Karlsruher Institut für Technologie

Fehleranalyse von auf trigonometrischen Integratoren basierenden Splittingverfahren für hochoszillatorische, semilineare Probleme

Zur Erlangung des akademischen Grades eines

DOKTOR DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Simone Franziska Buchholz

Tag der mündlichen Prüfung: 06. November 2018

Referentin: Prof. Dr. Marlis Hochbruck
Korreferent: PD Dr. Volker Grimm

DANKSAGUNG

An dieser Stelle möchte ich mich bei allen bedanken, die auf ihre Art zu dieser Arbeit beigetragen haben.

Mein Dank geht an erster Stelle an meine Doktormutter Prof. Dr. Marlis Hochbruck. Ihre Art, Mathematik zu lehren, hat mich bereits im Grundstudium für die numerische Mathematik begeistern können. Sie gab mir zuerst die Möglichkeit eine Diplomarbeit in ihrer Arbeitsgruppe zu schreiben und bot mir danach eine Stelle als Doktorandin an. Dieses Vertrauen in meine Fähigkeiten hat meine Promotionszeit stark geprägt. In Zeiten des familiären Nachwuches setzte sie alle Hebel in Bewegung, um die Vereinbarkeit zwischen Job, Dissertation und Familie zu gewährleisten. Für die exzellente Betreuung möchte ich mich ganz herzlich bedanken.

Weiterhin möchte ich mich bei meinem Zweitbetreuer Herrn PD Dr. Volker Grimm bedanken. Seine Tür stand immer offen (auch wenn man im neuen Mathebau klopfen musste). Für zahlreiche mathematische Diskussionen, seine ruhige Art und die vielen hilfreichen Anmerkungen zu meiner Forschung möchte ich mich herzlich bedanken.

Diese Arbeit entstand im Rahmen des GRK 1294 und des SFB 1173. Für die finanzielle Unterstützung durch die deutsche Forschungsgemeinschaft (DFG) möchte ich mich herzlich bedanken. Die Möglichkeit, Teil eines Sonderforschungsbereichs zu sein, hat mir einen guten Einblick in das Feld der mathematischen Forschung ermöglicht. Vielen Dank auch an alle Mitglieder des SFBs für alle fachlichen Diskussionen, aber auch das freundschaftliche Miteinander.

Ebenso geht mein Dank an Ludwig Gauckler und Tobias Jahnke. Es war eine tolle und sehr lehreiche Erfahrung mit euch zusammen eine Veröffentlichung zu schreiben.

Anschließend möchte ich mich bei allen bedanken, die Teile der Arbeit Korrektur gelesen haben und meine zahlreichen Rechtschreibfehler ertragen mussten. Vielen Dank an Markus, Patrick, Michaela, Katharina, Jonas, Sarah und Chaya, insbesondere auch an Benjamin.

Ein ganz besonderer Dank geht an meine Arbeitsgruppe in jeglicher Besetzung. Die super Arbeitsatmosphäre und die gemeinsam verbrachte Freizeit bei diversen Feiern und Ausflügen werde ich stets in guter Erinnerung behalten. Hervorheben möchte ich speziell Markus und Volker, aber auch den Sekretärinnen Sonja und Laurette - vielen Dank für die ganze Organisation im Hintergrund.

Und natürlich gebührt Christian und Mathias Dank für die geduldige Beantwortung aller technischen Fragen.

Für jede Menge unnützes Wissen aber auch so manche fachliche Diskussion möchte ich mich bei der Mittagsessensrunde mit allen ihren Teilnehmern bedanken. Ob Julian, Andreas, Jonas, Mathias, Michaela, Christian, David, Lukas, Patrick, Robin, Constantin, Bernhard, die zwei Jans, Benjamin, Philip oder Michael, ich hoffe, wir trinken alle nochmal ein Bier zusammen.

Zu guter Letzt möchte ich mich bei meiner Familie und der Familie meines Mannes für eure bedingungslose Liebe bedanken. Ihr wart und seid mein Rückhalt. Ohne euch hätte ich diese Arbeit nie geschrieben. Bei meinen Schwiegereltern Birgit und Gerhard und meiner Mutter Gabi möchte ich mich außerdem für die zahlreichen Babysitterdienste bedanken. Ich konnte mich immer darauf verlassen, dass ihr den Kleinen zum Lachen bringt.

Zu allerletzt geht mein Dank an Julius, den kleinen Forscher und Entdecker, der alle Prioritäten in die richtige Richtung verschiebt und den ich so unendlich liebe. Und an den Liebsten, Markus, meine Sonne, meine Sterne. Euch widme ich diese Arbeit.

INHALTSVERZEICHNIS

EINLEITUNG

„This theorem with its many proofs is a striking illustration of the fact that there is more than one way of establishing the same truth.“

Elisha S. Loomis, *The Pythagorean Proposition*, 1940, Seite 3

Der Satz des Pythagoras ist einer der fundamentalen Sätze der euklidischen Geometrie. Den vermutlich ersten Beweis für seine Aussage lieferte Pythagoras von Samos bereits 540 v.Chr. Mittlerweile gibt es über 400 verschiedene Beweise dieses zentralen Satzes. Einen neuen Beweis für eine bekannte Aussage zu entwickeln, ist also durchaus keine Seltenheit in der Mathematik. So liefert eine neue Beweistechnik häufig eine neue Sichtweise auf das Problem. Sie hilft dabei, die Struktur des Problems besser zu verstehen. Weiterhin ist es oft möglich, Beweistechniken auf andere Probleme zu übertragen. So können Antworten auf bislang offene Fragestellungen gefunden werden.

In der vorliegenden Arbeit wird eine neue Beweistechnik zur Konvergenzanalyse von Splittingverfahren für hochoszillatorische, semilineare Probleme vorgestellt.

Motivation

Die Simulation hochoszillatorischer Phänomene ist in vielen Bereichen der Naturwissenschaften und Technik von großer Bedeutung, beispielsweise in der Moleküldynamik, bei elektromagnetischen Wechselwirkungen oder in der Astrophysik. Unter einem hochoszillatorischen Problem versteht man das Auftreten von Daten und Signalen auf verschiedenen Zeitskalen. Beispielsweise wird eine langsame Bewegung von kleinen, sehr schnellen Oszillationen überlagert. Dabei ist das Ziel der Simulation meist, die globale, langsame Bewegung darzustellen. Die schnellen Oszillationen kleiner Amplitude sind häufig nicht von Interesse.

Zur mathematischen Modellierung hochoszillatorischer Phänomene werden meist Differentialgleichungen verwendet. Die Lösung einer solcher Gleichung lässt sich häufig nicht exakt berechnen. Daher werden numerische Integrationsverfahren (Integratoren) eingesetzt, um eine Approximation an die Lösung zu bestimmen. Klassische, explizite Verfahren wie das Störmer-Verlet-Verfahren beruhen auf der Taylorentwicklung der exakten Lösung, vergleiche Störmer (1907) oder Verlet (1967). Die Stabilitäts- und Fehleranalyse solcher Verfahren gelingt nur, wenn das Produkt aus verwendeter Schrittweite und größter Frequenz des Systems klein ist. Die größten Frequenzen eines hochoszillatorischen Problems sind jedoch typischerweise sehr viel größer als die gewünschte Schrittweite,

denn diese Frequenzen erzeugen die schnellen Oszillationen der Lösung. Da man jedoch primär an der globalen Bewegung auf der langsamen Zeitskala interessiert ist und diese effizient approximieren möchte, sind große Schrittweiten wünschenswert. Klassische, explizite Verfahren sind für solche Schrittweiten instabil. Um sinnvolle Approximationen mit einem expliziten, klassischen Verfahren zu berechnen, muss daher eine äußerst starke Schrittweiteneinschränkung in Kauf genommen werden, siehe beispielsweise Hairer, Lubich und Wanner (2006), Kapitel I.5.2.

Diese Instabilität klassischer Integrationsverfahren motivierte die Entwicklung neuer spezieller Verfahren für die numerische Zeitintegration von hochoszillatorischen Problemen. Dabei entstand die Klasse der trigonometrischen Integratoren, auch Gautschi-artige Verfahren oder Lange-Zeitschritt-Verfahren genannt, die in der Klasse der exponentiellen Integratoren enthalten sind, vergleiche Hairer et al. (2006), Kapitel XIII. Sie unterliegen keiner Schrittweiteneinschränkung. Erste Prototypen dieser Klasse zeigten jedoch numerische Resonanzen, wie das Verfahren von Gautschi (1961), das Verfahren von Deuflhard (1979) oder die Impulsmethode von Grubmüller, Heller, Windemuth und Schulten (1991) und Tuckerman, Berne und Martyna (1992). Unter einer numerischen Resonanz versteht man das Auftreten großer Fehler des Verfahrens für einzelne, speziell gewählte Schrittweiten. Abhilfe konnte durch Einsatz von Filterfunktionen geschaffen werden, die erstmals in García-Archilla, Sanz-Serna und Skeel (1999) verwendet werden. Diese „filtern“ kritische Fehlerterme, sodass diese sich nicht über die Zeit aufsummieren und somit numerische Resonanzen verhindert werden können.

Konvergenzanalysen verschiedener trigonometrischer Integratoren mit geeigneten Filterfunktionen wurden bereits in zahlreichen Arbeiten durchgeführt, wie beispielsweise in García-Archilla et al. (1999), Hochbruck und Lubich (1999), Grimm und Hochbruck (2006), Hairer et al. (2006), Sanz-Serna (2008) oder jüngst in Gauckler (2015). Dabei wird Konvergenz zweiter Ordnung in den Positionen gleichmäßig in den hohen Frequenzen und unter einer Finite-Energie-Bedingung an die Lösung gezeigt. Die Fehleranalysen basieren im Wesentlichen auf der Variation-der-Konstanten--Formel. Sie sind sehr aufwendig, da die kritischen Fehlerterme herausgearbeitet und geeignet behandelt werden müssen. Daraus ergeben sich hinreichende Bedingungen an die Filterfunktionen, mithilfe derer Konvergenz bewiesen werden kann. Jedoch war es bislang nicht möglich, notwendige Bedingungen an die Filterfunktionen zu formulieren.

In Buchholz, Gauckler, Grimm, Hochbruck und Jahnke (2018) gelang es, einen neuen Beweis für trigonometrische Integratoren angewandt auf lineare, hochoszillatorische Probleme aufzustellen. Die grundlegende, bereits zuvor bekannte Idee des Beweises beruht darauf, trigonometrische Integratoren in Form eines Strang-Splittingverfahrens angewandt auf eine modifizierte Gleichung zu schreiben. Neu ist dabei nun, Konvergenz mit Beweistechniken aus der Analyse von Splittingverfahren zu zeigen, wie beispielsweise die iterierten Kommutatoren oder eine Variante des Fächers der Lady Windermere, vergleiche Jahnke und Lubich (2000) oder Lubich (2008). Die neue Beweistechnik schließt damit die Lücke zwischen der Analyse von trigonometrischen Integratoren für hochoszillatorische Lösungen und der Analyse der Splittingverfahren für glatte Lösungen.

Ziele und Ergebnisse

Das zentrale Ziel der vorliegenden Arbeit ist es, den Beweis aus Buchholz et al. (2018) auf semilineare, hochoszillatorische Probleme zu erweitern. Untersucht wird dabei eine neue Klasse von trigonometrischen Integratoren, bei welcher das Splittingverfahren in vertauschter Reihenfolge ver-

wendet wird. Die neue Beweistechnik hilft somit, weitere Verfahren für hochoszillatorische Probleme zu entwickeln. Die Arbeit enthält weiterhin eine numerische Simulation des berühmten Fermi-Pasta-Ulam-Tsingou-Problems. Die Testergebnisse bekräftigen die Konvergenzaussage. Eine erweiterte Anwendung auf eine forschungsrelevante Problemstellung in der Laser-Plasma-Simulation wird ebenfalls diskutiert. Hier zeigt sich die Allgemeinheit der in der vorliegenden Arbeit betrachteten Problemstellung und des verwendeten Integrationsverfahrens.

Die Veröffentlichung Buchholz et al. (2018) stellt eine Vorarbeit meiner Promotion dar. Überträgt man den in Buchholz et al. (2018) geführte Beweis für lineare Probleme auf ein Splittingverfahren in geänderter Reihenfolge, so ist dieser Beweis ein Spezialfall des Konvergenzbeweises für semilineare Probleme der vorliegenden Arbeit und wird deshalb hier nicht separat geführt. Teile des Beweises für lineare Probleme werden allerdings als Motivation und zur Illustration der Beweistechnik aufgeführt.

Gliederung

Im ersten Kapitel wird die in dieser Arbeit betrachtete Differentialgleichung vorgestellt und dargelegt, weshalb ihre Lösung hochoszillatorisch ist. Weiterhin werden bereits bekannte Integrationsverfahren definiert und die Schwierigkeiten bei der numerischen Simulation hochoszillatorischer Probleme illustriert.

Das zweite Kapitel stellt die Grundlage für den zentralen Konvergenzbeweis dar. Zunächst werden einige Grundlagen zu Splittingverfahren wiederholt. Danach werden zwei Arbeiten zur Konvergenz von Splittingverfahren vorgestellt. Da der Konvergenzbeweis der vorliegenden Arbeit auf Beweistechniken dieser zwei Arbeiten beruht, werden sie kurz erläutert. Anschließend wird die historische Entwicklung der trigonometrischen Integratoren vorgestellt. Dieser Abschnitt liefert einerseits einen Überblick der bisherigen Arbeiten auf diesem Gebiet, andererseits wird auch hier auf die Beweistechniken eingegangen, um den Unterschied zur Konvergenzanalyse der vorliegenden Arbeit aufzuzeigen. Auch die Ergebnisse der Vorarbeit Buchholz et al. (2018) sind hier zusammengefasst.

Im dritten Kapitel werden trigonometrische Integratoren als Splittingverfahren dargestellt. Dazu wird zunächst eine modifizierte Gleichung eingeführt, welche durch die Filterfunktionen als Störung der ursprünglichen Gleichung aufgefasst wird. Das Kernstück der Arbeit befindet sich im vierten Kapitel. Hier wird die Fehleranalyse dieser auf trigonometrischen Integratoren basierenden Splittingverfahren durchgeführt. Die in diesem Kapitel vorgestellten Sätze und Beweise sind neu.

Die im vierten Kapitel gewonnenen Erkenntnisse werden im fünften Kapitel durch numerische Simulation des Fermi-Pasta-Ulam-Tsingou-Problems illustriert. Eine Übertragung des Konvergenzbeweises auf ein Verfahren zur numerischen Simulation der Wechselwirkungen zwischen einem Laser und hochdichtem Plasma findet im sechsten Kapitel statt.

KAPITEL 1

SEMILINEARE DIFFERENTIALGLEICHUNG ZWEITER ORDNUNG MIT HOCHOSZILLATORISCHER LÖSUNG

Zur Modellierung der Phänomene aus der Naturwissenschaft und technischer Systeme werden meist Differentialgleichungen verwendet. Das einfachste Beispiel ist $F = ma$, das Newton'sche Gesetz, das eine Beziehung zwischen der Masse m eines Körpers, seiner Beschleuigung a und der auf ihn wirkenden Kraft F herstellt. Die Position q des Körper ergibt sich nun aus zweimaliger Integration, denn die zweite zeitliche Ableitung der Positionen entspricht der Beschleuigung $q''(t) = a(t)$. Sind die auftretenden Kräfte linear, so erhält man eine Differentialgleichung in der Positionen q der Form, wie sie nachfolgend betrachtet wird.

Kompliziertere, physikalische Prozesse werden häufig durch partielle Differentialgleichungen beschrieben. Hier treten neben der zeitlichen Ableitung auch räumliche Ableitungen auf. Ein bekanntes Beispiel ist die inhomogene Wellengleichung in einer Raumdimension

$$\frac{\partial q}{\partial t^2}(x,t) = \frac{\partial q}{\partial x^2}(x,t) + f(q).$$

Durch räumliche Semidiskretisierung zum Beispiel mithilfe einer Spektralmethode kann man eine gewöhnliche Differentialgleichung erhalten, wie sie nachfolgend betrachtet wird (siehe dazu auch Abschnitt 2.2.6).

Die vorliegende Arbeit handelt jedoch primär davon, die Lösung $q : [0, t_{\text{end}}] \to \mathbb{R}^d$ der gewöhnlichen Differentialgleichung

$$\begin{aligned} q''(t) &= -\Omega^2 q(t) + g(q(t)), \quad 0 \le t \le t_{\text{end}} \\ q(0) &= q_0,\ q'(0) = q'_0. \end{aligned} \tag{1.1}$$

numerisch zu approximieren. In der folgenden Annahme werden die Voraussetzungen an die auftretende Matrix Ω, die Lösung q und die Funktion g zusammengefasst, die später für die Konvergenzanalyse in Kapitel 4 benötigt werden. Sie liefern auch Existenz der Lösung, vergleiche beispielsweise Satz 2.5.6 aus Aulbach (2004).

Annahme 1.1. *a) Die Matrix Ω sei symmetrisch und positiv semi-definit mit beliebig großer Norm $\|\Omega\| \gg 1$, wobei $\|\cdot\|$ die Euklidnorm oder die durch die Euklidnorm induzierte Matrixnorm bezeichnet.*

b) Die exakte Lösung $q(t)$ erfüllt die Finite-Energie-Bedingung

$$\|\Omega q(t)\|^2 + \left\|q'(t)\right\|^2 \leq K^2, \qquad 0 \leq t \leq t_{end}. \tag{1.2}$$

c) Es seien r_0, r_1 und r_2 positive reele Zahlen. Die Nichtlinearität $g \in C^2(\mathbb{R}^d)$ und ihre erste und zweite Ableitung seien auf Kugeln mit Radius $\widetilde{r}_0, \widetilde{r}_1$ beziehungsweise $\widetilde{r}_2$ beschränkt. Genauer gilt: Es existieren nichtnegative Konstanten $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$ und $\widetilde{C_{g,2}}$ mit

$$\|g(q)\| \leq \widetilde{C_{g,0}}, \qquad \text{für } \|q\| \leq \widetilde{r}_0 \tag{1.3a}$$
$$\left\|g'(q)\right\| \leq \widetilde{C_{g,1}}, \qquad \text{für } \|q\| \leq \widetilde{r}_1 \tag{1.3b}$$
$$\left\|g''(q)\right\| \leq \widetilde{C_{g,2}}, \qquad \text{für } \|q\| \leq \widetilde{r}_2 \tag{1.3c}$$

d) Es sei g Lipschitz-stetig mit Lipschitz-Konstante L_g:

$$\|g(v) - g(w)\| \leq L_g \|v - w\|. \tag{1.4}$$

Um in Kapitel 2 auf die Konvergenzanalysen in der Literatur einzugehen, werden die Annahmen leicht verschärft. Bedingung d) entfällt hier, da sie direkt aus $\|g'(q)\| \leq \widetilde{C_{g,1}}$ folgt.

Annahme 1.2. *Es gelten die Bedingungen a) und b) aus Annahme 1.1. Weiterhin seien die Nichtlinearität $g \in C^2(\mathbb{R}^d)$ und ihre erste und zweite Ableitung beschränkt für alle $q \in \mathbb{R}^d$, das heißt es existieren nichtnegative Konstanten $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$ und $\widetilde{C_{g,2}}$ mit*

$$\|g(q)\| \leq \widetilde{C_{g,0}}, \tag{1.5a}$$
$$\left\|g'(q)\right\| \leq \widetilde{C_{g,1}}, \tag{1.5b}$$
$$\left\|g''(q)\right\| \leq \widetilde{C_{g,2}}, \tag{1.5c}$$

Bemerkung 1.3. *Die Zahlen $\widetilde{r}_0$, $\widetilde{r}_1$ und $\widetilde{r}_2$ werden in obiger Annahme nicht genau bestimmt. Für die Konvergenzanalyse lassen sich jedoch maximale Kugelradien angeben, für die Konvergenz gezeigt werden kann, vergleiche* (4.21)*. Da im Konvergenzbeweis diese Größen von einer modifizierten Gleichung abhängen, dessen Lösung mit $\widetilde{u}$ bezeichnet werden wird, werden sie ebenfalls mit Tilde bezeichnet.*

Bemerkung 1.4. *Die Finite-Energie-Bedingung wird hier für alle Zeiten $0 \leq t \leq t_{end}$ gefordert. Es genügt allerdings, sie nur für $t = 0$ zu fordern, siehe dazu Lemma B.1. Für die meisten relevanten Probleme aus den Naturwissenschaften lässt sich eine solche Energiebedingung aus der Erhaltung der Gesamtenergie des Systems ableiten. Teilweise hat sie sogar physikalische Bedeutung, wie die harmonische Energie von Systemen der Moleküldynamik. Auch die numerischen Experimente in Kapitel 5 oder das Beispiel der Laser-Plasma-Simulation in Kapitel 6 zeigen, dass eine solche Bedingung physikalisch sinnvoll ist.*

Aufgrund der großen Eigenschwingungen der Matrix Ω und der niedrigen Regularitätsvoraussetzung an die Lösung q in Form der Finite-Energie-Bedingung bezeichnet man die Differentialgleichung (1.1) als oszillatorisch. Ihre Lösung q wird meist durch sehr schnelle Schwingungen kleiner

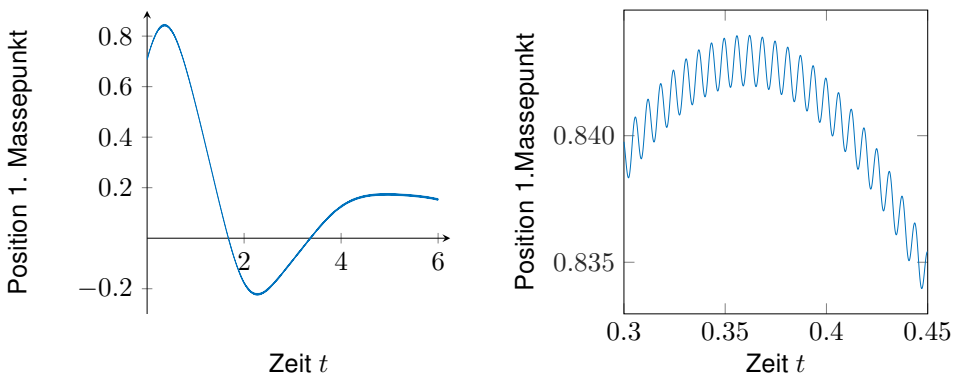

Abbildung 1.1: Darstellung der Simulation des ersten Massepunktes q_1 (hochoszillatorische Lösung) des Fermi-Pasta-Ulam-Tsingou-Problems für $t \in [0, 6]$ und in der Detailansicht (zur Modellbildung siehe Abschnitt 5.3).

Amplitude charakterisiert, wie sie beispielsweise in Abbildung 1.1 zu sehen sind. Im linken Bild erkennt man die glatte, langsame Bewegung der Lösung über sechs Zeiteinheiten. Betrachtet man die Lösung jedoch im rechten Bild mit höherer Auflösung, so erkennt man, dass diese langsame Bewegung von schnellen Oszillationen überlagert wird.

Solche Lösungen werden als hochoszillatorisch bezeichnet und sie sind zentrales Untersuchungsobjekt der vorliegenden Arbeit. Das oszillatorische Verhalten der Lösung erschwert die numerische Integration und erfordert die Entwicklung spezieller Verfahren zur effizienten, numerischen Approximation der Lösung. Einige bereits bekannte Integratoren werden in Kapitel 1.1 vorgestellt. Die Herausforderungen bei der numerischen Integration werden schließlich in Kapitel 1.2 erläutert.

1.1 Bekannte Integrationsverfahren für hochoszillatorische Probleme

In diesem Abschnitt wird zunächst das Störmer-Verlet-Verfahren und die Klasse trigonometrischer Integratoren vorgstellt. Die Klasse der trigonometrischen Integratoren ist in dieser Arbeit von zentraler Bedeutung. Von ihr leitet sich das Splittingverfahren ab, das in der vorliegenden Arbeit analysiert wird. Das Störmer-Verlet-Verfahren wiederum wird häufg in der Praxis mit sehr kleiner Schrittweite verwendet. Im letzten Abschnitt des Kapitels werden noch weitere Integrationsverfahren für hochoszillatorische Probleme vorgestellt, die sich aber wesentlich von der Klasse der trigonometrischen Integratoren unterscheiden.

1.1.1 Störmer-Verlet-Verfahren

Das Störmer-Verlet-Verfahren ist schon lange bekannt und sehr populär. Dieses Kapitel orientiert sich an der Darstellung in Hairer et al. (2006). Das Verfahren wurde von mehreren Personen unabhängig voneinander entwickelt. Verlet (1967) schlug es für Anwendungen in der Moleküldynamik vor, welche meist hochoszillatorische Probleme darstellen. In der Praxis wird es für die Simulation von Molekülen immer noch häufig verwendet, da es hervorragende Langzeiteigenschaften hat, einfach zu implementieren ist und jeder einzelne Zeitschritt günstig zu berechnen ist. Letztere Eigenschaft ist essentiell, denn das Verfahren ist nicht für alle Schrittweiten stabil. Es muss eine scharfe Schritt-

weiteneinschränkung in Kauf genommen werden, sodass sehr viele Zeitschritte zur Berechnung einer geeigneten Approximation notwendig sind, vergleiche Abschnitt 1.2.

In Störmer (1907) wurde diese Methode zur Untersuchung der Nordlichter vorgeschlagen. In der Astronomie schätzt man das Verfahren besonders für seine Symplektizität (siehe dazu Hairer et al. (2006), Theorem 3.4). Diese führt zu einem guten Langzeitverhalten, denn Erhaltungsgrößen wie die Energie des physikalischen Systems werden sehr gut approximiert. Die vorliegende Arbeit beschäftigt sich jedoch primär mit der Konvergenz der Verfahren in endlicher Zeit.

Das Störmer-Verlet-Verfahren berechnet numerische Approximationen $q_n \approx q(n\tau)$ der Lösung q der Differentialgleichung (1.1) mit Schrittweite τ durch

$$q_{n+1} - 2q_n + q_{n-1} = \tau^2(-\Omega^2 q_n + g(q_n)). \tag{1.6}$$

Man ersetzt hierfür die zweite Ableitung der Lösung q in (1.1) durch den zweiten zentralen Differenzenquotient. Es handelt sich also um ein Zweischrittverfahren, welches Startwerte q_0 und q_1 benötigt. Der zweite Startwert ergibt sich aus einer Taylorapproximation 2. Ordnung, also

$$q_1 = q_0 + \tau q_0' + \frac{\tau^2}{2}(-\Omega q_0 + g(q_0)).$$

Das Verfahren lässt sich auch als Einschrittverfahren schreiben. Dazu führt man folgende neue Variable zur Approximation der Ableitung $p = q'$ an entsprechenden Stellen ein:

$$p_n = \frac{q_{n+1} - q_{n-1}}{2\tau} \qquad \text{und} \qquad p_{n+1/2} = \frac{q_{n+1} - q_n}{\tau}$$

Daraus erhält man direkt

$$p_{n+1/2} + p_{n-1/2} = 2p_n$$

und mit (1.6)

$$p_{n+1/2} - p_{n-1/2} = \tau(-\Omega^2 q_n + g(q_n)).$$

Mithilfe dieser Umformungen ergibt sich nach Elimination von $p_{n-1/2}$ die folgende Einschritt-Formulierung:

$$\begin{aligned} p_{n+1/2} &= p_n + \frac{\tau}{2}(-\Omega^2 q_n + g(q_n)) \\ q_{n+1} &= q_n + \tau p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} + \frac{\tau}{2}(-\Omega q_{n+1} + g(q_{n+1})). \end{aligned} \tag{1.7}$$

1.1.2 Klasse der trigonometrischen Integratoren

Die hier beschriebene Klasse der trigonometrische Integratoren findet sich in dieser Form in Grimm und Hochbruck (2006). Zur geschichtlichen Entwicklung dieser Klasse von Integrationsverfahren sei auf Kapitel 2.2 verwiesen.

