum_{j=0}^{k}\alpha_j\sum_{i=j+1}^{k}\boldsymbol{\nu}_{n+1-i}(t_{n+2-i})\end{aligned}$$

mit (2.12), kann erkannt werden, dass stets Terme mit der Struktur

$$\sum_{j=0}^{k}\alpha_j\sum_{i=j+1}^{k}\sum_{\ell=0}^{k-1}a_\ell i^\ell \tag{3.29}$$

für verschiedene $a_l\in\mathbb{R}$ untersucht werden müssen. Auf die Angabe der Geschwindigkeit $\mathbf{v}(t_n)$ und ihrer Ableitungen sowie deren Kommutatoren wird zunächst verzichtet, da sie nicht von i oder j abhängen und aus den Summen ausfaktorisiert werden können. Außerdem kann unter Verwendung von [26]

$$\sum_{i=0}^{k}i^\ell = 1+\frac{1}{\ell+1}\sum_{i=0}^{\ell}(-1)^i\binom{\ell+1}{i}B_i k^{\ell+1-i},\quad \ell\in\mathbb{N},$$

mit den Bernoulli-Zahlen B_i aus (2.32) gezeigt werden, dass

$$\begin{aligned}\sum_{i=j+1}^{k}\sum_{\ell=0}^{k-1}a_\ell i^\ell &= \sum_{\ell=0}^{k-1}a_\ell\left(\sum_{i=0}^{k}i^\ell-\sum_{i=0}^{j}i^\ell\right)\\ &= \sum_{\ell=0}^{k-1}\frac{a_\ell}{\ell+1}\sum_{i=0}^{\ell}(-1)^i\binom{\ell+1}{i}B_i(k^{\ell+1-i}-j^{l+1-i})\end{aligned}$$

für $l<k$ immer ein Polynom in j vom maximalen Grad k ist. Jedoch ist für die Rechnung nur der Faktor vor j^1 interessant, da mit Lemma 1 alle anderen Terme nach Einsetzen in (3.29) null werden oder Terme höherer Ordnung sind, also in $\mathcal{O}(h^k)$ resultieren. Die Terme ungleich null können für $i=l$ erhalten werden. Mit Lemma 1 kann somit für $k\leq 6$

$$\begin{aligned}\sum_{j=0}^{k}\alpha_j\sum_{i=j+1}^{k}\sum_{\ell=0}^{k-1}a_\ell i^\ell &= \sum_{j=0}^{k}\alpha_j\left(\sum_{\ell=0}^{k-1}\frac{a_\ell}{\ell+1}\sum_{i=0}^{\ell}(-1)^i\binom{\ell+1}{i}B_i(k^{\ell+1-i}-j^{l+1-i})\right)\\ &= \sum_{j=0}^{k}\alpha_j\left(\sum_{\ell=0}^{k-1}\frac{a_\ell}{\ell+1}(-1)^\ell\binom{\ell+1}{\ell}B_\ell(-j)\right)\\ &= \left(\sum_{\ell=0}^{k-1}a_\ell(-1)^\ell B_\ell\right)\left(-\sum_{j=0}^{k}\alpha_j j\right)\\ &= \sum_{\ell=0}^{k-1}a_\ell(-1)^\ell B_\ell\end{aligned} \tag{3.30}$$

gezeigt werden. Die Terme aus (3.28) können durch (3.30) vereinfacht werden und es gilt zum Beispiel

$$\frac{1}{h}\sum_{j=0}^{k}\alpha_j\sum_{i=j+1}^{k}\left(\frac{7}{80}-\frac{31}{120}i+\frac{5}{16}i^2-\frac{7}{36}i^3+\frac{1}{16}i^4-\frac{1}{120}i^5\right)h^6\mathbf{v}^{(5)}(t_n)=\frac{h^5}{120}\mathbf{v}^{(5)}(t_n)$$

mit $a_0=\frac{7}{80}$, $a_1=-\frac{31}{120}$, $a_2=\frac{5}{16}$, $a_3=-\frac{7}{36}$, $a_4=\frac{1}{16}$ und $a_5=-\frac{1}{120}$. Wird für alle Terme auf diese Art und Weise vorgegangen, so kann die Gleichung

$$\begin{aligned}
&\frac{1}{h}\sum_{i=1}^{k}\gamma_i\boldsymbol{\nu}_{n+1-i}(t_{n+2-i})\\
&=\Bigg[\mathbf{v}+h\dot{\mathbf{v}}+\frac{h^2}{2}\ddot{\mathbf{v}}+\frac{h^3}{6}\dddot{\mathbf{v}}+\frac{h^4}{24}\mathbf{v}^{(4)}+\frac{h^5}{120}\mathbf{v}^{(5)}+\frac{h^2}{12}\widehat{\mathbf{v}}\dot{\mathbf{v}}+\frac{h^3}{12}\widehat{\mathbf{v}}\ddot{\mathbf{v}}+\frac{29}{720}h^4\widehat{\mathbf{v}}\,\dddot{\mathbf{v}}\\
&\quad+\frac{1}{720}h^4\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}}-\frac{1}{720}h^4\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}}+\frac{13}{360}h^4\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}}-\frac{1}{240}h^4\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}+\frac{1}{80}h^5\widehat{\mathbf{v}}\mathbf{v}^{(4)}+\frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\,\dddot{\mathbf{v}}\\
&\quad-\frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}}+\frac{1}{864}h^5\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}}-\frac{13}{8640}h^5\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}+\frac{1}{48}h^5\dot{\widehat{\mathbf{v}}}\,\dddot{\mathbf{v}}-\frac{11}{4320}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\ddot{\mathbf{v}}\\
&\quad-\frac{11}{8640}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}}-\frac{19}{4320}h^5\ddot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}\Bigg](t_n)+\mathcal{O}(h^k)
\end{aligned}$$

bewiesen werden. Die Behauptung folgt aufgrund von

$$\mathbf{v}(t_{n+1})=\mathbf{v}(t_n)+h\dot{\mathbf{v}}(t_n)+\frac{h^2}{2}\ddot{\mathbf{v}}(t_n)+\frac{h^3}{6}\dddot{\mathbf{v}}(t_n)+\frac{h^4}{24}\mathbf{v}^{(4)}(t_n)+\frac{h^5}{120}\mathbf{v}^{(5)}(t_n)+\mathcal{O}(h^6)$$

für $k\leq 6$.

■

Die zusätzlichen Terme in Lemma 5 im Vergleich zu Gleichung (3.26) müssen also durch den Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ repräsentiert werden. Dabei sollen möglichst wenig Kommutatoren bzw. $\widehat{\bullet}$-Operatoren pro Integrationsschritt berechnet werden. Eine Diskussion darüber, wie $\mathbf{L}_{h,n}^{(k)}$ strukturiert werden kann, erfolgt im nachfolgenden Abschnitt. Zuvor sollen jedoch die BLieDF-Verfahren definiert werden.

Definition 16 (k-Schritt BLieDF-Verfahren zur Lösung von (3.17))
Die *k-Schritt BLieDF-Verfahren* zur Lösung von (3.17) verwenden im Zeitschritt $t_n\to t_{n+1}:=t_n+h$ mit Schrittweite h die numerischen Lösungen q_n, $\boldsymbol{\omega}_0^{(n+1-i)}$, $(i=2,...,k)$, und $\mathbf{v}_{n+1-i}$, $(i=1,...,k)$, zur Berechnung von q_{n+1}, $\boldsymbol{\omega}_0^{(n)}$ und $\mathbf{v}_{n+1}$ anhand von

$$q_{n+1}=q_n\circ\exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big),\tag{3.31a}$$

$$\frac{1}{h}\sum_{i=1}^{k}\gamma_i\boldsymbol{\omega}_0^{(n+1-i)}=\mathbf{v}_{n+1}+\mathbf{L}_{h,n}^{(k)},\tag{3.31b}$$

$$\frac{1}{h}\mathbf{M}\big(q_{n+1}\big)\sum_{i=0}^{k}\alpha_i\mathbf{v}_{n+1-i}=-\mathbf{g}\big(t_{n+1},q_{n+1},\mathbf{v}_{n+1}\big),\tag{3.31c}$$

mit Parametern α_i, γ_i nach (2.7) und (2.16) und einem Korrekturterm (3.25) (vgl. auch Abschnitt 3.2.3 für eine konkrete numerische Beschreibung). In jedem Zeitschritt werden Variablen $q_{n+1}\approx q(t_{n+1})$, $\mathbf{v}_{n+1}\approx\mathbf{v}(t_{n+1})$ und $\boldsymbol{\omega}_0^{(n)}\approx\boldsymbol{\nu}_n(t_{n+1})$ bestimmt.

Wie zuvor ist es in praktischen Implementierungen nicht sinnvoll, die Inverse von $\mathbf{M}\big(q_{n+1}\big)$ zu berechnen, weshalb in (3.31c) von links mit $\mathbf{M}(q_{n+1})$ multipliziert wurde.

Numerische Auswertung des Korrekturterms $\mathbf{L}_{h,n}^{(k)}$

In diesem Abschnitt soll der Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ aus (3.25) und (3.31b) für unterschiedliche k genauer definiert werden. Dieser wird so gewählt, dass er in allen Variablen einer Lipschitzbedingung genügt und dabei möglichst wenig $\widehat{\bullet}$-Operatoren enthält. Außerdem muss der lokale Abbruchfehler (vgl. Definition A.3) von der Größe $\mathcal{O}(h^k)$ sein. Das bedeutet, er muss die im Vergleich zu (3.26) zusätzlichen Terme aus Lemma 5 bis zu entsprechender Ordnung approximieren. Für

$$\mathbf{L}_h^{(k)}(t_n) \coloneqq \mathbf{L}_h^{(k)}\big(\mathbf{v}(t_{n+1-k}), ..., \mathbf{v}(t_n), \mathbf{v}(t_{n+1}), \boldsymbol{\nu}_{n+1-k}(t_{n+2-k}), ..., \boldsymbol{\nu}_{n-1}(t_n), \boldsymbol{\nu}_n(t_{n+1})\big),$$

vgl. (3.25), muss somit die Gleichung

$$\mathbf{L}_h^{(k)}(t_n) = \frac{1}{h}\sum_{i=1}^{k} \gamma_i \boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) - \mathbf{v}(t_{n+1}) + \mathcal{O}(h^k),$$

erfüllt sein, also mit Lemma 5

$$\begin{aligned}\mathbf{L}_h^{(k)}(t_n) = \Bigg[&\frac{h^2}{12}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{h^3}{12}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \frac{29}{720}h^4\widehat{\mathbf{v}}\dddot{\mathbf{v}} + \frac{1}{720}h^4\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} - \frac{1}{720}h^4\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{13}{360}h^4\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}}\\
&- \frac{1}{240}h^4\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{1}{80}h^5\widehat{\mathbf{v}}\mathbf{v}^{(4)} + \frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dddot{\mathbf{v}} - \frac{1}{720}h^5\widehat{\mathbf{v}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} + \frac{1}{864}h^5\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\ddot{\mathbf{v}}\\
&- \frac{13}{8640}h^5\widehat{\mathbf{v}}\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}} + \frac{1}{48}h^5\dot{\widehat{\mathbf{v}}}\dddot{\mathbf{v}} - \frac{11}{4320}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\ddot{\mathbf{v}} - \frac{11}{8640}h^5\dot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\widehat{\mathbf{v}}\dot{\mathbf{v}}\\
&- \frac{19}{4320}h^5\ddot{\widehat{\mathbf{v}}}\widehat{\mathbf{v}}\dot{\mathbf{v}}\Bigg](t_n) + \mathcal{O}(h^k).\end{aligned} \tag{3.32}$$

Für $k = 1, 2$ ist diese Bedingung (3.32) trivialerweise für

$$\mathbf{L}_{h,n}^{(1)} = \mathbf{L}_{h,n}^{(2)} = \mathbf{0} \tag{3.33}$$

erfüllt. Um (3.32) für $3 \leq k \leq 6$ numerisch zu repräsentieren, können Linearkombinationen von $\mathbf{v}_{n+1-i}$, $(i = 0, 1, ..., k)$, und $\boldsymbol{\omega}_0^{(n+1-i)}$, $(i = 1, ..., k)$, verwendet werden. Der Vorteil der Verwendung der Inkremente $\boldsymbol{\omega}_0^{(n-i)}$ ist, dass sie Approximationen von $\boldsymbol{\nu}_{n-i}(t_{n+1-i})$ darstellen, welche Terme mit Kommutatoren bzw. $\widehat{\bullet}$-Operatoren enthalten, vgl. (2.33). Diese können verwendet werden, um mit (2.35) einige von ebendiesen $\widehat{\bullet}$-Operatoren zu approximieren, ohne sie direkt auszurechnen. Auf diese Weise können pro Integrationsschritt $\widehat{\bullet}$-Operatoren eingespart werden.
Eine allgemeine Struktur des Korrekturterms $\mathbf{L}_{h,n}^{(k)}$ für $k = 3, 4, 5, 6$ kann deshalb angegeben werden durch

$$\mathbf{L}_{h,n}^{(3)} \coloneqq \frac{1}{h}\widehat{\mathbf{lin}}_{n,1}^{(3)}\mathbf{lin}_{n,2}^{(3)}, \tag{3.34a}$$

$$\mathbf{L}_{h,n}^{(4)} \coloneqq \frac{1}{h}\widehat{\mathbf{lin}}_{n,1}^{(4)}\mathbf{lin}_{n,2}^{(4)}, \tag{3.34b}$$

$$\mathbf{L}_{h,n}^{(5)} \coloneqq \frac{1}{h}\widehat{\mathbf{lin}}_{n,1}^{(5)}\left(\mathbf{lin}_{n,2}^{(5)} + \widehat{\mathbf{lin}}_{n,3}^{(5)}\mathbf{lin}_{n,4}^{(5)}\right), \tag{3.34c}$$

$$\mathbf{L}_{h,n}^{(6)} \coloneqq \frac{1}{h}\widehat{\mathbf{lin}}_{n,1}^{(6)}\left(\mathbf{lin}_{n,2}^{(6)} + \widehat{\mathbf{lin}}_{n,3}^{(6)}\left(\mathbf{lin}_{n,4}^{(6)} + \widehat{\mathbf{lin}}_{n,5}^{(6)}\mathbf{lin}_{n,6}^{(6)}\right)\right) \tag{3.34d}$$

mit

$$\mathbf{lin}_{n,j}^{(k)} = h\sum_{i=0}^{k} a_{i,j}^{(k)}\mathbf{v}_{n+1-i} + \sum_{i=1}^{k} b_{i,j}^{(k)}\boldsymbol{\omega}_0^{(n+1-i)} \tag{3.35a}$$

$$\approx h\sum_{i=0}^{k} a_{i,j}^{(k)}\mathbf{v}(t_n+(1-i)h) + \sum_{i=1}^{k} b_{i,j}^{(k)}\boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) =: \mathbf{lin}_j^{(k)}(t_n) \tag{3.35b}$$

und Parametern $a_{i,j}^{(k)}, b_{i,j}^{(k)} \in \mathbb{R}$, die so gewählt sein müssen, dass die Konsistenzbedingung (3.32) erfüllt ist. Gilt

$$a_{0,j}^{(k)} = 0 = b_{1,j}^{(k)} \text{ für alle } j, \tag{3.36}$$

so ist der daraus resultierende Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ aus (3.34) explizit und muss in jedem Zeitschritt nur einmal ausgewertet werden. Ist (3.36) jedoch nicht erfüllt, resultiert ein impliziter Korrekturterm. Dieser muss folglich in jedem Newton-Iterationsschritt neu ausgewertet werden, was eine Vielzahl an Matrix-Kommutator-Berechnungen und damit einen großen Aufwand zur Folge hat. Für $k = 3$ und $k = 4$ wird in den numerischen Tests in Abschnitt 6.2.4 überprüft, ob sich dieser Mehraufwand lohnt. Für $k = 5$ und $k = 6$ sollen die Parameter stets so gewählt sein, dass (3.36) erfüllt ist.
Für $k = 3, 4$ können allgemeine Konsistenzbedingungen für die Parameter $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$ angegeben werden, weshalb in den numerischen Tests in Abschnitt 6.2.4 verschiedene Parametersätze verglichen werden können. Für $k = 5, 6$ ist eine allgemeine Beschreibung von Konsistenzbedingungen sehr aufwendig und auch die verwendeten Computer-Algebra-Programme sind dabei an ihre Grenzen gestoßen. Daher wird in diesem Fall nur ein einzelner Parametersatz vorgeführt und getestet. Da die weitere Vorgehensweise für $k = 3, 4$ und $k = 5, 6$ recht unterschiedlich ist, wird die Betrachtung an dieser Stelle unterteilt und beide Fälle werden einzeln näher untersucht.

Korrekturterme $\mathbf{L}_{h,n}^{(3)}$ und $\mathbf{L}_{h,n}^{(4)}$
Aus (3.32) folgt für $k = 3, 4$, dass der Korrekturterm die Bedingungen

$$\mathbf{L}_{h,n}^{(3)} \approx \mathbf{L}_h^{(3)}(t_n) = \frac{h^2}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^3), \tag{3.37a}$$

$$\begin{aligned}\mathbf{L}_{h,n}^{(4)} \approx \mathbf{L}_h^{(4)}(t_n) &= \frac{h^2}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \frac{h^3}{12}\widehat{\mathbf{v}}(t_n)\ddot{\mathbf{v}}(t_n) + \mathcal{O}(h^4)\\ &= \frac{h^2}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_{n+1}) + \mathcal{O}(h^4),\end{aligned} \tag{3.37b}$$

einhalten muss. Daher soll

$$\mathbf{lin}_1^{(3)}(t_n) = \frac{h}{12}\mathbf{v}(t_n) + \mathcal{O}(h^2), \tag{3.38a}$$

$$\mathbf{lin}_2^{(3)}(t_n) = h^2\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^3), \tag{3.38b}$$

$$\mathbf{lin}_1^{(4)}(t_n) = \frac{h}{12}\mathbf{v}(t_n) + \mathcal{O}(h^3), \tag{3.38c}$$

$$\mathbf{lin}_2^{(4)}(t_n) = h^2\dot{\mathbf{v}}(t_{n+1}) + \mathcal{O}(h^4) \tag{3.38d}$$

gefordert werden. Eine offensichtliche Wahl des Korrekturterms wäre somit

$$\mathbf{L}_{h,n}^{(3)} = \frac{h^2}{12}\widehat{\mathbf{v}}_n\dot{\mathbf{v}}_n = \frac{1}{h}\widehat{\left(\frac{1}{12}h\mathbf{v}_n\right)}\left(h^2\dot{\mathbf{v}}_n\right),$$

$$\mathbf{L}_{h,n}^{(4)} = \frac{h^2}{12}\widehat{\mathbf{v}}_n\dot{\mathbf{v}}_{n+1} = \frac{1}{h}\widehat{\left(\frac{1}{12}h\mathbf{v}_n\right)}\left(h^2\dot{\mathbf{v}}_{n+1}\right)$$

mit Approximationen $\dot{\mathbf{v}}_n \approx \dot{\mathbf{v}}(t_n) + \mathcal{O}(h)$ und $\dot{\mathbf{v}}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1}) + \mathcal{O}(h^2)$. In dieser Arbeit wurden

$$\dot{\mathbf{v}}_n \approx \dot{\mathbf{v}}(t_n) = \frac{3\mathbf{v}(t_n) - 4\mathbf{v}(t_{n-1}) + \mathbf{v}(t_{n-2})}{2h} + \mathcal{O}(h^2), \tag{3.39a}$$

$$\dot{\mathbf{v}}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1}) = \frac{7\mathbf{v}(t_n) - 7\mathbf{v}(t_{n-1}) - 3\mathbf{v}(t_{n-2}) + 3\mathbf{v}(t_{n-3})}{4h} + \mathcal{O}(h^2) \tag{3.39b}$$

gewählt. In (3.39a) wären auch andere Differenzenapproximationen mit hinreichender Ordnung $\mathcal{O}(h)$ denkbar. Durch die hier angegebene Variante konnte jedoch leicht eine Ordnung höher (also $\mathcal{O}(h^2)$) als nötig erhalten werden. Die Parameter $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$ für (3.34a) und (3.34b) aus (3.35a) wären daher gegeben durch

$$a_{1,1}^{(3)} = \frac{1}{12}, \quad a_{1,2}^{(3)} = \frac{3}{2}, \quad a_{2,2}^{(3)} = -2, \quad a_{3,2}^{(3)} = \frac{1}{2}, \tag{3.40a}$$

$$a_{1,1}^{(4)} = \frac{1}{12}, \quad a_{1,2}^{(4)} = \frac{7}{4}, \quad a_{2,2}^{(4)} = -\frac{7}{4}, \quad a_{3,2}^{(4)} = -\frac{3}{4}, \quad a_{4,2}^{(4)} = \frac{3}{4} \tag{3.40b}$$

und alle anderen Parameter sind null. Die numerischen Tests in Abschnitt 6.2.4 zeigen jedoch, dass die Wahl von anderen Parametern $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$ zu genaueren Ergebnissen führen können. Aus diesem Grund soll eine allgemeine Beschreibung des Korrekturterms $\mathbf{L}_{h,n}^{(k)}$ beibehalten und Konsistenzbedingungen für $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$ angegeben werden, die für das Einhalten von (3.37a) bzw. (3.37b) erfüllt sein müssen.

Lemma 6
Die Korrekturterme (3.34a) und (3.34b) approximieren (3.37a) bzw. (3.37b), wenn die Konsistenzbedingungen

$$\frac{1}{12} = \sum_{i=0}^{k} a_{i,1}^{(k)} + \sum_{i=1}^{k} b_{i,1}^{(k)}, \quad k \geq 3, \tag{3.41a}$$

$$0 = \sum_{i=0}^{k} a_{i,2}^{(k)} + \sum_{i=1}^{k} b_{i,2}^{(k)}, \quad k \geq 3, \tag{3.41b}$$

$$1 = \sum_{i=0}^{k} a_{i,2}^{(k)}(1-i) + \sum_{i=1}^{k} b_{i,2}^{(k)}\left(\frac{3}{2} - i\right), \quad k \geq 3, \tag{3.41c}$$

und

$$0 = \sum_{i=0}^{k} a_{i,1}^{(k)}(1-i) + \sum_{i=1}^{k} b_{i,1}^{(k)}\left(\frac{3}{2} - i\right), \quad k = 4, \tag{3.41d}$$

$$1 = \sum_{i=0}^{k} \frac{a_{i,2}^{(k)}}{2}(i-1)^2 + \sum_{i=1}^{k} b_{i,2}^{(k)}\left(\frac{7}{2} - \frac{3}{2}i + \frac{i^2}{2}\right), \quad k = 4, \tag{3.41e}$$

$$0 = \sum_{i=1}^{k} b_{i,2}^{(k)}, \quad k = 4, \tag{3.41f}$$

erfüllt sind mit Parametern aus (3.35a).

Beweis:
Die Verwendung der Taylorentwicklung

$$\mathbf{v}(t_n + (1-i)h) = \mathbf{v}(t_n) + (1-i)h\dot{\mathbf{v}}(t_n) + \frac{(i-1)^2h^2}{2}\ddot{\mathbf{v}}(t_n) + \mathcal{O}(h^3)$$

und (3.28) bis zu Termen der Größenordnung $\mathcal{O}(h^4)$

$$\begin{aligned}\boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) = h\mathbf{v}(t_n) + \left(\frac{3}{2} - i\right)h^2\dot{\mathbf{v}}(t_n) + \left(\frac{7}{6} - \frac{3}{2}i + \frac{1}{2}i^2\right)h^3\ddot{\mathbf{v}}(t_n)\\ + \frac{h^3}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^4),\end{aligned}$$

und Einsetzen in (3.35b) führt zu

$$\begin{aligned}\mathbf{lin}_{n,j}^{(k)} \approx &\left(\sum_{i=0}^{k} a_{i,j}^{(k)} + \sum_{i=1}^{k} b_{i,j}^{(k)}\right)h\mathbf{v}(t_n) + \left(\sum_{i=0}^{k} a_{i,j}^{(k)}(1-i) + \sum_{i=1}^{k} b_{i,j}^{(k)}\left(\frac{3}{2} - i\right)\right)h^2\dot{\mathbf{v}}(t_n)\\ &+ \left(\sum_{i=0}^{k} \frac{a_{i,j}^{(k)}}{2}(i-1)^2 + \sum_{i=1}^{k} b_{i,j}^{(k)}\left(\frac{7}{6} - \frac{3}{2}i + \frac{i^2}{2}\right)\right)h^3\ddot{\mathbf{v}}(t_n)\\ &+ \sum_{i=1}^{k} b_{i,j}^{(k)}\frac{h^3}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^4).\end{aligned}$$

Wird nun (3.38) gefordert, so folgen für $k = 3$ und $k = 4$ die Konsistenzbedingungen (3.41a)-(3.41c). Für $k = 4$ müssen zur Einhaltung von (3.38c)-(3.38d) zusätzlich die Bedingungen (3.41d)-(3.41f) erfüllt sein, wegen $h^2\dot{\mathbf{v}}(t_{n+1}) + \mathcal{O}(h^4) = h^2\dot{\mathbf{v}}(t_n) + h^3\ddot{\mathbf{v}}(t_n) + \mathcal{O}(h^4)$.

■

Korrekturterme $\mathbf{L}_{h,n}^{(5)}$ und $\mathbf{L}_{h,n}^{(6)}$

Um Konsistenzbedingungen an die Parameter $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$ für $k = 5$ und $k = 6$ anzugeben, ist eine Forderung der Form (3.38) nicht trivial formulierbar, da die benötigten Terme aus (3.32) für $k = 5, 6$ nicht allein mit $\widehat{\mathbf{v}}(t_n)$ beginnen, sondern zusätzlich mit $\dot{\widehat{\mathbf{v}}}(t_n)$ oder $\ddot{\widehat{\mathbf{v}}}(t_n)$. Um über eine Forderung der Form (3.38) Konsistenzbedingungen zu bestimmen, könnte in (3.34c) der alternative Ansatz

$$\mathbf{L}_{h,n}^{(5)} \coloneqq \frac{1}{h}\widehat{\mathbf{lin}}_{n,1}^{(5)}\left(\mathbf{lin}_{n,2}^{(5)} + \widehat{\mathbf{lin}}_{n,3}^{(5)}\mathbf{lin}_{n,4}^{(5)}\right) + \frac{1}{h}\widehat{\mathbf{lin}}_{n,5}^{(5)}\left(\mathbf{lin}_{n,6}^{(5)} + \widehat{\mathbf{lin}}_{n,7}^{(5)}\mathbf{lin}_{n,8}^{(5)}\right)$$

gewählt werden (analog für (3.34d)). Das würde aber zu doppelt so viel frei wählbaren Parametern führen. Daher wurden passend zu (3.34) die Gleichungen

$$\begin{aligned}\mathbf{L}_h^{(5)}(t_n) &= \frac{1}{h}\widehat{\mathbf{lin}}_1^{(5)}(t_n)\left(\mathbf{lin}_2^{(5)}(t_n) + \widehat{\mathbf{lin}}_3^{(5)}(t_n)\mathbf{lin}_4^{(5)}(t_n)\right),\\ \mathbf{L}_h^{(6)}(t_n) &= \frac{1}{h}\widehat{\mathbf{lin}}_1^{(6)}(t_n)\left(\mathbf{lin}_2^{(6)}(t_n) + \widehat{\mathbf{lin}}_3^{(6)}(t_n)\left(\mathbf{lin}_4^{(6)}(t_n) + \widehat{\mathbf{lin}}_5^{(6)}(t_n)\mathbf{lin}_6^{(6)}(t_n)\right)\right),\end{aligned}$$

mit (3.35b) verwendet, um die Bedingung (3.32) zu erfüllen. Dies führt jedoch auf nichtlineare Konsistenzbedingungen, die sich nicht so leicht, wie in (3.41), formulieren

lassen. Aus diesem Grund wird auf die Angabe dieser Konsistenzbedingungen verzichtet und nur eine mögliche Parameterwahl angegeben, so dass (3.32) durch (3.34c) bzw. (3.34d) approximiert wird. Die in Lemma 7 angegebenen Parameter sind dabei nicht optimiert. Es wurde lediglich darauf geachtet, dass (3.36) erfüllt ist, um eine erneute Berechnung der Kommutatoren in jedem Newton-Iterationsschritt zu vermeiden.

Lemma 7
Werden für $k = 5$ die Parameter

$$a_{1,1}^{(5)} = \frac{4927}{2943}, \quad a_{2,1}^{(5)} = -\frac{3109}{2943}, \quad a_{3,1}^{(5)} = \frac{16}{327}, \quad a_{1,2}^{(5)} = \frac{3161}{675}, \quad a_{2,2}^{(5)} = -\frac{1417}{1350},$$
$$a_{1,3}^{(5)} = -\frac{14}{75}, \quad a_{2,3}^{(5)} = \frac{53}{300}, \quad b_{2,1}^{(5)} = 1, \quad b_{2,4}^{(5)} = 1 \tag{3.42}$$

in (3.34c) verwendet, so approximiert der entsprechende Korrekturterm $\mathbf{L}_{h,n}^{(5)}$ die auftretenden Kommutatoren (3.32) bis zur fünften Ordnung. Für $k = 6$ approximiert $\mathbf{L}_{h,n}^{(6)}$ aus (3.34d) den Ausdruck (3.32) für Parameter

$$a_{1,1}^{(6)} = \frac{902305}{74412}, \quad a_{2,1}^{(6)} = -\frac{2080379}{37206}, \quad a_{3,1}^{(6)} = \frac{555107}{6201}, \quad a_{4,1}^{(6)} = -\frac{43513}{702},$$
$$a_{5,1}^{(6)} = \frac{1175851}{74412}, \quad a_{1,2}^{(6)} = -\frac{617837}{391500}, \quad a_{2,2}^{(6)} = -\frac{745049}{391500}, \quad a_{3,2}^{(6)} = \frac{31637}{195750},$$
$$a_{1,3}^{(6)} = \frac{8473853415487}{267658590240000}, \quad a_{2,3}^{(6)} = -\frac{7972785567043}{267658590240000},$$
$$a_{1,4}^{(6)} = \frac{17588486133607200}{103646212371559}, \quad a_{1,5}^{(6)} = \frac{1039128}{14053}, \quad a_{2,6}^{(6)} = -\frac{706805}{1039128}, \quad b_{2,1}^{(6)} = -\frac{89}{53},$$
$$b_{3,1}^{(6)} = 1, \quad b_{2,2}^{(6)} = \frac{432619}{130500}, \quad b_{2,3}^{(6)} = -\frac{236562797}{210423420000}, \quad b_{2,4}^{(6)} = -\frac{13202850459341111}{103646212371559},$$
$$b_{3,4}^{(6)} = -\frac{443096}{14053}, \quad b_{4,4}^{(6)} = 1, \quad b_{2,6}^{(6)} = 1. \tag{3.43}$$

Alle anderen Parameter sollen null sein.

Beweis:
Um zu beweisen, dass die Parameter aus Lemma 7 so gewählt sind, dass Gleichung (3.32) erfüllt ist, werden zunächst die zu betrachtenden Parameter (3.42) bzw. (3.43) in die Terme (3.35a) bzw. die Approximation (3.35b) eingesetzt. Für $k = 5$ wird dann der Term (3.34c) ausgewertet. Die ausgegebenen Terme stimmen bis zu h^4 mit (3.32) überein. Somit ist eben jene Bedingung (3.32) erfüllt. Damit sind die Parameter (3.42) eine gute Wahl für den Korrekturterm $\mathbf{L}_{h,n}^{(5)}$.
Für $k = 6$ wird (3.34d) ausgewertet. Es stimmen alle berechneten Terme bis zu h^5 mit denen aus (3.32) überein.

■

3.3 Zeitintegration für beschränkte mechanische Systeme

Dieser Abschnitt beschäftigt sich mit der Zeitintegration von beschränkten mechanischen Systemen, welche durch differential-algebraische Gleichungen dargestellt wer-

den. Differential-algebraische Gleichungen treten auch in anderen Anwendungsgebieten auf, zum Beispiel bei der Simulation von elektrischen Netzwerken, chemischer Reaktionskinetik, optimalen Steuerungsproblemen und der Mehrkörperdynamik [46].

3.3.1 Differential-algebraische Gleichungen vom Index 3 und beschränkte mechanische Mehrkörpersysteme

Definition 17 (Differential-algebraische Gleichungen (DAEs) [46])
Der Zusammenhang

$$\mathbf{F}\big(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)\big) = \mathbf{0}, \quad t \in [t_0, t_{\text{end}}],$$

zwischen der auf $[t_0, t_{\text{end}}]$ definierten stetig differenzierbaren Funktion $\mathbf{x}(t)$ und ihrer Ableitung $\dot{\mathbf{x}}(t)$ heißt *differential-algebraische Gleichung (DAE)*, wenn $\mathbf{F} : [t_0, t_{\text{end}}] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ stetig differenzierbar bezüglich $\dot{\mathbf{x}}$ ist, die Jacobimatrix $\frac{\partial \mathbf{F}}{\partial \dot{\mathbf{x}}}$ konstanten Rang hat und $0 < r := \operatorname{rank} \frac{\partial \mathbf{F}}{\partial \dot{\mathbf{x}}}(t, \mathbf{x}, \dot{\mathbf{x}}) < N$ gilt.

In der vorliegenden Arbeit werden nicht DAEs mit $\mathbf{x} \in \mathbb{R}^N$ aus dem euklidischen Raum untersucht, sondern Gleichungen in Lie-Gruppen G. Definition 17 lässt sich jedoch uneingeschränkt auf Gleichungen übertragen, bei denen $\mathbf{x}(t) \in \mathbb{R}^N$ ersetzt wird durch ein $q(t) \in G$ und ein $\mathbf{v}(t) \in \mathbb{R}^N$. Im Speziellen werden DAEs der Form [8]

$$\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\mathbf{v}}(t), \tag{3.44a}$$

$$\mathbf{M}\big(q(t)\big)\dot{\mathbf{v}}(t) = -\mathbf{g}\big(t, q(t), \mathbf{v}(t)\big) - \mathbf{B}^{\top}\big(q(t)\big)\boldsymbol{\lambda}(t), \tag{3.44b}$$

$$\mathbf{0} = \boldsymbol{\Phi}\big(q(t)\big) \tag{3.44c}$$

gelöst. Diese stellen beschränkte mechanische Mehrkörpersysteme dar. Hierbei sind $q \in G$ die Konfigurationsvariablen und $\mathbf{v} \in \mathbb{R}^N$ die Geschwindigkeiten. Weiterhin beschreibt $\mathbf{M}(q) \in \mathbb{R}^{N \times N}$ eine symmetrisch positiv definite Massematrix, $-\mathbf{g}(t, q, \mathbf{v}) \in \mathbb{R}^N$ den Kraftvektor, $\mathbf{B}(q) \in \mathbb{R}^{M \times N}$ die Ableitungsmatrix der holonomen Zwangsbedingung $\mathbf{0} = \boldsymbol{\Phi}\big(q(t)\big)$ definiert durch

$$D\boldsymbol{\Phi}(q) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{B}(q)\mathbf{w} \quad \text{für alle} \quad \mathbf{w} \in \mathbb{R}^N \tag{3.45}$$

und $\boldsymbol{\lambda} \in \mathbb{R}^M$ die Lagrange-Multiplikatoren. Die Matrix $\mathbf{B}(q)$ soll Vollrang haben, also $\operatorname{rank}\big(\mathbf{B}(q)\big) = M \leq N$ für alle $q \in G$.
Die *holonome Zwangsbedingung* (3.44c) impliziert *versteckte Zwangsbedingungen* auf Geschwindigkeits- und Beschleunigungsebene. Diese können durch Differentiation nach t erhalten werden. Für die versteckte Zwangsbedingung auf Geschwindigkeitsebene gilt daher [4]

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Phi}\big(q(t)\big) = D\boldsymbol{\Phi}\big(q(t)\big) \cdot \dot{q}(t) \overset{(3.44a)}{=} D\boldsymbol{\Phi}(q) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{v}}\big) \overset{(3.45)}{=} \mathbf{B}(q)\mathbf{v}. \tag{3.46}$$

Für die zweite Zeitableitung von (3.44c) wird $\boldsymbol{\Theta}(q, \mathbf{z}) := \mathbf{B}(q)\mathbf{z}$ definiert und für die partielle Ableitung gilt [4]

$$D_q\boldsymbol{\Theta}(q, \mathbf{z}) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{Z}(q)(\mathbf{z}, \mathbf{w}), \ (\mathbf{w} \in \mathbb{R}^N), \tag{3.47}$$

mit einer Bilinearform $\mathbf{Z}(q) : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}^N$, die die Krümmung der Zwangsmannigfaltigkeit $\{q : \ \boldsymbol{\Phi}(q) = \mathbf{0}\}$ repräsentiert. Deshalb kann die versteckte Zwangsbedingung auf Beschleunigungsebene beschrieben werden durch [4]

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\big(\mathbf{B}\big(q(t)\big)\mathbf{v}(t)\big) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Theta}\big(q(t), \mathbf{v}(t)\big) = \mathbf{B}\big(q(t)\big)\dot{\mathbf{v}}(t) + \mathbf{Z}\big(q(t)\big)\big(\mathbf{v}(t), \mathbf{v}(t)\big). \tag{3.48}$$

Aufgrund des Vollrangs von $\mathbf{B}(q)$ und der Definitheit von $\mathbf{M}(q)$ sind die Gleichungen (3.48) und (3.44c) nach $\dot{\mathbf{v}}$ und $\boldsymbol{\lambda}$ auflösbar. Daher kann Gleichung (3.44) durch erneutes Ableiten von (3.48) nach der Zeit t und algebraische Umformungen auf die Form

$$\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\mathbf{v}}(t), \tag{3.49a}$$

$$\dot{\mathbf{v}}(t) = \mathbf{M}^{-1}\big(q(t)\big)\big(- \mathbf{g}\big(t, q(t), \mathbf{v}(t)\big) - \mathbf{B}^{\top}\big(q(t)\big)\boldsymbol{\lambda}(t)\big), \tag{3.49b}$$

$$\dot{\boldsymbol{\lambda}}(t) = \boldsymbol{\varphi}\big(t, q(t), \mathbf{v}(t), \boldsymbol{\lambda}(t)\big) \tag{3.49c}$$

mit einer Funktion $\boldsymbol{\varphi} : [t_0, t_{\text{end}}] \times G \times \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^M$ gebracht werden. Da die Umformung von (3.44) in (3.49) drei Ableitungen von (3.44c) erforderte, ist der Differentiationsindex von (3.44) drei (vgl. für *Differentiationsindex* [46]).
Werden numerische Verfahren zur Lösung von (3.44) verwendet, müssen zwangsläufig nichtlineare Gleichungen gelöst werden. Dafür soll das Newton-Raphson-Verfahren (vgl. Gleichung (3.60)) verwendet werden, für das die lokale Dämpfungsmatrix (3.18) und die lokale Steifigkeitsmatrix (3.19) benötigt werden. Da in den Bewegungsgleichungen (3.44) auch der Term, der die Lagrange-Multiplikatoren enthält, beachtet werden muss, ergibt sich die Steifigkeitsmatrix $\mathbf{K}(t, q, \mathbf{v}, \dot{\mathbf{v}}, \boldsymbol{\lambda})$ zu

$$D_q\big(\mathbf{M}(q)\dot{\mathbf{v}} + \mathbf{g}(t, q, \mathbf{v}) + \mathbf{B}^{\top}(q)\boldsymbol{\lambda}\big) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{K}(t, q, \mathbf{v}, \dot{\mathbf{v}}, \boldsymbol{\lambda})\mathbf{w} \tag{3.50}$$

für alle $\mathbf{w} \in \mathbb{R}^N$ [4].

Benchmark: Schwerer Kreisel

Für den schweren Kreisel (Abbildung 3.1) sollen nun Bewegungsgleichungen in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ aufgestellt werden. Im Unterschied zur Beschreibung des schweren Kreisels aus Abschnitt 3.2.1 wird dazu eine redundante Konfigurationsvariable $q = (\mathbf{x}, \mathbf{R})$ mit dem Translationsvektor $\mathbf{x} \in \mathbb{R}^3$ und der Rotationsmatrix $\mathbf{R} \in SO(3)$ statt nur $q = \mathbf{R} \in SO(3)$ gewählt. Außerdem ist die holonome Zwangsbedingung (3.44c) explizit gegeben durch

$$\mathbf{0}_{3\times 1} = \boldsymbol{\Phi}(q) = -\mathbf{R}^{\top}\mathbf{x} + \mathbf{X}$$

mit dem Massenschwerpunkt $\mathbf{X}$ im körperfesten System.
Die Bewegungsgleichungen (3.44a) und (3.44b) für den schweren Kreisel in den Lie-Gruppen-Formulierungen $G = \mathbb{R}^3 \times SO(3)$ und $G = SE(3)$ sind in den nachfolgenden Bemerkungen angegeben.

Bemerkung 3 (Bewegungsgleichungen des schweren Kreisels für $\mathbb{R}^3 \times SO(3)$ [7])
Die Bewegungsgleichungen (3.44) in $\mathbb{R}^3 \times SO(3)$ für den schweren Kreisel sind gegeben durch

$$\dot{\mathbf{x}} = \mathbf{u}, \tag{3.51a}$$

$$\dot{\mathbf{R}} = \mathbf{R}\widetilde{\boldsymbol{\Omega}}, \tag{3.51b}$$

$$\mathbf{0}_{3\times 1} = m\dot{\mathbf{u}} - m\boldsymbol{\gamma} - \mathbf{R}\boldsymbol{\lambda}, \tag{3.51c}$$

$$\mathbf{0}_{3\times 1} = \mathbf{J}\dot{\boldsymbol{\Omega}} + \widetilde{\boldsymbol{\Omega}}\mathbf{J}\boldsymbol{\Omega} + \widetilde{\mathbf{X}}\boldsymbol{\lambda}, \tag{3.51d}$$

$$\mathbf{0}_{3\times 1} = -\mathbf{R}^{\top}\mathbf{x} + \mathbf{X}, \tag{3.51e}$$

wobei m die Masse des Körpers, $\mathbf{J}$ das Trägheitsmoment im körperfesten Bezugssystem und $\boldsymbol{\gamma}$ der Fallbeschleunigungsvektor ist. Die Geschwindigkeit $\mathbf{v}$ wird durch das

Paar $\mathbf{v} = (\mathbf{u}, \boldsymbol{\Omega})$ dargestellt mit der Winkelgeschwindigkeit $\boldsymbol{\Omega}$ und der Translationsgeschwindigkeit $\mathbf{u}$ im Inertialsystem. Speziell bedeutet dies in den Bezeichnungen aus (3.44), dass

$$\mathbf{M} = \mathbf{M}(q) = \begin{bmatrix} m\mathbf{I}_3 & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{J} \end{bmatrix}, \quad \mathbf{g}(t, q, \mathbf{v}) = \begin{bmatrix} -m\boldsymbol{\gamma} \\ \widetilde{\boldsymbol{\Omega}}\mathbf{J}\boldsymbol{\Omega} \end{bmatrix}, \quad \mathbf{B}(q) = \begin{bmatrix} -\mathbf{R}^\top & -\widetilde{\mathbf{X}} \end{bmatrix}$$

die Massematrix, den Vektor der äußeren und inneren Kräfte und die Ableitungsmatrix der Zwangsbedingungen darstellen. Die versteckte Zwangsbedingung (3.46), die sich durch die Ableitung von (3.51e) nach t berechnet, ist gegeben durch

$$\mathbf{0}_{3\times1} = -\mathbf{R}^\top\mathbf{u} - \widetilde{\mathbf{X}}\boldsymbol{\Omega}.$$

Für die Lösung der nichtlinearen Gleichungen innerhalb der Lie-Gruppen-Verfahren werden die Dämpfungsmatrix $\mathbf{D}$ aus (3.18) und die Steifigkeitsmatrix $\mathbf{K}$ aus (3.50) benötigt. Für (3.51) sind sie gegeben durch [52]

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \widetilde{\boldsymbol{\Omega}}\mathbf{J} - \widetilde{\mathbf{J}\boldsymbol{\Omega}} \end{bmatrix} \quad \text{und} \quad \mathbf{K} = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{R}\widetilde{\boldsymbol{\lambda}} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \end{bmatrix}.$$

Bemerkung 4 (Bewegungsgleichungen des schweren Kreisels für $SE(3)$ [7])
Die Bewegungsgleichungen (3.44) in $SE(3)$ für den schweren Kreisel sind gegeben durch

$$\dot{\mathbf{x}} = \mathbf{R}\mathbf{U}, \tag{3.52a}$$
$$\dot{\mathbf{R}} = \mathbf{R}\widetilde{\boldsymbol{\Omega}}, \tag{3.52b}$$
$$\mathbf{0}_{3\times1} = m\dot{\mathbf{U}} + m\widetilde{\boldsymbol{\Omega}}\mathbf{U} - m\mathbf{R}^\top\boldsymbol{\gamma} - \boldsymbol{\lambda}, \tag{3.52c}$$
$$\mathbf{0}_{3\times1} = \mathbf{J}\dot{\boldsymbol{\Omega}} + \widetilde{\boldsymbol{\Omega}}\mathbf{J}\boldsymbol{\Omega} + \widetilde{\mathbf{X}}\boldsymbol{\lambda}, \tag{3.52d}$$
$$\mathbf{0}_{3\times1} = -\mathbf{R}^\top\mathbf{x} + \mathbf{X}, \tag{3.52e}$$

wobei erneut m die Masse des Körpers, $\mathbf{J}$ das Trägheitsmoment im körperfesten Bezugssystem und $\boldsymbol{\gamma}$ der Fallbeschleunigungsvektor ist. Die Geschwindigkeit $\mathbf{v}$ wird durch das Paar $\mathbf{v} = (\mathbf{U}, \boldsymbol{\Omega})$ dargestellt mit der Winkelgeschwindigkeit $\boldsymbol{\Omega}$ und der Translationsgeschwindigkeit $\mathbf{U}$ im körperfesten Bezugssystem. Speziell bedeutet dies in den Bezeichnungen aus (3.44), dass

$$\mathbf{M} = \mathbf{M}(q) = \begin{bmatrix} m\mathbf{I}_3 & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{J} \end{bmatrix}, \quad \mathbf{g}(t, q, \mathbf{v}) = \begin{bmatrix} -m\mathbf{R}^\top\boldsymbol{\gamma} + m\widetilde{\boldsymbol{\Omega}}\mathbf{U} \\ \widetilde{\boldsymbol{\Omega}}\mathbf{J}\boldsymbol{\Omega} \end{bmatrix}, \tag{3.53a}$$

$$\mathbf{B}(q) = \begin{bmatrix} -\mathbf{I}_3 & -\widetilde{\mathbf{X}} \end{bmatrix} \tag{3.53b}$$

die Massematrix, den Vektor der äußeren und inneren Kräfte und die Ableitungsmatrix der Zwangsbedingungen darstellt. Die versteckte Zwangsbedingung (3.46), die sich durch die Ableitung von (3.52e) nach t berechnet wird, ist gegeben durch

$$\mathbf{0}_{3\times1} = -\mathbf{U} - \widetilde{\mathbf{X}}\boldsymbol{\Omega}.$$

Für die Lösung der nichtlinearen Gleichungen innerhalb der Lie-Gruppen-Verfahren werden die Dämpfungsmatrix $\mathbf{D}$ aus (3.18) und die Steifigkeitsmatrix $\mathbf{K}$ aus (3.50) benötigt. Für (3.52) sind sie gegeben durch [52]

$$\mathbf{D} = \begin{bmatrix} m\widetilde{\boldsymbol{\Omega}} & -m\widetilde{\mathbf{U}} \\ \mathbf{0}_{3\times3} & \widetilde{\boldsymbol{\Omega}}\mathbf{J} - \widetilde{\mathbf{J}\boldsymbol{\Omega}} \end{bmatrix} \quad \text{und} \quad \mathbf{K} = \begin{bmatrix} \mathbf{0}_{3\times3} & -m\widetilde{\mathbf{R}^\top\boldsymbol{\gamma}} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \end{bmatrix}.$$

Bemerkung 5 (Anfangswerte und Modellparameter für den schweren Kreisel)
Um eine eindeutige Lösung von (2.2a) zu erhalten, werden Anfangswerte (2.2b) benötigt. Dies ist auch bei dem speziellen System (3.44) bzw. (3.51) oder (3.52) der Fall. In dieser Arbeit werden die Anfangswerte des schweren Kreisels für die Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ wie in der Literatur üblich gewählt (vgl. [7]).
Dies bedeutet, $\mathbf{R}(0)$ und $\boldsymbol{\Omega}(0)$ werden wie in Bemerkung 2 festgelegt. Die anderen Anfangswerte sollen konsistent zu (3.51) bzw. (3.52) gewählt werden. Dafür wird im Allgemeinen die zweite Zeitableitung (3.48) der holonomen Zwangsbedingung (3.44c) mit Gleichung (3.44b) zu dem Gleichungssystem

$$\begin{bmatrix} \mathbf{M}(q(t)) & \mathbf{B}^T(q(t)) \\ \mathbf{B}(q(t)) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}(t) \\ \boldsymbol{\lambda}(t) \end{bmatrix} = \begin{bmatrix} -\mathbf{g}\big(t, q(t), \mathbf{v}(t)\big) \\ -\mathbf{Z}\big(q(t)\big)\big(\mathbf{v}(t), \mathbf{v}(t)\big) \end{bmatrix} \tag{3.54}$$

kombiniert und an der Stelle $t = t_0 = 0$ für die speziellen Gleichungen (3.51) bzw. (3.52) gelöst. Ausführliche Rechnungen dazu erfolgten in [51] und die Anfangswerte ergeben sich zu

$$\begin{aligned}
\mathbf{x}(0) &= \mathbf{X}, \\
\mathbf{u}(0) &= \mathbf{U}(0) = \widetilde{\boldsymbol{\Omega}}(0)\mathbf{X}, \\
\boldsymbol{\lambda}(0) &= m\dot{\widetilde{\boldsymbol{\Omega}}}(0)\mathbf{X} + m\widetilde{\boldsymbol{\Omega}}(0)\widetilde{\boldsymbol{\Omega}}(0)\mathbf{X} - m\boldsymbol{\gamma}, \\
\dot{\boldsymbol{\Omega}}(0) &= \big(\mathbf{J} - m\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}\big)^{-1}\big(- \widetilde{\boldsymbol{\Omega}}(0)\mathbf{J}\boldsymbol{\Omega}(0) + m\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\Omega}}(0)\widetilde{\mathbf{X}}\boldsymbol{\Omega}(0) + m\widetilde{\mathbf{X}}\boldsymbol{\gamma}\big),
\end{aligned}$$

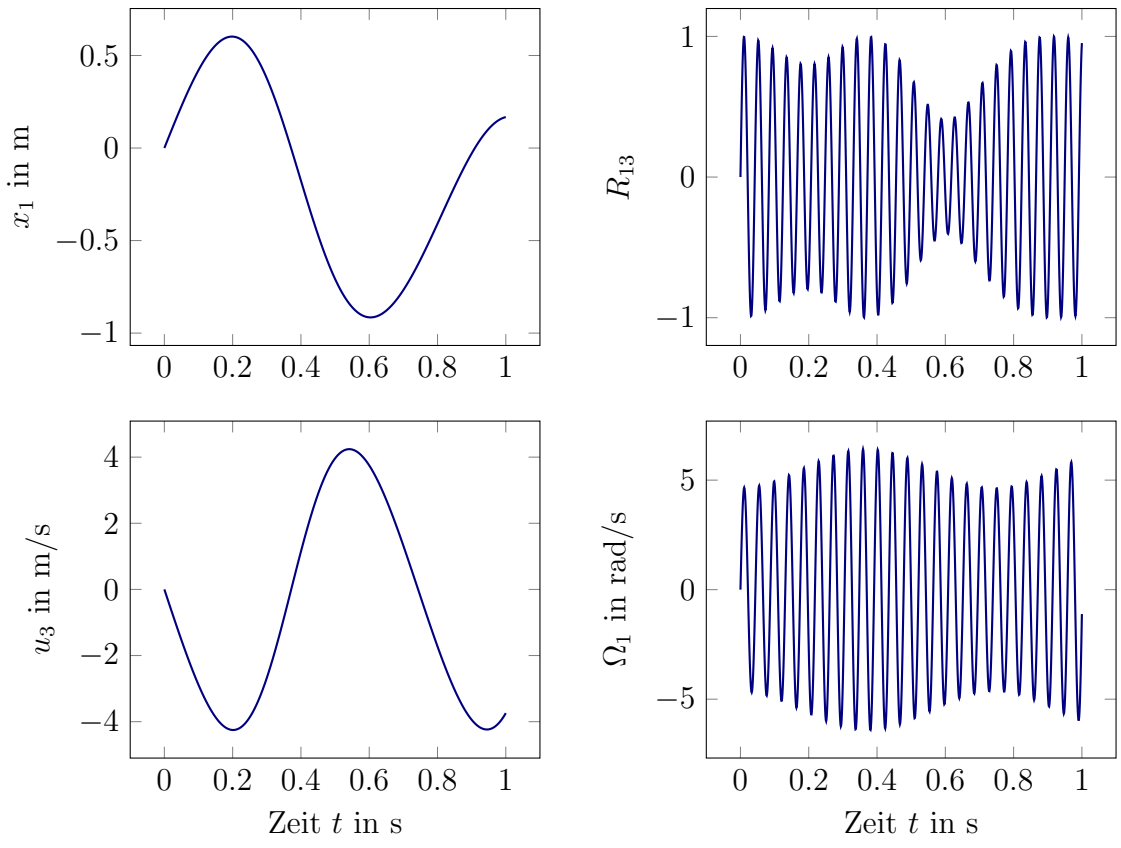

Abbildung 3.2: Verläufe von verschiedenen Variablen des schweren Kreisels in $\mathbb{R}^3 \times SO(3)$

und

$$\dot{\mathbf{u}}(0) = \dot{\widetilde{\mathbf{\Omega}}}(0)\mathbf{X} + \widetilde{\mathbf{\Omega}}(0)\widetilde{\mathbf{\Omega}}(0)\mathbf{X}$$

in $\mathbb{R}^3 \times SO(3)$ bzw.

$$\dot{\mathbf{U}}(0) = \dot{\widetilde{\mathbf{\Omega}}}(0)\mathbf{X}$$

in $SE(3)$.
Die Modellparameter werden als $m = 15$ kg, $\mathbf{J} = \mathrm{diag}(0.234375, 0.46875, 0.234375)$ kg $\cdot$ m^2, $\mathbf{X} = [0\ 1\ 0]^\top$ m und $\boldsymbol{\gamma} = [0\ 0\ -9.81]^\top$ m/s^2 festgesetzt.
Das Trägheitsmoment $\mathbf{J}$ unterscheidet sich hierbei von dem in Bemerkung 2. Dort wurde das Trägheitsmoment bezüglich des Fixpunktes und nicht bezüglich des körperfesten Koordinatensystems gewählt. Dabei wurde der Vorgehensweise aus [8] gefolgt.

Bemerkung 6 (Typischer Lösungsverlauf des schweren Kreisels)
Der typische Lösungsverlauf bis zu einem Endzeitpunkt $t_{\text{end}} = 1$s des Benchmarks „schwerer Kreisel“ in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ soll genauer vorgestellt werden.
In Abbildung 3.2 ist jeweils eine Komponente von $\mathbf{x}$, $\mathbf{R}$, $\mathbf{u}$ und $\mathbf{\Omega}$ in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ aufgezeichnet. Die Abbildungen zeigen, dass innerhalb von einer Sekunde in der Konfiguration (in den dargestellten Variablen vor allem in der Rotationsmatrix) eine starke Änderung auftritt, die durch die Oszillation der Variablen erkennbar ist. Auch in der Winkelgeschwindigkeit ist eine deutliche Oszillation zu erkennen. Dies veranschaulicht, dass der schwere Kreisel ein gutes Benchmarkproblem darstellt, da die Verfahren auch mit einer starken Schwankung innerhalb einzelner Variablen zurechtkommen und diese richtig darstellen müssen.

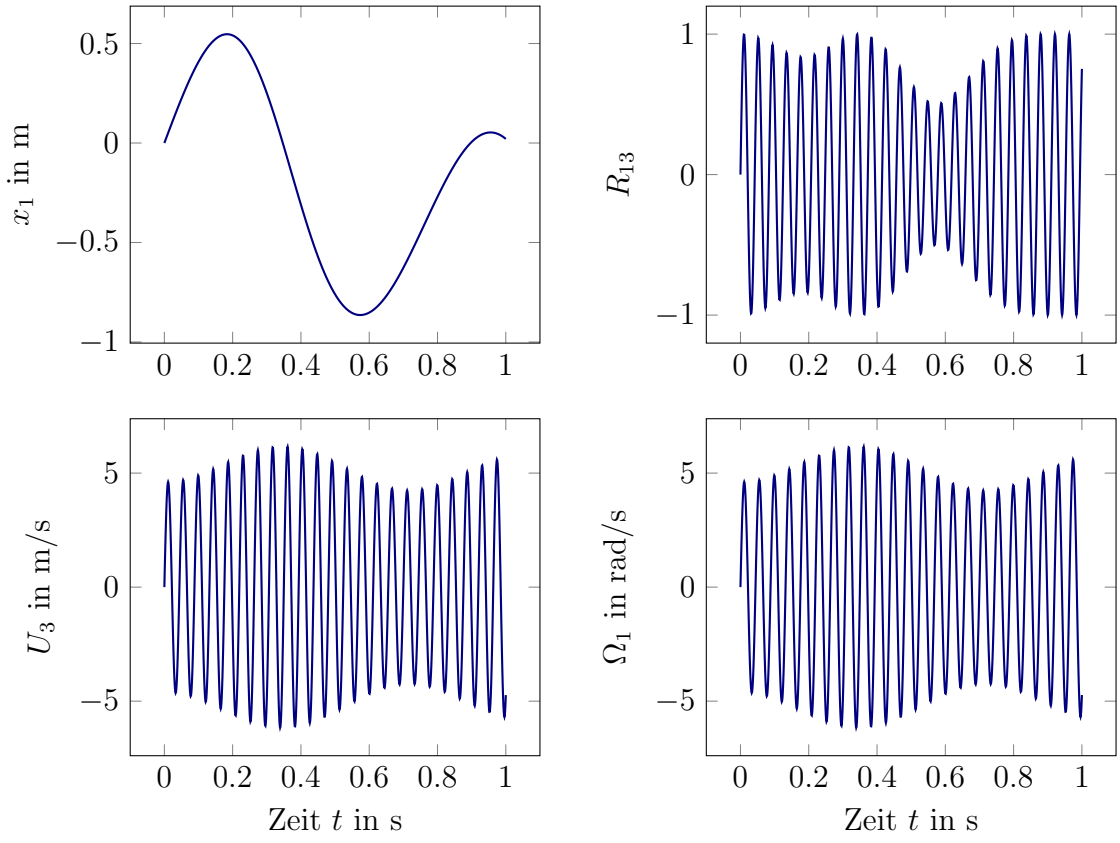

Abbildung 3.3: Verläufe von verschiedenen Variablen des schweren Kreisels in $SE(3)$

Ähnliches ist auch in Abbildung 3.3 in der Lie-Gruppen-Formulierung $SE(3)$ zu beobachten. In beiden Abbildungen sind die gleichen Komponenten dargestellt. Jedoch fällt auf, dass sich die Translationsgeschwindigkeiten $\mathbf{u}$ und $\mathbf{U}$ stark voneinander unterscheiden. Dies hängt mit der unterschiedlichen Repräsentation in verschiedenen Koordinatensystemen zusammen ($\mathbf{u}$ im Inertialsystem und $\mathbf{U}$ im körperfesten System). Es gibt also offensichtlich deutliche Unterschiede in beiden Formulierungen, weshalb die Verfahren an beiden Lie-Gruppen-Formulierungen getestet werden sollen.

3.3.2 Generalized-α-Verfahren

Definition 18 (Generalized-α-Verfahren zur Lösung von (3.44) [8])
Das *Generalized-α-Verfahren* zur Lösung von (3.44) verwendet im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit konstanter Schrittweite h die numerischen Lösungen q_n, $\mathbf{v}_n$ und $\mathbf{a}_n$ und ist gegeben durch

$$q_{n+1} = q_n \circ \exp\big(h\widetilde{\boldsymbol{\Delta}\mathbf{q}_n}\big), \tag{3.55a}$$
$$\boldsymbol{\Delta}\mathbf{q}_n = \mathbf{v}_n + (0.5-\beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1}, \tag{3.55b}$$
$$\mathbf{v}_{n+1} = \mathbf{v}_n + (1-\gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1}, \tag{3.55c}$$
$$(1-\alpha_m)\mathbf{a}_{n+1} + \alpha_m\mathbf{a}_n = (1-\alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n, \tag{3.55d}$$
$$\mathbf{M}\big(q_{n+1}\big)\dot{\mathbf{v}}_{n+1} = -\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) - \mathbf{B}^\top\big(q_{n+1}\big)\boldsymbol{\lambda}_{n+1}, \tag{3.55e}$$
$$\mathbf{0} = \boldsymbol{\Phi}\big(q_{n+1}\big). \tag{3.55f}$$

In jedem Zeitschritt werden Variablen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$, $\dot{\mathbf{v}}_{n+1} \approx \dot{\mathbf{v}}(t_{n+1})$, $\boldsymbol{\lambda}_{n+1} \approx \boldsymbol{\lambda}(t_{n+1})$ und die Beschleunigung $\mathbf{a}_{n+1}$ an einer Zwischenstelle (3.16) berechnet. Die Parameter sind durch (2.19) und (2.20) gegeben.

Für geeignete Startwerte q_0, $\mathbf{v}_0$, $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$ und $\boldsymbol{\lambda}_0$ ist das Verfahren (3.55) ein Zeitintegrationsverfahren zweiter Ordnung in den Variablen q, $\mathbf{v}$ und $\boldsymbol{\lambda}$, vgl. [2]. Die Konvergenz des Verfahrens wird außerdem in Kapitel 4 für den allgemeineren Fall des Generalized-α-Verfahrens mit variabler Schrittweite bewiesen. Allgemeine Implementierungsaspekte, wie die Angabe des Residuums und der Iterationsmatrix für das Newton-Raphson-Verfahren, erfolgten in [4].

Wahl der Startwerte

Zunächst werden anhand von gegebenen Anfangswerten $q_0 \coloneqq q(t_0)$ und $\mathbf{v}_0 \coloneqq \mathbf{v}(t_0)$ die Startwerte $\dot{\mathbf{v}}_0$ und $\boldsymbol{\lambda}_0$ konsistent zu (3.44) durch Lösung des Gleichungssystems (3.54) gewählt.
Die Struktur der Gleichungen (3.55) erfordert jedoch eine genauere Analyse der Startphase und eine spezielle Wahl der Startwerte, um in allen Zeitschritten eine zweite Ordnung in allen Variablen beobachten zu können [2]. Die Verwendung von Startwerten mit Funktionswerten der exakten Lösung in (3.55) kann zu einer Oszillation und Ordnungsreduktion in den Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ führen. Um diese Ordnungsreduktion zu vermeiden, welche aus der Nichteinhaltung der versteckten Zwangsbedingung (3.46) resultiert, müssen die Startwerte für die Geschwindigkeit $\mathbf{v}_0$ und die Beschleunigung $\mathbf{a}_0$ neu berechnet werden [2].

Bemerkung 7 (Anpassung der Startgeschwindigkeit [2])
In [2] wurde gezeigt, dass für die numerische Lösung der Geschwindigkeit zur Zeit t_0

gelten sollte

$$\mathbf{v}_0 \coloneqq \overline{\mathbf{v}}_0 + \boldsymbol{\Delta}_0^{\mathbf{v}},$$

wobei $\overline{\mathbf{v}}_0 \approx \mathbf{v}(t_0)$ konsistent bezüglich (3.44) ist. Der Term $\boldsymbol{\Delta}_0^{\mathbf{v}}$ berechnet sich durch Lösung des Gleichungssystems

$$\begin{bmatrix} \mathbf{M}(q_0) & \mathbf{B}^\top(q_0) \\ \mathbf{B}(q_0) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta}_0^{\mathbf{v}} \\ \boldsymbol{\Delta}_0^{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ \frac{1}{h}\mathbf{B}(q_0)\overline{\mathbf{l}}_0^q \end{bmatrix},$$

wobei $\overline{\mathbf{l}}_0^q$ eine Approximation des führenden Fehlerterms des lokalen Fehlers darstellt und gegeben ist durch

$$\overline{\mathbf{l}}_0^q \coloneqq \frac{h^3}{6}\left((1-6\beta-3\Delta_\alpha)\ddot{\mathbf{v}}_0 + \frac{1}{2}\widehat{\overline{\mathbf{v}}}_0\dot{\mathbf{v}}_0\right). \tag{3.56}$$

Dafür werden Approximationen $\dot{\mathbf{v}}_0 \approx \dot{\mathbf{v}}(t_0)$ konsistent zu (3.44) und $\ddot{\mathbf{v}}_0 \approx \ddot{\mathbf{v}}(t_0)$ mit

$$\ddot{\mathbf{v}}_0 \coloneqq \frac{\dot{\mathbf{v}}_{+h} - \dot{\mathbf{v}}_{-h}}{2h}$$

benötigt mit den Approximationen $\dot{\mathbf{v}}_{\pm h} \approx \dot{\mathbf{v}}(t_0 \pm h) + \mathcal{O}(h^2)$. Die Werte $\dot{\mathbf{v}}_{\pm h}$ können durch das Lösen des 2×2-Blocksystems

$$\begin{bmatrix} \mathbf{M}(q_{\pm h}) & \mathbf{B}^\top(q_{\pm h}) \\ \mathbf{B}(q_{\pm h}) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}_{\pm h} \\ \boldsymbol{\lambda}_{\pm h} \end{bmatrix} = \begin{bmatrix} -\mathbf{g}\big(t_{\pm h}, q_{\pm h}, \mathbf{v}_{\pm h}\big) \\ -\mathbf{Z}\big(q_{\pm h}\big)\big(\mathbf{v}_{\pm h}, \mathbf{v}_{\pm h}\big) \end{bmatrix}$$

bestimmt werden mit

$$\begin{aligned} q_{\pm h} &\coloneqq q_0 \circ \exp\big(\pm h\overline{\mathbf{v}}_0 + h^2\dot{\mathbf{v}}_0/2\big) \approx q(t_0 \pm h) + \mathcal{O}(h^2), \\ \mathbf{v}_{\pm h} &\coloneqq \overline{\mathbf{v}}_0 \pm h\dot{\mathbf{v}}_0 \approx \mathbf{v}(t_0 \pm h) + \mathcal{O}(h^2), \end{aligned}$$

das durch die Kombination der Bewegungsgleichungen (3.44) und der versteckten Zwangsbedingung auf Beschleunigungsebene (3.48) (ausgewertet an der Stelle $t = t_0 \pm h$) erhalten wird. Der Term $\boldsymbol{\Delta}_0^{\boldsymbol{\lambda}}$ wird in den nachfolgenden Rechnungen nicht weiter benötigt.