Zur Herleitung des Verfahrens schreibt man zunächst die Differentialgleichung (1.1) als System erster Ordnung. Man erhält Lösungen $q, q' : [0, t_{\text{end}}] \to \mathbb{R}^d$, welche die Gleichung erster Ordnung

$$\begin{bmatrix} q'(t) \\ q''(t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ -\Omega^2 & 0 \end{bmatrix} \begin{bmatrix} q(t) \\ q'(t) \end{bmatrix} + \begin{bmatrix} 0 \\ g(q(t)) \end{bmatrix} \tag{1.8}$$

erfüllen. Nun wendet man die Variation-der-Konstanten-Formel auf das System an:

$$\begin{bmatrix} q(t) \\ q'(t) \end{bmatrix} = \begin{bmatrix} \cos(t\Omega) & \Omega^{-1}\sin(t\Omega) \\ -\Omega\sin(t\Omega) & \cos(t\Omega) \end{bmatrix} \begin{bmatrix} q(0) \\ q'(0) \end{bmatrix} + \int_0^t \begin{bmatrix} \Omega^{-1}\sin\left((t-s)\Omega\right) \\ \cos\left((t-s)\Omega\right) \end{bmatrix} g(q(s))\, ds. \tag{1.9}$$

Die Klasse der trigonometrischen Integratoren erhält man nun, wenn man das Integral durch eine geeignete Approximation ersetzt:

$$\begin{bmatrix} q_{n+1} \\ q'_{n+1} \end{bmatrix} = \begin{bmatrix} \cos(\tau\Omega) & \Omega^{-1}\sin(\tau\Omega) \\ -\Omega\sin(\tau\Omega) & \cos(\tau\Omega) \end{bmatrix} \begin{bmatrix} q_n \\ q'_n \end{bmatrix} + \begin{bmatrix} \frac{1}{2}\tau^2\Psi g(\Phi q_n) \\ \frac{1}{2}\tau(\Psi_0 g(\Phi q_n) + \Psi_1 g(\Phi q_{n+1})) \end{bmatrix} \tag{1.10a}$$

mit

$$\Phi = \phi(\tau\Omega), \qquad \Psi = \psi(\tau\Omega), \qquad \Psi_0 = \psi_0(\tau\Omega), \qquad \Psi_1 = \psi_1(\tau\Omega). \tag{1.10b}$$

Zur Definition der hier auftretenden Matrixfunktionen siehe Abschnitt A.4. Die Funktionen ϕ, ψ, ψ_0 und ψ_1 werden Filterfunktionen genannt. Sie filtern kritische Fehlerterme heraus, die sonst zu einer Ordnungsreduktion des Verfahrens führen würden, näheres dazu findet man in Kapitel 1.2. Zunächst wird lediglich angenommen, dass die Filterfunktionen gerade und analytisch sind. Außerdem gelte

$$\phi(0) = \psi(0) = \psi_0(0) = \psi_1(0) = 0,$$

um die Konsistenz des Verfahrens zu gewährleisten. Man kann einfach nachweisen, dass der trigonometrische Integrator symmetrisch ist, wenn

$$\psi(z) = \operatorname{sinc}(z)\psi_1(z) \qquad \text{und} \qquad \psi_0(z) = \cos(z)\psi_1(z) \tag{1.11}$$

erfüllt sind, indem man $n \leftrightarrow n+1$ und $\tau \leftrightarrow -\tau$ vertauscht. Für ein symmetrisches Verfahren kann man eine Zweischritt-Formulierung unter Verwendung der Zeit-Umkehrbarkeit herleiten:

$$q_{n+1} - 2\cos(\tau\Omega)q_n + q_{n-1} = \tau^2\Psi g(\Phi q_n)$$

mit Startschritt

$$q_1 = \cos(\tau\Omega)q_0 + \Omega^{-1}\sin(\tau\Omega)q'_0 + \frac{1}{2}\tau^2\Psi g(\Phi q_0).$$

1.1.3 Weitere numerische Verfahren für hochoszillatorische Probleme

In diesem Abschnitt werden kurz weitere Integrationsverfahren vorgestellt, die sich für hochoszillatorische Probleme etabliert haben. Sie unterscheiden sich wesentlich von den in dieser Arbeit betrachteten trigonometrischen Intergratoren in den Voraussetzungen an die Differentialgleichung, der Idee des Integrationsverfahrens und den speziellen Bedingungen, unter denen Konvergenz garantiert werden kann. Eine gute Übersicht zu hochoszillatrosichen Problemen bieten Hairer et al. (2006) oder Engquist, Fokas, Hairer und Iserles (2009).

Für zeitabhängige, lineare Probleme wurden Magnus-Integratoren vorgeschlagen, vergleiche Magnus (1954) oder Hochbruck und Ostermann (2010) und Literaturreferenzen in letzterer Arbeit. Eine numerische Approximation wird hier mittels der Magnus-Reihe bestimmt. Wenn die zeitlichen

Ableitungen der zeitabhängigen rechten Seite der Differentialgleichung beschränkt sind, lässt sich Konvergenz dieser Verfahren zeigen, siehe beispielsweise Hochbruck und Lubich (1999) oder Hochbruck und Lubich (2003).

Neben trigonometrischen Integratoren und Magnus-Integratoren gibt es noch solche, die nicht die hochoszillatorische Lösung direkt approximieren, sondern lediglich das globale Verhalten dieser Lösung. Dafür wird zuerst eine Limitgleichung hergeleitet, deren Lösung dann das globale Verhalten wiedergibt und die einfacher zu approximieren ist. Solche Verfahren werden meist für spezielle partielle Differentialgleichung konstruiert, wohingegen sich die vorliegende Arbeit mit gewöhnlichen Differentialgleichung beschäftigt.

Zu nennen ist hier einerseits die Klasse der adiabatischen Integratoren, die man beispielsweise in Hairer et al. (2006), Abschnitt XIV.1, findet. Hier erhält man durch eine adiabatische Transformation eine Differentialgleichung, deren rechte Seite beschränkt ist (auch wenn sie stark oszilliert). Die Lösung dieser transformierten Gleichung lässt sich nun approximieren. In Hairer et al. (2006) findet man eine ausführliche Übersicht von Konvergenzresultaten solcher Verfahren und zahlreiche weiterführende Literatur. Diese Verfahren sind weiterhin Untersuchungsobjekt aktueller Forschung, siehe beispielsweise Jahnke und Mikl (2018), die adiabatische Integratoren für die „dispersion managment nonlinear Schrödinger equation“ (nichtlineare Schrödingergleichung zur Simulation der Streuung in Lichtwellenleitern) vorschlugen und Konvergenz dieser Verfahren bewiesen.

Ein vergleichbarere Ansatz wurde auch in Faou und Schratz (2014) und Krämer und Schratz (2017) verfolgt. Hier wurde eine Limitgleichung für die Klein-Gordon- beziehungsweise Maxwell-Klein-Gordon-Gleichung mit hochoszillatorischer Lösung hergeleitet. Das hochoszillatorische Verhalten hängt hier wesentlich von einem betragsmäßig großen Parameter der Gleichung ab. Durch einen Multiskalenansatz und modulierte Fourierentwicklung konnten Limitgleichungen hergeleitet werden, die unabhängig von diesem großen Parameter sind. Daher ist eine numerische Approximation mithilfe von Splittingverfahren möglich (für Splittingverfahren siehe Abschnitt 2.1, für modulierte Fourierentwicklung siehe Hairer, Lubich und Wanner (2003), Kapitel XIII.

Nachteil der Limitgleichungen ist häufig, dass der Ansatz nur dann gut funktioniert, wenn der Parameter der Gleichung tatsächlich sehr groß ist. In Baumstark, Faou und Schratz (2018) wurden exponentiell-artige Integratoren für die kubische Klein-Gordon-Gleichung entwicklet, die gleichmäßig konvergent sind. Das heißt, dass sie nicht nur hochoszillatorische Lösungen der Gleichung mit großem Parameter gut approximieren, sondern auch Lösungen der Gleichung mit kleinem Parameter. Hierbei wird ebenfalls eine Transformation ähnlich zur modulierten Fourierentwicklung eingesetzt, nur wird hier nicht der Grenzfall für große Parameter betrachtet. Es wird weiterhin die Gleichung gelöst, die das exakte Verhalten der Lösung wiederspiegeln.

Oben erwähnte Arbeiten machen deutlich, dass Limitgleichungen und ihre Approximation ein interessantes Forschungsgebiet im Bereich der hochoszillatorischen Probleme darstellen. Jedoch ist zur Herleitung der Limitgleichungen sehr viel Wissen um das konkrete Problem nötig. Der trigonometrische Integrator bietet demgegenüber die Möglichkeit, eine Vielzahl von Problemen mit vertretbarem Aufwand zu simulieren.

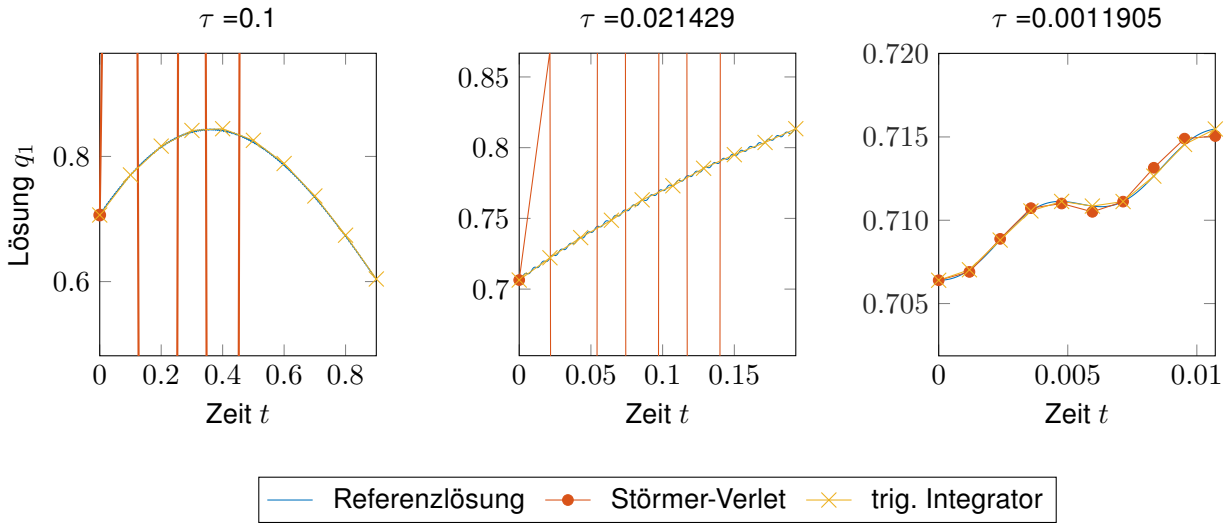

Abbildung 1.2: Referenzlösung q_1 und ihre numerischen Approximationen für verschiedene Schrittweiten τ

1.2 Schwierigkeiten bei der Simulation hochoszillatorischer Probleme

In der vorliegenden Arbeit sollen explizite Verfahren verwendet werden, um eine numerische Approximation an die Lösung q von (1.1) herzuleiten. Dabei unterliegen explizite Standardverfahren wie beispielsweise das Strömer-Verlet-Verfahren aus Kapitel 1.1.1 einer Stabilitätsbedingung, häufig CFL-Bedingung genannt (CFL steht dabei für die Mathematiker Courant, Friedrichs und Lewy, die diese Bedingungen definierten). Für das Störmer-Verlet-Verfahren angewandt auf (1.1) gilt beispielsweise

$$\tau\omega_{\max} \leq 2,$$

wobei $\omega_{\max}$ der größte Eigenwert der symmetrisch und positiv-definiten Matrix Ω ist und angenommen wird, dass die Funktion $g \equiv 0$ ist. Wählt man größere Schrittweiten, so wird das Verfahren instabil. Das bedeutet, dass sich die Approximationen sehr schnell von der exakten Lösung entfernen und somit die Lösung nicht mehr geeignet approximiert wird. Der Fehler wird sehr groß.

Dieses Verhalten wird in Abbildung 1.2 illustriert. Hier wurde die Bewegung q_1 des ersten Massepunktes im Fermi-Pasta-Ulam-Tsingou-Problem mit verschiedenen Schrittweiten jeweils über zehn Zeitschritte lang simuliert (für die Problemstellung siehe Kapitel 5.3). Das Störmer-Verlet-Verfahren berechnet nur für die Schrittweite $\tau = 0.0011905$ eine sinnvolle Approximation an die Referenzlösung. Für die beiden größeren Schrittweiten ist bereits die erste Nährung $y_1 \approx y(t_1) = y(\tau)$ schon weit von der Referenzlösung entfernt. Nach wenigen Schritte bricht das Verfahren ab, da die Approximationen den darstellbaren Zahlenbereich von MATLAB verlassen hat.

Die Instabilität des Verfahrens lässt sich folgendermaßen erklären: Betrachtet man die Einschritt-Formulierung des Störmer-Verlet-Verfahrens (1.7) angewandt die Testgleichung $q'' = -\omega^2 q$, so lässt sich leicht eine Iterationsmatrix M_{SV} angeben, für die

$$\begin{bmatrix} q_{n+1} \\ p_{n+1} \end{bmatrix} = M_{\text{SV}} \begin{bmatrix} q_n \\ p_n \end{bmatrix} = M_{\text{SV}}^{n+1} \begin{bmatrix} q_0 \\ p_0 \end{bmatrix}$$

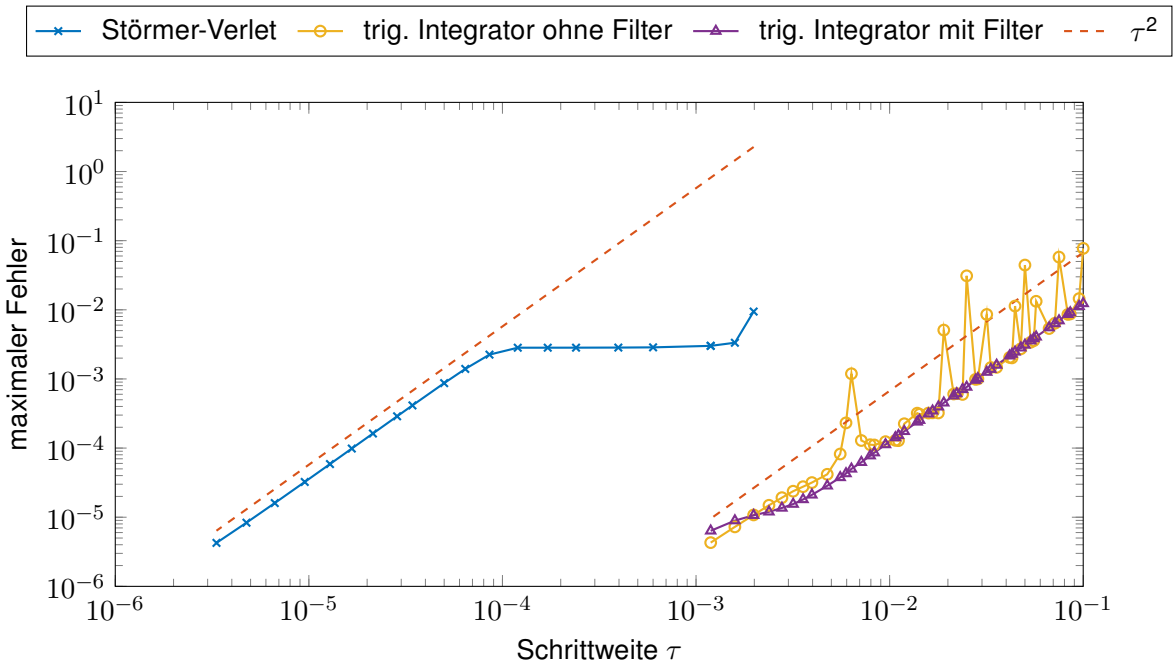

Abbildung 1.3: Maximaler Fehler der Positionen q über alle Zeitschritte mit Schrittweite τ bis $t = t_{\text{end}}$

gilt. Die Matrix M_{SV} hat jedoch für $\tau\omega > 2$ Eigenwerte größer als eins. Anteile des Startvektors in Richtung der Eigenvektoren für Eigenwerte größer als eins werden also mit jedem Schritt verstärkt. Für geeignete Startwerte werden die Approximationen also beliebig groß. Somit verstärkt das Verfahren für $\tau\omega > 2$ häufig auch Approximations- und Rundungsfehler.

Der große Fehler des Störmer-Verlet-Verfahren für große Schrittweiten entsteht jedoch auch dadurch, dass das Verfahren die zweite Ableitung der Lösung in der Differentialgleichung (1.1) durch eine Taylor-Approximation zweiter Ordnung diskretisiert. Der dabei auftretende Fehler hängt von der Ableitung der rechten Seite ab. Für glatte Probleme sind die rechte Seite und ihre Ableitungen mit kleiner Konstante beschränkt. Diese Annahme gilt jedoch nicht für hochoszillatorische Probleme. Hier geht man davon aus, dass $\tau\omega_{\text{max}} \gg 1$ ist, wobei ω_{max} der größte Eigenwert der Matrix Ω ist. Daher ist im oszillatorischen Fall der Fehler bei Approximation durch die Taylorentwicklung sehr groß.

Anschaulich berechnet das Störmer-Verlet-Verfahren die neue Approximation aus Information der Krümmung der exakten Lösung. Jedoch unterscheidet das Verfahren dabei nicht zwischen den schnellen Oszillationen und der globalen Bewegung der Lösung. Die feinen Oszillationen können aber nicht aufgelöst werden, wenn die Schrittweite zu groß ist. Das Verfahren verwendet somit die Krümmung der Oszillationen um das Verhalten der globalen Bewegung zu simulieren. Somit ist es mit dem Störmer-Verlet-Verfahren und großen Schrittweiten τ weder möglich die globale Bewegung der Lösung noch die schnellen Oszillationen aufzulösen.

Der Fehler des Störmer-Verlet-Verfahrens ist in Abbildung 1.3 aufgetragen. Hier wurde lediglich mit Schrittweiten gerechnet, für die das Störmer-Verlet-Verfahren stabil ist. Erst ab Schrittweiten kleiner als 10^{-4} konvergiert das Verfahren mit Ordnung zwei. Für Schrittweiten deutlich größer als 10^{-3} ist das Verfahren instabil.

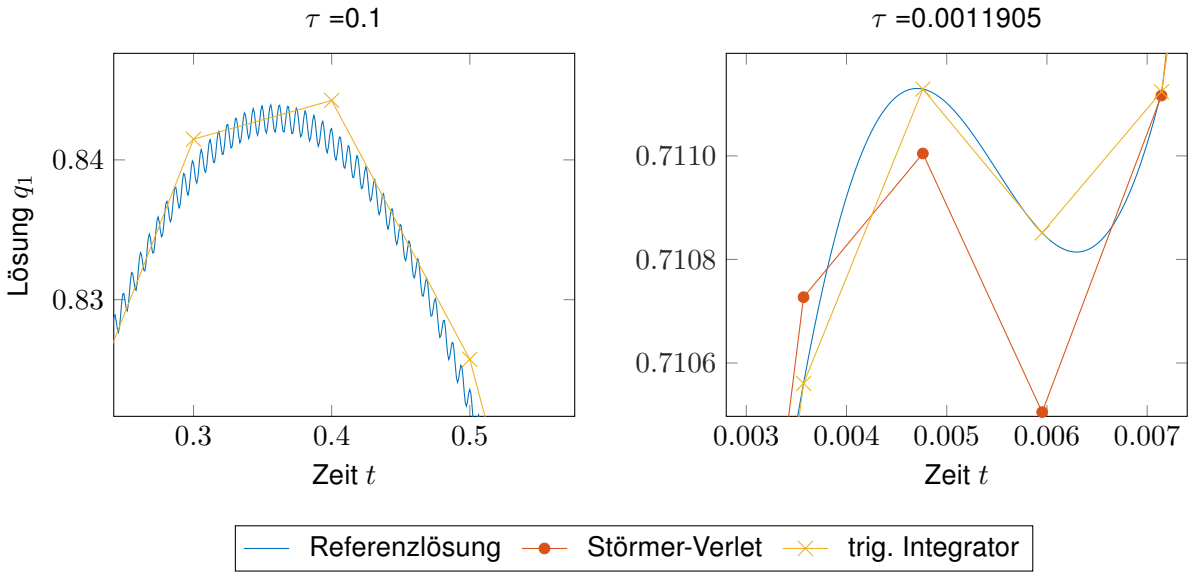

Abbildung 1.4: Referenzlösung q_1 und ihre numerischen Approximationen für verschiedene Schrittweiten τ für wenige Zeitschritte

Zusammenfassend zeigt obiges Beispiel, dass das Störmer-Verlet-Verfahren nur mit kleinen Schrittweiten sinnvoll auf hochoszillatorische Probleme angewandt werden darf. Jedoch ist die Berechnung eines Zeitschritts des Störmer-Verlet-Verfahrens sehr günstig, denn es wird nur eine Auswertung der rechten Seite der Differentialgleichung benötigt. Zudem ist es einfach zu implementieren. Daher wird dieses Verfahren in den Naturwissenschaften oft genutzt und der Nachteil der kleinen Schrittweite in Kauf genommen.

In den Abbildungen 1.2 und 1.3 sind zum Vergleich auch die numerischen Approximationen des trigonometrischen Integrators beziehungsweise ihr Fehler zur Referenzlösung aufgetragen. In beiden Abbildungen wurden die Filterfunktionen

$$\psi_1(z) = \operatorname{sinc}^2(z), \qquad \text{und} \qquad \phi(z) = \operatorname{sinc}(z)$$

verwendet. Die Filter ψ und ψ_0 wurden so gewählt, dass das entstehende Verfahren symmetrisch ist, vergleiche (1.11). In Abbildung 1.3 erkennt man zunächst, dass der trigonometrische Integrator mit größeren Schrittweiten gute Approximationen liefert. Außerdem erhält man einen Fehler der gleichen Größenordnung wie der Fehler des Störmer-Verlet-Verfahrens mit einer Schrittweite, die um einen Faktor 10^2 größer ist als beim Störmer-Verlet-Verfahren.

Weiterhin zeigt Abbildung 1.3, dass der Einsatz von Filterfunktionen sinnvoll ist. Ohne Filterfunktionen konvergiert der Integrator nicht mit Ordnung zwei. Die Fehlerkurve zeigt deutliche numerische Resonanzen, das heißt, dass der Fehler an einzelnen Stellen deutlich größer ist, als man bei Ordnung zwei und dem Fehler bei vergleichbar großen Schrittweiten erwarten würde. Somit führen diese Resonanzen zu einer Ordnungsreduktion im Bereich der große Schrittweiten. Das Verfahren konvergiert hier lediglich mit Ordnung eins oder gar Ordnung null, siehe dazu auch die Abbildungen der Fehler in Kapitel 5. Die Filterfunktionen „filtern“ Fehlerterme, die zu einer Ordnungsreduktion führen würden.

Betrachtet man Abbildung 1.2 und die Detailansicht in Abbildung 1.4, so kann man sich die Arbeitsweise des trigonometrischen Integrators anschaulich illustieren. Der trigonometrische Integrator liefert sinnvolle Approximationen für alle drei gewählten Schrittweiten. Der Vorteil des trigonometrischen Integrators ist, dass er die Oszillationen der Lösungen geeignet berücksichtigt. Das Verfahren ist sogar für $g \equiv 0$ exakt. Also verbleibt nur die Approximation der globalen Bewegung, für welche die Nichtlinearität g ausgewertet werden muss. Wie man beim Störmer-Verlet-Verfahren gut gesehen hat, ist die Punktauswertung bei hochoszillatorischen Problemen jedoch nicht sehr sinnvoll, da hier die Oszillationen einen zu großen Einfluss haben. Deshalb erhält man numerische Resonanzen, wenn der Integrator ohne Filterfunktionen verwendet wird. Die Filterfunktionen im Argument und als multiplikativer Faktor der Nichtlinearität helfen, die Punktauswertung zu verbessern. Wie man in Abbildung 1.4 in der Detailansicht sehen kann, ist der trigonometrische Integrator damit keinesfalls exakt für große Schrittweiten. Nur erlaubt der trigonometrische Integrator, die globale Bewegung trotz der großen Schrittweite nachzuvollziehen.

Der trigonometrische Integrator bietet also viele Vorteile gegenüber dem Störmer-Verlet-Verfahren bei der Integration hochoszillatorischer Probleme. Jedoch ist die Berechnung der Approximation deutlich aufwendiger als mit dem Störmer-Verlet-Verfahren. Es müssen mindestens die Matrixfunktionen $\sin(\tau\Omega)v$ und $\cos(\tau\Omega)v$, meist jedoch noch weitere Matrixfunktionen für die Filter Ψv, $\Psi_1 v$, $\Psi_0 v$ und Φv für einen Vektor $v \in \mathbb{R}^d$ ausgewertet werden. Weiterhin ist es nur im symmetrischen Fall möglich, eine Zwei-Schritt-Variante des Verfahrens herzuleiten, bei der lediglich die Positionen q berechnet werden. Im nicht-symmetrischen Fall müssen die Geschwindigkeiten q' ebenfalls approximiert werden. Es kann nur im Spezialfall entschieden werden, ob der Mehraufwand zur Berechnung des trigonometrischen Integrators geringer ist, als der Aufwand, deutlich mehr Zeitschritte mit dem Störmer-Verlet-Verfahren durchzuführen.

KAPITEL 2

SPLITTINGVERFAHREN UND TRIGONOMETRISCHE INTERGRATOREN - EIN GESCHICHTLICHER RÜCKBLICK

Das vorliegende Kapitel stellt die Entwicklung der trigonometrischen Integratoren für hochoszillatorische Probleme vor und liefert Grundlagenwissen zu Splittingverfahren für Differentialgleichungen mit genügend glatten Lösungen. Die in dieser Arbeit vorgestellte Analyse von auf trigonometrischen Integratoren basierenden Splittingverfahren beruht auf diesen beiden wichtigen Integrationsverfahren. Da hier die Konvergenzanalyse im Vordergrund steht, wird hier ebenfalls die Beweistechnik erläutert. Manche Konzepte aus den Konvergenzbeweisen für Splittingverfahren können später übernommen werden. Im Gegensatz dazu lässt sich ein deutlicher Unterschied zur bisherigen Analyse von trigonometrischen Integratoren feststellen.

2.1 Splittingverfahren

Dieser Abschnitt bietet eine Einführung zum Thema Splittingverfahren, wie sie für die vorliegende Arbeit relevant sind. Eine ausführlichere Darstellung solcher Verfahren und deren Eigenschaften findet sich beispielsweise in McLachlan und Quispel (2002), Hairer et al. (2006), Hundsdorfer und Verwer (2007), Holden, Karlsen, Lie und Risebro (2010) oder Blanes und Casas (2016).

Eine Vielzahl von Differentialgleichungen lassen sich in die folgende Form bringen

$$y' = f^{[1]}(y) + f^{[2]}(y), \qquad y(0) = y_0, \tag{2.1}$$

oder allgemeiner

$$y' = \sum_{j=1}^{N} f^{[j]}(y), \qquad y(0) = y_0, \tag{2.2}$$

wobei die Teilprobleme

$$y_j' = f^{[j]}(y_j), \qquad y_j(0) = y_{j,0}, \qquad j = 1, \dots N,$$

einfacher lösbar sind, die Lösung der Differentialgleichung (2.1) jedoch nicht bekannt ist. Bezeichnet man mit $\varphi_\tau^{[j]}(y_{j,0})$ den Fluss des Teilproblems $y_j' = f^{[j]}(y_j)$ mit Anfangswert $y_{j,0}$, so ergibt sich das

einfachste Splittingverfahren für (2.1) als Verkettung der Flüsse der Teilprobleme

$$\Phi_\tau = \varphi_\tau^{[2]} \circ \varphi_\tau^{[1]}, \qquad \Phi_\tau^* = \varphi_\tau^{[1]} \circ \varphi_\tau^{[2]}.$$

Zur Definition des Flusses einer Differentialgleichung siehe Abschnitt A.1. Dabei ist das eine Verfahren jeweils adjungiert zum anderen, zur Definition der Adjungierten siehe Definition A.1. Dieses erste Splittingverfahren wird als Lie-Trotter-Verfahren bezeichnet, siehe Trotter (1959). Die Verfahren liefern im Allgemeinen nicht die exakte Lösung. Dies zeigt sich bereits im linearen Fall, wenn $f^{[1]}(y) = Fy$ und $f^{[2]}(y) = Gy$ mit zwei Matrizen F und G gegeben sind, die nicht kommutieren, also $FG \neq GF$. Für diese ist auch

$$e^{\tau(F+G)} \neq e^{\tau G} e^{\tau F}.$$

Damit liefert das Lie-Trotter-Splittingverfahren im allgemeinen nicht der exakten Lösung. Es lässt sich jedoch mittels Taylorentwicklung für hinreichend glatte Lösungen zeigen, dass

$$e^{\tau(F+G)} y_0 = e^{\tau G} e^{\tau F} y_0 + \mathcal{O}(\tau^2),$$

und somit das Verfahren konvergent der Ordnung eins ist. Gleiches gilt auch im nichtlinearen Fall.

Neben dem Lie-Trotter-Verfahren ist das Strang-Splittingverfahren, vergleiche Strang (1968), wohl das bekannteste. Um das Strang-Splitting zu erhalten, kombiniert man das Lie-Trotter-Verfahren $\Phi_{\tau/2}$ und seine Adjungierte $\Phi_{\tau/2}^*$ mit halber Schrittweite:

$$\Phi_\tau^S = \Phi_{\tau/2} \circ \Phi_{\tau/2}^* = \varphi_{\tau/2}^{[2]} \circ \varphi_\tau^{[1]} \circ \varphi_{\tau/2}^{[2]}.$$

Das Strang-Splitting ist nach Konstruktion symmetrisch und hat als symmetrisches Verfahren Ordnung zwei bei ausreichend glatter Lösung. Zur Definition der Symmetrie eines Verfahrens siehe Definition A.1, zur Ordnung symmetrischer Verfahren siehe Hairer et al. (2006), Kapitel II, Theorem 3.2.

Solche Splittingverfahren lassen sich auch für Systeme mit mehr als zwei Teilsystemen wie (2.2) herleiten. Außerdem ist es auch möglich, ein Splittingverfahren zu konstruieren, wenn man anstelle der exakten Lösung der Teilprobleme nur eine numerische Approximation verwenden kann (siehe beispielsweise das gemittelte Impulsverfahren in (2.21)). Welche Eigenschaften wie Symmetrie oder Konvergenz sich übertragen lassen, hängt dabei stark von den gewählten Lösungsverfahren für die Teilprobleme ab.

Die Konvergenz von Splittingverfahren wurde bereits in zahlreichen Artikeln behandelt. Häufig werden dabei partielle Differentialgleichungen und ihre Semidiskretisierung in der Zeit betrachtet. Eine umfassende Analyse linearer Probleme lieferten Tobias Jahnke und Christian Lubich in Jahnke und Lubich (2000). Diese Theorie wurde bereits auf weitere, auch nichtlineare Probleme übertragen, beispielsweise in Lubich (2008), Hansen und Ostermann (2009), Thalhammer, Caliari und Neuhauser (2009), Koch und Lubich (2011), Holden, Lubich und Risebro (2013), Einkemmer und Ostermann (2014), Einkemmer und Ostermann (2015), Faou, Ostermann und Schratz (2015), Hochbruck, Jahnke und Schnaubelt (2015), Hansen und Ostermann (2016), Hansen, Ostermann und Schratz (2016), Auzinger, Kassebacher, Koch und Thalhammer (2017) oder Einkemmer und Ostermann (2018). Der Konvergenzbeweis der vorliegenden Arbeit orientiert sich an Jahnke und Lubich (2000) und Lubich (2008), deswegen werden die Artikel im Folgenden kurz vorgestellt.

2.1.1 Konvergenzanalyse für Splittingverfahren angewandt auf lineare Probleme

In Jahnke und Lubich (2000) wird das folgende Anfangswertproblem

$$u' = (A+B)u \qquad\qquad u(0) = u_0 \tag{2.3}$$

für lineare Operatoren A und B betrachtet und dessen Lösung numerisch mithilfe eines symmetrischen Strang-Splittingverfahrens

$$u_{n+1} = e^{\frac{1}{2}\tau B} e^{\tau A} e^{\frac{1}{2}\tau B} u_n$$

approximiert. Dabei ist $u_n \approx u(t_n)$. Für das Verfahren lässt sich durch Taylorentwicklung leicht Konvergenz zweiter Ordnung nachweisen, wenn die Operatoren A und B beschränkt sind. Hier wird aber speziell der Fall eines unbeschränkten Operators A betrachtet. Dies macht die Analyse für hochoszillatorische Probleme interessant. Neben dem im Folgenden vorgestellten Konvergenzresultat werden noch weitere Konvergenzsätze mit leicht veränderten Voraussetzungen bewiesen.

Die Gleichung (2.3) wird auf einem Banachraum X mit Norm $\|\cdot\|$ betrachtet. Der Operator A sei Generator einer stark-stetigen Halbgruppe e^{tA} auf X und B ein beschränkter, linearer Operator (zu stark-stetigen Halbgruppen siehe Pazy (1983)). Weiterhin sei

$$\left\|e^{tA}\right\| \le 1, \qquad\qquad \left\|e^{tB}\right\| \le 1, \qquad\qquad \left\|e^{t(A+B)}\right\| \le 1, \tag{2.4}$$

was durch geeignete Skalierung in einer geeigneten äquivalenten Norm erreicht werden kann. Außerdem sei $(-A)^\gamma$ wohldefiniert für beliebig großes positives γ, mit $\|v\| \le \|(-A)^\gamma v\|$ für alle $v \in D((-A)^\gamma)$. Es wird weiterhin $\|v\| \le \|(A+B)v\|$ für alle $v \in D(A+B)$ angenommen. Um Konvergenz zu zeigen, müssen die zwei folgenden Bedingungen mit $[A,B] = AB - BA$ und nichtnegativen α, β erfüllt sein:

$$\|[A,B]\,v\| \le c_1 \|(-A)^\alpha v\| \qquad \text{für alle } v \in D((-A)^\alpha), \tag{2.5}$$

$$\|[A,[A,B]]\,v\| \le c_2 \left\|(-A)^\beta v\right\| \qquad \text{für alle } v \in D((-A)^\beta). \tag{2.6}$$

Unter den gegebenen Voraussetzungen wird zunächst in Theorem 2.1 in Jahnke und Lubich (2000) gezeigt, dass für $\beta \ge 1 \ge \alpha$ der lokalen Fehler beschränkt ist durch

$$\left\|e^{\frac{1}{2}\tau B} e^{\tau A} e^{\frac{1}{2}\tau B} v - e^{\tau(A+B)} v\right\| \le C\tau^3 \left\|(-A)^\beta v\right\|, \qquad \text{für alle } v \in D((-A)^\beta), \tag{2.7}$$

wobei die Konstante C nur von c_1, c_2 und $\|B\|$ abhängt. Hat man dies gezeigt, so lässt sich mithilfe des Fächers der Lady Windermere der globale Fehler direkt beschränken durch

$$\|u_n - u(t_n)\| \le \tau^2 C t_{\mathsf{end}} \max_{0\le s\le t_{\mathsf{end}}} \left\|(-A)^\beta u(s)\right\|,$$

mit $0 \le t_n = n\tau \le t_{\mathsf{end}}$. Eine Fehlerrekursion kann dabei direkt aus der Teleskopsumme

$$u_n - u(n\tau) = \mathbb{S}^n u_0 - \mathbb{T}^n u_0 = \sum_{j=0}^{n-1} \mathbb{S}^{n-j-1}(\mathbb{S}-\mathbb{T})\mathbb{T}^j u_0$$

abgeleitet werden. Somit folgt die Abschätzung des globalen Fehlers direkt aus der Abschätzung des lokalen Fehlers. Mit etwas mehr Aufwand kann gezeigt werden, dass der globale Fehler nur von

der Norm des Anfangswerts $\|u_0\|$ anstelle von $\left\|(-A)^\beta u(s)\right\|$ abhängt, siehe Theorem 2.2 in Jahnke und Lubich (2000).

Aufwendiger als der globale Fehler, ist die Abschätzung des lokalen Fehlers (2.7). Er ist jedoch sehr anschaulich. Zunächst stellt man die exakte Lösung mithilfe der Variation-der-Konstanten-Formel

$$e^{\tau(A+B)}v = e^{\tau A}v + \int_0^\tau e^{sA}Be^{(\tau-s)(A+B)}v\,ds,$$

dar. Setzt man diese Formel erneut unter dem Integral ein, so erhält man

$$e^{\tau(A+B)}v = e^{\tau A}v + \int_0^\tau e^{sA}Be^{(\tau-s)A}v\,ds + R_1 v \tag{2.8}$$

mit einem Restterm R_1

$$R_1 = \int_0^\tau e^{sA}B\int_0^{\tau-s} e^{\sigma A}Be^{(\tau-s-\sigma)(A+B)}\,d\sigma\,ds.$$

Auch das numerische Verfahren kann mithilfe der Taylorentwicklung für $e^{\frac{1}{2}\tau B}$ genauer analysiert werden:

$$e^{\frac{1}{2}\tau B}e^{\tau A}e^{\frac{1}{2}\tau B}v = e^{\tau A}v + \frac{1}{2}(Be^{\tau A} + e^{\tau A}B)v + R_2 v, \tag{2.9}$$

mit Restterm R_2

$$R_2 = \frac{1}{8}\tau^2(B^2e^{\tau A} + 2Be^{\tau A}B + e^{\tau A}B^2) + \tilde{R}_2 = \frac{1}{8}\tau^2\left[B,\left[B,e^{\tau A}\right]\right] + \tilde{R}_2,$$

mit beschränktem Restterm $\left\|\tilde{R}_2\right\| \leq C\tau^3\|B\|^3$. Subtrahiert man die beiden Darstellungen (2.8) von (2.9) voneinander, so hebt sich der erste Term auf und es ergibt sich der Fehler

$$e^{\frac{1}{2}\tau B}e^{\tau A}e^{\frac{1}{2}\tau B}v - e^{\tau(A+B)}v = d + r$$

mit

$$d = \frac{1}{2}\left[B,e^{\tau A}\right]v - \int_0^\tau e^{sA}Be^{(\tau-s)A}v\,ds$$

und $r = R_2v - R_1v$. Diese lassen sich mittels Quadraturfehlern geeigneter Funktionen und der Kommutatorschranke aus (2.5) und (2.6) abschätzen, siehe Beweis von Theorem 2.1 in Jahnke und Lubich (2000).

Diese Analyse hat wesentlich die Analyse des linearen Problems in Buchholz et al. (2018) beeinflusst, siehe auch Abschnitt 2.2.7. Dort wird auch beschrieben, weshalb man die Analyse aus Jahnke und Lubich (2000) nicht direkt auf das hochoszillatorische Problem anwenden kann. Der Beweis wird anschließend modifiziert, um im hochoszillatorischen Fall ebenfalls Konvergenz zweiter Ordnung zu zeigen. Die hier bereits auftauchenden Kommutatoren helfen jedoch dabei, die Struktur des lokalen Fehlers auch im hochoszillatorischen Fall zu verstehen.

2.1.2 Konvergenzanalyse für Splittingverfahren angewandt auf nichtlineare Probleme am Beispiel der Schrödingergleichung

Die folgende Konvergenzanalyse des Strang-Splittingverfahrens stammt aus Lubich (2008). Das Verfahren wird auf die nichtlineare Schrödingergleichung

$$i\frac{\partial \psi}{\partial t} = -\Delta\psi + V\psi, \qquad x \in \mathbb{R}^3,\ t \geq 0, \tag{2.10}$$

mit kubischer Nichtlinearität

$$V = V(\psi) = \pm\,|\psi|^2$$

und asymptotischen Randbedingungen

$$\lim_{|x|\to\infty} \psi(x,t) = 0 \qquad\qquad \lim_{|x|\to\infty} V(x) = 0,$$

angewandt. Außerdem sei ein Anfangswert $\psi(x,0) = \psi_0(x)$ für $x \in \mathbb{R}^3$ gegeben.

Die Schrödingergleichung mit kubischer Nichtlinearität hat eine Vielzahl an Anwendungen in der Physik und ist deshalb Untersuchungsobjekt in vielen Bereichen der Mathematik. In Lubich (2008) wird die Konvergenz des Strang-Splittingverfahrens zur Semidiskretisierung in der Zeit untersucht. Es ergibt sich das numerische Verfahren

$$\psi^-_{n+1/2} = e^{\frac{i}{2}\tau\Delta}\psi_n\,, \tag{2.11}$$

$$\psi^+_{n+1/2} = e^{-i\tau V(\psi^-_{n+1/2})}\psi^-_{n+1/2}\,, \tag{2.12}$$

$$\psi_{n+1} = e^{\frac{i}{2}\tau\Delta}\psi^+_{n+1/2}\,. \tag{2.13}$$

Dabei bezeichnet $e^{\frac{i}{2}\tau\Delta}$ den Lösungsoperator der freien Schrödingergleichung, welcher mithilfe der schnellen Fouriertransformation (FFT - „Fast Fourier Transformation“, siehe beispielsweise Brigham (1982)) approximiert werden kann. Der Operator $e^{-i\tau V(\psi^-_{n+1/2})}$ verhält sich wie ein Multiplikationsoperator, denn es lässt sich durch Differenzieren leicht zeigen, dass $\left|\psi^+_{n+1/2}\right| = \left|\psi^-_{n+1/2}\right|$.

Das hier gezeigte Konvergenzresultat benötigt starke Regularität der Lösung. Im Folgenden wird mit $L_2 = L_2(\mathbb{R}^3)$ der Raum der Lebesgue-integrierbaren Funktionen und mit $H^k = H^k(\mathbb{R}^3)$ der Sobolevraum mit Funktionen in L_2 bezeichnet, deren erste k schwachen Ableitungen ebenfalls in L_2 sind, siehe beispielsweise Adams und Fournier (2003). Es wird angenommen, dass die Lösung $\psi(t)$ der Gleichung (2.10) im Sobolevraum H^4 für $0 \leq t \leq t_{\text{end}}$ liegt. Zusätzlich seien mit

$$m_k = \max_{0\leq t\leq t_{\text{end}}} \|\psi(t)\|_{H^k}\,, \qquad\qquad k \geq 4,$$

Schranken an die Normen der Lösung gegeben. Man erhält damit für das Strang-Splitting aus (2.11)

$$\begin{aligned}\|\psi_n - \psi(t_n)\|_{H^1} &\leq C(m_3, t_{\text{end}})\tau,\\ \|\psi_n - \psi(t_n)\|_{L_2} &\leq C(m_4, t_{\text{end}})\tau^2\end{aligned}$$

für $t_n = n\tau \leq t_{\text{end}}$. Das Verfahren konvergiert somit mit Ordnung eins in der H^1-Norm und man kann Konvergenz zweiter Ordnung in L_2 nachweisen. Dies ist das erste Konvergenzresultat zweiter Ordnung, welches keine Lipschitzbedingung an die Nichtlinearität benötigt.

Der Beweis ist in mehrere Teile unterteilt. Zuerst wird Stabilität in der L_2-Norm beziehungsweise der H^1-Norm des numerischen Verfahrens über einen Zeitschritt gezeigt, wobei Anfangswerte in H^1 vorausgesetzt werden. Damit lassen sich Fehlerschranken für den lokalen Fehler herleiten, genauer

$$\|\psi_1 - \psi(\tau)\|_{H^1} \leq C_3\tau^2,$$
$$\|\psi_1 - \psi(\tau)\|_{L_2} \leq C_4\tau^3,$$

wobei die Konstanten C_3 und C_4 nur von $\|\psi_0\|_{H^3}$ beziehungsweise $\|\psi_0\|_{H^4}$ abhängen. Interessant ist, dass die Fehlerabschätzung ähnlich zu der des lokalen Fehlers in Jahnke und Lubich (2000) ist, vergleiche Abschnitt 2.1.1. Die Lösung wird wieder mit der Variation-der-Konstanten-Formel dargestellt, jedoch diesmal mithilfe der nichtlinearen Variante unter Verwendung der Lie-Ableitung (vergleiche Lubich (2008), Abschnitt 4.3). Für die numerische Lösung wird mithilfe der Taylorentwicklung eine Darstellung hergeleitet. Die Fehlerabschätzung ergibt sich aus einer Abschätzung der entstehenden Quadraturfehler. Hierbei treten Lie-Kommutatoren auf, die beschränkt werden müssen. Dies ist dank der starken Regularitätsvoraussetzungen möglich. Daraus ergeben sich die Abschätzungen für den lokalen Fehler. Schlussendlich zeigt man, dass die numerische Lösung in H^2 bleibt, wenn der Anfangswert bereits in H^2 liegt und alle weiteren Iterierten in H^1 gleichmäßig beschränkt sind. Damit erhält man das Konvergenzresultat mithilfe des Lady Windermeres Fächer, vergleiche Abschnitt A.3.

Die Konvergenzanalyse der vorliegenden Arbeit verwendet ebenfalls Lie-Kommutatoren, um den Fehler zu analysieren. Der lokale Fehler kann auch, wie in Lubich (2008), mithilfe der Lie-Ableitung geschrieben werden. Andere Konzepte konnten leider nicht übernommen werden, da hier die starke Regularität der Lösung eine Rolle spielt, welche man für hochoszillatorische Probleme jedoch nicht voraussetzen kann.

2.2 Entwicklung der trigonometrischen Integratoren

Die Klasse der trigonometrischen Integratoren wurde bereits für hochoszillatorische Probleme der Form (1.1) in Abschnitt 1.1.2 definiert. Das vorliegende Kapitel erläutert die geschichtliche Entwicklung der Integrationsverfahren und erklärt, wie man Konvergenz dieser Verfahren nachweisen kann. Einen umfassenden Überblick findet man auch in Hairer et al. (2006), Kapitel XIII.

2.2.1 Gautschi-Verfahren

Walter Gautschi stellte bereits 1961 fest, dass klassische Verfahren häufig nicht geeignet sind, um das oszillatorische Verhalten der Lösung spezieller gewöhnlicher Differentialgleichungen korrekt wiederzugeben. Daher schlug er in Gautschi (1961) eine Reihe von Mehrschrittverfahren vor, die sich speziell für Probleme mit periodischen Lösungen eignen.

Zunächst führt er den Begriff der trigonometrischen Ordnung ein: Ein lineares Funktional L in $C^s([a,b])$ sei von algebraischer Ordnung p genau dann, wenn

$$Lt^r = 0, \qquad \text{für } r = 0, 1, \ldots, p, \tag{2.14}$$

ist. Es ist von trigonometrischer Ordnung p relativ zur Periode T genau dann, wenn

$$L\cos\left(r\frac{2\pi}{T}t\right) = L\sin\left(r\frac{2\pi}{T}t\right) = 0 \qquad \text{für } r = 0, 1, \ldots, p. \tag{2.15}$$

Als Funktional L wird später der Differenzenoperator des Verfahrens eingesetzt. Hierbei wird deutlich, dass die Periode T einen wesentlichen Einfluss auf die trigonometrische Ordnung im Sinne von Gautschi hat. Denn hier werden Kosinus- und Sinusschwingungen exakt integriert, deren Periode $\frac{r}{T}$ für $r = 1, \ldots, p$ ist (Zum Einfluss der Periode T, die oft im Vorfeld nicht bekannt ist, siehe Abschnitt 6 aus Gautschi (1961)).

Gautschi konstruierte nun numerische Verfahren, die eine möglichst hohe trigonometrische Ordnung haben. Zuerst untersuchte er dafür Systeme erster Ordnung. In Abschnitt 5 erweiterte er seine Theorie auf Differentialgleichung zweiter Ordnung der Form

$$q'' = f(t, q), \qquad q(t_0) = q_0, \quad q'(t_0) = q'_0, \tag{2.16}$$

wie sie für die vorliegende Arbeit relevant sind. Zur numerischen Approximation verwendet Gautschi lineare Mehrschrittverfahren, welche Approximationen $q_m \approx q(t_0 + m\tau)$ an die gesuchte Lösung der Gleichung (2.16) liefern. Die Approximationen q_m genügen dabei der Gleichung

$$q_{n+1} + \alpha_1 q_n + \cdots + \alpha_k q_{n+1-k} = \tau^2(\beta_0 q''_{n+1} + \beta_1 q''_n + \cdots + \beta_k q''_{n+1-k}) \tag{2.17}$$

für $n = k-1, k, \ldots$ mit

$$q''_m = f(t_0 + m\tau, q_m).$$

Man nennt das Verfahren k-stufig, falls α_k und β_k nicht gleichzeitig verschwinden. Bei einem k-stufigen Verfahren startet man mit k Startwerten q_{n+1-k} bis q_n um q_{n+1} zu berechnen. Für $\beta_0 = 0$ handelt es sich um ein explizites, für $\beta_0 \neq 0$ um ein implizites Verfahren. Man verwendet nun für L den linearen Differenzenoperator

$$L(q, t, \tau) = \sum_{i=0}^{k} \Big(\alpha_i q\big(t + (n+1-i)\tau\big) - \tau\beta_i q''\big(t + (n+1-i)\tau\big)\Big) \qquad (\alpha_0 = 1).$$

Die wesentliche Arbeit in Gautschi (1961) besteht darin, Bedingungen an die Koeffizienten α_k, die Schrittweite τ und die Periode T zu formulieren, sodass Koeffizienten β_k existieren, für die das Mehrschrittverfahren (2.17) konvergent von trigonometrischer Ordnung p relativ zur Periode T ist (vergleiche Theorem 2 und 3 aus Gautschi (1961)).

Nachdem er die Existenz nachgewiesen hat, werden in Abschnitt 5 explizit Koeffizienten β_k ausgerechnet und damit Mehrschrittverfahren konstruiert, die konvergent von trigonometrischer Ordnung p sind. Als einfachestes Verfahren stellt Gautschi ein explizites, zwei-stufiges Mehrschrittverfahren vor, welches heute als Gautschi-Verfahren bekannt ist. Das Verfahren hat zunächst die folgende Form

$$q_{n+1} + \alpha_1 q_n + \alpha_2 q_{n-1} = \tau^2 \beta_1 q_n''.$$

Das zugehörige Funktional lautet

$$Lq = q(t_0 + (n+1)\tau) + \alpha_1 q(t_0 + n\tau) + \alpha_2 q(t_0 + (n-1)\tau) - h^2 \beta_1 q''(t_0 + n\tau).$$

Es sollen nun die Koeffizienten α_1, α_2 und β_1 so gewählt werden, dass das Verfahren trigonometrische Ordnung eins relativ zur Periode T hat, das heißt

$$\begin{aligned} L1 &= 0, \\ L\cos\left(\frac{2\pi}{T}t\right) &= 0, \\ L\sin\left(\frac{2\pi}{T}t\right) &= 0. \end{aligned}$$

Es ergibt sich

$$\alpha_1 = -2, \qquad \alpha_2 = 1, \qquad \beta_1 = \operatorname{sinc}^2\left(\frac{2\pi}{T}\right).$$

Wendet man dies nun auf die in der vorliegenden Arbeit betrachtete Differentialgleichung (1.1) in einer Dimension ($d = 1$) an, also

$$q'' = -\omega^2 q + g(q)$$

an, so erhält man mit $T = \frac{2\pi}{\omega}$

$$\begin{aligned} & q_{n+1} - 2q_n + q_{n-1} = \tau^2 \operatorname{sinc}^2\left(\frac{\omega\tau}{2}\right)(-\omega^2 q_n + g(q_n)) \\ \Leftrightarrow \qquad & q_{n+1} - \left(2 - \sin^2\left(\frac{\omega\tau}{2}\right)\right) q_n + q_{n-1} = \tau^2 \operatorname{sinc}^2\left(\frac{\omega\tau}{2}\right) g(q_n) \\ \Leftrightarrow \qquad & q_{n+1} - \cos(\tau\omega) q_n + q_{n-1} = \tau^2 \operatorname{sinc}^2\left(\frac{\omega\tau}{2}\right) g(q_n) \end{aligned} \tag{2.18}$$

Spricht man in der Literatur vom *Gautschi-Verfahren*, bezeichnet man damit die Approximationen $q_n \approx q(t_0 + n\tau)$ aus (2.18). Die Wahl der Periode kann man dadurch motivieren, dass man die Gleichung als Störung des harmonischen Oszillators

$$q'' = -\omega^2 q$$

betrachtet, dessen Lösung die Periode $T = \frac{2\pi}{\omega}$ hat (zur Auswirkung der Approximation der Periode T auf den Begriff der trigonometrischen Ordnung, siehe Gautschi (1961), Abschnitt 6).

In Gautschi (1961) findet sich jedoch keine Fehleranalyse, die einen Zusammenhang der trigonometrischen Ordnung zur Konvergenz der Methode nachweist. Eine Fehleranalyse für die heute als Gautschi-Verfahren bekannte Methode findet sich in Hochbruck und Lubich (1999), vergleiche auch Abschnitt 2.2.4. Aus der Konstruktion ist jedoch klar, dass das Verfahren exakt ist, wenn die Lösung ein trigonometrisches Polynom vom Höchstgrad eins ist.

2.2.2 Deuflhard-Verfahren

Peter Deuflhard entdeckte 1979 in Deuflhard (1979) im Wesentlichen das Gautschi-Verfahren erneut, aber aus einer anderen Perspektive. In seinem Artikel untersuchte er den Zusammenhang zwischen der Differentialgleichung und der zugehörigen Differenzengleichung. Dazu multiplizierte er die rechte Seite der Differentialgleichung mit einer kleinen Störung ϵ. Dann konstruierte er Verfahren, deren Differenzengleichung keine parasitären Lösungen zulassen, das heißt, dass für $\epsilon = 0$ jede Lösung der Differenzengleichung auch Lösung der Differentialgleichung ist.

Interessant ist hier die Idee, die er in Lemma 2.2 aus dem genannten Artikel vorstellt. Hier werden speziell Mehrschrittverfahren betrachtet. Zunächst schreibt er die exakte Lösung als Linearkombination der Fundamentallösungen der homogenen Gleichung und einem Faltungsintegral mit der rechten Seite. Für dieses Faltungsintegral nutzt er schließlich die summierte Trapezregel, um ein Mehrschrittverfahren zu konstruieren. Zusätzlich führt er eine Fehleranalyse des vorgestellten Verfahrens durch (vergleiche Theorem 2.4 in Deuflhard (1979)), jedoch nur unter hohen Regularitätsvoraussetzungen an die rechte Seite der Differentialgleichung, weshalb diese Analyse für hochoszillatorische Probleme nicht verwendet werden kann.

Im dritten Kapitel erweitert er seine Theorie auch auf Differentialgleichungen zweiter Ordnung. Verwendet man speziell die in der vorliegenden Arbeit betrachtete Differentialgleichung (1.1) für Dimension $d = 1$ mit , so kann man die exakte Lösung zunächst durch die Variation-der-Konstanten--Formel darstellen:

$$\begin{bmatrix} q(t) \\ q'(t) \end{bmatrix} = \begin{bmatrix} \cos(t\Omega) & \Omega^{-1}\sin(t\Omega) \\ -\Omega\sin(t\Omega) & \cos(t\Omega) \end{bmatrix} \begin{bmatrix} q_0 \\ q_0' \end{bmatrix} + \int_0^t \begin{bmatrix} \Omega^{-1}\sin((t-s)\Omega) \\ \cos((t-s)\Omega) \end{bmatrix} g(q(s))\, ds.$$

Approximiert man das Faltungsintegral nun mittels der (Standard-) Trapezregel, so erhält man

$$\begin{bmatrix} q_{n+1} \\ q_{n+1}' \end{bmatrix} = \begin{bmatrix} \cos(\tau\Omega) & \Omega^{-1}\sin(\tau\Omega) \\ -\Omega\sin(\tau\Omega) & \cos(\tau\Omega) \end{bmatrix} \begin{bmatrix} q_n \\ q_n' \end{bmatrix} + \frac{\tau}{2} \begin{bmatrix} \tau\,\mathrm{sinc}(\tau\Omega) g_n \\ g_{n+1} + \cos(\tau\Omega) g_n \end{bmatrix}.$$

Dieses symmetrische Verfahren kennt man heute unter dem Namen *Verfahren nach Deuflhard*. Betrachtet man lediglich die Positionen q_n, so ist das Deuflhard-Verfahren äquivalent zum Gautschi-Verfahren (siehe (2.18)). Die Einschritt-Formulierung bietet zusätzlich Approximationen q_n' an die Ableitung. Ein Fehleranalyse führte Deuflhard jedoch nur für geeignet glatte rechte Seiten der Differentialgleichung durch. Das Verfahren ist nach Konstruktion für konstante Funktionen g exakt.