Bemerkung 8 (Anpassung der Startbeschleunigung [2])
In [2] wurde zudem gezeigt, dass auch die Beschleunigung eine spezielle Struktur aufweisen sollte. Da $\mathbf{a}_n$ die Beschleunigung an einer Zwischenstelle $\dot{\mathbf{v}}(t_n + \Delta_\alpha h)$ approximiert, wird für den Startwert

$$\mathbf{a}_0 \coloneqq \dot{\mathbf{v}}_0 + \Delta_\alpha h\ddot{\mathbf{v}}_0 \approx \dot{\mathbf{v}}(t_0) + \Delta_\alpha h\ddot{\mathbf{v}}(t_0) = \dot{\mathbf{v}}(t_n + \Delta_\alpha h) + \mathcal{O}(h^2)$$

gewählt, mit Δ_α aus (2.19) und den Approximationen $\dot{\mathbf{v}}_0$ und $\ddot{\mathbf{v}}_0$ aus Bemerkung 7.

3.3.3 Munthe-Kaas-BDF-Mehrschrittverfahren

Die Munthe-Kaas-BDF-Mehrschrittverfahren basieren auf den Verfahren aus [20], vgl. Abschnitt 3.1.3. Dort wurden sie jedoch nur zur Lösung von gewöhnlichen Differentialgleichungen verwendet. Die Verfahrensklasse wird im Folgenden auch auf differential-algebraische Gleichungen erweitert. Diese Erweiterung der Munthe-Kaas-BDF-Mehrschrittverfahren zur Lösung von (3.44) wurde für $k \leq 4$ in [55] eingeführt.

Definition 19 (k-Schritt Munthe-Kaas-BDF-Verfahren zur Lösung von (3.44))
Ein *k-Schritt Munthe-Kaas-BDF-Verfahren (MKBDF)* zur Lösung von (3.44) verwendet im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit Schrittweite h die numerischen Lösungen q_n, $\boldsymbol{\omega}_{i-1}^{(n-1)}$ und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, zur Berechnung von q_{n+1}, $\boldsymbol{\omega}_i^{(n)}$, $(i = 0, ..., k)$, $\mathbf{v}_{n+1}$ und $\boldsymbol{\lambda}_{n+1}$ anhand von

$$q_{n+1} = q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big), \tag{3.57a}$$

$$\boldsymbol{\omega}_i^{(n)} = \widehat{\mathrm{BCH}}_k\big(-\boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_{i-1}^{(n-1)}\big), \quad (i = 1, ..., k), \tag{3.57b}$$

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\omega}_i^{(n)} = \widehat{\mathrm{dexp}}_k^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big), \tag{3.57c}$$

$$\frac{1}{h}\mathbf{M}\big(q_{n+1}\big)\sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i} = -\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) - \mathbf{B}^{\top}\big(q_{n+1}\big)\boldsymbol{\lambda}_{n+1}, \tag{3.57d}$$

$$\mathbf{0} = \boldsymbol{\Phi}\big(q_{n+1}\big) \tag{3.57e}$$

mit Parametern α_i aus (2.7) und den Approximationen (3.11) und (3.13) der Abbildungen dexp^{-1} aus (2.31) und BCH aus (2.28). In jedem Zeitschritt werden Variablen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$, $\boldsymbol{\lambda}_{n+1} \approx \boldsymbol{\lambda}(t_{n+1})$ und $\boldsymbol{\omega}_i^{(n)} \approx \boldsymbol{\nu}_n(t_{n+1-i})$, $(i = 0, ..., k)$, berechnet.

In Kapitel 5 wird bewiesen, dass das Verfahren (3.57) für $2 \leq k \leq 6$ unter gewissen Voraussetzungen an die Startwerte, vgl. Voraussetzung 4, die Konvergenzordnung $p = k$ besitzt.

Wahl der Startwerte

Für die Wahl der Startwerte wird zunächst der Zeitpunkt t_0 betrachtet. Mit den gegebenen Anfangswerten $q_0 \coloneqq q(t_0)$ und $\mathbf{v}_0 \coloneqq \mathbf{v}(t_0)$ wird der Startwert $\boldsymbol{\lambda}_0$ konsistent zu (3.44) durch Lösung des Gleichungssystems (3.54) berechnet. Die Berechnung der Startwerte q_i, $\mathbf{v}_i$ und $\boldsymbol{\lambda}_i$, $(i = 1, ..., k-1)$, kann zum Beispiel durch ein geeignetes Einschrittverfahren erfolgen. In dieser Arbeit werden diese Startwerte jedoch anhand von der in MATLAB integrierten Funktion `ode15s` mit sehr kleinen Toleranzen berechnet. Dafür muss die zu lösende differential-algebraische-Gleichung jedoch zunächst in ein äquivalentes System von expliziten gewöhnlichen Differentialgleichungen umgeschrieben werden und überprüft werden, ob stets $q_i \in G$, $(i = 1, ..., k-1)$, erfüllt ist. Die Startwerte $\boldsymbol{\omega}_i^{(k-2)}$ für $i = 1, ..., k-1$ werden durch

$$\widetilde{\boldsymbol{\omega}}_i^{(k-2)} = \mathrm{BCH}_k\big(\widetilde{\boldsymbol{\omega}}_{i-1}^{(k-2)}, -\widetilde{\boldsymbol{\omega}}_0^{(k-1-i)}\big)$$

bzw.

$$\boldsymbol{\omega}_i^{(k-2)} = \widehat{\mathrm{BCH}}_k\big(\boldsymbol{\omega}_{i-1}^{(k-2)}, -\boldsymbol{\omega}_0^{(k-1-i)}\big)$$

bestimmt [20]. Der Wert $\boldsymbol{\omega}_0^{(k-2)}$ kann durch die Auswertung der Inversen der Exponentialabbildung von $q_i^{-1} \circ q_{i+1}$ erhalten werden, vgl. (3.9).
Die spezielle DAE-Struktur vom Index 3 der Gleichungen (3.44) erfordert, wie zuvor beim Generalized-α-Verfahren in Bemerkung 7, eine Anpassung der Anfangsgeschwindigkeiten $\mathbf{v}_i$, $(i = 0, ..., k-1)$, um eine Ordnungsreduktion in den ersten Zeitschritten zu vermeiden. Die Notwendigkeit wird in Kapitel 5 in Bemerkung 21 genauer diskutiert.

Bemerkung 9 (Anpassung der Startgeschwindigkeiten)
Die Startwerte der Geschwindigkeitsvariablen zu den Zeiten t_i, $(i = 0, ..., k-1)$, werden definiert als

$$\mathbf{v}_i \coloneqq \overline{\mathbf{v}}_i + \mathbf{\Delta}^{\mathbf{v}}_{k-1}, \tag{3.58}$$

wobei $\overline{\mathbf{v}}_i \approx \mathbf{v}(t_i)$ konsistent bezüglich (3.44) ist. Die Korrektur $\mathbf{\Delta}^{\mathbf{v}}_{k-1}$ berechnet sich durch die Lösung des Gleichungssystems

$$\begin{bmatrix} \mathbf{M}(q_{k-1}) & \mathbf{B}^\top(q_{k-1}) \\ \mathbf{B}(q_{k-1}) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \mathbf{\Delta}^{\mathbf{v}}_{k-1} \\ \mathbf{\Delta}^{\boldsymbol{\lambda}}_{k-1} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ \mathbf{B}(q_{k-1})\bar{\mathbf{l}}^{\boldsymbol{\omega}}_{k-1} \end{bmatrix}$$

mit einer Approximation $\bar{\mathbf{l}}^{\boldsymbol{\omega}}_{k-1}$ des führenden Fehlerterms des lokalen Abbruchfehlers $\mathbf{l}^{\boldsymbol{\omega}}_{k-1}$, vgl. Satz 6. Daher ist

$$\mathbf{\Delta}^{\mathbf{v}}_{k-1} = \left[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\right](q_{k-1})\bar{\mathbf{l}}^{\boldsymbol{\omega}}_{k-1}.$$

Wie in Bemerkung 7 werden zur Berechnung von $\bar{\mathbf{l}}^{\boldsymbol{\omega}}_{k-1}$ Approximationen der Ableitungen der Geschwindigkeiten $\mathbf{v}^{(j)}_{k-1} \approx \mathbf{v}^{(j)}(t_{k-1})$, $(j = 1, ..., k)$, benötigt. Dazu kann $\dot{\mathbf{v}}_{k-1}$ konsistent zu (3.44) gewählt werden. Für die restlichen Ableitungen werden Approximationen

$$\mathbf{v}^{(j)}(t_{k-1}) \approx \frac{1}{h^j}\sum_{i=0}^{j-1}\delta^{(j)}_i \mathbf{v}_{k-1-i} + \frac{j}{h^{j-1}}\dot{\mathbf{v}}_{k-1}$$

betrachtet und die Parameter $\delta^{(j)}_i$ durch Taylorentwicklung und Koeffizientenvergleich aus

$$\sum_{i=0}^{j-1}\delta^{(j)}_i \mathbf{v}(t_{k-1} - ih) + jh\dot{\mathbf{v}}(t_{k-1}) - h^j\mathbf{v}^{(j)}(t_{k-1}) = \mathcal{O}(h^{j+1})$$

bestimmt. Dies führt für festes j zu einem Vandermonde-System für $\delta^{(j)}_0, \delta^{(j)}_1, ..., \delta^{(j)}_{j-1}$, das eindeutig lösbar ist. Es folgen die Approximationen

$$\ddot{\mathbf{v}}_{k-1} \coloneqq \frac{-2\mathbf{v}_{k-1} + 2\mathbf{v}_{k-2}}{h^2} + \frac{2}{h}\dot{\mathbf{v}}_{k-1} \approx \ddot{\mathbf{v}}(t_{k-1}) + \mathcal{O}(h), \tag{3.59a}$$

$$\dddot{\mathbf{v}}_{k-1} \coloneqq \frac{-\frac{9}{2}\mathbf{v}_{k-1} + 6\mathbf{v}_{k-2} - \frac{3}{2}\mathbf{v}_{k-3}}{h^3} + \frac{3}{h^2}\dot{\mathbf{v}}_{k-1} \approx \dddot{\mathbf{v}}(t_{k-1}) + \mathcal{O}(h), \tag{3.59b}$$

$$\mathbf{v}^{(4)}_{k-1} \coloneqq \frac{-\frac{22}{3}\mathbf{v}_{k-1} + 12\mathbf{v}_{k-2} - 6\mathbf{v}_{k-3} + \frac{4}{3}\mathbf{v}_{k-4}}{h^4} + \frac{4}{h^3}\dot{\mathbf{v}}_{k-1} \approx \mathbf{v}^{(4)}(t_{k-1}) + \mathcal{O}(h), \tag{3.59c}$$

$$\begin{aligned}\mathbf{v}^{(5)}_{k-1} &\coloneqq \frac{-\frac{125}{12}\mathbf{v}_{k-1} + 20\mathbf{v}_{k-2} - 15\mathbf{v}_{k-3} + \frac{20}{3}\mathbf{v}_{k-4} - \frac{5}{4}\mathbf{v}_{k-5}}{h^5} + \frac{5}{h^4}\dot{\mathbf{v}}_{k-1} \\ &\approx \mathbf{v}^{(5)}(t_{k-1}) + \mathcal{O}(h),\end{aligned} \tag{3.59d}$$

$$\begin{aligned}\mathbf{v}^{(6)}_{k-1} &\coloneqq \frac{-\frac{137}{10}\mathbf{v}_{k-1} + 30\mathbf{v}_{k-2} - 30\mathbf{v}_{k-3} + 20\mathbf{v}_{k-4} - \frac{15}{2}\mathbf{v}_{k-5} + \frac{6}{5}\mathbf{v}_{k-6}}{h^6} + \frac{6}{h^5}\dot{\mathbf{v}}_{k-1} \\ &\approx \mathbf{v}^{(6)}(t_{k-1}) + \mathcal{O}(h).\end{aligned} \tag{3.59e}$$

Implementierungsaspekte

Bei der Verwendung der Munthe-Kaas-BDF-Lie-Gruppen-Verfahren (3.57) müssen nichtlineare Gleichungssysteme gelöst werden, was mit dem Newton-Raphson-Ver-

fahren (vgl. [18])

$$\boldsymbol{\xi}_{n+1}^{(k+1)} = \boldsymbol{\xi}_{n+1}^{(k)} + \boldsymbol{\Delta\xi}_{n+1}^{(k)} \quad \text{mit} \quad \frac{\partial \boldsymbol{\Psi}_{n,h}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}_{n+1}^{(k)})\boldsymbol{\Delta\xi}_{n+1}^{(k)} = -\boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}^{(k)}) \tag{3.60}$$

geschehen soll.
Das MKBDF-Verfahren wird dazu in die skalierte Form umgeschrieben zu

$$\mathbf{0} = \boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}) \coloneqq \begin{bmatrix} \frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\omega}_i^{(n)} - \widehat{\operatorname{dexp}}_k^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big) \\ \mathbf{r}_h\big(q(\boldsymbol{\omega}_0^{(n)}/h), \mathbf{v}_{n+1}, \dot{\mathbf{v}}(\mathbf{v}_{n+1}), h\boldsymbol{\lambda}_{n+1}, t_{n+1}\big) \\ \frac{1}{h}\boldsymbol{\Phi}\big(q(\boldsymbol{\omega}_0^{(n)}/h)\big) \end{bmatrix}$$

mit

$$\mathbf{r}_h(q, \mathbf{v}, \dot{\mathbf{v}}, h\boldsymbol{\lambda}, t) \coloneqq h\big(\mathbf{M}(q)\dot{\mathbf{v}} + \mathbf{g}(t, q, \mathbf{v})\big) + \mathbf{B}^\top(q) \cdot h\boldsymbol{\lambda}, \tag{3.61}$$

sowie $\boldsymbol{\xi}_{n+1} = \big(\frac{1}{h}(\boldsymbol{\omega}_0^{(n)})^\top, \mathbf{v}_{n+1}^\top, h\boldsymbol{\lambda}_{n+1}^\top\big)^\top$ und

$$q_{n+1} = q\big(\boldsymbol{\omega}_0^{(n)}/h\big) \coloneqq q_n \circ \exp\left(h\frac{\widetilde{\boldsymbol{\omega}}_0^{(n)}}{h}\right),$$

$$\dot{\mathbf{v}}(\mathbf{v}_{n+1}) \coloneqq \frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i}.$$

Die Iterationsmatrix hat die 3×3-Blockstruktur

$$\frac{\partial \boldsymbol{\Psi}_{n,h}}{\partial \boldsymbol{\xi}} = \begin{bmatrix} \alpha_0\mathbf{I}_N + hD_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_k^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big)\big) & -D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}}_k^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big)\big) & \mathbf{0}_{N\times M} \\ h^2\mathbf{KT} & \alpha_0\mathbf{M} + h\mathbf{D} & \mathbf{B}^\top \\ \mathbf{BT} & \mathbf{0}_{M\times N} & \mathbf{0}_{M\times M} \end{bmatrix}$$

mit der Massematrix $\mathbf{M} = \mathbf{M}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$, der Ableitungsmatrix der Zwangsbedingung $\mathbf{B} = \mathbf{B}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$ aus (3.45), dem Tangentialoperator $\mathbf{T} = \mathbf{T}\big(\boldsymbol{\omega}_0^{(n)}\big)$ aus (2.36), der Steifigkeitsmatrix $\mathbf{K} = \mathbf{K}\big(t_{n+1}, q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}_{n+1}, \dot{\mathbf{v}}(\mathbf{v}_{n+1}), \boldsymbol{\lambda}_{n+1}\big)$ aus (3.50) und der Dämpfungsmatrix $\mathbf{D} = \mathbf{D}\big(t_{n+1}, q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}_{n+1}\big)$ aus (3.18). Die partielle Ableitung von $\widehat{\operatorname{dexp}}_k^{-1}$ nach $\boldsymbol{\omega}_0^{(n)}$ wird durch $D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_k^{-1}\big(\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big)\big)$ dargestellt und die nach $\mathbf{v}_{n+1}$ durch $D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}}_k^{-1}\big(\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big)\big)$. Diese berechnen sich zu

$$\begin{aligned}
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_1^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) &\coloneqq \mathbf{0}_{N\times N}, \\
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_2^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) &\coloneqq \mathbf{0}_{N\times N}, \\
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_3^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) &\coloneqq \frac{1}{2}\widehat{\mathbf{w}}_2, \\
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_4^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) &\coloneqq D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_3^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) - \frac{1}{12}\left(\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_2 + \widehat{\widehat{\mathbf{w}}_1\mathbf{w}_2}\right), \\
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_5^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) &\coloneqq D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_4^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big), \\
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_6^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) &\coloneqq D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}}_5^{-1}\big(\mathbf{w}_1, \mathbf{w}_2\big)\big) + \frac{1}{720}\Big(\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_2 + \widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\widehat{\mathbf{w}}_1\mathbf{w}_2} \\
&\qquad + \widehat{\mathbf{w}}_1\widehat{\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2} + \widehat{\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\mathbf{w}_2}\Big)
\end{aligned}$$

und

$$
\begin{aligned}
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_1}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &\coloneqq \mathbf{I}_N,\\
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_2}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &\coloneqq \mathbf{I}_N,\\
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_3}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &\coloneqq D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_2}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) - \frac{1}{2}\widehat{\mathbf{w}}_1,\\
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_4}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &\coloneqq D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_3}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) + \frac{1}{12}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1,\\
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_5}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &\coloneqq D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_4}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big),\\
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_6}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &\coloneqq D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_5}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) - \frac{1}{720}\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1.
\end{aligned}
$$

In praktischen Implementierungen können durch geschicktes Zusammenfassen der $\widehat{\bullet}$-Operatoren einige von ebendiesen eingespart werden. Für beispielsweise $k = 6$ können dazu zunächst

$$
\begin{aligned}
\mathbf{A}(\mathbf{w}_1) &\coloneqq \widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_1 \in \mathbb{R}^{N\times N},\\
\mathbf{b}(\mathbf{w}_1,\mathbf{w}_2) &\coloneqq \widehat{\mathbf{w}}_1\mathbf{w}_2 \in \mathbb{R}^{N},\\
\mathbf{c}(\mathbf{w}_1,\mathbf{w}_2) &\coloneqq \widehat{\mathbf{w}}_1\mathbf{b}(\mathbf{w}_1,\mathbf{w}_2) \in \mathbb{R}^{N}
\end{aligned}
$$

gespeichert werden und dann werden die Ableitungen durch

$$
\begin{aligned}
D_{\boldsymbol{\omega}}\big(\widehat{\operatorname{dexp}_6}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &= \frac{1}{2}\widehat{\mathbf{w}}_2 + \left(-\frac{1}{12}\mathbf{I}_N + \mathbf{A}(\mathbf{w}_1)\right)\big(\widehat{\mathbf{w}}_1\widehat{\mathbf{w}}_2 + \mathbf{b}(\mathbf{w}_1,\mathbf{w}_2)\big)\\
&\quad + \frac{1}{720}\big(\widehat{\mathbf{w}}_1\widehat{\mathbf{c}}(\mathbf{w}_1,\mathbf{w}_2) + \widehat{\widehat{\mathbf{w}}_1\mathbf{c}}(\mathbf{w}_1,\mathbf{w}_2)\big),\\
D_{\mathbf{v}}\big(\widehat{\operatorname{dexp}_6}^{-1}(\mathbf{w}_1,\mathbf{w}_2)\big) &= \mathbf{I}_N - \frac{1}{2}\widehat{\mathbf{w}}_1\left(\mathbf{I}_N - \frac{1}{6}\widehat{\mathbf{w}}_1\left(\mathbf{I}_N + \frac{1}{60}\mathbf{A}(\mathbf{w}_1)\right)\right)
\end{aligned}
$$

bestimmt. So werden aus vorher 14 Matrix-Matrix- und 4 Matrix-Vektor-Multiplikationen nur noch 6 Matrix-Matrix- und 3 Matrix-Vektor-Multiplikationen für die Auswertung der beiden Ableitungen. Dabei ist jedoch nicht auszuschließen, dass noch weitere Einsparungen möglich wären.

3.3.4 BLieDF-Verfahren

Die BLieDF-Verfahren zur Lösung von (3.44) wurden in [54, 55] eingeführt.

Definition 20 (k-Schritt BLieDF-Verfahren zur Lösung von (3.44))
Die *k-Schritt BLieDF-Verfahren* zur Lösung von (3.44) verwenden im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h$ mit Schrittweite h die Eingangsgrößen q_n, $\boldsymbol{\omega}_0^{(n+1-i)}$, $(i = 2, ..., k)$, und $\mathbf{v}_{n+1-i}$, $(i = 1, ..., k)$, zur Berechnung von q_{n+1}, $\boldsymbol{\omega}_0^{(n)}$, $\mathbf{v}_{n+1}$ und $\boldsymbol{\lambda}_{n+1}$ anhand von

$$q_{n+1} = q_n \circ \exp\big(\widetilde{\boldsymbol{\omega}}_0^{(n)}\big), \tag{3.62a}$$

$$\frac{1}{h}\sum_{i=1}^{k}\gamma_i\boldsymbol{\omega}_0^{(n+1-i)} = \mathbf{v}_{n+1} + \mathbf{L}_{h,n}^{(k)}\big(\mathbf{v}_{n+1-k}, ..., \mathbf{v}_{n+1}, \boldsymbol{\omega}_0^{(n+1-k)}, ..., \boldsymbol{\omega}_0^{(n)}\big), \tag{3.62b}$$

$$\frac{1}{h}\mathbf{M}\big(q_{n+1}\big)\sum_{i=0}^{k}\alpha_i\mathbf{v}_{n+1-i} = -\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) - \mathbf{B}^{\top}\big(q_{n+1}\big)\boldsymbol{\lambda}_{n+1}, \tag{3.62c}$$

$$\mathbf{0} = \boldsymbol{\Phi}\big(q_{n+1}\big) \tag{3.62d}$$

mit Parametern (2.7) und (2.16) und Korrekturtermen (3.33) und (3.34) mit den Parametern aus den Lemmata 6 und 7.
In jedem Zeitschritt werden Variablen $q_{n+1} \approx q(t_{n+1})$, $\mathbf{v}_{n+1} \approx \mathbf{v}(t_{n+1})$, $\boldsymbol{\lambda}_{n+1} \approx \boldsymbol{\lambda}(t_{n+1})$ und $\boldsymbol{\omega}_0^{(n)} \approx \boldsymbol{\nu}_n(t_{n+1})$ berechnet.

In Kapitel 5 wird bewiesen, dass das Verfahren (3.62) für $2 \leq k \leq 6$ unter gewissen Voraussetzungen an die Startwerte, vgl. Voraussetzung 4, die Ordnung $p = k$ besitzt.

Bemerkung 10 (Vergleich der Verfahren (3.57) und (3.62))
Da beide Verfahren (3.57) und (3.62) BDF-Verfahren darstellen, ist ein Vergleich durchaus interessant. Solch ein Vergleich wurde in [55] für $k \leq 4$ durchgeführt. Es zeigt sich, dass die Verfahren für $k = 1, 2$ äquivalent sind. Für $k = 3$ und $k = 4$ benötigen die BLieDF-Verfahren (3.62) weniger Kommutatoren bzw. $\widehat{\bullet}$-Operatoren pro Integrationsschritt und sind daher effizienter. Numerische Tests dazu erfolgen in Abschnitt 6.2.

Wahl der Startwerte

Für die Wahl der Startwerte wird zunächst der Zeitpunkt t_0 betrachtet. Mit den gegebenen Anfangswerten $q_0 \coloneqq q(t_0)$ und $\mathbf{v}_0 \coloneqq \mathbf{v}(t_0)$ wird der Startwert $\boldsymbol{\lambda}_0$ konsistent zu (3.44) gewählt. Dazu wird das Gleichungssystem (3.54) an der Stelle $t = t_0$ gelöst. Die Berechnung der q_i, $\mathbf{v}_i$ und $\boldsymbol{\lambda}_i$, $(i = 1, ..., k-1)$, kann durch ein geeignetes Einschrittverfahren erfolgen. In dieser Arbeit werden diese Startwerte jedoch anhand von der in MATLAB integrierten Funktion `ode15s` berechnet. Dafür muss die zu lösende differential-algebraische-Gleichung jedoch zunächst in ein äquivalentes System von expliziten gewöhnlichen Differentialgleichungen umgeschrieben werden. Die $\boldsymbol{\omega}_0^{(i)}$, $(i = 0, ..., k-1)$, sind definiert durch

$$q_{i+1} = q_i \circ \exp\left(\widetilde{\boldsymbol{\omega}}_0^{(i)}\right)$$

und können daher durch die Auswertung der Inversen der Exponentialabbildung von $q_i^{-1} \circ q_{i+1}$ berechnet werden. Erneut erfordert die Struktur der DAEs vom Index 3 der Gleichungen (3.44), wie zuvor beim Generalized-α-Verfahren in Bemerkung 7, eine Anpassung der Anfangsgeschwindigkeiten $\mathbf{v}_i$, $(i = 0, ..., k-1)$, um eine Ordnungsreduktion in den ersten Zeitschritten zu vermeiden. Die Notwendigkeit dafür wird in Kapitel 5 in Bemerkung 21 genauer diskutiert.

Bemerkung 11 (Anpassung der Startgeschwindigkeiten)
Die Startwerte des Geschwindigkeitsvektors zu den Zeiten t_i, $(i = 0, ..., k-1)$, werden aus

$$\mathbf{v}_i \coloneqq \overline{\mathbf{v}}_i + \boldsymbol{\Delta}_{k-1}^{\mathbf{v}} \tag{3.63}$$

bestimmt, wobei $\overline{\mathbf{v}}_i \approx \mathbf{v}(t_i)$ konsistent bezüglich (3.44) ist. Die Korrektur $\boldsymbol{\Delta}_{k-1}^{\mathbf{v}}$ berechnet sich durch die Lösung des Gleichungssystems

$$\begin{bmatrix} \mathbf{M}(q_{k-1}) & \mathbf{B}^\top(q_{k-1}) \\ \mathbf{B}(q_{k-1}) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta}_{k-1}^{\mathbf{v}} \\ \boldsymbol{\Delta}_{k-1}^{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ \mathbf{B}(q_{k-1})\bar{\mathbf{l}}_{k-1}^{\boldsymbol{\omega}} \end{bmatrix}$$

mit einer Approximation $\bar{\mathbf{l}}_{k-1}^{\boldsymbol{\omega}}$ des führenden Fehlerterms des lokalen Abbruchfehlers $\mathbf{l}_{k-1}^{\boldsymbol{\omega}}$, welcher in Kapitel 5 in Satz 6 berechnet wird. Daher ist

$$\boldsymbol{\Delta}_{k-1}^{\mathbf{v}} = \left[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\right](q_{k-1})\bar{\mathbf{l}}_{k-1}^{\boldsymbol{\omega}}.$$

Zur Berechnung von $\bar{\mathbf{l}}^{\boldsymbol{\omega}}_{k-1}$ werden Approximationen der Ableitungen der Geschwindigkeiten $\mathbf{v}^{(j)}_{k-1} \approx \mathbf{v}^{(j)}(t_{k-1})$, $(j = 1, ..., k)$, benötigt. Diese werden durch (3.59) analog zu Bemerkung 9 bestimmt.

Implementierungsaspekte

Die Lösung der nichtlinearen Gleichungen in den BLieDF-Verfahren (3.62) erfolgt durch das Newton-Raphson-Verfahren (3.60). Der Aufwand und die Struktur hängt stark vom Korrekturterm $\mathbf{L}^{(k)}_{h,n} = \mathbf{L}^{(k)}_h\big(\mathbf{v}_{n+1-k}, ..., \mathbf{v}_n, \mathbf{v}_{n+1}, \boldsymbol{\omega}_0^{(n+1-k)}, ..., \boldsymbol{\omega}_0^{(n-1)}, \boldsymbol{\omega}_0^{(n)}\big)$ ab. Im allgemeinen Fall ist

$$\mathbf{0} = \boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}) \coloneqq \begin{bmatrix} \sum_{i=1}^{k} \gamma_i \frac{\boldsymbol{\omega}_0^{(n+1-i)}}{h} - \mathbf{v}_{n+1} - \mathbf{L}^{(k)}_{h,n} \\ \mathbf{r}_h\left(q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}_{n+1}, \dot{\mathbf{v}}(\mathbf{v}_{n+1}), h\boldsymbol{\lambda}_{n+1}, t_{n+1}\right) \\ \frac{1}{h}\boldsymbol{\Phi}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big) \end{bmatrix}$$

mit (3.61), $\boldsymbol{\xi}_{n+1} = \big(\frac{1}{h}(\boldsymbol{\omega}_0^{(n)})^\top, \mathbf{v}_{n+1}^\top, h\boldsymbol{\lambda}_{n+1}^\top\big)^\top$ und

$$q_{n+1} = q\big(\boldsymbol{\omega}_0^{(n)}/h\big) \coloneqq q_n \circ \exp\left(h\frac{\widetilde{\boldsymbol{\omega}}_0^{(n)}}{h}\right),$$

$$\dot{\mathbf{v}}(\mathbf{v}_{n+1}) \coloneqq \frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}_{n+1-i}.$$

Die Iterationsmatrix hat die 3×3-Blockstruktur

$$\frac{\partial \boldsymbol{\Psi}_{n,h}}{\partial \boldsymbol{\xi}} = \begin{bmatrix} \gamma_1 \mathbf{I}_N - D_{\boldsymbol{\omega}/h}\big(\mathbf{L}^{(k)}_{h,n}\big) & -\mathbf{I}_N - D_{\mathbf{v}}\big(\mathbf{L}^{(k)}_{h,n}\big) & \mathbf{0}_{N\times M} \\ h^2\mathbf{K}\mathbf{T} & \alpha_0\mathbf{M} + h\mathbf{D} & \mathbf{B}^\top \\ \mathbf{B}\mathbf{T} & \mathbf{0}_{M\times N} & \mathbf{0}_{M\times M} \end{bmatrix}$$

mit der Massematrix $\mathbf{M} = \mathbf{M}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$, der Ableitungsmatrix der Zwangsbedingung $\mathbf{B} = \mathbf{B}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$ aus (3.45), dem Tangentialoperator $\mathbf{T} = \mathbf{T}\big(\boldsymbol{\omega}_0^{(n)}\big)$ aus (2.36), der Steifigkeitsmatrix $\mathbf{K} = \mathbf{K}\big(t_{n+1}, q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}_{n+1}, \dot{\mathbf{v}}(\mathbf{v}_{n+1}), \boldsymbol{\lambda}_{n+1}\big)$ aus (3.50) und der Dämpfungsmatrix $\mathbf{D} = \mathbf{D}\big(t_{n+1}, q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}_{n+1}\big)$ aus (3.18). Dabei ist $D_{\boldsymbol{\omega}}\big(\mathbf{L}^{(k)}_{h,n}\big)$ die partielle Ableitung von $\mathbf{L}^{(k)}_{h,n}$ nach $\boldsymbol{\omega}_0^{(n)}$ mit

$$\begin{aligned} D_{\boldsymbol{\omega}}\big(\mathbf{L}^{(1)}_{h,n}\big) &= D_{\boldsymbol{\omega}}\big(\mathbf{L}^{(2)}_{h,n}\big) = \mathbf{0}_{N\times N}, \\ D_{\boldsymbol{\omega}}\big(\mathbf{L}^{(3)}_{h,n}\big) &= b^{(3)}_{1,2}\widehat{\mathbf{lin}}^{(3)}_{n,1} - b^{(3)}_{1,1}\widehat{\mathbf{lin}}^{(3)}_{n,2}, \\ D_{\boldsymbol{\omega}}\big(\mathbf{L}^{(4)}_{h,n}\big) &= b^{(4)}_{1,2}\widehat{\mathbf{lin}}^{(4)}_{n,1} - b^{(4)}_{1,1}\widehat{\mathbf{lin}}^{(4)}_{n,2} \end{aligned}$$

und $D_{\mathbf{v}}\big(\mathbf{L}^{(k)}_{h,n}\big)$ die partielle Ableitung von $\mathbf{L}^{(k)}_{h,n}$ nach $\mathbf{v}_{n+1}$ mit

$$\begin{aligned} D_{\mathbf{v}}\big(\mathbf{L}^{(1)}_{h,n}\big) &= D_{\mathbf{v}}\big(\mathbf{L}^{(2)}_{h,n}\big) = \mathbf{0}_{N\times N}, \\ D_{\mathbf{v}}\big(\mathbf{L}^{(3)}_{h,n}\big) &= a^{(3)}_{0,2}\widehat{\mathbf{lin}}^{(3)}_{n,1} - a^{(3)}_{0,1}\widehat{\mathbf{lin}}^{(3)}_{n,2}, \\ D_{\mathbf{v}}\big(\mathbf{L}^{(4)}_{h,n}\big) &= a^{(4)}_{0,2}\widehat{\mathbf{lin}}^{(4)}_{n,1} - a^{(4)}_{0,1}\widehat{\mathbf{lin}}^{(4)}_{n,2}. \end{aligned}$$

Hängt der Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ nicht von $\mathbf{v}_{n+1}$ und $\boldsymbol{\omega}_0^{(n)}$, $(a_{0,j}^{(k)} = b_{1,j}^{(k)} = 0,\, j = 1, 2, ...)$, ab, so muss die Gleichung (3.62b) nicht im Residuum $\boldsymbol{\Psi}_{n,h}$ berücksichtigt werden und kann explizit gelöst werden. Für $k = 5$ und $k = 6$ ist dies zum Beispiel für die Parameter aus Lemma 7 der Fall. In diesem einfacheren Fall werden die BLieDF-Verfahren (3.62) in die skalierte Form umgeschrieben zu

$$\mathbf{0} = \boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}) \coloneqq \begin{bmatrix} \mathbf{r}_h\left(q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}\big(\boldsymbol{\omega}_0^{(n)}/h\big), \dot{\mathbf{v}}\big(\boldsymbol{\omega}_0^{(n)}/h\big), h\boldsymbol{\lambda}_{n+1}, t_{n+1}\right) \\ \frac{1}{h}\boldsymbol{\Phi}\left(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\right) \end{bmatrix}$$

mit (3.61), $\boldsymbol{\xi}_{n+1} = \left(\frac{1}{h}(\boldsymbol{\omega}_0^{(n)})^\top, h\boldsymbol{\lambda}_{n+1}^\top\right)^\top$ und

$$\begin{aligned} q_{n+1} = q\big(\boldsymbol{\omega}_0^{(n)}/h\big) &\coloneqq q_n \circ \exp\left(h\frac{\boldsymbol{\omega}_0^{(n)}}{h}\right), \\ \mathbf{v}_{n+1} = \mathbf{v}\big(\boldsymbol{\omega}_0^{(n)}/h\big) &\coloneqq \frac{1}{h}\sum_{i=1}^{k}\gamma_i\boldsymbol{\omega}_0^{(n+1-i)} - \mathbf{L}_{h,n}^{(k)}, \\ \dot{\mathbf{v}}\big(\boldsymbol{\omega}_0^{(n)}/h\big) &\coloneqq \frac{1}{h}\left(\alpha_0\mathbf{v}\big(\boldsymbol{\omega}_0^{(n)}/h\big) + \sum_{i=1}^{k}\alpha_i\mathbf{v}_{n+1-i}\right). \end{aligned}$$

Die Iterationsmatrix hat die 2×2-Blockstruktur

$$\frac{\partial\boldsymbol{\Psi}_{n,h}}{\partial\boldsymbol{\xi}} = \begin{bmatrix} \alpha_0\gamma_1\mathbf{M} + h\gamma_1\mathbf{D} + h^2\mathbf{K}\mathbf{T} & \mathbf{B}^\top \\ \mathbf{B}\mathbf{T} & \mathbf{0}_{M\times M} \end{bmatrix}$$

mit der Massematrix $\mathbf{M} = \mathbf{M}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$, der Ableitungsmatrix der Zwangsbedingung $\mathbf{B} = \mathbf{B}\big(q\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$ aus (3.45), dem Tangentialoperator $\mathbf{T} = \mathbf{T}\big(\boldsymbol{\omega}_0^{(n)}\big)$ aus (2.36), der Dämpfungsmatrix $\mathbf{D} = \mathbf{D}\big(t_{n+1}, q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}\big(\boldsymbol{\omega}_0^{(n)}/h\big)\big)$ aus (3.18) und der Steifigkeitsmatrix $\mathbf{K} = \mathbf{K}\big(t_{n+1}, q\big(\boldsymbol{\omega}_0^{(n)}/h\big), \mathbf{v}\big(\boldsymbol{\omega}_0^{(n)}/h\big), \dot{\mathbf{v}}\big(\boldsymbol{\omega}_0^{(n)}/h\big), \boldsymbol{\lambda}_{n+1}\big)$ aus (3.50).

Kapitel 4

Konvergenzanalyse des Generalized-α-DAE-Integrationsverfahrens auf Lie-Gruppen für beschränkte mechanische Systeme

Das Ziel dieses Kapitels ist es, die Konvergenz des Generalized-α-Verfahrens für Konfigurationsräume mit Lie-Gruppen-Struktur für beschränkte mechanische Mehrkörpersysteme zu beweisen, welche differential-algebraische Gleichungen vom Index 3 darstellen. Ein Konvergenzbeweis für das Generalized-α-Verfahren (3.55) für konstante Schrittweiten erfolgte bereits in [2]. In vielen praktischen Anwendungen ist es jedoch sinnvoll, kein äquidistantes Punktgitter zu verwenden, sondern variable Schrittweiten h_n zuzulassen, vgl. (2.4). Dies ist zum Beispiel der Fall, wenn die Lösung eines Problems zu gewissen Zeitpunkten stark variiert (kleine Schrittweiten) und an anderen über lange Zeit fast identisch bleibt (große Schrittweiten). Um dies zu realisieren, wird eine Schrittweitensteuerung verwendet, die auf Grundlage des in einem Zeitschritt begangenen lokalen Fehlers eine optimale Schrittweite für den nächsten Schritt bestimmt bzw. im schlimmsten Fall den gerade ausgeführten Zeitschritt mit einer kleineren Schrittweite wiederholt [46]. Die Übertragung von Zeitintegrationsverfahren auf variable Schrittweiten ist jedoch nicht ohne Weiteres möglich. Dadurch entstehen weitere Einschränkungen, deren Nichtbeachtung zum Beispiel zu Stabilitätsproblemen führen können.
In diesem Kapitel soll das Generalized-α-Verfahren (3.55) auf variable Schrittweiten erweitert werden und die Konvergenzanalyse für dieses Generalized-α-Verfahren mit variabler Schrittweite ausgehend von [2] durchgeführt werden. Dazu muss das Generalized-α-Verfahren (3.55) für konstante Schrittweiten zunächst so erweitert werden, dass auch variable Schrittweiten h_n verwendet werden können. Solch eine Anpassung erfolgte bereits in [52], dabei wurde sich am Vorgehen von Jay und Negrut [34] für Hilber-Hughes-Taylor-α-Verfahren orientiert. In [52] wurde auch die Schrittweitensteuerung des Generalized-α-Verfahrens untersucht. Ein formaler Konvergenzbeweis wurde jedoch nicht geführt.
Die Konvergenzanalyse des Generalized-α-Verfahrens beginnt mit einer lokalen Abbruchfehleranalyse. Dabei wird die analytische Lösung in die Verfahrensvorschrift eingesetzt und der lokale Abbruchfehler unter Verwendung der Taylorentwicklung und

der Magnusentwicklung (vgl. (2.33)) abgeschätzt. Anschließend werden globale Fehlergleichungen aufgestellt. Dazu wird die Differenz aus der Verfahrensvorschrift mit der analytischen Lösung und der Verfahrensvorschrift mit der numerischen Lösung gebildet. Die so entstandenen Fehlerrekursionen werden zu einer gekoppelten Fehlerrekursion kombiniert. Durch solch eine gekoppelte Fehlerrekursion kann schließlich die Konvergenz zweiter Ordnung des Verfahrens bewiesen werden.
Die Konvergenz wird auf einem kompakten Zeitintervall unter der nachfolgenden Voraussetzung gezeigt.

Voraussetzung 1
Es sollen das abgeschlossene Zeitintervall $[t_0, t_{\text{end}}]$, die variablen Schrittweiten $h_n := t_{n+1} - t_n \in (0, \overline{h}]$ mit $t_{\text{end}} = t_0 + \sum_{n=0}^{N_{\text{end}}} h_n$, $N_{\text{end}} \in \mathbb{N}$, und einer von n unabhängigen Konstanten $\overline{h} > 0$ vorausgesetzt werden. Zudem wird angenommen, dass das Verhältnis aus maximaler Schrittweite $h_{\max} := \max_{n=0,\dots,N_{\text{end}}} h_n$ und minimaler Schrittweite $h_{\min} := \min_{n=0,\dots,N_{\text{end}}} h_n$ beschränkt bleibt, also

$$\frac{h_{\max}}{h_{\min}} \leq C_h \tag{4.1}$$

für ein $0 < C_h \in \mathbb{R}$. Daher bleiben auch die Schrittweitenverhältnisse $\sigma_n := h_n/h_{n-1}$ für alle $n \leq N_{\text{end}}$ beschränkt. Es gibt also positive Konstanten $\sigma_{\min} \in \mathbb{R}$ und $\sigma_{\max} \in \mathbb{R}$, so dass $\sigma_{\min} \leq \sigma_n \leq \sigma_{\max}$ gilt.

Unter dieser Voraussetzung gilt für das Schrittweitenverhältnis $\sigma_n := h_n/h_{n-1}$ die Beziehung

$$h_n = \sigma_n h_{n-1} = \mathcal{O}(h_{n-1}), \tag{4.2}$$

bzw. unter Verwendung von $h_n \leq h_{\max}$

$$h_n = \mathcal{O}(h_{\max})$$

für $n = 0, ..., N_{\text{end}}$.
Bevor der eigentliche Konvergenzbeweis begonnen werden kann, muss das Verfahren (3.55) für solche variablen Schrittweiten definiert werden.

4.1 Erweiterung des Generalized-α-Verfahrens auf variable Schrittweiten

Der Konvergenzbeweis des Generalized-α-Verfahrens (3.55) aus [2] hat gezeigt, dass speziell gewählte Startwerte wie in Abschnitt 3.3.2 vonnöten sind, um eine Ordnungsreduktion in den ersten Zeitschritten zu vermeiden. Dieses Verhalten hängt stark mit dem Umstand zusammen, dass das Verfahren quasi von einer Schrittweite null (also aus dem Ruhezustand) auf eine beliebige Schrittweite h wechselt. Man könnte somit sagen, dass im ersten Zeitschritt ein Schrittweitenwechsel vorgenommen wird, welcher durch die Anpassungen aus den Bemerkungen 7 und 8 kompensiert werden kann. Ähnliches muss nun auch vor jedem Schrittweitenwechsel beim Generalized-α-Verfahren für variable Schrittweiten geschehen. In den nachfolgenden Bemerkungen werden solche Anpassungen für die Geschwindigkeit $\mathbf{v}_n$ und die Beschleunigung $\mathbf{a}_n$ beschrieben. Dabei wird einem Ansatz von Jay und Negrut [34] gefolgt, die eine variable Schrittweitenformulierung für Hilber-Hughes-Taylor-α-Verfahren eingeführt

haben. Eine weitere Möglichkeit wäre die Verwendung variabler Verfahrensparameter $\beta = \beta(\sigma_n)$ und $\gamma = \gamma(\sigma_n)$, vgl. [6].

Bemerkung 12 (Anpassung der Beschleunigung, vgl. [52] und [34, für HHT-α])
Wird das Generalized-α-Verfahren (3.55) auf variable Schrittweiten $h_n \coloneqq t_{n+1} - t_n$ übertragen, hängt die Beschleunigung

$$\mathbf{a}_n \approx \dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_{n-1}) \tag{4.3}$$

von eben dieser Schrittweite $h_{n-1} = h_n/\sigma_n$ mit dem Schrittweitenverhältnis σ_n ab. Aus diesem Grund müssen weitere Änderungen an der im vorherigen Zeitschritt berechneten Variablen $\mathbf{a}_n$ vorgenommen werden, um auch im Fall variabler Schrittweiten die Konvergenz zweiter Ordnung nachweisen zu können.
Dabei wird $\mathbf{a}_n \approx \dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_{n-1})$ in (3.55) durch eine Approximation

$$\overline{\mathbf{a}}_n \approx \dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_n) = \dot{\mathbf{v}}(t_n + \Delta_\alpha h_{n-1}) + \Delta_\alpha h_n \left(1 - \frac{1}{\sigma_n}\right) \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^2) \tag{4.4}$$

substituiert. Die Approximation $\mathbf{a}_n \approx \dot{\mathbf{v}}(t_n + \Delta_\alpha h_{n-1})$ wird direkt durch das Generalized-α-Verfahren (3.55) berechnet. Eine Approximation von $\ddot{\mathbf{v}}(t_n)$ wird jedoch noch benötigt. Mit den verfügbaren Beschleunigungsvariablen kann diese im Allgemeinen durch

$$\ddot{\mathbf{v}}(t_n) = \frac{a_{1,n}\dot{\mathbf{v}}(t_n) + a_{2,n}\dot{\mathbf{v}}(t_{n+1}) + a_{3,n}\dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_{n-1}) + a_{4,n}\dot{\mathbf{v}}(t_{n+1} + \boldsymbol{\Delta}_\alpha h_n)}{h_n} + \mathcal{O}(h_n) \tag{4.5}$$

dargestellt werden, wobei

$$a_{1,n} = -1 + \left(\frac{\boldsymbol{\Delta}_\alpha}{\sigma_n} - 1\right) a_{3,n} + \boldsymbol{\Delta}_\alpha a_{4,n}, \tag{4.6a}$$

$$a_{2,n} = 1 - \frac{\boldsymbol{\Delta}_\alpha}{\sigma_n} a_{3,n} - (1 + \boldsymbol{\Delta}_\alpha) a_{4,n} \tag{4.6b}$$

und $a_{3,n}, a_{4,n} \in \mathbb{R}$ beliebig sind. Daher wird

$$\overline{\mathbf{a}}_n \coloneqq \mathbf{a}_n + \Delta_\alpha h_n \left(1 - \frac{1}{\sigma_n}\right) \ddot{\mathbf{v}}_n \tag{4.7}$$

gesetzt mit

$$\ddot{\mathbf{v}}_n \coloneqq \frac{a_{1,n}\dot{\mathbf{v}}_n + a_{2,n}\dot{\mathbf{v}}_{n+1} + a_{3,n}\mathbf{a}_n + a_{4,n}\mathbf{a}_{n+1}}{h_n}, \tag{4.8}$$

wobei $a_{1,n}$ und $a_{2,n}$ wie in (4.6) zu wählen sind. Es bleiben also zwei frei wählbare Parameter $a_{3,n}, a_{4,n} \in \mathbb{R}$ übrig. Diese sollen so gewählt werden, dass das Verfahren für möglichst viele Schrittweitenverhältnisse stabil bleibt (vgl. Satz 3). Beispielhaft werden die folgenden Näherungen untersucht, wobei jeweils zwei der $a_{i,n}$ null gesetzt wurden:

Näherung 1:

$$a_{1,n} = -\frac{\sigma_n}{\Delta_\alpha}, \quad a_{2,n} = 0, \quad a_{3,n} = \frac{\sigma_n}{\Delta_\alpha}, \quad a_{4,n} = 0, \tag{4.9a}$$

Näherung 2:

$$a_{1,n} = 0, \quad a_{2,n} = -\frac{\sigma_n}{\Delta_\alpha - \sigma_n}, \quad a_{3,n} = \frac{\sigma_n}{\Delta_\alpha - \sigma_n}, \quad a_{4,n} = 0, \tag{4.9b}$$

Näherung 3:

$$a_{1,n} = 0, \quad a_{2,n} = 0, \quad a_{3,n} = -\frac{\sigma_n}{(\sigma_n + (\sigma_n - 1)\Delta_\alpha)}, \quad a_{4,n} = \frac{\sigma_n}{(\sigma_n + (\sigma_n - 1)\Delta_\alpha)}, \tag{4.9c}$$

Näherung 4:

$$a_{1,n} = -1, \quad a_{2,n} = 1, \quad a_{3,n} = 0, \quad a_{4,n} = 0. \tag{4.9d}$$

Natürlich wären auch weitere Näherungen vorstellbar und können das Thema für weitere Untersuchungen sein. In [52] wurde lediglich Näherung 1 verwendet.
Ist die Schrittweite $h_n = h$ konstant, so gilt $\overline{\mathbf{a}}_n = \mathbf{a}_n$ und keine Änderung wurde vorgenommen.

Bemerkung 13 (Anpassung der Geschwindigkeit, vgl. [52])
Ähnlich zur Anpassung der Geschwindigkeiten in den Startwerten in Abschnitt 3.3.2 wird die Geschwindigkeit im Zeitschritt $t_{n-1} \to t_n$ von $\mathbf{v}_n \approx \mathbf{v}(t_n)$ in

$$\overline{\mathbf{v}}_n \coloneqq \mathbf{v}_n + \boldsymbol{\Delta}_n^{\mathbf{v}} \tag{4.10}$$

abgeändert. Dazu wird das Gleichungssystem

$$\begin{bmatrix} \mathbf{M}(q_n) & \mathbf{B}^\top(q_n) \\ \mathbf{B}(q_n) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta}_n^{\mathbf{v}} \\ \boldsymbol{\Delta}_n^{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ \mathbf{B}(q_n)\boldsymbol{\Delta}_n \mathbf{l}_n^q \end{bmatrix}$$

gelöst, wobei $\boldsymbol{\Delta}_n \mathbf{l}_n^q$ die Differenz zweier aufeinanderfolgender lokaler Abbruchfehler der Form (3.56) dividiert durch h_n bzw. h_{n-1} approximiert, also ist

$$\begin{aligned} \boldsymbol{\Delta}_n \mathbf{l}_n^q &\approx \frac{h_n^2}{6}\left((1 - 6\beta - 3\Delta_\alpha)\ddot{\mathbf{v}}(t_n) + \frac{1}{2}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n)\right) \\ &\quad - \frac{h_{n-1}^2}{6}\left((1 - 6\beta - 3\Delta_\alpha)\ddot{\mathbf{v}}(t_{n-1}) + \frac{1}{2}\widehat{\mathbf{v}}(t_{n-1})\dot{\mathbf{v}}(t_{n-1})\right) \\ &= \left(\frac{1}{6} - \frac{1}{2}\Delta_\alpha - \beta\right)(1 - 1/\sigma_n^2)h_n^2\ddot{\mathbf{v}}(t_n) + \frac{(1 - 1/\sigma_n^2)}{12}h_n^2\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h_{n-1}^3), \end{aligned}$$

mit (4.2). Daher wird

$$\boldsymbol{\Delta}_n \mathbf{l}_n^q = \left(\frac{1}{6} - \frac{1}{2}\Delta_\alpha - \beta\right)(1 - 1/\sigma_n^2)h_n^2\ddot{\mathbf{v}}_n + \frac{(1 - 1/\sigma_n^2)}{12}h_n^2\widehat{\mathbf{v}}_n\dot{\mathbf{v}}_n$$

gesetzt und $\ddot{\mathbf{v}}_n$ wie in Gleichung (4.8) gewählt mit $a_{1,n}$ und $a_{2,n}$ aus (4.6). Somit gilt

$$\boldsymbol{\Delta}_n^{\mathbf{v}} = \left[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\right](q_n)\boldsymbol{\Delta}_n \mathbf{l}_n^q.$$

Unabhängig von der Wahl der Näherung aus (4.9) ist stets $\boldsymbol{\Delta}_n \mathbf{l}_n^q = \mathbf{0}_{N\times 1}$ für $\sigma_n = 1$. Daher gilt $\overline{\mathbf{v}}_n = \mathbf{v}_n$, wenn sich in einem Zeitschritt die Schrittweite nicht verändert hat.

Werden die beiden vorherigen Bemerkungen beachtet und die Schrittweite h in (3.55) durch die variable $h_n \coloneqq t_{n+1} - t_n$ ersetzt, so kann das Generalized-α-Verfahren für variable Schrittweiten wie nachfolgend definiert werden.

Definition 21 (Generalized-α-Verfahren für variable Schrittweiten zur Lösung von (3.44))
Das *Generalized-α-Verfahren* zur Lösung von (3.44) verwendet die numerischen Lösungen q_n, $\mathbf{v}_n$ und $\mathbf{a}_n$ im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h_n$ mit Schrittweite h_n in

$$q_{n+1} = q_n \circ \exp\big(h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}_n}\big), \tag{4.11a}$$
$$\boldsymbol{\Delta}\mathbf{q}_n = \overline{\mathbf{v}}_n + (0.5-\beta)h_n\overline{\mathbf{a}}_n + \beta h_n \mathbf{a}_{n+1}, \tag{4.11b}$$
$$\mathbf{v}_{n+1} = \overline{\mathbf{v}}_n + (1-\gamma)h_n\overline{\mathbf{a}}_n + \gamma h_n \mathbf{a}_{n+1}, \tag{4.11c}$$
$$(1-\alpha_m)\mathbf{a}_{n+1} + \alpha_m\overline{\mathbf{a}}_n = (1-\alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n, \tag{4.11d}$$
$$\mathbf{M}\big(q_{n+1}\big)\dot{\mathbf{v}}_{n+1} = -\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) - \mathbf{B}^\top\big(q_{n+1}\big)\boldsymbol{\lambda}_{n+1}, \tag{4.11e}$$
$$\mathbf{0} = \boldsymbol{\Phi}\big(q_{n+1}\big) \tag{4.11f}$$

mit Parametern (2.20), $\overline{\mathbf{v}}_n$ aus (4.10) und $\overline{\mathbf{a}}_n$ aus (4.7) für $n \geq 1$. Ist $n = 0$, so wird $\overline{\mathbf{a}}_0 = \mathbf{a}_0$ und $\overline{\mathbf{v}}_0 = \mathbf{v}_0$ verwendet.

Bemerkung 14 (Startwerte des Generalized-α-Verfahrens (4.11))
Um die Ordnungsreduktion in den ersten Zeitschritten zu vermeiden, müssen vor dem ersten Zeitschritt die Anpassungen aus Bemerkungen 7 und 8 vorgenommen werden. Die Startwerte sind somit analog zu denen von (3.55) aus Abschnitt 3.3.2.

4.2 Lokale Abbruchfehler

Für die Konvergenzanalyse werden zunächst die lokalen Abbruchfehler bestimmt. Dazu wird die analytische Lösung in die Verfahrensvorschrift eingesetzt und der lokale Abbruchfehler durch die Taylorentwicklung bzw. die Magnusentwicklung abgeschätzt. Die Magnusentwicklung wird dabei für die Lösung der auf der Lie-Gruppe definierten gewöhnlichen Differentialgleichung $\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\mathbf{v}}(t)$ (vgl. Gleichung (2.25)) für vorgegebenes $\widetilde{\mathbf{v}}(t)$ verwendet.
Im nachfolgenden Beweis werden solche lokalen Abbruchfehler, teilweise mit matrixwertigen Funktionen multipliziert. Daher wird die Notation

$$\mathbf{l}_n^{(\mathbf{A}\bullet)} \coloneqq \mathbf{A}\big(t_n, q(t_n), \mathbf{v}(t_n), \boldsymbol{\lambda}(t_n)\big)\mathbf{l}_n^{(\bullet)}$$

für matrixwertige Funktionen $\mathbf{A} = \mathbf{A}(t, q, \mathbf{v}, \boldsymbol{\lambda})$ eingeführt.

Definition 22 (Lokale Abbruchfehler von (4.11))
Wegen (4.3) und (4.4) sind die lokalen Abbruchfehler des Generalized-α-Verfahrens für variable Schrittweiten (4.11) definiert durch

$$q(t_{n+1}) = q(t_n) \circ \exp\big(h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\big) \circ \exp\big(\widetilde{\mathbf{l}}_n^q\big), \tag{4.12a}$$
$$\boldsymbol{\Delta}\mathbf{q}(t_n) = \overline{\mathbf{v}}(t_n) + (0.5-\beta)h_n\dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_n) + \beta h_n\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h_n), \tag{4.12b}$$
$$\mathbf{v}(t_{n+1}) = \overline{\mathbf{v}}(t_n) + (1-\gamma)h_n\dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_n) + \gamma h_n\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h_n) + \mathbf{l}_n^{\mathbf{v}}, \tag{4.12c}$$
$$\begin{aligned}\mathbf{l}_n^{\mathbf{a}} = &-(1-\alpha_f)\dot{\mathbf{v}}(t_{n+1}) - \alpha_f\dot{\mathbf{v}}(t_n)\\ &+ \alpha_m\dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_n) + (1-\alpha_m)\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h_n),\end{aligned} \tag{4.12d}$$
$$\begin{aligned}\overline{\mathbf{v}}(t_n) = \mathbf{v}(t_n) + \mathbf{C}\big(q(t_n)\big)\Bigg(&\left(\frac{1}{6} - \frac{1}{2}\Delta_\alpha - \beta\right)(1 - 1/\sigma_n^2)h_n^2\ddot{\mathbf{v}}(t_n)\\ &+ \frac{(1-1/\sigma_n^2)}{12}h_n^2\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n)\Bigg)\end{aligned} \tag{4.12e}$$

mit $\mathbf{C}\big(q(t_n)\big) \coloneqq \big[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\big]\,\big(q(t_n)\big)$.

In (4.12a) wurde einer Vorgehensweise von Wensch [50] gefolgt, der die lokalen Fehler eines Lie-Gruppen-Integrators für Differentialgleichungen erster Ordnung in der Lie-Algebra $\mathfrak{g}$ untersucht hat. Gilt schließlich $\widetilde{\mathbf{l}}_n^q = \mathcal{O}(h^k)$ für $k \in \mathbb{N}$, so folgt Gleiches auch für das Äquivalent im euklidischen Raum $\mathbf{l}_n^q = \mathcal{O}(h^k)$.
Nun werden die Größenordnungen dieser lokalen Abbruchfehler bestimmt. Für konstante Schrittweiten erfolgte dies in [2, Lemma 1].

Satz 1
Die lokalen Abbruchfehler (4.12) des Generalized-α-Verfahrens mit variabler Schrittweite (4.11) genügen den Abschätzungen

$$\|\mathbf{l}_n^q\| = \mathcal{O}(h_n^3), \quad \|\mathbf{l}_n^{\mathbf{v}}\| = \mathcal{O}(h_n^2), \quad \|\mathbf{l}_n^{\mathbf{a}}\| = \mathcal{O}(h_n^2) \tag{4.13}$$

und

$$\left\|\mathbf{B}\big(q(t_n)\big)\left(\frac{\mathbf{l}_{n+1}^q}{h_{n+1}} - \frac{\mathbf{l}_n^q}{h_n} + \mathbf{l}_n^{\mathbf{v}}\right)\right\| = \mathcal{O}(h_n^3). \tag{4.14}$$

Beweis:

a) Das Einsetzen von (4.12e) in (4.12b) und die Taylorentwicklungen

$$\dot{\mathbf{v}}(t_n + \boldsymbol{\Delta}_\alpha h_n) = \dot{\mathbf{v}}(t_n) + \boldsymbol{\Delta}_\alpha h_n \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^2), \tag{4.15a}$$
$$\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h_n) = \dot{\mathbf{v}}(t_n) + (1 + \boldsymbol{\Delta}_\alpha) h_n \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^2) \tag{4.15b}$$

liefern

$$\begin{aligned}\boldsymbol{\Delta}\mathbf{q}(t_n) = \mathbf{v}(t_n) &+ \frac{h_n}{2}\dot{\mathbf{v}}(t_n) + \left(\frac{\Delta_\alpha}{2} + \beta\right) h_n^2 \ddot{\mathbf{v}}(t_n) \\ &+ \left(\frac{1}{6} - \frac{1}{2}\Delta_\alpha - \beta\right)(1 - 1/\sigma_n^2) h_n^2 \mathbf{C}\big(q(t_n)\big)\ddot{\mathbf{v}}(t_n) \\ &+ \frac{(1 - 1/\sigma_n^2)}{12} h_n^2 \mathbf{C}\big(q(t_n)\big)\widehat{\dot{\mathbf{v}}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^3).\end{aligned} \tag{4.16}$$

Wird (2.29) für $t = t_{n+1}$ und $m = n$ mit (4.12a) gleichgesetzt, so folgt [2]

$$q(t_n) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\big) = q(t_n) \circ \exp\big(h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\big) \circ \exp\big(\widetilde{\mathbf{l}}_n^q\big)$$

und daher

$$\exp\big(\widetilde{\mathbf{l}}_n^q\big) = \exp\big(-h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\big) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\big).$$

Dieses Produkt aus mehreren Matrixexponentiellen kann durch die Anwendung der Baker-Campbell-Hausdorff-Formel (2.28) untersucht werden, da die Abschätzungen $h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n) = \mathcal{O}(h_n)$ und $\widetilde{\boldsymbol{\nu}}_n(t_{n+1}) = \mathcal{O}(h_n)$ gelten (vgl. (2.33)). Der lokale Abbruchfehler kann somit durch [2]

$$\widetilde{\mathbf{l}}_n^q = \widetilde{\boldsymbol{\nu}}_n(t_{n+1}) - h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n) + \mathcal{O}(h_n)\left\|\widetilde{\boldsymbol{\nu}}_n(t_{n+1}) - h_n \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\right\|$$

abgeschätzt werden, weil

$$[\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2] = [\widetilde{\mathbf{w}}_1 + \widetilde{\mathbf{w}}_2, \widetilde{\mathbf{w}}_2] - \underbrace{[\widetilde{\mathbf{w}}_2, \widetilde{\mathbf{w}}_2]}_{=\mathbf{0}} = \mathcal{O}(h_n)\|\widetilde{\mathbf{w}}_1 + \widetilde{\mathbf{w}}_2\|$$

gilt, wenn $\widetilde{\mathbf{w}}_2 = \mathcal{O}(h_n)$ ist. Durch das Einsetzen von (2.33) und (4.16) und unter Verwendung von (2.35) folgt die Abschätzung

$$\mathbf{l}_n^q = \big(\mathbf{I} - \mathbf{C}\big(q(t_n)\big)(1 - 1/\sigma_n^2)\big) \left(\left(\frac{1}{6} - \frac{\Delta_\alpha}{2} - \beta \right) h_n^3 \ddot{\mathbf{v}}(t_n) + \frac{h_n^3}{12} \widehat{\mathbf{v}}(t_n) \dot{\mathbf{v}}(t_n) \right) + \mathcal{O}(h_n^4) \tag{4.17}$$

und damit die erste Behauptung in (4.13).

b) Aus Gleichung (4.12c), den Taylorentwicklungen (4.15) und

$$\mathbf{v}(t_{n+1}) = \mathbf{v}(t_n) + h_n \dot{\mathbf{v}}(t_n) + \frac{h_n^2}{2} \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^3)$$

folgt

$$\mathbf{l}_n^{\mathbf{v}} = -\mathbf{C}(q(t_n))(1 - 1/\sigma_n^2) \left(\left(\frac{1}{6} - \frac{1}{2}\Delta_\alpha - \beta \right) h_n^2 \ddot{\mathbf{v}}(t_n) + \frac{h_n^2}{12} \widehat{\mathbf{v}}(t_n) \dot{\mathbf{v}}(t_n) \right) + \mathcal{O}(h_n^3), \tag{4.18}$$

da $\left(\frac{1}{2} - \Delta_\alpha - \gamma\right) = 0$ mit (2.19) und (2.20) gilt.

c) Aus Gleichung (4.12d) und Taylorentwicklung folgt

$$\mathbf{l}_n^{\mathbf{a}} = -(\alpha_m - \alpha_f - \Delta_\alpha) h_n \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^2)$$

und dadurch die Behauptung mit (2.19).

d) Es gilt $\mathbf{B}\big(q(t_n)\big)\mathbf{C}\big(q(t_n)\big) = \mathbf{B}\big(q(t_n)\big)$ und damit

$$\mathbf{B}\big(q(t_n)\big)\mathbf{l}_n^q = \mathbf{B}\big(q(t_n)\big) \left(\left(\frac{1}{6} - \frac{\Delta_\alpha}{2} - \beta \right) \frac{h_n^3}{\sigma_n^2} \ddot{\mathbf{v}}(t_n) + \frac{h_n^3}{12\sigma_n^2} \widehat{\mathbf{v}}(t_n) \dot{\mathbf{v}}(t_n) \right) + \mathcal{O}(h_n^4)$$

mit dem lokalen Abbruchfehler (4.17). Daher folgt

$$\begin{aligned} \frac{\mathbf{B}\big(q(t_{n+1})\big)\mathbf{l}_{n+1}^q}{h_{n+1}} - \frac{\mathbf{B}\big(q(t_n)\big)\mathbf{l}_n^q}{h_n} &= \mathbf{B}\big(q(t_n)\big) \left(1 - \frac{1}{\sigma_n^2}\right) \left(\frac{1}{6} - \frac{1}{2}\Delta_\alpha - \beta \right) h_n^2 \ddot{\mathbf{v}}(t_n) \\ &\quad + \mathbf{B}\big(q(t_n)\big) \left(1 - \frac{1}{\sigma_n^2}\right) \frac{h_n^2}{12} \widehat{\mathbf{v}}(t_n) \dot{\mathbf{v}}(t_n) + \mathcal{O}(h_n^3) \\ &= -\mathbf{l}_n^{\mathbf{Bv}} + \mathcal{O}(h_n^3) \end{aligned}$$

mit den Gleichungen (4.18), (4.2), $\sigma_{n+1} = h_{n+1}/h_n$, $\mathbf{B}\big(q(t_{n+1})\big) = \mathbf{B}\big(q(t_n)\big) + \mathcal{O}(h_n)$ und $\mathbf{v}^{(i)}(t_{n+1}) = \mathbf{v}^{(i)}(t_n) + \mathcal{O}(h_n)$, $(i = 0, 1, 2)$.

■

Bemerkung 15 (Unterschiede zum Verfahren für konstante Schrittweiten)
Aufgrund der Änderung der Geschwindigkeit in (4.11c) tritt im Unterschied zum konstanten Fall ein zusätzlicher Fehlerterm im lokalen Abbruchfehler $\mathbf{l}_n^{\mathbf{v}}$ aus Gleichung (4.12c) auf. Dadurch ist eine Verringerung der Ordnung in ebendiesem zu beobachten. Für konstante Schrittweiten hätte in (4.13) daher $\|\mathbf{l}_n^{\mathbf{v}}\| = \mathcal{O}(h^3)$ bewiesen werden können. Für die Konvergenzanalyse ist diese Ordnungsreduktion jedoch unerheblich, da als weiterer Unterschied zum konstanten Fall der Ausdruck (4.14) von

Interesse ist und bei konstanter Schrittweite die Abschätzung $\|(\mathbf{l}_{n+1}^q - \mathbf{l}_n^q)/h\|$, vgl. [2, Lemma 1]. Durch die Hinzunahme von $\mathbf{B}\big(q(t_n)\big)\mathbf{l}_n^{\mathbf{v}}$ in (4.14) wird der zusätzliche Fehlerterm in $\mathbf{l}_n^{\mathbf{v}}$ kompensiert.
Außerdem wird in den nachfolgenden Betrachtungen der Projektor $\mathbf{P}(q) = \mathbf{I}_N - \big[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\big]\,(q)$ eingeführt. Durch diesen wird die Tangentialrichtung der Zwangsmannigfaltigkeit untersucht, vgl. Abschnitt 4.3.2. Die Multiplikation des lokalen Abbruchfehlers $\mathbf{l}_n^{\mathbf{v}}$ mit $\mathbf{P}(q)$ resultiert aufgrund von $[\mathbf{PC}](q) \equiv \mathbf{0}$ in $\|\mathbf{l}_n^{\mathbf{Pv}}\| = \mathcal{O}(h_n^3)$, vgl. (4.18), analog zum konstanten Fall.