2.2.3 Gemittelte Impulsmethode

Die Arbeit García-Archilla et al. (1999) von Bosco García-Archilla, Jesús María Sanz-Serna und Robert Skeel stellt einen Meilenstein in der Analyse von hochoszillatorischen Problemen und geeigneten Verfahren für Probleme solcher Art dar. Die Autoren versuchten ein Verfahren zu konstruieren, das die Stabilitätsprobleme herkömmlicher Verfahren angewandt auf dynamische Systeme mit verschiedenen Zeitskalen vermeidet. Hierfür kommt einerseits der Splittingansatz (Aufteilen in schnelle

und langsame Kräfte) zum Einsatz und es werden zum ersten Mal Filterfunktionen verwendet, um die Verstärkung kritischer Fehlerterme zu vermeiden. Das konstruierte Verfahren wird *gemittelte Impulsmethode* genannt. Der Artikel enthält zudem eine Stabilitäts- sowie Fehleranalyse.

Für die vorliegende Arbeit sollen speziell Verfahren für Gleichung (1.1) betrachtet werden. Zur numerischen Approximation wird zunächst die Impulsmethode (siehe auch Verlet-I und r-RESPA aus Grubmüller et al. (1991) und Tuckerman et al. (1992)) betrachtet. Die Lösung des reduzierten Problems $\tilde{q}'' = -\Omega^2 \tilde{q}$ ist gegeben als:

$$\begin{bmatrix} \tilde{q}(t) \\ \tilde{q}'(t) \end{bmatrix} = R(t\Omega) \begin{bmatrix} \tilde{q}_0 \\ \tilde{q}'_0 \end{bmatrix} \qquad \text{mit} \quad R(t\Omega) = \begin{bmatrix} \cos(t\Omega) & \Omega^{-1}\sin(t\Omega) \\ -\Omega\sin(t\Omega) & \cos(t\Omega) \end{bmatrix}. \tag{2.19}$$

Die Impulsmethode berechnet Approximationen $q_n \approx q(n\tau)$ und $p_n \approx p(n\tau) = q'(n\tau)$ durch

$$\begin{aligned} \textbf{Kick} \quad & p_n^+ = p_n + \tfrac{\tau}{2} g(q_n) \\ \textbf{Oszillator} \quad & \begin{bmatrix} q_{n+1} \\ p_{n+1}^- \end{bmatrix} = R(\tau\Omega) \begin{bmatrix} q_n \\ p_n^+ \end{bmatrix} \\ \textbf{Kick} \quad & p_{n+1} = p_{n+1}^- + \tfrac{\tau}{2} g(q_{n+1}). \end{aligned}$$

Die Impulsmethode ist eine Erweiterung des Störmer-Verlet-Verfahrens und wurde für die Molekülдynamik entwickelt, um die Effizienz zu steigern. Eine Konvergenzanalyse wurde jedoch weder in Grubmüller et al. (1991) noch in Tuckerman et al. (1992) durchgeführt. In Kapitel 3 aus García-Archilla et al. (1999) führen die Autoren zuerst eine lineare Stabilitätsanalyse durch, welche die Instabilitäten der Impulsmethode zeigt.

Die Impulsmethode soll im Folgenden modifiziert werden, um so ein stabiles Verfahren zu erhalten, das große Zeitschrittweiten zulässt. Dafür schlagen die Autoren vor, den Einfluss der Auswertung der Nichtlinearität $g(q_n)$ durch eine Mittelwertbildung zu ersetzen, denn die Punktauswertung wird zu stark von den lokalen Oszillationen beeinflusst. Dafür wird

$$g(q_n) \quad \rightarrow \quad \Phi g(\Phi q_n)$$

ersetzt. Die Matrix Φ wird als Filter bezeichnet und ist durch

$$\Phi = \frac{1}{\tau} \int_{-\infty}^{\infty} \phi\left(\frac{t}{\tau}\right) \cos(t\Omega)\, dt,$$

gegeben, wobei für ϕ zunächst lediglich

$$\int_{-\infty}^{\infty} \phi(s)\, ds = 1 \tag{2.20}$$

gefordert wird. Man erhält damit das gemitteltte Impulsverfahren, hier in Einschritt-Formulierung

$$\begin{bmatrix} q_n \\ p_n \end{bmatrix} = R \begin{bmatrix} q_{n-1} \\ p_{n-1} \end{bmatrix} + \frac{\tau}{2} R \begin{bmatrix} 0 \\ \Phi g(\Phi q_{n-1}) \end{bmatrix} + \frac{\tau}{2} \begin{bmatrix} 0 \\ \Phi g(\Phi q_n) \end{bmatrix}, \tag{2.21}$$

mit $r = R(\tau\Omega)$ aus (2.19). Dieses Verfahren liegt in der Klasse der trigonometrischen Integratoren, vergleiche (1.10a) mit spezieller Wahl der Filterfunktionen $\psi(\tau\Omega) = \operatorname{sinc}(\tau\Omega)\Phi$, $\psi_0(\tau\Omega) = \cos(\tau\Omega)\Phi$ und $\psi_1(\tau\Omega) = \Phi$. Aus spezieller Wahl von ϕ ergeben sich die folgenden, in García-Archilla et al.

(1999) diskutierten Verfahren

klassisches Impulsverfahren	$\Phi = I,$
ShortAverage	$\Phi = \operatorname{sinc}\left(\frac{\tau\Omega}{2}\right),$
LongAverage	$\Phi = \operatorname{sinc}\left(\frac{\tau\Omega}{2}\right)\cos\left(\frac{\tau\Omega}{2}\right),$
LinearAverage	$\Phi = \operatorname{sinc}^2\left(\frac{\tau\Omega}{2}\right).$

Neben einer umfangreichen linearen Stabilitätsanalyse bietet García-Archilla et al. (1999) ein Konvergenzresultat für das gemittelte Impulsverfahren zur numerischen Approximation der Lösung der Differentialgleichung (1.1). Dafür sei zunächst die Annahme 1.2 erfüllt. Zusätzlich muss $g'(q)$ ebenfalls global Lipschitz-stetig in q mit Konstante $\widetilde{C_{g,2}}$ sein. Die Bedingungen an den Filter Φ werden in Form von Bedingungen an die Funktion ϕ gegeben. Neben der Bedingung (2.20) sei $\phi(s) = 0$ für $|s| > 1$ und symmetrisch, also $\phi(s) = \phi(-s)$. Weiterhin sei ϕ beschränkt. Unter diesen Voraussetzungen ergibt sich für den Fehler zwischen den Approximationen (q_n, p_n) des gemittelten Impulsverfahrens (2.21) und der exakten Lösung $(q(t), p(t)) = (q(t), q'(t))$ der Differentialgleichung (1.1)

$$\begin{aligned} \|q_n - q(n\tau)\| &\leq C_0\tau^2, & 0 < n\tau \leq t_{\mathsf{end}},\\ \|p_n - p(n\tau)\| &\leq C_1\tau, & 0 < n\tau \leq t_{\mathsf{end}}. \end{aligned}$$

Die Fehlerkonstanten C_0, C_1 hängen dabei nur von der Endzeit t_{end}, den Schranken an die Funktion g und ihre Ableitungen L_g, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,2}}$ und der Finiten-Energie K ab. Diese Fehleranalyse lässt also hochoszillatorische Probleme zu, da weder die Norm von Ω noch höhere Regularitätsanforderungen als die Finite-Energie-Bedingung an die Lösung in die Fehlerkonstante eingehen.

Der Konvergenzbeweis beruht im Wesentlichen darauf, eine Fehlerrekursion herzuleiten, indem man die exakte Lösung mittels der Variation-der-Konstanten-Formel darstellt und die Differenz zur numerischen Lösung betrachtet. Mit $\epsilon_n = q_n - q(n\tau)$, $\delta_n = p_n - p(n\tau)$ und $\Lambda_n = g(\Phi q_n) - g(\Phi q(n\tau))$ ergibt sich

$$\begin{bmatrix} \epsilon_n \\ \delta_n \end{bmatrix} = R \begin{bmatrix} \epsilon_{n-1} \\ \delta_{n-1} \end{bmatrix} + \frac{\tau}{2} R \begin{bmatrix} 0 \\ \Phi\Lambda_{n-1} \end{bmatrix} + \frac{\tau}{2} \begin{bmatrix} 0 \\ \Phi\Lambda_n \end{bmatrix} + \sigma_n$$

mit dem Quadraturfehler

$$\sigma_n = \frac{\tau}{2} R \begin{bmatrix} 0 \\ g(\Phi q\big((n-1)\tau\big)) \end{bmatrix} + \frac{\tau}{2} \begin{bmatrix} 0 \\ \Phi g(\Phi q(n\tau)) \end{bmatrix} - \int_{(n-1)\tau}^{n\tau} \begin{bmatrix} \Omega^{-1}\sin((n\tau - s)\Omega) \\ \cos((n\tau - s)\Omega) \end{bmatrix} g(q(s))\, ds.$$

Durch sukzessives Einsetzen erhält man letztendlich eine Fehlerdarstellung, welche man mit einer speziellen Variante des Lemmas von Gronwall (siehe Lemma 2 in García-Archilla et al. (1999)) abschätzen kann. Die Schwierigkeit besteht im Nachweis des Lemmas von Gronwall und in der Abschätzung des Quadraturfehlers σ_n. Um Letzteren zu beschränken, werden die Anteile in den Geschwindigkeiten p und in den Positionen q separat betrachtet. Weiterhin wird der Fehler in mehrere Teile aufgeteilt und die Voraussetzungen sowie zahlreiche trigonometrische Identitäten verwendet, um eine geeignete Abschätzung zu erzielen.

B. García-Archilla, J. M. Sanz-Serna und R. Skeel liefern damit ein erstes Verfahren, für das sie unter gewissen Voraussetzungen Konvergenz zweiter Ordnung in den Positionen für hochoszillatorische Probleme nachweisen können. Sie konnten die Stabilitätsprobleme der Impulsmethode überwinden. Zusätzlich führten sie die Filterfunktionen ein, die sich für hochoszillatorische Probleme etabliert haben. Auch die Finite-Energie-Bedingung taucht hier als sinnvolle Regularitätsvoraussetzung an die Lösung auf.

Erweiterung der gemittelten Impulsmethode

Jesús María Sanz-Serna schlug in Sanz-Serna (2008) eine Erweiterung der gemittelten Impulsmethode zu sogenannten (ϕ, ψ)-Verfahren vor. Gegenüber der gemittelten Impulsmethode aus García-Archilla et al. (1999) werden hier zwei Filterfunktionen benötigt: Der Filter ψ glättet das Verfahren, der Filter ϕ bewirkt eine Mittelwertbildung in der Auswertung der langsamen Kräfte. Er liefert weiterhin eine Fehleranalyse für Probleme mit linearen schnellen Kräften.

Hier wird nun die Anwendung der Theorie auf Differentialgleichungen der Form (1.1) betrachtet. Es gelten die Annahmen 1.2. Das Verfahren wird nun ähnlich zu García-Archilla et al. (1999) mithilfe von

$$\bar{g}_n = \Psi g(\Phi q_n)$$

definiert:

$$\begin{aligned} \textbf{Kick} \quad & p_n^+ = p_n + \tfrac{\tau}{2}\bar{g}_n \\ \textbf{Oszillator} \quad & \begin{bmatrix} q_{n+1} \\ p_{n+1}^- \end{bmatrix} = R(\tau\Omega) \begin{bmatrix} q_n \\ p_n^+ \end{bmatrix} \\ \textbf{Kick} \quad & p_{n+1} = p_{n+1}^- + \tfrac{\tau}{2}\bar{g}_{n+1}. \end{aligned}$$

Dabei sind die Filter $\Phi = \hat{\phi}(\tau\Omega)$ und $\Psi = \hat{\psi}(\tau\Omega)$ gegeben durch

$$\hat{\phi}(\omega) = \int_{-\infty}^{\infty} \cos(\omega s)\phi(s)\, ds, \qquad \hat{\psi}(\omega) = \int_{-\infty}^{\infty} \cos(\omega s)\psi(s)\, ds,$$

für integrierbare Funktionen ϕ und ψ.

Eine Fehleranalyse wird nur für den Fall linearer, schneller Kräfte gegeben. Es sei die Annahme 1.2 erfüllt. Die Funktionen ψ und ϕ seien beschränkt, integrierbar, reellwertig und haben kompakten Träger. Weiterhin seien die Funktionen gerade, $\chi(-t) \equiv \chi(t)$, und

$$\int_{-\infty}^{\infty} \chi(s)\, ds = 1, \tag{2.22}$$

für $\chi = \phi,\ \psi$. Außerdem sei eine der folgenden äquivalenten Bedingungen erfüllen:

$$\sum_{j=-\infty}^{\infty} \psi(t-j) \equiv 1 \qquad \Leftrightarrow \qquad \hat{\psi}(2\pi n) = 0, \qquad n = \pm 1, \pm 2, \ldots, \tag{2.23}$$

wobei $\hat{\psi}$ die Fourierkoeffizienten der Funktion

$$\mathbf{\Psi}(t) = \sum_{j=-\infty}^{\infty} \psi(t-j)$$

sind. Unter diesen Annahmen wird Konvergenz der Ordnung eins für p_n und q_n gezeigt. Dies geschieht im Wesentlichen mit Techniken aus García-Archilla et al. (1999). Hier wird auch begründet, weshalb die obigen äquivalenten Bedingungen (2.23) an ψ nicht nur hinreichend, sondern sogar notwendig sind. Darunter findet sich auch eine Erklärung, wie man Ordnung zwei in den Positionen q_n erreichen kann, wenn man einen Filter $\phi \neq \delta$ wählt, wobei mit δ die Standard-Diracfunktion gemeint ist.

Diese Erweiterung der gemittelten Impulsmethode stellt also eine Fehleranalyse vor, die sich von der in Grimm und Hochbruck (2006) oder der in Hairer et al. (2006) darin unterscheidet, dass die Filterfunktionen als Gewichte eingesetzt werden und daraus Bedingungen an die Filter abgeleitet werden, vergleiche Abschnitt 2.2.5. Gegenüber García-Archilla et al. (1999) werden hier zwei verschiedene Filterfunktionen eingeführt.

2.2.4 Fehleranalyse des Gautschi-Verfahrens

Wie in Abschnitt 2.2.1 bereits erwähnt, wurde in Gautschi (1961) keine Fehleranalyse für das Gautschi-Verfahren angegeben. Diese lieferten Marlis Hochbruck und Christian Lubich in Hochbruck und Lubich (1999). In dem Artikel wird die Differentialgleichung zweiter Ordnung (1.1) betrachtet. Die Autoren geben zunächst eine andere Herleitung für das Gautschi-Verfahren an, als Gautschi selbst. Sie orientiert sich an der Herleitung aus Deuflhard (1979). Die Variation-der-Konstanten-Formel liefert für die exakte Lösung die Darstellung

$$q(t+\tau) = \cos(\tau\Omega)q(t) + \Omega^{-1}\sin(\tau\Omega)q'(t) + \int_0^\tau \Omega^{-1}\sin((\tau-s)\Omega)g(q(t+s))\,ds. \tag{2.24}$$

Ist die Inhomogenität $g(q) \equiv g$ konstant, so erhält man

$$q(t+\tau) - 2q(t) + q(t-\tau) = \tau^2\psi((\tau\Omega)^2)(-\Omega^2 q(t) + g) \tag{2.25}$$

mit

$$\psi(\xi^2) = \operatorname{sinc}^2\left(\frac{\xi}{2}\right).$$

Daraus lässt sich das Gautschi-Verfahren für allgemeine Funktionen g herleiten, indem man g während eines Zeitschritts durch eine geeignete Approximation g_n ersetzt. Man erhält das Verfahren

$$q_{n+1} - 2q_n + q_{n-1} = \tau^2\Psi((\tau\Omega)^2)(-\Omega^2 q_n + g_n), \tag{2.26}$$

wobei $q_n \approx q(t_n)$ für $t_n = n\tau$. Wählt man

- $g_n = g(q_n)$, so ist (2.26) ein Gautschi-Verfahren, vgl Gautschi (1961).
- $g_n = g(\phi((\tau\Omega)^2)q_n)$, so erhält man ein Verfahren nach Art von Gautschi.

Letzteres ist die favorisierte Wahl. Hier werden wie bei García-Archilla et al. (1999) Filterfunktionen verwendet, um die Konvergenz des Verfahrens zu gewährleisten (siehe Abschnitt 2.2.3). Als Startvektor für das Zweischrittverfahren 2.26 wird

$$q_1 = \cos(\tau\Omega)q_0 + \Omega^{-1}\sin(\tau\Omega)q_0' + \frac{1}{2}\tau^2\psi((\tau\Omega)^2)g_0 \tag{2.27}$$

gewählt. Das Verfahren lässt sich so auch in ein Einschrittverfahren überführen

$$v_{n+\frac{1}{2}} = v_n + \frac{1}{2}\tau\psi((\tau\Omega)^2)(-\Omega^2 q_n + g_n), \tag{2.28a}$$

$$q_{n+1} = q_n + \tau v_{n+\frac{1}{2}}, \tag{2.28b}$$

$$v_{n+1} = v_{n+\frac{1}{2}} + \frac{1}{2}\tau\psi((\tau\Omega)^2)(-\Omega^2 q_{n+1} + g_{n+1}), \tag{2.28c}$$

mit $v_n = (q_{n+1} - q_{n-1})/(2\tau)$, was der Diskretisierung einer gemittelten Geschwindigkeit

$$\bar{v}(t) = \frac{1}{2\tau}\int_{-\tau}^{\tau} q'(t+\tau)\,d\tau,$$

entspricht. Die Formulierungen (2.26) und (2.28) sind äquivalent, wenn man

$$v_0 = \sigma((\tau\Omega)^2)y_0'$$

wählt. Ist man an Approximationen der Geschwindigkeiten q' interessiert, so lassen sich diese nachträglich aus

$$q_{n+1}' = q_n' + 2\tau\sigma((\tau\Omega)^2)(-\Omega^2 q_n + g_n)$$

berechnen. Man erkennt hier, dass das Verfahren nach Art von Gautschi mehr einer gefilterten Variante des Strömer-Verlet-Verfahrens entspricht anstelle der gemittelten Impulsmethode (vergleiche (2.28) mit (1.7) und (2.21)). Während bei der gemittelten Impulsmethode weiterhin nur ein Filter auftritt, der sich ausschließlich auf die Inhomogenität auswirkt, treten hier zwei Filterfunktionen auf, wobei der Filter Ψ auch auf die Matrix Ω^2 und nicht nur auf die Inhomogenität g wirkt.

Kern des Artikels ist die Fehleranalyse des Verfahrens. Dazu werden einige Voraussetzungen benötigt: Zunächst sei die Annahme 1.2 erfüllt. Die Filterfunktion $\phi(x)$ und ihre ersten zwei Ableitungen sollen für $x > 0$ beschränkt sein. Weiterhin soll

$$\phi(0) = 1, \qquad \phi(k^2\pi^2) = 0, \quad k = 1, 2, 3, \ldots \tag{2.29}$$

und

$$|\phi(x)| \leq 1, \qquad x \geq 0 \tag{2.30}$$

gelten. Unter diesen Voraussetzungen folgt für den globalen Fehler

$$\|q_n - q(t_n)\| \leq \tau^2 C\,\ell(n,d), \qquad 0 < t_n = n\tau \leq t_{\text{end}},$$

wobei die Fehlerkonstante C nur von K, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$ und t_{end} abhängt, jedoch nicht von τ oder gar $\|\Omega\|$. Für ℓ gilt $\ell(n,d) \leq \log(n+1)\log(d+1)$ und auch $\ell(n,d) \leq \sqrt{d}$, wobei d die Dimension der Matrix Ω repräsentiert. Die Autoren vermuteten bereits, dass dieser Anteil vermieden werden kann. Dies wurde später in Grimm (2005a) gezeigt. Für die Geschwindigkeiten q' lässt sich zeigen, dass der Fehler $\|q_n' - q'(t_n)\|$ sich proportional zu τ mit vergleichbarer Fehlerkonstante wie bei den Positionen q beschränken lässt.

Der Beweis ist aufwendig und daher in mehrere Teile aufgeschlüsselt worden. Zentral ist dabei die Herleitung einer Fehlerrekursion in Lemma 2 des Artikels. Dafür ersetzt man zunächst die Approximationen q_n in der Verfahrensvorschrift (2.26) durch $q(t_n)$:

$$q(t_{n+1}) - 2q(t_n) + q(t_{n-1}) = \tau^2\Psi((\tau\Omega)^2)(-\Omega^2 y(t_n) + g(\phi((\tau\Omega)^2)q(t_n))) + d_n. \tag{2.31}$$

Der hierbei entstehende Fehler wird d_n genannt und in Lemma 1 analysiert. Die Darstellung beruht im Wesentlichen auf der Variation-der-Konstanten-Formel. Zur Abschätzung werden die Eigenschaften der Funktion g verwendet, die helfen zu zeigen, dass d_n in Terme der Größenordnung $\mathcal{O}(\tau^3)$ und $\mathcal{O}(\tau^4)$ aufgeteilt werden kann. Um nun die Fehlerrekursion herzuleiten, bildet man die Differenz zwischen der Verfahrensvorschrift (2.26) und (2.31). Die Rekursion aufzulösen bedarf einigen Aufwand, da die n-te Potenz der entstehenden Rekursionsmatrix in der Einschritt-Formulierung berechnet werden muss. Man erhält für den Fehler $e_n = q_n - q(t_n)$

$$e_{n+1} = -W_{n-1}e_0 + W_n e_1 + \sum_{j=1}^{n} W_{n-j}(\tau^2 G_j e_j - d_j)$$

mit $W_n = (\sin((n+1)\tau\Omega))(\sin(\tau\Omega))^{-1}$ und Matrizen G_j, die durch $\|G_j\| \leq \widetilde{C_{g,1}}$ beschränkt sind. Nun soll das diskrete Lemma von Gronwall B.3 verwendet werden, um e_n geeignet abzuschätzen (vergleiche Beweis von Theorem 1 aus dem Artikel). Dafür muss man für die Summe die Schranke

$$\left\| \sum_{j=1}^{n} W_{n-j} d_j \right\| \leq C\tau^2$$

zeigen, welches den aufwendigsten Teil des Beweises darstellt. Der Grund dafür ist, dass eine direkte Abschätzung der Anteile von Größenordnung $\mathcal{O}(\tau^3)$ in d_n nicht möglich ist. Nach Einsetzen der Darstellung von d_n wird die Summe in drei Teile aufgeteilt. Die dabei entstandenen Terme werden einzeln analysiert. Hierfür wird zunächst eine Orthogonaltransformation verwendet, um die Matrix Ω zu diagonalisieren. Nach mehreren Umformungen und einer partiellen Summation lässt sich der Fehler als komponentenweises Produkt zweier Matrizen darstellen, welches später (siehe Lemma 5 des Artikels) geeignet beschränkt werden kann. Dabei entsteht der Fehlerterm $\ell(n, d)$. Diese Abschätzungen sind extrem aufwendig und erfordern eine Vielzahl an trigonometrischen Identitäten und erneut partielle Summation.

2.2.5 Fehleranalyse der trigonometrische Integratoren

Nachdem García-Archilla et al. (1999) und Hochbruck und Lubich (1999) bereits erste Fehleranalysen für trigonometrische Verfahren angewandt auf hochoszillatorische Probleme lieferten und Filterfunktionen zur Vermeidung von kritischen Fehlertermen etablierten, fehlte es an einer Analyse, die jegliche bisher bekannte Verfahren einschließt und allgemeine Bedingungen an die Filterfunktionen aufstellt, die Konvergenz zweiter Ordnung garantieren. Weiterhin wurde in García-Archilla et al. (1999) nur eine Filterfunktion verwendet, während Hochbruck und Lubich (1999) zwei Filterfunktionen einführten, wobei hier die Filterfunktion ψ im Vorfeld festgelegt worden ist.

Volker Grimm und Marlis Hochbruck entwickelten in Grimm und Hochbruck (2006) eine Analyse für trigonometrische Integratoren mit vier verschiedenen Filterfunktionen, an die Bedingungen gestellt werden, um Konvergenz zweiter Ordnung zu garantieren. Alle bisher vorgestellten Verfahren sind in dieser Analyse enthalten. Die Untersuchungen des Lang-Zeitverhaltens aus Hairer und Lubich (2000) motivierte diese Analyse ebenfalls, da sowohl die gemittelte Impulsmethode als auch das modifizierte Gautschi-Verfahren kein optimales Langzeit-Verhalten zeigten. Eine Fehleranalyse für trigonometrische Integratoren mit Filterfunktionen angewandt auf Gleichungen der Form (1.1) mit einer allgemeinen, symmetrischen, positiv definiten Matrix Ω, die gutes Langzeit-Verhalten zeigten, fehlte bislang.

In Grimm und Hochbruck (2006) wird die die Konvergenz der Klasse von trigonometrische Integratoren (1.10a) angewandt auf Differentialgleichungen zweiter Ordnung der Form (1.1) analysiert, welche in der vorliegenden Arbeit bereits in Abschnitt 1.1.2 vorgestellt wurden. Wählt man die Filter entsprechend, lassen sich alle Verfahren darstellen, die bereits erwähnt worden sind. Da es sich dabei um symmetrische Verfahren handelt, genügt es, die Filterfunktionen ϕ und ψ festzulegen. Die Filter ψ_1 und ψ_0 ergeben sich dann aus (1.11).

$$\begin{array}{ll} \text{Gautschi (1961) :} & \psi(\xi) = \operatorname{sinc}^2(\frac{1}{2}\xi), \quad \phi(\xi) = 1, \\ \text{Deuflhard (1979) :} & \psi(\xi) = \operatorname{sinc}(\xi) \quad \phi(\xi) = 1, \\ \text{García-Archilla et al. (1999) :} & \psi(\xi) = \operatorname{sinc}(\xi)\phi(\xi), \quad \phi(\xi) = \operatorname{sinc}(\xi), \\ \text{Hochbruck und Lubich (1999) :} & \psi(\xi) = \operatorname{sinc}^2(\frac{1}{2}\xi), \quad \phi(\xi) = \operatorname{sinc}(\xi)\left(1 + \frac{1}{3}\sin^2\left(\frac{1}{2}\xi\right)\right), \\ \text{Hairer und Lubich (2000) :} & \psi(\xi) = \operatorname{sinc}^2(\xi), \quad \phi(\xi) = 1. \\ \text{Grimm und Hochbruck (2006) :} & \psi(\xi) = \operatorname{sinc}^3(\xi), \quad \phi(\xi) = \operatorname{sinc}(\xi), \end{array} \tag{2.32}$$

Letzteres Verfahren wird hier zusätzlich vorgeschlagen, denn diese neue Methode erfüllt alle nachfolgend aufgeführten Bedingungen an die Filterfunktionen, um ein konvergentes Verfahren von der Ordnung zwei zu erhalten. Weiterhin liefert sie ein gutes Langzeit-Verhalten (siehe Hairer und Lubich (2000)).