4.3 Gleichungen für die globalen Fehler

Für die Konvergenzanalyse müssen zunächst die globalen Fehler der einzelnen Variablen definiert werden. Dabei bietet es sich an, wie zuvor beim lokalen Abbruchfehler $\mathbf{l}_n^q$ in (4.12a), den globalen Fehler in $q \in G$ als Element der entsprechenden Lie-Algebra $\widetilde{\mathbf{e}}_n^q \in \mathfrak{g}$ zu definieren [50]. Die restlichen Variablen liegen in linearen Räumen und die zugehörigen globalen Fehler können wie in Gleichung (A.6) definiert werden. Es folgt [2]

$$q(t_n) = q_n \circ \exp\left(\widetilde{\mathbf{e}}_n^q\right), \tag{4.19a}$$
$$\mathbf{v}(t_n) = \mathbf{v}_n + \mathbf{e}_n^{\mathbf{v}}, \tag{4.19b}$$
$$\boldsymbol{\lambda}(t_n) = \boldsymbol{\lambda}_n + \mathbf{e}_n^{\boldsymbol{\lambda}}, \tag{4.19c}$$
$$\dot{\mathbf{v}}(t_n) = \dot{\mathbf{v}}_n + \mathbf{e}_n^{\dot{\mathbf{v}}}, \tag{4.19d}$$
$$\dot{\mathbf{v}}(t_n + \Delta_\alpha h_{n-1}) = \mathbf{a}_n + \mathbf{e}_n^{\mathbf{a}}, \tag{4.19e}$$
$$\boldsymbol{\Delta}\mathbf{q}(t_n) = \boldsymbol{\Delta}\mathbf{q}_n + \mathbf{e}_n^{\boldsymbol{\Delta}\mathbf{q}}. \tag{4.19f}$$

Im nachfolgenden Beweis werden solche globalen Fehler teilweise mit matrixwertigen Funktionen multipliziert. Daher wird die Notation

$$\mathbf{e}_n^{(\mathbf{A}\bullet)} \coloneqq \mathbf{A}\big(t_n, q(t_n), \mathbf{v}(t_n), \boldsymbol{\lambda}(t_n)\big)\mathbf{e}_n^{(\bullet)}$$

für matrixwertige Funktionen $\mathbf{A} = \mathbf{A}(t, q, \mathbf{v}, \boldsymbol{\lambda})$ eingeführt.
Um die Gleichungen für die globalen Fehler aufzustellen, wird die Differenz aus der Verfahrensvorschrift mit der analytischen Lösung (4.12) und der Verfahrensvorschrift mit der numerischen Lösung (4.11) gebildet. Zur Abschätzung von Termen höherer Ordnung wird eine für Verfahren höherer Ordnung übliche Annahme getroffen [29, Theorem VII.3.5]. Dabei wird vorausgesetzt, dass die numerische Lösung q_n in einer Umgebung der analytischen Lösung $q(t_n)$ der Größe $\mathcal{O}(h_n)$ verbleibt. Die numerischen Lösungen $\mathbf{v}_n$, $\boldsymbol{\lambda}_n$ und $\mathbf{a}_n$ sollen in einer Umgebung der analytischen Lösungen $\mathbf{v}(t_n)$, $\boldsymbol{\lambda}(t_n)$ und $\dot{\mathbf{v}}(t_n + \Delta_\alpha h_{n-1})$ bleiben. Aus diesem Grund wird die nachfolgende technische Voraussetzung getroffen.

Voraussetzung 2
Es gibt positive Konstanten $\overline{h}$ und C_T, so dass für alle $h_n \in (0, \overline{h}]$ und $t_0 + \sum_{i=0}^{r} h_i \in [t_0, t_{\text{end}}]$ mit $n, r \in \mathbb{N}$ die globalen Fehler durch

$$\|\mathbf{e}_r^q\| \leq C_T h_r,\ \|\mathbf{e}_r^{\mathbf{v}}\| + \|\mathbf{e}_r^{\boldsymbol{\lambda}}\| + \|\mathbf{e}_r^{\mathbf{a}}\| \leq C_T$$

beschränkt bleiben.

Die Annahme ist vertretbar, da am Ende des Konvergenzbeweises gezeigt wird, dass alle genannten globalen Fehler mindestens mit einer Ordnung höher konvergieren. Um die Terme höherer Ordnung zusammenfassen zu können, wird die Definition

$$\epsilon_n \coloneqq \|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| + h_n\|\mathbf{e}_n^{\mathbf{a}}\| + h_n\|\mathbf{e}_n^{\boldsymbol{\lambda}}\|$$

eingeführt.
In den nachfolgenden Lemmata und Sätzen sollen die Voraussetzungen 1 und 2 stets erfüllt sein.

4.3.1 Elimination von $\mathbf{e}_n^{\dot{\mathbf{v}}}$, $\mathbf{e}_{n+1}^{\dot{\mathbf{v}}}$ und $\mathbf{e}_n^{\boldsymbol{\Delta q}}$

Aufgrund der Definitionen der globalen Fehler in (4.19) kann davon ausgegangen werden, dass Gleichungen berechnet werden, die $\mathbf{e}_n^q$, $\mathbf{e}_n^{\mathbf{v}}$, $\mathbf{e}_n^{\boldsymbol{\lambda}}$, $\mathbf{e}_n^{\dot{\mathbf{v}}}$, $\mathbf{e}_n^{\mathbf{a}}$ und $\mathbf{e}_n^{\boldsymbol{\Delta q}}$ enthalten. Einige dieser globalen Fehler sollen jedoch eliminiert werden. Dies sind $\mathbf{e}_n^{\dot{\mathbf{v}}}$ bzw. $\mathbf{e}_{n+1}^{\dot{\mathbf{v}}}$ im nachfolgenden Lemma 8 und $\mathbf{e}_n^{\boldsymbol{\Delta q}}$ in Lemma 10. Eine Abschätzung für erstere ergibt sich analog zum Verfahren für konstante Schrittweiten und soll an dieser Stelle nicht erneut bewiesen werden. Dabei wird die Gleichgewichtsgleichung (3.44b) für $t = t_n$ und ihr Pendant im Generalized-α-Verfahren (4.11e) verwendet [2, Lemma 3].

Lemma 8 ([2, Lemma 3])
Für die globalen Fehler $\mathbf{e}_n^{\dot{\mathbf{v}}}$ des Generalized-α-Verfahrens (4.11) gelten die Fehlerabschätzungen

$$\begin{aligned}
\mathbf{e}_n^{\dot{\mathbf{v}}} &= -\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} + \mathcal{O}(1)\epsilon_n, \\
\mathbf{e}_{n+1}^{\dot{\mathbf{v}}} + \mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} &= \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3).
\end{aligned}$$

Für den globalen Fehler $\mathbf{e}_n^{\boldsymbol{\Delta q}}$ müssen im Unterschied zum konstanten Fall die Anpassungen von $\mathbf{v}_n$ und $\mathbf{a}_n$ aus den Bemerkungen 12 und 13 vor jedem Zeitschritt beachtet werden. Daher schätzt das nachfolgende Lemma die Fehler ab, die durch die Änderung von $\mathbf{v}_n$ in $\overline{\mathbf{v}}_n$ und $\mathbf{a}_n$ in $\overline{\mathbf{a}}_n$ entstehen.

Lemma 9
Gegeben sei das Generalized-α-Verfahren (4.11), dann gelten die globalen Fehlergleichungen

$$\ddot{\mathbf{v}}(t_n) - \ddot{\mathbf{v}}_n = \frac{1}{h_n}\big(a_{1,n}\mathbf{e}_n^{\dot{\mathbf{v}}} + a_{2,n}\mathbf{e}_{n+1}^{\dot{\mathbf{v}}} + a_{3,n}\mathbf{e}_n^{\mathbf{a}} + a_{4,n}\mathbf{e}_{n+1}^{\mathbf{a}}\big) + \mathcal{O}(h_n), \quad (4.20a)$$

$$\begin{aligned}
\dot{\mathbf{v}}(t_n + \Delta_\alpha h_n) - \overline{\mathbf{a}}_n = {} & -\Delta_\alpha a_{1,n}\left(1 - 1/\sigma_n\right)\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} - \Delta_\alpha a_{2,n}\left(1 - 1/\sigma_n\right)\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} \\
& + \left(1 + \Delta_\alpha a_{3,n}\left(1 - 1/\sigma_n\right)\right)\mathbf{e}_n^{\mathbf{a}} + \Delta_\alpha a_{4,n}\left(1 - 1/\sigma_n\right)\mathbf{e}_{n+1}^{\mathbf{a}} \\
& + \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2), \quad (4.20b)
\end{aligned}$$

$$\begin{aligned}
\overline{\mathbf{v}}(t_n) - \overline{\mathbf{v}}_n = {} & \mathbf{e}_n^{\mathbf{v}} + \left(1 - 1/\sigma_n^2\right)\left(\frac{1}{6} - \frac{\Delta_\alpha}{2} - \beta\right)h_n\Big(-a_{1,n}\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} \\
& - a_{2,n}\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} + a_{3,n}\mathbf{e}_n^{\mathbf{Ca}} + a_{4,n}\mathbf{e}_{n+1}^{\mathbf{Ca}}\Big) + \mathcal{O}(h_n)\epsilon_n \\
& + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3). \quad (4.20c)
\end{aligned}$$

Beweis:

a) Unter Verwendung von (4.19d), (4.5), (4.8) und (4.19e) gilt Gleichung (4.20a).

b) Mit den Gleichungen (4.19d), (4.4), (4.7), (4.19e) und (4.20a) kann

$$\begin{aligned}\dot{\mathbf{v}}(t_n+\Delta_\alpha h_n)-\overline{\mathbf{a}}_n = {}& \Delta_\alpha a_{1,n}\,(1-1/\sigma_n)\,\mathbf{e}_n^{\dot{\mathbf{v}}}+\Delta_\alpha a_{2,n}\,(1-1/\sigma_n)\,\mathbf{e}_{n+1}^{\dot{\mathbf{v}}}\\ &+(1+\Delta_\alpha a_{3,n}\,(1-1/\sigma_n))\mathbf{e}_n^{\mathbf{a}}+\Delta_\alpha a_{4,n}\,(1-1/\sigma_n)\,\mathbf{e}_{n+1}^{\mathbf{a}}\\ &+\mathcal{O}(h_n^2)\end{aligned}$$

bewiesen werden. Gleichung (4.20b) folgt mit Lemma 8.

c) Die Gleichungen (4.19), (4.10), (4.12e) und (4.19e) liefern

$$\begin{aligned}\overline{\mathbf{v}}(t_n)-\overline{\mathbf{v}}_n &= \mathbf{e}_n^{\mathbf{v}}+\mathbf{C}\big(q(t_n)\big)(1-1/\sigma_n^2)\left(\left(\frac{1}{6}-\frac{\Delta_\alpha}{2}-\beta\right)h_n^2(\ddot{\mathbf{v}}(t_n)-\ddot{\mathbf{v}}_n)\right.\\ &\quad\left.+\frac{h_n^2}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n)-\frac{h_n^2}{12}\widehat{\overline{\mathbf{v}}(t_n)-\mathbf{e}_n^{\mathbf{v}}}(\dot{\mathbf{v}}(t_n)-\mathbf{e}_n^{\dot{\mathbf{v}}})\right)+\mathcal{O}(h_n^2)\epsilon_n\\ &= \mathbf{e}_n^{\mathbf{v}}+\mathbf{C}\big(q(t_n)\big)(1-1/\sigma_n^2)\left(\frac{1}{6}-\frac{\Delta_\alpha}{2}-\beta\right)h_n\left(a_{1,n}\mathbf{e}_n^{\dot{\mathbf{v}}}\right.\\ &\quad\left.+a_{2,n}\mathbf{e}_{n+1}^{\dot{\mathbf{v}}}+a_{3,n}\mathbf{e}_n^{\mathbf{a}}+a_{4,n}\mathbf{e}_{n+1}^{\mathbf{a}}\right)+\mathcal{O}(h_n^2)\|\mathbf{e}_n^{\dot{\mathbf{v}}}\|+\mathcal{O}(h_n^2)\epsilon_n\end{aligned}$$

mit (4.20a) und $\mathbf{C}(q_n)=\mathbf{C}\big(q(t_n)\circ\exp\big(-\widetilde{\mathbf{e}}_n^q\big)\big)=\mathbf{C}\big(q(t_n)\big)+\mathcal{O}(1)\epsilon_n$. Die Behauptung folgt mit Lemma 8 und der Feststellung, dass

$$\left[\mathbf{C}\mathbf{M}^{-1}\mathbf{B}^\top\right](q)=\left[\mathbf{M}^{-1}\mathbf{B}^\top\right](q) \tag{4.21}$$

gilt.

■

Lemma 10
Für den globalen Fehler $\mathbf{e}_n^{\mathbf{\Delta q}}$ des Generalized-α-Verfahrens (4.11) gilt die globale Fehlergleichung

$$\begin{aligned}\mathbf{e}_n^{\mathbf{\Delta q}} = {}& \mathbf{e}_n^{\mathbf{v}}-\big(b_{1,n}+c_{1,n}\big)h_n\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}-\big(b_{2,n}+c_{2,n}\big)h_n\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}\\ &+\big(b_{3,n}\mathbf{C}\big(q(t_n)\big)+c_{3,n}+0.5-\beta\big)h_n\mathbf{e}_n^{\mathbf{a}}+\big(b_{4,n}\mathbf{C}\big(q(t_n)\big)+c_{4,n}+\beta\big)h_n\mathbf{e}_{n+1}^{\mathbf{a}}\\ &+\mathcal{O}(h_n)\epsilon_n+\mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\|+\|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big)+\mathcal{O}(h_n^3)\end{aligned} \tag{4.22}$$

mit

$$b_{i,n}\coloneqq\left(1-\frac{1}{\sigma_n^2}\right)\left(\frac{1}{6}-\frac{\Delta_\alpha}{2}-\beta\right)a_{i,n},\quad(i=1,2,3,4), \tag{4.23a}$$

$$c_{i,n}\coloneqq\left(\frac{1}{2}-\beta\right)\left(1-\frac{1}{\sigma_n}\right)\Delta_\alpha a_{i,n},\quad(i=1,2,3,4), \tag{4.23b}$$

und somit die Abschätzung

$$\|\mathbf{e}_n^{\mathbf{\Delta q}}\|=\mathcal{O}(1)(\epsilon_n+\epsilon_{n+1})+\mathcal{O}(h_n)\|\mathbf{e}_{n+1}^{\mathbf{a}}\|+\mathcal{O}(h_n^3).$$

Beweis:
Gleichung (4.22) folgt aus der Differenz der Gleichungen (4.12b) und (4.11b) mit (4.19), Lemma 9 und unter Einführung der Bezeichnungen (4.23).

■

4.3.2 Gleichungen der differentiellen Komponenten $\mathbf{e}_n^q$ und $\mathbf{e}_n^{\mathbf{v}}$

Zunächst werden die Gleichungen für die differentiellen Komponenten, also für die globalen Fehler $\mathbf{e}_n^q$ und $\mathbf{e}_n^{\mathbf{v}}$, benötigt. Aufgrund der Besonderheit, dass $\widetilde{\mathbf{e}}_n^q$ in der Lie-Algebra $\mathfrak{g}$ definiert ist, kann im Unterschied zu den anderen globalen Fehlern nicht die Differenz von (4.12a) und (4.11a) gebildet werden, um eine Fehlergleichung für $\mathbf{e}_n^q$ zu erhalten. Die untenstehende Vorgehensweise der Abschätzung stammt aus dem Beweis für konstante Schrittweiten in [2, Lemma 2] und wird an dieser Stelle auch für Verfahren mit variablen Schrittweiten verwendet. Dabei spielt die Baker-Campbell-Hausdorff-Formel eine entscheidende Rolle.

Lemma 11
Für das Generalized-α-Verfahren (4.11) gilt die Fehlergleichung

$$\mathbf{e}_{n+1}^q = \mathbf{e}_n^q + \mathcal{O}(h_n)\epsilon_n + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3) \tag{4.24}$$

und für das skalierte Inkrement der globalen Fehler $\mathbf{e}_n^q$ folgt die Abschätzung

$$\begin{aligned}\boldsymbol{\Delta}_{h_n}\mathbf{e}_n^q = \mathbf{e}_n^{\mathbf{v}} &+ \widehat{\mathbf{e}}_n^q\mathbf{v}(t_n) - \big(b_{1,n} + c_{1,n}\big)h_n\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} \\ &- \big(b_{2,n} + c_{2,n}\big)h_n\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} + \big(b_{3,n}\mathbf{C}\big(q(t_n)\big) + c_{3,n} + 0.5 - \beta\big)h_n\mathbf{e}_n^{\mathbf{a}} \\ &+ \big(b_{4,n}\mathbf{C}\big(q(t_n)\big) + c_{4,n} + \beta\big)h_n\mathbf{e}_{n+1}^{\mathbf{a}} + \frac{1}{h_n}\mathbf{l}_n^q + \mathcal{O}(h_n)\epsilon_n \\ &+ \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3),\end{aligned} \tag{4.25}$$

wobei $\boldsymbol{\Delta}_{h_n}\mathbf{e}_n^q \coloneqq (\mathbf{e}_{n+1}^q - \mathbf{e}_n^q)/h_n$ ist.

Beweis:
Da hierfür nur der Zeitschritt $t_n \to t_{n+1} = t_n + h_n$ betrachtet werden muss, würde das Ergebnis analog zu [2, Lemma 2] folgen, wenn h durch h_n ersetzt wird. Jedoch müssen die Veränderungen von $\mathbf{v}_n$ in $\overline{\mathbf{v}}_n$ und $\mathbf{a}_n$ in $\overline{\mathbf{a}}_n$ beachtet werden (vgl. Lemma 9). Daher soll der Beweis in einer etwas gekürzten Version an dieser Stelle mit den genannten Besonderheiten erneut geführt werden.
Mit (4.19a), (4.11a) und (4.12a) folgt [2]

$$\begin{aligned}\exp\big(\widetilde{\mathbf{e}}_{n+1}^q\big) &= q_{n+1}^{-1} \circ q(t_{n+1}) \\ &= \exp\big(-h_n\widetilde{\boldsymbol{\Delta}\mathbf{q}}_n\big) \circ q_n^{-1} \circ q(t_n) \circ \exp\big(h_n\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\big) \circ \exp\big(\widetilde{\mathbf{l}}_n^q\big) \\ &= \exp\big(h_n\widetilde{\mathbf{e}}_n^{\boldsymbol{\Delta}\mathbf{q}} - h_n\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\big) \circ \exp\big(\widetilde{\mathbf{e}}_n^q\big) \circ \exp\big(h_n\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\big) \circ \exp\big(\widetilde{\mathbf{l}}_n^q\big).\end{aligned}$$

Unter Voraussetzung 2 und mit Satz 1 kann die Baker-Campbell-Hausdorff-Formel (2.28) angewendet werden, da die Zusammenhänge $\widetilde{\mathbf{e}}_n^q = \mathcal{O}(h_n)$, $h_n\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n) = \mathcal{O}(h_n)$, $\widetilde{\mathbf{l}}_n^q = \mathcal{O}(h_n)$ und mit Lemma 10 und Gleichung (4.12b)

$$\begin{aligned}h_n\mathbf{e}_n^{\boldsymbol{\Delta}\mathbf{q}} - h_n\boldsymbol{\Delta}\mathbf{q}(t_n) &= -h_n\mathbf{v}(t_n) + \mathcal{O}(h_n)\epsilon_n + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2) \\ &= \mathcal{O}(h_n)\end{aligned}$$

gelten. Daher folgt für das Argument der Exponentialabbildung (vgl. [2])

$$\mathbf{e}_{n+1}^q = \mathbf{e}_n^q + h_n\mathbf{e}_n^{\boldsymbol{\Delta}\mathbf{q}} + h_n\widehat{\mathbf{e}}_n^q\mathbf{v}(t_n) + \mathbf{l}_n^q + \mathcal{O}(h_n^2)\big(\epsilon_n + h_n\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big)\big) + \mathcal{O}(h_n^4).$$

Mit Lemma 10 kann der Zusammenhang

$$\begin{aligned}\mathbf{e}_{n+1}^{q} = \mathbf{e}_n^q + h_n\mathbf{e}_n^{\mathbf{v}} + h_n\widehat{\mathbf{e}}_n^{q}\mathbf{v}(t_n) - \big(b_{1,n}+c_{1,n}\big)h_n^2\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} \\ - \big(b_{2,n}+c_{2,n}\big)h_n^2\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} + \big(b_{3,n}\mathbf{C}\big(q(t_n)\big)+c_{3,n}+0.5-\beta\big)h_n^2\mathbf{e}_n^{\mathbf{a}} \\ + \big(b_{4,n}\mathbf{C}\big(q(t_n)\big)+c_{4,n}+\beta\big)h_n^2\mathbf{e}_{n+1}^{\mathbf{a}} + \mathbf{l}_n^q + \mathcal{O}(h_n^2)\epsilon_n \\ + \mathcal{O}(h_n^3)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^4)\end{aligned}$$

gezeigt werden, womit Gleichung (4.25) erfüllt ist. Gleichung (4.24) folgt mit Satz 1.

■

Um optimale Fehlerabschätzungen zu erhalten, erfolgt die Fehlerrekursion separat in Tangential- und Normalenrichtung der Zwangsmannigfaltigkeit $\mathfrak{M} = \{q \in G : \mathbf{\Phi}(q) = \mathbf{0}\}$, vgl. [2, 29]. Die Tangentialrichtung kann durch Multiplikation mit dem Projektor

$$\mathbf{P}(q) \coloneqq \mathbf{I}_N - \left[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\right](q) \tag{4.26}$$

erhalten werden, da $\mathbf{P}(q)$ in den Tangentialraum $T_q\mathfrak{M} = \ker\mathbf{B}(q)$ projiziert und es gilt $[\mathbf{PP}](q) = \mathbf{P}(q)$ und $[\mathbf{BP}](q) = \left[\mathbf{B} - (\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\right](q) = \mathbf{0}$ [2]. Der Fehler in Normalenrichtung kann durch Multiplikation mit $\mathbf{B}(q)$ erhalten werden [2]. Dies bedeutet, dass der globale Fehler in $\mathbf{v}$ durch $\mathbf{e}_n^{\mathbf{v}}$, $\mathbf{e}_n^{\mathbf{Pv}}$ und $\mathbf{e}_n^{\mathbf{Bv}}$ untersucht wird.

Lemma 12
Es gelten die Gleichungen der globalen Fehler für das Generalized-α-Verfahren (4.11)

$$\mathbf{e}_{n+1}^{\mathbf{v}} = \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2), \tag{4.27a}$$

$$\begin{aligned}\mathbf{e}_{n+1}^{\mathbf{Pv}} = \mathbf{e}_n^{\mathbf{Pv}} + (d_{3,n}+1-\gamma)h_n\mathbf{e}_n^{\mathbf{Pa}} + \big(d_{4,n}+\gamma\big)h_n\mathbf{e}_{n+1}^{\mathbf{Pa}} + \mathcal{O}(h_n)\epsilon_n \\ + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3)\end{aligned} \tag{4.27b}$$

$$\begin{aligned}\mathbf{e}_{n+1}^{\mathbf{Bv}} = \mathbf{e}_n^{\mathbf{Bv}} + \big(b_{3,n}+d_{3,n}+1-\gamma\big)h_n\mathbf{e}_n^{\mathbf{Ba}} + \big(b_{4,n}+d_{4,n}+\gamma\big)h_n\mathbf{e}_{n+1}^{\mathbf{Ba}} \\ - \big(b_{1,n}+d_{1,n}\big)h_n\mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}} - \big(b_{2,n}+d_{2,n}\big)h_n\mathbf{e}_{n+1}^{\mathbf{S}\boldsymbol{\lambda}} + \mathbf{l}_n^{\mathbf{Bv}} \\ + \mathcal{O}(h_n)\epsilon_n + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3)\end{aligned} \tag{4.27c}$$

mit (4.23), $\mathbf{S}(q) \coloneqq \mathbf{B}(q)\mathbf{M}^{-1}(q)\mathbf{B}^\top(q)$ und

$$d_{i,n} \coloneqq (1-\gamma)\Delta_\alpha a_{i,n}\left(1-\frac{1}{\sigma_n}\right), \quad (i=1,2,3,4). \tag{4.28}$$

Beweis:
Unter Verwendung von Lemma 9 folgt aus der Differenz von (4.12c) und (4.11c) mit den Bezeichnungen (4.23) und (4.28)

$$\begin{aligned}\mathbf{e}_{n+1}^{\mathbf{v}} = \mathbf{e}_n^{\mathbf{v}} + \big(b_{3,n}\mathbf{C}\big(q(t_n)\big)+d_{3,n}+1-\gamma\big)h_n\mathbf{e}_n^{\mathbf{a}} + \big(b_{4,n}\mathbf{C}\big(q(t_n)\big)+d_{4,n}+\gamma\big)h_n\mathbf{e}_{n+1}^{\mathbf{a}} \\ - \big(b_{1,n}+d_{1,n}\big)h_n\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} - \big(b_{2,n}+d_{2,n}\big)h_n\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} \\ + \mathbf{l}_n^{\mathbf{v}} + \mathcal{O}(h_n)\epsilon_n + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^3).\end{aligned} \tag{4.29}$$

Mit Satz 1 gilt (4.27a). Durch Linksmultiplikation von (4.29) mit $\mathbf{P}\big(q(t_{n+1})\big)$ wird durch $\mathbf{P}\big(q(t_{n+1})\big) = \mathbf{P}\big(q(t_n)\big) + \mathcal{O}(h_n)$, $[\mathbf{PC}](q) \equiv \mathbf{0}$ und

$$\left[\mathbf{P}\mathbf{M}^{-1}\mathbf{B}^\top\right](q) \equiv \mathbf{0} \tag{4.30}$$

Gleichung (4.27b) unter Beachtung von $\|\mathbf{l}_n^{\mathbf{Pv}}\| = \mathcal{O}(h_n^3)$ (vgl. (4.18) und Bemerkung 15) erhalten.
Wird (4.29) von links mit $\mathbf{B}\big(q(t_{n+1})\big)$ multipliziert, so kann Gleichung (4.27c) gezeigt werden, da $[\mathbf{BC}](q) = \mathbf{B}(q)$ und

$$\mathbf{B}\big(q(t_{n+1})\big) = \mathbf{B}\big(q(t_n)\big) + \mathcal{O}(h_n)$$

gelten.

■

Bemerkung 16 (Unterschiedliche Verwendung von $\mathbf{e}_n^{\mathbf{v}}$ für konstante und variable Schrittweiten)
Im Beweis des Verfahrens für konstante Schrittweiten wird der Fehler $\mathbf{e}_n^{\mathbf{v}}$ nur in Normalenrichtung, also $\mathbf{e}_n^{\mathbf{Bv}}$, näher untersucht und ansonsten die ursprüngliche Fehlergleichung verwendet. In dieser Gleichung hätte eine Transformation in Tangentialrichtung durch Multiplikation mit $\mathbf{P}$ (vgl. (4.26)) die Abschätzung nicht wesentlich verändert. Wie das vorausgehende Lemma gezeigt hat, ist dies für variable Schrittweiten anders, da der lokale Abbruchfehler $\mathbf{l}_n^{\mathbf{v}}$ aufgrund der Änderung der Geschwindigkeit $\mathbf{v}_n$ in $\overline{\mathbf{v}}_n$ nur in Tangentialrichtung die Ordnung drei besitzt. Daher müssen im Fall von variablen Schrittweiten Gleichungen in beiden Richtungen für die globalen Fehler $\mathbf{e}_n^{\mathbf{Bv}}$ und $\mathbf{e}_n^{\mathbf{Pv}}$ aufgestellt werden.

4.3.3 Gleichungen der algebraischen Komponenten $\mathbf{e}_n^{\boldsymbol{\lambda}}$ und $\mathbf{e}_n^{\mathbf{a}}$

Um Gleichungen für die algebraischen Komponenten für DAEs vom Index 3 aufzustellen, werden Differenzenapproximationen der (versteckten) Zwangsbedingungen mit geeigneten Schätzern für die Approximationsfehler in linearen Räumen kombiniert, vgl. [1]. Ähnliche Ergebnisse werden auch für Konfigurationsräume mit Lie-Gruppen-Struktur benötigt und es können Abschätzungen für die Produkte von $\mathbf{B}(q)$ mit den Fehlertermen $\mathbf{e}_n^q$ und $\boldsymbol{\Delta}_h\mathbf{e}_n^q$ erhalten werden.

Lemma 13 ([2, Lemma 4], [4, Lemma 4.7])
Die globalen Fehler $\mathbf{e}_n^q$ erfüllen die Abschätzung

$$\begin{aligned}\mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_{h_n}\mathbf{e}_n^q + \mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) = \mathcal{O}(h_n)\epsilon_n + \mathcal{O}(h_n^2)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big)\\ + \mathcal{O}(h_n^3).\end{aligned} \tag{4.31}$$

Beweis:
Der Beweis erfolgt analog zu [4, Lemma 4.7]. Lediglich h wird durch die variable Schrittweite h_n ersetzt. Dabei wird zunächst mit dem Hauptsatz der Differential- und Integralrechnung der Zusammenhang

$$\begin{aligned}\mathbf{0}_N &= -\big(\boldsymbol{\Phi}\big(q_n\big) - \boldsymbol{\Phi}\big(q(t_n)\big)\big)\\ &= -\big(\boldsymbol{\Phi}\big(q(t_n)\circ\exp\big(-1\cdot\widetilde{\mathbf{e}}_n^q\big)\big) - \boldsymbol{\Phi}\big(q(t_n)\circ\exp\big(-0\cdot\widetilde{\mathbf{e}}_n^q\big)\big)\big)\\ &= -\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\vartheta}\boldsymbol{\Phi}\big(q(t_n)\circ\exp\big(-\vartheta\widetilde{\mathbf{e}}_n^q\big)\big)\mathrm{d}\vartheta\\ &= \int_0^1 \mathbf{B}\big(q(t_n)\circ\exp\big(-\vartheta\widetilde{\mathbf{e}}_n^q\big)\big)\mathbf{e}_n^q\mathrm{d}\vartheta\\ &= \mathbf{B}\big(q(t_n)\big)\mathbf{e}_n^q + \mathcal{O}(h_n)\|\mathbf{e}_n^q\|\end{aligned}$$

unter Verwendung von (4.11f), (3.44c), (4.19a) und (3.45) gezeigt. Eine erneute Anwendung des Hauptsatzes der Differential- und Integralrechnung führt in ähnlicher Weise auf

$$\mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_{h_n}\mathbf{e}_n^q + \mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) = \mathcal{O}(h_n)\big(\|\mathbf{e}_n^q\| + \|\boldsymbol{\Delta}_{h_n}\mathbf{e}_n^q\|\big)$$

mit dem Krümmungsterm $\mathbf{Z}$ aus (3.47), der die Ableitung von $\mathbf{B}$ darstellt.
Da mit (4.25) und Satz 1

$$\|\boldsymbol{\Delta}_{h_n}\mathbf{e}_n^q\| = \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2) \tag{4.32}$$

gilt, folgt die Behauptung. ■

Bemerkung 17 (Vernachlässigung des Approximationsfehlers im Newton-Raphson–Verfahren)
Im Beweis der Konvergenz des Generalized-α-Verfahrens für konstante Schrittweiten h in [2, Lemma 4] wurden die Gleichungen

$$\begin{aligned} -\boldsymbol{\Phi}(q_n) &= \mathbf{B}\big(q(t_n)\big)\mathbf{e}_n^q + \mathcal{O}(h)\|\mathbf{e}_n^q\|, \\ -\frac{\boldsymbol{\Phi}(q_{n+1}) - \boldsymbol{\Phi}(q_n)}{h} &= \mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_h\mathbf{e}_n^q + \mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) + \mathcal{O}(h)\big(\|\mathbf{e}_n^q\| + \|\boldsymbol{\Delta}_h\mathbf{e}_n^q\|\big) \end{aligned}$$

bewiesen. Da mit (3.55f) jedoch stets $\boldsymbol{\Phi}(q_i) = \mathbf{0}$ für $i > 0$ gilt, können ebenso die Gleichungen aus Lemma 13 bzw. [4, Lemma 4.7] verwendet werden. In diesem Fall wird der Fehler, der durch den Abbruch des Newton-Raphson-Verfahrens nach endlich vielen Iterationsschritten entsteht, vernachlässigt.

Analog zu den Geschwindigkeiten $\mathbf{v}$ in Lemma 12 wird der Fehler in den Beschleunigungen $\mathbf{a}$ in Tangential- und Normalenrichtung untersucht. Für den konstanten Fall erfolgte dies in [2, Lemma 5].

Lemma 14
Die globalen Fehler $\mathbf{e}_n^{\mathbf{a}}$ und $\mathbf{e}_n^{\boldsymbol{\lambda}}$ des Generalized-α-Verfahrens (4.11) genügen den Abschätzungen

$$\begin{aligned} &\big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)\mathbf{e}_{n+1}^{\mathbf{Pa}} + \alpha_m(1 + \Delta_\alpha a_{3,n}\,(1 - 1/\sigma_n))\mathbf{e}_n^{\mathbf{Pa}} \\ &= \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2), \qquad (4.33a) \\ &\big(1 - \alpha_f - \alpha_m\Delta_\alpha a_{2,n}\,(1 - 1/\sigma_n)\,\big)\mathbf{e}_{n+1}^{\mathbf{S}\boldsymbol{\lambda}} + \big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)\mathbf{e}_{n+1}^{\mathbf{Ba}} \\ &= \big(-\alpha_f + \alpha_m\Delta_\alpha a_{1,n}\,(1 - 1/\sigma_n)\,\big)\mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}} - \alpha_m(1 + \Delta_\alpha a_{3,n}\,(1 - 1/\sigma_n))\mathbf{e}_n^{\mathbf{Ba}} + \mathcal{O}(1)\epsilon_n \\ &\quad + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2). \qquad (4.33b) \end{aligned}$$

Beweis:
Durch die Differenz von (4.12d) mit (4.11d) ergibt sich mit den Lemmata 8, 9 und Satz 1

$$\begin{aligned} &\big(1 - \alpha_f - \alpha_m\Delta_\alpha a_{2,n}\,(1 - 1/\sigma_n)\,\big)\mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} + \big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)\mathbf{e}_{n+1}^{\mathbf{a}} \\ &= \big(-\alpha_f + \alpha_m\Delta_\alpha a_{1,n}\,(1 - 1/\sigma_n)\,\big)\mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} - \alpha_m(1 + \Delta_\alpha a_{3,n}\,(1 - 1/\sigma_n))\mathbf{e}_n^{\mathbf{a}} + \mathcal{O}(1)\epsilon_n \\ &\quad + \mathcal{O}(h_n)\big(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|\big) + \mathcal{O}(h_n^2). \qquad (4.34) \end{aligned}$$

Die Behauptung folgt durch Linksmultiplikation mit $\mathbf{P}\big(q(t_{n+1})\big)$ bzw. $\mathbf{B}\big(q(t_{n+1})\big)$ aufgrund von (4.30), $\mathbf{P}\big(q(t_{n+1})\big) = \mathbf{P}\big(q(t_n)\big) + \mathcal{O}(h_n)$ und $\mathbf{B}\big(q(t_{n+1})\big) = \mathbf{B}\big(q(t_n)\big) + \mathcal{O}(h_n)$.
■

Die globalen Fehler der algebraischen Komponenten zur Zeit t_{n+1} lassen sich durch die Komponenten zur Zeit t_n abschätzen.

Lemma 15 (vgl. [2, Corollary 1])
Gegeben sei das Generalized-α-Verfahren (4.11). Die skalierten globalen Fehler in den algebraischen Komponenten sind beschränkt durch

$$h_n\big(\|\mathbf{e}^{\mathbf{a}}_{n+1}\| + \|\mathbf{e}^{\dot{\mathbf{v}}}_{n+1}\| + \|\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}\|\big) = \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2).$$

Beweis:
Aus Gleichung (4.34) folgt durch Multiplikation mit h_n

$$\begin{aligned}
&\big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)h_n\mathbf{e}^{\mathbf{a}}_{n+1}\\
&= -\big(1 - \alpha_f - \alpha_m\Delta_\alpha a_{2,n}\,(1 - 1/\sigma_n)\,\big)h_n\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1} + \mathcal{O}(1)\epsilon_n\\
&\quad + \mathcal{O}(h_n^2)\big(\|\mathbf{e}^{\mathbf{a}}_{n+1}\| + \|\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}\|\big) + \mathcal{O}(h_n^3).
\end{aligned}$$

Für $1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big) \neq 0$ ist somit

$$\begin{aligned}
h_n\mathbf{e}^{\mathbf{a}}_{n+1} = &-\frac{\big(1 - \alpha_f - \alpha_m\Delta_\alpha a_{2,n}\,(1 - 1/\sigma_n)\,\big)}{\big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)}h_n\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1} + \mathcal{O}(1)\epsilon_n\\
&+ \mathcal{O}(h_n^2)\|\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}\| + \mathcal{O}(h_n^3)
\end{aligned} \tag{4.35}$$

für hinreichend kleines $h_n > 0$ erfüllt. Wird dies in (4.25) eingesetzt, so folgt

$$\begin{aligned}
&\boldsymbol{\Delta}_{h_n}\mathbf{e}^q_n\\
&= -\left(b_{2,n} + c_{2,n} + \big(b_{4,n} + c_{4,n} + \beta\big)\frac{\big(1 - \alpha_f - \alpha_m\Delta_\alpha a_{2,n}\,(1 - 1/\sigma_n)\,\big)}{\big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)}\right)h_n\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1}\\
&\quad + \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2)\|\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}\| + \mathcal{O}(h_n^2)
\end{aligned}$$

wegen (4.21) und Satz 1. Durch das Einsetzen in (4.31) kann

$$\begin{aligned}
&-\left(b_{2,n} + c_{2,n} + \big(b_{4,n} + c_{4,n} + \beta\big)\frac{\big(1 - \alpha_f - \alpha_m\Delta_\alpha a_{2,n}\,(1 - 1/\sigma_n)\,\big)}{\big(1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}\,(1 - 1/\sigma_n)\,\big)\big)}\right)h_n\mathbf{e}^{\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1}\\
&= \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2)\|\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}\| + \mathcal{O}(h_n^2)
\end{aligned}$$

gezeigt werden. Mit dem Satz über die implizite Funktion ist diese Gleichung lokal eindeutig nach $h_n\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}$ auflösbar, da $\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top$ regulär ist. Daher folgt $h_n\|\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}\| = \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2)$ und somit mit Lemma 8 und Gleichung (4.35) die Behauptung.

■

Eine weitere Fehlerschranke wird durch die diskrete Approximation (4.31) der versteckten Zwangsbedingung erhalten. Dabei wird der Term $\mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_{h_n}\mathbf{e}^q_n$ aus (4.31) durch $\mathbf{B}\big(q(t_n)\big)\mathbf{r}_n$ substituiert mit dem Vektor $\mathbf{r}_n \in \mathbb{R}^N$ definiert durch

$$\begin{aligned}
h_n\mathbf{r}_n \coloneqq\ &\boldsymbol{\Delta}_{h_n}\mathbf{e}^q_n + \big(b_{1,n} + c_{1,n}\big)h_n\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_n + \big(b_{2,n} + c_{2,n}\big)h_n\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1}\\
&- \big(b_{3,n} + c_{3,n} + 0.5 - \beta\big)h_n\mathbf{e}^{\mathbf{a}}_n - \big(b_{4,n} + c_{4,n} + \beta\big)h_n\mathbf{e}^{\mathbf{a}}_{n+1}.
\end{aligned} \tag{4.36}$$

Im konstanten Fall wird dies in [2, Lemma 4.8] bewiesen.

Lemma 16
Unter Verwendung der Notation

$$\mathbf{r}_n^{\mathbf{B}} \coloneqq \mathbf{B}\big(q(t_{n+1})\big)\mathbf{r}_n + \frac{1}{h_n}\mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) \tag{4.37}$$

ergeben sich für das Generalized-α-Verfahren (4.11) die Abschätzungen

$$\begin{aligned}\mathbf{r}_n^{\mathbf{B}} &= \big(b_{1,n} + c_{1,n}\big)\mathbf{e}_n^{\mathbf{S\lambda}} + \big(b_{2,n} + c_{2,n}\big)\mathbf{e}_{n+1}^{\mathbf{S\lambda}} - \big(b_{3,n} + c_{3,n} + 0.5 - \beta\big)\mathbf{e}_n^{\mathbf{Ba}} \\ &\quad - \big(b_{4,n} + c_{4,n} + \beta\big)\mathbf{e}_{n+1}^{\mathbf{Ba}} + \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2), \qquad (4.38a)\\ \sigma_{n+1}\mathbf{r}_{n+1}^{\mathbf{B}} - \mathbf{r}_n^{\mathbf{B}} &= \big(b_{3,n} + d_{3,n} + 1 - \gamma\big)\mathbf{e}_n^{\mathbf{Ba}} + \big(b_{4,n} + d_{4,n} + \gamma\big)\mathbf{e}_{n+1}^{\mathbf{Ba}} - \big(b_{1,n} + d_{1,n}\big)\mathbf{e}_n^{\mathbf{S\lambda}} \\ &\quad - \big(b_{2,n} + d_{2,n}\big)\mathbf{e}_{n+1}^{\mathbf{S\lambda}} + \mathcal{O}(1)\big(\epsilon_n + \epsilon_{n+1}\big) + \mathcal{O}(h_n^2). \qquad (4.38b)\end{aligned}$$

Beweis:

a) Mit Hilfe der Lemmata 13 und 15 lässt sich $\mathbf{r}_n^{\mathbf{B}}$ aus den Definitionen (4.36) und (4.37) als (4.38a) schreiben.

b) Durch Skalierung mit h_n folgt aus (4.27c) und Lemma 15

$$\begin{aligned}\frac{\mathbf{e}_{n+1}^{\mathbf{Bv}} - \mathbf{e}_n^{\mathbf{Bv}}}{h_n} &= \big(b_{3,n} + d_{3,n} + 1 - \gamma\big)\mathbf{e}_n^{\mathbf{Ba}} + \big(b_{4,n} + d_{4,n} + \gamma\big)\mathbf{e}_{n+1}^{\mathbf{Ba}} - \big(b_{2,n} + d_{2,n}\big)\cdot \\ &\quad \cdot \mathbf{e}_{n+1}^{\mathbf{S\lambda}} - \big(b_{1,n} + d_{1,n}\big)\mathbf{e}_n^{\mathbf{S\lambda}} + \frac{1}{h_n}\mathbf{l}_n^{\mathbf{Bv}} + \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2). \qquad (4.39)\end{aligned}$$

Um die linke Seite dieser Gleichung abzuschätzen, wird Gleichung (4.25) von links mit $\mathbf{B}\big(q(t_n)\big)$ multipliziert und es folgt erneut durch Lemma 15

$$\begin{aligned}\mathbf{e}_n^{\mathbf{Bv}} &= -\mathbf{B}\big(q(t_n)\big)\frac{\mathbf{l}_n^q}{h_n} + h_n\mathbf{r}_n^{\mathbf{B}} - \mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) - \mathbf{B}\big(q(t_n)\big)\widehat{\mathbf{e}}_n^q\mathbf{v}(t_n) \\ &\quad + \mathcal{O}(h_n)\epsilon_n + \mathcal{O}(h_n^3). \qquad (4.40)\end{aligned}$$

Wird von (4.40) die Differenz für zwei aufeinanderfolgende Schritte skaliert durch $1/h_n$ betrachtet, so folgt

$$\begin{aligned}\frac{\mathbf{e}_{n+1}^{\mathbf{Bv}} - \mathbf{e}_n^{\mathbf{Bv}}}{h_n} &= -\mathbf{B}\big(q(t_{n+1})\big)\frac{\mathbf{l}_{n+1}^q}{h_n h_{n+1}} + \mathbf{B}\big(q(t_n)\big)\frac{\mathbf{l}_n^q}{h_n^2} + \sigma_{n+1}\mathbf{r}_{n+1}^{\mathbf{B}} - \mathbf{r}_n^{\mathbf{B}} \\ &\quad + \mathcal{O}(1)\big(\epsilon_n + \epsilon_{n+1}\big) + \mathcal{O}(h_n^2), \qquad (4.41)\end{aligned}$$

da die Zusammenhänge

$$\begin{aligned}&\frac{\mathbf{Z}\big(q(t_{n+1})\big)\big(\mathbf{e}_{n+1}^q, \mathbf{v}(t_{n+1})\big) - \mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big)}{h_n} \\ &= \mathbf{Z}\big(q(t_n)\big)\left(\frac{\mathbf{e}_{n+1}^q - \mathbf{e}_n^q}{h_n}, \mathbf{v}(t_n)\right) + \mathcal{O}(1)\epsilon_{n+1} \\ &= \mathcal{O}(1)\big(\epsilon_{n+1} + \|\mathbf{\Delta}_{h_n}\mathbf{e}_n^q\|\big) = \mathcal{O}(1)\big(\epsilon_n + \epsilon_{n+1}\big) + \mathcal{O}(h_n^2)\end{aligned}$$

und

$$\begin{aligned}&\frac{\mathbf{B}\big(q(t_{n+1})\big)\widehat{\mathbf{e}}_{n+1}^q\mathbf{v}(t_{n+1}) - \mathbf{B}\big(q(t_n)\big)\widehat{\mathbf{e}}_n^q\mathbf{v}(t_n)}{h_n} \\ &= \mathbf{B}\big(q(t_n)\big)\left(\widehat{\frac{\mathbf{e}_{n+1}^q - \mathbf{e}_n^q}{h_n}}\right)\mathbf{v}(t_n) + \mathcal{O}(1)\epsilon_{n+1} \\ &= \mathcal{O}(1)\big(\epsilon_{n+1} + \|\mathbf{\Delta}_{h_n}\mathbf{e}_n^q\|\big) = \mathcal{O}(1)\big(\epsilon_n + \epsilon_{n+1}\big) + \mathcal{O}(h_n^2)\end{aligned}$$

mit $q(t_{n+1}) = q(t_n) + \mathcal{O}(h_n)$, $\mathbf{v}(t_{n+1}) = \mathbf{v}(t_n) + \mathcal{O}(h_n)$ und

$$\|\mathbf{\Delta}_{h_n}\mathbf{e}_n^q\| = \mathcal{O}(1)\epsilon_n + \mathcal{O}(h_n^2)$$

nach Gleichung (4.32) und Lemma 15 gelten. Gleichsetzen von (4.39) und (4.41) ergibt (4.38b) aufgrund von $\frac{1}{h_n}\left\|\mathbf{B}\big(q(t_n)\big)\left(\frac{\mathbf{l}_{n+1}^q}{h_{n+1}} - \frac{\mathbf{l}_n^q}{h_n} + \mathbf{l}_n^{\mathbf{v}}\right)\right\| = \mathcal{O}(h_n^2)$ (vgl. Satz 1).

■

Bemerkung 18 (Zusammenhang zum Verfahren für konstante Schrittweiten)
Die Stabilitätsanalyse von Chung und Hulbert [15] für $\ddot{q} + \omega^2 q = 0$ zeigt, dass die Stabilität des Generalized-α-Verfahrens mit konstanter Schrittweite h für (sehr) steife Probleme durch eine gekoppelte Fehlerrekursion von drei Komponenten auf Lage-, Geschwindigkeits- und Beschleunigungsebene charakterisiert wird, wobei die 3×3-Fehlerfortpflanzungsmatrix für $h\omega \to \infty$ einen dreifachen Eigenwert $-\rho_\infty$ mit geometrischer Vielfachheit 1 hat. Für $\omega \to \infty$, das heißt $1/\omega \to 0$, ergibt sich hieraus für Systeme mit Zwangsbedingungen eine gekoppelte Fehlerrekursion für $\mathbf{e}_n^{\boldsymbol{\lambda}}$ und diejenigen Komponenten von $\mathbf{e}_n^{\mathbf{v}}$ und $\mathbf{e}_n^{\mathbf{a}}$, die in Normalenrichtung zur Zwangsmannigfaltigkeit $\mathfrak{M} = \{q : \ \mathbf{\Phi}(q) = \mathbf{0}\}$ im Punkt $q = q(t_n)$ liegen und durch $\mathbf{e}_n^{\mathbf{Bv}}$ und $\mathbf{e}_n^{\mathbf{Ba}}$ gegeben sind, vgl. [2, Abschnitt 3.2]. Letztendlich zeigt sich, dass $\mathbf{e}_n^{\mathbf{Bv}}$ hierbei mit $1/h_n$ zu skalieren ist und weitere Terme ergänzt werden müssen (vgl. (4.36) und (4.37)). Entsprechend ist im allgemeinen Fall $\mathbf{e}_n^{\boldsymbol{\lambda}}$ durch $\mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}}$ zu ersetzen.

4.3.4 Gekoppelte Fehlerrekursion

In diesem Abschnitt sollen die berechneten Gleichungen für die globalen Fehler zu einer gekoppelten Fehlerrekursion der Form

$$\|\mathbf{E}_{n+1}^{\mathbf{y}} - \mathbf{T}_{\mathbf{y},n}\mathbf{E}_n^{\mathbf{y}}\| \le L_0 h_n\big(\|\mathbf{E}_n^{\mathbf{y}}\| + \|\mathbf{E}_{n+1}^{\mathbf{y}}\| + \|\mathbf{E}_n^{\mathbf{z}}\| + \|\mathbf{E}_{n+1}^{\mathbf{z}}\|\big) + h_n M_0, \tag{4.42a}$$

$$\|\mathbf{E}_{n+1}^{\mathbf{z}} - \mathbf{T}_{\mathbf{z},n}\mathbf{E}_n^{\mathbf{z}}\| \le L_0\big(\|\mathbf{E}_n^{\mathbf{y}}\| + \|\mathbf{E}_{n+1}^{\mathbf{y}}\| + h_n\|\mathbf{E}_n^{\mathbf{z}}\| + h_n\|\mathbf{E}_{n+1}^{\mathbf{z}}\|\big) + M_0 \tag{4.42b}$$

kombiniert werden. Die positiven Konstanten L_0 und M_0 sind unabhängig von n und der Schrittweite h_n. $(\mathbf{E}_n^{\mathbf{y}})_{n\ge 0}$ und $(\mathbf{E}_n^{\mathbf{z}})_{n\ge 0}$ sind vektorwertige Folgen. $\mathbf{T}_{\mathbf{y},n}$ und $\mathbf{T}_{\mathbf{z},n}$ sind Matrizen abhängig vom aktuellen Zeitschritt $t_n \to t_{n+1}$. Die Gleichungen des Generalized-α-Verfahrens für variable Schrittweiten (4.11) für die differentiellen Fehlerkomponenten $\mathbf{e}_n^q$, $\mathbf{e}_n^{\mathbf{v}}$ und die algebraischen Komponenten $\mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}}$, $\mathbf{e}_n^{\mathbf{Ba}}$, $\mathbf{r}_n^{\mathbf{B}}$, $\mathbf{e}_n^{\mathbf{Pa}}$ können, wie im nachfolgenden Satz gezeigt, in der Form (4.42) geschrieben werden.

Satz 2
Das Generalized-α-Verfahren (4.11) erfüllt die gekoppelte Fehlerrekursion (4.42) mit

$$\mathbf{E}_n^{\mathbf{y}} = \begin{bmatrix} \mathbf{e}_n^q \\ \mathbf{e}_n^{\mathbf{Pv}} \end{bmatrix}, \quad \mathbf{E}_n^{\mathbf{z}} := \begin{bmatrix} \mathbf{e}_n^{\mathbf{Pa}} \\ \mathbf{E}_n^{\mathbf{r}} \end{bmatrix}, \quad \mathbf{E}_n^{\mathbf{r}} := \begin{bmatrix} \mathbf{r}_n^{\mathbf{B}} \\ \mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}} \\ \mathbf{e}_n^{\mathbf{Ba}} \end{bmatrix} \tag{4.43a}$$

und den Matrizen

$$\mathbf{T}_{\mathbf{y}} = \mathbf{T}_{\mathbf{y},n} = \mathbf{I}_{2N}, \tag{4.43b}$$

$$\mathbf{T}_{\mathbf{z},n} = \text{blockdiag}\left(-\frac{\alpha_m(1 + \Delta_\alpha a_{3,n}(1 - 1/\sigma_n))}{1 - \alpha_m\big(1 - \Delta_\alpha a_{4,n}(1 - 1/\sigma_n)\big)}, \big(\mathbf{T}_{+,n}^{-1}\mathbf{T}_{0,n}\big)\right) \otimes \mathbf{I}_N, \tag{4.43c}$$

wobei

$$\mathbf{T}_{+,n} \coloneqq \begin{bmatrix} 0 & \big(b_{2,n}+c_{2,n}\big) & -\big(b_{4,n}+c_{4,n}+\beta\big) \\ \sigma_{n+1} & \big(b_{2,n}+d_{2,n}\big) & -\big(b_{4,n}+d_{4,n}+\gamma\big) \\ 0 & \big(1-\alpha_f-\alpha_m\Delta_\alpha a_{2,n}(1-1/\sigma_n)\big) & \big(1-\alpha_m\big(1-\Delta_\alpha a_{4,n}(1-1/\sigma_n)\big)\big) \end{bmatrix} \tag{4.44a}$$

und

$$\mathbf{T}_{0,n} \coloneqq \begin{bmatrix} 1 & -\big(b_{1,n}+c_{1,n}\big) & \big(b_{3,n}+c_{3,n}+\frac{1}{2}-\beta\big) \\ 1 & -\big(b_{1,n}+d_{1,n}\big) & \big(b_{3,n}+d_{3,n}+1-\gamma\big) \\ 0 & \big(-\alpha_f+\alpha_m\Delta_\alpha a_{1,n}(1-1/\sigma_n)\big) & -\alpha_m(1+\Delta_\alpha a_{3,n}(1-1/\sigma_n)) \end{bmatrix} \tag{4.44b}$$

gelten.

Beweis:

a) Die Kombination von den Gleichungen (4.24) und (4.27b) mit Lemma 15 ergibt

$$\begin{bmatrix} \mathbf{e}_{n+1}^{q} \\ \mathbf{e}_{n+1}^{\mathbf{Pv}} \end{bmatrix} - \begin{bmatrix} \mathbf{e}_{n}^{q} \\ \mathbf{e}_{n}^{\mathbf{Pv}} \end{bmatrix} = \mathcal{O}(h_n)\big(\epsilon_n+\epsilon_{n+1}\big) + \mathcal{O}(h_n)\big(\|\mathbf{e}_n^{\mathbf{Pa}}\|+\|\mathbf{e}_{n+1}^{\mathbf{Pa}}\|\big) + \mathcal{O}(h_n^3).$$

Damit ist (4.42a) für geeignete positive Konstanten L_0 und M_0 und mit den Bezeichnungen (4.43) erfüllt.

b) Die Kombination von den Gleichungen (4.33a), (4.33b) und (4.38b) mit Lemma 15 ergibt

$$\begin{aligned} &\begin{bmatrix} \big(1-\alpha_m\big(1-\Delta_\alpha a_{4,n}(1-1/\sigma_n)\big)\big)\mathbf{I}_N & \mathbf{0}_{N\times 3N} \\ \mathbf{0}_{3N\times N} & \mathbf{T}_{+,n}\otimes\mathbf{I}_N \end{bmatrix} \begin{bmatrix} \mathbf{e}_{n+1}^{\mathbf{Pa}} \\ \mathbf{E}_{n+1}^{\mathbf{r}} \end{bmatrix} \\ &= \begin{bmatrix} -\alpha_m(1+\Delta_\alpha a_{3,n}(1-1/\sigma_n))\mathbf{I}_N & \mathbf{0}_{N\times 3N} \\ \mathbf{0}_{3N\times N} & \mathbf{T}_{0,n}\otimes\mathbf{I}_N \end{bmatrix} \begin{bmatrix} \mathbf{e}_{n}^{\mathbf{Pa}} \\ \mathbf{E}_{n}^{\mathbf{r}} \end{bmatrix} \\ &\quad + \mathcal{O}(1)(\epsilon_n+\epsilon_{n+1}) + \mathcal{O}(h_n^2) \end{aligned}$$

und somit folgt (4.42b) mit den Bezeichnungen (4.43).

■

4.3.5 Stabilität

Erste Stabilitätsuntersuchungen mit der parameterunabhängigen Testgleichung für das Generalized-α-Verfahren wurden schon von Chung und Hulbert [15] veröffentlicht. Untersuchungen zur Null-Stabilität für differential-algebraische Gleichungen vom Index 3 und konstante Schrittweiten erfolgten in [2]. In [53] wurde zudem die Stabilität des Generalized-α-Verfahrens für variable Schrittweiten untersucht.
Dazu muss die Matrix $\mathbf{T}_{\mathbf{z},n}$ aus (4.43c) betrachtet werden. Das Verfahren (4.11) ist

Tabelle 4.1: Obere und untere Schranken der Schrittweitenverhältnisse σ_n

	Näherung 1		Näherung 2		Näherung 3		Näherung 4	
ρ_∞	$\sigma_{\min}$	$\sigma_{\max}$	$\sigma_{\min}$	$\sigma_{\max}$	$\sigma_{\min}$	$\sigma_{\max}$	$\sigma_{\min}$	$\sigma_{\max}$
0.1	0.7872	1.2688	0.6357	1.6048	0.6287	1.6473	0.7264	1.2038
0.2	0.8491	1.1587	0.6326	1.5897	0.6134	1.1064	0.7974	1.1302
0.3	0.9029	1.0842	0.6632	1.5525	0.6368	1.0254	0.8625	1.0934
0.4	0.9419	1.0454	0.7495	1.4931	0.7112	1.0619	0.91	1.0651
0.5	0.9682	1.0234	0.8375	1.3491	0.7930	1.0818	0.9436	1.0434
0.6	0.9845	1.011	0.9131	1.088	0.874	1.0812	0.9669	1.0269
0.7	0.9951	1.0043	0.9633	1.0433	0.9519	1.0609	0.9826	1.0149
0.8	0.999	1.0012	0.9885	1.0154	0.9839	1.0279	0.9935	1.0062
0.9	1	1.0001	0.998	1.0023	0.9965	1.0057	0.9988	1.0012

instabil, wenn das Produkt der Matrizen $\mathbf{T}_{\mathbf{z},n}$ nicht beschränkt ist. Daher muss eine Norm $\|.\|_*$ gefunden werden, so dass

$$\|\mathbf{T}_{\mathbf{z},n+l} \cdot \ldots \cdot \mathbf{T}_{\mathbf{z},n+1}\mathbf{T}_{\mathbf{z},n}\|_* < C_S$$

für alle $n, l \in \mathbb{N}$ und einer Konstanten $C_S > 0$ gilt, vgl. [28]. Diese Bedingung ist immer erfüllt, wenn

$$\|\mathbf{T}_{\mathbf{z},i}\|_* < 1$$

für alle $i \in \mathbb{N}$ eingehalten wird. Mit Hilfe des Generalized-α-Verfahrens (4.11) angewendet auf konstante Schrittweiten und der entsprechenden Matrix $\mathbf{T}_{\mathbf{z},n}$ aus (4.43c) wird eine Norm $\|.\|_*$ definiert, mit der obere und untere Schranken $\sigma_{\max} = \sigma_{\max}(\rho_\infty)$ bzw. $\sigma_{\min} = \sigma_{\min}(\rho_\infty)$ für Schrittweitenverhältnisse $\sigma_{\min} \leq \sigma_n \leq \sigma_{\max}$, $(n \geq 0)$, konstruiert werden können.

Satz 3
Für die Näherungen 1-4 aus (4.9), spezielle ρ_∞ und die resultierende Matrix $\mathbf{T}_{\mathbf{z},n}$ aus (4.43c) gibt es eine Norm $\|.\|_*$, so dass $\|\mathbf{T}_{\mathbf{z},n}\|_* < 1$ ist, wenn die Schrittweitenverhältnisse innerhalb der Grenzen aus Tabelle 4.1 liegen.

Beweis:
Für $\sigma_n = 1$, $(n \geq 0)$, ist die Matrix

$$\mathbf{T}_{\mathbf{z}} \coloneqq \mathbf{T}_{\mathbf{z},n} = \begin{bmatrix} -\frac{\alpha_m}{1-\alpha_m} & 0 & 0 & 0 \\ 0 & 1-\frac{\gamma}{\beta} & 0 & 1-\frac{\gamma}{2\beta} \\ 0 & \frac{1-\alpha_m}{\beta(1-\alpha_f)} & -\frac{\alpha_f}{1-\alpha_f} & \frac{1-\alpha_m-2\beta}{2\beta(1-\alpha_f)} \\ 0 & -\frac{1}{\beta} & 0 & 1-\frac{1}{2\beta} \end{bmatrix} \otimes \mathbf{I}_N$$

konstant für alle Näherungen aus (4.9). Zu dieser Matrix $\mathbf{T}_{\mathbf{z}}$ mit Parametern (2.20) gibt es eine Jordansche Normalform, welche sich aufstellen lässt zu und einer Block-

diagonalmatrix

$$\mathbf{D} = \mathbf{V}^{-1}\mathbf{T}_\mathbf{z}\mathbf{V} = \begin{bmatrix} 2+\frac{3}{\rho_\infty-2} & 0 & 0 & 0 \\ 0 & -\rho_\infty & 1 & 0 \\ 0 & 0 & -\rho_\infty & 1 \\ 0 & 0 & 0 & -\rho_\infty \end{bmatrix} \otimes \mathbf{I}_N,$$

in der auf der ersten Nebendiagonalen Einsen stehen, und mit der Matrix

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2(1+\rho_\infty)^2} & \frac{\rho_\infty}{2(\rho_\infty-1)(1+\rho_\infty)^3} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\rho_\infty^2} & \frac{2-\rho_\infty}{(\rho_\infty^2-1)^2} \end{bmatrix} \otimes \mathbf{I}_N.$$

Um die Nichtdiagonalelemente der Größe $\mathcal{O}(1)$ zu eliminieren, wird eine Matrix

$$\mathbf{E} = \mathrm{diag}(1, \epsilon, \epsilon^2, ...)$$

mit $0 < \epsilon := \frac{1-\rho_\infty}{2} < 1$ verwendet und eine Norm $\|.\|_*$ wird durch

$$\|\mathbf{T}_\mathbf{z}\|_* = \left\|\mathbf{E}^{-1}\mathbf{V}^{-1}\mathbf{T}_\mathbf{z}\mathbf{V}\mathbf{E}\right\|_\infty \tag{4.45}$$

definiert. Für konstante Schrittweiten ist der Wert dieser Norm gegeben durch

$$\|\mathbf{T}_\mathbf{z}\|_* = \max\{|2 + 3/(\rho_\infty - 2)|, 0.5(1 + \rho_\infty)\}.$$

Somit gilt stets $\|\mathbf{T}_\mathbf{z}\|_* < 1$ für $\rho_\infty \in (0, 1)$.
Die Norm (4.45) wird nun für die Näherungen 1-4 aus (4.9) auf die Matrix für variable Schrittweiten $\mathbf{T}_{\mathbf{z},n}$ angewendet und numerisch überprüft, für welche σ_n die Bedingung $\|\mathbf{T}_{\mathbf{z},n}\|_* < 1$ erfüllt ist. Entsprechende obere und untere Schranken $\sigma_{\min} \leq \sigma_n \leq \sigma_{\max}$ für die Schrittweitenverhältnisse σ_n können für spezielle ρ_∞ aus Tabelle 4.1 abgelesen werden. Die Rechnungen hierzu erfolgten mithilfe von Mathematica. Damit ist die Behauptung des Satzes bewiesen.

■

Bemerkung 19 (Interpretation der Ergebnisse von Satz 3)
In Tabelle 4.1 sind nur Schranken für Schrittweitenverhältnisse für spezielle Spektralradien angegeben. Theoretisch können auch für andere $\rho_\infty \in [0, 1)$ solche Schranken berechnet werden. Aufgrund der komplexen Rechnungen wird jedoch in dieser Arbeit darauf verzichtet.
Weiterhin kann aus Tabelle 4.1 abgelesen werden, dass Näherung 2 und 3 die größten Schrittweitenänderungen zulassen und Näherung 1 die wenigsten. Praktische Untersuchungen zu den verschiedenen Näherungen erfolgen in Abschnitt 6.3.1.
Die Schranken aus Satz 3 erweisen sich in praktischen Rechnungen als sehr pessimistisch. Unter sinnvollen Voraussetzungen, wie nicht zu schnelle Änderung der Schrittweite oder einige Schritte mit konstanter Schrittweite, können auch andere Schrittweitenverhältnisse verwendet werden. Zudem sind die gegebenen Schranken lediglich hinreichend, was ebenso ein Indiz dafür ist, dass größere Schrittweitenverhältnisse möglich sind. Bessere Ergebnisse können zum Beispiel erzielt werden, wenn anstatt (3.44) die stabilisierte Index-2-Formulierung verwendet wird [53].

4.4 Von gekoppelter Fehlerrekursion zur Konvergenz

Die gekoppelte Fehlerrekursion (4.42) soll nun so umgeschrieben werden, dass die Konvergenz bewiesen werden kann. Dazu wird ein Lemma aus [4, Lemma 4.12] verwendet.

Lemma 17 ([4, Lemma 4.12])
Es seien $(v_n)_{n\geq 0}$ und $(w_n)_{n\geq 0}$ zwei Folgen mit nichtnegativen Zahlen, die

$$v_{n+1} \leq (1 + Lh)v_n + Lh\kappa^n e_0 + hM, \tag{4.46a}$$
$$w_{n+1} \leq (\kappa + Lh)w_n + Lh\kappa^n e_0 + M \tag{4.46b}$$

erfüllen. Hierbei sind die Konstanten $L > 0$, $M \geq 0$, $e_0 \geq 0$, $\kappa \in [0, 1)$ alle unabhängig von $h > 0$ und $n \geq 0$. Dann gelten für alle $n \geq 0$ die Abschätzungen

$$v_n \leq e^{Lnh}\left(v_0 + \frac{hLe_0}{1-\kappa}\right) + \frac{e^{Lnh}-1}{L}M, \tag{4.47a}$$
$$w_n \leq (\kappa + Lh)^n w_0 + \frac{hLe_0}{1-\kappa} + \frac{M}{1-(\kappa+Lh)} \tag{4.47b}$$

für alle $h \in (0, \overline{h}]$ mit einer hinreichend kleinen Konstanten $\overline{h}$, so dass $\kappa + L\overline{h} < 1$ gilt.