Um nun Konvergenz zweiter Ordnung nachzuweisen, müssen die Filterfunktionen die folgenden Bedingungen erfüllen:

Annahme 2.1. *Die geraden und analytischen Filterfunktionen ϕ, ψ, ψ_0 und ψ_1 erfüllen die folgenden Bedingungen:*

$$\chi(0) = 1 \qquad \textit{für } \chi = \phi, \psi, \psi_0, \psi_1, \tag{2.33}$$

$$|\chi(\xi)| \leq m_1 \qquad \textit{für } \chi = \phi, \psi, \psi_0, \psi_1, \tag{2.34}$$

$$\left| \frac{1}{\sin\left(\frac{\xi}{2}\right)} \left(\operatorname{sinc}^2\left(\frac{\xi}{2}\right) - \psi(\xi) \right) \right| \leq m_3, \tag{2.35}$$

$$\left| \frac{1}{\xi \sin\left(\frac{\xi}{2}\right)} \left(\operatorname{sinc}(\xi) - \chi(\xi) \right) \right| \leq m_4 \qquad \textit{für } \chi = \phi, \psi_0, \psi_1, \tag{2.36}$$

für alle $\xi \geq 0$.

Aus den Bedingungen an die Filterfunktionen folgt unmittelbar

$$\left| \frac{\phi(\xi) - 1}{\xi} \right| \leq m_2 \qquad \text{für alle } \xi \geq 0.$$

Konvergenz erster Ordnung in den Geschwindigkeiten erhält man unter den folgenden zusätzlichen Annahmen an die Filterfunktionen:

Annahme 2.2. *Die geraden und analytischen Filterfunktionen ψ, ψ_0 und ψ_1 erfüllen die folgenden*

Bedingungen:

$$|\xi\psi(\xi)| \leq m_5, \tag{2.37}$$

$$\left| \frac{\xi}{\sin\left(\frac{\xi}{2}\right)} \left(\operatorname{sinc}^2\left(\frac{\xi}{2}\right) - \psi(\xi) \right) \right| \leq m_6, \tag{2.38}$$

$$\left| \frac{1}{\sin\left(\frac{\xi}{2}\right)} \left(\operatorname{sinc}(\xi) - \chi(\xi) \right) \right| \leq m_7 \qquad \text{für } \chi = \psi_0, \psi_1, \tag{2.39}$$

für alle $\xi \geq 0$.

Das Konvergenzresultat lautet wie folgt: Es seien die Annahmen 1.2 und 2.1 erfüllt, dann ist

$$\|q(t_n) - q_n\| \leq C\tau^2, \qquad t_0 \leq t_n = t_0 + n\tau \leq t_{\text{end}}.$$

Die Konstante C hängt dabei nur von t_{end}, K, m_i, $i = 1, \ldots 4$, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$ und $\widetilde{C_{g,2}}$ ab. Ist zusätzlich auch die Annahme 2.2 erfüllt, so folgt

$$\left\|q'(t_n) - q'_n\right\| \leq C\tau, \qquad t_0 \leq t_n = t_0 + n\tau \leq t_{\text{end}},$$

wobei die Konstante C nur von t_{end}, K, m_i, $i = 1, \ldots 7$, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$ und $\widetilde{C_{g,2}}$ abhängt.

Der Beweis ist aufwendig und wird daher in mehrere Teile zerlegt. Zentraler Schritt ist die Herleitung einer Fehlerrekursion. Dazu ersetzt man q_n durch $q(t_n)$ in der Verfahrensvorschrift (1.10a)

$$\begin{bmatrix} q(t_{n+1}) \\ q'(t_{n+1}) \end{bmatrix} = R(\tau\Omega) \begin{bmatrix} q(t_n) \\ q'(t_n) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \tau^2\Psi g(\Phi q(t_n)) \\ \tau(\Psi_0 g(\Phi q(t_n)) + \Psi_1 g(\Phi q(t_{n+1})) \end{bmatrix} + \begin{bmatrix} d_n \\ d'_n \end{bmatrix},$$

mit R aus (2.19). Der dabei auftretende Fehler wird d_n genannt. Subtrahiert man nun die Verfahrensvorschrift (1.10a) und löst die Rekursion durch Einsetzen der entstandenen Formel in sich selbst, so erhält man für $e_n = q(t_n) - q_n$ und $e'_n = q'(t_n) - q'_n$

$$\begin{bmatrix} e_{n+1} \\ e'_{n+1} \end{bmatrix} = R^{n+1} \begin{bmatrix} e_0 \\ e'_0 \end{bmatrix} + \frac{1}{2} \sum_{j=0}^{n} R^{n-j} \begin{bmatrix} \tau^2\Psi G_j e_j \\ \tau(\Psi_0 G_j e_j + \Psi_1 G_{j+1} e_{j+1} \end{bmatrix} + \begin{bmatrix} D_n \\ D'_n \end{bmatrix}, \tag{2.40}$$

mit

$$G_n = \int_0^1 g_y(\Phi(q_n + se_n))\, ds \cdot \Phi,$$

und

$$\begin{bmatrix} D_n \\ D'_n \end{bmatrix} = \sum_{j=0}^{n} R^{n-j} \begin{bmatrix} d_j \\ d'_j \end{bmatrix}.$$

Der Fehler lässt sich nun mithilfe des diskreten Lemmas von Gronwall B.3 abschätzen, wenn man

$$\|D_n\| \leq C\tau^2, \qquad \left\|D'_n\right\| \leq C\tau,$$

zeigen kann. Dies stellt den aufwendigste Teil des Beweises dar. Man verwendet hierfür zuerst die Variation-der-Konstanten-Formel, um die Fehler d_n und d'_n darzustellen und weiterhin Taylorentwicklung der Nichtlinearität g, um den Fehler in Terme der Größenordnung $\mathcal{O}(\tau^2)$ und $\mathcal{O}(\tau^3)$ aufzuteilen

(Lemma 1 und Lemma 2 des Artikels). Während man die Anteile, die sich proportional zu τ^3 verhalten, direkt abschätzen kann, muss man eine sorgfältigere Abschätzung bei der Summation der verbliebenden Terme vornehmen. Dabei wird der Fehler D_n nicht nur in seine Komponenten in d_j und d'_j aufgeteilt, sondern d'_j in verschiedene Fehlerterme zerlegt. Man braucht mehrere partielle Summationen, um die Beschränktheit der Summe D_n proportional zu τ^2 zu garantieren, die jeweils eine Bedingung an die Filterfunktion liefern (siehe Lemma 3 und Lemma 4 des Artikels). Gleiches geschieht für D'_n (vergleiche Lemma 5 und Lemma 6).

Die Filterbedingungen ergeben sich im Wesentlichen aus der Beweistechnik. Sie sind hinreichend für Konvergenz zweiter Ordnung. Ob jede von ihnen auch notwendig ist, ist nicht geklärt.

2.2.6 Polynomielle Nichtlinearitäten

Alle bisherigen Beweise nahmen an, dass die Nichtlinearität zumindest eine Lipschitzbedingung erfüllt. Ludwig Gauckler betrachtet in Gauckler (2015) speziell die semilineare Wellengleichung mit polynomieller Nichtlinearität

$$w_{tt} = w_{xx} + w^p \qquad \text{für } w = w(x,t), \tag{2.41}$$

mit $p \geq 2$ und 2π-periodischen Randbedingungen in einer Raumdimension ($\mathbb{T} = \mathbb{R}\backslash(2\pi\mathbb{Z})$). Für die Anfangswerte soll

$$w(\cdot, t_0) \in H^{s+1} \qquad \text{und} \qquad w_t(\cdot, t_0) \in H^s \qquad \text{für } s \geq 0$$

gelten, wobei H^s den Sobolevraum $H^s(\mathbb{T})$ darstellt. Dabei soll s klein sein. Speziell $s = 0$ liefert Lösungen, die eine Finite-Energie-Bedingung erfüllen.

Bevor man den trigonometrischen Integrator anwenden kann, wird die partielle Differentialgleichung (2.41) im Raum diskretisiert, indem man die Lösung durch das trigonometrische Polynom

$$w_{\mathcal{K}}(x,t) = \sum_{j \in \mathcal{K}} q_j(t) e^{ijx} \qquad \text{mit } \mathcal{K} = \{-K, \dots, K-1\} \tag{2.42}$$

approximiert. Dabei sind $q_j(t)$ die Fourierkoeffizienten der Lösung w. Setzt man diese Approximation in die (2.41) ein, und wertet sie jeweils in den Kollokationspunkten $x_k = \pi k/K$ mit $k \in \mathcal{K}$ aus, so erhält man eine Differentialgleichungin Form von (1.1) mit dem Vektor $q(t) = (q_j(t))_{j \in \mathcal{K}}$ ($q(t) \in \mathbb{C}^{\mathcal{K}}$). Dabei ist Ω eine nichtnegative Diagonalmatrix

$$\Omega = \operatorname{diag}(\omega_j)_{j \in \mathcal{K}} \qquad \text{mit } \omega_j = |j|\,, \tag{2.43}$$

und die Nichtlinearität g ist durch die diskrete Faltung $\star$

$$g(y) = \underbrace{q \star \cdots \star q}_{p \text{ mal}} \qquad \text{mit} \qquad (u \star v)_j = \sum_{k+l \equiv j \bmod 2K} u_k v_l, \quad j \in \mathcal{K} \tag{2.44}$$

definiert. Für die in Gauckler (2015) präsentierte Analyse ist die spezielle Form der Nichtlinearität g und der Matrix Ω entscheidend, siehe dazu auch Abschnitt 4 aus Gauckler (2015). Die Anfangswerte sind durch die Anfangswerte $w(\cdot, t_0)$ beziehungsweise $w_t(\cdot, t_0)$ festgelegt durch

$$q_j(t_0) = \sum_{k \in \mathbb{Z}, k \equiv j \bmod 2K} w_k(t_0) \qquad \text{und} \qquad \dot{q}(t_0) = \sum_{k \in \mathbb{Z}, k \equiv j \bmod 2K} \dot{w}_k(t_0), \qquad \text{für } j \in \mathcal{K},$$

wobei $w_k(t)$ beziehungsweise $\dot{w}_k$ die Fourierkoeffizienten von $w(\cdot, t)$ und $w_t(\cdot, t_0)$ sind. Diese Definition ist genau dann wohldefiniert, wenn die Reihe der Fourierkoeffizienten konvergiert und entspricht dann einer trigonometrischen Interpolation der Anfangswerte $w(\cdot, t_0)$ beziehungsweise $w_t(\cdot, t_0)$ in den Kollokationspunkten. Wenn die Anfangswerte bereits durch ihre Fourierkoeffizienten definiert worden sind, so ist die Wahl

$$q_j(t_0) = w_j(t_0), \qquad \text{und} \qquad \dot{q}_j(t_0) = \dot{w}_j(t_0), \qquad \text{für } j \in \mathcal{K}$$

für die Berechnung vorteilhaft.

Als numerischer Integrator wird die Klasse der trigonometrischen Integratoren (vergleiche (1.10a)) vorgeschlagen. Um Konvergenz zu zeigen, werden folgende Annahmen benötigt:

Annahme 2.3. *Für ein gegebenes* $-1 \leq \beta \leq 1$ *existiere eine Konstante* c*, sodass für alle* $\xi = h\omega_j$ *mit* $j \in \mathcal{K}$ *und* $\omega_j \neq 0$*,*

$$\begin{aligned} |\phi(\xi)| &\leq c, \\ |\psi(\xi)| &\leq c\xi^{\beta}, && \text{falls } -1 \leq \beta \leq 0, \\ |1 - \psi(\xi)| &\leq c\xi^{\beta} && \text{falls } 0 < \beta \leq 1, \\ |1 - \chi(\xi)| &\leq c\xi^{1+\beta} && \text{für } \chi = \phi, \psi_0, \psi_1, \end{aligned}$$

gilt.

Diese Annahmen werden von vielen bekannten trigonometrischen Integratoren erfüllt, beispielsweise dem Deuflhard-Verfahren (vergleiche Abschnitt 2.2.2), der gemittelten Impulsmethode aus García-Archilla et al. (1999) (vergleiche Abschnitt 2.2.3) oder dem Verfahren von Grimm und Hochbruck (2006) (vergleiche Abschnitt 2.2.5).

Der Fehler wird in der Norm

$$\|q\|_s = \left(\sum_{j \in \mathcal{K}} \langle j \rangle^{2s} |q_j|^2 \right)^{\frac{1}{2}} \qquad \text{mit } \langle j \rangle = \max(1, |j|),$$

für $q \in \mathbb{C}^{\mathcal{K}}$ gemessen. Diese Norm ist äquivalent zu der Sobolev H^s-Norm der trigonometrischen Polynome $\sum_{j \in \mathcal{K}} q_j e^{ijx}$. Für $q = q^n$ und $x = x_j$ ist dies eine Approximation der Lösung $w(x_j, t_n)$ der nichtlinearen Wellengleichung.

Man erhält das folgende Konvergenzresultat: Sei $c \geq 1$ und $s \geq 0$. Die exakte Lösung $(q(t), \dot{q}(t))$ der räumlichen Diskretisierung (1.1) mit der Matrix Ω aus (2.43) und der Nichtlinearität g aus (2.44) der nichtlinearen Wellengleichung erfülle

$$\|q(t)\|_{s+1} + \|\dot{q}(t)\|_s \leq K_s \qquad \text{für } 0 \leq t - t_0 \leq t_{\text{end}}. \tag{2.46}$$

Dann existiert eine Schrittweitenschranke τ_0, sodass für alle Schrittweiten $\tau < \tau_0$ die folgenden Fehlerschranken für die numerische Lösung $(q^n, \dot{q}^n)$, berechnet mit Hilfe des trigonometrischen Integrators, gelten:

Wenn Annahme 2.3 mit einer Konstante c für $\beta = 0$ und $\beta = \alpha$ für $-1 \leq \alpha \leq 1$ erfüllt ist, dann gilt

$$\|q(t_n) - q^n\|_{s+1-\alpha} + \|\dot{q}(t_n) - \dot{q}^n\|_{s-\alpha} \leq C\tau^{1+\alpha} \qquad \text{für } 0 \leq t_n - 0 = n\tau \leq t_{\text{end}}.$$

Die Konstanten C und τ_0 hängen dabei einzig von K_s und s aus (2.46), dem Exponenten p der Nichtlinearität, der Endzeit t_{end} und der Filterkonstante c ab.

Für $s = 0$ entspricht die Annahme (2.46) im Wesentlichen einer Finite-Energie-Bedingung an die Lösung der nichtlinearen Wellengleichung und seiner räumlichen Diskretisierung. In diesem Fall erhalten wir Fehlerschranken zweiter Ordnung an q in L^2 ($\alpha = 1$) und Fehlerschranken erster Ordnung für $\dot{q}$ in L^2 ($\alpha = 0$).

Der Beweis ist aufwendig, jedoch von anderer Art als die Konvergenzbeweise, die in den vorherigen Abschnitten beschrieben worden sind. Zuerst muss die Norm von $g(q)$ und die zeitliche Ableitung dieser Funktion in Abhängigkeit von $\|q\|_s$ beschränkt werden. Hier spielt die Semidiskretisierung und damit verbunden die Darstellung der Sobolevnorm eine entscheidende Rolle. Der Beweis wird danach in zwei Teile unterteilt: Zuerst werden Fehlerschranken von niedriger Ordnung in Sobolevräumen höherer Ordnung gezeigt. Im zweiten Teil zeigt man Fehlerschranken höherer Ordnung in Sobolevräumen niedrigerer Ordnung.

Dafür kann die klassiche Beweistechnik des Fächers der Lady Windermere verwendet werden, bei welcher der lokale Fehler und die Fehlerfortpflanzung kontrolliert werden müssen, siehe auch Abschnitt A.3. Dies gelingt, da man zunächst nur Fehlerschranken niedriger Ordnung nachweisen möchte. Der lokale Fehler beispielsweise muss durch $C\tau^{2+\alpha}$ für $-1 \leq \alpha \leq 0$ beschränkt werden, also sind für $\alpha = 0$ Fehlerterme der Größenordnung $\mathcal{O}(\tau^2)$ unkritisch. Entscheidend ist, dass im Konvergenzbeweis der Fall $\alpha = 0$ zuerst behandelt wird. Hier kann mittels Induktion das Konvergenzresultat bewiesen werden. Man erhält daraus aber die Stabilität der numerischen Lösung, also

$$\|q_n\|_{s+1}^2 + \|\dot{q}_n\|_s^2 \leq 2K_s, \qquad \text{für } 0 \leq t_n = t_0 + n\tau \leq t_{\text{end}}.$$

Die Regularität der Lösung ist entscheidend, um Lady Windermeres Fächer für $-1 \leq \alpha < 0$ und $0 < \alpha \leq 1$ anwenden zu können. Um Fehlerschranken höherer Ordnung für den lokalen Fehler und die Fehlerfortpflanzung in Sobolevräumen niedrigerer Ordnung zu beweisen, geht wesentlich die spezielle Wahl der Semidiskretisierung, der diskreten Sobolevnorm und die Bedingungen an die Filterfunktionen ein.

An der Beweistechnik überrascht zunächst, dass die klassische Methode des Fächers der Lady Windermere kombiniert mit der Stabilität der numerischen Lösung ausreichen, um Konvergenz nachzuweisen. Betrachtet man den Beweis genauer, erkennt man, dass die Struktur der Nichtlinearität g, die gewählte Semidiskretisierung und die diskrete Sobolevnorm eine entscheidende Rolle spielen. Möchte man also hochoszillatorische Probleme betrachten, die aus einer Semidiskretisierung entstammen, scheint es vorteilhaft, die konkrete Semidiskretisierung und die diskreten Normen in die Analyse einzubeziehen.

2.2.7 Fehleranalyse von auf trigonometrischen Integratoren basierenden Splittingverfahren für lineare hochoszillatorische Probleme

Die Idee für die vorliegende Arbeit entstand bei einem Workshop 2015, an dem Ludwig Gauckler, Tobias Jahnke, Marlis Hochbruck und ich beteiligt waren. Die Ergebnisse wurden in Buchholz et al. (2018) veröffentlicht. Ein Auszug daraus ist auch in Buchholz, Gauckler, Grimm, Hochbruck und Jahnke (2016) zu finden. Die Arbeit stellt eine direkte Vorarbeit zur vorliegenden Arbeit dar.

Es war bereits lange bekannt, dass sich die Klasse der trigonometrischen Integratoren als Splittingverfahren auffassen lässt. Eine Fehleranalyse, die auf Splittingverfahren beruht, ist aber meines Wissens nach bis zur Veröffentlichung von Buchholz et al. (2018) nicht durchgeführt worden. Mit einer solchen Analyse erhofft man sich einen tieferen Einblick in die Struktur der hochoszillatorischen Probleme und ihrer numerischen Integration zu erhalten. Wie in Grimm und Hochbruck (2006), Hairer et al. (2006) und Gauckler (2015) gesehen, entstehen die Bedingungen an die Filterfunktionen im Wesentlichen aus der Fehleranalyse, vergleiche Abschnitt 2.2.5 und 2.2.6. Eine neue Konvergenzanalyse liefert also eventuell neue Bedingungen an die Filterfunktionen und erweitert so die Klasse der Verfahren, die bewiesenermaßen mit Ordnung zwei konvergieren.

In Buchholz et al. (2018) wird die gewöhnliche, lineare Differentialgleichung (1.1) mit $g(q) = Gq$ mit einer Matrix G betrachtet, wobei angenommen wird, dass $\|\Omega\| \gg 1$ ist. An das System werden Voraussetzungen gestellt, vergleiche Assumption 2.1 und 2.2 aus Buchholz et al. (2018). Diese lauten wie folgt:

Annahme 2.4. *Es gelten die Bedingungen a) und b) aus Annahme 1.1. Da g linear ist, reduziert sich die Bedingungen c) und d) auf die Forderung, dass die Norm der Matrix G unabhängig von Ω beschränkt ist.*

Zusätzlich wird gefordert, dass die Inverse der Matrix Ω existiert und

$$\|\Omega^{-1}\| \leq C_{inv} \tag{2.47}$$

gilt.

Zur Bedingung (2.47) siehe zusätzlich Bemerkung 3.2 im folgenden Kapitel. Die Gleichung (1.1) kann in ein System erster Ordnung geschrieben werden. Mithilfe einer Transformation mit Ω^{-1} ergeben sich die neuen Variablen

$$u = \begin{bmatrix} q \\ \Omega^{-1}q' \end{bmatrix}, \qquad u_0 = \begin{bmatrix} q_0 \\ \Omega^{-1}q_0' \end{bmatrix}.$$

Diese erfüllen die Differentialgleichung

$$u' = Au + Bu, \qquad u(0) = u_0, \tag{2.48}$$

mit Matrizen

$$A = \begin{bmatrix} 0 & \Omega \\ -\Omega & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ \Omega^{-1}G & 0 \end{bmatrix}.$$

Die Finite-Energie-Bedingung (1.2) für q, q' ist dann äquivalent zu

$$\|Au(t)\| \leq K, \qquad 0 \leq t \leq t_{\text{end}}. \tag{2.49}$$

Für ein System der Form (2.48) bietet sich ein Splittingverfahren zur numerischen Integration an. Verwendet man das Strang-Splitting, so wird schnell klar, dass der entstandene Integrator dem symmetrischen trigonometrischen Integrator (1.10a) entspricht, jedoch mit den Filtern $\phi(s) \equiv \psi_1(s) \equiv 1$. Mit diesen Filterfunktionen ist der Integrator allerdings nicht konvergent von Ordnung zwei, vergleiche Abschnitt 1.2 und die Fehleranalyse in Abschnitt 2.2.5. Daher müssen nun zuerst Filter eingebracht werden, um ein konvergentes Verfahren zu erhalten. Es seien also ϕ und ψ_S reellwertige Filterfunktionen, die Annahme 4.1 erfüllen. Es wird hier der Index S bei der Filterfunktion Ψ_S verwendet, um Verwechslungen mit der Filterfunktion ψ der trigonometrischen Integrators (1.10a) zu vermeiden. Wie man später sehen wird, entspricht der Filter ψ_S dem Filter ψ_1 des trigonometrischen Integrators. Man definiert nun

$$\widetilde{B} = \begin{bmatrix} 0 & 0 \\ \Omega^{-1}\widetilde{G} & 0 \end{bmatrix}, \tag{2.50}$$

wobei die Matrizen $\widetilde{G}$, Φ und Ψ_S durch

$$\widetilde{G} = \Psi_S G \Phi, \qquad \Phi = \phi(\tau\Omega), \qquad \Psi_S = \psi_S(\tau\Omega), \tag{2.51}$$

definiert sind. Damit lässt sich die modifizierte Gleichung

$$\widetilde{u}' = A\widetilde{u} + \widetilde{B}\widetilde{u}, \qquad \widetilde{u}(0) = \widetilde{u}_0 = u_0, \tag{2.52}$$

aufstellen. Wendet man ein Strang-Splittingverfahren zur Approximation der Lösung dieser Gleichung an, erhält man

$$u_{n+1} = S_{\mathsf{lin}} u_n, \qquad S_{\mathsf{lin}} = e^{\frac{\tau}{2}\widetilde{B}} e^{\tau A} e^{\frac{\tau}{2}\widetilde{B}}. \tag{2.53}$$

Dabei ist $v(t) = e^{tA} v_0$ mit

$$e^{tA} = \begin{bmatrix} \cos(t\Omega) & \sin(t\Omega) \\ -\sin(t\Omega) & \cos(t\Omega) \end{bmatrix} \tag{2.54}$$

die Lösung des Teilproblems $v'(t) = Av(t)$ mit Anfangswert v_0. Da $\widetilde{B}$ nilpotent ist, folgt für die Lösung $w(t) = e^{t\widetilde{B}} w_0$ des zweiten Teilproblems $w'(t) = Bw(t)$ mit Anfangswert w_0

$$e^{\tau\widetilde{B}} = I + \tau\widetilde{B}. \tag{2.55}$$

Man erhält daher ein Verfahren der Form

$$u_{n+1} = S_{\mathsf{lin}} u_n = e^{\tau A} u_n + \frac{\tau}{2} \begin{bmatrix} \sin(\tau\Omega)\Omega^{-1}\widetilde{G} q_n \\ \cos(\tau\Omega)\Omega^{-1}\widetilde{G} q_n + \Omega^{-1}\widetilde{G} q_{n+1} \end{bmatrix}. \tag{2.56}$$

Hier erkennt man, dass das Splittingverfahren dem trigonometrischen Integrator (1.10a) entspricht, wenn man $\psi_S = \psi_1$ wählt und die anderen Filterfunktionen so gewählt werden, dass das Verfahren symmetrisch ist. Wegen der Symmetrie des Verfahrens gilt außerdem

$$q_{n+1} - 2\cos(\tau\Omega) q_n + q_{n-1} = \tau \sin(\tau\Omega)\Omega^{-1}\widetilde{G} q_n, \tag{2.57a}$$

mit Startschritt

$$q_1 = \cos(\tau\Omega) q_0 + \sin(\tau\Omega)\Omega^{-1} q_0' + \frac{\tau}{2}\sin(\tau\Omega)\Omega^{-1}\widetilde{G} q_0. \tag{2.57b}$$

Damit lässt sich das folgende Hauptresultat formulieren. Hier wird auf die Annahme 4.1 vorgegriffen, die später auch für die Konvergenzanalyse des semilinearen Problems benötigt wird:

Satz 2.5. *(vergleiche Theorem 3.2 aus Buchholz et al. (2018)) Es gelten die Annahmen 2.4 und 4.1. Dann ist der globale Fehler der Iterierten u_n des Splittingverfahrens* (2.56) *zur Lösung u von* (2.48) *beschränkt durch*

$$\|u_n - u(t_n)\| \leq C\tau^2, \qquad 0 \leq t_n = n\tau \leq t_{end},$$

wobei die Konstante C nur von C_{inv}, $\|G\|$, K, M_j, $j = 1, \ldots, 4$, und t_{end} abhängt, aber nicht von $\|\Omega\|$.

Der Beweis wird an dieser Stelle nicht angegeben, denn er ist als Spezialfall in der semilinearen Analyse der vorliegenden Arbeit enthalten, vergleiche Abschnitt 4.2. Teile des Beweises werden allerdings in Kapitel 4 als Motivation für den Konvergenzbeweis aufgeschrieben, vergleiche dazu die jeweiligen Abschnitte nach Lemma 4.5 und Satz 4.8).

Bemerkung 2.6. *Für eine Gleichung der Form* (2.48)*, auf die ein Strang-Splittingverfahren angewandt wird, wäre es zunächst naheliegend gewesen, das Konvergenzresultat aus Jahnke und Lubich (2000) anzuwenden, vergleiche Abschnitt 2.1.1. Um jedoch Konvergenz zweiter Ordnung nachzuweisen, muss man die Abschätzungen an den Kommutatoren* (2.5) *erfüllen (für $\beta = \alpha = 1$, damit die exakte Lösung lediglich eine Finite-Energie-Bedingung erfüllen muss). Jedoch ist die zweite Bedingung nicht erfüllt, denn*

$$[A, [A, B]]\, u = \begin{bmatrix} -2\Psi_S G \Phi \Omega v \\ -\Omega \Psi_S G \Phi q + \Psi \Omega^{-1} G \Phi \Omega^2 q \end{bmatrix} \qquad u = \begin{bmatrix} q \\ v \end{bmatrix}.$$

Dieser Kommutator muss nun durch eine Konstante multipliziert mit $\|Au\|$ beschränkt werden, vergleiche (2.5)*. Das ist jedoch nicht möglich, da hier sowohl Ω^2 als auch $\Omega\Psi_S G\Phi$ auftaucht und Ω und G im allgemeinen nicht kommutieren. Den lokalen Fehler proportional zu τ^3 zu beschränken, ist somit nicht möglich. Wie sich durch die Analyse in Buchholz et al. (2018) herausstellt, kann eine solche Schranke auch nicht gefunden werden und eine genauere Analyse des lokalen Fehlers und der Summationen für den globalen Fehler sind notwendig.*

2.2.8 Weiterführende Arbeiten zu trigonometrischen Integratoren

Ernst Hairer und Christian Lubich untersuchten in Hairer und Lubich (2000) Erhaltungsgrößen und deren Entwicklung über lange Zeitintervalle für die bis dorthin bekannten trigonometrischen Integratoren für hochoszillatorische Probleme. Dafür nutzten sie eine Analysetechnik, die modulierte Fourierentwicklung („Modulated Fourier Expansion“) genannt wird. In Hairer et al. (2006) wird diese Technik erläutert und zusätzlich verwendet, um Konvergenz des Verfahrens für eine spezielle Wahl der Matrix Ω nachzuweisen. Die modulierte Fourierentwicklung wurde zahlreich zur Untersuchung des Langzeitverhaltens eingesetzt, siehe Cohen, Hairer und Lubich (2005), Cohen (2006) oder Cohen, Gauckler, Hairer und Lubich (2015). Weiterhin wird sie verwendet um Verfahren für hochoszillatorische Probleme, die von einem betragsmäßig großen Parameter abhängen, zu analysieren, siehe Abschnitt 1.1.3.

Für Systeme der Form

$$q''(t) = -\Omega^2(t, q)q + g(t, q), \qquad q(t_0) = q_0, \quad q'(t_0) = q_0'$$

mit einer zeit- und Lösungs-abhängigen, symmetrischen und positiv semidefiniten Matrix $\Omega^2(t, y)$ von beliebig großer Norm wurden von Volker Grimm in Grimm (2002) (für $\Omega(t, y) = \Omega(t)$) und in

Grimm (2005b) betrachtet. Hier wurde jeweils auch eine Fehleranalyse für trigonometrische Integratoren, die für zeitabhängige Frequenzen erweitert worden sind, durchgeführt, wobei man annimmt, dass die erste und zweite partielle Ableitung von $\Omega(t, y)$ beschränkt ist. Allerdings taucht in der Fehlerkonstanten der Faktor $l(n, d)$ auf, der bereits in der Analyse von Hochbruck und Lubich (1999) (siehe Abschnitt 2.2.4) vorkam. Die Fehlerkonstante ist somit nicht unabhängig von der Anzahl der Schritte des Verfahrens, beziehungsweise der Dimension der Matrix Ω. Der Faktor kann durch Techniken aus Grimm (2005a) vermieden werden.

Es gibt weiterhin Arbeiten, die höhere Regularität an die Lösung fordern, um beispielsweise Konvergenz der Verfahren mit höherer Ordnung nachzuweisen, siehe Cano und Moreta (2010) und Cano und Moreta (2013), oder um mit Nichtlinearitäten umzugehen, die nicht Lipschitz-stetig sind, siehe Dong (2014). Auch wurde der trigonometrische Integrator auf stochastische Differentialgleichungen erweitert, siehe Cohen, Larsson und Sigg (2013).

KAPITEL 3

DARSTELLUNG DER TRIGONOMETRISCHEN INTEGRATOREN IN FORM EINES SPLITTINGVERFAHREN

In Buchholz et al. (2018) wurde bereits für den linearen Fall vorgeschlagen, den trigonometrischen Integrator (1.10a) angewandt auf ein System von Differentialgleichungen der Form (1.8) als Splitting-Verfahren angewandt auf eine modifizierte Gleichung zu analysieren. In diesem Kapitel soll diese Analyse nun auf den semilinearen Fall übertragen werden. Hierfür wird zunächst die folgende zusätzliche Annahme an die Differentialgleichung (1.1) gestellt:

Annahme 3.1. *Die Inverse von Ω sei mit einer kleinen Konstante C_{inv} beschränkt:*

$$\left\|\Omega^{-1}\right\| \leq C_{inv}. \tag{3.1}$$

Bemerkung 3.2. *Die Invertierbarkeit der Matrix Ω wurde bei der Fehleranalyse des Gautschi-Verfahrens oder der trigonometrischen Integratoren in Grimm und Hochbruck (2006) nicht verwendet, siehe dazu auch die Abschnitte 2.2.4 und 2.2.5. Hier wird lediglich gefordert, dass die Matrix Ω positiv semidefinit ist.*

Liegt ein System der Form (1.1) *mit positiv semi-definiter Matrix Ω vor, so betrachtet man das äquivalente System*

$$\begin{aligned} q''(t) &= -\widehat{\Omega}^2 q(t) + \widehat{g}(q(t)), \quad 0 \leq t \leq t_{end}, \\ q(0) &= q_0,\ q'(0) = q_0'. \end{aligned}$$

mit

$$\widehat{\Omega}^2 = \Omega^2 + \gamma I, \qquad\qquad \widehat{g}(q) = g(q) + \gamma q,$$

und einem geeigneten skalaren Shift $\gamma > 0$ so, dass $\widehat{\Omega}$ positiv definit ist. Wichtig ist, dass die Norm von $\widehat{g}$ im äquivalenten System nicht von der Norm von Ω abhängt. Ein Beispiel für einen solchen Shift, sieht man im Kapitel 5.3.

Um das Splittingverfahren zu konstruieren, geht man wie im linearen Fall vor. Zunächst schreibt man (1.1) um in ein System erster Ordnung. Zusätzlich wird eine Variablentransformation in der Ableitung q' durchgeführt. Die neuen Variable u mit passendem Anfangswert

$$u = \begin{bmatrix} q \\ \Omega^{-1} q' \end{bmatrix} \qquad \text{mit} \qquad u_0 = \begin{bmatrix} q_0 \\ \Omega^{-1} q_0' \end{bmatrix}$$

erfüllt dann die Differentialgleichung

$$u'(t) = Au(t) + b(u(t)), \qquad u(0) = u_0 \tag{3.2a}$$

mit

$$A = \begin{bmatrix} 0 & \Omega \\ -\Omega & 0 \end{bmatrix}, \qquad \text{und} \qquad b\left(\begin{bmatrix} q \\ v \end{bmatrix}\right) = \begin{bmatrix} 0 \\ \Omega^{-1} g(q) \end{bmatrix}. \tag{3.2b}$$

Man kann nachrechnen, dass $\|A\| = \|\Omega\|$ ist (und damit auch entsprechend groß ist) und auch $\left\|A^{-1}\right\| = \left\|\Omega^{-1}\right\| \leq C_{\text{inv}}$ gilt. Die Lösung u erfüllt die Finite-Energie-Bedingung in Form von

$$\|Au(t)\| \leq K, \qquad 0 \leq t \leq t_{\text{end}}, \tag{3.3}$$

wenn die Lösung q der Differentialgleichung (1.1) die Annahme 1.1 b) erfüllt.

Für (3.2a) ist es naheliegend ein Splittingverfahren zu verwenden, um eine numerische Approximation an die Lösung zu berechnen. Die Lösung der Teilprobleme

$$\begin{aligned} v'(t) &= Av(t) && \text{mit } v(0) = v_0, \\ w'(t) &= b(w(t)) && \text{mit } w(0) = w_0, \end{aligned}$$

ist gegeben durch

$$v(t) \;=\; e^{tA}\, v_0, \qquad \text{mit} \quad e^{tA} = \begin{bmatrix} \cos(t\Omega) & \sin(t\Omega) \\ -\sin(t\Omega) & \cos(t\Omega) \end{bmatrix}$$

und

$$w(t) = \varphi_b\,(t, w_0) = \begin{bmatrix} w_{0,1} \\ \Omega^{-1} g(w_{0,1})t + w_{0,2} \end{bmatrix} = w_0 + b(w_0)t \qquad \text{für} \quad w_0 = \begin{bmatrix} w_{0,1} \\ w_{0,2} \end{bmatrix} \tag{3.4}$$

Da A schief-symmetrisch ist, ist e^{tA} unitär und damit

$$\left\|e^{tA}\right\| = 1, \qquad t \in \mathbb{R}. \tag{3.5}$$

Es soll nun ein Strang-Splittingverfahren verwendet werden, um die Lösung zu approximieren. Zunächst sei

$$u_{n+1} = \widehat{T}(u_n) = \varphi_b\left(\frac{\tau}{2}, e^{\tau A}\varphi_b\left(\frac{\tau}{2}, u_n\right)\right)$$

gegeben. Setzt man nun den Fluss $\varphi_b\left(\frac{\tau}{2}, \cdot\right)$ ein, so erhält man für $u_n = [q_n^T, v_n^T]^T$ wie im linearen Fall

$$\begin{aligned} \widehat{T}(u_n) &= e^{\tau A}\varphi_b\left(\frac{\tau}{2}, u_n\right) + \frac{\tau}{2} b(e^{\tau A}\varphi_b\left(\frac{\tau}{2}, u_n\right)) \\ &= e^{\tau A}\left(u_n + \frac{\tau}{2} b(u_n)\right) + \frac{\tau}{2} b(e^{\tau A}\varphi_b\left(\frac{\tau}{2}, u_n\right)) \\ &= e^{\tau A} u_n + \frac{\tau}{2}\begin{bmatrix} \sin(\tau\Omega)\Omega^{-1} g(q_n) \\ \cos(\tau\Omega)\Omega^{-1} g(q_n) + \Omega^{-1} g(q_{n+1}) \end{bmatrix}. \end{aligned}$$

Dabei wurde im letzten Schritt verwendet, dass die erste Block-Komponenten von $e^{\tau A}\varphi_{\widetilde{b}}\left(\frac{\tau}{2}, u_n\right)$ gerade q_{n+1} entspricht. Man erkennt nun, dass das Splittingverfahren $\widehat{T}$ genau dem trigonometrischen Integrator (1.10a) angewandt auf das transformierte Problem (3.2a) mit Filterfunktionen $\Phi \equiv \Psi_1 \equiv I$, $\Psi = \operatorname{sinc}(\tau\Omega)$ und $\Psi_0 = \cos(\tau\Omega)$ entspricht.

Wie jedoch in Abschnitt 1.2 und der Konvergenzanalyse in Abschnitt 2.2.5 gesehen, ist der Intergator $\widehat{T}$ nicht konvergent von Ordnung zwei. Es müssen geeignete Filterfunktionen eingesetzt werden oder der Zeitschritt τ einer starken CFL-Bedingung unterliegen, um Konvergenz des Verfahrens zu garantieren. Deswegen werden im folgenden Abschnitt Filterfunktionen hinzugefügt, um ein Verfahren zu erhalten, welches konvergent von Ordnung zwei ohne starke Schrittweiteneinschränkung ist.

3.1 Modifizierte Gleichung

Wie bereits im linearen Fall, vergleiche Abschnitt 2.2.7, betrachtet man nun die modifizierte Gleichung

$$\widetilde{u}'(t) = A\widetilde{u} + \widetilde{b}(\widetilde{u}), \qquad \widetilde{u}(0) = u_0, \tag{3.6}$$

mit

$$\widetilde{b}(u) = \begin{bmatrix} 0 \\ \Omega^{-1}\Psi_S g(\Phi q) \end{bmatrix}, \qquad u = \begin{bmatrix} q \\ v \end{bmatrix}. \tag{3.7}$$

Man beachte, dass die Lösung $\widetilde{u}$ der modifizierten Gleichung von der gewählten Schrittweite τ abhängt. Um die Notation kurz zu halten, schreiben wir lediglich $\widetilde{u}(t)$ anstelle von $\widetilde{u}(t, \tau)$ oder $\widetilde{u}_\tau(t)$. In den folgenden Kapiteln wird weiterhin die Kurzschreibweise

$$\widetilde{b}(u) = \widetilde{\Psi} b(\widetilde{\Phi} u)$$

mit den Blockmatrizen

$$\widetilde{\Psi} = \begin{bmatrix} \Psi_S & 0 \\ 0 & \Psi_S \end{bmatrix} \qquad \text{und} \qquad \widetilde{\Phi} = \begin{bmatrix} \Phi & 0 \\ 0 & \Phi \end{bmatrix} \tag{3.8}$$

verwendet.

Definiert man nun die exakte Lösung $\widetilde{w}$ von

$$\widetilde{w}' = \widetilde{b}(\widetilde{w}), \qquad \text{mit } \widetilde{w}(0) = \widetilde{w}_0,$$

analog zu $\varphi_b(t, w)$ in (3.4) über

$$\widetilde{w}(t) = \varphi_{\widetilde{b}}(t, \widetilde{w}_0) = \widetilde{w}_0 + \widetilde{b}(\widetilde{w}_0)t, \tag{3.9}$$

so ist der trigonometrische Integrator (1.10a) äquivalent zum Strang-Splittingverfahren angewandt auf die modifizierte Gleichung (3.6), das heißt

$$u_{n+1} = T(u_n) = \varphi_{\widetilde{b}}\left(\frac{\tau}{2}, e^{\tau A}\varphi_{\widetilde{b}}\left(\frac{\tau}{2}, u_n\right)\right).$$

Wählt man mit $\psi_S = \psi_1$ und ψ und ψ_0 so, dass das Verfahren symmetrisch ist, so entspricht dies dem trigonometrischen Integrator (1.10a). In der vorliegenden Arbeit wird allerdings das Strang-Splittingverfahren in vertauschter Reihenfolge

$$u_{n+1} = S(u_n) = e^{\frac{\tau}{2}A}\varphi_{\widetilde{b}}\left(\tau, e^{\frac{\tau}{2}A}u_n\right) \tag{3.10}$$

analysiert. Durch die geänderte Reihenfolge ist pro Zeitschritt nur eine Funktionsauswertung der Nichtlinearität g nötig, was sich in der Analyse als vorteilhaft herausgestellt hat. Implementiert man die Verfahren, kommt T mit $n+1$ und S mit n Funktionsauswertungen aus, da für den Integrator T eine Funktionsauswertung im zweiten Schritt wieder verwendet werden kann. Der Zusammenhang zwischen beiden Reihenfolgen wird in Abschnitt 3.2 untersucht. Man kann den Konvergenzbeweis in Buchholz et al. (2018) auch in der veränderten Reihenfolge des Splittingverfahrens durchführen und erhält unter identischen Voraussetzungen Konvergenz zweiter Ordnung.

Zur Implementierung des Splittingverfahrens (3.10) ist häufig die folgende Schreibweise günstig:

$$\begin{aligned}
\begin{bmatrix} q_1 \\ v_1 \end{bmatrix} &= e^{\frac{\tau}{2}A}\varphi_{\widetilde{b}}\left(\tau, e^{\frac{\tau}{2}A}\begin{bmatrix} q_0 \\ v_0 \end{bmatrix}\right) \\
&= e^{\frac{\tau}{2}A}\left(e^{\frac{\tau}{2}A}\begin{bmatrix} q_0 \\ v_0 \end{bmatrix} + \tau\widetilde{b}\left(e^{\frac{\tau}{2}A}\begin{bmatrix} q_0 \\ v_0 \end{bmatrix}\right)\right) \\
&= e^{\tau A}\begin{bmatrix} q_0 \\ v_0 \end{bmatrix} + \tau e^{\frac{\tau}{2}A}\begin{bmatrix} 0 \\ \Omega^{-1}\Psi_S g\left(\Phi\left(\cos\left(\frac{\tau}{2}\Omega\right)q_0 + \sin\left(\frac{\tau}{2}\Omega\right))v_0\right)\right) \end{bmatrix} \\
&= e^{\tau A}\begin{bmatrix} q_0 \\ v_0 \end{bmatrix} + \tau\begin{bmatrix} \Omega^{-1}\Psi_S \sin(\frac{\tau}{2}\Omega)\tilde{g} \\ \Omega^{-1}\Psi_S \cos(\frac{\tau}{2}\Omega)\tilde{g} \end{bmatrix},
\end{aligned}$$

mit $\tilde{g} = g\left(\Phi\left(\cos\left(\frac{\tau}{2}\Omega\right)q_0 + \sin\left(\frac{\tau}{2}\Omega\right))v_0\right)\right)$.

3.2 Reihenfolge des Splittingverfahrens

Im Folgenden sollen die beiden numerischen Verfahren

$$\phi_\tau(u) = \varphi_{\widetilde{b}}\left(\tau, e^{\tau A}u\right) \qquad \text{und} \qquad \phi^*_\tau(u) = e^{\tau A}\varphi_{\widetilde{b}}\left(\tau, u\right) \tag{3.11}$$

betrachtet werden, die beide Kompositionen zweier Flüsse von Differentialgleichungen sind. Wegen $\varphi_{\widetilde{b}}(t,\cdot)^{-1} = \varphi_{\widetilde{b}}(-t,\cdot)$ ist die zweite Methode die Adjungierte der ersten. Die beiden Verfahren (3.11) werden als Lie-Trotter-Splittingverfahren bezeichnet, siehe Trotter (1959).

Mithilfe ϕ_τ und ϕ^*_τ lassen sich die beiden, in Abschnitt 3.1 vorgestellten Operatoren S und T konstruieren. Es ist

$$T(u) = \phi_{\tau/2} \circ \phi^*_{\tau/2}(u), \qquad \text{und} \qquad S(u) = \phi^*_{\tau/2} \circ \phi_{\tau/2}(u).$$

Nun lässt sich leicht zeigen, dass beide Operatoren T und S symmetrische Verfahren definieren. Weiterhin lässt sich folgende Beziehungen zwischen den Operatoren herstellen

$$\begin{aligned}
T \circ \phi_{\tau/2}(u) = \phi_{\tau/2} \circ S(u) \qquad &\Leftrightarrow \qquad S(u) = \phi^*_{-\tau/2} \circ T \circ \phi_{\tau/2}(u), \\
S \circ \phi^*_{\tau/2}(u) = \phi^*_{\tau/2} \circ T(u) \qquad &\Leftrightarrow \qquad T(u) = \phi_{-\tau/2} \circ S \circ \phi^*_{\tau/2}(u).
\end{aligned}$$

Für die n-te Iterierte folgt daher

$$S^n(u_0) = \phi^*_{-\tau/2} \circ T^n \circ \phi_{\tau/2}(u_0), \qquad T^n(u_0) = \phi_{-\tau/2} \circ S^n \circ \phi^*_{\tau/2}(u_0).$$

Die beiden Operatoren T und S entscheiden sich also jeweils nur um einen Halbschritt zu Beginn und am Schluss der Simulation. Da das Verfahren mit dem Operator T äquivalent zum trigonometrischen Integrator in Grimm und Hochbruck (2006) angewandt auf das transformierte Problem (3.2a) ist, weiß man, dass es konvergent von Ordnung zwei in den Positionen und konvergent der Ordnung eins in den (untransformierten) Geschwindigkeiten ist. Aus der linearen Analyse in Buchholz et al. (2018) folgt außerdem, dass es für eine lineare Funktion g konvergent der Ordnung zwei in den Positionen und den transformierten Geschwindigkeiten ist. Es stellt sich die Frage, ob man aus den Beziehungen zwischen S und T und der Konvergenz des Verfahrens mit Operator T direkt schließen kann, dass auch das Verfahren mit dem Operator S konvergent der Ordnung zwei ist?

Hierzu betrachtet man zunächst die Verfahren mit den Operatoren T und S an einem numerischen Beispiel. In Abbildung 3.1 wird der Fehler der Verfahren, der bei Simulation des nichtlinearen Fermi-Pasta-Ulam-Tsingou-Problems auftritt, vergleiche Kapitel 5, dargestellt. Dieses Beispiel untermauert die Vermutung, dass die Reihenfolge des Splittingverfahrens keinen wesentlichen Einfluss auf die Größe des Fehlers hat. Beide Verfahren konvergieren mit zweiter Ordnung, die Fehlerkonstanten variieren nur leicht.

Es ist außerdem möglich, den Konvergenzbeweis aus Buchholz et al. (2018) für das Verfahren mit dem Operator S aufzuschreiben. Konvergenz zweiter Ordnung dieses Verfahrens für semilineare Probleme (1.1) konnte jedoch bislang nur durch den Konvergenzbeweis in Kapitel 4 gezeigt werden und nicht direkt aus der Beziehung zu T hergeleitet werden.

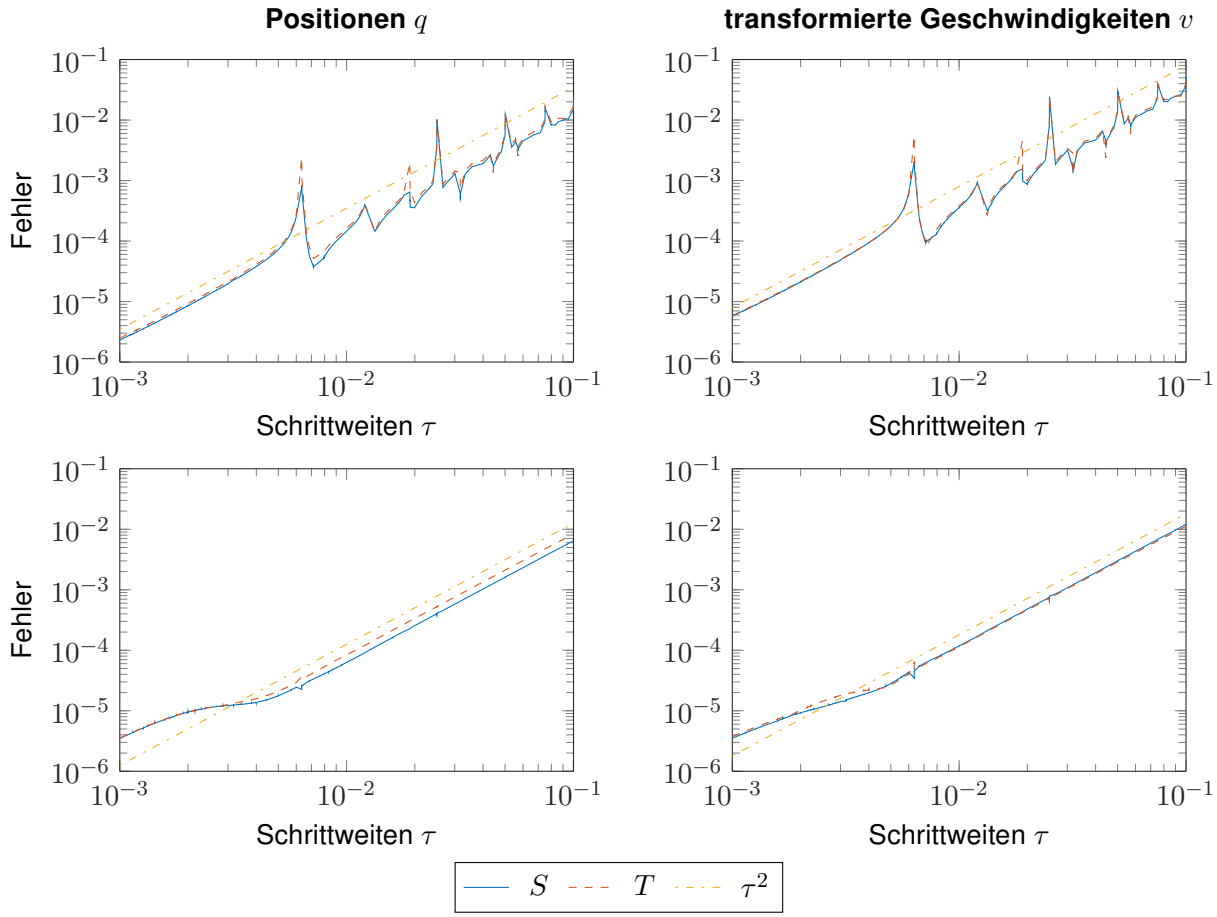

Abbildung 3.1: Vergleich der Fehler der Verfahren mit den Operatoren T und S ohne Filterfunktionen (erste Zeile) und mit Filterfunktionen (zweite Zeile)

KAPITEL 4

KONVERGENZRESULTAT UND FEHLERANALYSE DES SPLITTINGVERFAHRENS

Zunächst sei $\tau < 1$ vorausgesetzt. Um zu beweisen, dass das Splittingverfahren S konvergent von Ordnung zwei ist, benötigen wir folgende Annahme an die Filterfunktionen Ψ_S bzw. Φ:

Annahme 4.1. *Die Filterfunktionen $\chi = \phi$ oder $\chi = \psi_S$ seien gerade und analytisch. Weiterhin seien die folgenden Bedingungen erfüllt*

$$\chi(0) = 1, \tag{4.1a}$$
$$|\chi(\xi)| \leq M_1, \tag{4.1b}$$
$$|\xi\chi(\xi)| \leq M_2, \tag{4.1c}$$
$$\left|\cot\left(\frac{\xi}{2}\right)\xi\chi(\xi)\right| \leq M_3 \tag{4.1d}$$

mit Konstanten M_j, die gleichmäßig für alle $\xi \in \mathbb{R}$ gelten.

Eine Funktion, die alle Bedingungen erfüllt, ist beispielsweise $\chi(\xi) = \operatorname{sinc}(\xi)$, die auch für trigonometrische Integratoren verwendet wurde, siehe beispielsweise Kapitel 2 und die Filter aus (2.32). Da χ gerade und analytisch ist, folgt aus Bedingung (4.1a) und (4.1b)

$$\left|\xi^{-2}(1-\chi(\xi))\right| \leq M_4, \tag{4.2}$$

mit einer Konstante M_4, für alle $\xi \in \mathbb{R}$ gilt. Diese Bedingung wird im Konvergenzbeweis an mehreren Stellen benötigt und hier deshalb zusätzlich aufgeführt.

Im folgenden Abschnitt soll gezeigt werden, dass für den Fehler des Splittingverfahrens (3.10) angewandt auf (3.2a) gilt

$$\|u_n - u(t_n)\| \leq C\tau^2, \qquad 0 \leq t_n = n\tau \leq t_{\text{end}}.$$

Dabei steht C hier und nachfolgend für eine generische Konstante, die an verschiedenen Stellen unterschiedliche Werte annehmen kann. Die Konstante C hängt jedoch nicht von $\|\Omega\|$, τ oder n ab.

Der Beweis dieser Aussage ist aufwendig und wird daher in mehrere Teile unterteilt. Zunächst wird gezeigt, dass der Fehler zwischen der exakten Lösung u der Ausgangsgleichung (3.2a) und der

Lösung $\widetilde{u}$ der modifizierten Gleichung (3.6) sich wie τ^2 verhält, siehe Satz 4.2. Außerdem erfüllt die Lösung $\widetilde{u}$ ebenfalls eine Finite-Energie-Bedingung, vergleiche Lemma 4.3. Diese Abschätzungen sind notwendig, damit man nachfolgend nur den Fehler der Strang-Splittingverfahrens u_n aus (3.10) zur Lösung $\widetilde{u}$ betrachten muss. Die modifizierte Gleichung stellt somit eine τ^2-Störung der Gleichung (1.1) dar.

Anschließend wird der lokale Fehler zwischen dem Strang-Splittingverfahren u_n und der Lösung $\widetilde{u}$ in eine geeignete Form gebracht, um den globalen Fehler abzuschätzen, vergleiche Lemma 4.5 und Lemma 4.7. Insgesamt lässt sich so zunächst zeigen, dass der Fehler zwischen der numerischen Lösung und der Lösung der modifizierten Gleichung 3.6 von der Größenordnung $\mathcal{O}(\tau^2)$ ist. Die obige Konvergenzaussage ergibt sich aus Kombination der erzielten Ergebnisse.

4.1 Fehleranalyse

Um den Konvergenzbeweis durchzuführen, benötigt man die Ableitungen von $\varphi_b\,(t, u_0)$ nach dem Anfangswert u_0. Aus der Darstellung (3.4) des Flusses ergibt sich

$$\partial_2\varphi_b\,(t, u_0) = \begin{bmatrix} I & 0 \\ \Omega^{-1}g'(q_0)t & I \end{bmatrix}, \qquad \text{für } u_0 = \begin{bmatrix} q_0 \\ v_0 \end{bmatrix}. \tag{4.3}$$

Mithilfe der Jacobi-Matrix von b

$$J_b(u) = \begin{bmatrix} 0 & 0 \\ \Omega^{-1}g'(q) & 0 \end{bmatrix}, \qquad \text{für } u = \begin{bmatrix} q \\ v \end{bmatrix}, \tag{4.4}$$

ergibt sich

$$\big(\partial_2\varphi_b\,(t, u_0)\,\big)(u) = \big(I + tJ_b(q_0)\big)u. \tag{4.5}$$

Für die zweite Ableitung nach dem Anfangswert definiert man den Vektor

$$H_b(u)(y, z) = \begin{bmatrix} 0 \\ \Omega^{-1}g''(q_u)(q_y, q_z) \end{bmatrix}, \qquad \text{für } u = \begin{bmatrix} q_u \\ v_u \end{bmatrix},\ y = \begin{bmatrix} q_y \\ v_y \end{bmatrix},\ z = \begin{bmatrix} q_z \\ v_z \end{bmatrix}. \tag{4.6}$$

Analog folgen für das modifizierte System

$$J_{\widetilde{b}}(u) = \begin{bmatrix} 0 & 0 \\ \Omega^{-1}\Psi_S g'(\Phi q)\Phi & 0 \end{bmatrix} \quad \text{und} \quad H_{\widetilde{b}}(u)(y, z) = \begin{bmatrix} 0 \\ \Omega^{-1}\Psi_S g''(\Phi q_u)(\Phi q_y, \Phi q_z) \end{bmatrix}.$$

4.1.1 Eigenschaften der modifizierten Gleichung

Zunächst untersucht man den Fehler zwischen der Lösung u der Ausgangsgleichung (3.2a) und der Lösung $\widetilde{u}$ des modifizierte Systems (3.6). Wie in Abschnitt 3.1 erklärt, hängt die Lösung $\widetilde{u}$ von der Zeitschrittweite τ ab. Man zeigt nun, dass $\widetilde{u}$ gegen die Lösung u konvergiert, wenn τ gegen 0 geht.