Im Unterschied zu [4, Lemma 4.12.] wurde hierbei nicht die Notation

$$t_n := t_0 + nh \tag{4.48}$$

verwendet. Da im Fall der variablen Schrittweiten die Fehlerrekursion (4.42) auf Gleichungen der Form (4.46) mit $h = h_{\max}$ führt, wäre (4.48) offensichtlich nicht erfüllt. Unter Verwendung von Lemma 17 kann nun, ausgehend von der gekoppelten Fehlerrekursion (4.42), eine Fehlerabschätzung für DAEs mit variabler Schrittweite unter Voraussetzung 1 erfolgen. Der Beweis orientiert sich erneut an der Variante für konstante Schrittweiten aus [4, Theorem 4.16].

Satz 4
Es seien $(\mathbf{E}_n^{\mathbf{y}})_{n\geq 0}$ und $(\mathbf{E}_n^{\mathbf{z}})_{n\geq 0}$ Folgen von Vektoren, die die gekoppelte Fehlerrekursion (4.42) erfüllen mit Matrizen $\mathbf{T}_{\mathbf{y},n}$, $\mathbf{T}_{\mathbf{z},n}$ und positiven Konstanten L_0, M_0, die unabhängig von $h_n > 0$ und $n \geq 0$ sind. Wenn es Normen $\|.\|_{\mathbf{y},\rho}$, $\|.\|_{\mathbf{z},\rho}$ gibt, so dass $\|\mathbf{T}_{\mathbf{y},n}\|_{\mathbf{y},\rho} \leq 1$ und $\|\mathbf{T}_{\mathbf{z},n}\|_{\mathbf{z},\rho} < 1$ gilt, dann impliziert (4.42) für Schrittweitenfolgen, die Voraussetzung 1 erfüllen mit hinreichend kleinem $\overline{h} > 0$, die Fehlerschranken

$$\|\mathbf{E}_n^{\mathbf{y}}\| \leq e^{\overline{L}_0 C_h (t_n - t_0)}\left(\|\mathbf{E}_0^{\mathbf{y}}\| + \overline{C}_0 h_{\max}\|\mathbf{E}_0^{\mathbf{z}}\| + \frac{e^{\overline{L}_0 C_h (t_n - t_0)} - 1}{\overline{L}}\overline{M}_0\right), \tag{4.49a}$$

$$\|\mathbf{E}_n^{\mathbf{z}} - \mathbf{T}_{\mathbf{z},n}^n \mathbf{E}_0^{\mathbf{z}}\| \leq \overline{C}_0 e^{\overline{L}_0 C_h (t_n - t_0)}\big(\|\mathbf{E}_0^{\mathbf{y}}\| + h_{\max}\|\mathbf{E}_0^{\mathbf{z}}\| + \overline{M}_0\big). \tag{4.49b}$$

Die Konstanten $\overline{C}_0$, $\overline{L}_0$, und $\overline{M}_0$ sollen positiv sein und hängen von den Konstanten L_0, M_0 in (4.42) und von den Normen ab. Die Konstante C_h beschreibt das Verhältnis aus maximaler und minimaler Schrittweite, vgl. (4.1).

Beweis:
Dieser Satz ist eine Folgerung aus [4, Theorem 4.16] und Voraussetzung 1. Dafür werden die Variablen h_n, $\kappa_{\mathbf{y},n}$ und $\kappa_{\mathbf{z},n}$ durch $h_{\max}$, $\kappa_{\mathbf{y},\max}$ und $\kappa_{\mathbf{z},\max}$ mit

$$\kappa_{\mathbf{y},\max} = \max_{n=0,\dots,N_{\text{end}}} \kappa_{\mathbf{y},n} \leq 1 \quad \text{und} \quad \kappa_{\mathbf{z},\max} = \max_{n=0,\dots,N_{\text{end}}} \kappa_{\mathbf{z},n} < 1$$

ersetzt und zum Abschluss des Beweises wird nicht $nh = (t_n - t_0)$ sondern

$$nh_{\max} \leq nC_h h_{\min} \leq C_h \sum_{i=0}^{n} h_i = C_h(t_n - t_0)$$

verwendet. Der ausführliche Beweis wird in Anhang D geführt.

■

Schließlich folgt der Konvergenzsatz für das Generalized-α-Verfahren (4.11).

Satz 5
Es sei das Generalized-α-Verfahren für variable Schrittweiten (4.11) mit einer der Näherungen aus (4.9) gegeben. Wenn die Startwerte q_0, $\mathbf{v}_0$, $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$ und $\boldsymbol{\lambda}_0$ die Bedingungen

$$\|\mathbf{M}(q_0)\dot{\mathbf{v}}_0 + \mathbf{g}(q_0, \mathbf{v}_0, t_0) + \mathbf{B}^\top(q_0)\boldsymbol{\lambda}_0\| = \mathcal{O}(h_0^2), \quad \|\boldsymbol{\Phi}(q_0)\| = \mathcal{O}(h_0^4), \tag{4.50a}$$

$$\|\mathbf{e}_0^q\| + \left\|\mathbf{e}_0^{\mathbf{Bv}} + \frac{1}{h_0}\mathbf{B}(q(t_0))\mathbf{l}_0^q\right\| = \mathcal{O}(h_0^3), \tag{4.50b}$$

$$\|\mathbf{e}_0^{\mathbf{v}}\| + \|\mathbf{e}_0^{\mathbf{Pv}}\| + \|\mathbf{e}_0^{\mathbf{Ba}}\| + h_0\|\mathbf{e}_0^{\dot{\mathbf{v}}}\| + h_0\|\mathbf{e}_0^{\mathbf{Pa}}\| = \mathcal{O}(h_0^2) \tag{4.50c}$$

erfüllen, dann sind die globalen Fehler beschränkt durch

$$\|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| \leq C_0 e^{\overline{L}C_h(t_n - t_0)} h_{\max}^2,$$
$$\|\mathbf{e}_n^{\boldsymbol{\lambda}}\| \leq C_0 e^{\overline{L}C_h(t_n - t_0)} h_{\max}^2$$

für Schrittweitenfolgen, die Voraussetzung 1 erfüllen. Dabei sind die positiven Konstanten C_0, C_h, $\overline{L}$ unabhängig von n und $h_{\max}$.

Beweis:
Mit Satz 2 erfüllt das Generalized-α-Verfahren (4.11) die gekoppelte Fehlerrekursion mit (4.43). Weiterhin kann nach Satz 3 mit Voraussetzung 1 eine Norm $\|.\|$ gefunden werden, sodass die Bedingung $\|\mathbf{T}_{\mathbf{z},n}\| < 1$ erfüllt ist. Da $\mathbf{T}_{\mathbf{y},n} = \mathbf{I}_{2N}$ ist, vgl. (4.43c), gilt $\|\mathbf{T}_{\mathbf{y},n}\| = 1$. Aus diesem Grund kann Satz 4 angewendet werden und es folgt (4.49). Mit (4.43a) und (4.50) ist $\|\mathbf{E}_0^{\mathbf{y}}\| = \mathcal{O}(h_0^2)$. Außerdem ist mit (4.40)

$$\mathbf{r}_0^{\mathbf{B}} = \frac{1}{h_0}\left(\mathbf{e}_0^{\mathbf{Bv}} + \mathbf{B}(q(t_0))\frac{\mathbf{l}_0^q}{h_0}\right) + \frac{1}{h_0}\left(\mathbf{Z}\big(q(t_0)\big)\big(\mathbf{e}_0^q, \mathbf{v}(t_0)\big) + \mathbf{B}\big(q(t_0)\big)\widehat{\mathbf{e}}_0^q\mathbf{v}(t_0)\right)$$
$$+ \mathcal{O}(1)\epsilon_0 + \mathcal{O}(h_0^2)$$

erfüllt und daher $\|\mathbf{r}_0^{\mathbf{B}}\| = \mathcal{O}(h_0^2)$ mit (4.50). Außerdem folgt mit (4.43a) und (4.50), dass $\|\mathbf{E}_0^{\mathbf{r}}\| = \mathcal{O}(h_0^2)$ und $\|\mathbf{E}_0^{\mathbf{z}}\| = \mathcal{O}(h_0)$ sind. Der Fehlerterm $\overline{M}_0$ repräsentiert eine obere Schranke für die lokalen Fehler und somit folgt mit Satz 1 die Behauptung.

■

Kapitel 5

Konvergenzanalyse der BDF-Verfahren für Konfigurationsräume mit Lie-Gruppen-Struktur für differential-algebraische Gleichungen vom Index 3

Das Ziel dieses Kapitels ist es, die Konvergenz von BDF-Zeitintegrationsverfahren für Konfigurationsräume mit Lie-Gruppen-Struktur für mechanische Mehrkörpersysteme zu beweisen, wobei die Bewegungsgleichungen differential-algebraische Gleichungen vom Index 3 darstellen. Für lineare Konfigurationsräume haben Lötstedt und Petzold [36] die Konvergenz von BDF-Verfahren für DAEs in Hessenbergform gezeigt. Weiterhin befasst sich [2] mit der Konvergenzanalyse des Generalized-α-Verfahrens für DAEs vom Index 3 für Konfigurationsräume mit Lie-Gruppen-Struktur. Im folgenden Kapitel werden beide Vorgehensweisen kombiniert, um einen Konvergenzbeweis für $2 \leq k \leq 6$ für das Munthe-Kaas-BDF-Verfahren (3.57) und das BLieDF-Verfahren (3.62) zu erhalten. Der Konvergenzbeweis für das BLieDF-Verfahren wurde bereits in [55] veröffentlicht. In Abschnitt 5.2 wird das BLieDF-Verfahren (3.62) schließlich auf variable Schrittweiten übertragen und die Konvergenz für $2 \leq k \leq 3$ bewiesen.

5.1 Konvergenz der BDF-Verfahren für konstante Schrittweiten

Wie für das Generalized-α-Verfahren beginnt der Konvergenzbeweis mit einer lokalen Fehleranalyse. Im Anschluss werden Gleichungen für die globalen Fehler aufgestellt, die zu einer gekoppelten Fehlerrekursion kombiniert werden. Diese führt schließlich für geeignete Startwerte auf die Konvergenz der BDF-Verfahren (3.57) und (3.62).

5.1.1 Lokale Abbruchfehler

Zunächst werden die lokalen Abbruchfehler der Munthe-Kaas-BDF-Verfahren (3.57) und der BLieDF-Verfahren (3.62) berechnet, indem die analytische Lösung in die Verfahrensvorschrift eingesetzt und durch die Taylor- und Magnusentwicklung abgeschätzt wird.

BLieDF-Verfahren (3.62)

Die lokalen Abbruchfehler für die Variablen q und $\mathbf{v}$ sollen wie in Definition A.3 definiert und durch Taylorentwicklung berechnet werden. Durch das Einsetzen der analytischen Lösungen $q(t_n) \approx q_n$, $q(t_{n+1}) \approx q_{n+1}$ und $\boldsymbol{\nu}_n(t_{n+1}) \approx \boldsymbol{\omega}_0^{(n)}$ in (3.62a) wird jedoch Gleichung (2.29) für $t = t_{n+1}$ und $m = n$ erhalten. Daher wird, anders als beim Generalized-α-Verfahren (3.55), nicht der lokale Abbruchfehler $\mathbf{l}_n^q$ aus (4.12a) verwendet, sondern der, der beim Einsetzen der analytischen Lösung in (3.62b) entsteht.

Definition 23 (Lokale Abbruchfehler von (3.62))
Die lokalen Abbruchfehler der BLieDF-Verfahren (3.62) sind definiert durch

$$q(t_{n+1}) = q(t_n) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\big), \tag{5.1a}$$

$$\frac{1}{h}\sum_{i=1}^{k} \gamma_i \boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) = \mathbf{v}(t_{n+1}) + \mathbf{L}_h^{(k)}(t_n) + \mathbf{l}_n^{\boldsymbol{\omega}}, \tag{5.1b}$$

$$\begin{aligned}\mathbf{M}\big(q(t_{n+1})\big)\frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}(t_{n+1-i}) = &-\mathbf{g}\big(t_{n+1}, q(t_{n+1}), \mathbf{v}(t_{n+1})\big) \\ &- \mathbf{B}^\top\big(q(t_{n+1})\big)\boldsymbol{\lambda}(t_{n+1}) + \mathbf{l}_n^{\mathbf{v}},\end{aligned} \tag{5.1c}$$

$$\mathbf{0} = \boldsymbol{\Phi}\big(q(t_{n+1})\big). \tag{5.1d}$$

Satz 6
Die lokalen Abbruchfehler (5.1) der BLieDF-Verfahren (3.62) erfüllen für $1 \le k \le 6$ die Abschätzungen

$$\|\mathbf{l}_n^{\boldsymbol{\omega}}\| = \mathcal{O}(h^k), \quad \left\|\frac{\mathbf{l}_{n+1}^{\boldsymbol{\omega}} - \mathbf{l}_n^{\boldsymbol{\omega}}}{h}\right\| = \mathcal{O}(h^k), \quad \|\mathbf{l}_n^{\mathbf{v}}\| = \mathcal{O}(h^k). \tag{5.2}$$

Beweis:

a) Die lokalen Abbruchfehler $\mathbf{l}_n^{\boldsymbol{\omega}}$ sind definiert durch (5.1b). Mit Lemma 5 und Gleichung (3.32) kann die Abschätzung $\|\mathbf{l}_n^{\boldsymbol{\omega}}\| = \mathcal{O}(h^k)$ direkt gezeigt werden.

b) Es seien $\mathbf{l}_n^{\boldsymbol{\omega}}$ die lokalen Abbruchfehler des k-Schritt BLieDF-Verfahrens (3.62) für den Zeitschritt $t_n \to t_{n+1}$. Der Beweis von Lemma 5 zeigt, dass diese Fehler durch $\mathbf{l}_n^{\boldsymbol{\omega}} = -\frac{1}{k+1}h^k\mathbf{v}^{(k)}(t_n) + h^k\mathbf{F}\big(\mathbf{v}(t_n), ..., \mathbf{v}^{(k-1)}(t_n)\big) + \mathcal{O}(h^{k+1})$ (vgl. [28, Table III.2.1]) gegeben sind mit einer Funktion $\mathbf{F}$, die aus Matrix-Kommutatoren besteht. Die Behauptung folgt wegen

$$\begin{aligned}\frac{\mathbf{l}_{n+1}^{\boldsymbol{\omega}} - \mathbf{l}_n^{\boldsymbol{\omega}}}{h} = &\frac{-\frac{1}{k+1}h^k\mathbf{v}^{(k)}(t_{n+1}) + h^k\mathbf{F}\big(\mathbf{v}(t_{n+1}), ..., \mathbf{v}^{(k-1)}(t_{n+1})\big)}{h} \\ &- \frac{-\frac{1}{k+1}h^k\mathbf{v}^{(k)}(t_n) + h^k\mathbf{F}\big(\mathbf{v}(t_n), ..., \mathbf{v}^{(k-1)}(t_n)\big)}{h} + \mathcal{O}(h^k) = \mathcal{O}(h^k).\end{aligned}$$

c) Gleichung (5.1c) ist in linearen Räumen definiert. Deshalb ist $\mathbf{l}_n^{\mathbf{v}} = \mathcal{O}(h^k)$ erfüllt, vgl. [28].

■

Munthe-Kaas-BDF-Verfahren (3.57)

Die Munthe-Kaas-BDF-Verfahren werden in [20] so definiert, dass die Approximationen BCH_k und dexp_k^{-1} bis zu dem Reihenglied ausgewertet werden, das nötig ist, um eine Konvergenzordnung $p = k$ zu erhalten. Eine Berechnung der lokalen Abbruchfehler ist daher für die Gleichungen (3.57b)-(3.57c) theoretisch nicht nötig. Um in der anschließenden Konvergenzanalyse für DAEs vom Index 3 festzustellen, an welchen Stellen die näherungsweise Auswertung von BCH und dexp^{-1} einen Einfluss hat, werden in der nachfolgenden Definition trotzdem lokale Abbruchfehler eingeführt.

Definition 24 (Lokale Abbruchfehler von (3.57))
Die lokalen Abbruchfehler der Munthe-Kaas-BDF-Verfahren (3.57) sind durch

$$q(t_{n+1}) = q(t_n) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\big), \tag{5.3a}$$

$$\begin{aligned}\boldsymbol{\nu}_n(t_{n+1-i}) = {}& \widehat{\mathrm{BCH}}_k\big(-\boldsymbol{\nu}_{n-1}(t_n), \boldsymbol{\nu}_{n-1}(t_{n-i+1})\big) \\ & + \mathbf{l}_{n,i}^{\mathrm{BCH}}, \quad (i = 1, ..., k),\end{aligned} \tag{5.3b}$$

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \boldsymbol{\nu}_n(t_{n+1-i}) = \widehat{\mathrm{dexp}_k}^{-1}\big(-\boldsymbol{\nu}_n(t_{n+1}), \mathbf{v}(t_{n+1})\big) + \mathbf{l}_n^{\boldsymbol{\omega}}, \tag{5.3c}$$

$$\begin{aligned}\mathbf{M}\big(q(t_{n+1})\big)\frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{v}(t_{n+1-i}) = {}& -\mathbf{g}\big(t_{n+1}, q(t_{n+1}), \mathbf{v}(t_{n+1})\big) \\ & - \mathbf{B}^{\top}\big(q(t_{n+1})\big)\boldsymbol{\lambda}(t_{n+1}) + \mathbf{l}_n^{\mathbf{v}},\end{aligned} \tag{5.3d}$$

$$\mathbf{0} = \boldsymbol{\Phi}\big(q(t_{n+1})\big) \tag{5.3e}$$

definiert.

Satz 7
Die lokalen Abbruchfehler (5.3) der Munthe-Kaas-BDF-Verfahren (3.57) erfüllen für $1 \le k \le 6$ die Abschätzungen (5.2) und zusätzlich

$$\big\|\mathbf{l}_{n,i}^{\mathrm{BCH}}\big\| = \mathcal{O}(h^{k+1}). \tag{5.4}$$

Beweis:
Die Abschätzungen (5.4) und $\|\mathbf{l}_n^{\boldsymbol{\omega}}\| = \mathcal{O}(h)$ folgen direkt aus dem Konstruktionsprinzip der Munthe-Kaas-BDF-Verfahren in [20]. Die Abschätzung $\left\|\frac{\mathbf{l}_{n+1}^{\boldsymbol{\omega}} - \mathbf{l}_n^{\boldsymbol{\omega}}}{h}\right\| = \mathcal{O}(h^k)$ gilt analog zum Beweisteil b) von Satz 6 und $\mathbf{l}_n^{\mathbf{v}} = \mathcal{O}(h^k)$ folgt wie in linearen Räumen, vgl. [28].

■

5.1.2 Gleichungen für die globalen Fehler

Um Gleichungen für die globalen Fehler aufzustellen, werden zunächst die globalen Fehler wie in Gleichung (A.6) definiert. Diese sind gegeben durch die Gleichungen

(4.19a)-(4.19c) und

$$\boldsymbol{\nu}_n(t_{n+1-i}) = \boldsymbol{\omega}_i^{(n)} + \mathbf{e}_{n,i}^{\boldsymbol{\omega}}, \quad (i = 0, ..., k). \tag{5.5}$$

Für die BLieDF-Verfahren (3.62) wird in (5.5) nur der Fall $i = 0$ von Interesse sein. Diese zusätzlichen globalen Fehler $\mathbf{e}_{n,i}^{\boldsymbol{\omega}}$ (im Vergleich zum Generalized-α-Verfahren) müssen in der benötigten technischen Voraussetzung (analog zu Voraussetzung 2) berücksichtigt werden.

Voraussetzung 3
Es gibt positive Konstanten $\overline{h}$ und C_T, so dass für alle $h \in (0, \overline{h}]$ und alle r mit $t_0 + rh \in [t_0, t_{\text{end}}]$ die globalen Fehler durch

$$\|\mathbf{e}_r^q\| + \|\mathbf{e}_{r,i}^{\boldsymbol{\omega}}\| \leq C_T h, \quad \|\mathbf{e}_r^{\mathbf{v}}\| + \|\mathbf{e}_r^{\boldsymbol{\lambda}}\| \leq C_T, \quad (i = 0, ..., k),$$

beschränkt bleiben.

Die Annahme ist für $2 \leq k \leq 6$ vertretbar, da am Ende des Konvergenzbeweises gezeigt wird, dass alle genannten globalen Fehler mindestens mit einer Ordnung höher konvergieren. Für $k = 1$ kann diese Voraussetzung jedoch nicht verifiziert werden, weshalb in diesem Fall formal auf die folgende Art und Weise keine Konvergenz bewiesen werden kann, vgl. Bemerkung 20.
Aufgrund der Mehrschrittstruktur der BDF-Verfahren werden, anders als beim Generalized-α-Verfahren (3.55), weitere Voraussetzungen benötigt. Zum einen müssen die Startwerte von einer entsprechend hohen Ordnung sein (vgl. Voraussetzung 4) und zum anderen sollte sich die Lösung des Anfangswertproblems stetig über die linke Intervallgrenze hinaus fortsetzen lassen (vgl. Voraussetzung 5).

Voraussetzung 4
Die Startwerte $q_0, ..., q_{k-1}$, $\boldsymbol{\omega}_0^{(0)}, ..., \boldsymbol{\omega}_0^{(k-2)}$, $\mathbf{v}_0, ..., \mathbf{v}_{k-1}$ und $\boldsymbol{\lambda}_0, ..., \boldsymbol{\lambda}_{k-1}$ zur Lösung des Anfangswertproblems (3.44) mit $q(t_0) = q_0$ und $\mathbf{v}(t_0) = \mathbf{v}_0$ sollen die Bedingungen

$$\sum_{i=0}^{k-1} \|\mathbf{e}_i^q\| + \left\|\mathbf{e}_i^{\mathbf{B}\mathbf{v}} + \mathbf{B}(q(t_{k-1}))\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\right\| = \mathcal{O}(h^{k+1}), \quad \sum_{i=0}^{k-2} \|\mathbf{e}_{i,0}^{\boldsymbol{\omega}}\| = \mathcal{O}(h^{k+1}),$$

$$\sum_{i=0}^{k-1} \|\mathbf{e}_i^{\mathbf{v}}\| + \|\mathbf{e}_i^{\boldsymbol{\lambda}}\| = \mathcal{O}(h^k), \quad \max_{0 \leq i \leq k-1} \|\boldsymbol{\Phi}(q_i)\| = \mathcal{O}(h^{k+2}),$$

erfüllen.

Voraussetzung 5
Das Anfangswertproblem (3.44) mit $q(t_0) = q_0$ und $\mathbf{v}(t_0) = \mathbf{v}_0$ ist eindeutig lösbar auf $[t_0 - \overline{t}, t_{\text{end}}]$ für ein $\overline{t} > 0$ mit einer Lösung $\big(q(t), \mathbf{v}(t)\big)$, die hinreichend oft differenzierbar ist.

Ist Voraussetzung 5 erfüllt, dann sind $q_j \coloneqq q(t_0 - jh)$ und $\mathbf{v}_j \coloneqq \mathbf{v}(t_0 - jh)$ wohldefiniert für $j = 1, ..., k$, wenn $h \in (0, \overline{h}]$ ist mit einer positiven Konstanten $\overline{h} \leq \overline{t}/k$ und $k \in \mathbb{N}$.
Um die Konvergenzanalyse zu beginnen, werden nun, wie zuvor beim Generalized-α-Verfahren (3.55), globale Fehlerrekursionen für jede der zu betrachtenden Variablen benötigt. Dies sind im Speziellen Ungleichungen, die Aussagen über $\mathbf{e}_n^q$, $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$, $\mathbf{e}_n^{\mathbf{v}}$ und $\mathbf{e}_n^{\boldsymbol{\lambda}}$ liefern. In allen folgenden Sätzen und Lemmata werden die Voraussetzungen 3-5

als gegeben angenommen. Außerdem wird zum Zusammenfassen der Terme höherer Ordnung die Notation

$$\epsilon_n := \|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| + h\|\mathbf{e}_n^{\boldsymbol{\lambda}}\|$$

eingeführt.

Globale Fehlergleichung für $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$

Die globale Fehlerrekursion von $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$ wird in den beiden BDF-Verfahren sehr unterschiedlich bestimmt. Bei den BLieDF-Verfahren (3.62) ist es eine direkte Konsequenz aus den Gleichungen (3.62b) und (5.1b) sowie der Lipschitzstetigkeit von $\mathbf{L}_{h,n}^{(k)}$. Für die Munthe-Kaas-BDF-Verfahren (3.57) ist solch eine Fehlergleichung schwieriger zu beweisen. Die Differenz von (3.57c) und (5.3c) führt lediglich auf eine gewichtete Summe der $\mathbf{e}_{n,i}^{\boldsymbol{\omega}}$, $(i = 0, ..., k)$. Diese muss schließlich durch geeignete Methoden auf eine gewichtete Summe über $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$ zurückgeführt werden.

Satz 8
Der globale Fehler der BLieDF-Verfahren (3.62) erfüllt die Fehlerrekursion

$$\frac{1}{h}\sum_{i=1}^{k}\gamma_i\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}} = \mathbf{e}_{n+1}^{\mathbf{v}} + \mathbf{l}_n^{\boldsymbol{\omega}} + \mathcal{O}(h)\left(\sum_{i=0}^{k}\epsilon_{n+1-i} + \frac{1}{h}\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right).$$

Beweis:
Wird (3.62b) von (5.1b) subtrahiert, so folgt die Behauptung mit (4.19) und (5.5), weil der Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ einer Lipschitzbedingung genügt und daher

$$\mathbf{L}_{h,n}^{(k)} - \mathbf{L}_h^{(k)}(t_n) = \mathcal{O}(h)\sum_{i=0}^{k}\|\mathbf{e}_{n+1-i}^{\mathbf{v}}\| + \mathcal{O}(1)\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|$$

gilt.

■

Satz 9
Der globale Fehler der Munthe-Kaas-BDF-Verfahren (3.57) erfüllt die Fehlerrekursion

$$\begin{aligned}\frac{1}{h}\sum_{i=1}^{k}\gamma_i\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}} = \mathbf{e}_{n+1}^{\mathbf{v}} - \widehat{\mathbf{v}}(t_n)\mathbf{e}_n^q + \widehat{\mathbf{v}}(t_n)\sum_{i=1}^{k}\gamma_i\mathbf{e}_{n+1-i}^q + \mathbf{l}_n^{\boldsymbol{\omega}} + \mathcal{O}(h)\sum_{i=0}^{k}\epsilon_{n+1-i}\\ + \mathcal{O}(1)\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|.\end{aligned}$$

Beweis:
Ausgehend von (4.19a) für $n+1-i$, $(i = 0, ..., k)$, folgt durch das Einsetzen von (3.9) und (2.29) für $t = t_{n+1-i}$ und $m = n$

$$\begin{aligned}\exp\left(\widetilde{\mathbf{e}}_{n+1-i}^q\right) &= q_{n+1-i}^{-1}\circ q(t_{n+1-i}) = \exp\left(-\widetilde{\boldsymbol{\omega}}_i^{(n)}\right)\circ q_n^{-1}\circ q(t_n)\circ\exp\left(\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})\right)\\ &= \exp\left(-\widetilde{\boldsymbol{\omega}}_i^{(n)}\right)\circ\exp\left(\widetilde{\mathbf{e}}_n^q\right)\circ\exp\left(\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})\right)\\ &= \exp\left(-\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i}) + \widetilde{\mathbf{e}}_{n,i}^{\boldsymbol{\omega}}\right)\circ\exp\left(\widetilde{\mathbf{e}}_n^q\right)\circ\exp\left(\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})\right),\end{aligned}$$

vgl. [2]. Aufgrund von Voraussetzung 3 und da für $i = 0, ..., k$ gilt

$$\|\boldsymbol{\nu}_n(t_{n+1-i})\| + \|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| + \|\mathbf{e}^q_n\| = \mathcal{O}(h)$$

mit (2.33), können die Kompositionen der Exponentialabbildungen durch zweimaliges Anwenden von (2.28) untersucht werden. Es folgt

$$\begin{aligned}\exp\left(\widetilde{\mathbf{e}}^q_{n+1-i}\right) &= \exp\big(-\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i}) + \widetilde{\mathbf{e}}^{\boldsymbol{\omega}}_{n,i} + \widetilde{\mathbf{e}}^q_n - \frac{1}{2}[\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i}), \widetilde{\mathbf{e}}^q_n] + \mathcal{O}(h)\|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| \\ &\qquad + \mathcal{O}(h^2)\|\mathbf{e}^q_n\|\big) \circ \exp\left(\widetilde{\boldsymbol{\nu}}_n(t_{n+1-i})\right) \\ &= \exp\left(\widetilde{\mathbf{e}}^q_n + \widetilde{\mathbf{e}}^{\boldsymbol{\omega}}_{n,i} - h(1-i)[\widetilde{\mathbf{v}}(t_n), \widetilde{\mathbf{e}}^q_n] + \mathcal{O}(h)\|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| + \mathcal{O}(h^2)\|\mathbf{e}^q_n\|\right),\end{aligned}$$

wobei die Terme höherer Ordnung durch Voraussetzung 3 abgeschätzt wurden und unter Verwendung von $\boldsymbol{\nu}_n(t_{n+1-i}) = h(1-i)\mathbf{v}(t_n) + \mathcal{O}(h^2)$ (siehe (2.33)) sowie $[\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2] = -[\widetilde{\mathbf{w}}_2, \widetilde{\mathbf{w}}_1]$ für $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^N$, vgl. [2]. Die Betrachtung des Argumentes der Exponentialabbildung liefert die Gleichung

$$\mathbf{e}^q_{n+1-i} = \mathbf{e}^q_n + \mathbf{e}^{\boldsymbol{\omega}}_{n,i} - h(1-i)\widehat{\mathbf{v}}(t_n)\mathbf{e}^q_n + \mathcal{O}(h)\|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| + \mathcal{O}(h^2)\epsilon_n, \tag{5.6}$$

mit der Abbildung (2.35). Mit Gleichung (5.6) gilt daher für $i = 2, ..., k$

$$\begin{aligned}\mathbf{e}^{\boldsymbol{\omega}}_{n,i} = (\mathbf{e}^q_{n+1-i} - \mathbf{e}^q_{n+2-i}) + (\mathbf{e}^q_{n+2-i} - \mathbf{e}^q_{n+3-i}) + ... + (\mathbf{e}^q_{n-1} - \mathbf{e}^q_n) + h(1-i)\widehat{\mathbf{v}}(t_n)\mathbf{e}^q_n \\ + \mathcal{O}(h)\|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| + \mathcal{O}(h^2)\epsilon_n\end{aligned}$$

und $\mathbf{e}^{\boldsymbol{\omega}}_{n,1} = \mathbf{0}$, da $\boldsymbol{\nu}_n(t_n) = \boldsymbol{\omega}^{(n)}_1 = \mathbf{0}$. Die Verwendung von Gleichung (5.6) im Zeitschritt $t_{n+1-j} \to t_{n+2-j}$ für $i = 0$ und $j = 2, ..., i$ führt mit $\mathbf{v}(t_{n+1-j}) = \mathbf{v}(t_n) + \mathcal{O}(h)$ auf

$$\begin{aligned}\mathbf{e}^{\boldsymbol{\omega}}_{n,i} = &-\sum_{j=2}^{i} \mathbf{e}^{\boldsymbol{\omega}}_{n+1-j,0} + h\widehat{\mathbf{v}}(t_n)\sum_{j=2}^{i}\mathbf{e}^q_{n+1-j} + \mathcal{O}(h^2)\sum_{j=2}^{i}\left(\epsilon_{n+1-j} + \frac{1}{h}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-j,0}\|\right) \\ &+ h(1-i)\widehat{\mathbf{v}}(t_n)\mathbf{e}^q_n + \mathcal{O}(h)\|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| + \mathcal{O}(h^2)\epsilon_n\end{aligned}$$

und im Speziellen auf $\|\mathbf{e}^{\boldsymbol{\omega}}_{n,i}\| = \mathcal{O}(h)\left(\sum_{j=1}^{k}\epsilon_{n+1-j} + \frac{1}{h}\sum_{j=2}^{i}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-j,0}\|\right)$. Für die gewichtete Summe über $\mathbf{e}^{\boldsymbol{\omega}}_{n,i}$ gilt daher

$$\begin{aligned}\frac{1}{h}\sum_{i=0}^{k}\alpha_i\mathbf{e}^{\boldsymbol{\omega}}_{n,i} &= \frac{\alpha_0}{h}\mathbf{e}^{\boldsymbol{\omega}}_{n,0} + \frac{1}{h}\sum_{i=2}^{k}\alpha_i\mathbf{e}^{\boldsymbol{\omega}}_{n,i} \\ &= \frac{1}{h}\sum_{i=1}^{k}\gamma_i\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0} + \widehat{\mathbf{v}}(t_n)\mathbf{e}^q_n - \widehat{\mathbf{v}}(t_n)\sum_{i=1}^{k}\gamma_i\mathbf{e}^q_{n+1-i} \\ &\quad + \mathcal{O}(h)\left(\sum_{i=1}^{k}\epsilon_{n+1-i} + \frac{1}{h}\sum_{i=1}^{k}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\|\right)\end{aligned} \tag{5.7}$$

aufgrund von

$$-\sum_{i=2}^{k}\alpha_i\sum_{j=2}^{i}\mathbf{e}^{\boldsymbol{\omega}}_{n+1-j,0} = \sum_{i=2}^{k}\left(-\sum_{j=i}^{k}\alpha_j\right)\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}$$

und

$$-\sum_{j=i}^{k}\alpha_j = \sum_{j=0}^{i-1}\alpha_j = \gamma_i.$$

Die Differenz von (3.57c) und (5.3c) führt auf

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i\mathbf{e}^{\boldsymbol{\omega}}_{n,i} = \mathbf{e}^{\mathbf{v}}_{n+1} + \mathbf{l}^{\boldsymbol{\omega}}_{n} + \mathcal{O}(h)\left(\epsilon_{n+1} + \frac{1}{h}\|\mathbf{e}^{\boldsymbol{\omega}}_{n,0}\|\right) \tag{5.8}$$

mit (4.19) und (5.5) und da mit Voraussetzung 3 und der Lipschitzstetigkeit von dexp_k^{-1}

$$\begin{aligned}&\widehat{\mathrm{dexp}_k}^{-1}\big(-\boldsymbol{\nu}_n(t_{n+1}), \mathbf{v}(t_{n+1})\big) - \widehat{\mathrm{dexp}_k}^{-1}\big(-\boldsymbol{\omega}_0^{(n)}, \mathbf{v}_{n+1}\big)\\ &= \mathbf{e}^{\mathbf{v}}_{n+1} + \mathcal{O}(h)\left(\|\mathbf{e}^{\mathbf{v}}_{n+1}\| + \frac{1}{h}\|\mathbf{e}^{\boldsymbol{\omega}}_{n,0}\|\right)\end{aligned}$$

gilt. Gleichsetzen von (5.7) und (5.8) liefert die Behauptung.

■

Globale Fehlergleichung für $\mathbf{e}^q_n$

In diesem Abschnitt soll die Fehlerrekursion von $\mathbf{e}^q_n$ für die BDF-Integratoren (3.57) und (3.62) bewiesen werden. Dabei wird wie im Beweis von [2, Lemma 2] begonnen, jedoch muss die andere Definition des lokalen Abbruchfehlers $\mathbf{l}^{\boldsymbol{\omega}}_n$ im Unterschied zu $\mathbf{l}^q_n$ sowie die Mehrschrittstruktur der BDF-Verfahren beachtet werden.

Satz 10
Die BDF-Integratoren (3.57) und (3.62) erfüllen für $n \geq k-1$ und $2 \leq k \leq 6$ die Fehlerabschätzung

$$\begin{aligned}\frac{1}{h}\sum_{i=0}^{k}\alpha_i\mathbf{e}^q_{n+1-i} = \mathbf{e}^{\mathbf{v}}_{n+1} + \mathbf{b}\big(t_n, \mathbf{e}^q_n, ..., \mathbf{e}^q_{n+1-k}\big) + \mathbf{l}^{\boldsymbol{\omega}}_n\\ + \mathcal{O}(h)\left(\sum_{i=0}^{k}\epsilon_{n+1-i} + \frac{1}{h}\sum_{i=1}^{k}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\|\right),\end{aligned} \tag{5.9}$$

wobei $\mathbf{b}\big(t_n, \mathbf{e}^q_n, ..., \mathbf{e}^q_{n+1-k}\big)$ eine vektorwertige Funktion darstellt, die linear in $\mathbf{e}^q_{n+1-i}$, $(i = 1, ..., k)$, ist. Diese ist gegeben

a) für die Munthe-Kaas-BDF-Verfahren (3.57) durch

$$\mathbf{b}\big(t_n, \mathbf{e}^q_n, ..., \mathbf{e}^q_{n+1-k}\big) = -\widehat{\mathbf{v}}(t_n)\mathbf{e}^q_n.$$

b) für die BLieDF-Verfahren (3.62) durch

$$\mathbf{b}\big(t_n, \mathbf{e}^q_n, ..., \mathbf{e}^q_{n+1-k}\big) = -\widehat{\mathbf{v}}(t_n)\sum_{i=1}^{k}\gamma_i\mathbf{e}^q_{n+1-i}.$$

Beweis:
Begonnen wird wie in [2, Lemma 2] bzw. wie im Beweis von Satz 9. Im Vergleich zu diesem Satz wird jedoch nur $i = 0$ betrachtet. In diesem Fall ist (3.9) durch (3.62a) bzw. (3.57a) gegeben und es folgt

$$\mathbf{e}^q_{n+1} = \mathbf{e}^q_n + \mathbf{e}^{\boldsymbol{\omega}}_{n,0} - h\widehat{\mathbf{v}}(t_n)\mathbf{e}^q_n + \mathcal{O}(h)\|\mathbf{e}^{\boldsymbol{\omega}}_{n,0}\| + \mathcal{O}(h^2)\epsilon_n. \tag{5.10}$$

Nun wird die gewichtete Summe $\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^q_{n+1-i}$ betrachtet. Mit Lemma 2 gilt der Zusammenhang

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^q_{n+1-i} = \frac{1}{h}\sum_{i=1}^{k}\gamma_i(\mathbf{e}^q_{n+2-i} - \mathbf{e}^q_{n+1-i}). \tag{5.11}$$

Das Einsetzen von (5.10) in (5.11) liefert

$$\begin{aligned}\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^q_{n+1-i} = \frac{1}{h}\sum_{i=1}^{k}\gamma_i \mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0} - \sum_{i=1}^{k}\gamma_i \widehat{\mathbf{v}}(t_n)\mathbf{e}^q_{n+1-i} + \mathcal{O}(1)\sum_{i=1}^{k}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\| \\ + \mathcal{O}(h)\sum_{i=1}^{k}\epsilon_{n+1-i}\end{aligned}$$

mit $\mathbf{v}(t_{n+1-i}) = \mathbf{v}(t_n) + \mathcal{O}(h)$ und die Behauptung folgt mit den Sätzen 8 bzw. 9.

■

Globale Fehlergleichung für $\mathbf{e}^{\mathbf{v}}_n$

Die Berechnung der globalen Fehlerrekursion für $\mathbf{e}^{\mathbf{v}}_n$ ist unkompliziert, da die entsprechenden Gleichungen nur Rechnungen im linearen Raum enthalten, wie der folgende Satz zeigt. Es wird jedoch wie zuvor in Kapitel 4 zwischen der Normalen- und Tangentialrichtung der Zwangsmannigfaltigkeit durch Multiplikation mit $\mathbf{B}(q)$ bzw. $\mathbf{P}(q)$ unterschieden.

Satz 11
Die globalen Fehler $\mathbf{e}^{\mathbf{v}}_n$ der Munthe-Kaas-BDF-Verfahren (3.57) und der BLieDF-Verfahren (3.62) erfüllen

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^{\mathbf{v}}_{n+1-i} = -\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1} + \mathcal{O}(1)\epsilon_{n+1} + \mathbf{l}^{\mathbf{M}^{-1}\mathbf{v}}_n, \tag{5.12a}$$

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^{\mathbf{B}\mathbf{v}}_{n+1-i} = -\mathbf{e}^{\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1} + \mathcal{O}(1)\sum_{i=0}^{k}\epsilon_{n+1-i} + \mathbf{l}^{\mathbf{B}\mathbf{M}^{-1}\mathbf{v}}_n, \tag{5.12b}$$

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^{\mathbf{P}\mathbf{v}}_{n+1-i} = \mathcal{O}(1)\sum_{i=0}^{k}\epsilon_{n+1-i} + \mathbf{l}^{\mathbf{P}\mathbf{M}^{-1}\mathbf{v}}_n. \tag{5.12c}$$

Beweis:
Wird (3.62c) bzw. (3.57d) von links mit $\mathbf{M}(q_{n+1})^{-1}$ multipliziert und (5.1c) bzw. (5.3d) von links mit $\mathbf{M}(q(t_{n+1}))^{-1}$ multipliziert und beides voneinander subtrahiert, so folgt

$$\begin{aligned}\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{e}^{\mathbf{v}}_{n+1-i} = -\mathbf{M}\big(q(t_{n+1})\big)^{-1}\big(\mathbf{g}\big(t_{n+1}, q(t_{n+1}), \mathbf{v}(t_{n+1})\big) + \mathbf{B}^\top\big(q(t_{n+1})\big)\boldsymbol{\lambda}(t_{n+1})\big) \\ + \mathbf{M}\big(q_{n+1}\big)^{-1}\big(\mathbf{g}\big(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}\big) + \mathbf{B}^\top\big(q_{n+1}\big)\boldsymbol{\lambda}_{n+1}\big) + \mathbf{l}^{\mathbf{M}^{-1}\mathbf{v}}_n.\end{aligned}$$

Das Einsetzen von (4.19) und Taylorentwicklung liefert Gleichung (5.12a), da

$$q(t_{n+1}) = q_{n+1} \circ \exp\big(\widetilde{\mathbf{e}}^q_{n+1}\big) = q_{n+1} + \mathcal{O}(1)\|\mathbf{e}^q_{n+1}\|$$

gilt, wenn die Voraussetzung 3 erfüllt ist (siehe (2.26)).
Die Gleichungen (5.12b) und (5.12c) folgen durch Multiplikation mit $\mathbf{B}(q(t_{n+1}))$ bzw. $\mathbf{P}(q(t_{n+1}))$, da $[\mathbf{P}\mathbf{M}^{-1}\mathbf{B}^\top](q) = \mathbf{0}$.

■

Globale Fehlergleichung für $\mathbf{e}_n^{\boldsymbol{\lambda}}$

In diesem Abschnitt soll die Fehlerrekursion von $\mathbf{e}_n^{\boldsymbol{\lambda}}$ untersucht werden. Erneut wird mit den Abschätzungen für die Produkte von $\mathbf{B}(q)$ mit den Fehlertermen $\mathbf{e}_n^q$ begonnen.

Lemma 18
Wenn $n \geq k-1$ und $k \geq i \geq 1$ ist, dann folgt die Abschätzung

$$\begin{aligned}\mathbf{B}\big(q(t_{n+1-i})\big)\frac{\mathbf{e}_{n+2-i}^q - \mathbf{e}_{n+1-i}^q}{h} + \mathbf{Z}\big(q(t_{n+1-i})\big)\big(\mathbf{e}_{n+1-i}^q, \mathbf{v}(t_{n+1-i})\big)\\ = \mathcal{O}\left(\frac{1}{h}\right)\max_r \|\boldsymbol{\Phi}(q_r)\| + \mathcal{O}(h)\left(\|\mathbf{e}_{n+1-i}^q\| + \frac{1}{h}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right)\end{aligned}$$

mit dem Krümmungsterm $\mathbf{Z}$ aus (3.47).

Beweis:
Dieses Lemma ist eine direkte Folgerung aus [2, Lemma 4], da

$$\frac{\mathbf{e}_{n+2-i}^q - \mathbf{e}_{n+1-i}^q}{h} = \frac{1}{h}\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}} + \mathcal{O}(h)\left(\|\mathbf{e}_{n+1-i}^q\| + \frac{1}{h}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right)$$

erfüllt ist, siehe (5.6).

■

Wie zuvor beim Beweis für das Generalized-α-Verfahren (3.55) wird der globale Fehler $\mathbf{e}_n^{\mathbf{v}}$ in Normalenrichtung durch Multiplikation mit $\mathbf{B}(q)$ untersucht. Um dabei in den nachfolgenden Fehlerrekursionen nicht zwischen der Start- und Laufphase unterscheiden zu müssen, wird ein Term $\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}$ eingeführt, für den

$$\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}} = \mathcal{O}(h^{k+1}), \quad (n+1 = -k, -k+1, ..., 0, 1, ..., k-1), \tag{5.13a}$$

$$\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}} = \mathbf{0}_{N\times 1}, \quad (n+1 \geq k), \tag{5.13b}$$

gilt. $\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}$ ist somit nur in den ersten Zeitschritten ungleich null. Das nachfolgende Lemma kombiniert somit die Voraussetzung an die Startwerte 4 mit einer Folgerung aus Lemma 18.

Lemma 19
Unter Voraussetzung 4 gibt es einen Term $\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}$ mit (5.13) und es gilt für die BDF-Verfahren (3.57) und (3.62) die Abschätzung

$$\begin{aligned}\mathbf{e}_{n+1}^{\mathbf{Bv}} + \mathbf{B}\big(q(t_{n+1})\big)\mathbf{l}_n^{\boldsymbol{\omega}} = \mathcal{O}\left(\frac{1}{h}\right)\max_r \|\boldsymbol{\Phi}(q_r)\| - \mathbf{B}\big(q(t_{n+1})\big)\mathbf{b}\big(t_n, \mathbf{e}_n^q, ..., \mathbf{e}_{n+1-k}^q\big)\\ - \mathbf{Z}\big(q(t_n)\big)\left(\sum_{i=1}^{k}\gamma_i\mathbf{e}_{n+1-i}^q, \mathbf{v}(t_n)\right) + \boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}\\ + \mathcal{O}(h)\left(\sum_{i=0}^{k}\epsilon_{n+1-i} + \frac{1}{h}\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right)\end{aligned} \tag{5.14}$$

für $n+1 \geq -k$.

Beweis:
Für $n+1 < k$ ist diese Fehlerabschätzung eine direkte Konsequenz aus Voraussetzung 4 an die Startwerte und unter Verwendung von $t_{n+1} = t_{k-1} + \mathcal{O}(h)$ und $\mathbf{l}^{\boldsymbol{\omega}}_n = \mathbf{l}^{\boldsymbol{\omega}}_{k-1} + (\mathbf{l}^{\boldsymbol{\omega}}_{k-2} - \mathbf{l}^{\boldsymbol{\omega}}_{k-1}) + ... + (\mathbf{l}^{\boldsymbol{\omega}}_n - \mathbf{l}^{\boldsymbol{\omega}}_{n+1}) = \mathbf{l}^{\boldsymbol{\omega}}_{k-1} + \mathcal{O}(h^{k+1})$.
Für $n+1 \geq k$ wird die Fehlerrekursion (5.9) von links mit $\mathbf{B}\big(q(t_{n+1})\big)$ multipliziert. Die Behauptung folgt direkt aus Lemma 18 wegen (5.11) und $\mathbf{B}\big(q(t_{n+1})\big) = \mathbf{B}\big(q(t_{n+1-i})\big) + \mathcal{O}(h)$.

■

Somit kann eine Abschätzung für den globalen Fehler $\mathbf{e}^{\boldsymbol{\lambda}}_{n+1}$ bewiesen werden.

Satz 12
Die globalen Fehler $\mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1}$ der BDF-Verfahren (3.57) und (3.62) erfüllen die Abschätzung

$$\begin{aligned}\mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1} = &-\frac{1}{h}\sum_{i=0}^{k}\alpha_i \boldsymbol{\Delta}^{\mathbf{B}\mathbf{v}}_{n-i+1} + \mathcal{O}(1)\sum_{i=1}^{k}\left\|\frac{\mathbf{l}^{\boldsymbol{\omega}}_{n+1-i} - \mathbf{l}^{\boldsymbol{\omega}}_{n-i}}{h}\right\| + \mathcal{O}(1)\sum_{i=0}^{k}\|\mathbf{l}^{\boldsymbol{\omega}}_{n-i}\| + \mathcal{O}(1)\|\mathbf{l}^{\mathbf{v}}_n\| \\ &+ \mathcal{O}\left(\frac{1}{h^2}\right)\max_r\|\boldsymbol{\Phi}(q_r)\| + \mathcal{O}(1)\left(\sum_{i=0}^{2k}\epsilon_{n+1-i} + \frac{1}{h}\sum_{i=1}^{2k}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\|\right) \qquad (5.15)\end{aligned}$$

mit $\mathbf{S} \coloneqq \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top$.

Beweis:
Aus Gleichung (5.12b) folgt

$$\begin{aligned}\mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1} = &-\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{B}\big(q(t_{n+1-i})\big)\big(\mathbf{e}^{\mathbf{v}}_{n+1-i} + \mathbf{l}^{\boldsymbol{\omega}}_{n-i}\big) + \frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{B}\big(q(t_{n+1-i})\big)\mathbf{l}^{\boldsymbol{\omega}}_{n-i} \\ &+ \mathcal{O}(1)\sum_{i=0}^{k}\epsilon_{n+1-i} + \mathbf{l}^{\mathbf{B}\mathbf{M}^{-1}\mathbf{v}}_n.\end{aligned}$$

Durch das Einsetzen von (5.14) wird die Abschätzung

$$\begin{aligned}\mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1} = &\frac{1}{h}\mathbf{B}\big(q(t_n)\big)\sum_{i=0}^{k}\alpha_i \mathbf{b}(t_{n-i}, \mathbf{e}^q_{n-i}, ..., \mathbf{e}^q_{n+1-i-k}) - \frac{1}{h}\sum_{i=0}^{k}\alpha_i \boldsymbol{\Delta}^{\mathbf{B}\mathbf{v}}_{n-i+1} \\ &+ \frac{1}{h}\mathbf{Z}\big(q(t_n)\big)\left(\sum_{i=0}^{k}\alpha_i\sum_{j=1}^{k}\gamma_j \mathbf{e}^q_{n-i+1-j}, \mathbf{v}(t_n)\right) + \mathcal{O}(1)\sum_{i=1}^{k}\left\|\frac{\mathbf{l}^{\boldsymbol{\omega}}_{n+1-i} - \mathbf{l}^{\boldsymbol{\omega}}_{n-i}}{h}\right\| \\ &+ \mathcal{O}(1)\sum_{i=0}^{k}\|\mathbf{l}^{\boldsymbol{\omega}}_{n-i}\| + \mathcal{O}(1)\|\mathbf{l}^{\mathbf{v}}_n\| + \mathcal{O}\left(\frac{1}{h^2}\right)\max_r\|\boldsymbol{\Phi}(q_r)\| \\ &+ \mathcal{O}(1)\left(\sum_{i=0}^{2k}\epsilon_{n+1-i} + \frac{1}{h}\sum_{i=1}^{2k}\|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\|\right)\end{aligned}$$

erhalten mit

$$\frac{1}{h}\sum_{i=0}^{k}\alpha_i \mathbf{B}\big(q(t_{n+1-i})\big)\mathbf{l}^{\boldsymbol{\omega}}_{n-i} = \mathbf{B}\big(q(t_{n+1})\big)\sum_{i=1}^{k}\gamma_i\frac{\mathbf{l}^{\boldsymbol{\omega}}_{n+1-i} - \mathbf{l}^{\boldsymbol{\omega}}_{n-i}}{h} + \mathcal{O}(1)\sum_{i=0}^{k}\|\mathbf{l}^{\boldsymbol{\omega}}_{n-i}\|$$

$$= \mathcal{O}(1) \sum_{i=1}^{k} \left\| \frac{\mathbf{l}^{\boldsymbol{\omega}}_{n+1-i} - \mathbf{l}^{\boldsymbol{\omega}}_{n-i}}{h} \right\| + \mathcal{O}(1) \sum_{i=0}^{k} \|\mathbf{l}^{\boldsymbol{\omega}}_{n-i}\|.$$

Nach Voraussetzung 5 sind $\epsilon_i = 0$ und $\mathbf{e}^{\boldsymbol{\omega}}_{i,0} = \mathbf{0}$, $(i = -k, ..., -1)$. Für die Munthe-Kaas-BDF-Verfahren (3.57) ist

$$\begin{aligned}
&\frac{1}{h}\mathbf{B}\big(q(t_n)\big) \sum_{i=0}^{k} \alpha_i \mathbf{b}(t_{n-i}, \mathbf{e}^q_{n-i}, ..., \mathbf{e}^q_{n+1-i-k}) \\
&= -\frac{1}{h}\mathbf{B}\big(q(t_n)\big) \sum_{i=0}^{k} \alpha_i \widehat{\mathbf{v}}(t_{n-i}) \mathbf{e}^q_{n-i} \\
&= -\frac{1}{h}\mathbf{B}\big(q(t_n)\big)\widehat{\mathbf{v}}(t_n) \sum_{i=0}^{k} \alpha_i \mathbf{e}^q_{n-i} + \mathcal{O}(1) \sum_{i=1}^{k+1} \epsilon_{n+1-i} \\
&= -\mathbf{B}\big(q(t_n)\big)\widehat{\mathbf{v}}(t_n) \sum_{i=1}^{k} \gamma_i \frac{\mathbf{e}^q_{n+1-i} - \mathbf{e}^q_{n-i}}{h} + \mathcal{O}(1) \sum_{i=1}^{k+1} \epsilon_{n+1-i} \\
&= \mathcal{O}(1) \left(\sum_{i=1}^{k+1} \epsilon_{n+1-i} + \frac{1}{h} \sum_{i=2}^{k+1} \|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\| \right)
\end{aligned}$$

und für die BLieDF-Verfahren (3.62) ist

$$\begin{aligned}
&\frac{1}{h}\mathbf{B}\big(q(t_n)\big) \sum_{i=0}^{k} \alpha_i \mathbf{b}(t_{n-i}, \mathbf{e}^q_{n-i}, ..., \mathbf{e}^q_{n+1-i-k}) \\
&= -\frac{1}{h}\mathbf{B}\big(q(t_n)\big) \sum_{i=0}^{k} \alpha_i \widehat{\mathbf{v}}(t_{n-i}) \sum_{j=1}^{k} \gamma_j \mathbf{e}^q_{n+1-i-j} \\
&= -\frac{1}{h}\mathbf{B}\big(q(t_n)\big)\widehat{\mathbf{v}}(t_n) \sum_{i=0}^{k} \alpha_i \sum_{j=1}^{k} \gamma_j \mathbf{e}^q_{n+1-i-j} + \mathcal{O}(1) \sum_{i=0}^{k} \sum_{j=1}^{k} \left\|\mathbf{e}^q_{n+1-i-j}\right\|
\end{aligned}$$

erfüllt, vgl. Gleichung (5.6). Die Behauptung folgt aufgrund von

$$\begin{aligned}
\frac{1}{h} \sum_{i=0}^{k} \sum_{j=1}^{k} \alpha_i \gamma_j \mathbf{e}^q_{n+1-i-j} &= \sum_{i=1}^{k} \sum_{j=1}^{k} \gamma_i \gamma_j \frac{\mathbf{e}^q_{n+2-i-j} - \mathbf{e}^q_{n+1-i-j}}{h} \\
&= \mathcal{O}(1) \left(\sum_{i=1}^{2k} \epsilon_{n+1-i} + \frac{1}{h} \sum_{i=2}^{2k} \|\mathbf{e}^{\boldsymbol{\omega}}_{n+1-i,0}\| \right).
\end{aligned}$$

■

5.1.3 Zwei-Term-Fehlerrekursion und Konvergenz

Die Gleichungen für die globalen Fehler $\mathbf{e}^q_n$, $\mathbf{e}^{\mathbf{v}}_n$, $\mathbf{e}^{\boldsymbol{\omega}}_{n,0}$ und $\mathbf{e}^{\boldsymbol{\lambda}}_n$ aus dem vorherigen Abschnitt können nun zu der Zwei-Term-Fehlerrekursion

$$\|\mathbf{E}^{\mathbf{y}}_{n+1} - \mathbf{T}_{\mathbf{y}}\mathbf{E}^{\mathbf{y}}_n\| \leq L_0 h\big(\|\mathbf{E}^{\mathbf{y}}_n\| + \|\mathbf{E}^{\mathbf{y}}_{n+1}\| + \|\mathbf{E}^{\mathbf{z}}_n\| + \|\mathbf{E}^{\mathbf{z}}_{n+1}\|\big) + hM_0, \tag{5.16a}$$

$$\|\mathbf{E}^{\mathbf{z}}_{n+1} - \mathbf{T}_{\mathbf{z}}\mathbf{E}^{\mathbf{z}}_n\| \leq L_0\big(\|\mathbf{E}^{\mathbf{y}}_n\| + \|\mathbf{E}^{\mathbf{y}}_{n+1}\| + h\|\mathbf{E}^{\mathbf{z}}_n\| + h\|\mathbf{E}^{\mathbf{z}}_{n+1}\|\big) + M_0 \tag{5.16b}$$

mit positiven Konstanten L_0, M_0, die unabhängig von $h > 0$ und $n \geq 0$ sind, kombiniert werden.

Lemma 20
Die globalen Fehler der Munthe-Kaas-BDF-Verfahren (3.57) und der BLieDF-Verfahren (3.62) erfüllen für $n \geq k-1$ die gekoppelte Fehlerrekursion (5.16) mit

$$\mathbf{E}_n^{\mathbf{y}} := \begin{bmatrix} \mathbf{e}_n^q \\ \vdots \\ \mathbf{e}_{n+1-2k}^q \\ \mathbf{e}_n^{\mathbf{Pv}} \\ \vdots \\ \mathbf{e}_{n+1-2k}^{\mathbf{Pv}} \end{bmatrix}, \quad \mathbf{E}_n^{\mathbf{z}} := \begin{bmatrix} \frac{1}{h}\mathbf{e}_{n-1,0}^{\boldsymbol{\omega}} \\ \vdots \\ \frac{1}{h}\mathbf{e}_{n+1-2k,0}^{\boldsymbol{\omega}} \\ \mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}} \\ \vdots \\ \mathbf{e}_{n+1-2k}^{\mathbf{S}\boldsymbol{\lambda}} \end{bmatrix}$$

und

$$\mathbf{T}_{\mathbf{y}} := \begin{bmatrix} \mathbf{T}_\alpha & \mathbf{0}_{2k\times 2k} \\ \mathbf{0}_{2k\times 2k} & \mathbf{T}_\alpha \end{bmatrix} \otimes \mathbf{I}_N, \quad \mathbf{T}_{\mathbf{z}} = \begin{bmatrix} \mathbf{T}_\gamma & \mathbf{0}_{(2k-1)\times 2k} \\ \mathbf{0}_{2k\times(2k-1)} & \mathbf{J}_{2k} \end{bmatrix} \otimes \mathbf{I}_N,$$

mit

$$\begin{aligned} \mathbf{T}_\alpha &= -\frac{1}{\alpha_0}\mathbf{e}_{1,2k}\cdot(\alpha_1,\alpha_2,...,\alpha_k,0,...,0) + \mathbf{J}_{2k} \in \mathbb{R}^{2k\times 2k}, \\ \mathbf{T}_\gamma &= -\frac{1}{\gamma_1}\mathbf{e}_{1,2k-1}\cdot(\gamma_2,\gamma_3,...,\gamma_k,0,...,0) + \mathbf{J}_{2k-1} \in \mathbb{R}^{(2k-1)\times(2k-1)}, \end{aligned}$$

wobei $\mathbf{e}_{1,r} \in \mathbb{R}^r$ der erste Einheitsvektor ist und $\mathbf{J}_r \in \mathbb{R}^{r\times r}$ mit $\mathbf{J}_r = \left(j_{il}^{(r)}\right)_{i,l=1}^r$ und $j_{il}^{(r)} = \delta_{i+1,l}$.

Beweis:
Aus Gleichung (5.12) kann

$$\begin{aligned} &\begin{bmatrix} \mathbf{e}_{n+1}^{\mathbf{Pv}} \\ \vdots \\ \mathbf{e}_{n+2-2k}^{\mathbf{Pv}} \end{bmatrix} - \begin{bmatrix} -\frac{\alpha_1}{\alpha_0}\mathbf{I}_N & -\frac{\alpha_2}{\alpha_0}\mathbf{I}_N & \cdots & -\frac{\alpha_k}{\alpha_0}\mathbf{I}_N & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{I}_N & \mathbf{0}_{N\times N} & & \cdots & & & \mathbf{0}_{N\times N} \\ & & \ddots & \ddots & & & \vdots \\ & & & & & \mathbf{I}_N & \mathbf{0}_{N\times N} \end{bmatrix} \begin{bmatrix} \mathbf{e}_n^{\mathbf{Pv}} \\ \vdots \\ \mathbf{e}_{n+1-2k}^{\mathbf{Pv}} \end{bmatrix} \\ =&\mathcal{O}(h)\left(\|\mathbf{E}_{n+1}^{\mathbf{y}}\| + \|\mathbf{E}_n^{\mathbf{y}}\| + \|\mathbf{E}_{n+1}^{\mathbf{z}}\| + \|\mathbf{E}_n^{\mathbf{z}}\|\right) + \mathcal{O}(h^{k+1}) \end{aligned} \tag{5.17}$$

erhalten werden mit den Sätzen 6 bzw. 7. Ebenso führt (5.9) mit den Sätzen 6 bzw. 7 auf

$$\begin{aligned} &\begin{bmatrix} \mathbf{e}_{n+1}^{q} \\ \vdots \\ \mathbf{e}_{n+2-2k}^{q} \end{bmatrix} - \begin{bmatrix} -\frac{\alpha_1}{\alpha_0}\mathbf{I}_N & -\frac{\alpha_2}{\alpha_0}\mathbf{I}_N & \cdots & -\frac{\alpha_k}{\alpha_0}\mathbf{I}_N & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{I}_N & \mathbf{0}_{N\times N} & & \cdots & & & \mathbf{0}_{N\times N} \\ & & \ddots & \ddots & & & \vdots \\ & & & & & \mathbf{I}_N & \mathbf{0}_{N\times N} \end{bmatrix} \begin{bmatrix} \mathbf{e}_n^{q} \\ \vdots \\ \mathbf{e}_{n+1-2k}^{q} \end{bmatrix} \\ =&\mathcal{O}(h)\left(\|\mathbf{E}_{n+1}^{\mathbf{y}}\| + \|\mathbf{E}_{n+1}^{\mathbf{z}}\|\right) + \mathcal{O}(h^{k+1}). \end{aligned} \tag{5.18}$$

Die Kombination von (5.17) und (5.18) führt auf (5.16a). Aus den Sätzen 8, 9 und 7 folgt

$$\begin{bmatrix} \frac{1}{h}\mathbf{e}^{\boldsymbol{\omega}}_{n,0} \\ \vdots \\ \frac{1}{h}\mathbf{e}^{\boldsymbol{\omega}}_{n+2-2k,0} \end{bmatrix} - \begin{bmatrix} -\frac{\gamma_2}{\gamma_1}\mathbf{I}_N & -\frac{\gamma_3}{\gamma_1}\mathbf{I}_N & \cdots & -\frac{\gamma_k}{\gamma_1}\mathbf{I}_N & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{I}_N & \mathbf{0}_{N\times N} & & \cdots & & & \mathbf{0}_{N\times N} \\ & & \ddots & \ddots & & & \vdots \\ & & & & & \mathbf{I}_N & \mathbf{0}_{N\times N} \end{bmatrix} \begin{bmatrix} \frac{1}{h}\mathbf{e}^{\boldsymbol{\omega}}_{n-1,0} \\ \vdots \\ \frac{1}{h}\mathbf{e}^{\boldsymbol{\omega}}_{n+1-2k,0} \end{bmatrix}$$

$$= \mathcal{O}(1)\big(\|\mathbf{E}^{\mathbf{y}}_{n+1}\| + \|\mathbf{E}^{\mathbf{y}}_{n}\| + h\|\mathbf{E}^{\mathbf{z}}_{n+1}\| + h\|\mathbf{E}^{\mathbf{z}}_{n}\|\big) + \mathcal{O}(h^k) \tag{5.19}$$

und aus (5.15) mit (5.13), den Sätzen 6 bzw. 7, Voraussetzung 4 folgt

$$\begin{bmatrix} \mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1} \\ \vdots \\ \mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+2-2k} \end{bmatrix} - \begin{bmatrix} \mathbf{0}_{N\times N} & \cdots & & \mathbf{0}_{N\times N} \\ \mathbf{I}_N & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ & \ddots & \ddots & \vdots \\ & & \mathbf{I}_N & \mathbf{0}_{N\times N} \end{bmatrix} \begin{bmatrix} \mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n} \\ \vdots \\ \mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1-2k} \end{bmatrix}$$

$$= \mathcal{O}(1)\big(\|\mathbf{E}^{\mathbf{y}}_{n+1}\| + \|\mathbf{E}^{\mathbf{y}}_{n}\| + h\|\mathbf{E}^{\mathbf{z}}_{n+1}\| + h\|\mathbf{E}^{\mathbf{z}}_{n}\|\big) + \mathcal{O}(h^k). \tag{5.20}$$

Die Kombination aus (5.19) und (5.20) liefert Gleichung (5.16b).

■

Anhand der Fehlerrekursion (5.16) können durch das nachfolgende Lemma die globalen Fehler zur Zeit t_{n+1} auf t_0 zurückgeführt werden.

Lemma 21 ([4, Theorem 4.16])
Es seien $(\mathbf{E}^{\mathbf{y}}_n)_{n\geq 0}$ und $(\mathbf{E}^{\mathbf{z}}_n)_{n\geq 0}$ vektorwertige Folgen, die die gekoppelte Fehlerrekursion (5.16) erfüllen mit positiven Konstanten L_0, M_0, die unabhängig von $h > 0$ und $n \geq 0$ sind. Wenn es Normen $\|.\|_{\mathbf{y}}$ und $\|.\|_{\mathbf{z}}$ gibt, so dass $\|\mathbf{T}_{\mathbf{y}}\|_{\mathbf{y}} = 1$ und $\|\mathbf{T}_{\mathbf{z}}\|_{\mathbf{z}} < 1$ gilt, dann impliziert (5.16) für Zeitschrittweiten $h \in (0, \overline{h}]$ die Fehlerschranken

$$\|\mathbf{E}^{\mathbf{y}}_n\| \leq e^{\overline{L}_0(t_n - t_0)}\big(\|\mathbf{E}^{\mathbf{y}}_0\| + \overline{C}_0 h\|\mathbf{E}^{\mathbf{z}}_0\|\big) + \frac{e^{\overline{L}_0(t_n-t_0)} - 1}{\overline{L}_0}\overline{M}_0, \tag{5.21a}$$

$$\|\mathbf{E}^{\mathbf{z}}_n - \mathbf{T}^n_{\mathbf{z}}\mathbf{E}^{\mathbf{z}}_0\| \leq \overline{C}_0 e^{\overline{L}_0(t_n-t_0)}\big(\|\mathbf{E}^{\mathbf{y}}_0\| + h\|\mathbf{E}^{\mathbf{z}}_0\| + \overline{M}_0\big) \tag{5.21b}$$

mit $t_n = t_0 + nh$, $(n \geq 0)$ und gewissen positiven Konstanten $\overline{h}$, $\overline{C}_0$, $\overline{L}_0$ und $\overline{M}_0$. Diese hängen von den Konstanten L_0 und M_0 aus (5.16) und von den Normen $\|.\| = \|.\|_{\mathbf{y}}$ und $\|.\| = \|.\|_{\mathbf{z}}$ für $\mathbf{E}^{\mathbf{y}}_n$ und $\mathbf{E}^{\mathbf{z}}_n$ ab und sind unabhängig von n und h.

Das vorausgehende Lemma kann nun verwendet werden, um die Konvergenz der BDF-Verfahren (3.57) und (3.62) für das gesamte Zeitintervall zu beweisen.

Satz 13
Unter den Voraussetzungen 3-5 gibt es positive Konstanten C_0, $\overline{L}$ und $\overline{h}$ unabhängig von n und h, so dass für alle $h \in (0, \overline{h}]$ und alle $n \geq 0$ mit $t_0 + nh \leq t_{\text{end}} - h$ die globalen Fehlerabschätzungen der Munthe-Kaas-BDF-Verfahren (3.57) und der BLieDF-Verfahren (3.62) die Bedingungen

$$\|\mathbf{e}^{q}_n\| + \|\mathbf{e}^{\mathbf{v}}_n\| + \|\mathbf{e}^{\boldsymbol{\lambda}}_n\| \leq C_0 e^{\overline{L}(t_n - t_0)} h^k$$

erfüllen und die k-Schritt-BDF-Integratoren (3.57) und (3.62) die Konvergenzordnung $p = k$ für $2 \leq k \leq 6$ besitzen.

Beweis:
Mit Lemma 20 ist für die Verfahren (3.57) und (3.62) die gekoppelte Fehlerrekursion (5.16) erfüllt. Außerdem gibt es eine Norm $\|.\|_{\mathbf{y}}$ mit $\|\mathbf{T}_{\mathbf{y}}\|_{\mathbf{y}} = 1$ (vgl. [28, Lemma III.4.4]). Für $k \leq 6$ sind BDF-Verfahren nullstabil und die Wurzeln des charakteristischen Polynoms $p_\alpha(\zeta) = \sum_{i=0}^{k} \alpha_i \zeta^{k-i}$ erfüllen $\zeta_1 = 1$ und $|\zeta_i| < 1$, $(i = 2, ..., k)$, vgl. [28]. Wegen $\det\left(\zeta \mathbf{I}_{2k-1} - \mathbf{T}_\gamma\right) = \zeta^k \sum_{i=1}^{k} \gamma_i \zeta^{k-i}/\gamma_1$ und

$$(\zeta - 1) \sum_{i=1}^{k} \gamma_i \zeta^{k-i} = \sum_{i=1}^{k} \gamma_i (\zeta^{k+1-i} - \zeta^{k-i}) = \sum_{i=0}^{k} \alpha_i \zeta^{k-i} = p_\alpha(\zeta)$$

sind alle Eigenwerte von $\mathbf{T}_\gamma$ innerhalb des Einheitskreises und es gibt eine Norm $\|.\|_{\mathbf{z}}$ mit $\|\mathbf{T}_{\mathbf{z}}\|_{\mathbf{z}} < 1$. Mit Lemma 21 können die Abschätzungen (5.21) erhalten werden für Schrittweiten $h \in (0, \overline{h}]$, $t_n = t_0 + nh$ und positiven Konstanten $\overline{h}$, $\overline{C}_0$, $\overline{L}_0$, $\overline{M}_0$. Mit den gegebenen Voraussetzungen folgen $\|\mathbf{E}_{k-1}^{\mathbf{y}}\| = \mathcal{O}(h^k)$ und $\|\mathbf{E}_{k-1}^{\mathbf{z}}\| = \mathcal{O}(h^k)$ und daher die Behauptung.

■

Bemerkung 20 (Sonderfall: Konvergenz erster Ordnung)
Für $k = 1$ kann die technische Voraussetzung 3 nicht angewendet werden. Gefordert werden könnte, dass es positive Konstanten $\overline{h}$ und C_T gibt, so dass für alle $h \in (0, \overline{h}]$ und alle r mit $t_0 + rh \in [t_0, t_{\text{end}}]$ die globalen Fehler durch

$$\|\mathbf{e}_r^q\| + \|\mathbf{e}_{r,i}^{\boldsymbol{\omega}}\| \leq C_T, \ \|\mathbf{e}_r^{\mathbf{v}}\| + \|\mathbf{e}_r^{\boldsymbol{\lambda}}\| \leq C_T, \quad (i = 0, ..., k),$$

beschränkt bleiben. Diese Voraussetzung reicht jedoch nicht aus, um die Konvergenz des Verfahrens auf vorherige Weise zu beweisen. Ähnliche Probleme traten auch in anderen Quellen wie zum Beispiel [36] auf. Dort konnte das Problem für diesen Sonderfall gelöst werden. Jedoch wurden in der genannten Quelle lineare Räume betrachtet. In den vorliegenden Konfigurationsräumen mit Lie-Gruppen-Struktur sind zusätzliche Einschränkungen vorhanden, die sich nicht einfach lösen lassen. Zum Beispiel wird im Beweis von Satz 10 die Kombination der Exponentialfunktionen durch die Baker-Campbell-Hausdorff-Formel (2.28) bestimmt. Diese kann jedoch nur für Argumente innerhalb der Exponentialfunktionen angewendet werden, die gegen null streben. Die Argumente, die im Fall $k = 1$ betrachtet werden müssten, sind jedoch nur beschränkt. Eine Abschätzung mit $\mathcal{O}(h)$ wäre hier also nicht möglich und es müssten alle entstehenden Reihenglieder aufgeführt werden, was die theoretische Berechnung um einiges erschwert.
Aus diesem Grund wurde entschieden, den Fall $k = 1$ in die theoretischen Untersuchungen nicht weiter einzubeziehen. Die genannten Probleme scheinen jedoch nur von theoretischer und technischer Art zu sein, denn in den numerischen Tests in Kapitel 6 kann trotzdem eine Konvergenz erster Ordnung beobachtet werden.
Dies zu beweisen wird Gegenstand von nachfolgenden Arbeiten.

Bemerkung 21 (Wahl der Startwerte)
Um in den BDF-Verfahren eine Konvergenz der Ordnung $p = k$ zu erhalten, müssen die Startwerte die Voraussetzung 4 erfüllen. Die korrigierten Startwerte aus Abschnitt 3.3.4 für die BLieDF-Verfahren und aus Abschnitt 3.3.3 für die Munthe-Kaas-BDF-Verfahren erfüllen diese Bedingungen, Startwerte, die mit Funktionswerten der

exakten Lösung initialisiert werden, jedoch nicht, wie diese Bemerkung zeigen soll. Werden konsistente Startwerte bezüglich (3.44) gewählt, so gilt

$$\sum_{i=0}^{k-1} \|\mathbf{e}_i^q\| = 0, \quad \sum_{i=0}^{k-1} \|\mathbf{e}_i^{\mathbf{v}}\| + \|\mathbf{e}_i^{\boldsymbol{\lambda}}\| = 0, \quad \max_{0\le i\le k-1} \|\boldsymbol{\Phi}(q_i)\| = 0 \tag{5.22}$$

und alle Bedingungen aus Voraussetzung 4 sind erfüllt außer

$$\sum_{i=0}^{k-1} \left\|\mathbf{e}_i^{\mathbf{Bv}} + \mathbf{B}\big(q(t_{k-1})\big)\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\right\| = \mathcal{O}(h^{k+1}), \tag{5.23}$$

denn es gilt

$$\|\mathbf{e}_i^{\mathbf{Bv}} + \mathbf{B}\big(q(t_{k-1})\big)\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\| \overset{\text{Satz } 6}{=} \|\mathbf{B}\big(q(t_i)\big)\mathbf{e}_i^{\mathbf{v}} + \mathcal{O}(h^k)\| \overset{(5.22)}{=} \mathcal{O}(h^k).$$

Deshalb kann für Startwerte (5.22) bei numerischer Rechnung eine Ordnungsreduktion in den ersten Zeitschritten auftreten (siehe Kapitel 6).
Um dies zu verhindern, werden die Startwerte $\mathbf{v}_j$, $(j = 0, ..., k-1)$, um den Korrekturterm $\boldsymbol{\Delta}_{k-1}^{\mathbf{v}}$ ergänzt, vgl. (3.58) und (3.63). Mit Hilfe dieses Terms gelten mit den Sätzen 6 und 7 für $i = 0, ..., k-1$

$$\|\mathbf{e}_i^{\mathbf{v}}\| \overset{(4.19c)}{=} \|\mathbf{v}(t_i) - (\mathbf{v}(t_i) + \boldsymbol{\Delta}_{k-1}^{\mathbf{v}})\| = \|\boldsymbol{\Delta}_{k-1}^{\mathbf{v}}\| = \mathcal{O}(1)\|\bar{\mathbf{l}}_{k-1}^{\boldsymbol{\omega}}\| = \mathcal{O}(h^k) \tag{5.24}$$

und

$$\begin{aligned}
\|\mathbf{e}_i^{\mathbf{Bv}} + \mathbf{B}\big(q(t_{k-1})\big)\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\| &\overset{(4.19c)}{=} \| - \mathbf{B}\big(q(t_i)\big)\boldsymbol{\Delta}_{k-1}^{\mathbf{v}} + \mathbf{B}\big(q(t_{k-1})\big)\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\| \\
&\overset{(5.24)}{=} \| - \mathbf{B}\big(q(t_{k-1})\big)\boldsymbol{\Delta}_{k-1}^{\mathbf{v}} + \mathbf{B}\big(q(t_{k-1})\big)\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\| + \mathcal{O}(h^{k+1}) \\
&\overset{\text{Satz } 6}{=} \| - \mathbf{B}\big(q(t_{k-1})\big)\boldsymbol{\Delta}_{k-1}^{\mathbf{v}} + \mathbf{B}\big(q(t_{k-1})\big)\bar{\mathbf{l}}_{k-1}^{\boldsymbol{\omega}}\| + \mathcal{O}(h^{k+1}) \\
&= \mathcal{O}(h^{k+1}),
\end{aligned}$$

wobei $\bar{\mathbf{l}}_{k-1}^{\boldsymbol{\omega}}$ der führende Fehlerterm von $\mathbf{l}_{k-1}^{\boldsymbol{\omega}}$ ist. Für die korrigierten Startwerte ist somit die Voraussetzung 4 erfüllt und in den praktischen Tests in Kapitel 6 ist im gesamten Zeitintervall die Ordnung $p = k$ zu beobachten.

Bemerkung 22 (Lokaler Fehler in der Lie-Gruppen-Formulierung $SE(3)$ für den schweren Kreisel)
Ist die Ableitungsmatrix der holonomen Zwangsbedingung $\mathbf{B} = \mathbf{B}(q)$ konstant, wie für den schweren Kreisel in der Lie-Gruppen-Formulierung $SE(3)$ (vgl. (3.53)), so ist

$$\mathbf{B}\mathbf{v}^{(i)}(t) = \mathbf{0} \text{ für alle } i \ge 0 \tag{5.25}$$

erfüllt, vgl. [4]. Für $i \ge 0$ folgt somit für $\mathbf{v}^{(i)}(t) = [\mathbf{U}^{(i)}(t)^\top \ \boldsymbol{\Omega}^{(i)}(t)^\top]^\top$

$$\mathbf{0} = \mathbf{B}\mathbf{v}^{(i)}(t) \overset{(3.53)}{=} \begin{bmatrix} -\mathbf{I}_3 & -\widetilde{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \mathbf{U}^{(i)}(t) \\ \boldsymbol{\Omega}^{(i)}(t) \end{bmatrix} = -\mathbf{U}^{(i)}(t) - \widetilde{\mathbf{X}}\boldsymbol{\Omega}^{(i)}(t). \tag{5.26}$$

Außerdem gilt für $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^3$

$$\widetilde{\widetilde{\mathbf{w}}_1\mathbf{w}_2} = \widetilde{\widehat{\mathbf{w}}_1\mathbf{w}_2} = \widetilde{\mathbf{w}}_1\widetilde{\mathbf{w}}_2 - \widetilde{\mathbf{w}}_2\widetilde{\mathbf{w}}_1, \tag{5.27}$$

vgl. (2.35) und (2.38).
Die Munthe-Kaas-BDF-Verfahren (3.57) und die BLieDF-Verfahren (3.62) sind für $k = 2$ identisch, vgl. Gleichung (6.5). Daher berechnet sich der lokale Abbruchfehler $\mathbf{l}_n^{\boldsymbol{\omega}}$ nach Lemma 5 für $k = 2$ zu

$$\mathbf{l}_n^{\boldsymbol{\omega}} = -\frac{h^2}{3}\ddot{\mathbf{v}}(t_n) + \frac{h^2}{12}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^3).$$

Wird nun $\mathbf{B}\mathbf{l}_n^{\boldsymbol{\omega}}$ betrachtet, so folgt mit (5.25) und

$$\begin{aligned}
\mathbf{B}\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) &\overset{(3.53),(2.41)}{=} \begin{bmatrix} -\mathbf{I}_3 & -\widetilde{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{\Omega}} & \widetilde{\dot{\mathbf{U}}} \\ \mathbf{0}_{3\times 3} & \widetilde{\boldsymbol{\Omega}} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{U}} \\ \dot{\boldsymbol{\Omega}} \end{bmatrix} = \begin{bmatrix} -\widetilde{\boldsymbol{\Omega}} & -\widetilde{\dot{\mathbf{U}}} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\Omega}} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{U}} \\ \dot{\boldsymbol{\Omega}} \end{bmatrix} \\
&\overset{(5.26)}{=} \begin{bmatrix} -\widetilde{\boldsymbol{\Omega}} & \widetilde{\widetilde{\mathbf{X}}\boldsymbol{\Omega}} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\Omega}} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{U}} \\ \dot{\boldsymbol{\Omega}} \end{bmatrix} \overset{(5.27)}{=} \begin{bmatrix} -\widetilde{\boldsymbol{\Omega}} & -\widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{U}} \\ \dot{\boldsymbol{\Omega}} \end{bmatrix} \\
&= -\widetilde{\boldsymbol{\Omega}}\dot{\mathbf{U}} - \widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{X}}\dot{\boldsymbol{\Omega}} = \widetilde{\boldsymbol{\Omega}}(-\dot{\mathbf{U}} - \widetilde{\mathbf{X}}\dot{\boldsymbol{\Omega}}) \\
&\overset{(5.26)}{=} \mathbf{0},
\end{aligned}$$

dass für den führenden Fehlerterm $\bar{\mathbf{l}}_n^{\boldsymbol{\omega}}$ dieses lokalen Abbruchfehlers gilt $\mathbf{B}\bar{\mathbf{l}}_n^{\boldsymbol{\omega}} = \mathbf{0}$. Dies bedeutet jedoch auch, dass die Bedingung

$$\sum_{i=0}^{k-1} \|\mathbf{e}_i^q\| + \left\|\mathbf{e}_i^{\mathbf{Bv}} + \mathbf{B}(q(t_{k-1}))\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\right\| = \mathcal{O}(h^{k+1})$$

für $k = 2$ auch für die analytische Lösung als Startwert, also $\mathbf{v}_j = \mathbf{v}(t_j)$, gilt. Für den schweren Kreisel in der Lie-Gruppen-Formulierung $SE(3)$ ist in den ersten Zeitschritten somit auch für die analytische Lösung als Startwert für $k = 2$ keine Ordnungsreduktion zu beobachten. Die numerischen Tests in Kapitel 6 weisen weiterhin darauf hin, dass auch für $k > 2$ die analytische Lösung als Startwert keinen Einfluss auf die Ordnung hat, auf einen formalen analytischen Beweis soll jedoch an dieser Stelle verzichtet werden.

5.2 Konvergenz der BLieDF-Verfahren für variable Schrittweiten

In diesem Abschnitt sollen die BLieDF-Verfahren (3.62) auf variable Schrittweiten erweitert und untersucht werden. Wie zuvor beim Generalized-α-Verfahren in Abschnitt 4.1 wird dafür zunächst das Verfahren (3.62) umgeschrieben. Dafür werden, anders als beim Generalized-α-Verfahren, aber wie für BDF-Verfahren üblich [11, 32], variable Parameter $\alpha_{i,n}$ und $\gamma_{i,n}$ verwendet. Der Fokus der Konvergenzanalyse liegt dabei auf den lokalen Abbruchfehlern und den Gleichungen für die globalen Fehler. Auf einen Beweis der Nullstabilität wird verzichtet, jedoch wird davon ausgegangen, dass es Schranken für die Schrittweitenverhältnisse gibt, so dass die Nullstabilität bewiesen werden könnte, vgl. [12].
Wie zuvor soll für die untersuchten Schrittweitenfolgen die Voraussetzung 1 erfüllt sein.

5.2.1 Übertragung der BLieDF-Verfahren auf variable Schrittweiten

Für die Übertragung von BDF-Verfahren auf variable Schrittweiten gibt es in der Literatur mehrere Varianten. Häufig verwendete Techniken, um die Schrittweite zu variieren, sind Interpolation und variable Koeffizienten. Bei ersterer wird in jedem Zeitschritt die Methode mit fester Schrittweite angewendet und wenn nötig werden die fehlende Werte der vorherigen Zeitschritte durch Interpolation berechnet. Solche Verfahren wurden zum Beispiel von Gear [22] und Hindmarsh [31] implementiert. In der zweiten Variante müssen die vergangenen Werte nicht neu berechnet bzw. interpoliert werden. Dafür werden Parameter $\alpha_{i,n} = \alpha_{i,n}(\sigma_n) \in \mathbb{R}$ verwendet, die vom Schrittweitenverhältnis σ_n abhängen. Implementierungen dazu wurden von Byrne und Hindmarsh entwickelt [11, 32]. Diese Schrittweitenverhältnisse können jedoch nicht beliebig gewählt werden; wie zuvor beim Generalized-α-Verfahren in Abschnitt 4.3 kann die Nullstabilität nur für σ_n in gewissen Schranken bewiesen werden. Solche Untersuchungen wurden von Calvo et al. [12] und Butcher et al. [10] vorgenommen. In dieser Arbeit soll die Variante der variablen Parameter verwendet werden.
Da die Parameter und der benötigte Korrekturterm für wachsende Ordnung aufgrund der benötigten Schrittweitenverhältnisse σ_{n-i}, $(i = 0, ..., k-2)$, sehr schnell kompliziert und unübersichtlich werden, sollen die BLieDF-Verfahren für variable Schrittweiten in dieser Arbeit nur für $2 \leq k \leq 3$ vorgestellt werden. Der nachfolgende Konvergenzbeweis lässt sich jedoch ohne Probleme auch auf $4 \leq k \leq 6$ anwenden, wenn die Parameter und der Korrekturterm entsprechend berechnet wurden.
Für $k \leq 3$ sind die Parameter der BDF-Verfahren wie im linearen Fall gegeben durch (vgl. [10, 12, 28])

$$k = 1: \quad \alpha_{0,n} = 1,\ \alpha_{1,n} = -1, \tag{5.28a}$$

$$k = 2: \quad \alpha_{0,n} = \frac{1+2\sigma_n}{1+\sigma_n},\ \alpha_{1,n} = -1-\sigma_n,\ \alpha_{2,n} = \frac{\sigma_n^2}{1+\sigma_n}, \tag{5.28b}$$

$$\begin{aligned} k = 3: \quad \alpha_{0,n} &= \frac{1+\sigma_{n-1}+\sigma_n(2+\sigma_{n-1}(4+3\sigma_n))}{(1+\sigma_n)(1+\sigma_{n-1}+\sigma_n\sigma_{n-1})}, \\ \alpha_{1,n} &= -\frac{(1+\sigma_n)(1+\sigma_{n-1}+\sigma_n\sigma_{n-1})}{1+\sigma_{n-1}}, \\ \alpha_{2,n} &= \frac{\sigma_n^2(1+\sigma_{n-1}+\sigma_n\sigma_{n-1})}{1+\sigma_n}, \\ \alpha_{3,n} &= -\frac{\sigma_n^2(1+\sigma_n)\sigma_{n-1}^3}{(1+\sigma_{n-1})(1+\sigma_{n-1}+\sigma_n\sigma_{n-1})}. \end{aligned} \tag{5.28c}$$

In den BLieDF-Verfahren (3.62) werden die BDF-Verfahren in Inkrementform für die Variable $q_n \in G$ bzw. $\boldsymbol{\omega}_n^{(0)} \in \mathbb{R}^N$ verwendet. Für die Parameter $\gamma_{i,n}$ dieser Inkrementvariante gilt analog zu (2.12)

$$\gamma_{k,n} \coloneqq \sum_{i=0}^{k-1} \alpha_{i,n}. \tag{5.29}$$

Aufgrund der DAE-Struktur der zu lösenden Gleichungen (3.44) sind die variablen Parameter (5.28) nicht ausreichend, um eine Konvergenz der Ordnung $p = k$ zu beweisen. Wie zuvor beim Generalized-α-Verfahren (4.11) muss vor jedem Zeitschritt die Geschwindigkeit $\mathbf{v}_n$ angepasst werden.

Bemerkung 23 (Anpassung der Geschwindigkeiten)
Bei den BLieDF-Verfahren für variable Schrittweiten muss vor jedem Zeitschritt $t_n \to t_{n+1}$ (außer dem ersten) die Geschwindigkeit durch

$$\overline{\mathbf{v}}_{n+1-i} = \mathbf{v}_{n+1-i} + \mathbf{\Delta}^{\mathbf{v}}_{n,i}, \quad (i = 1, ..., k), \tag{5.30}$$

ersetzt werden. Der Korrekturterm $\mathbf{\Delta}^{\mathbf{v}}_{n,i}$ kann durch die Lösung des Gleichungssystems

$$\begin{bmatrix} \mathbf{M}(q_{n+1-i}) & \mathbf{B}^\top(q_{n+1-i}) \\ \mathbf{B}(q_{n+1-i}) & \mathbf{0}_{M\times M} \end{bmatrix} \begin{bmatrix} \mathbf{\Delta}^{\mathbf{v}}_{n,i} \\ \mathbf{\Delta}^{\boldsymbol{\lambda}}_{n,i} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ -\frac{\gamma_{i,n}}{\alpha_{i,n}} \mathbf{B}(q_{n+1-i}) \mathbf{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i} \end{bmatrix}, \tag{5.31}$$

erhalten werden, wobei $\mathbf{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}$ die Differenz zweier aufeinanderfolgender lokaler Abbruchfehler $\mathbf{l}^{\boldsymbol{\omega}}_n$ ist und stets $\alpha_{i,n} \neq 0$ für $\sigma_n, \sigma_{n-1} > 0$ gilt. Dieser lokale Abbruchfehler $\mathbf{l}^{\boldsymbol{\omega}}_n$ wird später in Satz 14 berechnet. Es gilt für $i = 1, ..., k$ und $k = 2$

$$\begin{aligned}
\mathbf{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i} &\approx \left[\left(-\left(\frac{1}{\sigma_{n+1-i}} + 1 \right) \frac{h^2_{n+1-i}}{6} \ddot{\mathbf{v}} + \frac{2\sigma_{n+1-i} - 1}{12\sigma_{n+1-i}} h^2_{n+1-i} \widehat{\mathbf{v}}\dot{\mathbf{v}} \right) \right] (t_{n+1-i}) \\
&\quad + \left[\left(\left(\frac{1}{\sigma_{n-i}} + 1 \right) \frac{h^2_{n-i}}{6} \ddot{\mathbf{v}} - \frac{2\sigma_{n-i} - 1}{12\sigma_{n-i}} h^2_{n-i} \widehat{\mathbf{v}}\dot{\mathbf{v}} \right) \right] (t_{n-i}) \\
&= \left(\frac{1 + \sigma_{n-i}}{\sigma^2_{n+1-i}\sigma_{n-i}} - \frac{1 + \sigma_{n+1-i}}{\sigma_{n+1-i}} \right) \frac{h^2_{n+1-i}}{6} \ddot{\mathbf{v}}(t_n) \\
&\quad + \left(\frac{2\sigma_{n+1-i} - 1}{\sigma_{n+1-i}} - \frac{2\sigma_{n-i} - 1}{\sigma_{n-i}\sigma_{n+1-i}} \right) \frac{h^2_{n+1-i}}{12} \widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^3_n)
\end{aligned}$$

und für $k = 3$

$$\begin{aligned}
\mathbf{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i} &\approx \left[\left(-\frac{1}{24} \left(\frac{1 + \sigma_{n+1-i}}{\sigma^2_{n+1-i}\sigma_{n-i}} \right) h^3_{n+1-i} \dddot{\mathbf{v}} + s(\sigma_{n+1-i}, \sigma_{n-i}) h^3_{n+1-i} \widehat{\mathbf{v}}\ddot{\mathbf{v}} \right) \right] (t_{n+1-i}) \\
&\quad + \left[\left(\frac{1}{24} \left(\frac{1 + \sigma_{n-i}}{\sigma^2_{n-i}\sigma_{n-i-1}} \right) h^3_{n-i} \dddot{\mathbf{v}} - s(\sigma_{n-i}, \sigma_{n-i-1}) h^3_{n-i} \widehat{\mathbf{v}}\ddot{\mathbf{v}} \right) \right] (t_{n-i}) \\
&= \frac{h^2_{n+1-i}}{24} \left(\frac{1 + \sigma_{n-i}}{\sigma^3_{n+1-i}\sigma^2_{n-i}\sigma_{n-i-1}} - \frac{1 + \sigma_{n+1-i}}{\sigma^2_{n+1-i}\sigma_{n-i}} \right) \dddot{\mathbf{v}}(t_n) \\
&\quad + h^2_{n+1-i} \left(s(\sigma_{n+1-i}, \sigma_{n-i}) - \frac{s(\sigma_{n-i}, \sigma_{n-i-1})}{\sigma^2_{n+1-i}} \right) \widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) + \mathcal{O}(h^3_n)
\end{aligned}$$

mit

$$s(\sigma_n, \sigma_{n-1}) \coloneqq \frac{-1 + \sigma_n(-1 + \sigma_{n-1}(1 + \sigma_{n-1} + \sigma_n(2 + (4 + 3\sigma_n)\sigma_{n-1})))}{24\sigma^2_n\sigma_{n-1}(1 + \sigma_{n-1} + \sigma_n\sigma_{n-1})} \tag{5.32}$$

und $\mathbf{v}^{(j)}(t_{n-i}) = \mathbf{v}^{(j)}(t_n) + \mathcal{O}(h_n)$. Daher wird

$$\begin{aligned}
k = 2: \quad \mathbf{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i} &= \frac{h^2_{n+1-i}}{6} \left(\frac{1 + \sigma_{n-i}}{\sigma^2_{n+1-i}\sigma_{n-i}} - \frac{1 + \sigma_{n+1-i}}{\sigma_{n+1-i}} \right) \ddot{\mathbf{v}}_n \\
&\quad + \frac{h^2_{n+1-i}}{12} \left(\frac{2\sigma_{n+1-i} - 1}{\sigma_{n+1-i}} - \frac{2\sigma_{n-i} - 1}{\sigma_{n-i}\sigma_{n+1-i}} \right) \widehat{\mathbf{v}}_n\dot{\mathbf{v}}_n, \qquad (5.33a) \\
k = 3: \quad \mathbf{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i} &= \frac{h^3_{n+1-i}}{24} \left(\frac{1 + \sigma_{n-i}}{\sigma^3_{n+1-i}\sigma^2_{n-i}\sigma_{n-i-1}} - \frac{1 + \sigma_{n+1-i}}{\sigma^2_{n+1-i}\sigma_{n-i}} \right) \dddot{\mathbf{v}}_n \\
&\quad + h^3_{n+1-i} \left(s(\sigma_{n+1-i}, \sigma_{n-i}) - \frac{s(\sigma_{n-i}, \sigma_{n-i-1})}{\sigma^2_{n+1-i}} \right) \widehat{\mathbf{v}}_n\ddot{\mathbf{v}}_n \qquad (5.33b)
\end{aligned}$$

definiert. Wie in Abschnitt 4.1 beim Generalized-α-Verfahren (4.11) könnten allgemeine Approximationen für $\mathbf{v}_n^{(i)} \approx \mathbf{v}^{(i)}(t_n) + \mathcal{O}(h_n)$, $(i = 1, 2, 3)$, gewählt werden ($\mathbf{v}_n$ ergibt sich aus dem zuvor berechneten Zeitschritt). Da an dieser Stelle aber nur ein Einblick in die Erweiterung auf variable Schrittweiten gegeben werden soll, wird auf eine allgemeine Darstellung verzichtet und ein spezieller Fall untersucht. Es werden Approximationen

$$\dot{\mathbf{v}}_n = \frac{\sigma_n \mathbf{v}_n - \sigma_n \mathbf{v}_{n-1}}{h_n}, \tag{5.34a}$$

$$\ddot{\mathbf{v}}_n = \frac{2\sigma_n^2 \sigma_{n-1}}{(1+\sigma_{n-1})} \frac{(\mathbf{v}_n - (1+\sigma_{n-1})\mathbf{v}_{n-1} + \sigma_{n-1}\mathbf{v}_{n-2})}{h_n^2}, \tag{5.34b}$$

$$\begin{aligned}\dddot{\mathbf{v}}_n = \frac{6\sigma_n^3 \sigma_{n-1}^2 \sigma_{n-2}}{h_n^3} \Bigg(& \frac{1}{(1+\sigma_{n-1})(1+\sigma_{n-2}+\sigma_{n-1}\sigma_{n-2})} \mathbf{v}_n - \frac{1}{(1+\sigma_{n-2})} \mathbf{v}_{n-1} \\ & + \frac{\sigma_{n-1}}{(1+\sigma_{n-1})} \mathbf{v}_{n-2} - \frac{\sigma_{n-1}\sigma_{n-2}^2}{(1+\sigma_{n-2})(1+\sigma_{n-2}+\sigma_{n-2}\sigma_{n-1})} \mathbf{v}_{n-3} \Bigg)\end{aligned} \tag{5.34c}$$

verwendet mit $\mathbf{v}_{n-i} \approx \mathbf{v}(t_{n-i}) + \mathcal{O}(h_n^k)$.
Durch die Lösung des Gleichungssystems (5.31) folgt

$$\boldsymbol{\Delta}_{n,i}^{\mathbf{v}} = -\left[\mathbf{M}^{-1}\mathbf{B}^\top (\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}\right](q_{n+1-i}) \frac{\gamma_{i,n}}{\alpha_{i,n}} \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}$$

für $i = 1, ..., k$.
Für $k = 1$ kann die Konvergenz wie für konstante Schrittweiten (Bemerkung 20) nicht bewiesen werden, daher wird dieser Fall nicht näher untersucht.

Die BLieDF-Verfahren können nun unter Berücksichtigung der Parameter (5.28) und Bemerkung 23 auf variable Schrittweiten $h_n \coloneqq t_{n+1} - t_n$ erweitert werden.

Definition 25 (BLieDF-Verfahren für variable Schrittweiten zur Lösung von (3.44))
Die *k-Schritt BLieDF-Verfahren mit variablen Schrittweiten* zur Lösung von (3.44) verwenden im Zeitschritt $t_n \to t_{n+1} \coloneqq t_n + h_n$ mit Schrittweite h_n die Lösungen q_n, $\boldsymbol{\omega}_0^{(n+1-i)}$, $(i = 2, ..., k)$, $\mathbf{v}_{n+1-i}$ und $\boldsymbol{\Delta}_{n+1-i}^{\mathbf{v}}$, $(i = 1, ..., k)$, zur Berechnung von q_{n+1}, $\boldsymbol{\omega}_0^{(n)}$, $\mathbf{v}_{n+1}$ und $\boldsymbol{\lambda}_{n+1}$ anhand von

$$q_{n+1} = q_n \circ \exp\left(\widetilde{\boldsymbol{\omega}}_0^{(n)}\right), \tag{5.35a}$$

$$\frac{1}{h_n} \sum_{i=1}^{k} \gamma_{i,n} \boldsymbol{\omega}_0^{(n+1-i)} = \mathbf{v}_{n+1} + \mathbf{L}_{h_n,n}^{(k)}, \tag{5.35b}$$

$$\frac{1}{h_n} \mathbf{M}(q_{n+1}) \left(\alpha_{0,n} \mathbf{v}_{n+1} + \sum_{i=1}^{k} \alpha_{i,n} \overline{\mathbf{v}}_{n+1-i} \right) = -\mathbf{g}(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}) - \mathbf{B}^\top(q_{n+1}) \boldsymbol{\lambda}_{n+1}, \tag{5.35c}$$

$$\mathbf{0} = \boldsymbol{\Phi}(q_{n+1}) \tag{5.35d}$$

mit $\overline{\mathbf{v}}_{n+1-i}$, $(i = 1, ..., k)$, aus (5.30), Parametern (5.28) und (5.29) und einem Korrekturterm $\mathbf{L}_{h_n,n}^{(k)}$, der durch

$$\mathbf{L}_{h_n,n}^{(1)} = \mathbf{L}_{h_n,n}^{(2)} = \mathbf{0}, \tag{5.36a}$$

$$\mathbf{L}_{h_n,n}^{(3)} = -\frac{h_n}{4} \left(\sigma_n - 1 + \frac{1}{1 + \sigma_{n-1} + \sigma_n \sigma_{n-1}} \right) \widehat{\mathbf{v}}_n \mathbf{v}_{n-1} \tag{5.36b}$$

gegeben ist.

Der Korrekturterm (5.36) wird analog zu Lemma 5 für den konstanten Fall motiviert und soll die gleichen Voraussetzungen erfüllen. Für $k = 3$ wird dabei zunächst von

$$\mathbf{L}^{(3)}_{h_n,n} = \frac{h_n^2}{4}\left(\sigma_n - 1 + \frac{1}{1+\sigma_{n-1}+\sigma_n\sigma_{n-1}}\right)\widehat{\mathbf{v}}_n\dot{\mathbf{v}}_n$$

ausgegangen. Unter Verwendung von $\dot{\mathbf{v}}_n = (\mathbf{v}_n - \mathbf{v}_{n-1})/h_n$ und $\widehat{\mathbf{v}}_n\mathbf{v}_n = \mathbf{0}$ folgt die Darstellung (5.36b). Die folgende Fehleranalyse, im Speziellen die lokale Abbruchfehleranalyse, wird beweisen, dass der hier angegebene Korrekturterm seinen Zweck erfüllt.

Bemerkung 24 (Startwerte und erste Zeitschritte der BLieDF-Verfahren (5.35))
Die Startwerte der BLieDF-Verfahren (5.35) sollen äquidistant gewählt werden und sind damit analog zu Abschnitt 3.3.4 zu wählen. Für die ersten Zeitschritte wird bei der Anpassung aus Bemerkung 23 daher

$$\boldsymbol{\Delta}^{\mathbf{v}}_{n,i} \coloneqq \boldsymbol{\Delta}^{\mathbf{v}}_{k,i}$$

für $0 \leq n < k$ gesetzt.

Mit dieser Modifikation der Verfahren (5.35) im Vergleich zu (3.62) kann die Konvergenz der BLieDF-Verfahren für variable Schrittweiten bewiesen werden.

5.2.2 Lokale Abbruchfehler

Die lokale Fehleranalyse wird als erstes durchgeführt. Dazu werden die lokalen Abbruchfehler wie in Definition A.3 definiert und durch die Taylorentwicklung berechnet.

Definition 26 (Lokale Abbruchfehler von (5.35))
Die lokalen Abbruchfehler der BLieDF-Verfahren (5.35) sind gegeben durch

$$q(t_{n+1}) = q(t_n) \circ \exp\big(\widetilde{\boldsymbol{\nu}}_n(t_{n+1})\big), \tag{5.37a}$$

$$\frac{1}{h_n}\sum_{i=1}^{k}\gamma_{i,n}\boldsymbol{\nu}_{n+1-i}(t_{n+2-i}) = \mathbf{v}(t_{n+1}) + \mathbf{L}^{(k)}_{h_n}(t_n) + \mathbf{l}^{\boldsymbol{\omega}}_n, \tag{5.37b}$$

$$\begin{aligned}\mathbf{M}\big(q(t_{n+1})\big)\frac{\alpha_{0,n}}{h_n}\mathbf{v}(t_{n+1}) = &-\mathbf{M}\big(q(t_{n+1})\big)\frac{1}{h_n}\sum_{i=1}^{k}\alpha_{i,n}\overline{\mathbf{v}}(t_{n+1-i}) - \mathbf{B}^{\top}\big(q(t_{n+1})\big)\boldsymbol{\lambda}(t_{n+1})\\ &- \mathbf{g}\big(t_{n+1}, q(t_{n+1}), \mathbf{v}(t_{n+1})\big) + \mathbf{l}^{\mathbf{v}}_n,\end{aligned} \tag{5.37c}$$

$$\overline{\mathbf{v}}(t_{n+1-i}) = \mathbf{v}(t_{n+1-i}) - \mathbf{C}\big(q(t_{n+1-i})\big)\frac{\gamma_{i,n}}{\alpha_{i,n}}\boldsymbol{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}(t_n), \quad (i = 1, ..., k), \tag{5.37d}$$

$$\mathbf{0} = \boldsymbol{\Phi}\big(q(t_{n+1})\big). \tag{5.37e}$$

mit

$$\begin{aligned}\boldsymbol{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}(t_n) = &\frac{h^2_{n+1-i}}{6}\left(\frac{1+\sigma_{n-i}}{\sigma^2_{n+1-i}\sigma_{n-i}} - \frac{1+\sigma_{n+1-i}}{\sigma_{n+1-i}}\right)\ddot{\mathbf{v}}(t_n)\\ &+ \frac{h^2_{n+1-i}}{12}\left(\frac{2\sigma_{n+1-i}-1}{\sigma_{n+1-i}} - \frac{2\sigma_{n-i}-1}{\sigma_{n-i}\sigma_{n+1-i}}\right)\widehat{\dot{\mathbf{v}}}(t_n)\dot{\mathbf{v}}(t_n)\end{aligned} \tag{5.38a}$$

für $k = 2$ und

$$\begin{aligned}\boldsymbol{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}(t_n) = \frac{h^3_{n+1-i}}{24}\left(\frac{1+\sigma_{n-i}}{\sigma^3_{n+1-i}\sigma^2_{n-i}\sigma_{n-i-1}} - \frac{1+\sigma_{n+1-i}}{\sigma^2_{n+1-i}\sigma_{n-i}}\right)\dddot{\mathbf{v}}(t_n)\\ + h^3_{n+1-i}\left(s(\sigma_{n+1-i},\sigma_{n-i}) - \frac{s(\sigma_{n-i},\sigma_{n-i-1})}{\sigma^2_{n+1-i}}\right)\widehat{\mathbf{v}}(t_n)\dot{\mathbf{v}}(t_n) \qquad (5.38b)\end{aligned}$$

für $k = 3$ mit $s(\sigma_n, \sigma_{n-1})$ aus (5.32).

Wie zuvor im Generalized-α-Verfahren ist im Unterschied zum konstanten Fall in Abschnitt 5.1 nicht allein die Differenz $\|(\mathbf{l}^q_{n+1} - \mathbf{l}^q_n)/h\|$ von Interesse. Der lokale Abbruchfehler $\mathbf{l}^{\mathbf{v}}_n$ muss miteinbezogen werden. Aufgrund der Mehrschrittstruktur wird dazu der Term

$$\left\|\mathbf{l}^{\mathbf{BM}^{-1}\mathbf{v}}_n + \frac{1}{h_n}\mathbf{B}(q(t_n))\sum_{i=0}^{k}\alpha_{i,n}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}\right\|$$

untersucht.

Satz 14
Die lokalen Abbruchfehler der BLieDF-Verfahren (5.35) erfüllen für $1 \le k \le 3$ die Abschätzungen

$$\|\mathbf{l}^{\boldsymbol{\omega}}_n\| = \mathcal{O}(h^k_n), \quad \|\mathbf{l}^{\mathbf{v}}_n\| = \mathcal{O}(h^{k-1}_n), \quad \|\mathbf{l}^{\mathbf{PM}^{-1}\mathbf{v}}_n\| = \mathcal{O}(h^k_n),$$

und

$$\left\|\mathbf{l}^{\mathbf{BM}^{-1}\mathbf{v}}_n + \frac{1}{h_n}\mathbf{B}\big(q(t_n)\big)\sum_{i=0}^{k}\alpha_{i,n}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}\right\| = \mathcal{O}(h^k_n).$$

Beweis:

a) Der lokale Abbruchfehler $\mathbf{l}^{\boldsymbol{\omega}}_n$ wird anhand von Gleichung (5.37b) unter Verwendung der Taylor- und Magnusentwicklung (2.33) abgeschätzt. Die führenden Fehlerterme $\bar{\mathbf{l}}^{\boldsymbol{\omega}}_n$ der lokalen Abbruchfehler sind gegeben durch

$$\begin{aligned}k = 1 &: \bar{\mathbf{l}}^{\boldsymbol{\omega}}_n = -\frac{1}{2}h_n\dot{\mathbf{v}}(t_n),\\ k = 2 &: \bar{\mathbf{l}}^{\boldsymbol{\omega}}_n = \frac{1}{12\sigma_n}h^2_n\left((-1+2\sigma_n)\left[\mathbf{v}(t),\dot{\mathbf{v}}(t)\right] - 2(1+\sigma_n)\ddot{\mathbf{v}}(t)\right),\\ k = 3 &: \bar{\mathbf{l}}^{\boldsymbol{\omega}}_n = h^3_n/\left(24\sigma^2_n\sigma_{n+1}(1+\sigma_{n+1}+\sigma_n\sigma_{n+1})\right)\cdot(-(1+\sigma_n)\cdot\\ &\qquad \cdot(1+\sigma_{n+1}+\sigma_n\sigma_{n+1})^2\dddot{\mathbf{v}}(t) + (-1+\sigma_n(-1+\sigma_{n+1}(1+\sigma_{n+1}\\ &\qquad +\sigma_n(2+(4+3\sigma_n)\sigma_{n+1}))))\left[\mathbf{v}(t),\ddot{\mathbf{v}}(t)\right])\end{aligned}$$

und die Behauptung ist damit bestätigt.

b) Gleichung (5.37c) ist in linearen Räumen definiert. Deshalb wäre $\mathbf{l}^{\mathbf{v}}_n = \mathcal{O}(h^k_n)$ erfüllt, wenn $\overline{\mathbf{v}}(t_{n+1-i}) = \mathbf{v}(t_{n+1-i})$, $(i = 0, ..., k)$, gelten würde. Durch die Korrektur mit $\boldsymbol{\Delta}^{\mathbf{v}}_{n,i}$ entsteht ein neuer führender Fehlerterm, der aus genau dieser Korrektur besteht. Daher ist

$$\mathbf{l}^{\mathbf{M}^{-1}\mathbf{v}}_n = -\frac{1}{h_n}\sum_{i=1}^{k}\gamma_{i,n}\mathbf{C}\big(q(t_{n+1-i})\big)\boldsymbol{\Delta}\mathbf{l}^{\boldsymbol{\omega}}_{n-i}(t_n) = \mathcal{O}(h^{k-1}_n),$$

vgl. (5.38). Wird diese Gleichung mit $\mathbf{P}(q(t_{n+1}))$ multipliziert, so folgt wegen $[\mathbf{PC}](q) \equiv \mathbf{0}$ direkt $\|\mathbf{l}^{\mathbf{PM}^{-1}\mathbf{v}}_n\| = \mathcal{O}(h^k_n)$.

c) Mit Teil b), $\mathbf{B}\big(q(t_n)\big)\mathbf{C}\big(q(t_n)\big) = \mathbf{B}\big(q(t_n)\big)$ und $\mathbf{B}\big(q(t_{n+1-i})\big) = \mathbf{B}\big(q(t_n)\big) + \mathcal{O}(h_n)$ ist

$$\mathbf{l}_n^{\mathbf{BM}^{-1}\mathbf{v}} = -\frac{1}{h_n}\mathbf{B}\big(q(t_n)\big)\sum_{i=1}^{k}\gamma_{i,n}\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n).$$

Weiterhin gilt

$$\frac{1}{h_n}\mathbf{B}\big(q(t_n)\big)\sum_{i=0}^{k}\alpha_{i,n}\mathbf{l}_{n-i}^{\boldsymbol{\omega}} = \frac{1}{h_n}\mathbf{B}\big(q(t_n)\big)\sum_{i=1}^{k}\gamma_{i,n}(\mathbf{l}_{n+1-i}^{\boldsymbol{\omega}} - \mathbf{l}_{n-i}^{\boldsymbol{\omega}}).$$

Da $\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) \approx \mathbf{l}_{n+1-i}^{\boldsymbol{\omega}} - \mathbf{l}_{n-i}^{\boldsymbol{\omega}}$ ist, folgt die Behauptung.

■

5.2.3 Stabilität und Konvergenz

Für variable Schrittweiten können die globalen Fehler analog zu (4.19a)-(4.19c) und (5.5) für $i = 0$ definiert werden. Außerdem werden in allen nachfolgenden Sätzen und Lemmata die Voraussetzungen 1 und 5 (hier jedoch mit $h := h_0$) sowie die folgenden Voraussetzungen analog zu den Voraussetzungen 3 und 4 angenommen.

Voraussetzung 6
Es gibt positive Konstanten $\overline{h}$ und C_T, so dass für alle $h_n \in (0, \overline{h}]$ und $t_0 + \sum_{i=0}^{r} h_i \in [t_0, t_{\text{end}}]$ mit $n, r \in \mathbb{N}$ die globalen Fehler durch

$$\|\mathbf{e}_r^q\| + \|\mathbf{e}_{r,0}^{\boldsymbol{\omega}}\| \leq C_T h_r, \quad \|\mathbf{e}_r^{\mathbf{v}}\| + \|\mathbf{e}_r^{\boldsymbol{\lambda}}\| \leq C_T$$

beschränkt bleiben.

Voraussetzung 7
Die Startwerte $q_0, ..., q_{k-1}, \boldsymbol{\omega}_0^{(0)}, ..., \boldsymbol{\omega}_0^{(k-2)}, \mathbf{v}_0, ..., \mathbf{v}_{k-1}$ und $\boldsymbol{\lambda}_0, ..., \boldsymbol{\lambda}_{k-1}$ zur Lösung des Anfangswertproblems (3.44) mit $q(t_0) = q_0$ und $\mathbf{v}(t_0) = \mathbf{v}_0$ sollen die Bedingungen

$$\sum_{i=0}^{k-1}\|\mathbf{e}_i^q\| + \|\mathbf{e}_i^{\mathbf{Bv}} + \mathbf{B}(q(t_{k-1}))\mathbf{l}_{k-1}^{\boldsymbol{\omega}}\| = \mathcal{O}(h_0^{k+1}), \quad \sum_{i=0}^{k-2}\|\mathbf{e}_i^{\boldsymbol{\omega}}\| = \mathcal{O}(h_0^{k+1}),$$

$$\sum_{i=0}^{k-1}\|\mathbf{e}_i^{\mathbf{v}}\| + \|\mathbf{e}_i^{\mathbf{Pv}}\| + \|\mathbf{e}_i^{\boldsymbol{\lambda}}\| = \mathcal{O}(h_0^k), \quad \max_{0\leq i\leq k-1}\|\boldsymbol{\Phi}(q_i)\| = \mathcal{O}(h_0^{k+2}),$$

erfüllen.

Um die Terme höherer Ordnung zusammenzufassen, wird

$$\epsilon_n = \|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| + h_n\|\mathbf{e}_n^{\boldsymbol{\lambda}}\|$$

gewählt.
Der Konvergenzbeweis kann analog zum Fall konstanter Schrittweiten geführt werden. Innerhalb der Beweise muss h durch h_n ersetzt werden und beachtet werden, dass die Parameter $\alpha_{i,n}$ und $\gamma_{i,n}$ vom aktuellen Zeitschritt abhängen. Weiterhin wird der Zusammenhang

$$\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{i,n}\mathbf{e}_{n+1-i}^{(\bullet)} = \frac{1}{h_n}\sum_{i=1}^{k}\gamma_{i,n}\big(\mathbf{e}_{n+2-i}^{(\bullet)} - \mathbf{e}_{n+1-i}^{(\bullet)}\big) \tag{5.39}$$

benötigt, um zwischen den beiden Parametervarianten zu wechseln. In manchen Beweisteilen ist es von Bedeutung, dass die Schrittweite im Nenner zu den globalen Fehlern im Zähler passt. Dazu kann (5.39) zu

$$\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{i,n}\mathbf{e}_{n+1-i}^{(\bullet)} = \sum_{i=1}^{k}\gamma_{i,n}\left(\prod_{j=0}^{i-2}\frac{1}{\sigma_{n-j}}\right)\frac{\mathbf{e}_{n+2-i}^{(\bullet)} - \mathbf{e}_{n+1-i}^{(\bullet)}}{h_{n+1-i}} \tag{5.40}$$

umgeschrieben werden mit

$$\frac{h_{n+1-i}}{h_n} = \prod_{j=0}^{i-2}\frac{1}{\sigma_{n-j}}. \tag{5.41}$$

Außerdem muss die Änderung von $\mathbf{v}_{n+1-i}$ in $\overline{\mathbf{v}}_{n+1-i}$, $(i = 1, ..., k)$, beachtet werden. Dazu wird analog zum Generalized-α-Verfahren der Fehler abgeschätzt, der durch diese Änderung entsteht (vgl. Lemma E.1). Da die Berechnung der Gleichungen für den Beweis der Konvergenz der BLieDF-Verfahren (5.35) bis auf die genannten Änderungen keine neuen Beweisideen enthalten, wird an dieser Stelle auf eine ausführliche Erläuterung verzichtet. Genauere Angaben, wie die gekoppelte Fehlerrekursion (4.42) für die BLieDF-Verfahren (5.35) aufgestellt wird, sind jedoch im Anhang E zu finden.
Für die Konvergenz der BLieDF-Verfahren (5.35) muss die Nullstabilität gegeben sein. In [10, 12] wurden Schranken für die Schrittweitenverhältnisse σ_n berechnet, so dass die Stabilität garantiert werden kann. Dabei wurde untersucht, für welche Schrittweitenverhältnisse σ_n es eine von n unabhängige Norm $\|.\|$ gibt, so dass die Norm der Produkte

$$\|\mathbf{A}_{n+j}\cdot\mathbf{A}_{n+j-1}\cdot ...\cdot\mathbf{A}_n\| \leq C_A$$

beschränkt bleibt mit

$$\mathbf{A}_n := -\frac{1}{\alpha_{0,n}}\mathbf{e}_{1,k}\cdot(\alpha_{1,n}, ..., \alpha_{k,n}) + \mathbf{J}_k$$

für alle n und $j \geq 0$ sowie einer Konstanten $C_A > 0$.
In dem Konvergenzbeweis für DAEs wird jedoch eine von σ_n unabhängige Norm $\|.\|_{\mathbf{y}}$ gesucht, so dass jede einzelne Matrix die Bedingung

$$\|\mathbf{T}_{\mathbf{y},n}\|_{\mathbf{y}} = 1 \tag{5.42}$$

für alle $n \in \mathbb{N}$ erfüllt. Die Resultate aus [10, 12] lassen sich somit nicht direkt auf den DAE-Fall übertragen und es müsste untersucht werden, für welche σ_n die Bedingung (5.42) erfüllt ist. Da in der vorliegenden Arbeit nur ein Einblick in die Erweiterung des BLieDF-Verfahrens auf variable Schrittweiten gegeben werden soll, wird an dieser Stelle auf die Untersuchung verzichtet, wann (5.42) erfüllt ist. Dies könnte ein Thema für nachfolgende Arbeiten sein. Trotzdem wird davon ausgegangen, dass es $\sigma_{\min}$ und $\sigma_{\max}$ gibt, so dass (5.42) gilt. Daher wird die nachfolgende Voraussetzung getroffen.

Voraussetzung 8
Es gibt zwei von n unabhängige Normen $\|.\|_{\mathbf{y}}$ und $\|.\|_{\mathbf{z}}$ sowie zwei positive Konstanten $\sigma_{\min} \leq 1$ und $\sigma_{\max} \geq 1$, so dass für alle $n \in \mathbb{N}$ die Bedingungen

$$\|\mathbf{T}_{\mathbf{y},n}\|_{\mathbf{y}} = 1 \text{ und } \|\mathbf{T}_{\mathbf{z},n}\|_{\mathbf{z}} < 1$$

mit $\sigma_{\min} \leq \sigma_n = h_n/h_{n-1} \leq \sigma_{\max}$ erfüllt sind.

Die numerischen Tests aus Kapitel 6 legen nahe, dass diese Voraussetzung 8 für $k = 2$ und $k = 3$ erfüllbar ist. Nun kann die Konvergenz der BLieDF-Verfahren (5.35) für das gesamte Zeitintervall bewiesen werden.

Satz 15
Unter den Voraussetzungen 1, 5, 6, 7 und 8 gibt es positive Konstanten C_0, $\overline{L}$ und $\overline{h}$ unabhängig von n und h_n, so dass für alle $h_n \in (0, \overline{h}]$ und alle $n \geq 0$ mit $t_0 + \sum_{i=0}^{n} h_i \leq t_{\text{end}}$ die globalen Fehlerabschätzungen der BLieDF-Verfahren (5.35) die Bedingungen

$$\|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| + \|\mathbf{e}_n^{\boldsymbol{\lambda}}\| \leq C_0 e^{\overline{L}(t_n - t_0)} h_{\max}^k$$

erfüllen und die k-Schritt-BLieDF-Integratoren (5.35) die Konvergenzordnung $p = k$ für $2 \leq k \leq 3$ besitzen.

Beweis:
Mit Lemma E.4 ist für die Verfahren (5.35) die gekoppelte Fehlerrekursion (4.42) erfüllt. Mit Voraussetzung 8 gibt es eine Norm $\|.\|_{\mathbf{y}}$ mit $\|\mathbf{T}_{\mathbf{y},n}\|_{\mathbf{y}} = 1$ und eine Norm $\|.\|_{\mathbf{z}}$ mit $\|\mathbf{T}_{\mathbf{z},n}\|_{\mathbf{z}} < 1$. Mit Satz 4 können die Schätzungen (4.49) erhalten werden. Mit den gegebenen Voraussetzungen folgen $\|\mathbf{E}_{k-1}^{\mathbf{y}}\| = \mathcal{O}(h_0^k)$ und $\|\mathbf{E}_{k-1}^{\mathbf{z}}\| = \mathcal{O}(h_0^k)$ und daher die Behauptung. ■

Mit dem vorhergehenden Satz konnte die Konvergenz der BLieDF-Verfahren (5.35) für variable Schrittweiten mit der Ordnung $p = k$ für $2 \leq k \leq 3$ bewiesen werden. Werden die Koeffizienten $\alpha_{i,n}$ und der Korrekturterm $\mathbf{L}_{h_n,n}^{(k)}$ für $4 \leq k \leq 6$ analog zu den vorgestellten Fällen berechnet, so kann die Konvergenz in gleicher Vorgehensweise auch für $4 \leq k \leq 6$ bewiesen werden, wenn die Schrittweitenverhältnisse den Bedingungen der BDF-Verfahren in linearen Räumen (vgl. [46]) genügen. Für $k = 1$ ist es, wie im konstanten Fall (Bemerkung 20), mit den vorgestellten Beweismitteln formal nicht möglich, die Konvergenz theoretisch zu beweisen.

Kapitel 6

Implementierung und numerische Tests

Die numerischen Tests sollen die theoretischen Resultate verifizieren. Dabei werden die Verfahren zunächst für konstante Schrittweiten untersucht. Anschließend werden das Generalized-α-Verfahren (4.11) und die BLieDF-Verfahren (5.35) für variable Schrittweiten getestet. Zunächst sollen jedoch allgemeine Implementierungsaspekte vorgestellt werden.

6.1 Allgemeine Implementierungsaspekte

In diesem Abschnitt sollen allgemeine Aspekte für die Implementierung in MATLAB geklärt werden.

Bemerkung 25 (Referenzlösung)
Wenn innerhalb der numerischen Tests von Referenzlösung q_{ref}, $\mathbf{v}_{\text{ref}}$, $\dot{\mathbf{v}}_{\text{ref}}$ oder $\boldsymbol{\lambda}_{\text{ref}}$ die Rede ist, dann wurde eine Lösung des Benchmarkproblems für die entsprechenden Variablen mit Hilfe der in MATLAB integrierten Funktion ode15s mit sehr kleinen Toleranzen ATOL $= 1 \cdot 10^{-16}$ und RTOL $= 2.22045 \cdot 10^{-14}$ berechnet.

Bemerkung 26 (Arten von Fehlern)
In den Tests werden zwei Arten von Fehlern in den Variablen untersucht. Dies sind zum einen der absolute Fehler im Endzeitpunkt und zum anderen der maximale absolute Fehler.
Sei o.B.d.A. die untersuchte Variable die Geschwindigkeit $\mathbf{v}$. Dann wird zunächst eine Referenzlösung $\mathbf{v}_{\text{ref},n}$ und eine numerische Lösung $\mathbf{v}_n$ in allen Zeitpunkten $t_n := t_0 + nh$ bestimmt. Der absolute Fehler im Endzeitpunkt $t_{\text{end}} = t_{N_{\text{end}}}$ berechnet sich aus

$$\text{ERR}_{1,\mathbf{v}} := \|\mathbf{v}_{\text{ref},N_{\text{end}}} - \mathbf{v}_{N_{\text{end}}}\|_2$$

mit der euklidischen Norm $\|.\|_2$.
Der maximale absolute Fehler ist gegeben durch

$$\text{ERR}_{2,\mathbf{v}} := \|\max_{i=0,\ldots,N_{\text{end}}} (\mathbf{v}_{\text{ref},n} - \mathbf{v}_n)\|_2.$$

Die Fehler können ebenso für die anderen Variablen untersucht werden. Für die Konfigurationsvariable q werden die Komponenten dazu jedoch zunächst in einem Vektor

angeordnet. Das heißt

$$q = \mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \in SO(3)$$

wird zum Beispiel zunächst zu

$$q_{\text{vec}} = [R_{11}\ R_{12}\ R_{13}\ R_{21}\ R_{22}\ R_{23}\ R_{31}\ R_{32}\ R_{33}]^T$$

umgeschrieben. Die Fehler in q werden somit nicht in der Lie-Algebra gemessen, sondern in dem höherdimensionalen Raum, indem die Lie-Algebra eingebettet ist.

Bemerkung 27 (Effizienzuntersuchungen)
Um die Effizienz der Verfahren zu charakterisieren, wird wie folgt vorgegangen. Dazu wird die Rechenzeit bestimmt, die für jeweils einen Zeitschritt benötigt wird, ohne jegliche Anpassung der Startwerte, Initialisierung oder Ähnliches. Die Zeit für alle diese Zeitschritte wurde addiert und über den absoluten Fehler aufgetragen. Dabei wurden die Rechnungen über einen Rechenserver ausgeführt und stets auf den Abend gelegt, um ein aussagekräftiges Ergebnis zu erhalten. Natürlich können trotzdem andere Faktoren, wie die Art und Weise der Implementierung, dazu beitragen, dass eines der untersuchten Verfahren langsamer rechnet als die anderen – unabhängig von der eigentlichen Laufzeit des Verfahrens. Es wurde aber sorgfältig versucht, möglichst viele Programmierbausteine in allen Verfahren gleich zu verwenden.

Bemerkung 28 (Erstellung von Schrittweitenfolgen mit unterschiedlichen Schrittweiten in jedem Zeitschritt)
Das Generalized-α-Verfahren (4.11) und die BLieDF-Verfahren (5.35) sollen für variable Schrittweiten getestet werden. Dazu werden den Programmen vorgefertigte Schrittweitenfolgen $(h_n)_{n=0}^{N_{\text{end}}}$ mit festem $N_{\text{end}} \in \mathbb{N}$ übergeben. Zur Erstellung solcher Schrittweitenfolgen werden zunächst die maximale $h_{\max}$ und minimale $h_{\min}$ Schrittweite und das maximale $\sigma_{\max}$ und minimale $\sigma_{\min}$ Schrittweitenverhältnis für eine Folge festgelegt. Außerdem werden durch eine in MATLAB vorhandene Zufallsfunktion zufällig Schrittweitenverhältnisse $\sigma_n \in [\sigma_{\min}, \sigma_{\max}]$, $(i = 1, ..., N_{\text{end}})$, berechnet. Eine Schrittweitenfolge ergibt sich dann iterativ durch

$$h_{i+1} = \max\{h_{\min}, \min\{\sigma_{i+1} h_i, h_{\max}\}\}, \quad i = 0, ..., N_{\text{end}} - 1. \tag{6.1}$$

Da die Konvergenzordnung der Verfahren von Interesse ist, werden verschiedene Schrittweitenfolgen für unterschiedliche Intervalle $[h_{\min}, h_{\max}]$ benötigt. Dazu wird, nachdem eine Schrittweitenfolge bestimmt wurde, $h_{\min,\text{neu}} = h_{\min,\text{alt}} \cdot 10^{-1/10}$ und $h_{\max,\text{neu}} = h_{\max,\text{alt}} \cdot 10^{-1/10}$ gewählt. Um mit dieser Folge etwa bis zum gleichen Endzeitpunkt zu rechnen, wird zudem die Anzahl der Schrittweitenglieder N_{alt} zu $N_{\text{neu}} = N_{\text{alt}}/10^{-1/10}$ verändert. Dadurch können Schrittweitenfolgen bestimmt werden, deren durchschnittliche Schrittweite unterschiedlich ist.
Wird eine Schrittweitenfolge benötigt, deren Glieder sich für N_{konst} Anzahl von Zeitschritten nicht erneut verändern, so kann analog vorgegangen werden. Jedoch muss dann zusätzlich $\sigma_{i+1} = 1$ für alle i gefordert werden, die nicht ohne Rest durch N_{konst} teilbar sind.

6.2 Konstante Schrittweiten

In diesem Abschnitt sollen numerische Tests für die vorgestellten Verfahren für konstante Schrittweiten durchgeführt werden. Es soll überprüft werden, ob die theoretisch bestimmten Ordnungen auch numerisch zu beobachten sind und Vergleiche zwischen den Verfahren bzgl. der Genauigkeit und der Effektivität durchgeführt werden. Für die BDF-Verfahren (3.57) und (3.62) soll die Wahl der Startwerte gerechtfertigt werden. Für die BLieDF-Verfahren (3.62) wird die Wahl der freien Parameter im Korrekturterm $\mathbf{L}_{n,h}^{(k)}$ überprüft.