Um den Beweis durchzuführen, wird die Annahme 1.1 an die Differentialgleichung gestellt. Zunächst werden dafür Radien

$$r_0 = M_1 C_{\text{inv}} K, \qquad r_1 = (1 + M_1)C_{\text{inv}} K, \qquad r_2 = (1 + M_1)C_{\text{inv}} K, \tag{4.7}$$

gewählt, für welche die Funktion g und ihre Ableitungen zunächst die Schranken

$$\begin{aligned} \|g(q)\| &\leq C_{g,0} && \text{für } \|q\| \leq r_0, \\ \|g'(q)\| &\leq C_{g,1} && \text{für } \|q\| \leq r_1, \\ \|g''(q)\| &\leq C_{g,2} && \text{für } \|q\| \leq r_2, \end{aligned}$$

erfüllen. Mit diesen Schranken lässt sich Satz 4.2 und Lemma 4.3 beweisen. Anschließend werden die Radien vergrößert, für die man auch die Darstellung des lokalen Fehlers in Lemma 4.5 und die Konvergenz des globalen Fehlers in Satz 4.8 nachweisen kann. Diese zweistufige Definition ist notwendig, da die vergößerten Radien aus (4.21) von der in Lemma 4.3 definierten Schranke für die Finite-Energie der modifizierten Gleichung abhängen.

Satz 4.2. *Es seien die Annahme 1.1 mit r_0, r_1 und r_2 aus* (4.7) *und die Annahme 3.1 erfüllt. Weiterhin erfüllen die Filterfunktionen die Annahmen* (4.1a) *und* (4.1b)*. Dann gilt für den Fehler zwischen der Lösung u von* (3.2a) *und der Lösung $\widetilde{u}$ der modifizierten Gleichung* (3.2a)

$$\|u(t) - \widetilde{u}(t)\| \leq C_{av}\tau^2, \qquad 0 \leq t \leq t_{end},$$

mit einer Konstante C_{av}, die nur von $C_{g,0}$, $C_{g,1}$, $C_{g,2}$, C_{inv}, K, L_g, M_1, M_4 und t_{end} abhängt.

Der Beweis orientiert sich am Beweis von Theorem 4.1 aus Buchholz et al. (2018).
Beweis. Um den Fehler darzustellen, schreibt man die Lösungen mithilfe der Variation-der-Konstanten-Formel

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A} b(u(s))\, ds, \tag{4.8a}$$

$$\widetilde{u}(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A} \widetilde{b}(\widetilde{u}(s))\, ds. \tag{4.8b}$$

Es ist somit

$$\begin{aligned} u(t) - \widetilde{u}(t) &= \int_0^t e^{(t-s)A}(b(u(s)) - \widetilde{b}(\widetilde{u}(s)))\, ds \\ &= I_1(t) + I_2(t) + \int_0^t e^{(t-s)A}\widetilde{\Psi}\big(b(\widetilde{\Phi}u(s)) - b(\widetilde{\Phi}\widetilde{u}(s))\big)\, ds, \end{aligned}$$

mit

$$\begin{aligned} I_1(t) &= \int_0^t e^{(t-s)A}(I - \widetilde{\Psi})b(u(s))\, ds \\ I_2(t) &= \int_0^t e^{(t-s)A}\widetilde{\Psi}\big(b(u(s)) - b(\widetilde{\Phi}u(s))\big)\, ds. \end{aligned}$$

Man zeigt nun, dass beide Terme I_1 und I_2 proportional zu τ^2 beschränkt sind. Zunächst sei festgehalten, dass man die Matrix A und die Filterfunktionen $\widetilde{\Phi}$ und $\widetilde{\Psi}_S$ vertauschen darf, da diese Matrizen nur von Ω abhängen. Für I_1 ergibt sich mithilfe partieller Integration und Ausklammern von τ^2:

$$\begin{aligned} I_1(t) &= \Big[- A^{-1}e^{(t-s)A}(I - \widetilde{\Psi})b(u(s))\Big]_0^t - \int_0^t (-A)^{-1}e^{(t-s)A}(I - \widetilde{\Psi})J_b(u(s))u'(s)\, ds, \\ &= \Big[- \tau^2 e^{(t-s)A}(\tau A)^{-2}(I - \widetilde{\Psi})Ab(u(s))\Big]_0^t + \tau^2 \int_0^t e^{(t-s)A}(\tau A)^{-2}(I - \widetilde{\Psi})AJ_b(u(s))u'(s)\, ds. \end{aligned}$$

Unter Verwendung der Annahmen (1.3) mit den Radien aus (4.7), (3.1) und (3.3) folgt zunächst

$$\|u(s)\| = \left\|A^{-1}Au(s)\right\| \leq C_{\text{inv}}K \leq \min(r_0, r_1), \tag{4.9}$$
$$\left\|u'(s)\right\| = \|Au(s) + b(u(s))\| \leq K + C_{\text{inv}}C_{g,0}, \tag{4.10}$$

Daraus ergibt sich mithilfe von (3.5) (4.1b), (4.2)

$$\|Ab(u(s))\| = \left\| \begin{bmatrix} 0 \\ g(q_u) \end{bmatrix} \right\| \leq C_{g,0}, \qquad u(s) = [q_u^T, v_u^T]^T \text{ und } \|u(s)\| \leq r_0, \tag{4.11}$$
$$\|AJ_b(u(s))\| = \left\| \begin{bmatrix} 0 & 0 \\ g'(q_u) & 0 \end{bmatrix} \right\| \leq C_{g,1}, \qquad u(s) = [q_u^T, v_u^T]^T \text{ und } \|u(s)\| \leq r_1 \tag{4.12}$$

und so ingesamt

$$\|I_1(t)\| \leq \left(2M_4C_{g,0} + t_{\text{end}}M_4C_{g,1}\,(K + C_{\text{inv}}C_{g,0})\right)\tau^2 \leq C_1\tau^2, \tag{4.13}$$

mit einer Konstante C_1, die nur von M_4, $C_{g,0}$, $C_{g,1}$, C_{inv}, K und t_{end} abhängt.

Um I_2 zu beschränken, wendet man zunächst den Hauptsatz der Differential- und Integralrechnung auf die Differenz $b(u(s)) - b(\widetilde{\Phi}u(s))$ an. Anschließend wird die Variation-der-Konstanten-Formel (4.8a) verwendet, um $u(s)$ zu ersetzen. Es folgt

$$\begin{aligned} I_2(t) &= \int_0^t e^{(t-s)A}\widetilde{\Psi} \int_0^1 J_b\Big(\sigma u(s) + (1-\sigma)\widetilde{\Phi}u(s)\Big)(I - \widetilde{\Phi})u(s)\,d\sigma\,ds \\ &= I_{21}(t) + I_{22}(t) \end{aligned}$$

mit

$$\begin{aligned} I_{21}(t) &= \int_0^t e^{(t-s)A}\widetilde{\Psi} \int_0^1 J_b\Big(\sigma u(s) + (1-\sigma)\widetilde{\Phi}u(s)\Big)(I - \widetilde{\Phi})e^{sA}u_0\,d\sigma\,ds, \\ I_{22}(t) &= \int_0^t e^{(t-s)A}\widetilde{\Psi} \int_0^1 J_b\Big(\sigma u(s) + (1-\sigma)\widetilde{\Phi}u(s)\Big)(I - \widetilde{\Phi}) \int_0^s e^{(s-\theta)A}b(u(\theta))\,d\theta\,d\sigma\,ds. \end{aligned}$$

Im nächsten Abschnitt wird gezeigt, dass die Normen von I_{21} und I_{22} ebenfalls proportional zu τ^2 beschränkt werden können. Zuerst verwendet man partielle Integration nach s für I_{21} (integriert wird

dabei lediglich e^{sA}) und danach wird τ^2 ausgeklammert. Es folgt

$$\begin{aligned}
I_{21}(t) &= \left[e^{(t-s)A}\widetilde{\Psi}\int_0^1 J_b\Big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\Big)(I-\widetilde{\Phi})A^{-1}e^{sA}u_0\,d\sigma\right]_{s=0}^{t}\\
&\quad-\int_0^t e^{(t-s)A}(-A)\widetilde{\Psi}\int_0^1 J_b\Big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\Big)(I-\widetilde{\Phi})A^{-1}e^{sA}u_0\,d\sigma\,ds\\
&\quad-\int_0^t e^{(t-s)A}\widetilde{\Psi}\int_0^1 H_b\big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\big)\\
&\qquad\qquad\Big(\sigma u'(s)+(1-\sigma)\widetilde{\Phi}u'(s),\ (I-\widetilde{\Phi})A^{-1}e^{sA}u_0\Big)\,d\sigma\,ds\\
&= \tau^2\left[e^{(t-s)A}\widetilde{\Psi}\int_0^1 J_b\Big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\Big)(I-\widetilde{\Phi})(\tau A)^{-2}e^{sA}Au_0\,d\sigma\right]_{s=0}^{t}\\
&\quad+\tau^2\int_0^t e^{(t-s)A}\widetilde{\Psi}\int_0^1 AJ_b\Big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\Big)(I-\widetilde{\Phi})(\tau A)^{-2}e^{sA}Au_0\,d\sigma\,ds\\
&\quad-\tau^2\int_0^t e^{(t-s)A}\widetilde{\Psi}\int_0^1 H_b\big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\big)\\
&\qquad\qquad\Big(\sigma u'(s)+(1-\sigma)\widetilde{\Phi}u'(s),\ (I-\widetilde{\Phi})(\tau A)^{-2}e^{sA}Au_0\Big)\,d\sigma\,ds,
\end{aligned}$$

wobei man ausnutzen kann, dass A, e^{tA} und Ψ_S kommutieren. Mit Hilfe von (3.1), (1.3), (3.5), (4.1b), (4.2), (4.9), (4.10), (4.12) und

$$\Big\|\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\Big\| \le (1+M_1)C_{\mathsf{inv}}K \le \min(r_1,r_2) \tag{4.14}$$

folgt zunächst

$$\Big\|J_b\Big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\Big)\Big\| \le C_{\mathsf{inv}}C_{g,1}, \tag{4.15}$$

$$\Big\|H_b\big(\sigma u(s)+(1-\sigma)\widetilde{\Phi}u(s)\big)(v,w)\Big\| \le C_{\mathsf{inv}}C_{g,2}\ \|v\|\,\|w\|\,. \tag{4.16}$$

Insgesamt erhält man

$$\begin{aligned}
\|I_{21}(t)\| &\le \left(2M_1C_{\mathsf{inv}}C_{g,1}M_4K + t_{\mathsf{end}}M_1\Big(C_{g,1}M_4K + C_{\mathsf{inv}}C_{g,2}(1+M_1)(K+C_{\mathsf{inv}}C_{g,0})M_4K\Big)\right)\tau^2\\
&\le C_{21}\tau^2,
\end{aligned} \tag{4.17}$$

mit einer Konstante C_{21}, die nur von M_1, M_4, $C_{g,0}$ $C_{g,1}$, $C_{g,2}$, C_{inv}, K und t_{end} abhängt.

Um den zweiten Term I_{22} abzuschätzen, wird zuerst partiell nach θ integriert

$$\begin{aligned}
(I-\widetilde{\Phi})\int_0^s e^{(s-\theta)A}b(u(\theta))\,d\theta &= (I-\widetilde{\Phi})(-A)^{-1}e^{(s-\theta)A}b(u(\theta))\Big|_{\theta=0}^{s}\\
&\quad+(I-\widetilde{\Phi})\int_0^s A^{-1}e^{(s-\theta)A}J_b(u(\theta))u'(\theta)\,d\theta\\
&= -\tau^2(I-\widetilde{\Phi})(\tau A)^{-2}e^{(s-\theta)A}Ab(u(\theta))\Big|_{\theta=0}^{s}\\
&\quad+\tau^2(I-\widetilde{\Phi})(\tau A)^{-2}\int_0^s e^{(s-\theta)A}AJ_b(u(\theta))u'(\theta)\,d\theta.
\end{aligned}$$

Mit (3.5), (4.1b), (4.2), (4.9), (4.11), (4.10), (4.12), (4.15) und (4.14) ergibt sich folgende obere Schranke an I_{22}:

$$\|I_{22}(t)\| \leq t_{\mathsf{end}} M_1 C_{\mathsf{inv}} C_{g,1} M_4 \Big(2C_{g,0} + t_{\mathsf{end}} C_{g,1}(K + C_{\mathsf{inv}} C_{g,0})\Big)\tau^2 \leq C_{22}\tau^2 \tag{4.18}$$

mit einer Konstante C_{22}, die nur von M_1, M_4, $C_{g,0}$ $C_{g,1}$, C_{inv}, K und t_{end} abhängt.

Der Fehler lässt sich mittels der drei Schranken (4.13), (4.17) und (4.18) und der Lipschitz-Bedingung (1.4) wie folgt abschätzen:

$$\begin{aligned}\|u(t) - \widetilde{u}(t)\| &\leq \|I_1(t)\| + \|I_{21}(t)\| + \|I_{22}(t)\| + \int_0^t M_1 \left\| b(\widetilde{\Phi}u(s)) - b(\widetilde{\Phi}\widetilde{u}(s)) \right\| ds \\ &\leq (C_1 + C_{21} + C_{22})\tau^2 + M_1^2 C_{\mathsf{inv}} L_g \int_0^t \|u(s) - \widetilde{u}(s)\| \, ds\end{aligned}$$

Die Behauptung folgt dann mit dem Lemma von Gronwall B.2. □

Satz 4.2 ermöglicht es, im Konvergenzbeweis (siehe Abschnitt 4.1.2) direkt den Fehler zwischen numerischer Approximation u_n und der Lösung $\widetilde{u}$ der modifizierten Gleichung (3.6) zu betrachten. Mit dem folgenden Lemma wird nachgewiesen, dass auch für $\widetilde{u}$ eine vergleichbare Finite-Energie--Bedingung wie (3.3) gilt.

Lemma 4.3. *Es seien die Annahme 1.1 mit r_0, r_1 und r_2 aus* (4.7)*, die Annahme 3.1 und die Filterbedingung* (4.1b) *erfüllt. Dann erfüllt die Lösung $\widetilde{u}$ der modifizierten Gleichung* (3.6) *die Finite-Energie-Bedingung*

$$\|A\widetilde{u}(t)\| \leq \widetilde{K}, \qquad\qquad 0 \leq t \leq t_{end}, \tag{4.19}$$

mit einer Konstante $\widetilde{K}$, die nur von K, C_{inv}, $C_{g,0}$ L_g, C_{av}, M_1 und t_{end} abhängt.

Beweis. Aus der Variation-der-Konstanten-Formel (4.8b) folgt

$$\begin{aligned}\|A\widetilde{u}(t)\| &= \left\| Ae^{tA}u_0 + \int_0^t Ae^{(t-s)A}\widetilde{b}\big(\widetilde{u}(s)\big)\, ds \right\| \\ &\leq \|Au_0\| + \int_0^t \left\| A\widetilde{b}\big(\widetilde{u}(s)\big) \right\| ds \\ &\leq \|Au_0\| + \int_0^t \left\| A\widetilde{b}\big(u(s)\big) \right\| + \left\| A\Big(\widetilde{b}\big(\widetilde{u}(s)\big) - \widetilde{b}\big(u(s)\big)\Big) \right\| ds \\ &\leq K + \int_0^t M_1 C_{g,0} + M_1^2 L_g \, \|\widetilde{u}(s) - u(s)\| \, ds \\ &\leq K + M_1 t_{\mathsf{end}} \left(C_{g,0} + M_1 L_g C_{\mathsf{av}} \tau^2\right) \\ &\leq \widetilde{K}.\end{aligned}$$

Hierfür wurde die Aussage aus Satz 4.2

$$\|\widetilde{u}(s) - u(s)\| \leq C_{\mathsf{av}}\tau^2$$

zusammen mit $\tau < 1$, (1.3), (1.4) (3.3) und

$$\|\Phi u(s)\| \leq M_1 C_{\mathsf{inv}} K \leq r_0 \tag{4.20}$$

verwendet. □

Im nächsten Abschnitt wird der lokale Fehler analysiert und in eine geeignete Form gebracht, um die Abschätzung des globalen Fehlers zu erleichtern. Bevor man jedoch mit dem Beweis fortfahren kann, müssen die Konstanten r_0, r_1 und r_2 angepasst werden. Bis zu diesem Zeitpunkt wurden diese nach (4.7) gewählt. Es seien nun

$$\widetilde{r}_0 = \max(r_0, M_1 C_{\mathsf{inv}}\widetilde{K}), \qquad \widetilde{r}_1 = \max(r_1, 2M_1 C_{\mathsf{inv}}\widetilde{K}), \qquad \widetilde{r}_2 = \max(r_2, M_1 C_{\mathsf{inv}}\widetilde{K}). \tag{4.21}$$

und für die Funktion g und ihre Ableitungen gelte

$$\begin{aligned} \|g(q)\| &\leq \widetilde{C_{g,0}} & \|q\| &\leq \widetilde{r}_0, \\ \|g'(q)\| &\leq \widetilde{C_{g,1}} & \|q\| &\leq \widetilde{r}_1, \\ \|g''(q)\| &\leq \widetilde{C_{g,2}} & \|q\| &\leq \widetilde{r}_2. \end{aligned}$$

Die Radien $\widetilde{r}_0$, $\widetilde{r}_1$ und $\widetilde{r}_2$ hängen von $\widetilde{K}$ ab. Die Größe der Finiten-Energie der modifizierten Gleichung steht meist nicht vor der Simulation zur Verfügung. Sie kann wie oben gesehen geschätzt werden. Jedoch wird die Größe dabei vermutlich weit überschätzt. Eine andere Möglichkeit besteht darin, die Finite-Energie der numerischen Approximation zu jeden Zeitschritt auszuwerten um diese Größe abschätzen zu können.

Mithilfe dieser neuen Radien kann nun der Konvergenzbeweis durchgeführt werden. Sind die Funktion g und ihre Ableitungen wie in Annahme 1.1 mit den Radien aus (4.21) beschränkt, so gilt ebenfalls Satz 4.2, denn die Radien wurden vergrößert. Die Radien aus (4.7) dienten lediglich dazu die Größe $\widetilde{K}$ sinnvoll abzuschätzen.

4.1.2 Transformation des lokalen Fehlers

Das folgende Lemma analysiert den lokalen Fehler, also die Differenz zwischen der exakten Lösung

$$\varphi_{A+\widetilde{b}}(\tau, \widetilde{u}(t_n)) = e^{\tau A}\widetilde{u}(t_n) + \int_0^\tau e^{(\tau-s)A} b(\widetilde{u}(t_n + s))\, ds, \tag{4.22}$$

der modifizierten Differentialgleichung (3.6) und dem Splittingverfahren (3.10). Dieser umfasst Terme der Größenordnung $\mathcal{O}(\tau^3)$, für die der globale Fehler direkt mit Hilfe der Technik des *Lady Windermeres Fächer* abgeschätzt werden kann, zu Lady Windermeres Fächer siehe auch Abschnitt A.3. Allerdings treten auch Terme der Größenordnung $\mathcal{O}(\tau^2)$ auf, die bei der Analyse des globalen Fehlers gesondert behandelt werden müssen.

In diesem Kapitel werden Kommutatoren verwendet, um den Fehler zu analysieren:

Definition 4.4. *Für zwei Vektorfelder $F,\ G : \mathbb{R}^d \to \mathbb{R}^d$ sei der Kommutator $[F, G]$ definiert durch*

$$[F, G]\,(v) = F(G(v)) - G(F(v)).$$

Der Lie-Kommutator $[F, G]_L$ ist gegeben durch

$$[F, G]_L\,(v) = F'(v)G(v) - G'(v)F(v)$$

wobei F' und G' die jeweilige Jacobi-Matrix bezeichnen.

Der Lie-Kommutator wird häufig verwendet, um Ordnungsbedingungen von Splittingverfahren herzuleiten beziehungsweise Konvergenz nachzuweisen, siehe zum Beispiel Hairer et al. (2006), Kapitel III.5, Hundsdorfer und Verwer (2007), Kapitel IV.1.4 oder Lubich (2008). Der lokale Fehler des in dieser Arbeit betrachteten Splittingverfahrens lässt sich auch mittels der Lie-Ableitung ausdrücken.

Lemma 4.5. *(Darstellung des lokalen Fehlers)*
Es seien die Annahme 1.1 mit den Radien $\widetilde{r}_0$, $\widetilde{r}_1$ und $\widetilde{r}_2$ aus (4.21)*, die Annahmen 3.1 und 4.1 erfüllt. Dann ist der lokale Fehler zur Zeit $t_n = n\tau$, $0 \le t_n \le t_{end} - \tau$, des Splittingverfahrens* (3.10) *angewandt auf die modifizierte Gleichung* (3.6) *gegeben als*

$$\delta_n = S\,\widetilde{u}(t_n) - \varphi_{A+\widetilde{b}}\,(\tau, \widetilde{u}(t_n)) = \delta_n^{(1)} + \delta_n^{(2)} + D_n,$$

mit

$$\delta_n^{(1)} = -\frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A} A^2 \widetilde{b} \left(e^{\frac{\theta}{2}A} \varphi_{A+\widetilde{b}}\,(\tau - \xi, \widetilde{u}(t_n))) \right) d\theta\, d\sigma\, d\xi, \tag{4.23}$$

$$\delta_n^{(2)} = -\frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A} J_{\widetilde{b}} \left(e^{\frac{\theta}{2}A} \varphi_{A+\widetilde{b}}\,(\tau - \xi, \widetilde{u}(t_n))) \right) A^2 e^{\left(\frac{\theta}{2}+\tau-\xi\right)A} \widetilde{u}(t_n)\, d\theta\, d\sigma\, d\xi, \tag{4.24}$$

und $\|D_n\| \le C\tau^3$. Die Konstante C hängt dabei von C_{inv}, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$, $\widetilde{K}$, M_1 und M_2 ab.

Wie bereits in Kapitel 2 erklärt, beruhten bisherige Darstellungen des lokalen Fehlers auf der Variation-der-Konstanten-Formel. Im Gegensatz dazu verwendet die folgende Darstellung den Hauptsatz der Differential- und Integralrechung. Um die Idee zu illustieren, wird hier zunächst der lineare Fall betrachtet, vergleiche Buchholz et al. (2018) beziehungsweise die Notation aus Abschnitt 2.2.7. Sei dafür $g(u) = Gu$ für eine Matrix G mit kleiner Norm. Der lokale Fehler ist im linearen Fall ebenfalls als Differenz zwischen einem Schritt mit dem Splittingverfahren (2.53) und der exakten Lösung der modifizierten Gleichung (2.52) gegeben. Er kann mithilfe des Hauptsatzes der Differential- und Integralrechung geschrieben werden als

$$\begin{aligned}
\delta_n &= e^{\frac{\tau}{2}A} e^{\tau\widetilde{B}} e^{\frac{\tau}{2}A} \widetilde{u}(t_n) - e^{\tau(A+\widetilde{B})} \widetilde{u}(t_n) \\
&= \int_0^\tau \frac{d}{d\xi} \left(e^{\frac{\xi}{2}A} e^{\xi\widetilde{B}} e^{\frac{\xi}{2}A} e^{(\tau-\xi)(A+\widetilde{B})} \right) \widetilde{u}(t_n)\, d\xi \\
&= \frac{1}{2}\int_0^\tau e^{\frac{\xi}{2}A} \left(\left[A, e^{\xi\widetilde{B}}\right] e^{\frac{\xi}{2}A} + 2e^{\xi\widetilde{B}} \left[\widetilde{B}, e^{\frac{\xi}{2}A}\right] \right) e^{(\tau-\xi)(A+\widetilde{B})} \widetilde{u}(t_n)\, d\xi.
\end{aligned}$$

Im linearen Fall erkennt man direkt, dass die Kommutatoren $\left[A, e^{\xi\widetilde{B}}\right]$ und $\left[\widetilde{B}, e^{\frac{\xi}{2}A}\right]$ auftreten, ähnlich wie auch in der Darstellung des lokalen Fehlers bei der Analyse von Splittingverfahren für glatte

Lösungen (vergleiche mit Abschnitt 2.1.1). Wie in Buchholz et al. (2018) gezeigt, kann man die auftretenden Differenzen in Form der Kommutatoren wiederum mit Hilfe des Hauptsatzes der Differential- und Integralrechung darstellen. Nach dreimaliger Anwedung erhält man schlussendlich eine Darstellung mit Doppelkommutator, im linearen Fall ist das

$$\begin{aligned}\delta_n = \frac{1}{2}\int_0^\tau \int_\theta^\xi \int_0^\xi e^{\frac{\xi}{2}A} e^{\xi\widetilde{B}} e^{\frac{\xi-\sigma}{2}A} \left[\left[A,\widetilde{B}\right],A\right] e^{\frac{\sigma}{2}A} e^{(\tau-\xi)(A+\widetilde{B})}\widetilde{u}(t_n)\, d\sigma\, d\theta\, d\xi \\ + \frac{1}{2}\int_0^\tau \int_0^\theta \int_0^\xi e^{\frac{\xi}{2}A} e^{(\xi-\sigma)\widetilde{B}} \left[\left[A,\widetilde{B}\right],\widetilde{B}\right] e^{\sigma\widetilde{B}} e^{\frac{\xi}{2}A} e^{(\tau-\xi)(A+\widetilde{B})}\widetilde{u}(t_n)\; d\sigma\, d\theta\, d\xi.\end{aligned} \tag{4.25}$$

Die Untersuchung der Doppelkommutatoren ermöglichst es, zwischen Termen der Größenordnung $\mathcal{O}(\tau^2)$ und $\mathcal{O}(\tau^3)$ zu unterscheiden, die im globalen Fehler getrennt behandelt werden müssen. Hier ist beispielsweise

$$\left[\left[A,\widetilde{B}\right],A\right] = -A^2\widetilde{B} - \widetilde{B}A^2 + 2A\widetilde{B}A.$$

An der Struktur von $\widetilde{B}$ erkennt man, dass $A\widetilde{B}$ beschränkt ist, vergleiche mit (2.50). Ein Auftreten von A in Kombination mit $\widetilde{u}(t_n)$ kann durch die Finite-Energie-Bedingung beschränkt werden. Deswegen lässt sich der Anteil $2A\widetilde{B}A$ im Integral proportional zu τ^3 beschränken. Für die beiden ersten Terme ist dies nicht möglich. Es lässt sich lediglich eine Schranke der Ordnung $\mathcal{O}(\tau^2)$ herleiten, wenn man die Filterbedingung (4.1c) berücksichtigt.

Ein vergleichbares Vorgehen wird nun auch im semilinearen Fall angestrebt. Die Differenzen werden mithilfe des Hauptsatzes der Differential- und Integralrechung dargestellt und später die Größe der vermutlich auftauchenden Terme des Kommutators untersucht. Durch die Nichtlinearität g treten nun zusätzlich die in Definition 4.4 vorgestellten Lie-Kommutatoren auf. Mit Hilfe der Lie-Kommutatoren ist es möglich, den lokalen Fehler in eine mit (4.25) vergleichbare Form zu bringen (vergleiche (4.28)). Dies ermöglicht eine präzise Aufspaltung des lokalen Fehlers in Terme $\delta_n^{(1)}$ und $\delta_n^{(2)}$ der Größenordnung $\mathcal{O}(\tau^2)$ und den Term D_n der Größe $\mathcal{O}(\tau^3)$. Die Beweistechnik aus Buchholz et al. (2018) kann also adaptiert werden.