6.2.1 Wahl der Startwerte in den BDF-Verfahren

In diesem Abschnitt soll gezeigt werden, dass die Wahl der korrigierten Startwerte aus den Abschnitten 3.3.3 (Munthe-Kaas-BDF-Verfahren) und 3.3.4 (BLieDF-Verfahren) gerechtfertigt ist. Diese wurden benötigt, um die Bedingung (5.23) einzuhalten und damit einer Ordnungsreduktion vorzubeugen, welche bei einer annähernd analytischen Lösung als Startwerte zu beobachten sein sollte.
Dazu wird der schwere Kreisel in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ für $h \in [10^{-4}, 10^{-3}]$ bis zu einem Zeitpunkt $t_{\text{end}} = 1$ gelöst und der maxi-

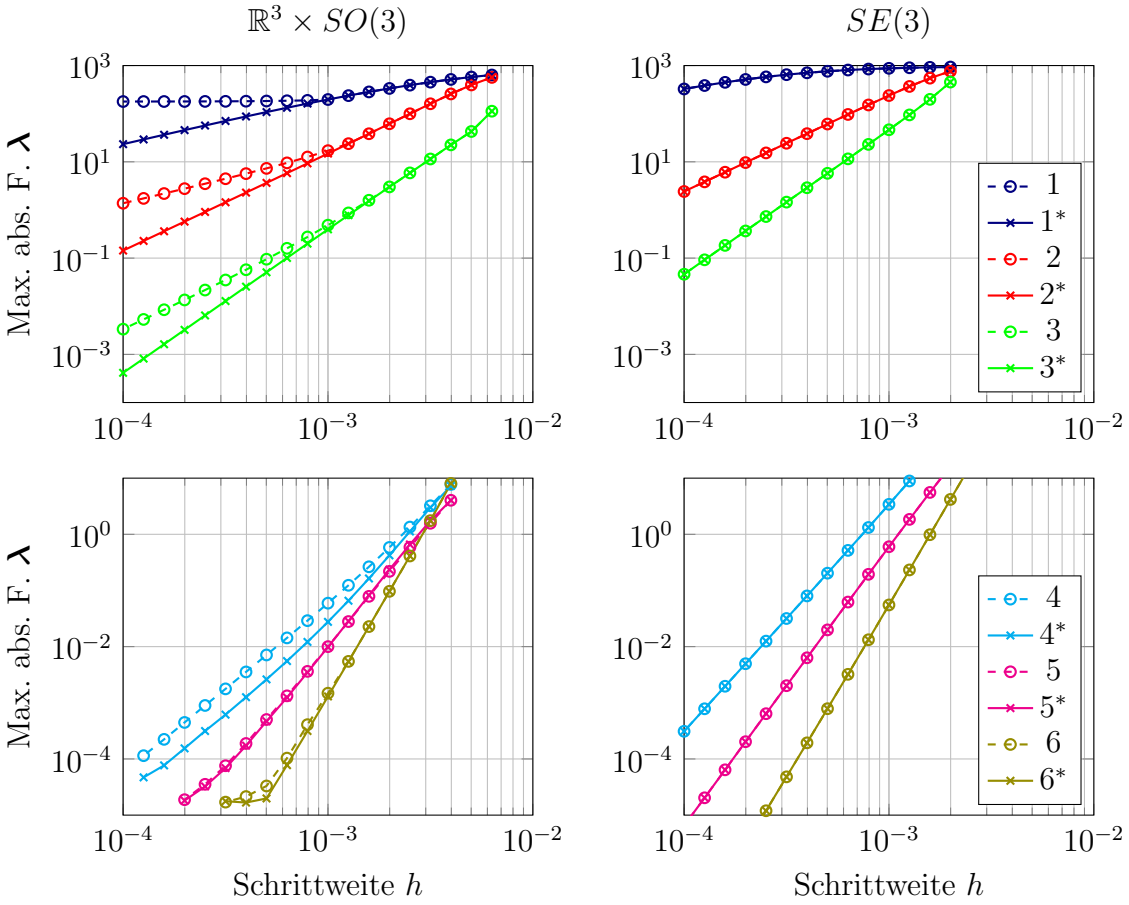

Abbildung 6.1: Berechnung des schweren Kreisels in $\mathbb{R}^3 \times SO(3)$ (links) und $SE(3)$ (rechts) mit den MKBDF-Verfahren (3.57) für $k = 1, 2, 3$ (oben) und $k = 4, 5, 6$ (unten) mit einer Initialisierung der Startwerte mit Funktionswerten der exakten Lösung (k ohne *) und korrigierten Startwerten (k mit *)

male absolute Fehler der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ berechnet. Die Ergebnisse sind in Abbildung 6.1 für die Munthe-Kaas-BDF-Verfahren (3.57) und in Abbildung 6.2 für die BLieDF-Verfahren (3.62) dargestellt. Werden die Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ (links) und eine Initialisierung der Startwerte mit Funktionswerten der exakten Lösung (ohne *) verwendet, so ist stets eine Konvergenzordnung weniger, also $p = k - 1$, im maximalen absoluten Fehler zu beobachten. Hier tritt somit in beiden BDF-Verfahren (3.57) und (3.62) wie erwartet eine Ordnungsreduktion auf. Werden hingegen die korrigierten Startwerte verwendet, so ist stets die Ordnung $p = k$ erkennbar. Mit der Wahl der angegebenen Startwerte kann die auftretende Ordnungsreduktion somit vermieden werden.
Bei der Verwendung der Lie-Gruppen-Formulierung $SE(3)$ (rechts) ist die Situation anders. Hier spielt die Wahl der Startwerte eine untergeordnete Rolle, denn sowohl für eine Initialisierung der Startwerte mit Funktionswerten der exakten Lösung als auch für die korrigierten ist die Ordnung $p = k$ zu beobachten und somit keine Ordnungsreduktion vorhanden. Dieses Verhalten hängt stark mit der konstanten Ableitungsmatrix $\mathbf{B}$ der Zwangsbedingung für $SE(3)$ und damit der Einhaltung der versteckten Zwangsbedingungen zusammen, vgl. Bemerkung 22. In diesem Fall ist daher die Bedingung (5.23) auch für Startwerte nahe der analytischen Lösung erfüllt. Im Konvergenzbeweis aus Kapitel 5 zeigte sich, dass die Bedingung (5.23) vor allem den globalen Fehler $\mathbf{e}_n^{\boldsymbol{\lambda}}$ beeinflusst, vgl. Lemma 19 und Satz 12. Aus diesem Grund

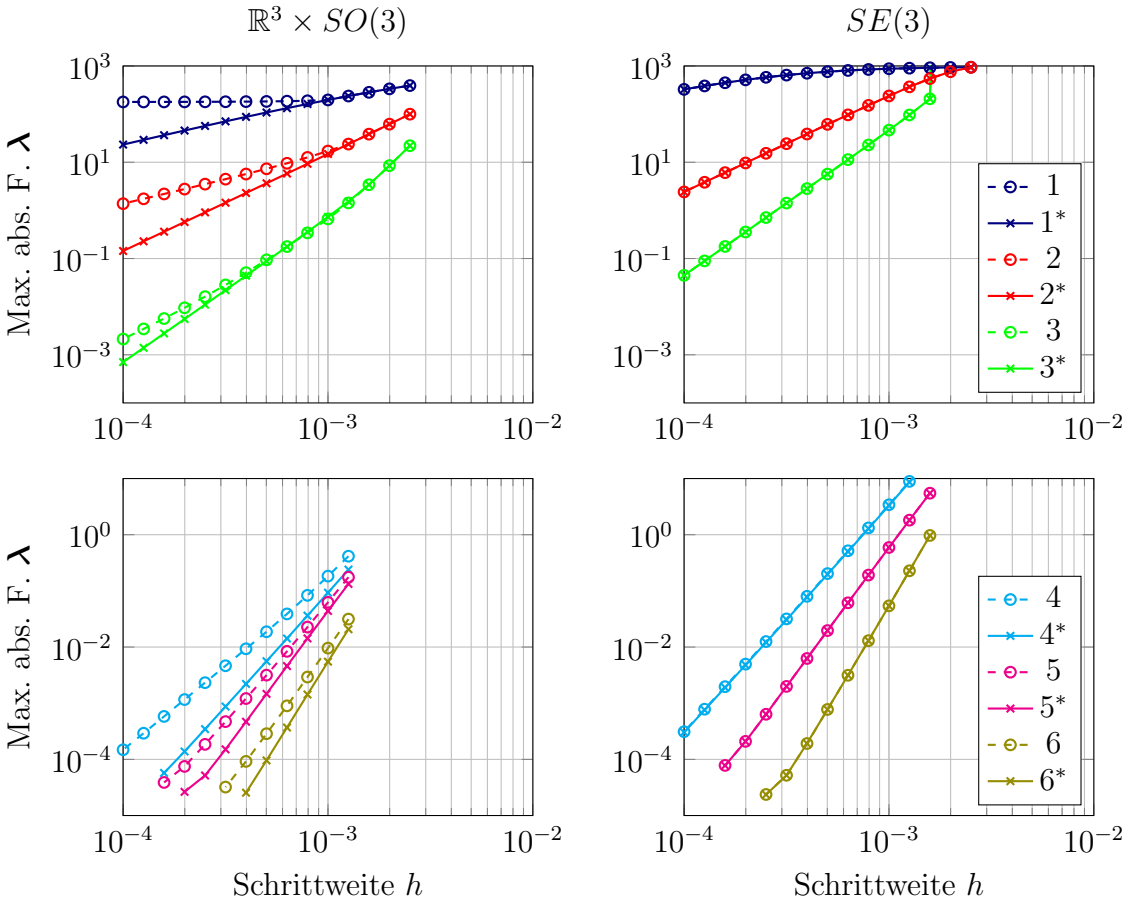

Abbildung 6.2: Berechnung des schweren Kreisels in $\mathbb{R}^3 \times SO(3)$ (links) und $SE(3)$ (rechts) mit den BLieDF-Verfahren (3.62) für $k = 1, 2, 3$ (oben) und $k = 4, 5, 6$ (unten) mit Startwerten mit Funktionswerten der analytischen Lösung (k ohne *) und korrigierten Startwerten (k mit *)

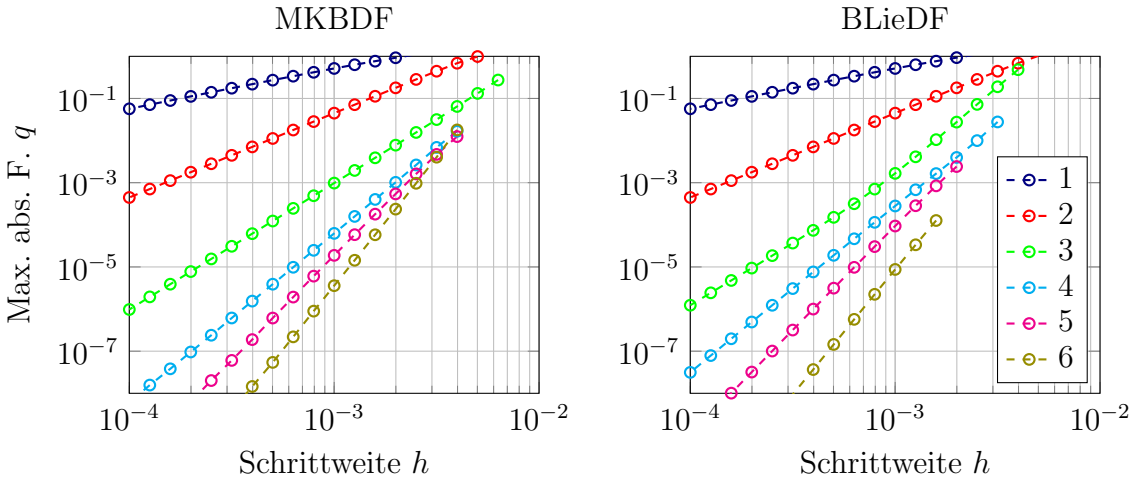

Abbildung 6.3: Maximaler absoluter Fehler in q für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit den MKBDF-Verfahren (links) und den BLieDF-Verfahren (rechts)

ist die Ordnungsreduktion in den Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ zu beobachten. Auf die anderen Variablen hat dies keinen Einfluss, wie auch in Abbildung 6.3 zu sehen ist. Hier wurde der maximale absolute Fehler in der Konfigurationsvariablen q mit einer Initialisierung der Startwerte mit Funktionswerten der exakten Lösung in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ für die Munthe-Kaas-BDF-Verfahren (3.57) (links) und die BLieDF-Verfahren (3.62) (rechts) dargestellt. Es zeigt sich, dass in dieser Variablen stets die Konvergenzordnung $p = k$ zu erkennen ist. Die Nichteinhaltung der Bedingung (5.23) hat somit keine Auswirkung auf die Konfigurationsvariable q. Für die Geschwindigkeit $\mathbf{v}$ konnte Gleiches beobachtet werden.

Fazit

Die numerischen Tests zeigen, dass die Ordnung $p = k$ auch in der Lie-Gruppen-Formulierung für differential-algebraische Gleichungen vom Index 3 in allen Komponenten erreicht wird, was die Untersuchungen aus Kapitel 5 verifiziert.
Es konnte bestätigt werden, dass eine Ordnungsreduktion in den Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ vermieden wird, wenn die korrigierten Startwerte aus den Abschnitten 3.3.3 (Munthe-Kaas-BDF-Verfahren) und 3.3.4 (BLieDF-Verfahren) anstelle der analytischen Lösung als Startwerte verwendet werden. Die Ordnungsreduktion hat keinen unmittelbaren Einfluss auf die Konvergenz der Konfigurations- oder der Geschwindigkeitsvariablen.

6.2.2 Vergleich aller untersuchten Verfahren

In diesem Abschnitt sollen alle Verfahren, die in dieser Arbeit vorgestellt wurden, miteinander verglichen werden. Dazu werden die Bewegungsgleichungen (3.20) in $SO(3)$ des schweren Kreisels aus dem Abschnitt 3.2.1 als Beispiel für ein mechanisches System ohne Zwangsbedingungen (3.17) gelöst. Es werden das Crouch-Grossman-Verfahren (3.4), das kommutatorfreie Lie-Gruppen-Verfahren (3.7), das Generalized-α-Verfahren (3.21), die Munthe-Kaas-BDF-Verfahren (3.23) und die BLieDF-Verfahren (3.31) verwendet.
Außerdem erfolgt ein weiterer Vergleich für das Generalized-α-Verfahren (3.55), die

Munthe-Kaas-BDF-Verfahren (3.57) und die BLieDF-Verfahren (3.62) zur Lösung des Benchmarks schwerer Kreisel mit den Bewegungsgleichungen (3.51) beispielhaft in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$.
Der Vergleich wird für die unterschiedlichen Konvergenzordnungen separat durchgeführt. Es wird erwartet, dass für kleine Konvergenzordnungen die Effizienz der Verfahren ähnlich ist. Mit steigendem p bzw. k unterscheiden sich die Auswertungen der Exponentialabbildungen und die Anzahl der Additionen bzw. Multiplikationen innerhalb der Verfahren.
Im Vergleich Crouch-Grossman-Verfahren (3.4) und kommutatorfreies Lie-Gruppen-Verfahren (3.7) werden in Ersterem mehr Exponentialabbildungen ausgewertet, wohingegen in Zweiterem die „fehlenden“ Exponentialabbildungen durch zusätzliche Additionen und Multiplikationen kompensiert werden. Somit sollte das kommutatorfreie Lie-Gruppen-Verfahren effizienter sein. Im Vergleich Munthe-Kaas-BDF-Verfahren (3.23) und BLieDF-Verfahren (3.31) müssen theoretisch in Ersterem pro Zeitschritt mehr Matrix-Kommutatoren ausgewertet werden, weshalb die BLieDF-Verfahren effizienter sein sollten. Wie sich das Generalized-α-Verfahren einordnet, zeigen die nachfolgenden numerischen Tests.

Vergleich für $\mathbf{p = 1}$

Die Verfahren (3.4), (3.7), (3.23) und (3.31) zur Lösung von (3.20) für $p = 1$ haben alle die Gestalt

$$\mathbf{R}_{n+1} = \mathbf{R}_n \circ \exp\left(h\widetilde{\mathbf{\Omega}}_{n+1}\right), \tag{6.2a}$$

$$\mathbf{\Omega}_{n+1} = \mathbf{\Omega}_n - h\mathbf{J}^{-1}\left(\widetilde{\mathbf{\Omega}}_{n+1}\mathbf{J}\mathbf{\Omega}_{n+1} - \widetilde{\mathbf{X}}\mathbf{R}_{n+1}^{\top} m\boldsymbol{\gamma}\right). \tag{6.2b}$$

Es handelt sich also in jedem Fall um das gleiche Verfahren. Der maximale absolute Fehler der Konfigurationsvariablen q und der Geschwindigkeit $\mathbf{v}$ ist in Abbildung 6.4 in doppelt logarithmischer Skala über die Schrittweite $h \in [10^{-5}, 10^{-4}]$ dargestellt. Dabei ist für beide Variablen die Konvergenz erster Ordnung erkennbar. Da nur ein Verfahren (6.2) betrachtet wird, ist ein Effizienzvergleich nicht sinnvoll.
Werden die Munthe-Kaas-BDF-Verfahren (3.57) und die BLieDF-Verfahren (3.62)

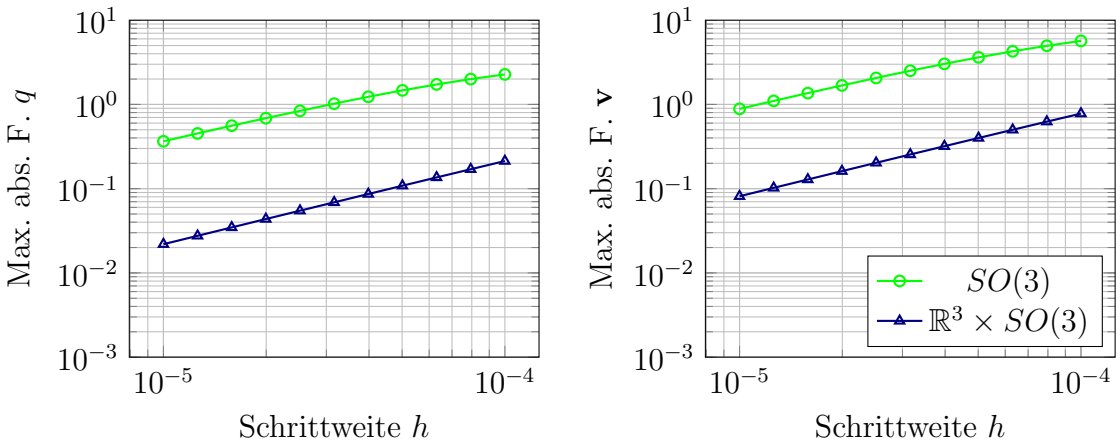

Abbildung 6.4: Maximaler absoluter Fehler von q und $\mathbf{v}$ für den schweren Kreisel in $SO(3)$ mit Bewegungsgl. (3.20) und in $\mathbb{R}^3 \times SO(3)$ mit Bewegungsgl. (3.51) für $p = 1$

zur Lösung von (3.44) untersucht, so ergibt sich das Verfahren

$$\begin{aligned} q_{n+1} &= q_n \circ \exp(h\widetilde{\mathbf{v}}_{n+1}), \\ \frac{1}{h}\mathbf{M}(q_{n+1})(\mathbf{v}_{n+1} - \mathbf{v}_n) &= -\mathbf{g}(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}) - \mathbf{B}^\top(q_{n+1})\boldsymbol{\lambda}_{n+1}, \\ \mathbf{0} &= \boldsymbol{\Phi}(q_{n+1}). \end{aligned}$$

Auch für dieses Verfahrens ist in Abbildung 6.4 die Konvergenzordnung eins in beiden Variablen erkennbar.

Vergleich für p = 2

Für $p = 2$ ist das Crouch-Grossman-Verfahren (3.4) zur Lösung von (3.20) gegeben durch

$$\mathbf{R}_{n+1} = \mathbf{R}_n \circ \exp\left(\frac{h}{2}\widetilde{\boldsymbol{\Omega}}_n\right) \exp\left(\frac{h}{2}\widetilde{\boldsymbol{\Omega}}_{n+1}\right), \tag{6.3a}$$

$$\boldsymbol{\Omega}_{n+1} = \boldsymbol{\Omega}_n - \frac{h}{2}\mathbf{J}^{-1}\left(\widetilde{\boldsymbol{\Omega}}_{n+1}\mathbf{J}\boldsymbol{\Omega}_{n+1} - \widetilde{\mathbf{X}}\mathbf{R}_{n+1}^\top m\boldsymbol{\gamma}\right) - \frac{h}{2}\mathbf{J}^{-1}\left(\widetilde{\boldsymbol{\Omega}}_n\mathbf{J}\boldsymbol{\Omega}_n - \widetilde{\mathbf{X}}\mathbf{R}_n^\top m\boldsymbol{\gamma}\right) \tag{6.3b}$$

und das kommutatorfreie Lie-Gruppen-Mehrschrittverfahren (3.7) durch

$$\mathbf{R}_{n+1} = \mathbf{R}_n \circ \exp\left(\frac{h}{2}\widetilde{\boldsymbol{\Omega}}_n + \frac{h}{2}\widetilde{\boldsymbol{\Omega}}_{n+1}\right), \tag{6.4a}$$

$$\boldsymbol{\Omega}_{n+1} = \boldsymbol{\Omega}_n - \frac{h}{2}\mathbf{J}^{-1}\left(\widetilde{\boldsymbol{\Omega}}_{n+1}\mathbf{J}\boldsymbol{\Omega}_{n+1} - \widetilde{\mathbf{X}}\mathbf{R}_{n+1}^\top m\boldsymbol{\gamma}\right) - \frac{h}{2}\mathbf{J}^{-1}\left(\widetilde{\boldsymbol{\Omega}}_n\mathbf{J}\boldsymbol{\Omega}_n - \widetilde{\mathbf{X}}\mathbf{R}_n^\top m\boldsymbol{\gamma}\right). \tag{6.4b}$$

Der Unterschied besteht somit darin, dass im Crouch-Grossman-Verfahren (6.3) zwei Exponentialabbildungen und im kommutatorfreien Verfahren (6.4) nur eine Exponentialabbildung verwendet werden, wofür aber im Argument eine Addition mehr auftaucht.
Die Munthe-Kaas-BDF-Verfahren (3.23) und die BLieDF-Verfahren (3.31) sind für $p = k = 2$ zur Lösung von (3.20) identisch gegeben durch

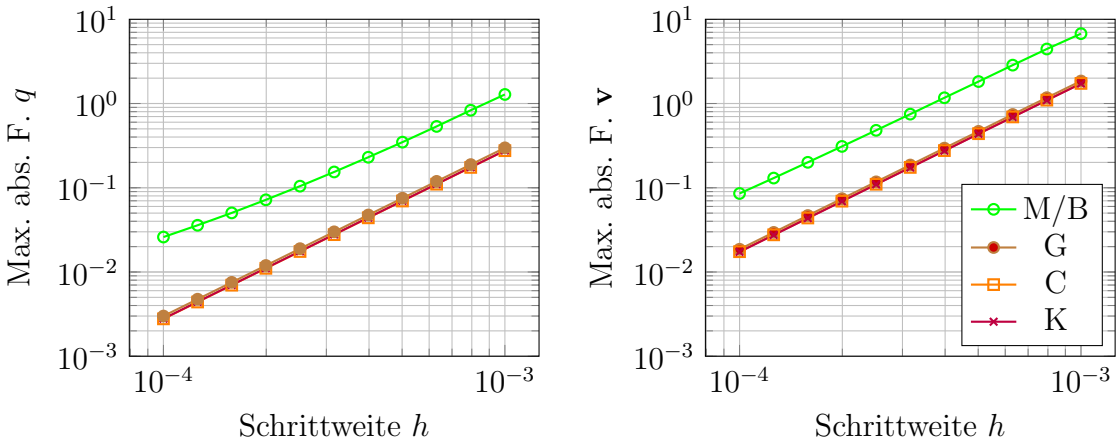

Abbildung 6.5: Maximaler absoluter Fehler von q und $\mathbf{v}$ für den schweren Kreisel in $SO(3)$, vgl. (3.20) für $p = 2$, für die Verfahren M: Munthe-Kaas, B: BLieDF, C: Crouch-Grossman, K: kommutatorfrei, G: Generalized-α

$$\mathbf{R}_{n+1} = \mathbf{R}_n \circ \exp\left(\widetilde{\boldsymbol{\omega}}_0^{(n)}\right), \tag{6.5a}$$

$$\frac{3}{2}\boldsymbol{\omega}_0^{(n)} - \frac{1}{2}\boldsymbol{\omega}_0^{(n-1)} = \boldsymbol{\Omega}_{n+1}, \tag{6.5b}$$

$$\frac{3}{2}\boldsymbol{\Omega}_{n+1} - 2\boldsymbol{\Omega}_n + \frac{1}{2}\boldsymbol{\Omega}_{n-1} = -h\mathbf{J}^{-1}\left(\widetilde{\boldsymbol{\Omega}}_{n+1}\mathbf{J}\boldsymbol{\Omega}_{n+1} - \widetilde{\mathbf{X}}\mathbf{R}_{n+1}^{\top} m\boldsymbol{\gamma}\right). \tag{6.5c}$$

Interessant ist nun, welches der Verfahren (6.3), (6.4) und (6.5) das effizienteste zur Lösung von (3.20) ist und wie sich das Generalized-α-Verfahren (3.21) einordnet.
Dazu ist zunächst in Abbildung 6.5 der maximale absolute Fehler der Konfigurationsvariablen q und der Geschwindigkeit $\mathbf{v}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in doppelt logarithmischer Skala dargestellt. Dabei zeigt sich, dass die Verfahren (6.3) und (6.4) in beiden Variablen die genauesten Ergebnisse liefern. Das Generalized-α-Verfahren (3.21) ist minimal ungenauer. Das Verfahren (6.5) (also MKBDF- und BLieDF-Verfahren) hat die größte Fehlerkonstante.
Trotzdem ist es möglich, dass (6.5) effizienter als die anderen Verfahren rechnet. Daher ist im linken Teil der Abbildung 6.6 der maximale absolute Fehler von q über die Rechenzeit dargestellt, die die Verfahren für alle Zeitschritte benötigen, vgl. Bemerkung 27. Dabei zeigt sich, dass das Generalized-α-Verfahren am effizientesten ist. Die anderen drei Verfahren (6.3), (6.4) und (6.5) liegen in Hinblick auf die Effizienz nah beieinander. Ein leichter Nachteil für das Verfahren (6.5) ist jedoch aufgrund der größeren Fehlerkonstanten erkennbar. Dies liegt daran, dass für $k = 2$ auch im Crouch-Grossman-Verfahren (6.3) lediglich zwei Exponentialabbildungen pro Zeitschritt ausgewertet werden müssen, was rechentechnisch vertretbar ist; dazu kommt die sehr gute Genauigkeit des Verfahrens. Für wachsendes p oder andere Testprobleme, bei denen die Auswertung der Exponentialabbildung aufwendiger ist, könnte sich dies jedoch ändern.
Zur Lösung von (3.44) ergeben sich die MKBDF-Verfahren (3.57) und die BLieDF-Verfahren (3.62) zu

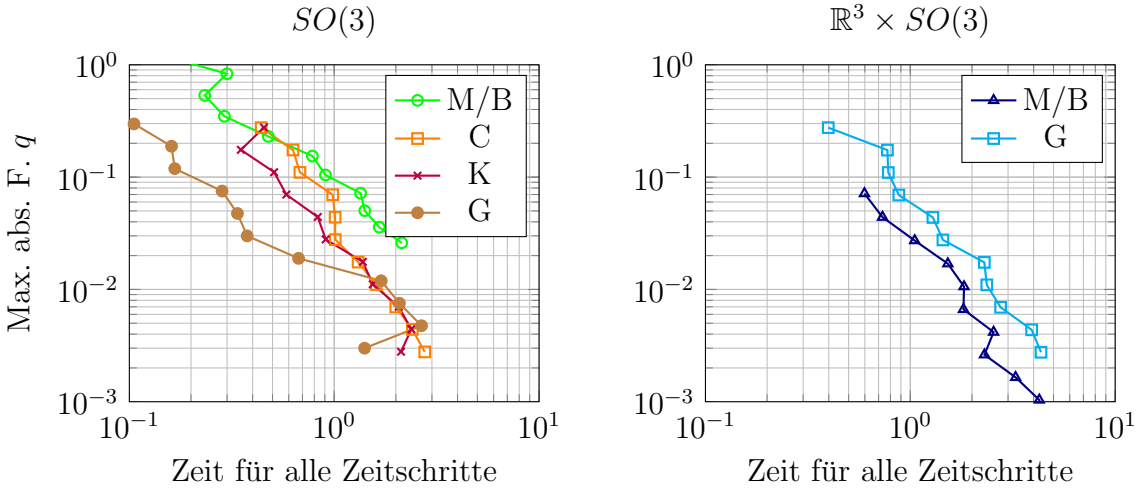

Abbildung 6.6: Zeit für alle Zeitschritte für den schweren Kreisel in $SO(3)$ mit den Bewegungsgl. (3.20) (links) in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) (rechts) für $p = 2$ und die Verfahren M: Munthe-Kaas, B: BLieDF, C: Crouch-Grossman, K: kommutatorfrei, G: Generalized-α

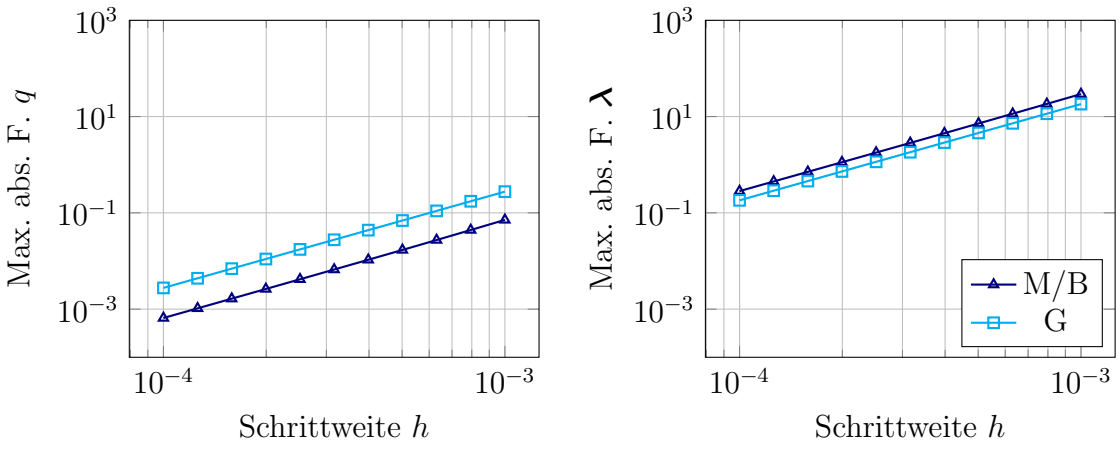

Abbildung 6.7: Maximaler absoluter Fehler von q und $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) für $p = 2$ und die Verfahren M: Munthe-Kaas, B: BLieDF, G: Generalized-α

$$\mathbf{q}_{n+1} = \mathbf{q}_n \circ \exp\left(\widetilde{\boldsymbol{\omega}}_0^{(n)}\right), \tag{6.6a}$$

$$\frac{3}{2}\boldsymbol{\omega}_0^{(n)} - \frac{1}{2}\boldsymbol{\omega}_0^{(n-1)} = \mathbf{v}_{n+1}, \tag{6.6b}$$

$$\frac{1}{h}\mathbf{M}(q_{n+1})\left(\frac{3}{2}\mathbf{v}_{n+1} - 2\mathbf{v}_n + \frac{1}{2}\mathbf{v}_{n-1}\right) = -\mathbf{g}(t_{n+1}, q_{n+1}, \mathbf{v}_{n+1}) - \mathbf{B}^\top(q_{n+1})\boldsymbol{\lambda}_{n+1}, \tag{6.6c}$$

$$\mathbf{0} = \boldsymbol{\Phi}(q_{n+1}). \tag{6.6d}$$

In Abbildung 6.7 ist der maximale absolute Fehler der Konfigurationsvariable q und der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in doppelt logarithmischer Skala dargestellt für die Lösung von (3.44). Dabei ist zu erkennen, dass das Verfahren (6.6) und das Generalized-α-Verfahren (3.55) zur Lösung von (3.44) in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ in beiden Variablen mit zweiter Ordnung konvergieren. In der Konfigurationsvariable q liefert das Verfahren (6.6) genauere Ergebnisse als das Generalized-α-Verfahren (3.55). Werden die Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ untersucht, ist jedoch das Generalized-α-Verfahren (3.55) genauer. Der Vergleich im rechten Teil der Abbildung 6.6 zeigt zudem, dass das Verfahren (6.6) bei der Lösung von (3.44) effizienter ist. Hier ist der maximale absolute Fehler von q über die Zeit dargestellt, die die Verfahren für alle Zeitschritte benötigen, vgl. Bemerkung 27.

Vergleich für p = 3

Von nun an werden die Verfahren (3.4), (3.7), (3.23) und (3.31) zur Lösung von (3.20) nicht mehr für das jeweilige p explizit angegeben, da sie mit größer werdendem p auch immer umfangreicher werden.
Zunächst wird überprüft, ob alle betrachteten Verfahren die gewünschte Konvergenzordnung aufweisen. Dazu ist in Abbildung 6.8 der maximale absolute Fehler der Konfigurationsvariablen q und der Geschwindigkeit $\mathbf{v}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in doppelt logarithmischer Skala dargestellt. Dabei zeigt sich erneut, dass das Crouch-Grossman-Verfahren (3.4) und das kommutatorfreie Lie-Gruppen-Verfahren (3.7) die genauesten Ergebnisse liefern. Die beiden BDF-Verfahren sind ungenauer.
Beim Effizienztest im linken Teil der Abbildung 6.9 sind die Ergebnisse der BLieDF-

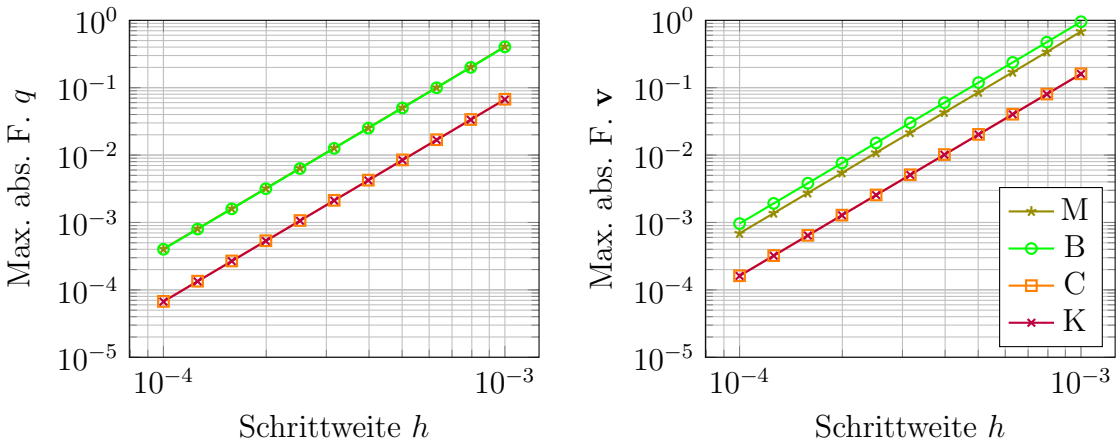

Abbildung 6.8: Maximaler absoluter Fehler von q und $\mathbf{v}$ für den schweren Kreisel in $SO(3)$ mit den Bewegungsgl. (3.20) für $p = 3$ und die Verfahren M: Munthe-Kaas, B: BLieDF, C: Crouch-Grossman, K: kommutatorfrei

Verfahren (3.31), des Crouch-Grossman-Verfahrens (3.4) und des kommutatorfreien Lie-Gruppen-Verfahrens (3.7) sehr eng beieinander. Aufgrund von Schwankungen bei der Zeitmessung ist es schwierig einzuschätzen, welches dieser drei Verfahren am effizientesten eine Lösung bestimmt. Die BLieDF-Verfahren konnten somit trotz deutlich schlechterer Genauigkeit durch einen guten Rechenaufwand aufholen. Die MKBDF-Verfahren (3.57) sind in Hinblick auf die Effizienz etwas abgeschlagen, was vermutlich an der häufigen Auswertung der Matrix-Kommutatoren in jedem Zeitschritt liegt.
Für die Lösung des beschränkten Mehrkörpersystems (3.44) ist in Abbildung 6.10 der maximale absolute Fehler der Konfigurationsvariable q und der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in doppelt logarithmischer Skala dargestellt. Beide Verfahren konvergieren in beiden Variablen mit der Ordnung drei. In jedem Fall sind die MKBDF-Verfahren (3.57) genauer als die BLieDF-Verfahren (3.62). Mit dem rechten Teil der Abbildung 6.9 wird die Effizienz überprüft. In den MKBDF-Verfahren (3.57) werden pro Zeitschritt mehr Matrix-Kommutatoren ausgewertet als

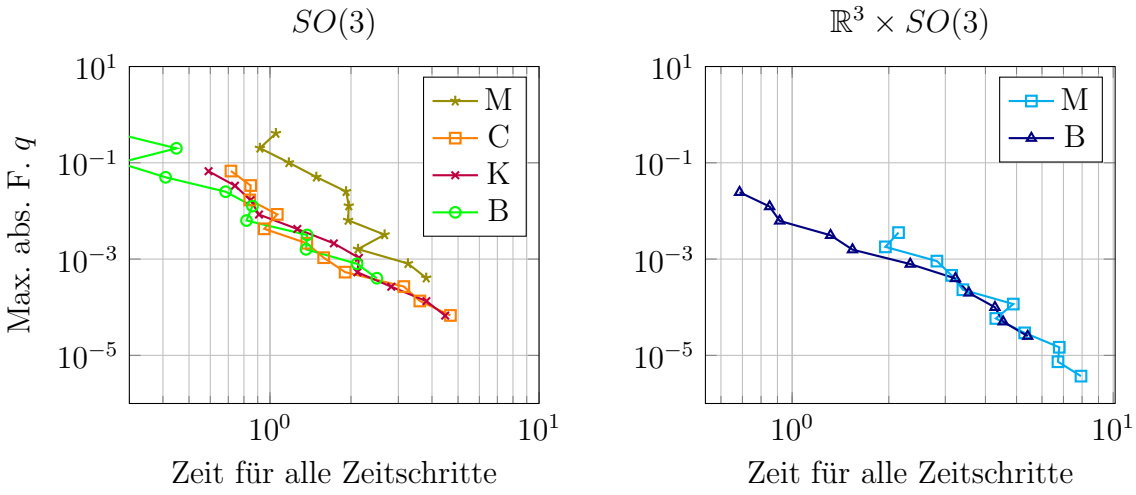

Abbildung 6.9: Zeit für alle Zeitschritte für den schweren Kreisel in $SO(3)$ mit den Bewegungsgl. (3.20) (links) und in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) (rechts) für $p = 3$ und die Verfahren M: Munthe-Kaas, B: BLieDF, C: Crouch-Grossman, K: kommutatorfrei

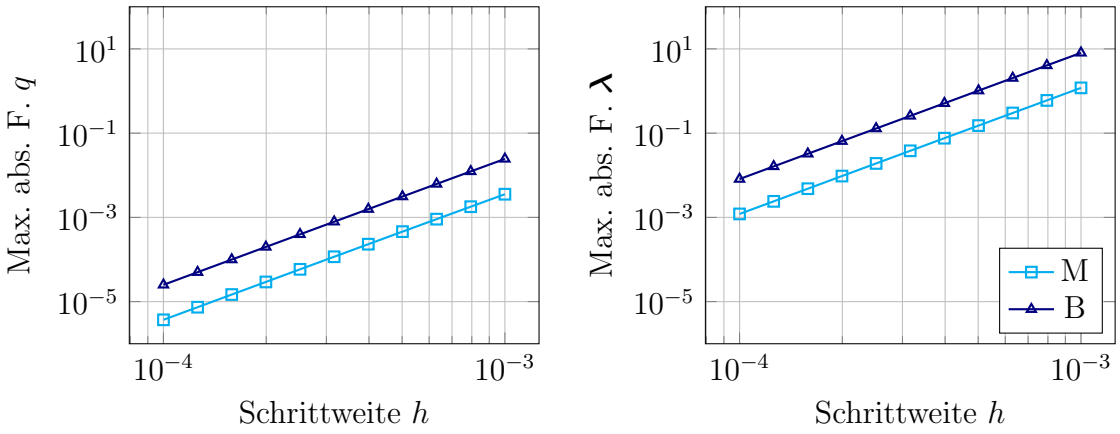

Abbildung 6.10: Maximaler absoluter Fehler von q und $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) für $p = 3$ und die Verfahren M: Munthe-Kaas, B: BLieDF

bei den BLieDF-Verfahren (3.62), trotzdem ist in Hinblick auf die Effizienz kein großer Unterschied bemerkbar, weil die MKBDF-Verfahren (3.57) genauere Ergebnisse liefern, vgl. Abbildung 6.10.

Vergleich für p = 4

Zunächst wird überprüft, ob alle betrachteten Verfahren die gewünschte Konvergenzordnung vier aufweisen. Dazu ist in Abbildung 6.11 der maximale absolute Fehler der Konfigurationsvariable q und der Geschwindigkeit $\mathbf{v}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in doppelt logarithmischer Skala dargestellt. Es zeigt sich, dass das Crouch-Grossman-Verfahren (3.4) und das kommutatorfreie Lie-Gruppen-Verfahren (3.7) die höchste Genauigkeit erreichen. Die beiden BDF-Verfahren (3.57) und (3.62) sind ungenauer.
Beim Effizienztest in Abbildung 6.12 liegen die BLieDF-Verfahren (3.31) ganz vorne gefolgt vom kommutatorfreien Lie-Gruppen-Verfahren (3.7) und dem Crouch-Grossman-Verfahren (3.4). Die MKBDF-Verfahren (3.57) schneiden schlechter ab. Trotz der schlechteren Genauigkeit konnten die BLieDF-Verfahren (3.31) in Hinblick

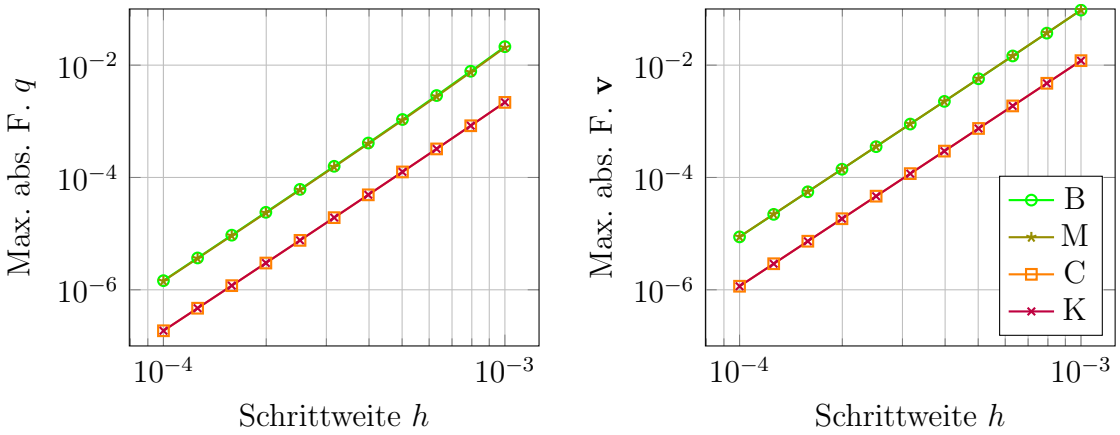

Abbildung 6.11: Maximaler absoluter Fehler von q und $\mathbf{v}$ für den schweren Kreisel in $SO(3)$ mit den Bewegungsgl. (3.20) für $p = 4$ und die Verfahren M: Munthe-Kaas, B: BLieDF, C: Crouch-Grossman, K: kommutatorfrei

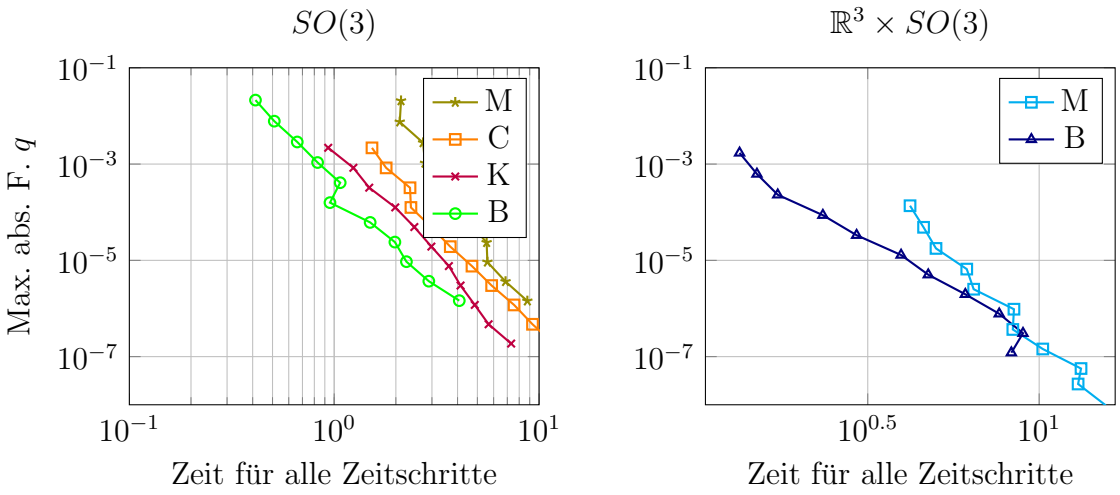

Abbildung 6.12: Zeit für alle Zeitschritte für den schweren Kreisel in $SO(3)$ mit Bewegungsgleichungen (3.20) (links) und in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) (rechts) für $p = 4$ und die Verfahren M: Munthe-Kaas, B: BLieDF, C: Crouch-Grossman, K: kommutatorfrei

auf die Effizienz Vorteile aufzeigen.
Nun soll die Lösung von (3.44) durch die MKBDF-Verfahren (3.57) und die BLieDF-Verfahren (3.62) untersucht werden. In Abbildung 6.13 ist der maximale absolute Fehler der Konfigurationsvariable q und der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in doppelt logarithmischer Skala dargestellt. Dabei können in $\mathbb{R}^3 \times SO(3)$ für die MKBDF-Verfahren (3.57) genauere Ergebnisse als in den BLieDF-Verfahren (3.62) erhalten werden.
Zur Überprüfung der Effizienz wurde im rechten Teil der Abbildung 6.12 der maximale absolute Fehler in q über die benötigte Zeit für alle Zeitschritte dargestellt. Es zeigt sich, dass aufgrund der geringeren Anzahl an Matrix-Kommutatorauswertungen die BLieDF-Verfahren effizienter eine Lösung bestimmen.

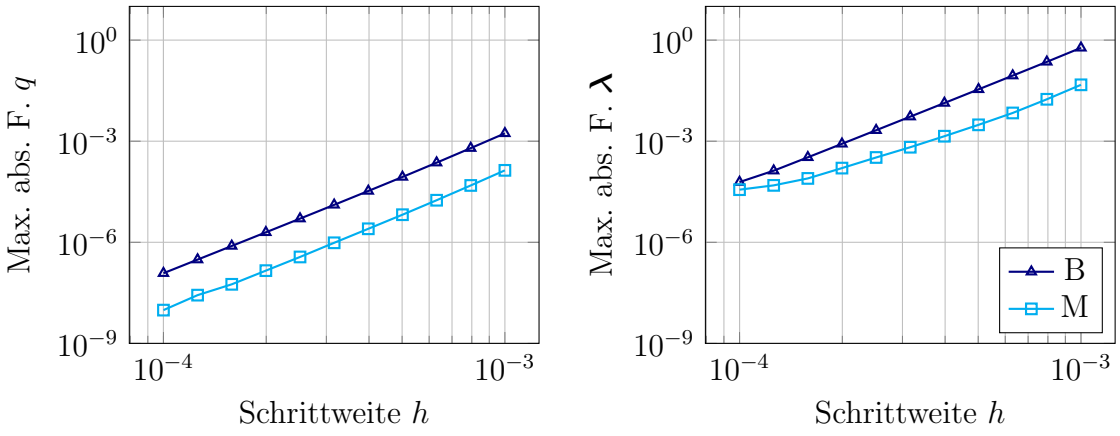

Abbildung 6.13: Maximaler absoluter Fehler von q und $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) für $p = 4$ und die Verfahren M: Munthe-Kaas, B: BLieDF

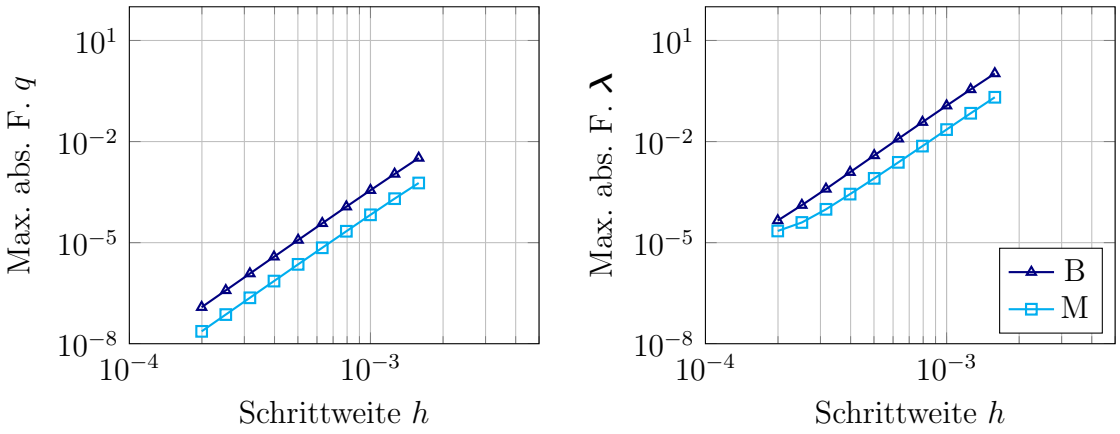

Abbildung 6.14: Maximaler absoluter Fehler von q und $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) für $p = 5$ und die Verfahren M: Munthe-Kaas, B: BLieDF

Vergleich für p = 5

Für $p > 4$ wurden in dieser Arbeit keine Parameter für das Crouch-Grossman-Verfahren (3.4) und dem kommutatorfreien Lie-Gruppen-Verfahren (3.7) angegeben. Der Vergleich beschränkt sich daher im Folgenden auf die Munthe-Kaas-BDF-Verfahren (3.57) und die BLieDF-Verfahren (3.62). In Abbildung 6.14 ist der maximale absolute Fehler der Konfigurationsvariablen q und der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ zur Lösung von (3.51) in doppelt logarithmischer Skala dargestellt. Dabei schneidet erneut das Munthe-Kaas-BDF-Verfahren (3.57) besser ab. Beide Verfahren besitzen die gewünschte Konvergenzordnung $p = k = 5$.
Wird die Effizienz der Verfahren in Abbildung 6.15 in den Lie-Gruppen-Formulierungen $SO(3)$ und $\mathbb{R}^3 \times SO(3)$ verglichen, so ist das BLieDF-Verfahren (3.31) überlegen.

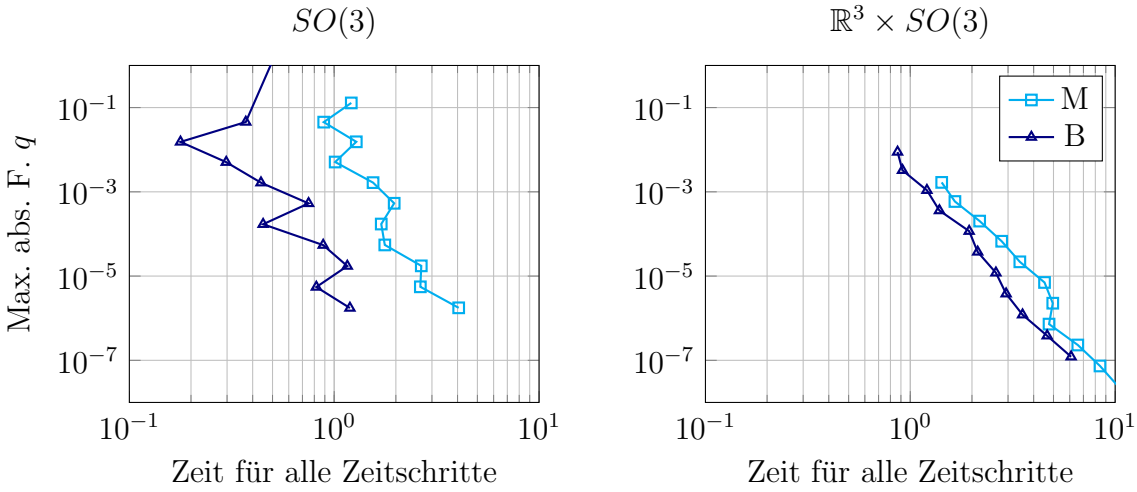

Abbildung 6.15: Zeit für alle Zeitschritte für den schweren Kreisel in $SO(3)$ mit Bewegungsgl. (3.20) (links) und in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) (rechts) für $p = 5$ und die Verfahren M: Munthe-Kaas, B: BLieDF

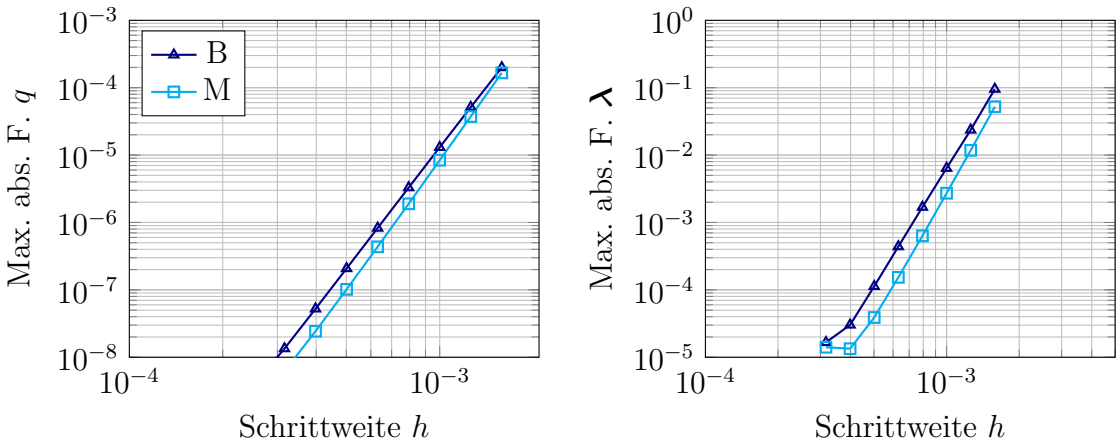

Abbildung 6.16: Maximaler absoluter Fehler von q und $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) für $p = 6$ und die Verfahren M: Munthe-Kaas, B: BLieDF

Vergleich für p = 6

In Abbildung 6.16 ist der maximale absolute Fehler der Konfigurationsvariablen q und der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ über die Schrittweite $h \in [10^{-4}, 10^{-3}]$ in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ in doppelt logarithmischer Skala dargestellt. Dazu wurde eine Lösung von (3.51) mit den MKBDF-Verfahren (3.57) und den BLieDF-Verfahren (3.62) bestimmt. Die gewünschte Konvergenzordnung $p = k = 6$ ist in beiden Variablen und Verfahren zu erkennen, jedoch rechnen die Munthe-Kaas-BDF-Verfahren genauer.
Wird die Effizienz der Verfahren für die Lie-Gruppen-Formulierungen $SO(3)$ und $\mathbb{R}^3 \times SO(3)$ in Abbildung 6.17 verglichen, so schneiden erneut die BLieDF-Verfahren (3.62) besser ab.

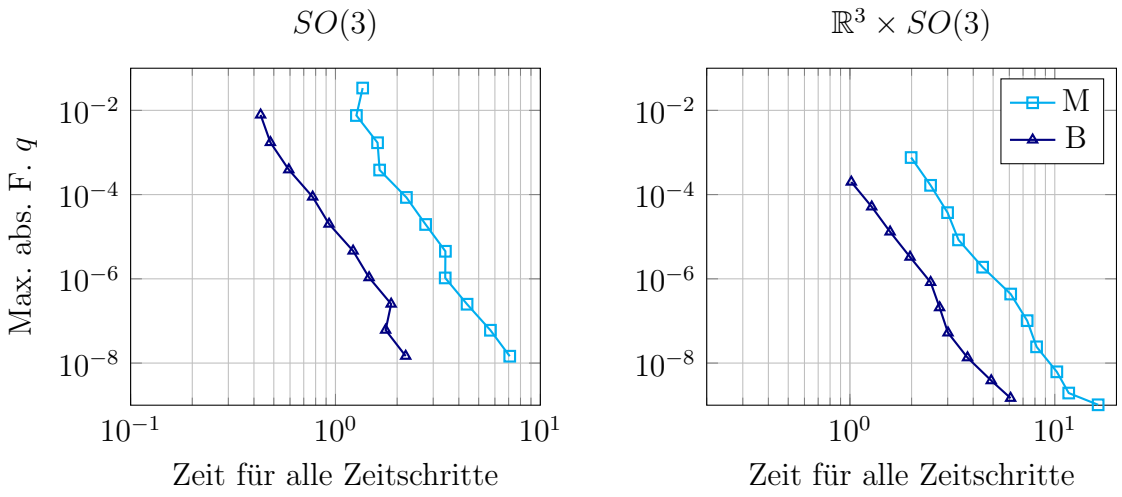

Abbildung 6.17: Zeit für alle Zeitschritte für den schweren Kreisel in $SO(3)$ mit Bewegungsgl. (3.20) (links) und in $\mathbb{R}^3 \times SO(3)$ mit den Bewegungsgl. (3.51) (rechts) für $p = 6$ und die Verfahren M: Munthe-Kaas, B: BLieDF

Fazit

Im Aufwand-Nutzen-Vergleich der MKBDF-Verfahren (3.57) und BLieDF-Verfahren (3.62) waren für die untersuchten Testprobleme des schweren Kreisels mit und ohne Zwangsbedingungen die BLieDF-Verfahren überlegen. Beide Verfahren stimmen jedoch für $1 \leq k = p \leq 2$ überein.
Ab einer Konvergenzordnung drei war zu erkennen, dass das kommutatorfreie-Lie-Gruppen-Verfahren (10) effizienter eine Lösung bestimmt als das Crouch-Grossman-Verfahren (3.4). Für $k = 1$ stimmen beide Verfahren überein und für $k = 2$ waren die Verfahren etwa gleich gut, da auch im Crouch-Grossman-Verfahren lediglich zwei Exponentialabbildungen pro Zeitschritt ausgewertet werden mussten. Der direkte Vergleich aller vier genannten Verfahren zeigt, dass mit wachsender Konvergenzordnung die BLieDF-Verfahren und das kommutatorfreie Lie-Gruppen-Verfahren am effizientesten die Lösung bestimmten. Jedoch zeigte sich, dass für $p = 2$ das Generalized-α-Verfahren (2.18) sehr effizient rechnete.
Trotz der guten Effizienz der BLieDF-Verfahren im Allgemeinen war die Genauigkeit bei gleicher Schrittweite h dieses Verfahrens immer schlechter als die des kommutatorfreien Lie-Gruppen-Verfahrens und des Crouch-Grossman-Verfahrens. Die Munthe-Kaas-BDF-Verfahren lagen in Hinblick auf die Genauigkeit für das Problem mit Zwangsbedingungen vor den BLieDF-Verfahren.

6.2.3 Notwendigkeit des Korrekturterms

In Abschnitt 5.1.1 wurde gezeigt, dass der lokale Abbruchfehler $\mathbf{l}_n^{\boldsymbol{\omega}}$ der BLieDF-Verfahren nur von Ordnung $\mathcal{O}(h^2)$ ist, wenn der Korrekturterm $\mathbf{L}_{h,n}^{(k)} \equiv \mathbf{0}$ ist. Dies soll durch numerische Tests in diesem Abschnitt verifiziert werden. Dazu werden alle Verfahrensparameter $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$ für $k = 3, 4, 5, 6$ auf null gesetzt. Numerisch sollte in jedem Fall eine zweite Ordnung zu beobachten sein.
In Abbildung 6.18 ist der absolute Fehler in q zum Endzeitpunkt $t_{\text{end}} = 1$ über die Schrittweite in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ aufgetragen. Wie zu erwarten war, ist für $3 \leq k \leq 6$ nur eine zweite Ordnung zu beobachten und somit

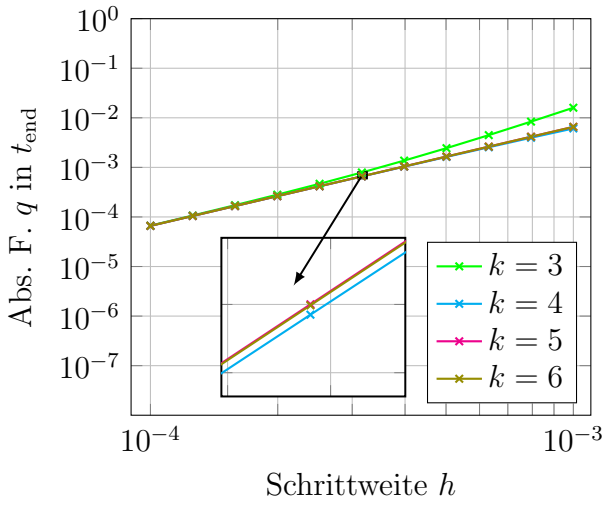

Abbildung 6.18: Absoluter Fehler von q in t_{end} berechnet mit den BLieDF-Verfahren mit Korrekturterm $\mathbf{L}_{h,n}^{(k)} \equiv \mathbf{0}$ zur Lösung von (3.62)

eine Ordnungsreduktion vorhanden. Dies entspricht den Untersuchungen aus Abschnitt 3.2.3.

6.2.4 Wahl der freien Parameter in den BLieDF-Verfahren

Die Definition des Korrekturterms $\mathbf{L}_{h,n}^{(k)}$ (3.34) bei den BLieDF- Verfahren (3.62) enthält freie Parameter $a_{i,j}^{(k)}$ und $b_{i,j}^{(k)}$. Für $k = 3$ und $k = 4$ ist mit (3.40) eine offensichtliche Wahl für solche Parameter durch die Verwendung einer Differenzenapproximation getroffen wurden. Diese können in jedem Fall verwendet werden, wenn für ein gegebenes Testproblem keine Daten für Parameterkombinationen vorhanden sind. In diesem Abschnitt sollen jedoch auch andere Parameterkombinationen untersucht werden, die für das spezielle Benchmark schwerer Kreisel zu genaueren Ergebnissen führen.
Numerische Testrechnungen für das Benchmark schwerer Kreisel in $\mathbb{R}^3 \times SO(3)$ zeigen, dass betragsmäßig große Parameter in $\mathbf{L}_{h,n}^{(k)}$ zur numerischen Instabilität der BLieDF-Verfahren (3.62) mit $k \geq 3$ führen können. Deshalb werden alle nachfolgenden Rechnungen auf Verfahrensparameter $a_{i,j}^{(k)},\ b_{i,j}^{(k)} \in (-10, 10)$ für $(i, j = 0, ..., k)$ beschränkt.
Für die BLieDF-Verfahren (3.62) sind dazu zunächst für $j = 1, 2$ Parameter $a_{i,j}^{(k)}$, $(i = 0, ..., k)$, $b_{i,j}^{(k)}$, $(i = 1, ..., k)$, zu bestimmen, die die Konsistenzbedingungen (3.41) erfüllen müssen. Dabei gibt es zwei Gruppen von Korrekturtermen. Ein expliziter Korrekturterm ist durch Parameter definiert, die zusätzlich zu den Konsistenzbedingungen (3.41) auch $a_{0,j}^{(k)} = b_{1,j}^{(k)} = 0$ für $j = 1, 2$ und $k = 3, 4$ erfüllen. Gegenüber den im allgemeinen implizit gegebenen Korrekturtermen haben explizite Korrekturterme praktisch den Vorteil, dass der Korrekturterm im Newton-Verfahren nicht in jedem Iterationsschritt erneut ausgewertet werden muss, vgl. Abschnitt 3.3.4. Dies kann aufgrund der nichtlinearen Struktur des Korrekturterms zu einer enormen Rechenzeiteinsparung führen. Daher soll die Effizienz der BLieDF-Verfahren (3.62) für $k = 3, 4$ mit explizitem und implizitem Korrekturterm untersucht werden. Insgesamt werden drei verschiedene Varianten von Parameterkombinationen verglichen.

Verfahren mit k = 3

Zunächst wird aus einer Stichprobe von je 100 zufällig bestimmten konsistenten Parametersätzen mit $a_{i,j}^{(3)} \in (-10, 10)$, $(i = 0, 1, 2, 3)$, $b_{i,j}^{(3)} \in (-10, 10)$, $(i = 1, 2, 3)$, für $j = 1, 2$ derjenige Parametersatz ausgewählt, für den in einem numerischen Test der absolute Fehler im Endpunkt $t_{\text{end}} = 1$ von q mit der Schrittweite $h = 1 \cdot 10^{-3}$ für das Benchmark schwerer Kreisel in der Formulierungen $\mathbb{R}^3 \times SO(3)$ minimal wird. Das bedeutet natürlich nicht, dass nicht noch genauere Ergebnisse mit anderen Parameterkombinationen erzielt werden könnten. Diese Verfahrensparameter sind im expliziten Korrekturterm-Fall gegeben durch

$$a_{1,1}^{(3)} = -1.37872190956776, \quad a_{2,1}^{(3)} = -2.95414965850481, \tag{6.7a}$$

$$a_{3,1}^{(3)} = -4.46630667652555, \quad a_{1,2}^{(3)} = 6.43118898334207, \tag{6.7b}$$

$$a_{2,2}^{(3)} = -0.251965339229852, \quad a_{3,2}^{(3)} = 6.04133559544059, \tag{6.7c}$$

$$b_{2,1}^{(3)} = 4.86465799288703, \quad b_{3,1}^{(3)} = 4.01785358504442, \tag{6.7d}$$

$$b_{2,2}^{(3)} = -5.50013300767788, \quad b_{3,2}^{(3)} = -6.72042623187493. \tag{6.7e}$$

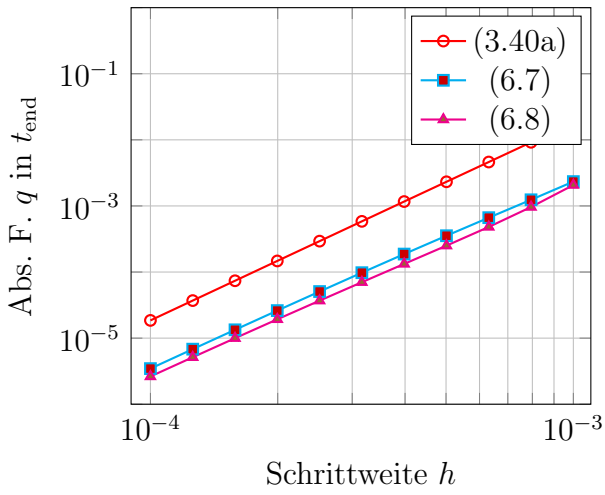

Abbildung 6.19: Absoluter Fehler von q in t_{end} berechnet mit den BLieDF-Verfahren mit $k = 3$ für unterschiedliche Verfahrensparametersätze im Korrekturterm $\mathbf{L}_{h,n}^{(3)}$ zur Lösung von (3.62) in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$

Wird ein impliziter Korrekturterm zugelassen, so war der Fehler minimal bei Verfahrensparametern

$$a_{0,1}^{(3)} = 5.24281797269331, \quad a_{1,1}^{(3)} = 2.80233526987551, \tag{6.8a}$$
$$a_{2,1}^{(3)} = -3.85724425749725, \quad a_{3,1}^{(3)} = 3.30539890486575, \tag{6.8b}$$
$$a_{0,2}^{(3)} = 0.501916816184551, \quad a_{1,2}^{(3)} = -3.37875139908466, \tag{6.8c}$$
$$a_{2,2}^{(3)} = -1.13188143412463, \quad a_{3,2}^{(3)} = 6.44554348846025, \tag{6.8d}$$
$$b_{1,1}^{(3)} = -3.61957377496834, \quad b_{2,1}^{(3)} = 0.108068150385785, \tag{6.8e}$$
$$b_{3,1}^{(3)} = -3.89846893202143, \quad b_{1,2}^{(3)} = 3.14090074432045, \tag{6.8f}$$
$$b_{2,2}^{(3)} = 2.32024603081716, \quad b_{3,2}^{(3)} = -7.89797424657311. \tag{6.8g}$$

Ein weiterer Verfahrensparametersatz wurde durch die Wahl mit Parametern aus (3.40a) definiert, die durch die Verwendung eines Differenzenquotienten bestimmt wurden.
Der direkte Vergleich der drei Parametervarianten (3.40a), (6.7) und (6.8) ist in Abbildung 6.19 dargestellt. Wie die Konvergenzanalyse in Kapitel 5 bewiesen hat, konvergieren die BLieDF-Verfahren für $k = 3$ für alle Verfahrensparametersätze mit dritter Ordnung. Die Parameter aus (3.40a) schneiden am schlechtesten ab. Die genausten Ergebnisse liefert das Verfahren mit einem impliziten Korrekturterm (Parameter (6.8)). Da in diesem Fall die Newton-Iteration jedoch am aufwendigsten ist, vgl. Abschnitt 3.3.4, wird in Abbildung 6.20 die Genauigkeit über die Rechenzeit, die für einen Zeitschritt benötigt wird, für die beiden Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ abgetragen. Dabei kann festgestellt werden, dass die Berechnung mit explizitem Korrekturterm (Parameter (6.7) und (3.40b)) stets effizienter ist.

Verfahren mit k = 4

Für $k = 4$ wird analog zu $k = 3$ vorgegangen.
Zunächst wird aus einer Stichprobe von je 100 zufällig bestimmten konsistenten Pa-

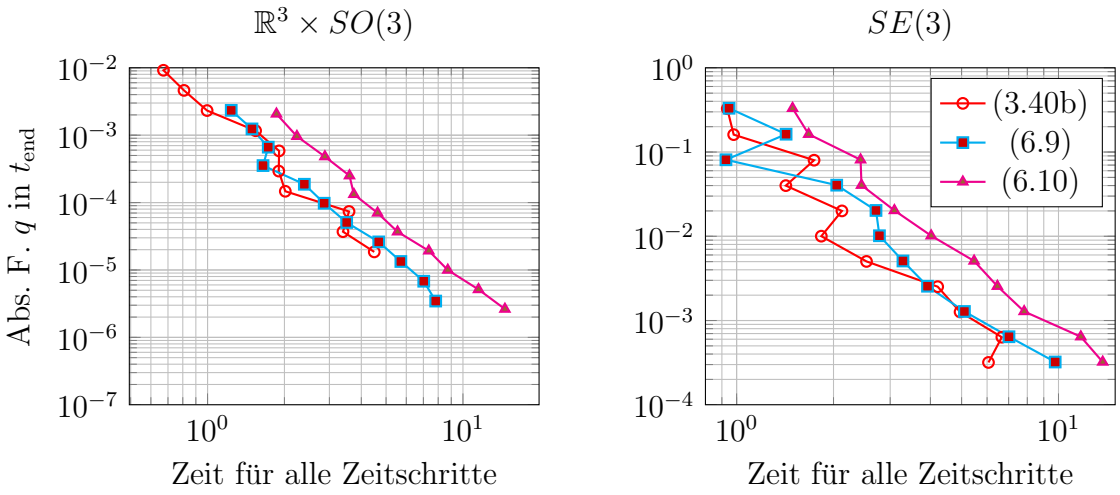

Abbildung 6.20: Absoluter Fehler von q in t_{end} berechnet mit den BLieDF-Verfahren in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ für $k = 3$ und unterschiedliche Verfahrensparametersätze im Korrekturterm $\mathbf{L}_{h,n}^{(3)}$

rametersätzen mit $a_{i,j}^{(4)} \in (-10, 10)$, $(i = 0, ..., 4)$, $b_{i,j}^{(4)} \in (-10, 10)$, $(i = 1, ..., 4)$, für $j = 1, 2$ derjenige Parametersatz ausgewählt, für den in einem numerischen Test der absolute Fehler im Endpunkt $t_{\text{end}} = 1$ von q mit der Schrittweite $h = 1 \cdot 10^{-3}$ für das Benchmark schwerer Kreisel in der Formulierung $\mathbb{R}^3 \times SO(3)$ minimal wird. Das bedeutet natürlich nicht, dass nicht noch genauere Ergebnisse mit anderen Parameterkombinationen erzielt werden könnten. Diese Verfahrensparameter sind im expliziten Korrekturterm-Fall gegeben durch

$$a_{1,1}^{(4)} = -1.67876614106032, \quad a_{2,1}^{(4)} = 3.07122613520328, \tag{6.9a}$$
$$a_{3,1}^{(4)} = -2.50907130087646, \quad a_{1,2}^{(4)} = 6.9760443494959, \tag{6.9b}$$
$$a_{2,2}^{(4)} = -8.18750698866965, \quad a_{3,2}^{(4)} = 2.21000229945281, \tag{6.9c}$$

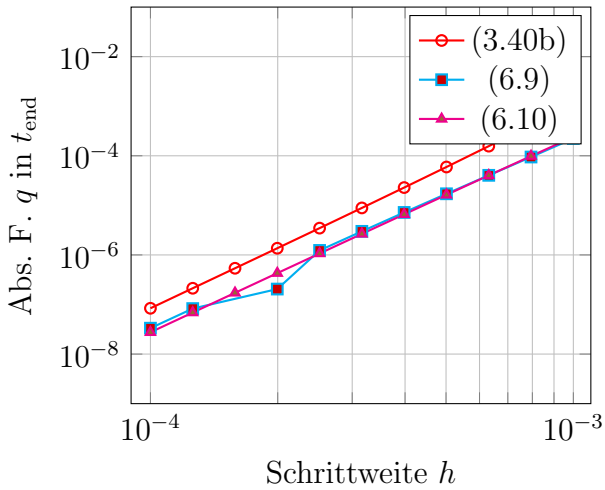

Abbildung 6.21: Absoluter Fehler von q in t_{end} berechnet mit den BLieDF-Verfahren mit $k = 4$ für unterschiedliche Verfahrensparametersätze im Korrekturterm $\mathbf{L}_{h,n}^{(4)}$ zur Lösung von (3.62) in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$

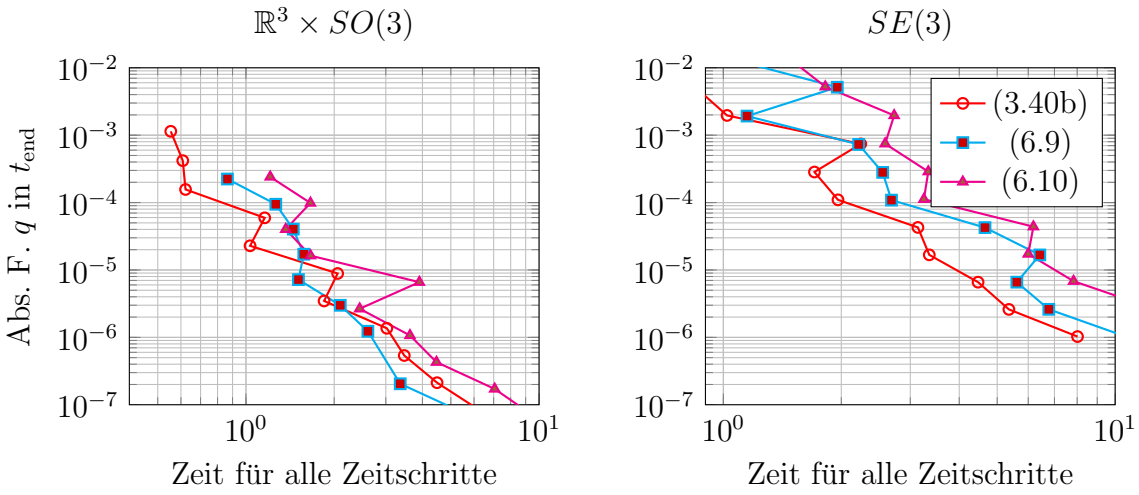

Abbildung 6.22: Absoluter Fehler von q in t_{end} berechnet mit den BLieDF-Verfahren in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ (links) und $SE(3)$ (rechts) für $k = 4$ und unterschiedliche Verfahrensparametersätze im Korrekturterm $\mathbf{L}_{h,n}^{(4)}$

$$b_{2,1}^{(4)} = -3.36166983338703, \quad b_{3,1}^{(4)} = 9.60778111160441, \tag{6.9d}$$
$$b_{2,2}^{(4)} = -6.35906537451745, \quad b_{3,2}^{(4)} = 6.95500937843368. \tag{6.9e}$$

Wird ein impliziter Korrekturterm zugelassen, so war der Fehler minimal bei Verfahrensparametern

$$a_{0,1}^{(4)} = 5.5259797909541, \quad a_{1,1}^{(4)} = -5.74722666327503, \tag{6.10a}$$
$$a_{2,1}^{(4)} = 4.27407568275581, \quad a_{3,1}^{(4)} = 2.67226789875473, \tag{6.10b}$$
$$a_{0,2}^{(4)} = 5.9482882956157, \quad a_{1,2}^{(4)} = -8.19039669944764, \tag{6.10c}$$
$$a_{2,2}^{(4)} = 3.70086989589663, \quad a_{3,2}^{(4)} = -2.16080421168301, \tag{6.10d}$$
$$b_{1,1}^{(4)} = -2.62867609640801, \quad b_{2,1}^{(4)} = -3.31210875450453, \tag{6.10e}$$
$$b_{3,1}^{(4)} = 0.389431322962876, \quad b_{1,2}^{(4)} = -3.00106876837373, \tag{6.10f}$$
$$b_{2,2}^{(4)} = -0.584753934011666, \quad b_{3,2}^{(4)} = 6.70981550891437. \tag{6.10g}$$

Ein weiterer Verfahrensparametersatz wurde durch die Wahl mit Parametern aus (3.40b) definiert, die durch die Verwendung eines Differenzenquotienten bestimmt wurden.
Der direkte Vergleich der drei Parametervarianten (3.40b), (6.9) und (6.10) ist in Abbildung 6.21 dargestellt. Wie die Konvergenzanalyse in Kapitel 5 bewiesen hat, konvergieren die BLieDF-Verfahren für $k = 4$ für alle Verfahrensparametersätze mit vierter Ordnung. Dabei schneiden die Parameter aus (3.40b) am schlechtesten ab. Die anderen beiden Parametersätze liefern ähnliche Ergebnisse. Im Fall des impliziten Korrekturterms ist die Newton-Iteration jedoch am aufwendigsten, vgl. Abschnitt 3.3.4. Daher wird in Abbildung 6.22 die Genauigkeit über die Zeit, die für einen Zeitschritt benötigt wird, für die beiden Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ abgetragen. Dabei kann festgestellt werden, dass die Berechnung mit explizitem Korrekturterm (Parameter (6.9)) effizienter ist. Jedoch liegen alle Parametervarianten für $k = 4$ näher beieinander.

Fazit

Die Untersuchungen dieses Absatzes zeigten, dass für das Testproblem schwerer Kreisel mit anderen Parametervarianten als die in (3.40) definierten noch genauere Ergebnisse erzielt werden können. Sind für ein gegebenes Testproblem jedoch keine Daten zu optimalen Parametern vorhanden, so können die Parameter (3.40) verwendet werden.
Zudem zeigt sich, dass für unser Benchmarkproblem ein expliziter Korrekturterm $\mathbf{L}_{h,n}^{(k)}$ dem impliziten aus Effizienzgründen vorzuziehen ist. Aufgrund des Mehraufwandes in der Newton-Iteration bei einem impliziten Korrekturterm, kann davon ausgegangen werden, dass dies auch für andere Testprobleme der Fall ist.

6.3 Variable Schrittweiten

In diesem Abschnitt sollen numerische Tests für das Generalized-α-Verfahren (4.11) und die BLieDF-Verfahren (5.35) für variable Schrittweiten durchgeführt werden. Dabei wird zunächst überprüft, was bei einmaliger Änderung der Schrittweite passiert. Der Fokus hierbei liegt darauf zu bestätigen, dass nach dieser Änderung der Schrittweite keine Ordnungsreduktion zu beobachten ist. Damit kann bestätigt werden, dass die Modifikation der Verfahren auf variable Schrittweiten korrekt ausgeführt wurde. Im Anschluss wird die Schrittweite in jedem Zeitschritt verändert, wobei die Schrittweitenverhältnisse σ_n nah bei 1 gewählt werden. Dieser Test soll damit die Konvergenzbeweise aus Kapitel 4 und Abschnitt 5.2 verifizieren. Laut den Konvergenzbeweisen sind für eine häufige Änderung der Schrittweite geringe Schrittweitenänderungen möglich, was praktisch nicht die Nutzung von variablen Schrittweiten rechtfertigt. Schnelle und große Änderungen sind praktisch sinnvoll, um besser auf signifikante Variationen in der Lösung reagieren zu können. Daher sollen schließlich auch andere Schrittweitenverhältnisse in den Verfahren verwendet werden, jedoch wird die Schrittweite für eine gewisse Anzahl an Zeitschritten nicht erneut verändert. Die Tests für das Generalized-α-Verfahren (4.11) erfolgen in Abschnitt 6.3.1 und jene für die BLieDF-Verfahren (5.35) in Abschnitt 6.3.2.

6.3.1 Generalized-α-Verfahren

Das Generalized-α-Verfahren (4.11) für variable Schrittweiten wurde bereits in [52] eingeführt. Dort wurden auch einige numerische Tests für den schweren Kreisel in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ durchgeführt. Unter anderem wurde der schwere Kreisel gelöst, wobei etwa in der Mitte des untersuchten Zeitintervalls die Schrittweite einmal verändert wurde. Dabei fiel auf, dass mit den beschriebenen Modifikationen aus den Bemerkungen 12 und 13 in allen Variablen eine Konvergenzordnung von zwei zu beobachten ist. Ohne die Modifikationen kam es in den Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ zu einer Ordnungsreduktion. Da diese Testergebnisse bereits vorliegen, soll der Test an dieser Stelle nicht wiederholt werden, auch wenn in [52] nur die Näherung 1 aus (4.9) verwendet wurde.
Im nachfolgenden Abschnitt wird daher untersucht, was bei häufiger Schrittweitenänderung mit Schrittweitenverhältnissen aus Tabelle 4.1 passiert. Im Anschluss werden auch andere Schrittweitenverhältnisse zugelassen, jedoch soll dann die Schritt-

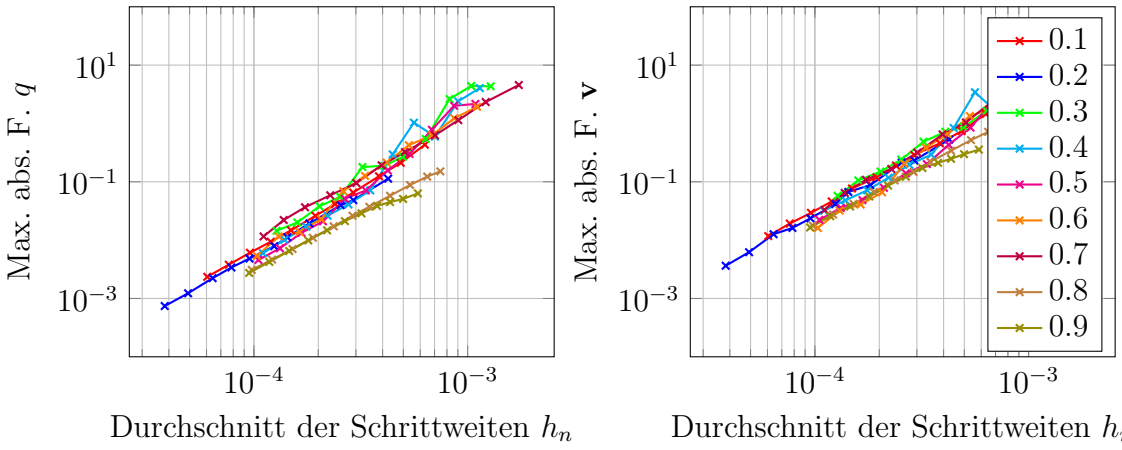

Abbildung 6.23: Maximaler absoluter Fehler von q (links) und $\mathbf{v}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten unter Verwendung der Näherung 1 aus (4.9) und Schrittweitenwechsel in jedem Zeitschritt für unterschiedliche ρ_∞

weite über mehrere Zeitschritte konstant bleiben.

Änderung der Schrittweite in jedem Zeitschritt

In Satz 3 wurde gezeigt, dass für gewisse Schrittweitenverhältnisse $\sigma_n = h_n/h_{n-1}$ die Nullstabilität des Generalized-α-Verfahrens (4.11) mit Modifikationen aus den Bemerkungen 12 und 13 für die Näherungen 1-4 aus (4.9) garantiert werden kann. Daher sollte unter Verwendung dieser Schrittweitenverhältnisse die Konvergenz zweiter Ordnung beobachtbar und keine Stabilitätsprobleme erkennbar sein. Dieser Aspekt wird in diesem Abschnitt untersucht.
Dazu wurden für jeden Spektralradius $\rho_\infty = 0.1, 0.2, ..., 0.9$ und jede Näherung 1-4 aus (4.9) jeweils Schrittweitenfolgen wie in Bemerkung 28 bestimmt und mit diesen das Benchmarkproblem schwerer Kreisel in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$

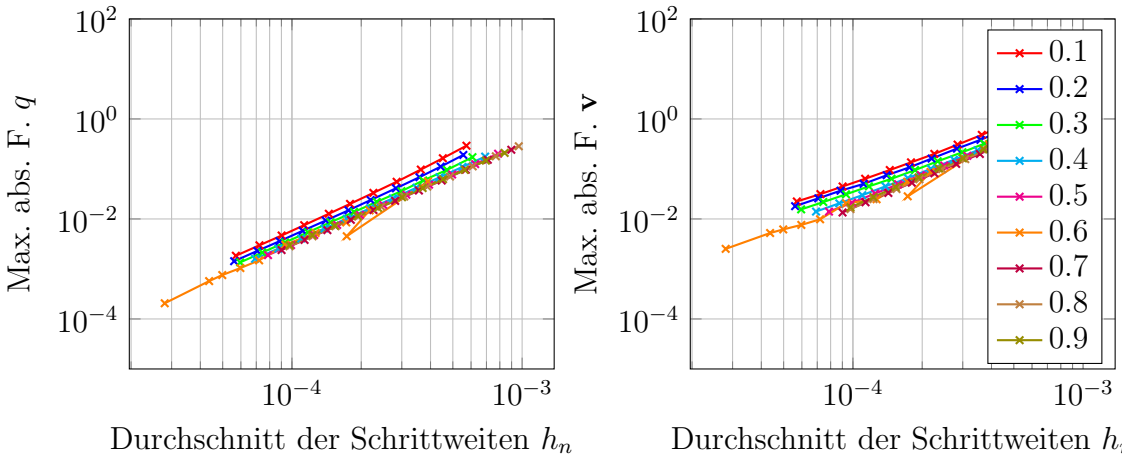

Abbildung 6.24: Maximaler absoluter Fehler von q (links) und $\mathbf{v}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten unter Verwendung der Näherung 2 aus (4.9) und Schrittweitenwechsel in jedem Zeitschritt für unterschiedliche ρ_∞

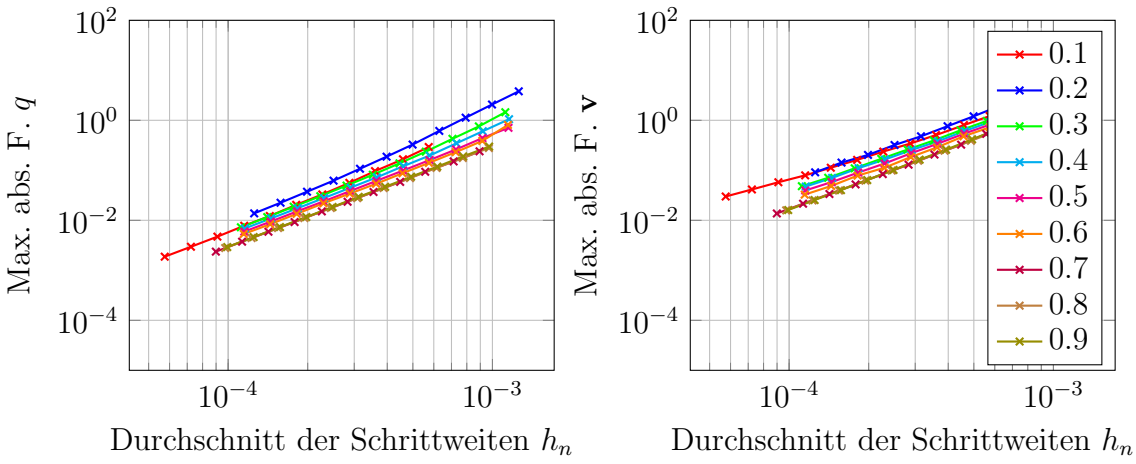

Abbildung 6.25: Maximaler absoluter Fehler von q (links) und $\mathbf{v}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten unter Verwendung der Näherung 3 aus (4.9) und Schrittweitenwechsel in jedem Zeitschritt für unterschiedliche ρ_∞

gelöst. In Abbildung 6.23 ist der maximale absolute Fehler der Konfigurationsvariablen q und der Geschwindigkeit $\mathbf{v}$ über dem Durchschnitt der verwendeten Schrittweiten in doppelt logarithmischer Skala unter Verwendung der Näherung 1 dargestellt. In den Abbildungen 6.24, 6.25 und 6.26 ist Gleiches für die Näherungen 2-4 aufgetragen. Dabei zeigt sich, dass in jedem Fall in den beiden Variablen q und $\mathbf{v}$ die Konvergenzordnung zwei ablesbar ist. Dies bestätigt die theoretischen Untersuchungen der Konvergenzanalyse aus Kapitel 4. Außerdem fällt auf, dass mit wachsendem ρ_∞ die berechnete Lösung genauer wird. Jedoch konnten bei größerem ρ_∞ nur geringere Schrittweitenänderungen vorgenommen werden (vgl. Satz 3), daher ist die Verwendung von z.B. $\rho_\infty = 0.9$ trotz der besseren Genauigkeit nicht zwangsläufig zu empfehlen, wenn variable Schrittweiten gewünscht sind. Würde derselbe numerische Test mit der Lie-Gruppen-Formulierung $SE(3)$ ausgeführt werden, so würden vergleichbare Ergebnisse folgen.
Nachdem nun gezeigt wurde, dass die Konvergenzordnung für alle untersuchten Näh-

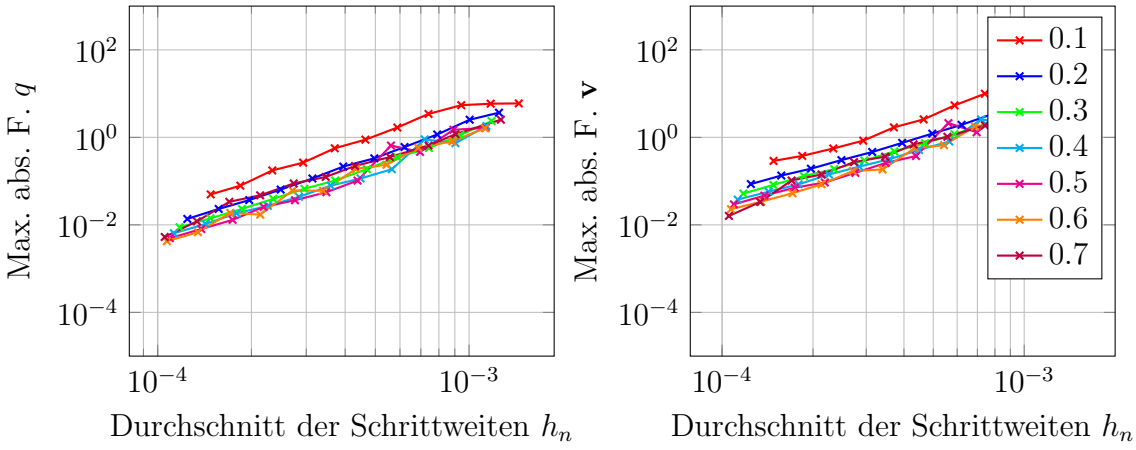

Abbildung 6.26: Maximaler absoluter Fehler von q (links) und $\mathbf{v}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten unter Verwendung der Näherung 4 aus (4.9) und Schrittweitenwechsel in jedem Zeitschritt für unterschiedliche ρ_∞

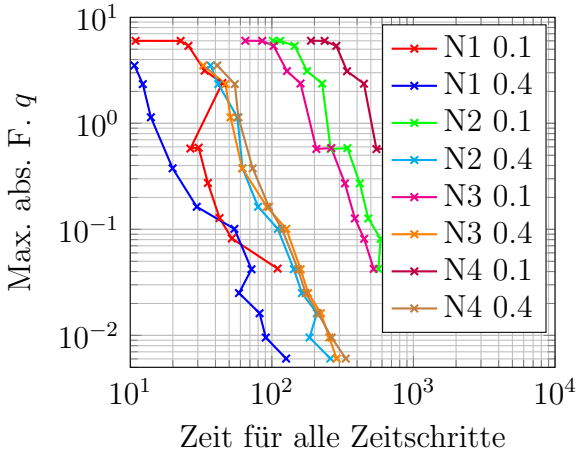

Abbildung 6.27: Maximaler absoluter Fehler von q für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten und Schrittweitenwechsel in jedem Zeitschritt für unterschiedliche Näherungen aus (4.9) und ρ_∞

erungen auch numerisch erhalten wird, ist ein Vergleich dieser Näherungen 1-4 aus (4.9) interessant. Hierbei ist zu beachten, dass je nach Näherung unterschiedliche Schrittweitenänderungen möglich sind. Um daher einen gerechteren Vergleich zu erhalten, werden Schrittweitenfolgen verwendet, die laut Satz 3 für alle Näherungen möglich sind. Für $\sigma_{\min}$ wird somit immer der größte Wert einer Zeile aus Tabelle 4.1 gewählt und für $\sigma_{\max}$ der kleinste. Außerdem wird der Test nur für $\rho_\infty = 0.1$ und $\rho_\infty = 0.4$ ausgeführt. In Abbildung 6.27 ist dazu die Zeit, die für alle Zeitschritte benötigt wird (vgl. Bemerkung 27), über den maximalen absoluten Fehler in q in doppelt logarithmischer Skala für die Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ aufgezeichnet. Dabei zeigt sich, dass die Näherung 1 aus (4.9) für beide Spektralradien im Unendlichen ρ_∞ am effizientesten zu Ergebnissen führt. Dies hängt damit zusammen, dass in Näherung 1 die Anpassungen nur die Beschleunigungen $\dot{\mathbf{v}}_n$ und $\mathbf{a}_n$ aus bereits berechneten Zeitschritten verwenden. In den restlichen Näherungen 2-4 aus (4.9) werden auch $\dot{\mathbf{v}}_{n+1}$ bzw. $\mathbf{a}_{n+1}$ verwendet, welche im Newton-Raphson-Verfahren (vgl. (3.60)) beachtet und in jedem Iterationsschritt ausgewertet werden müssen. Dies führt zu den großen Rechenzeitunterschieden in Abbildung 6.27. Für $\rho_\infty = 0.4$ liegen die Näherungen 2-4 sehr nah beieinander. Wird der Spektralradius $\rho_\infty = 0.1$ verwendet ist Näherung 2 am effizientesten und Näherung 4 berechnet am langsamsten die Lösungen. Für die Rechnungen ist stets abzuwägen, ob größere Schrittweitenänderungen nötig sind (wie in den Näherungen 2 und 3) oder eine effizientere Rechnung von Bedeutung ist (wie bei Näherung 1).

Änderung der Schrittweite nach einer gewissen Anzahl an Zeitschritten

Im vorherigen Abschnitt wurde gezeigt, dass das Generalized-α-Verfahren (4.11) für variable Schrittweiten mit Schrittweitenverhältnissen aus Tabelle 4.1 von zweiter Ordnung konvergiert. Die möglichen Schrittweitenverhältnisse aus diesem Abschnitt lassen jedoch nur sehr kleine Schrittweitenänderungen zu. Diese Schranken sind jedoch nur hinreichend. In praktischen Implementierungen wären zudem schnellere Schrittweitenänderungen von Vorteil. Daher soll in diesem Abschnitt überprüft werden, ob beispielhaft auch Schrittweitenverhältnisse aus den Intervallen $\sigma_n \in [0.5, 2]$ oder

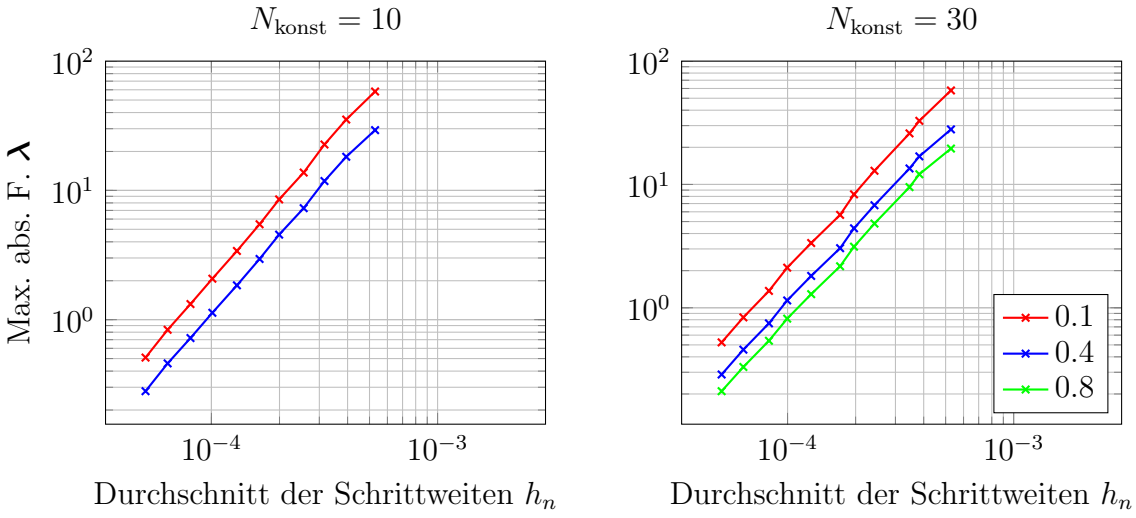

Abbildung 6.28: Maximaler absoluter Fehler von $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten unter Verwendung von Näherung 1 aus (4.9) und unterschiedlichen ρ_∞ mit $\sigma_n \in [0.5, 2]$ und Schrittweitenwechsel nach $N_{\text{konst}} = 10$ (links) und $N_{\text{konst}} = 30$ (rechts)

$\sigma_n \in [0.2, 5]$ verwendet werden könnten, solange die geänderte Schrittweite dann einige Zeitschritte konstant bleibt, bevor sie erneut verändert wird. Dazu werden Schrittweitenfolgen berechnet, bei der sich die Schrittweite erst nach einer festen Anzahl N_{konst} von Zeitschritten erneut ändert, wobei die Schrittweitenverhältnisse zufällig aus den entsprechenden Intervallen ausgewählt werden. Mit diesen Schrittweitenfolgen wird der schwere Kreisel in der Lie-Gruppen-Formulierung $\mathbb{R}^3 \times SO(3)$ gelöst. Da der Fehler bei gleicher Schrittweitenfolge für die vier Näherungen aus (4.9) sehr nah beieinander ist, wird nur Näherung 1 verwendet. Da dort nur kleine Schrittweitenänderungen möglich sind, sollten auch die anderen Näherungen keine Probleme zeigen, wenn die Rechnung für Näherung 1 funktioniert. Außerdem werden die Spektralradien im Unendlichen $\rho_\infty = 0.1, 0.4, 0.8$ verwendet.

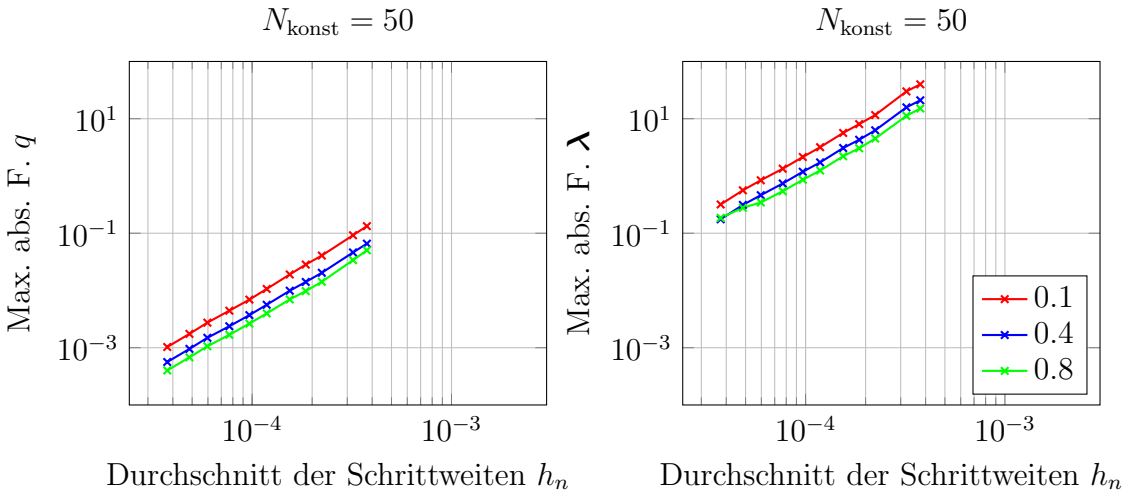

Abbildung 6.29: Maximaler absoluter Fehler von q (links) und $\boldsymbol{\lambda}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ mit dem Generalized-α-Verfahren (4.11) für variable Schrittweiten unter Verwendung von Näherung 1 aus (4.9) und unterschiedlichen ρ_∞ mit $\sigma_n \in [0.2, 5]$ und Schrittweitenwechsel nach $N_{\text{konst}} = 50$ Schritten

In Abbildung 6.28 ist der maximale absolute Fehler von $\boldsymbol{\lambda}$ über den Durchschnitt der verwendeten Schrittweiten h_n in doppelt logarithmischer Skala aufgezeichnet für $N_{\text{konst}} = 10$ (links) und $N_{\text{konst}} = 30$ (rechts) für $\sigma_n \in [0.5, 2]$. Für $\rho_\infty = 0.8$ und $N_{\text{konst}} = 10$ ist keine Gerade in der Abbildung zu erkennen. Das liegt daran, dass in diesem Fall Stabilitätsprobleme auftraten und dadurch der Fehler viel größer ist als der hier angegebene Bereich. Mit wachsendem N_{konst} kann jedoch auch für $\rho_\infty = 0.8$ eine Lösung erhalten werden, die mit zweiter Ordnung in $\boldsymbol{\lambda}$ konvergiert. In Abbildung 6.29 wurden $\sigma_n \in [0.2, 5]$ und $N_{\text{konst}} = 50$ gewählt und der maximale absolute Fehler in q und $\boldsymbol{\lambda}$ über den Durchschnitt der Schrittweiten h_n in doppelt logarithmischer Skala aufgezeichnet. Auch hier ist in allen Fällen die Konvergenz zweiter Ordnung ablesbar.
Dies zeigt, dass unter gewissen Voraussetzungen also auch Schrittweitenverhältnisse außerhalb der in Satz 3 gegebenen Intervalle verwendet werden können.

Fazit

In diesem Abschnitt konnte das Resultat aus Satz 5 für das Generalized-α-Verfahren mit variablen Schrittweiten (4.11) verifiziert werden. Es konnte zudem gezeigt werden, dass über das hinreichende Kriterium für die Schrittweitenverhältnisse aus Satz 3 hinausgegangen werden kann, im Speziellen wenn geeignete Einschränkungen an die Änderungen der Schrittweiten gewählt werden. In den obigen numerischen Tests war die Einschränkung dadurch gegeben, dass nach einer Änderung der Schrittweite diese eine gewisse Anzahl von Zeitschritten konstant gehalten wird. Die Testergebnisse zeigen somit, dass das Generalized-α-Verfahren (4.11) praktisch mit variablen Schrittweiten verwendet werden kann.

6.3.2 BLieDF-Verfahren

Die BLieDF-Verfahren (5.35) für variable Schrittweiten sollen nun numerisch getestet werden. Dazu wird im nachfolgenden Abschnitt die Schrittweite zunächst im gesamten Zeitintervall nur einmal verändert. Dabei soll die Notwendigkeit der Modifikation aus Bemerkung 23 gezeigt werden. Im Anschluss wird wie beim Generalized-α-Verfahren (4.11) untersucht, was bei häufiger Schrittweitenänderung mit Schrittweitenverhältnissen σ_n nah an 1 passiert. Schließlich werden auch andere Schrittweitenverhältnisse zugelassen, jedoch soll dabei die Schrittweite über mehrere Zeitschritte konstant bleiben.

Einmalige Änderung der Schrittweite

In diesem Abschnitt sollen die Bewegungsgleichungen des schweren Kreisels in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ durch die BLieDF-Verfahren (5.35) gelöst werden, wobei zunächst eine Schrittweite h_0 verwendet wird und ab einem Zeitpunkt von $t = 1$ mit $2h_0$ weiter gerechnet wird. In Abbildung 6.30 ist dazu der maximale absolute Fehler von q über die durchschnittliche Schrittweite h_n in doppelt logarithmischer Skala angegeben. In beiden Lie-Gruppen-Formulierungen ist sowohl mit als auch ohne die Modifizierung aus Bemerkung 23 die Konvergenzordnung $p = k$ für $k = 2$ und $k = 3$ zu beobachten.
Diese Feststellung ändert sich jedoch, wenn in Abbildung 6.31 der maximale absolute Fehler von $\boldsymbol{\lambda}$ betrachtet wird. Dort ist bei der Verwendung der Lie-Gruppen-

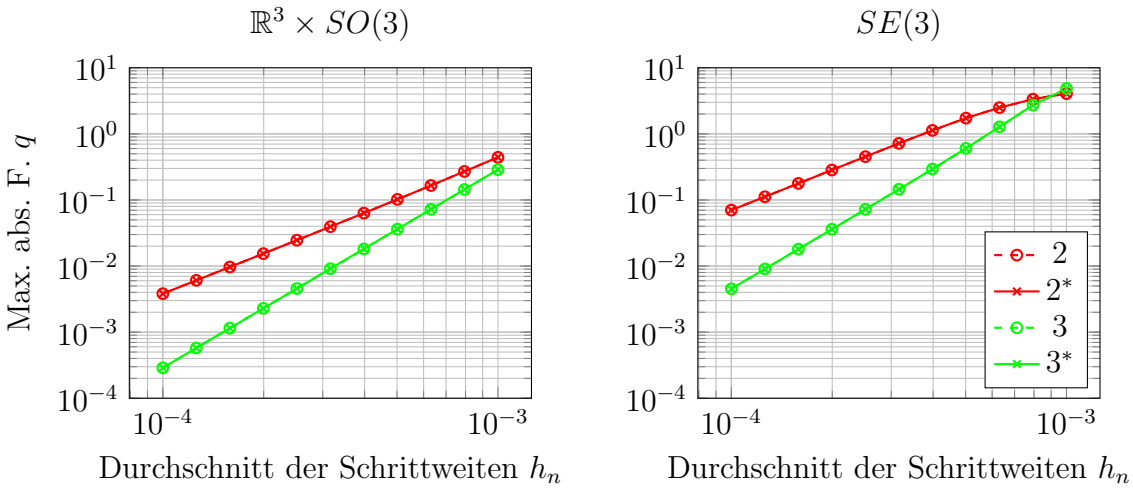

Abbildung 6.30: Maximaler absoluter Fehler von q für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ (links) und $SE(3)$ (rechts) für $k = 2$ und $k = 3$ bei einmaliger Schrittweitenänderung mit der Modifizierung aus Bemerkung 23 (k mit *) und ohne diese (k ohne *)

Formulierung $\mathbb{R}^3 \times SO(3)$ eine Ordnungsreduktion zu verzeichnen, wenn die Modifikation aus Bemerkung 23 nicht angewendet wird. Ähnlich zu den Beobachtungen zu den Anpassungen der Startwerte aus Abschnitt 6.2.1 ist für die Lie-Gruppen-Formulierung $SE(3)$ solch eine Ordnungsreduktion nicht erkennbar.

Änderung der Schrittweite in jedem Zeitschritt

In Voraussetzung 8 wurde davon ausgegangen, dass es für gewisse Schrittweitenverhältnisse $\sigma_n = h_n/h_{n-1}$ Schranken $\sigma_{\min}$ und $\sigma_{\max}$ gibt, die die Nullstabilität der BLieDF-Verfahren (5.35) garantieren. Daher sollte unter Verwendung von Schrittweitenverhältnissen, die nah an 1 liegen, die Konvergenz der Ordnung $p = k$ für $k = 2$

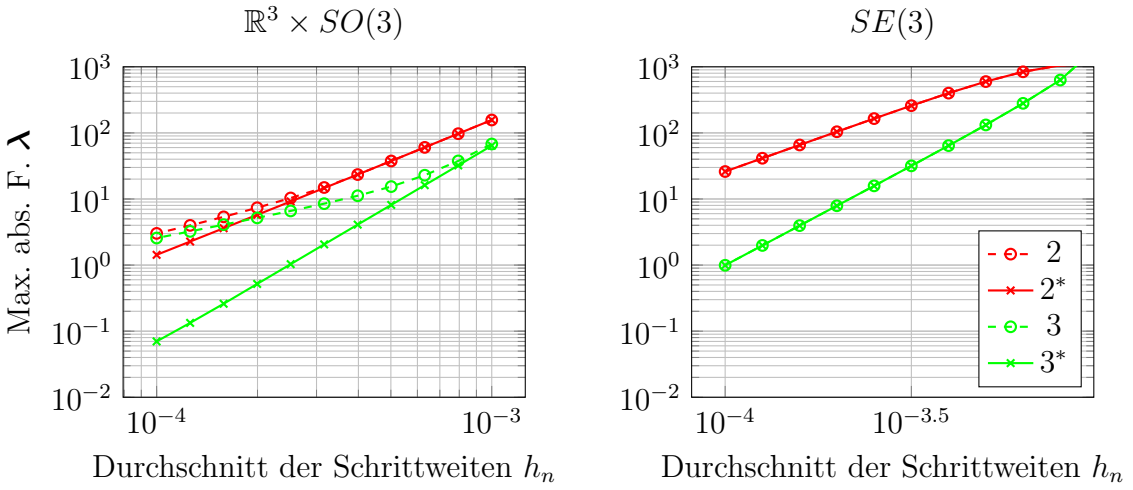

Abbildung 6.31: Maximaler absoluter Fehler von $\boldsymbol{\lambda}$ für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ (links) und $SE(3)$ (rechts) für $k = 2$ und $k = 3$ bei einmaliger Schrittweitenänderung mit der Modifizierung aus Bemerkung 23 (k mit *) und ohne diese (k ohne *)

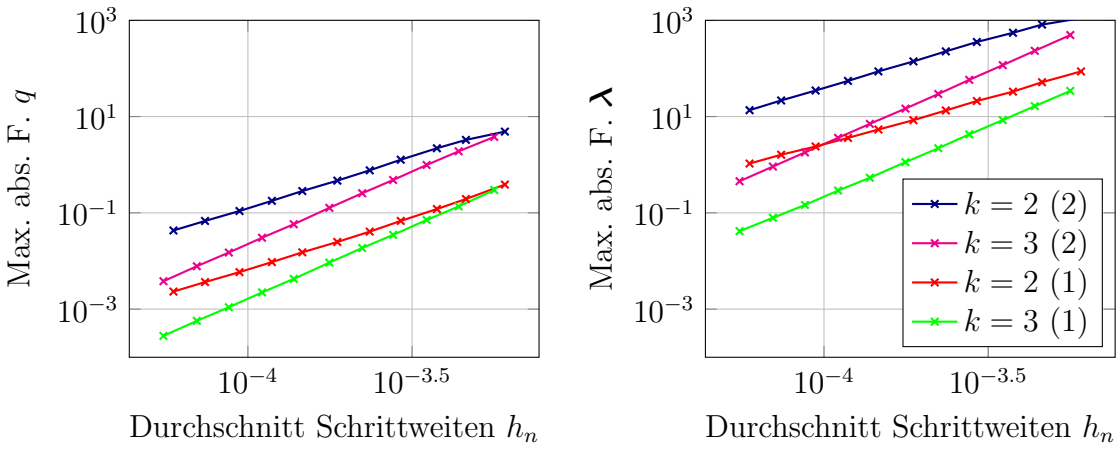

Abbildung 6.32: Maximaler absoluter Fehler von q (links) und $\boldsymbol{\lambda}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ (1) und $SE(3)$ (2) für $k = 2$ und $k = 3$

und $k = 3$ beobachtbar und keine Stabilitätsprobleme erkennbar sein. Dies soll in diesem Abschnitt untersucht werden.
Dazu wurden für $k = 2$ und $k = 3$ jeweils Schrittweitenfolgen wie in Bemerkung 28 bestimmt und mit diesen das Benchmarkproblem schwerer Kreisel in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ gelöst. In Abbildung 6.32 ist der maximale absolute Fehler der Konfigurationsvariable q (links) und der Lagrange-Multiplikatoren $\boldsymbol{\lambda}$ (rechts) über dem Durchschnitt der verwendeten Schrittweiten in doppelt logarithmischer Skala dargestellt. Dabei zeigt sich, dass in jedem Fall in den beiden Variablen q und $\boldsymbol{\lambda}$ die Konvergenzordnung $p = k$ ablesbar ist. Dies bestätigt die theoretischen Untersuchungen der Konvergenzanalyse aus Abschnitt 5.2.

Änderung der Schrittweite erst nach einer gewissen Anzahl an Zeitschritten

Im vorherigen Abschnitt konnte gezeigt werden, dass die BLieDF-Verfahren (5.35) für variable Schrittweiten mit Schrittweitenverhältnissen nah bei 1 mit der Ordnung $p = k$ konvergieren. In praktischen Implementierungen wären jedoch größere Schrittweitenänderungen von Vorteil. Daher soll in diesem Abschnitt überprüft werden, ob beispielhaft auch Schrittweitenverhältnisse aus dem Intervall $\sigma_n \in [0.5, 2]$ verwendet werden könnten, solange die geänderte Schrittweite dann einige Zeitschritte konstant bleibt, bevor sie erneut verändert wird.
Dazu werden Schrittweitenfolgen berechnet, bei denen sich die Schrittweite erst nach 10 Zeitschritten erneut ändert, vgl. Bemerkung 28, wobei die Schrittweitenverhältnisse zufällig aus dem Intervall $\sigma_n \in [0.5, 2]$ ausgewählt werden. Mit diesen Schrittweitenfolgen wird der schwere Kreisel in den Lie-Gruppen-Formulierungen $\mathbb{R}^3 \times SO(3)$ und $SE(3)$ gelöst.
In Abbildung 6.33 ist der maximale absolute Fehler von q und $\boldsymbol{\lambda}$ über den Durchschnitt der verwendeten Schrittweiten h_n in doppelt logarithmischer Skala aufgezeichnet. Dabei ist sowohl in q als auch in $\boldsymbol{\lambda}$ die Konvergenz der Ordnung $p = k$ für $k = 2$ und $k = 3$ in beiden Lie-Gruppen-Formulierungen zu beobachten. Dies zeigt, dass unter gewissen Voraussetzungen auch größere Schrittweitenfolgen verwendet werden können.

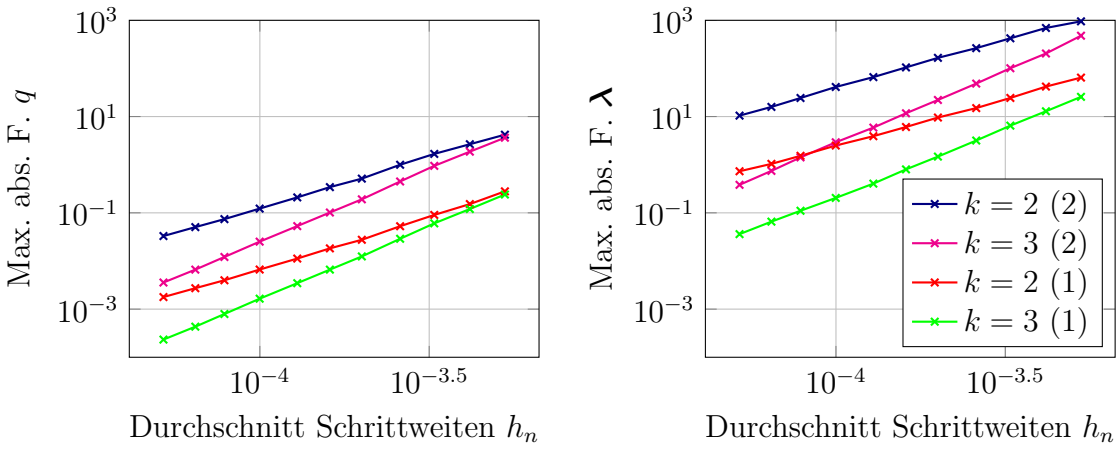

Abbildung 6.33: Maximaler absoluter Fehler von q (links) und $\boldsymbol{\lambda}$ (rechts) für den schweren Kreisel in $\mathbb{R}^3 \times SO(3)$ (1) und $SE(3)$ (2) für $k = 2$ und $k = 3$ für Schrittweitenänderungen erst nach 10 Zeitschritten

Fazit

In diesem Abschnitt konnte das Resultat aus Satz 15 für das BLieDF-Verfahren mit variablen Schrittweiten (5.35) für $k = 2, 3$ verifiziert werden. Durch die Modifikation aus Abschnitt 5.2 konnte praktisch die Vermeidung einer Ordnungsreduktion bestätigt werden. Zudem wurde gezeigt, dass bei geeigneten Einschränkungen größere Schrittweitenverhältnisse verwendet werden können. In den obigen numerischen Tests war die Einschränkung dadurch gegeben, dass nach einer Änderung der Schrittweite diese nur eine gewisse Anzahl von Zeitschritten konstant gehalten wird. Die Testergebnisse zeigen somit, dass das BLieDF-Verfahren (5.35) praktisch mit variablen Schrittweiten verwendet werden kann.

Kapitel 7

Zusammenfassung

In der vorliegenden Arbeit wurden Zeitintegrationsverfahren zur Lösung der Bewegungsgleichungen von Mehrkörpersystemen ohne Zwangsbedingungen und von beschränkten Mehrkörpersystemen für Konfigurationsräume mit Lie-Gruppen-Struktur vorgestellt. Dabei wurden das Generalized-α-Verfahren und lineare implizite Mehrschrittverfahren fokussiert. Das Generalized-α-Verfahren ist ein aus der Literatur bekanntes Verfahren zur Lösung von mechanischen Mehrkörpersystemen auch in der Lie-Gruppen-Formulierung. Für die linearen Mehrschrittverfahren wurden Ideen von bereits in der Literatur bekannten Lie-Gruppen-Verfahren wie die Crouch-Grossman-Verfahren und die kommutatorfreien Lie-Gruppen-Verfahren aufgegriffen und verwendet, um Adams-Moulton-Verfahren für Mehrkörpersystemmodelle ohne Zwangsbedingungen in Lie-Gruppen-Formulierung herzuleiten. Außerdem wurden BDF-Verfahren näher untersucht. Dazu wurde zum einen ein bekanntes Lie-Gruppen-Mehrschrittverfahren, das auf Faltinsen et al. [20] zurückgeht, vorgestellt und speziell in der BDF-Variante untersucht. Zum anderen wurde das neue BLieDF-Verfahren vorgestellt. Bei diesem Verfahren wird die Nichtlinearität des Konfigurationsraums durch einen speziell eingeführten Korrekturterm berücksichtigt. Beide Verfahren wurden auf Mehrkörpersysteme ohne Zwangsbedingungen und auf beschränkte Mehrkörpersysteme angewendet. Außerdem konnte ausgehend von dem bereits bekannten Konvergenzbeweis des Generalized-α-Verfahrens für beschränkte mechanische Systeme die Konvergenz der Ordnung $p = k$ für $2 \leq k \leq 6$ für die beiden BDF-Lie-Gruppen-Verfahren bewiesen werden.
Die Anwendung der Verfahren mit variablen Schrittweiten erfolgte lediglich für das Generalized-α-Verfahren und die BLieDF-Verfahren. Dabei sind einige Modifikationen, wie z.B. die Anpassung der Geschwindigkeit vor jedem Zeitschritt, notwendig, um dieselbe Konvergenzordnung wie im Fall konstanter Schrittweiten zu erhalten. Die Konvergenz dieser beiden Verfahren konnte ebenfalls bewiesen werden.
Die Arbeit schloss mit einigen numerischen Tests, in denen die theoretischen Konvergenzresultate verifiziert und Vergleiche zwischen den Verfahren gezogen wurden. Alles in Allem zeigte sich, dass die neuen BLieDF-Verfahren mit den bereits bekannten Lie-Gruppen-Mehrschrittverfahren mithalten konnten und somit viel Potenzial für zukünftige Implementierungen aufweist. In weiterführenden Arbeiten wäre eine Verwendung der BLieDF-Verfahren für umfangreichere Testprobleme durchaus interessant. So könnte überprüft werden, ob sich die positiven Ergebnisse dieser Arbeit bestätigen.

Anhang A

Differentialgleichungen erster Ordnung und weitere wichtige Begriffe zu linearen Mehrschrittverfahren

In diesem Anhang sollen grundlegende Begriffe für Differentialgleichungen erster Ordnung definiert und erklärt werden. Alle diese Begriffe lassen sich auch auf die zu betrachtenden Systeme zweiter Ordnung (2.3) anwenden.

Definition A.1 (Differentialgleichung erster Ordnung [46])
Ein System *gewöhnlicher Differentialgleichungen (ODEs) 1. Ordnung* ist von der Form

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}) \tag{A.1}$$

mit einer gegebenen Funktion $\mathbf{f} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$. Eine Funktion $\mathbf{x}(t)$ wird *Lösung* der Differentialgleichung (A.1) genannt, falls für alle $t \in \mathbb{R}$ der Zusammenhang

$$\mathbf{x}'(t) = \mathbf{f}\big(t, \mathbf{x}(t)\big) \tag{A.2a}$$

gilt. Ist die Funktion $\mathbf{f}$ lipschitzstetig, so ist die Lösung $\mathbf{x}(t)$ eindeutig durch den *Anfangswert*

$$\mathbf{x}(t_0) = \mathbf{x}_0 \tag{A.2b}$$

mit $t_0 \in \mathbb{R}$ festgelegt. Ein *Anfangswertproblem* ist die Suche nach einer Lösung $\mathbf{x}(t)$ von (A.2a) mit den Anfangswerten (A.2b).

In der Literatur sind zwei Bereiche von Methoden besonders von Interesse: die Einschritt- und die linearen Mehrschrittverfahren. In dieser Arbeit werden vor allem Mehrschrittverfahren untersucht und deshalb auch nur die Begriffe für ebendiese genauer vorgestellt.

Definition A.2 (Lineare Mehrschrittverfahren [28])
Ein *lineares Mehrschrittverfahren* mit k Schritten zur Lösung von (A.2) auf dem Zeitintervall $t \in [t_0, t_{N_{\text{end}}}]$ hat die Gestalt

$$\frac{1}{h}\sum_{i=0}^{k} \alpha_i \mathbf{x}_{n+1-i} = \sum_{i=0}^{k} \beta_i \mathbf{f}(t_{n+1-i}, \mathbf{x}_{n+1-i}), \quad (n = 0, ..., N_{\text{end}} - k), \tag{A.3a}$$

$$\mathbf{x}(t_i) = \mathbf{x}_i, \quad (i = 0, ..., k-1), \tag{A.3b}$$

mit Parametern $\alpha_i, \beta_i \in \mathbb{R}$, $(i = 0, ..., k)$, $\alpha_0 \neq 0$, $|\alpha_k| + |\beta_k| \neq 0$ und der konstanten Schrittweite $h \coloneqq t_{n+1} - t_n$. Das Verfahren (A.3a) heißt *explizit*, wenn $\beta_0 = 0$ ist, und ansonsten *implizit*.

Ein Zeitintegrationsverfahren soll die Lösung der Differentialgleichung möglichst genau approximieren. Dazu ist zunächst wichtig, welchen Fehler die Verfahren nach einem Schritt im Vergleich zur analytischen Lösung erzeugt haben. Um diesen Fehler qualitativ zu beschreiben, wird der lokale Abbruchfehler verwendet. Dieser kann durch das Einsetzen der analytischen Lösung in das numerische Verfahren erhalten werden.

Definition A.3 (Lokaler Abbruchfehler [25, 47])
Der *lokale Abbruch- oder Diskretisierungsfehler* eines linearen Mehrschrittverfahrens (A.3a) ist definiert durch

$$\mathbf{le}_n^{\mathbf{x}} \coloneqq \frac{1}{h} \sum_{i=0}^{k} \alpha_i \mathbf{x}(t_{n+1-i}) - \sum_{i=0}^{k} \beta_i \mathbf{f}\big(t_{n+1-i}, \mathbf{x}(t_{n+1-i})\big). \tag{A.4}$$

Definition A.4 (Konsistenz [46])
Ein lineares Mehrschrittverfahren (A.3a) ist *konsistent*, wenn für jedes Anfangswertproblem (A.2) gilt

$$\lim_{h \to 0} \frac{\|\mathbf{le}_n^{\mathbf{x}}\|}{h} = 0 \quad \text{für} \quad n = 0, ..., N_{\text{end}} - k.$$

Es hat die *Konsistenzordnung* $p \in \mathbb{N}$, wenn für den lokalen Abbruchfehler (A.4) und die genügend oft stetig differenzierbare Funktion $\mathbf{f}$ gilt

$$\|\mathbf{le}_n^{\mathbf{x}}\| \leq C_{\text{le}} h^{p+1} \quad \text{für alle} \quad h \in (0, \overline{h}] \quad \text{und} \quad n = 0, ..., N_{\text{end}} - k,$$

mit von h unabhängigen Konstanten $C_{\text{le}} > 0$ und $\overline{h} > 0$.

Der lokale Abbruchfehler ist ein Maß für den Fehler des Verfahrens in einem einzelnen Schritt. Praktisch ist jedoch das Langzeitverhalten eines Verfahrens von Interesse. Dazu wird der globale Fehler und die Konvergenzordnung eingeführt. Einen Zusammenhang zwischen Konsistenz- und Konvergenzordnung für lineare Mehrschrittverfahren liefert der nachfolgende Satz. Dabei müssen jedoch zusätzlich die Startwerte des Verfahrens berücksichtigt werden.

Definition A.5 (Konvergenz von linearen Mehrschrittverfahren [46])
Ein lineares Mehrschrittverfahren (A.3a) heißt *konvergent* im Zeitintervall $[t_0, t_{N_{\text{end}}}]$, wenn für alle Anfangswertprobleme (A.2) mit stetig differenzierbarem $\mathbf{f}$ und für alle Startwerte $\mathbf{x}(t_i) = \mathbf{x}_i$, $(i = 0, ..., k-1)$, mit

$$\|\mathbf{x}(t_i) - \mathbf{x}_i\| \to 0 \quad \text{für} \quad h \to 0$$

gilt

$$\|\mathbf{e}_n^{\mathbf{x}}\| \to 0, \quad h \to 0, \quad (n = 0, ..., N_{\text{end}} - k).$$

Das Mehrschrittverfahren (A.3a) besitzt die Konvergenzordnung p, wenn für alle Anfangswertprobleme (A.2) mit p-mal stetig differenzierbarem $\mathbf{f}$ und für alle Startwerte mit

$$\|\mathbf{x}(t_i) - \mathbf{x}_i\| \leq C_S h^p, \quad h \in (0, \overline{h}], \tag{A.5}$$

gilt

$$\|\mathbf{e}_n^{\mathbf{x}}\| \leq C_G h^p, \quad h \in (0, \overline{h}].$$

Der globale Fehler $\mathbf{e}_n^{\mathbf{x}}$ ist dabei definiert durch

$$\mathbf{e}_n^{\mathbf{x}} \coloneqq \mathbf{x}(t_n) - \mathbf{x}_n. \tag{A.6}$$

Ein notwendiges Kriterium für die Konsistenz und Konvergenz von linearen Mehrschrittverfahren (A.3a) ist durch den nachfolgenden Satz gegeben. Die darin enthaltenen Bedingungen werden auch Konsistenzbedingungen genannt.

Satz A.1 ([28])
Ein lineares Mehrschrittverfahren (A.3a) mit k Schritten konvergiert nur dann mit Ordnung p, wenn die Konsistenzbedingungen

$$\sum_{i=0}^{k} \alpha_i = 0, \tag{A.7a}$$

$$\sum_{i=0}^{k} \alpha_i (k-i)^\ell = \ell \sum_{i=0}^{k} \beta_i (k-i)^{\ell-1}, \quad (\ell = 1, ..., p), \tag{A.7b}$$

erfüllt sind.

Würden Einschrittverfahren untersucht werden, so würde die Konsistenzordnung p direkt auch eine Konvergenzordnung von p zur Folge haben [46]. Für lineare Mehrschrittverfahren wird jedoch die zusätzliche Bedingung der Nullstabilität benötigt. Durch diese Nullstabilität eines Verfahrens wird erreicht, dass die Lösungen nicht unbeschränkt wachsen können.

Definition 27 (Nullstabilität [46])
Ein lineares Mehrschrittverfahren (A.3a) heißt *nullstabil*, wenn alle Lösungen der homogenen Differenzengleichung

$$\sum_{i=0}^{k} \alpha_i \mathbf{x}_{n+1-i} = \mathbf{0}$$

beschränkt bleiben.

Satz A.2 ([46])
Sei $\mathbf{f}$ hinreichend oft stetig differenzierbar auf dem Intervall $[t_0, t_{N_{\text{end}}}]$ und sei das lineare Mehrschrittverfahren (A.3a) nullstabil und konsistent mit der Ordnung p, dann besitzt es auch die Konvergenzordnung p.

Anhang B

Lösung der kinematischen Gleichung in Lie-Gruppen-Formulierung unter Verwendung der Linkstranslation

In diesem Anhang soll die Lösung der kinematischen Gleichung (2.25) analog zu [27] berechnet werden. In [27] wurde anstelle der Linkstranslation die Rechtstranslation verwendet. Dazu werden zunächst vier Lemmata benötigt.

Lemma B.1
Für den adjungierten Operator gilt

$$\mathrm{ad}^i_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = (-1)^i \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})$$

für alle $\widetilde{\mathbf{u}}, \widetilde{\mathbf{w}} \in \mathfrak{g}$ und $i \in \mathbb{N}$.

Beweis:
Der Beweis wird durch vollständige Induktion geführt.

IA: Ist $i = 0$, so folgt

$$\mathrm{ad}^0_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = \widetilde{\mathbf{u}} = (-1)^0 \mathrm{ad}^0_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}).$$

IV: Die Behauptung gelte für ein $i \in \mathbb{N}$.

IS: Die Behauptung ist auch für $i+1$ erfüllt, da

$$\begin{aligned}\mathrm{ad}^{i+1}_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) &= \left[-\widetilde{\mathbf{w}}, \mathrm{ad}^i_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\right] \overset{IV}{=} -\left[\widetilde{\mathbf{w}}, (-1)^i \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\right] = (-1)^{i+1}\left[\widetilde{\mathbf{w}}, \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\right] \\ &= (-1)^{i+1} \mathrm{ad}^{i+1}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\end{aligned}$$

gilt.

■

Lemma B.2
Die Identität

$$\widetilde{\mathbf{w}}^k \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = \sum_{j=0}^{k} \binom{k}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j}$$

ist für alle $\widetilde{\mathbf{u}}, \widetilde{\mathbf{w}} \in \mathfrak{g}$ und $i, k \in \mathbb{N}$ erfüllt.

Beweis:
Der Beweis erfolgt durch vollständige Induktion nach k.

IA: Für $k = 0$ ist

$$\widetilde{\mathbf{w}}^0 \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = \sum_{j=0}^{0} \binom{0}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{0-j}$$

und für $k = 1$

$$\begin{aligned}
\widetilde{\mathbf{w}}^1 \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) &= \widetilde{\mathbf{w}} \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) - \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}} + \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}} \\
&= \left[\widetilde{\mathbf{w}}, \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\right] + \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}} \\
&= \binom{1}{0} \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^1 + \binom{1}{1} \mathrm{ad}^{i+1}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^0 \\
&= \sum_{j=0}^{1} \binom{1}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{1-j}
\end{aligned}$$

für alle $i \in \mathbb{N}$ erfüllt.

IV: Die Behauptung gelte für ein $k \in \mathbb{N}$ und alle $i \in \mathbb{N}$.

IS: Die Behauptung ist auch für $k + 1$ erfüllt, da

$$\begin{aligned}
\widetilde{\mathbf{w}}^{k+1} \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = \widetilde{\mathbf{w}} \widetilde{\mathbf{w}}^k \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) &\overset{IV}{=} \widetilde{\mathbf{w}} \sum_{j=0}^{k} \binom{k}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j} \\
&= \sum_{j=0}^{k} \binom{k}{j} \widetilde{\mathbf{w}} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j} \\
&\overset{IV}{=} \sum_{j=0}^{k} \binom{k}{j} \left(\mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}} + \mathrm{ad}^{i+j+1}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\right) \widetilde{\mathbf{w}}^{k-j} \\
&= \sum_{j=0}^{k} \binom{k}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j+1} + \sum_{j=0}^{k} \binom{k}{j} \mathrm{ad}^{i+j+1}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j} \\
&= \mathrm{ad}^i_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k+1} + \sum_{j=1}^{k} \binom{k}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j+1} \\
&\quad + \sum_{j=1}^{k} \binom{k}{j-1} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j+1} + \mathrm{ad}^{i+k+1}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \\
&= \sum_{j=0}^{k+1} \binom{k+1}{j} \mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \widetilde{\mathbf{w}}^{k-j+1}
\end{aligned}$$

für alle $i \in \mathbb{N}$ folgt.

■

Lemma B.3 (vgl. [27, Abschnitt III.4.1])
Die Ableitung erfüllt

$$\left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{w}}} \widetilde{\mathbf{w}}^k\right) \widetilde{\mathbf{u}} = \sum_{i=0}^{k-1} \binom{k}{i+1} \widetilde{\mathbf{w}}^{k-i-1} \mathrm{ad}^i_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})$$

für alle $\widetilde{\mathbf{u}}, \widetilde{\mathbf{w}} \in \mathfrak{g}$.

Beweis:
Es gilt mit den Lemmata B.1 und B.2

$$\begin{aligned}
&\sum_{i=0}^{k-1}\binom{k}{i+1}\widetilde{\mathbf{w}}^{k-i-1}\mathrm{ad}^{i}_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\\
&=\sum_{i=0}^{k-1}\binom{k}{i+1}\widetilde{\mathbf{w}}^{k-i-1}(-1)^{i}\mathrm{ad}^{i}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\\
&=\sum_{i=0}^{k-1}(-1)^{i}\binom{k}{i+1}\sum_{j=0}^{k-i-1}\binom{k-i-1}{j}\mathrm{ad}^{i+j}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\widetilde{\mathbf{w}}^{k-i-j-1}.
\end{aligned}$$

Durch die Änderung des Indexes in der zweiten Summe mit $\ell = i + j$ folgt

$$\begin{aligned}
&\sum_{i=0}^{k-1}\binom{k}{i+1}\widetilde{\mathbf{w}}^{k-i-1}\mathrm{ad}^{i}_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\\
&=\sum_{i=0}^{k-1}(-1)^{i}\binom{k}{i+1}\sum_{\ell=i}^{k-1}\binom{k-i-1}{\ell-i}\mathrm{ad}^{\ell}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\widetilde{\mathbf{w}}^{k-\ell-1}\\
&=\sum_{\ell=0}^{k-1}\mathrm{ad}^{\ell}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\widetilde{\mathbf{w}}^{k-1-\ell}\sum_{i=0}^{\ell}(-1)^{i}\binom{k}{i+1}\binom{k-i-1}{\ell-i}
\end{aligned}$$

unter Vertauschung der Summen. Das Umschreiben von

$$\begin{aligned}
\sum_{i=0}^{\ell}(-1)^{i}\binom{k}{i+1}\binom{k-i-1}{\ell-i} &= \binom{k}{\ell+1}\sum_{i=0}^{\ell}(-1)^{i}\binom{l+1}{i+1}\\
&= -\binom{k}{\ell+1}\sum_{i=1}^{\ell+1}(-1)^{i}\binom{l+1}{i} = \binom{k}{\ell+1}
\end{aligned}$$

wegen

$$\binom{k}{i+1}\binom{k-i-1}{\ell-i} = \binom{k}{\ell+1}\binom{l+1}{i+1}, \quad \text{für alle } k \geq l+1 \geq i+1 \geq 0,$$

und

$$\sum_{i=0}^{\ell+1}(-1)^{i}\binom{\ell+1}{i} = 0, \quad \ell \geq 0 \in \mathbb{N},$$

vgl. [26], liefert die Behauptung, da nach [27, Section III.4.1] die Gleichung

$$\left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{w}}}\widetilde{\mathbf{w}}^{k}\right)\widetilde{\mathbf{u}} = \sum_{\ell=0}^{k-1}\binom{k}{\ell+1}\mathrm{ad}^{\ell}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\widetilde{\mathbf{w}}^{k-\ell-1}$$

gültig ist.

■

Lemma B.4 (vgl. [27, Lemma 4.1])
Die Ableitung der Exponentialabbildung (2.26) ist gegeben durch

$$\left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{w}}}\exp(\widetilde{\mathbf{w}})\right)\widetilde{\mathbf{u}} = \exp(\widetilde{\mathbf{w}})\left(\mathrm{dexp}_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})\right)$$

mit $\widetilde{\mathbf{u}}, \widetilde{\mathbf{w}} \in \mathfrak{g}$ und

$$\operatorname{dexp}_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) = \sum_{k\geq 0} \frac{1}{(k+1)!} \operatorname{ad}^k_{\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}).$$

Beweis:
Der Beweis ist analog zu [27, Lemma III.4.1]. Es gilt mit Gleichung (2.26) und Lemma B.3

$$\begin{aligned}
\left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{w}}} \exp(\widetilde{\mathbf{w}})\right) \widetilde{\mathbf{u}} &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{w}}} \widetilde{\mathbf{w}}^k\right) \widetilde{\mathbf{u}} \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i=0}^{k-1} \binom{k}{i+1} \widetilde{\mathbf{w}}^{k-i-1} \operatorname{ad}^i_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \\
&= \sum_{i=0}^{\infty} \left(\sum_{k=i+1}^{\infty} \frac{1}{(i+1)!(k-i-1)!} \widetilde{\mathbf{w}}^{k-i-1} \right) \operatorname{ad}^i_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}).
\end{aligned}$$

Wird der Index in der zweiten Summe durch $\ell = k - i - 1$ ersetzt, so ist

$$\begin{aligned}
\left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\mathbf{w}}} \exp(\widetilde{\mathbf{w}})\right) \widetilde{\mathbf{u}} &= \sum_{i=0}^{\infty} \left(\sum_{\ell=0}^{\infty} \frac{1}{(i+1)!\ell!} \widetilde{\mathbf{w}}^{\ell} \right) \operatorname{ad}^i_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}}) \\
&= \exp(\widetilde{\mathbf{w}}) \operatorname{dexp}_{-\widetilde{\mathbf{w}}}(\widetilde{\mathbf{u}})
\end{aligned}$$

erfüllt.

■

Lemma B.5 (vgl. [27, Theorem IV.7.1])
Die Lösung der Differentialgleichung (2.25) mit $q(t_m) = q_m$ ist in einer Umgebung von t_m gegeben durch

$$q(t) = q_m \circ \exp\big(\widetilde{\boldsymbol{\nu}}_m(t)\big), \tag{B.1}$$

wobei $\widetilde{\boldsymbol{\nu}}_m(t)$ der Differentialgleichung

$$\dot{\widetilde{\boldsymbol{\nu}}}_m(t) = \operatorname{dexp}^{-1}_{-\widetilde{\boldsymbol{\nu}}_m(t)}\big(\widetilde{\mathbf{v}}(t)\big)$$

genügt.

Beweis:
Der Beweis erfolgt analog zu [27, Theorem IV.7.1]. Durch das Ableiten von (B.1) nach t folgt

$$\begin{aligned}
\dot{q}(t) &= DL_{q_m}(e) \cdot \left(\left(\frac{\mathrm{d}}{\mathrm{d}\widetilde{\boldsymbol{\nu}}_m} \exp\big(\widetilde{\boldsymbol{\nu}}_m(t)\big) \right) \cdot \dot{\widetilde{\boldsymbol{\nu}}}_m(t) \right) \\
&= DL_{q_m}(e) \cdot \Big(\exp\big(\widetilde{\boldsymbol{\nu}}_m(t)\big) \operatorname{dexp}_{-\widetilde{\boldsymbol{\nu}}_m(t)}\big(\dot{\widetilde{\boldsymbol{\nu}}}_m(t)\big) \Big)
\end{aligned} \tag{B.2}$$

mit Lemma B.4. Weiterhin gilt durch Einsetzen von (B.1) in (2.25)

$$\dot{q}(t) = DL_{q_m \circ \exp(\widetilde{\boldsymbol{\nu}}_m(t))}(e) \cdot \widetilde{\mathbf{v}}(t). \tag{B.3}$$

Das Gleichsetzen von (B.2) und (B.3) ergibt

$$\widetilde{\mathbf{v}}(t) = \operatorname{dexp}_{-\widetilde{\boldsymbol{\nu}}_m(t)}\big(\dot{\widetilde{\boldsymbol{\nu}}}_m(t)\big)$$

und die Behauptung folgt durch Anwendung des inversen Operators $\operatorname{dexp}^{-1}_{-\widetilde{\boldsymbol{\nu}}_m(t)}$.

■

Anhang C

Reihenglieder der Magnus-Entwicklung

Nach [40] berechnet sich $\widetilde{\boldsymbol{\nu}}(t)$ aus (2.29) durch die Formel

$$\widetilde{\boldsymbol{\nu}}(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \widetilde{\boldsymbol{\mu}}_r(t) \tag{C.1}$$

mit

$$\widetilde{\boldsymbol{\mu}}_r(t) = \widetilde{\mathbf{v}}^{(r-1)}(t) - (r-1)! \sum_{j=1}^{r-1} \frac{1}{(j-1)!} \sum_{k=1}^{r-j} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_{r-j}^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_j(t).$$

Hierbei ist Π_j^k die Menge der geordneten Partitionen von j der Länge k, $\pi \in \Pi_j^k$ ist eine geordnete Menge von nicht-negativen natürlichen Zahlen $\pi = \{\pi_1, ..., \pi_k\}$, so dass $\pi_1 + ... + \pi_k = j$ und $\mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} = \mathrm{ad}_{\widetilde{\mathbf{Y}}_{\pi_1}} \cdot ... \cdot \mathrm{ad}_{\widetilde{\mathbf{Y}}_{\pi_k}}$ sind, wobei $\widetilde{\mathbf{Y}}_i = \frac{1}{i!} \widetilde{\boldsymbol{\mu}}_i$ gilt. Mit [40] sind die Werte bis fünfter Ordnung gegeben durch

$$\widetilde{\boldsymbol{\mu}}_0(t) = \mathbf{0}, \tag{C.2a}$$

$$\widetilde{\boldsymbol{\mu}}_1(t) = \widetilde{\mathbf{v}}(t), \tag{C.2b}$$

$$\widetilde{\boldsymbol{\mu}}_2(t) = \dot{\widetilde{\mathbf{v}}}(t), \tag{C.2c}$$

$$\widetilde{\boldsymbol{\mu}}_3(t) = \ddot{\widetilde{\mathbf{v}}}(t) + \frac{1}{2}[\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t)], \tag{C.2d}$$

$$\widetilde{\boldsymbol{\mu}}_4(t) = \dddot{\widetilde{\mathbf{v}}}(t) + [\widetilde{\mathbf{v}}(t), \ddot{\widetilde{\mathbf{v}}}(t)], \tag{C.2e}$$

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_5(t) = \widetilde{\mathbf{v}}^{(4)}(t) &+ \frac{3}{2}[\widetilde{\mathbf{v}}(t), \dddot{\widetilde{\mathbf{v}}}(t)] + [\dot{\widetilde{\mathbf{v}}}(t), \ddot{\widetilde{\mathbf{v}}}(t)] - \frac{1}{2}[\dot{\widetilde{\mathbf{v}}}(t), [\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t)]] \\ &+ \frac{1}{6}[\widetilde{\mathbf{v}}(t), [\widetilde{\mathbf{v}}(t), \ddot{\widetilde{\mathbf{v}}}(t)]] - \frac{1}{6}[\widetilde{\mathbf{v}}(t), [\widetilde{\mathbf{v}}(t), [\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t)]]]. \end{aligned} \tag{C.2f}$$

Die Vorzeichen vor den Kommutatoren mit gerader Anzahl an Einträgen unterscheiden sich von denen in [40] gegebenen, da dort die Rechtstranslation verwendet wurde. Die Ergebnisse für $r = 6$ und $r = 7$ sollen nun vorgestellt werden. Dabei wurde die Formel (C.1) aus [40] verwendet, $\widetilde{\boldsymbol{\mu}}_6(t)$ und $\widetilde{\boldsymbol{\mu}}_7(t)$ sind dort jedoch nicht angegeben. Es folgt

$$\widetilde{\boldsymbol{\mu}}_6(t) = \widetilde{\mathbf{v}}^{(5)}(t) - 5! \sum_{j=1}^{5} \frac{1}{(j-1)!} \sum_{k=1}^{6-j} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_{6-j}^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_j(t)$$

(C.3)

$$= \widetilde{\mathbf{v}}^{(5)}(t) - 5! \sum_{k=1}^{5} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_5^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_1(t) - 5! \sum_{k=1}^{4} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_4^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t)$$
$$- 60 \sum_{k=1}^{3} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_3^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_3(t) - 20 \sum_{k=1}^{2} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_2^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_4(t)$$
$$+ \frac{5}{2} \sum_{\pi \in \Pi_1^1} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_5(t). \tag{C.4}$$

Um die Rechnungen nicht zu sehr ausufern zu lassen, wird nachfolgend nur eine der Summen aus (C.4) ausführlich berechnet. Die Berechnung der restlichen Summen folgt jedoch analog. Somit ist

$$\begin{aligned} -5! \sum_{k=1}^{4} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_4^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) &= 60 \sum_{\pi \in \Pi_4^1} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) - 20 \sum_{\pi \in \Pi_4^2} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) \\ &\quad + 5 \sum_{\pi \in \Pi_4^3} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) - \sum_{\pi \in \Pi_4^4} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) \\ &= -20 \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \mathrm{ad}_{\widetilde{\mathbf{Y}}_3} \widetilde{\boldsymbol{\mu}}_2(t) - 20 \mathrm{ad}_{\widetilde{\mathbf{Y}}_3} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \widetilde{\boldsymbol{\mu}}_2(t) \\ &\quad + 5 \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \mathrm{ad}_{\widetilde{\mathbf{Y}}_2} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \widetilde{\boldsymbol{\mu}}_2(t) + 5 \mathrm{ad}_{\widetilde{\mathbf{Y}}_2} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \widetilde{\boldsymbol{\mu}}_2(t) \\ &\quad - \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \mathrm{ad}_{\widetilde{\mathbf{Y}}_1} \widetilde{\boldsymbol{\mu}}_2(t), \end{aligned}$$

da $\mathrm{ad}_{\widetilde{\mathbf{Y}}_2} \widetilde{\boldsymbol{\mu}}_2(t) = \frac{1}{2!}[\widetilde{\boldsymbol{\mu}}_2(t), \widetilde{\boldsymbol{\mu}}_2(t)] = \mathbf{0}$ ist bzw. im Allgemeinen

$$\mathrm{ad}_{\widetilde{\mathbf{Y}}_i} \widetilde{\boldsymbol{\mu}}_i(t) = \frac{1}{i!}[\widetilde{\boldsymbol{\mu}}_i(t), \widetilde{\boldsymbol{\mu}}_i(t)] = \mathbf{0}$$

für alle $i \in \mathbb{N}$ gilt. Durch das Einsetzen der $\widetilde{\mathbf{Y}}_i$ und Gleichung (C.2) folgt dann

$$\begin{aligned} &-5! \sum_{k=1}^{4} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_4^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) \\ &= -\frac{5}{2} \left[\dot{\widetilde{\mathbf{v}}}(t), \dddot{\widetilde{\mathbf{v}}}(t) \right] - \frac{5}{2} \left[\dot{\widetilde{\mathbf{v}}}(t), \left[\widetilde{\mathbf{v}}(t), \ddot{\widetilde{\mathbf{v}}}(t) \right] \right] + \frac{10}{3} \left[\widetilde{\mathbf{v}}(t), \left[\dot{\widetilde{\mathbf{v}}}(t), \ddot{\widetilde{\mathbf{v}}}(t) \right] \right] \\ &\quad + \frac{25}{6} \left[\widetilde{\mathbf{v}}(t), \left[\dot{\widetilde{\mathbf{v}}}(t), \left[\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t) \right] \right] \right] - \frac{10}{3} \left[\ddot{\widetilde{\mathbf{v}}}(t), \left[\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t) \right] \right] \\ &\quad + \frac{5}{2} \left[\dot{\widetilde{\mathbf{v}}}(t), \left[\widetilde{\mathbf{v}}(t), \left[\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t) \right] \right] \right] - \left[\widetilde{\mathbf{v}}(t), \left[\widetilde{\mathbf{v}}(t), \left[\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t) \right] \right] \right]. \end{aligned}$$

Wird jede der Summen aus (C.4) auf diese Art untersucht, gilt insgesamt

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_6(t) = \Bigg[&\widetilde{\mathbf{v}}^{(5)} + 2[\widetilde{\mathbf{v}}, \widetilde{\mathbf{v}}^{(4)}] + \frac{1}{2}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dddot{\widetilde{\mathbf{v}}}]] - \frac{1}{2}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]]] + \frac{1}{3}[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, \ddot{\widetilde{\mathbf{v}}}]] \\ &- \frac{7}{12}[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] - \frac{5}{6}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]] - \frac{5}{12}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] + \frac{5}{2}[\dot{\widetilde{\mathbf{v}}}, \dddot{\widetilde{\mathbf{v}}}] \\ &- \frac{5}{3}[\ddot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]] \Bigg] (t) \end{aligned}$$

bzw. mit der Jacobi-Identität (2.22)

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_6(t) = \Bigg[&\widetilde{\mathbf{v}}^{(5)} + 2[\widetilde{\mathbf{v}}, \widetilde{\mathbf{v}}^{(4)}] + \frac{1}{2}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dddot{\widetilde{\mathbf{v}}}]] - \frac{1}{2}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]]] + 2[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, \ddot{\widetilde{\mathbf{v}}}]] \\ &- \frac{7}{12}[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] - \frac{5}{2}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]] - \frac{5}{12}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] + \frac{5}{2}[\dot{\widetilde{\mathbf{v}}}, \dddot{\widetilde{\mathbf{v}}}] \Bigg] (t). \end{aligned}$$

Analog erfolgt die Rechnung für $\widetilde{\boldsymbol{\mu}}_7$. Daher ist

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_7(t) &= \widetilde{\mathbf{v}}^{(6)}(t) - 6! \sum_{j=1}^{6} \frac{1}{(j-1)!} \sum_{k=1}^{7-j} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_{7-j}^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_j(t) \\
&= \widetilde{\mathbf{v}}^{(6)}(t) - 6! \sum_{k=1}^{6} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_6^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_1(t) - 6! \sum_{k=1}^{5} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_5^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_2(t) \\
&\quad - 360 \sum_{k=1}^{4} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_4^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_3(t) - 120 \sum_{k=1}^{3} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_3^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_4(t) \\
&\quad - 30 \sum_{k=1}^{2} \frac{(-1)^k}{(k+1)!} \sum_{\pi \in \Pi_2^k} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_5(t) + 3 \sum_{\pi \in \Pi_1^1} \mathrm{ad}_{\widetilde{\mathbf{Y}}}^{\pi} \widetilde{\boldsymbol{\mu}}_6(t).
\end{aligned}$$

Schließlich folgt mit der Jacobi-Identität (2.22)

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_7(t) = \Bigg[& \widetilde{\mathbf{v}}^{(6)} + \frac{5}{2}[\widetilde{\mathbf{v}}, \widetilde{\mathbf{v}}^{(5)}] + [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \widetilde{\mathbf{v}}^{(4)}]] - [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dddot{\widetilde{\mathbf{v}}}]]] - \frac{1}{6}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]]]] \\
& + \frac{1}{6}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]]]] + [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, \ddot{\widetilde{\mathbf{v}}}]]] + \frac{7}{24}[\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]]] + 5[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, \dddot{\widetilde{\mathbf{v}}}]] \\
& - \frac{15}{4}[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]]] - \frac{1}{24}[\widetilde{\mathbf{v}}, [\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]]] + \frac{5}{12}[[\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}], [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]] \\
& - \frac{5}{12}[[\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}], [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] + \frac{9}{2}[\dot{\widetilde{\mathbf{v}}}, \widetilde{\mathbf{v}}^{(4)}] - \frac{9}{2}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \dddot{\widetilde{\mathbf{v}}}]] - \frac{7}{4}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]]] \\
& + \frac{1}{4}[\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]]] - \frac{1}{2}[\dot{\widetilde{\mathbf{v}}}, [\dot{\widetilde{\mathbf{v}}}, \ddot{\widetilde{\mathbf{v}}}]] - [\dot{\widetilde{\mathbf{v}}}, [\dot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] + \frac{5}{2}[\ddot{\widetilde{\mathbf{v}}}, \dddot{\widetilde{\mathbf{v}}}] - \frac{25}{6}[\ddot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, \ddot{\widetilde{\mathbf{v}}}]] \\
& - \frac{5}{6}[\ddot{\widetilde{\mathbf{v}}}, [\widetilde{\mathbf{v}}, [\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]]] \Bigg] (t).
\end{aligned}$$

Anhang D

Beweis von Satz 4

In diesem Anhang soll Satz 4 bewiesen werden. Dieser Satz ist eine Folgerung aus [4, Theorem 4.16] und Voraussetzung 1. Dafür werden die Variablen h_n, $\kappa_{\mathbf{y},n}$ und $\kappa_{\mathbf{z},n}$ durch $h_{\max}$, $\kappa_{\mathbf{y},\max}$ und $\kappa_{\mathbf{z},\max}$ mit

$$\kappa_{\mathbf{y},\max} = \max_{n=0,\dots,N_{\text{end}}} \kappa_{\mathbf{y},n} \leq 1 \quad \text{und} \quad \kappa_{\mathbf{z},\max} = \max_{n=0,\dots,N_{\text{end}}} \kappa_{\mathbf{z},n} < 1$$

ersetzt und zum Abschluss des Beweises wird nicht $nh = (t_n - t_0)$ sondern

$$nh_{\max} \leq nC_h h_{\min} \leq C_h \sum_{i=0}^{n} h_i = C_h(t_n - t_0) \tag{D.1}$$

verwendet.

Beweis:

a) Zunächst soll gezeigt werden, dass die Wahl der Norm keine Rolle spielt, solange jeweils eine Norm existiert, die die gegebenen Voraussetzungen erfüllt. Dies kann analog zu [4, Beweisteil a) in Theorem 4.16] erfolgen und soll aus diesem Grund nicht weiter vertieft werden. Zur Vereinfachung der Notation soll die folgende Fehleranalyse auf ein Normenpaar $(\|.\|_{\mathbf{y},\rho}, \|.\|_{\mathbf{z},\rho})$ mit $\kappa_{\mathbf{y},n} \coloneqq \|\mathbf{T}_{\mathbf{y},n}\|_{\mathbf{y}} \leq 1$ und $\kappa_{\mathbf{z},n} \coloneqq \|\mathbf{T}_{\mathbf{z},n}\|_{\mathbf{z}} < 1$ beschränkt werden. Außerdem wird im Folgenden auf die Indizes $\mathbf{y}$ und $\mathbf{z}$ am Normsymbol $\|.\|$ verzichtet.

b) Dieser Teil des Beweises erfolgt in Anlehnung an [4, Beweisteil b) Theorem 4.16]. Der einzige Unterschied besteht darin, dass sowohl h_n als auch $\kappa_{\mathbf{z},n}$ vom gewählten Zeitschritt abhängen.
Zunächst werden zwei Folgen $(u_n)_{n\geq 0}$ und $(w_n)_{n\geq 0}$ definiert durch

$$u_n \coloneqq \left\| \mathbf{E}_n^{\mathbf{y}} - \prod_{i=0}^{n-1} \mathbf{T}_{\mathbf{y},i} \mathbf{E}_0^{\mathbf{y}} \right\|, \quad u_0 \coloneqq \|\mathbf{E}_0^{\mathbf{y}} - \mathbf{E}_0^{\mathbf{y}}\| = 0, \tag{D.2a}$$

$$w_n \coloneqq \left\| \mathbf{E}_n^{\mathbf{z}} - \prod_{i=0}^{n-1} \mathbf{T}_{\mathbf{z},i}^{n} \mathbf{E}_0^{\mathbf{z}} \right\|, \quad w_0 \coloneqq \|\mathbf{E}_0^{\mathbf{z}} - \mathbf{E}_0^{\mathbf{z}}\| = 0. \tag{D.2b}$$

Weiter seien

$$\kappa_{\mathbf{y},n} \coloneqq \|\mathbf{T}_{\mathbf{y},n}\| \quad \text{und} \quad \kappa_{\mathbf{z},n} \coloneqq \|\mathbf{T}_{\mathbf{z},n}\|. \tag{D.3}$$

Laut Voraussetzungen gilt

$$\max_{n=0,\dots,N_{\text{end}}} \kappa_{\mathbf{y},n} \leq 1 \quad \text{und} \quad \max_{n=0,\dots,N_{\text{end}}} \kappa_{\mathbf{z},n} = \kappa_{\mathbf{z},\max} < 1. \tag{D.4}$$

Mit Hilfe der Dreiecksungleichung sind die Gleichungen

$$\begin{aligned}\|\mathbf{E}_n^{\mathbf{y}}\| &\leq \left\|\mathbf{E}_n^{\mathbf{y}} - \prod_{i=0}^{n-1}\mathbf{T}_{\mathbf{y},i}\mathbf{E}_0^{\mathbf{y}}\right\| + \left\|\prod_{i=0}^{n-1}\mathbf{T}_{\mathbf{y},i}\mathbf{E}_0^{\mathbf{y}}\right\| \leq u_n + \prod_{i=0}^{n-1}\kappa_{\mathbf{y},i}\|\mathbf{E}_0^{\mathbf{y}}\| \\ &\overset{(D.4)}{\leq} u_n + \|\mathbf{E}_0^{\mathbf{y}}\| \end{aligned} \tag{D.5a}$$

und analog

$$\|\mathbf{E}_n^{\mathbf{z}}\| \leq w_n + \prod_{i=0}^{n-1}\kappa_{\mathbf{z},i}\|\mathbf{E}_0^{\mathbf{z}}\| \leq w_n + \kappa_{\mathbf{z},\max}^n\|\mathbf{E}_0^{\mathbf{z}}\| \tag{D.5b}$$

erfüllt. Werden (D.5a) und (D.5b) in (4.42) eingesetzt, so folgen die Abschätzungen

$$\begin{aligned}\|\mathbf{E}_{n+1}^{\mathbf{y}} - \mathbf{T}_{\mathbf{y},n}\mathbf{E}_n^{\mathbf{y}}\| \leq\ & L_0h_n(u_n + u_{n+1} + w_n + w_{n+1}) \\ & + L_0h_n\left(2\|\mathbf{E}_0^{\mathbf{y}}\| + \kappa_{\mathbf{z},\max}^n(1+\kappa_{\mathbf{z},\max})\|\mathbf{E}_0^{\mathbf{z}}\|\right) + h_nM_0,\end{aligned} \tag{D.6a}$$

$$\begin{aligned}\|\mathbf{E}_{n+1}^{\mathbf{z}} - \mathbf{T}_{\mathbf{z},n}\mathbf{E}_n^{\mathbf{z}}\| \leq\ & L_0(u_n + u_{n+1} + h_n(w_n + w_{n+1})) \\ & + L_0\left(2\|\mathbf{E}_0^{\mathbf{y}}\| + h_n\kappa_{\mathbf{z},\max}^n(1+\kappa_{\mathbf{z},\max})\|\mathbf{E}_0^{\mathbf{z}}\|\right) + M_0.\end{aligned} \tag{D.6b}$$

Außerdem gilt mit (D.2a), (D.3), (D.4), (D.6a) und $h_n \leq h_{\max}$

$$\begin{aligned}u_{n+1} &\leq \|\mathbf{E}_{n+1}^{\mathbf{y}} - \mathbf{T}_{\mathbf{y},n}\mathbf{E}_n^{\mathbf{y}}\| + \|\mathbf{T}_{\mathbf{y},n}\|\left\|\mathbf{E}_n^{\mathbf{y}} - \prod_{i=0}^{n-1}\mathbf{T}_{\mathbf{y},i}\mathbf{E}_0^{\mathbf{y}}\right\| \\ &\leq (1 + L_0h_{\max})u_n + L_0h_{\max}(u_{n+1} + w_n + w_{n+1}) \\ &\quad + L_0h_{\max}\left(2\|\mathbf{E}_0^{\mathbf{y}}\| + \kappa_{\mathbf{z},\max}^n(1+\kappa_{\mathbf{z},\max})\|\mathbf{E}_0^{\mathbf{z}}\|\right) + h_{\max}M_0\end{aligned}$$

und analog

$$\begin{aligned}w_{n+1} \leq\ & L_0(u_n + u_{n+1}) + (h_{\max}L_0 + \kappa_{\mathbf{z},\max})w_n + h_{\max}L_0w_{n+1} \\ & + L_0\left(2\|\mathbf{E}_0^{\mathbf{y}}\| + h_{\max}\kappa_{\mathbf{z},\max}^n(1+\kappa_{\mathbf{z},\max})\|\mathbf{E}_0^{\mathbf{z}}\|\right) + M_0.\end{aligned}$$

Daraus lässt sich das Ungleichungssystem

$$\begin{aligned}&\underbrace{\begin{bmatrix}1 - L_0h_{\max} & -L_0h_{\max} \\ -L_0 & 1 - L_0h_{\max}\end{bmatrix}}_{=:\mathbf{L}}\underbrace{\begin{bmatrix}u_{n+1} \\ w_{n+1}\end{bmatrix}}_{=:\mathbf{v}_{n+1}} \\ &\leq \underbrace{\begin{bmatrix}1 + L_0h_{\max} & L_0h_{\max} \\ L_0 & L_0h_{\max} + \kappa_{\mathbf{z},\max}\end{bmatrix}\begin{bmatrix}u_n \\ w_n\end{bmatrix} + \mathbf{R}}_{=:\mathbf{b}_n}\end{aligned}$$

mit

$$\mathbf{R} \coloneqq \begin{bmatrix}L_0h_{\max}\left(2\|\mathbf{E}_0^{\mathbf{y}}\| + \kappa_{\mathbf{z},\max}^n(1+\kappa_{\mathbf{z},\max})\|\mathbf{E}_0^{\mathbf{z}}\|\right) + h_{\max}M_0 \\ L_0\left(2\|\mathbf{E}_0^{\mathbf{y}}\| + h_{\max}\kappa_{\mathbf{z},\max}^n(1+\kappa_{\mathbf{z},\max})\|\mathbf{E}_0^{\mathbf{z}}\|\right) + M_0\end{bmatrix}$$

formulieren. Die Matrix $\mathbf{L}$ ist inversmonoton (vgl. [35]), da $0 < h_{\max} \leq \overline{h} \leq \frac{1}{4L_0+2L_0^2}$ und $L_0 > 0$ den Zusammenhang

$$\begin{aligned}\begin{bmatrix} 1 - L_0 h_{\max} & -L_0 h_{\max} \\ -L_0 & 1 - L_0 h_{\max} \end{bmatrix}^{-1} &= \frac{1}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} \cdot \\ &\quad \cdot \begin{bmatrix} 1 - h_{\max} L_0 & h_{\max} L_0 \\ L_0 & 1 - h_{\max} L_0 \end{bmatrix} \\ &\geq \mathbf{0}_{4\times 4}\end{aligned}$$

liefern. Deshalb folgt aus $\mathbf{0}_{4\times 4} \leq (\mathbf{b}_n - \mathbf{L}\mathbf{v}_{n+1})$ direkt $\mathbf{0}_{4\times 4} \leq \mathbf{L}^{-1}(\mathbf{b}_n - \mathbf{L}\mathbf{v}_{n+1})$ und somit

$$\begin{aligned}&\mathbf{v}_{n+1} \leq \mathbf{L}^{-1}\mathbf{b}_n = \\ &\begin{bmatrix} \frac{1 + h_{\max} L_0^2 - h_{\max}^2 L_0^2}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} & \frac{h_{\max} L_0 (1 + \kappa_{\mathbf{z},\max})}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} \\ \frac{L_0(1 + \kappa_{\mathbf{z},\max}) + 2h_{\max} L_0^2}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} & \frac{h_{\max} L_0^2 + h_{\max}^2 L_0^2 + 2h_{\max} L_0 \kappa_{\mathbf{z},\max} + \kappa_{\mathbf{z},\max}^2}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} \end{bmatrix} \\ &\cdot \begin{bmatrix} u_n \\ \\ w_n \end{bmatrix} + \mathbf{L}^{-1}\mathbf{R}.\end{aligned}$$

Weiterhin sind für $0 < h_{\max} \leq \overline{h} \leq \frac{1}{4L_0+2L_0^2}$, $L_0 > 0$ und $\overline{L}_0 \coloneqq 8L_0 + 4L_0^2$ die Ungleichungen

$$\begin{aligned}\frac{1 + h_{\max} L_0^2 - h_{\max}^2 L_0^2}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} &\leq 1 + (8L_0 + 4L_0^2)h_{\max} \leq 1 + \overline{L}_0 h_{\max}, \\ \frac{h_{\max} L_0 (1 + \kappa_{\mathbf{z},\max})}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} &\leq 2L_0(1 + \kappa_{\mathbf{z},\max})h_{\max} \leq \overline{L}_0 h_{\max}, \\ \frac{L_0(1 + \kappa_{\mathbf{z},\max}) + 2h_{\max} L_0^2}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} &\leq 4L_0 \leq \overline{L}_0, \\ \frac{h_{\max}(L_0^2 + 2L_0\kappa_{\mathbf{z},\max}) + h_{\max}^2 L_0^2 + \kappa_{\mathbf{z},\max}^2}{1 - h_{\max}(2L_0 + L_0^2) + h_{\max}^2 L_0^2} &\leq \kappa_{\mathbf{z},\max} + (8L_0 + 4L_0^2)h_{\max} \\ &\leq \kappa_{\mathbf{z},\max} + \overline{L}_0 h_{\max}\end{aligned}$$

erfüllt und es ergibt sich das Ungleichungssystem

$$\begin{bmatrix} u_{n+1} \\ w_{n+1} \end{bmatrix} \leq \underbrace{\begin{bmatrix} 1 + \overline{L}_0 h_{\max} & \overline{L}_0 h_{\max} \\ \overline{L}_0 & \kappa_{\mathbf{z},\max} + \overline{L}_0 h_{\max} \end{bmatrix}}_{\mathbf{W}(h_{\max})} \begin{bmatrix} u_n \\ w_n \end{bmatrix} + \overline{\mathbf{R}} \tag{D.7}$$

mit $\overline{M}_0 \coloneqq 2(1 + L_0)$ und

$$\overline{\mathbf{R}} \coloneqq \begin{bmatrix} \overline{L}_0 h_{\max} \|\mathbf{E}_0^{\mathbf{y}}\| + \kappa_{\mathbf{z},\max}^n \overline{L}_0 h_{\max} \|\mathbf{E}_0^{\mathbf{z}}\| + h_{\max}\overline{M}_0 \\ \overline{L}_0 \|\mathbf{E}_0^{\mathbf{y}}\| + \kappa_{\mathbf{z},\max}^n \overline{L}_0 h_{\max} \|\mathbf{E}_0^{\mathbf{z}}\| + \overline{M}_0 \end{bmatrix}.$$

c) Da h_n durch die maximale Schrittweite $h_{\max}$ abgeschätzt wurde, hängen weder $\mathbf{W}(h_{\max})$ noch $\overline{\mathbf{R}}$ vom betrachteten Zeitschritt n ab. Damit kann von nun an genau wie in [4, Beweisteil c) Theorem 4.16.] weiter vorgegangen werden. Dabei wird durch eine Eigenwertanalyse die Matrix $\mathbf{W}(h_{\max})$ auf Diagonalgestalt gebracht mit

$$\begin{aligned}&\mathbf{V}^{-1}(h_{\max})\mathbf{W}(h_{\max})\mathbf{V}(h_{\max})\\ &= \begin{bmatrix} 1+\overline{L}_0(L_\zeta+1)h_{\max} & 0 \\ \overline{L}_0 & \kappa_{\mathbf{z},\max}+\overline{L}_0(1-L_\zeta)h_{\max}\end{bmatrix},\end{aligned}$$

wobei

$$\mathbf{V}(h_{\max}) \coloneqq \begin{bmatrix} 1 & -L_\zeta h_{\max} \\ 0 & 1\end{bmatrix}, \quad \mathbf{V}^{-1}(h_{\max}) = \begin{bmatrix} 1 & L_\zeta h_{\max} \\ 0 & 1\end{bmatrix}$$

und $\mathbf{L}_\zeta \coloneqq \frac{\overline{L}_0}{1-\kappa_{\mathbf{z},\max}} + \mathcal{O}(h_{\max})$ sind.

Mit einer Folge $(v_n)_{n\geq 0}$ von nichtnegativen Zahlen v_n definiert durch

$$\begin{bmatrix} v_n \\ w_n\end{bmatrix} = \mathbf{V}^{-1}(h_{\max}) \begin{bmatrix} u_n \\ w_n\end{bmatrix}$$

kann (D.7) umgeschrieben werden zu

$$\begin{aligned} v_{n+1} &\leq \big(1+\overline{L}_0\big(L_\zeta+1\big)h_{\max}\big)v_n + \overline{L}_0\big(L_\zeta h_{\max}+1\big)h_{\max}\kappa_{\mathbf{z},\max}^n\|\mathbf{E}_0^{\mathbf{z}}\| \\ &\quad + \big(L_\zeta+1\big)h_{\max}\big(\overline{M}_0+\overline{L}_0\|\mathbf{E}_0^{\mathbf{y}}\|\big), \qquad \text{(D.8a)} \\ w_{n+1} &\leq \overline{L}_0 v_n + \big(\kappa_{\mathbf{z},\max}+\overline{L}_0\big(1-L_\zeta\big)\big)w_n + \overline{L}_0 h_{\max}\kappa_{\mathbf{z},\max}^n\|\mathbf{E}_0^{\mathbf{z}}\| + \overline{M}_0 + \overline{L}_0\|\mathbf{E}_0^{\mathbf{y}}\|. \qquad \text{(D.8b)}\end{aligned}$$

d) Auch dieser Teil des Beweises kann analog zu [4, Beweisteil d) Theorem 4.16] erfolgen. Da die rechte Seite von (D.8a) nichtlinear von $h_{\max}$ abhängt, denn $L_\zeta = L_\zeta(h_{\max})$, wird L_ζ für hinreichend kleines $h_{\max}$ durch $\overline{L}_\zeta$ mit

$$L_\zeta = \frac{\overline{L}_0}{1-\kappa_{\mathbf{z},\max}} + \mathcal{O}(h_{\max}) \leq \frac{2\overline{L}_0}{1-\kappa_{\mathbf{z},\max}} =: \overline{L}_\zeta$$

substituiert. Für $L \coloneqq \overline{L}_0\big(\overline{L}_\zeta \max\{1,\overline{h}\}+1\big)$, $e_0 = \|\mathbf{E}_0^{\mathbf{z}}\|$, $M \coloneqq \big(\overline{L}_\zeta+1\big)\overline{M}_0 + L\|\mathbf{E}_0^{\mathbf{y}}\|$, $\kappa \coloneqq \kappa_{\mathbf{z},\max}$ und $h \coloneqq h_{\max}$ kann für Gleichung (D.8a) das Lemma 17 angewendet werden, so dass aus Gleichung (4.47a) die Abschätzung

$$v_n \leq \mathrm{err}_n - \|\mathbf{E}_0^{\mathbf{y}}\| \qquad \text{(D.9)}$$

mit

$$\mathrm{err}_n \coloneqq e^{Lnh_{\max}}\left(\|\mathbf{E}_0^{\mathbf{y}}\| + \frac{h_{\max}L}{1-\kappa_{\mathbf{z},\max}}\|\mathbf{E}_0^{\mathbf{z}}\|\right) + \frac{e^{Lnh_{\max}}-1}{L}(\overline{L}_\zeta+1)\overline{M}_0$$

folgt. Aus den Gleichungen (D.1), (D.5a), (D.9) und $u_n = v_n - L_{\zeta,n}w_n \leq v_n$ kann daher

$$\|\mathbf{E}_n^{\mathbf{y}}\| \leq e^{\overline{L}_0 C_h(t_n-t_0)}\left(\|\mathbf{E}_0^{\mathbf{y}}\| + \overline{C}_0 h_{\max}\|\mathbf{E}_0^{\mathbf{z}}\| + \frac{e^{\overline{L}_0 C_h(t_n-t_0)}-1}{\overline{L}}\overline{M}_0\right)$$

gezeigt werden.
Wird in (D.8b) die Abschätzung (D.9) eingesetzt, so kann für $\kappa \coloneqq \kappa_{\mathbf{z},\max}$, $\overline{L}_0(1-L_{\zeta,n}) \leq \overline{L}_0 \leq L$, $e_0 = \|\mathbf{E}_0^{\mathbf{z}}\|$, $M \coloneqq \overline{M}_0 + \overline{L}_0 \mathrm{err}_n$ und $h \coloneqq h_{\max}$ Lemma 17 angewendet werden, da $(\mathrm{err}_n)_{n\geq 0}$ monoton wachsend ist. Es folgt für hinreichend kleines $h_{\max} \in (0, \overline{h})$ aus (4.47b), (D.1), (D.2b) und $1 \leq e^x$ für $x \geq 0$ die Gleichung

$$\|\mathbf{E}_n^{\mathbf{z}} - \mathbf{T}_{\mathbf{z},n}^n \mathbf{E}_0^{\mathbf{z}}\| \leq \overline{C}_0 e^{\overline{L}_0 C_h (t_n - t_0)} \big(\|\mathbf{E}_0^{\mathbf{y}}\| + h_{\max} \|\mathbf{E}_0^{\mathbf{z}}\| + \overline{M}_0 \big).$$

Damit folgt die Behauptung.

■

Anhang E

Gleichungen für die globalen Fehler und gekoppelte Fehlerrekursion der BLieDF-Verfahren mit variablen Schrittweiten

In Kapitel 5 wurde die Konvergenz der BLieDF-Verfahren (5.35) bewiesen. Ein Teil der Berechnungen wurde jedoch ausgelassen, da keine neuen Beweisideen enthalten sind. Die fehlenden Beweisschritte sollen in diesem Anhang angegeben werden. Dazu werden die Gleichungen der globalen Fehler $\mathbf{e}_n^q$, $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$, $\mathbf{e}_n^{\mathbf{v}}$ und $\mathbf{e}_n^{\boldsymbol{\lambda}}$ aufgestellt und zu der gekoppelten Fehlerrekursion (4.42) kombiniert.

E.1 Globale Fehlergleichung für $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$

Um eine globale Fehlergleichung für $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$ zu erhalten, wird wie in Satz 8 vorgegangen.

Satz E.1
Der globale Fehler $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$ der BLieDF-Verfahren (5.35) erfüllt die Fehlerrekursion

$$\frac{1}{h_n}\sum_{i=1}^{k}\gamma_{i,n}\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}} = \mathbf{e}_{n+1}^{\mathbf{v}} + \mathbf{l}_n^{\boldsymbol{\omega}} + \mathcal{O}(h_n)\left(\sum_{i=0}^{k}\epsilon_{n+1-i} + \frac{1}{h_n}\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right). \quad \text{(E.1)}$$

Beweis:
Da der Korrekturterm $\mathbf{L}_{h_n,n}^{(k)}$ einer Lipschitzbedingung

$$\mathbf{L}_{h_n,n}^{(k)} - \mathbf{L}_{h_n}^{(k)}(t_n) = \mathcal{O}(h_n)\sum_{i=0}^{k}\|\mathbf{e}_{n+1-i}^{\mathbf{v}}\| + \mathcal{O}(1)\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|$$

genügt, folgt die Behauptung analog zu Satz 8 mit den Gleichungen (5.35b) und (5.37b).

■

E.2 Globale Fehlergleichung für $\mathbf{e}_n^q$

Die Berechnung der Fehlerrekursion für $\mathbf{e}_n^q$ erfolgt analog zum konstanten Fall in Satz 10.

Satz E.2
Die globalen Fehler $\mathbf{e}_n^q$ der BLieDF-Verfahren (5.35) erfüllen die Abschätzung

$$\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{i,n}\mathbf{e}_{n+1-i}^q = \mathbf{e}_{n+1}^{\mathbf{v}} + \mathbf{l}_n^{\boldsymbol{\omega}} - \sum_{i=1}^{k}\gamma_{i,n}\left(\prod_{j=0}^{i-2}\frac{1}{\sigma_{n-j}}\right)\widehat{\mathbf{v}}(t_n)\mathbf{e}_{n+1-i}^q + \mathcal{O}(h_n)\left(\sum_{i=0}^{k}\epsilon_{n+1-i} + \frac{1}{h_n}\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right). \tag{E.2}$$

Beweis:
Der Beweis erfolgt analog zu Satz 10 für $h = h_n$ unter Beachtung von (5.39) und (5.41).

■

E.3 Globale Fehlerrekursion für $\mathbf{e}_n^{\mathbf{v}}$

Für den globalen Fehler $\mathbf{e}_n^{\mathbf{v}}$ muss zunächst, wie zuvor in Generalized-α-Verfahren (vgl. Lemma 9), die Differenz von $\overline{\mathbf{v}}(t_n)$ mit $\overline{\mathbf{v}}_n$ untersucht werden.

Lemma E.1
Es gilt für $i = 1, ..., k$

$$\overline{\mathbf{v}}(t_{n+1-i}) - \overline{\mathbf{v}}_{n+1-i} = \mathbf{e}_{n+1-i}^{\mathbf{v}} - \mathbf{C}\big(q(t_{n+1-i})\big)\frac{\gamma_{i,n}}{\alpha_{i,n}}\big(\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}\big) + \mathcal{O}(h_n^{k+1})$$

mit

$$\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}} = \frac{1}{6}\frac{2\sigma_n^2\sigma_{n-1}}{(1+\sigma_{n-1})}\left(\prod_{j=0}^{i-1}\frac{1}{\sigma_{n-j}^2}\right)\left(\frac{1+\sigma_{n-i-1}}{\sigma_{n-i}^2\sigma_{n-i-1}} - \frac{1+\sigma_{n-i}}{\sigma_{n-i}}\right)\cdot \big(\mathbf{e}_n^{\mathbf{v}} - (1+\sigma_{n-1})\mathbf{e}_{n-1}^{\mathbf{v}} + \sigma_{n-1}\mathbf{e}_{n-2}^{\mathbf{v}}\big) + \mathcal{O}(h_n)\sum_{i=1}^{k}\epsilon_{n+1-i} + \mathcal{O}(h_n^3) \tag{E.3a}$$

für $k = 2$ und

$$\begin{aligned}\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}} = \frac{1}{4}\sigma_n^3\sigma_{n-1}^2\sigma_{n-2}\prod_{j=0}^{i-1}\frac{1}{\sigma_{n-j}^3}\left(\frac{1+\sigma_{n-i-1}}{\sigma_{n-i}^3\sigma_{n-i-1}^2\sigma_{n-i-2}} - \frac{1+\sigma_{n-i}}{\sigma_{n-i}^2\sigma_{n-i-1}}\right)\cdot \\ \cdot\left(\frac{1}{(1+\sigma_{n-1})(1+\sigma_{n-2}+\sigma_{n-1}\sigma_{n-2})}\mathbf{e}_n^{\mathbf{v}} - \frac{1}{(1+\sigma_{n-2})}\mathbf{e}_{n-1}^{\mathbf{v}}\right. \\ \left. + \frac{\sigma_{n-1}}{(1+\sigma_{n-1})}\mathbf{e}_{n-2}^{\mathbf{v}} - \frac{\sigma_{n-1}\sigma_{n-2}^2}{(1+\sigma_{n-2})(1+\sigma_{n-2}+\sigma_{n-2}\sigma_{n-1})}\mathbf{e}_{n-3}^{\mathbf{v}}\right) \\ + \mathcal{O}(h_n)\sum_{i=1}^{k}\epsilon_{n+1-i} + \mathcal{O}(h_n^4)\end{aligned} \tag{E.3b}$$

für $k = 3$.

Beweis:
Aufgrund von (5.34) sind

$$\dot{\mathbf{v}}(t_n) - \dot{\mathbf{v}}_n = \frac{\sigma_n \mathbf{e}_n^{\mathbf{v}} - \sigma_n \mathbf{e}_{n-1}^{\mathbf{v}}}{h_n} + \mathcal{O}(h_n), \tag{E.4a}$$

$$\ddot{\mathbf{v}}(t_n) - \ddot{\mathbf{v}}_n = \frac{2\sigma_n^2 \sigma_{n-1}}{(1+\sigma_{n-1})} \frac{(\mathbf{e}_n^{\mathbf{v}} - (1+\sigma_{n-1})\mathbf{e}_{n-1}^{\mathbf{v}} + \sigma_{n-1}\mathbf{e}_{n-2}^{\mathbf{v}})}{h_n^2} + \mathcal{O}(h_n), \tag{E.4b}$$

$$\begin{aligned}\dddot{\mathbf{v}}(t_n) - \dddot{\mathbf{v}}_n = \frac{6\sigma_n^3 \sigma_{n-1}^2 \sigma_{n-2}}{h_n^3} \Bigg(& \frac{1}{(1+\sigma_{n-1})(1+\sigma_{n-2}+\sigma_{n-1}\sigma_{n-2})} \mathbf{e}_n^{\mathbf{v}} - \frac{1}{(1+\sigma_{n-2})} \mathbf{e}_{n-1}^{\mathbf{v}} \\ & + \frac{\sigma_{n-1}}{(1+\sigma_{n-1})} \mathbf{e}_{n-2}^{\mathbf{v}} - \frac{\sigma_{n-1}\sigma_{n-2}^2}{(1+\sigma_{n-2})(1+\sigma_{n-2}+\sigma_{n-2}\sigma_{n-1})} \mathbf{e}_{n-3}^{\mathbf{v}} \Bigg) + \mathcal{O}(h_n)\end{aligned} \tag{E.4c}$$

erfüllt. Mit den Gleichungen (5.30), (5.33) und (5.37d) folgt für $i = 1, ..., k$

$$\begin{aligned}\overline{\mathbf{v}}(t_{n+1-i}) - \overline{\mathbf{v}}_{n+1-i} = \mathbf{e}_{n+1-i}^{\mathbf{v}} - \mathbf{C}\big(q(t_{n+1-i})\big) \frac{\gamma_{i,n}}{\alpha_{i,n}} \left(\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}\right) \\ + \mathcal{O}(h_{n+1-i}) \|\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}\|,\end{aligned} \tag{E.5}$$

da $\mathbf{C}\big(q_n\big) = \mathbf{C}\big(q(t_n)\big) + \mathcal{O}(h_n)$ ist (vgl. (4.19a) und Voraussetzung 6). Wegen (5.38) und (5.33) ist mit (E.4) die Abschätzung (E.3) erfüllt. Einsetzen in (E.5) liefert mit (5.33) die Behauptung.

■

Nun können globale Fehlerrekursion für $\mathbf{e}_n^{\mathbf{v}}$ analog zu Satz 11 bestimmt werden.

Satz E.3
Die globalen Fehlerrekursionen von $\mathbf{e}_n^{\mathbf{v}}$ für die BLieDF-Verfahren (5.35) sind gegeben durch

$$\begin{aligned}\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{i,n} \mathbf{e}_{n+1-i}^{\mathbf{v}} = & -\frac{1}{h_n} \sum_{i=1}^{k} \gamma_{i,n} \mathbf{C}\big(q(t_{n+1-i})\big) \left(\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}\right) - \mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top \boldsymbol{\lambda}} \\ & + \mathcal{O}(1)\epsilon_{n+1} + \mathbf{l}_n^{\mathbf{M}^{-1}\mathbf{v}} + \mathcal{O}(h_n^k),\end{aligned} \tag{E.6a}$$

$$\begin{aligned}\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{i,n} \mathbf{e}_{n+1-i}^{\mathbf{B}\mathbf{v}} = & -\frac{1}{h_n} \sum_{i=1}^{k} \gamma_{i,n} \mathbf{B}\big(q(t_{n+1-i})\big) \left(\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}\right) - \mathbf{e}_{n+1}^{\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top \boldsymbol{\lambda}} \\ & + \mathcal{O}(1) \sum_{i=0}^{k+1} \epsilon_{n+1-i} + \mathbf{l}_n^{\mathbf{B}\mathbf{M}^{-1}\mathbf{v}} + \mathcal{O}(h_n^k),\end{aligned} \tag{E.6b}$$

$$\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{i,n} \mathbf{e}_{n+1-i}^{\mathbf{P}\mathbf{v}} = \mathcal{O}(1) \sum_{i=0}^{k+1} \epsilon_{n+1-i} + \mathcal{O}(h_n^k), \tag{E.6c}$$

Beweis:
Die Gleichung (E.6a) folgt aus der Differenz von (5.37c) mit (5.35c) und unter Verwendung von Lemma E.1. Durch Multiplikation mit $\mathbf{B}(q(t_{n+1}))$ bzw. $\mathbf{P}(q(t_{n+1}))$ folgen (E.6b) und (E.6c), da $[\mathbf{BC}](q) = \mathbf{B}(q)$ und $[\mathbf{PC}](q) = \mathbf{0}$ sind und unter Verwendung von (E.3) und Satz 14.

■

E.4 Globale Fehlerrekursion für $\mathbf{e}_n^{\boldsymbol{\lambda}}$

Auch für die globale Fehlerrekursion für $\mathbf{e}_n^{\boldsymbol{\lambda}}$ wird wie im konstanten Fall vorgegangen. Dazu kann zunächst das nachfolgende Lemma bewiesen werden, das analog zu Lemma 18 folgt.

Lemma E.2
Wenn $n \geq k-1$ und $k \geq i \geq 1$ ist, dann folgt die Abschätzung

$$\begin{aligned}&\mathbf{B}\big(q(t_{n+1-i})\big)\frac{\mathbf{e}_{n+2-i}^q - \mathbf{e}_{n+1-i}^q}{h_{n+1-i}} + \mathbf{Z}\big(q(t_{n+1-i})\big)\big(\mathbf{e}_{n+1-i}^q, \mathbf{v}(t_{n+1-i})\big)\\ &= \mathcal{O}\left(\frac{1}{h_{n+1-i}}\right)\max_r \|\boldsymbol{\Phi}(q_r)\| + \mathcal{O}(h_{n+1-i})\left(\|\mathbf{e}_{n+1-i}^q\| + \frac{1}{h_{n+1-i}}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right)\end{aligned}$$

mit dem Krümmungsterm $\mathbf{Z}$ aus (3.47).

Nun wird das Pendant zu Lemma 19 untersucht. Dafür wird ein Term $\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}$ eingeführt, für den

$$\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}} = \mathcal{O}(h_0^{k+1}), \quad (n+1 = -k, -k+1, ..., 0, 1, ..., k-1), \tag{E.7a}$$
$$\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}} = \mathbf{0}_{N\times 1}, \quad (n+1 \geq k), \tag{E.7b}$$

gilt.

Lemma E.3
Unter Voraussetzung 7 gibt es einen Term $\boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}$ mit (E.7) und es gilt für die BLieDF-Verfahren (5.35) die Abschätzung

$$\begin{aligned}\mathbf{e}_{n+1}^{\mathbf{Bv}} + \mathbf{B}\big(q(t_{n+1})\big)\mathbf{l}_n^{\boldsymbol{\omega}} = \mathcal{O}\left(\frac{1}{h_n}\right)\max_r \|\boldsymbol{\Phi}(q_r)\| &- \mathbf{B}\big(q(t_n)\big)\widehat{\mathbf{v}}(t_n)\sum_{i=1}^{k}\gamma_{i,n}\prod_{j=0}^{i-2}\frac{1}{\sigma_{n-j}}\mathbf{e}_{n+1-i}^q\\ &- \mathbf{Z}\big(q(t_n)\big)\left(\sum_{i=1}^{k}\gamma_{i,n}\prod_{j=0}^{i-2}\frac{1}{\sigma_{n-j}}\mathbf{e}_{n+1-i}^q, \mathbf{v}(t_n)\right) + \boldsymbol{\Delta}_{n+1}^{\mathbf{Bv}}\\ &+ \mathcal{O}(h_n)\left(\sum_{i=0}^{k}\epsilon_{n+1-i} + \frac{1}{h_n}\sum_{i=1}^{k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right)\end{aligned} \tag{E.8}$$

für $n+1 \geq -k$.

Beweis:
Der Beweis erfolgt analog zu Lemma 19 unter Verwendung der Startwerte aus Voraussetzung 7, Gleichung (E.2), Lemma E.2 und (5.40).

■

Schließlich kann eine Abschätzung für den globalen Fehler $\mathbf{e}_{n+1}^{\mathbf{S}\boldsymbol{\lambda}}$ bewiesen werden.

Satz E.4
Die globalen Fehler $\mathbf{e}_{n+1}^{\mathbf{S}\boldsymbol{\lambda}}$ der BLieDF-Verfahren (5.35) mit $2 \leq k \leq 3$ erfüllen die

Abschätzung

$$\mathbf{e}_{n+1}^{\mathbf{S}\boldsymbol{\lambda}} = -\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{i,n}\boldsymbol{\Delta}_{n-i+1}^{\mathbf{B}\mathbf{v}} + \mathcal{O}\left(\frac{1}{h_n^2}\right)\max_r\|\boldsymbol{\Phi}(q_r)\| + \mathcal{O}(1)\left(\sum_{i=0}^{2k}\epsilon_{n+1-i} + \frac{1}{h_n}\sum_{i=1}^{2k}\|\mathbf{e}_{n+1-i,0}^{\boldsymbol{\omega}}\|\right) + \mathcal{O}(h^k)$$

mit $\mathbf{S} \coloneqq \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top$.

Beweis:
Der Beweis gilt wie in Satz 12 mit den Gleichungen (E.6b) und (E.8), den Zusammenhängen (5.39) und (5.40) und Satz 14. Der zusätzliche Term

$$-\frac{1}{h_n}\sum_{i=1}^{k}\gamma_{i,n}\mathbf{B}\big(q(t_{n+1-i})\big)\big(\boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}(t_n) - \boldsymbol{\Delta}\mathbf{l}_{n-i}^{\boldsymbol{\omega}}\big),$$

vgl. Lemma E.1, enthält globale Fehler $\mathbf{e}_n^{\mathbf{B}\mathbf{v}}$ und wird daher analog zu dem Term $\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{i,n}\mathbf{e}_{n+1-i}^{\mathbf{B}\mathbf{v}}$ mithilfe von (E.8) eliminiert.

■

E.5 Gekoppelter Fehlerrekursion

Die Gleichungen für die globalen Fehler $\mathbf{e}_n^q$, $\mathbf{e}_n^{\mathbf{v}}$, $\mathbf{e}_{n,0}^{\boldsymbol{\omega}}$ und $\mathbf{e}_n^{\boldsymbol{\lambda}}$ aus den vorherigen Abschnitten können nun zu der Zwei-Term-Fehlerrekursion (4.42) kombiniert werden.

Lemma E.4
Die globalen Fehler der BLieDF-Verfahren (5.35) erfüllen für $n \geq k-1$ unter den Voraussetzungen 1, 5, 6 und 7 die gekoppelte Fehlerrekursion (4.42) mit

$$\mathbf{E}_n^{\mathbf{y}} \coloneqq \begin{bmatrix} \mathbf{e}_n^q \\ \vdots \\ \mathbf{e}_{n+1-2k}^q \\ \mathbf{e}_n^{\mathbf{P}\mathbf{v}} \\ \vdots \\ \mathbf{e}_{n+1-2k}^{\mathbf{P}\mathbf{v}} \end{bmatrix}, \quad \mathbf{E}_n^{\mathbf{z}} \coloneqq \begin{bmatrix} \frac{1}{h_{n-1}}\mathbf{e}_{n-1,0}^{\boldsymbol{\omega}} \\ \vdots \\ \frac{1}{h_{n+1-2k}}\mathbf{e}_{n+1-2k,0}^{\boldsymbol{\omega}} \\ \mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}} \\ \vdots \\ \mathbf{e}_{n+1-2k}^{\mathbf{S}\boldsymbol{\lambda}} \end{bmatrix},$$

$$\mathbf{T}_{\mathbf{y},n} \coloneqq \begin{bmatrix} \mathbf{T}_{\alpha,n} & \mathbf{0}_{2k\times 2k} \\ \mathbf{0}_{2k\times 2k} & \mathbf{T}_{\alpha,n} \end{bmatrix} \otimes \mathbf{I}_N, \quad \mathbf{T}_{\mathbf{z},n} = \begin{bmatrix} \mathbf{T}_{\gamma,n} & \mathbf{0}_{(2k-1)\times 2k} \\ \mathbf{0}_{2k\times(2k-1)} & \mathbf{J}_{2k} \end{bmatrix} \otimes \mathbf{I}_N,$$

mit

$$\mathbf{T}_{\alpha,n} = -\frac{1}{\alpha_{0,n}}\mathbf{e}_{1,2k}\cdot(\alpha_{1,n},\alpha_{2,n},...,\alpha_{k,n},0,...,0) + \mathbf{J}_{2k} \in \mathbb{R}^{2k\times 2k},$$

$$\mathbf{T}_{\gamma,n} = -\frac{1}{\gamma_{1,n}}\mathbf{e}_{1,2k-1}\cdot\left(\gamma_{2,n}\frac{1}{\sigma_n}, \gamma_{3,n}\prod_{j=0}^{1}\frac{1}{\sigma_{n-j}}, ..., \gamma_{k,n}\prod_{j=0}^{k-1}\frac{1}{\sigma_{n-j}}, 0, ..., 0\right) + \mathbf{J}_{2k-1} \in \mathbb{R}^{(2k-1)\times(2k-1)},$$

wobei $\mathbf{e}_{1,r} \in \mathbb{R}^r$ der erste Einheitsvektor ist und $\mathbf{J}_r \in \mathbb{R}^{r \times r}$ mit $\mathbf{J}_r = \left(j_{il}^{(r)}\right)_{i,l=1}^{r}$ und $j_{il}^{(r)} = \delta_{i+1,l}$.

Beweis:
Der Beweis folgt analog zu dem von Lemma 20, wobei die entsprechenden Gleichungen aus der konstanten Variante durch die variable Variante ersetzt werden. Lediglich bei Gleichung (E.1) muss der Zusammenhang (5.41) gelten.

■

Literaturverzeichnis

[1] M. Arnold and O. Brüls. Convergence of the generalized-α scheme for constrained mechanical systems. *Multibody System Dynamics*, 18:185–202, 2007.

[2] M. Arnold, O. Brüls, and A. Cardona. Error analysis of generalized-α Lie group time integration methods for constrained mechanical systems. *Numerische Mathematik*, 129:149–179, 2015.

[3] M. Arnold, B. Burgermeister, C. Führer, G. Hippmann, and G. Rill. Numerical methods in vehicle system dynamics: state of the art and current developments. *Vehicle System Dynamics*, 49:1159–1207, 2011.

[4] M. Arnold, A. Cardona, and O. Brüls. A Lie algebra approach to Lie group time integration of constrained systems. In P. Betsch, editor, *Structure-preserving Integrators in Nonlinear Structural Dynamics and Flexible Multibody Dynamics*, pages 91–158, Cham, 2016. Springer International Publishing.

[5] A. Baker. *Matrix Groups: An Introduction to Lie Group Theory*. Springer undergraduate mathematics series. Springer, London, United Kingdom, 2002.

[6] O. Brüls and M. Arnold. The Generalized-α Scheme as a Linear Multistep Integrator: Towards a General Mechatronic Simulator. *Journal of Computational and Nonlinear Dynamics*, 3:041007, 2008.

[7] O. Brüls, M. Arnold, and A. Cardona. Two Lie Group Formulations for Dynamic Multibody Systems with Large Rotations. In *Proceedings of the ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, pages 85–94, 2011.

[8] O. Brüls and A. Cardona. On the Use of Lie group Time Integrators in Multibody Dynamics. *Journal of Computational and Nonlinear Dynamics*, 5:031002, 2010.

[9] O. Brüls, A. Cardona, and M. Arnold. Lie group generalized-α time integration of constrained flexible multibody systems. *Mechanism and Machine Theory*, 48:121–137, 2012.

[10] J. Butcher and A. Heard. Stability of Numerical Methods for Ordinary Differential Equations. *Numerical Algorithms*, 31:59–73, 2002.

[11] G. D. Byrne and A. C. Hindmarsh. A Polyalgorithm for the Numerical Solution of Ordinary Differential Equations. *ACM Trans. Math. Softw.*, 1(1):71–96, 1975.

[12] M. Calvo, T. Grande, and R. D. Grigorieff. On the zero stability of the variable order variable stepsize BDF-Formulas. *Numerische Mathematik*, 57:39–50, 1990.

[13] E. Celledoni, A. Marthinsen, and B. Owren. Commutator-free Lie group methods. *Future Generation Computer Systems*, 19(3):341–352, 2003.

[14] D. Chevallier and J. Lerbet. *Multi-Body Kinematics and Dynamics with Lie Groups*. 2017.

[15] J. Chung and G. M. Hulbert. A Time Integration Algorithm for Structural Dynamics With Improved Numerical Dissipation: The Generalized-α Method. *Journal of Applied Mechanics*, 60(2):371–375, 1993.

[16] P. E. Crouch and R. Grossman. Numerical integration of ordinary differential equations on manifolds. *Journal of Nonlinear Science*, 3:1–33, 1993.

[17] C. F. Curtiss and J. O. Hirschfelder. Integration of Stiff Equations. In *Proceedings of the National Academy of Science of the United States of America*, volume 38, pages 235–243, 1952.

[18] P. Deuflhard and A. Hohmann. *Numerische Mathematik: Eine algorithmisch orientierte Einführung*, volume 1. De Gruyter, Berlin, 2008.

[19] P. Dobrinski, G. Krakau, and A. Vogel. *Physik für Ingenieure*. Teuber, Stuttgart, 4 edition, 1976.

[20] S. Faltinsen, A. Marthinsen, and H. Munthe-Kaas. Multistep methods integrating ordinary differential equations on manifolds. *Applied Numerical Mathematics*, 39:349–365, 2001.

[21] T. Fließbach. *Mechanik: Lehrbuch zur Theoretischen Physik I*. Springer-Verlag, Berlin Heidelberg, 8 edition, 2020.

[22] C. W. Gear. Algorithm 407: DIFSUB for solution of ordinary differential equations [D2]. *Commun. ACM*, 14(3):185–190, 1971.

[23] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall PTR, USA, 1971.

[24] M. Géradin and A. Cardona. *Flexible Multibody Dynamics: A Finite Element Approach*. John Wiley & Sons, Ltd., Chichester, 2001.

[25] K. Gopal, R. Sacks-Davis, and P. E. Tescher. A Review of Recent Developments in Solving ODEs. *ACM Comput. Surv.*, 17(1):5–47, 1985.

[26] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, 1989.

[27] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag, Berlin Heidelberg New York, 2 edition, 2006.

[28] E. Hairer, S. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer-Verlag, Berlin Heidelberg New York, 2 edition, 1993.

[29] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin Heidelberg New York, 2 edition, 1996.

[30] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Vieweg + Teubner Verlag, 2009.

[31] A. C. Hindmarsh. LSODE and LSODI, two new initial value ordinary differential equation solvers. *SIGNUM Newsl.*, 15(4):10–11, 1980.

[32] A. C. Hindmarsh and G. D. Byrne. Applications of EPISODE: An Effective Package for the Integration of Systems of Ordinary Differential Equations. In L. Lapidus and W. E. Schiesser, editors, *Numerical Methods for Differential Systems*, pages 147–166, New York, 1976.

[33] A. Iserles, H. Munthe-Kaas, S. Nørsett, and A. Zanna. Lie-Group Methods. *Acta Numerica*, 9:215–365, 2000.

[34] L. Jay and D. Negrut. Extensions of the HHT-α method to differential-algebraic equations in mechanics. *Electronic Transactions on Numerical Analysis*, 26:190–208, 2007.

[35] P. Knabner and L. Angermann. *Numerik partieller Differentialgleichungen: Eine anwendungsorientierte Einführung*. Masterclass. Springer Berlin Heidelberg, 2013.

[36] P. Lötstedt and L. Petzold. Numerical Solution of Nonlinear Differential Equations with Algebraic Constraints I: Convergence Results for Backward Differentiation Formulas. *Mathematics of Computation*, 46(174):491–516, 1986.

[37] C. Lunk and B. Simeon. Solving constrained mechanical systems by the family of Newmark and α-methods. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 86(10):772–784, 2006.

[38] W. Magnus. On the exponential solution of differential equations for a linear operator. *Communications on Pure and Applied Mathematics*, 7:649–673, 1954.

[39] J. Marsden and T. Ratiu. *Einführung in die Mechanik und Symmetrie: Eine grundlegende Darstellung klassischer mechanischer Systeme*. Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, 2001. übersetzt von S. Hackmann und U. Krähmer.

[40] A. Müller. Approximation of finite rigid body motions from velocity fields. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 90:514–521, 2010.

[41] A. Müller. Screw and Lie group theory in multibody dynamics: Recursive algorithms and equations of motion of tree-topology systems. *Multibody System Dynamics*, 43:37–70, 2018.

[42] H. Munthe-Kaas. Runge-Kutta methods on Lie groups. *BIT*, 38:92–111, 1998.

[43] H. Munthe-Kaas. High order Runge-Kutta methods on manifolds. *Applied Numerical Mathematics*, 29:115–127, 1999.

[44] H. Munthe-Kaas and B. Owren. Computations in a free Lie algebra. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357:957–981, 1999.

[45] I. Newton. *The Method of Fluxions and Infinite Series: With Its Application to the Geometry of Curve-lines*. Gedruckt durch Henry Woodfall; und verkauft durch John Nourse, London, 1736.

[46] K. Strehmel, R. Weiner, and H. Podhaisky. *Numerik gewöhnlicher Differentialgleichungen: Nichtsteife, steife und differential-algebraische Gleichungen*. Viweg+Teubner Verlag, 2 edition, 2012.

[47] E. Süli and D. Mayers. *An Introduction to Numerical Analysis*. Cambridge Univ. Press, Cambridge, 2003.

[48] Z. Terze, A. Müller, and D. Zlatar. Lie-group integration method for constrained multibody systems in state space. *Multibody System Dynamics*, 34:275–305, 2015.

[49] V. S. Varadarajan. *Lie groups, Lie algebras, and their representations*. Springer Verlag, New York, Berlin, Heidelberg, Tokyo, 1992.

[50] J. Wensch. Extrapolation methods in Lie groups. *Numerische Mathematik*, 89:591–604, 2001.

[51] V. Wieloch. *Analytisch äquivalente Lie-Gruppen-Beschreibungen: Ein numerischer Vergleich*. Bachelorarbeit, Martin-Luther-Universität Halle-Wittenberg, Institut für Mathematik, 2013.

[52] V. Wieloch. *Schrittweitensteuerung eines Generalized-α Lie-Gruppen-Integrators*. Masterarbeit, Martin-Luther-Universität Halle-Wittenberg, Institut für Mathematik, 2016.

[53] V. Wieloch and M. Arnold. Stability bounds for step size ratios in variable time step implementations of Newmark integrators. In *Proceedings of the ECCOMAS Thematic Conference on Multibody Dynamics*, Prague, 2017.

[54] V. Wieloch and M. Arnold. BLieDF2$^{\text{nd}}$ - a k-step BDF integrator for constrained mechanical systems on Lie groups. In *Proceedings of the 5th Joint International Conference on Multibody System Dynamics*, Lisbon, 2018.

[55] V. Wieloch and M. Arnold. BDF integrators for constrained mechanical systems on Lie groups. *Journal of Computational and Applied Mathematics*, 387:112517, 2021.

Lebenslauf

Persönliche Angaben

Name:	Wieloch
Vorname:	Victoria
Geburtsdatum:	11. Februar 1992
Geburtsort:	Halle (Saale)

Ausbildung

10/13 - 01/16:	Masterstudium Mathematik mit Anwendungsfach (Wirtschaft)
	Martin-Luther-Universität Halle-Wittenberg
	Abschluss: Master of Science in Mathematik
10/10 - 10/13:	Bachelorstudium Mathematik mit Anwendungsfach (Physik)
	Martin-Luther-Universität Halle-Wittenberg
	Abschluss: Bachelor of Science in Mathematik
07/04 - 07/10:	Schulausbildung am Gymnasium Landsberg
	Abschluss: Abitur

Beruflicher Werdegang

seit 10/20:	Prozessentwicklerin
	ComTS Finance
02/16-09/20:	Wissenschaftliche Mitarbeiterin am Institut für Mathematik
	Martin-Luther-Universität Halle-Wittenberg

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werke wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht.

Halle (Saale), den 28.08.2021 Victoria Wieloch