Beweis von Lemma 4.5. Aufgrund der einfachen Struktur des Flusses $\varphi_{\widetilde{b}}$ (vergleiche (3.9)) ist

$$\partial_1\varphi_{\widetilde{b}}(t,u_0) = \frac{\partial}{\partial t}\varphi_{\widetilde{b}}(t,u_0) = b(u_0).$$

Weiterhin sei die Abkürzung $v(\sigma,\xi) = e^{\frac{\sigma}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n)))$ definiert. Mithilfe des Hauptsatzes der Differential- und Integralrechung kann der lokale Fehler folglich geschrieben werden als

$$\begin{aligned}\delta_n &= S\big(\widetilde{u}(t_n)\big) - \varphi_{A+\widetilde{b}}(\tau,\widetilde{u}(t_n)) \\ &= e^{\frac{\tau}{2}A}\varphi_{\widetilde{b}}\left(\tau, e^{\frac{\tau}{2}A}\widetilde{u}(t_n)\right) - \varphi_{A+\widetilde{b}}(\tau,\widetilde{u}(t_n)) \\ &= \int_0^\tau \frac{d}{d\xi}\left(e^{\frac{\xi}{2}A}\varphi_{\widetilde{b}}(\xi,v(\xi,\xi))\right)\, d\xi \\ &= \int_0^\tau e^{\frac{\xi}{2}A}(\tfrac{1}{2}A)\varphi_{\widetilde{b}}(\xi,v(\xi,\xi))\; d\xi + \int_0^\tau e^{\frac{\xi}{2}A}\partial_1\varphi_{\widetilde{b}}(\xi,v(\xi,\xi))\; d\xi \\ &\quad + \int_0^\tau e^{\frac{\xi}{2}A}\partial_2\varphi_{\widetilde{b}}(\xi,v(\xi,\xi))\, e^{\frac{\xi}{2}A}\left(\tfrac{1}{2}Av(0,\xi) - Av(0,\xi) - \widetilde{b}(v(0,\xi))\right)\, d\xi.\end{aligned}$$

Umsortieren der Terme ergibt

$$\begin{aligned}\delta_n = \frac{1}{2}\int_0^\tau e^{\frac{\xi}{2}A}\left(A\varphi_{\widetilde{b}}(\xi, v(\xi,\xi)) - \partial_2\varphi_{\widetilde{b}}(\xi, v(\xi,\xi))\, Av(\xi,\xi)\right)\, d\xi \\ + \int_0^\tau e^{\frac{\xi}{2}A}\left(\widetilde{b}(v(\xi,\xi)) - \partial_2\varphi_{\widetilde{b}}(\xi, v(\xi,\xi))\, e^{\frac{\xi}{2}A}\widetilde{b}(v(0,\xi))\right)\, d\xi.\end{aligned}$$

Verwendet man nun die Darstellung (4.3) in Form von $\partial_2\varphi_{\widetilde{b}}(t,w) = I + tJ_{\widetilde{b}}(w)$ für das zweite Integral und die Definition 4.4, so erhält man

$$\begin{aligned}\delta_n = \frac{1}{2}\int_0^\tau e^{\frac{\xi}{2}A}\left[A, \varphi_{\widetilde{b}}(\xi,\cdot)\right]_L (v(\xi,\xi))\, d\xi + \int_0^\tau e^{\frac{\xi}{2}A}\left[\widetilde{b}, e^{\frac{\xi}{2}A}\right](v(0,\xi))\, d\xi \\ - \int_0^\tau \xi e^{\frac{\xi}{2}A} J_{\widetilde{b}}(v(\xi,\xi))\, e^{\frac{\xi}{2}A}\widetilde{b}(v(0,\xi))\, d\xi.\end{aligned}$$

Hierzu sei vermerkt, dass Matrizen im (Lie-)Kommutator stets für ihre linearen Abbildungen stehen. Bezeichnet man beispielsweise mit F_A die lineare Abbildung $F_A : x \to Ax$, so entspricht

$$\left[A, \varphi_{\widetilde{b}}(\xi,\cdot)\right]_L = \left[F_A, \varphi_{\widetilde{b}}(\xi,\cdot)\right]_L.$$

Der Lie-Kommutator im ersten Integral lässt sich mit Hilfe der Definition des Flusses $\varphi_{\widetilde{b}}$ aus (3.9) und der Ableitung $\partial_2\varphi_{\widetilde{b}}(\tau,u)$ in (4.3) und $v = v(\xi,\xi)$ vereinfachen zu

$$\begin{aligned}\left[A, \varphi_{\widetilde{b}}(\xi,\cdot)\right]_L (v) &= A\varphi_{\widetilde{b}}(\xi,v) - \partial_2\varphi_{\widetilde{b}}(\xi,v)\, Av \\ &= A\big(v + \xi\widetilde{b}(v)\big) - \big(I + \xi J_{\widetilde{b}}(v)\big)Av \\ &= \xi\big(A\widetilde{b}(v) - J_{\widetilde{b}}(v)Av\big) \\ &= \xi\left[A, \widetilde{b}\right]_L (v). \end{aligned} \tag{4.26}$$

Die Differenz im Kommutator $\left[\widetilde{b}, e^{\frac{\xi}{2}A}\right]$ im zweiten Integral lässt sich wieder mit Hilfe des Hauptsatzes der Differential- und Integralrechung darstellen:

$$\begin{aligned}\left[\widetilde{b}, e^{\frac{\xi}{2}A}\right](v(0,\xi)) &= \widetilde{b}(v(\xi,\xi)) - e^{\frac{\xi}{2}A}\widetilde{b}(v(0,\xi)) \\ &= \int_0^\xi \frac{\partial}{\partial\sigma}\left(e^{\frac{\xi-\sigma}{2}A}\widetilde{b}\,(v(\sigma,\xi))\right)\, d\sigma \\ &= \int_0^\xi e^{\frac{\xi-\sigma}{2}A}\left(-\frac{1}{2}A\widetilde{b}\,(v(\sigma,\xi)) + J_{\widetilde{b}}\,(v(\sigma,\xi))\,\frac{1}{2}Av(\sigma,\xi)\right)\, d\sigma \\ &= -\frac{1}{2}\int_0^\xi e^{\frac{\xi-\sigma}{2}A}\left[A, \widetilde{b}\right]_L (v(\sigma,\xi))\, d\sigma, \end{aligned} \tag{4.27}$$

wobei die Ableitung $\frac{\partial}{\partial\sigma}v(\sigma,\xi) = \frac{\partial}{\partial\sigma}\left(e^{\frac{\sigma}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi, \widetilde{u}(t_n))\right) = \frac{1}{2}Av(\sigma,\xi)$ entspricht. Setzt man nun die Darstellungen (4.26) und (4.27) in die Darstellung des lokalen Fehlers ein, ergibt sich

$$\begin{aligned}\delta_n &= \frac{1}{2}\int_0^\tau \xi e^{\frac{\xi}{2}A}\left[A, \widetilde{b}\right]_L (v(\xi,\xi))\, d\xi - \frac{1}{2}\int_0^\tau\int_0^\xi e^{\frac{\xi}{2}A}e^{\frac{\xi-\sigma}{2}A}\left[A, \widetilde{b}\right]_L (v(\sigma,\xi))\, d\sigma\, d\xi \\ &\quad - \int_0^\tau \xi e^{\frac{\xi}{2}A} J_{\widetilde{b}}(v(\xi,\xi))\, e^{\frac{\xi}{2}A}\widetilde{b}(v(0,\xi))\, d\xi \\ &= \frac{1}{2}\int_0^\tau\int_0^\xi e^{\frac{\xi}{2}A}\left(\left[A, \widetilde{b}\right]_L (v(\xi,\xi)) - e^{\frac{\xi-\sigma}{2}A}\left[A, \widetilde{b}\right]_L (v(\sigma,\xi))\right)\, d\sigma\, d\xi \\ &\quad - \int_0^\tau \xi e^{\frac{\xi}{2}A} J_{\widetilde{b}}(v(\xi,\xi))\, e^{\frac{\xi}{2}A}\widetilde{b}(v(0,\xi))\, d\xi.\end{aligned}$$

Die Differenz im ersten Integral wird wie oben (vgl (4.27)) behandelt:

$$\begin{aligned}
&\left[A,\widetilde{b}\right]_L\big(v(\xi,\xi)\big) - e^{\frac{\xi-\sigma}{2}A}\left[A,\widetilde{b}\right]_L\big(v(\sigma,\xi)\big)\\
&\quad= \int_\sigma^\xi \frac{\partial}{\partial\theta}\left(e^{\frac{\xi-\theta}{2}A}\left[A,\widetilde{b}\right]_L\big(v(\theta,\xi)\big)\right)\,d\theta\\
&\quad= \int_\sigma^\xi e^{\frac{\xi-\theta}{2}A}\left(-\frac{1}{2}A\left[A,\widetilde{b}\right]_L\big(v(\theta,\xi)\big) + \left(\frac{d}{dv}\left[A,\widetilde{b}\right]_L\big(v(\theta,\xi)\big)\right)\frac{1}{2}Av(\theta,\xi)\right)\,d\theta\\
&\quad= \frac{1}{2}\int_\sigma^\xi e^{\frac{\xi-\theta}{2}A}\left[\left[A,\widetilde{b}\right]_L,A\right]_L\big(v(\theta,\xi)\big)\,d\theta.
\end{aligned}$$

Insgesamt erhält man

$$\begin{aligned}
\delta_n = &\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{\frac{\xi}{2}A}e^{\frac{\xi-\theta}{2}A}\left[\left[A,\widetilde{b}\right]_L,A\right]_L\big(v(\theta,\xi)\big)\,d\theta\,d\sigma\,d\xi\\
&-\int_0^\tau \xi e^{\frac{\xi}{2}A}J_{\widetilde{b}}\big(v(\xi,\xi)\big)e^{\frac{\xi}{2}A}\widetilde{b}\,(v(0,\xi))\,d\xi.
\end{aligned}\tag{4.28}$$

Die Darstellung ähnelt derjenigen des lokalen Fehlers aus (4.25), hat also die gleiche Form wie in Gleichung (5.5) in Buchholz et al. (2018) bei geänderter Reihenfolge des Splittings. Im folgenden Abschnitt wird der iterierte Lie-Kommutator untersucht, um den lokalen Fehler in Terme der Größenordnung $\mathcal{O}(\tau^2)$ und $\mathcal{O}(\tau^3)$ zu zerlegen. Es ist

$$\begin{aligned}
\frac{d}{dv}\left[A,\widetilde{b}\right]_L(v)w &= \frac{d}{dv}\big(A\widetilde{b}(v) - J_{\widetilde{b}}(v)Av\big)w\\
&= AJ_{\widetilde{b}}(v)w - H_{\widetilde{b}}(v)(Av,w) - J_{\widetilde{b}}(v)Aw
\end{aligned}$$

und damit gilt

$$\left[\left[A,\widetilde{b}\right]_L,A\right]_L(v) = 2AJ_{\widetilde{b}}(v)Av - H_{\widetilde{b}}(v)(Av,Av) - J_{\widetilde{b}}(v)A^2v - A^2\widetilde{b}(v).$$

Der lokale Fehler ist somit gegeben durch

$$\delta_n = \delta_n^{(1)} + \delta_n^{(2)} + D_n^{(1)} + D_n^{(2)} + D_n^{(3)} + D_n^{(4)},$$

mit

$$\begin{aligned}
\delta_n^{(1)} &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}A^2\widetilde{b}\,(v(\theta,\xi))\,d\theta\,d\sigma\,d\xi,\\
\delta_n^{(2)} &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}J_{\widetilde{b}}\,(v(\theta,\xi))\,A^2e^{\frac{\theta}{2}A}e^{(\tau-\xi)A}\widetilde{u}(t_n)\,d\theta\,d\sigma\,d\xi,\\
D_n^{(1)} &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}J_{\widetilde{b}}\,(v(\theta,\xi))\,A^2e^{\frac{\theta}{2}A}\int_0^{\tau-\xi}e^{(\tau-\xi-\nu)A}\widetilde{b}(\widetilde{u}(t_n+\nu))\,d\nu\,d\theta\,d\sigma\,d\xi,\\
D_n^{(2)} &= \frac{1}{2}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}AJ_{\widetilde{b}}\,(v(\theta,\xi))\,Av(\theta,\xi)\,d\theta\,d\sigma\,d\xi,\\
D_n^{(3)} &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}H_{\widetilde{b}}\,(v(\theta,\xi))\,(Av(\theta,\xi),Av(\theta,\xi))\,d\theta\,d\sigma\,d\xi,\\
D_n^{(4)} &= -\int_0^\tau \xi e^{\frac{\xi}{2}A}J_{\widetilde{b}}\,(v(\xi,\xi))\,e^{\frac{\xi}{2}A}\widetilde{b}\,(v(0,\xi))\,d\xi,
\end{aligned}$$

wobei $\delta_n^{(2)} + D_n^{(1)}$ mittels der Variation-der-Konstanten-Formel (4.8b) angewandt auf

$$v(\theta,\xi) = e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))) = e^{\frac{\theta}{2}A}\left(e^{(\tau-\xi)A}\widetilde{u}(t_n) + \int_0^{\tau-\xi} e^{(\tau-\xi-\nu)A}\widetilde{b}(\widetilde{u}(t_n+\nu))\,d\nu\right)$$

entstanden sind. Mithilfe der Annahmen 1.1 und 4.1 lässt sich zeigen, dass die Terme $\delta_n^{(1)}$ und $\delta_n^{(2)}$ von der Größenordnung $\mathcal{O}(\tau^2)$ sind, vergleiche Bemerkung 4.6. Sie müssen deshalb bei der Analyse des globalen Fehlers separat betrachtet werden. Die Terme $D_n^{(i)}$ für $i = 1,\dots,4$ sind von der Größenordnung $\mathcal{O}(\tau^3)$, wie nun nachgewiesen wird:

Beginnend lässt sich $D_n^{(1)}$ schreiben als

$$D_n^{(1)} = -\frac{1}{4\tau}\int_0^{\tau}\int_0^{\xi}\int_{\sigma}^{\xi}\int_0^{\tau-\xi} e^{(\xi-\frac{\theta}{2})A}J_{\widetilde{b}}\left(v(\theta,\xi)\right)(\tau A)e^{(\frac{\theta}{2}+\tau-\xi-\nu)A}A\widetilde{b}(\widetilde{u}(t_n+\nu))\,d\nu\,d\theta\,d\sigma\,d\xi.$$

Mit Hilfe der Abschätzungen (1.3) und den Radien aus (4.21), (3.5), (4.1b), (4.1c), (4.20) und

$$\|v(\theta,\xi)\| = \left\|\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right\| = \|\widetilde{u}(t_n+\tau-\xi)\| = \left\|A^{-1}A\widetilde{u}(t_n+\tau-\xi)\right\| \le C_{\mathsf{inv}}\widetilde{K}, \quad (4.29)$$

für alle $\xi > 0$ mit $t_n + \tau - \xi \le t_{\mathsf{end}}$. Damit folgt mit $C_{\mathsf{inv}}\widetilde{K} \le \widetilde{r}_0$ zunächst

$$\left\|J_{\widetilde{b}}(v(\theta,\xi))(\tau A)\right\| = \left\|\begin{bmatrix} 0 & 0 \\ 0 & \Omega^{-1}\Psi_S g'(\Phi q_v)\Phi(\tau\Omega) \end{bmatrix}\right\| \le C_{\mathsf{inv}}M_1\widetilde{C_{g,1}}M_2.$$

Wegen $\|\widetilde{u}(t_n+\nu)\| \le C_{\mathsf{inv}}\widetilde{K} \le \widetilde{r}_0$ folgt außerdem

$$\left\|A\widetilde{b}(\widetilde{u}(t_n+\nu))\right\| = \left\|\begin{bmatrix} \Psi_S g(\Phi q_u) \\ 0 \end{bmatrix}\right\| \le M_1\widetilde{C_{g,0}}.$$

Insgesamt ergibt sich

$$\left\|D_n^{(1)}\right\| \le \frac{1}{4\tau}\frac{\tau^4}{24}M_1C_{\mathsf{inv}}\widetilde{C_{g,1}}M_2M_1\widetilde{C_{g,0}} = \frac{1}{96}M_1^2M_2C_{\mathsf{inv}}\widetilde{C_{g,1}}\widetilde{C_{g,0}}\tau^3.$$

Weiter folgen mit (1.3) und den Radien aus (4.21), (3.1) (3.5), (4.1b), (4.19), (4.29) und

$$\left\|AJ_{\widetilde{b}}(v(\theta,\xi))\right\| = \left\|\begin{bmatrix} \Psi_S g'(\Phi q_v)\Phi & 0 \\ 0 & 0 \end{bmatrix}\right\| \le M_1^2\widetilde{C_{g,1}}, \quad (4.30)$$

$$\left\|H_{\widetilde{b}}(v(\theta,\xi))(y,z)\right\| = \left\|\begin{bmatrix} 0 \\ \Psi_s\Omega^{-1}g''(\Phi q_v)(\Phi q_y,\Phi q_z) \end{bmatrix}\right\| \le C_{\mathsf{inv}}M_1^3\widetilde{C_{g,2}}\,\|y\|\,\|z\|, \quad (4.31)$$

für $x = [q_x^T, v_x^T]^T$ mit $x = v, y, z$ die folgenden Schranken für $D_n^{(2)}$ und $D_n^{(3)}$:

$$\left\|D_n^{(2)}\right\| \le \frac{1}{2}\frac{\tau^3}{6}M_1^2\widetilde{C_{g,1}}\widetilde{K} = \frac{1}{12}M_1^2\widetilde{K}\widetilde{C_{g,1}}\tau^3,$$
$$\left\|D_n^{(3)}\right\| \le \frac{1}{4}\frac{\tau^3}{6}C_{\mathsf{inv}}M_1^3\widetilde{C_{g,2}}\widetilde{K}\widetilde{K} = \frac{1}{24}C_{\mathsf{inv}}M_1^3\widetilde{K}^2\widetilde{C_{g,2}}\tau^3.$$

Nach Definition ist $J_{\widetilde{b}}(u)\widetilde{b}(w) = 0$ unabhängig von u und w. Fügt man diesen Term zusätzlich in $D_n^{(4)}$ ein, so erhält man

$$\begin{aligned} D_n^{(4)} &= \int_0^\tau \xi e^{\frac{\xi}{2}A} J_{\widetilde{b}}\left(v(\xi,\xi)\right)\left(\widetilde{b}\left(v(\xi,\xi)\right) - e^{\frac{\xi}{2}A}\widetilde{b}\left(v(0,\xi)\right)\right) d\xi \\ &= \int_0^\tau \xi e^{\frac{\xi}{2}A} J_{\widetilde{b}}\left(v(\xi,\xi)\right)\left[\widetilde{b}, e^{\frac{\xi}{2}A}\right]\left(v(0,\xi)\right) d\xi. \end{aligned}$$

Der Kommutator $\left[\widetilde{b}, e^{\frac{\xi}{2}A}\right]\left(v(0,\xi)\right)$ lässt sich wie in (4.27) durch ein Integral darstellen. Es folgt

$$D_n^{(4)} = -\frac{1}{2}\int_0^\tau \int_0^\xi \xi e^{\frac{\xi}{2}A} J_{\widetilde{b}}\left(v(\xi,\xi)\right) e^{\frac{\xi-\theta}{2}A}\left[A,\widetilde{b}\right]_L\left(v(\theta,\xi)\right) d\theta\, d\xi.$$

Wegen (1.3),(3.5), (4.1b), (4.29) folgt zunächst

$$\|v(\xi,\xi)\| \le C_{\mathsf{inv}}\widetilde{K} \le \widetilde{r}_1, \qquad \text{und} \qquad \|v(\theta,\xi)\| \le C_{\mathsf{inv}}\widetilde{K} \le \min(\widetilde{r}_0,\widetilde{r}_1),$$

und daraus

$$\left\|J_{\widetilde{b}}(v(\xi,\xi))\right\| = \left\|\begin{bmatrix} 0 & 0 \\ \Omega^{-1}\Psi_S g'(\Phi q_v)\Phi & 0 \end{bmatrix}\right\| \le C_{\mathsf{inv}} M_1^2 \widetilde{C_{g,1}}, \tag{4.32}$$

$$\left\|\left[A,\widetilde{b}\right]_L\left(v(\theta,\xi)\right)\right\| = \left\|A\widetilde{b}(v(\theta,\xi)) - J_{\widetilde{b}}(v(\theta,\xi))Au\right\| \le M_1\widetilde{C_{g,0}} + C_{\mathsf{inv}}M_1^2\widetilde{C_{g,1}}\widetilde{K}. \tag{4.33}$$

Damit lässt sich $D_n^{(4)}$ beschränken durch

$$\begin{aligned} \left\|D_n^{(4)}\right\| &\le \frac{1}{2}\frac{\tau^3}{3}M_1^2 C_{\mathsf{inv}}\widetilde{C_{g,1}}\left(M_1\widetilde{C_{g,0}} + C_{\mathsf{inv}}M_1^2\widetilde{C_{g,1}}1\widetilde{K}\right) \\ &= \frac{1}{6}M_1^3 C_{\mathsf{inv}}\widetilde{C_{g,1}}\left(\widetilde{C_{g,0}} + C_{\mathsf{inv}}M_1\widetilde{K}\widetilde{C_{g,1}}\right)\tau^3. \end{aligned}$$

Mit $D_n = D_n^{(1)} + D_n^{(2)} + D_n^{(3)} + D_n^{(4)}$ ist die Aussage des Lemmas bewiesen. □

Bemerkung 4.6. *Die Terme $\delta_n^{(1)}$ und $\delta_n^{(2)}$ sind proportional zu τ^2 beschränkt. Um dies zu zeigen, schreibt man*

$$\delta_n^{(1)} = -\frac{1}{4\tau}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}(\tau A\widetilde{\Psi}_S)Ab\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right) d\theta\, d\sigma\, d\xi,$$

$$\delta_n^{(2)} = -\frac{1}{4\tau}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}\widetilde{\Psi}_S J_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right)(\widetilde{\Phi}\tau A)e^{\left(\frac{\theta}{2}+\tau-\xi\right)A}A\widetilde{u}(t_n)\, d\theta\, d\sigma\, d\xi.$$

Da $\left\|\Phi e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right\| \le M_1 C_{inv}\widetilde{K} \le \min(\widetilde{r}_0,\widetilde{r}_1)$ mit den Radien $\widetilde{r}_0$ und $\widetilde{r}_1$ aus (4.21) *ist, folgt mit* (4.1c), (4.11)

$$\left\|\delta_n^{(1)}\right\| \le \frac{1}{4\tau}\frac{\tau^3}{6}M_2\widetilde{C_{g,0}} = \frac{1}{24}M_2\widetilde{C_{g,0}}\tau^2,$$

$$\left\|\delta_n^{(2)}\right\| \le \frac{1}{4\tau}\frac{\tau^3}{6}M_1\widetilde{C_{g,1}}M_2\widetilde{K} = \frac{1}{24}M_1M_2\widetilde{C_{g,1}}\widetilde{K}\tau^2.$$

Mit Lemma 4.5 wurde der lokale Fehler in die Teile $\delta_n^{(1)}$ und $\delta_n^{(2)}$ der Größenordnung $\mathcal{O}(\tau^2)$ und einen Teil D_n der Größenordnung $\mathcal{O}(\tau^3)$ aufgeteilt. Damit das vorgestellte Splittingverfahren (3.10) konvergent der Ordnung zwei ist, muss bei der Abschätzung des globalen Fehlers nun beachtet werden, dass sich die Terme $\delta_n^{(1)}$ und $\delta_n^{(2)}$ nicht aufsummieren. Um die Notation kurz zu halten, führt das folgende Lemma eine verkürzte Darstellung der Terme $\delta_n^{(1)}$ und $\delta_n^{(2)}$ ein und liefert obere Schranken, die für den Beweis der Beschränktheit des globalen Fehlers benötigt werden.

Lemma 4.7. *Unter den Voraussetzungen von Lemma 4.5 genügt der dominante Anteil des lokalen Fehlers δ_n definiert in* (4.23) *and* (4.24) *der Darstellung*

$$\begin{aligned}\delta_n^{(1)} &= A\widetilde{\Psi}_S z_1(\widetilde{u}(t_n)),\\ \delta_n^{(2)} &= Z_2(\widetilde{u}(t_n))\widetilde{\Phi}A^2\widetilde{u}(t_n),\end{aligned}$$

mit

$$\begin{aligned}\|z_1(\widetilde{u}(t_n))\| &\le C_1\tau^3, & \|Z_2(\widetilde{u}(t_n))\| &\le C_3\tau^3,\\ \left\|\frac{1}{\tau}\left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))\right)\right\| &\le C_2\tau^3, & \|AZ_2(\widetilde{u}(t_n))\| &\le C_4\tau^3,\\ & & \left\|\frac{1}{\tau}\left(Z_2(\widetilde{u}(t_{j-1})) - Z_2(\widetilde{u}(t_j))\right)\right\| &\le C_5\tau^3,\end{aligned}$$

wobei die Konstanten C_i für $i = 1,\dots,5$ nur von C_{inv}, $\widetilde{K}$, M_1, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$ abhängig sind.

Beweis. Aus Lemma 4.5, erhält man

$$\begin{aligned}\delta_n^{(1)} &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}A^2\widetilde{b}\left(e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right)\,d\theta\,d\sigma\,d\xi,\\ \delta_n^{(2)} &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}J_{\widetilde{b}}\left(e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right)A^2e^{(\frac{\theta}{2}+\tau-\xi)A}\widetilde{u}(t_n)\,d\theta\,d\sigma\,d\xi.\end{aligned}$$

Verwendet man nun die Darstellung mittels der Blockmatrizen $\widetilde{\Phi}$ und $\widetilde{\Psi}$ aus (3.8), so erhält man

$$A^2\widetilde{b}(u) = (A\widetilde{\Psi}_S)Ab(\widetilde{\Phi}u) \qquad \text{und} \qquad J_{\widetilde{b}}(u)A^2 = \widetilde{\Psi}_S J_b(\widetilde{\Phi}u)\widetilde{\Phi}A^2.$$

Setzt man diese ein, so ergibt sich

$$\delta_n^{(1)} = A\widetilde{\Psi}_S z_1(\widetilde{u}(t_n)) \qquad \text{und} \qquad \delta_n^{(2)} = Z_2(\widetilde{u}(t_n))\widetilde{\Phi}A^2\widetilde{u}(t_n),$$

mit

$$\begin{aligned}z_1(\widetilde{u}(t_n)) &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}Ab\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right)\,d\theta\,d\sigma\,d\xi,\\ Z_2(\widetilde{u}(t_n)) &= -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A}\widetilde{\Psi}_S J_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right)e^{(\frac{\theta}{2}+\tau-\xi)A}\,d\theta\,d\sigma\,d\xi.\end{aligned}$$

Man beachte, dass z_1 ein Vektor und Z_2 eine Matrix ist.

Nachfolgend werden die im Lemma 4.7 geforderten oberen Schranken an z_1 und Z_2 gezeigt. Beginnend mit z_1 zeigt man mit Hilfe von (1.3) mit den Radien aus (4.21), (3.5), (4.1b), (4.11), (4.29) und

$$\left\|\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_n))\right\| \le M_1C_{\mathsf{inv}}\widetilde{K} \le \widetilde{r}_0,$$

die obere Schranke

$$\|z_1(\widetilde{u}(t_n))\| \leq \frac{1}{4}\frac{\tau^3}{6}\widetilde{C_{g,0}} = \frac{1}{24}\widetilde{C_{g,0}}\tau^3.$$

Als Nächstes soll die Differenz

$$\begin{aligned}\frac{1}{\tau}\left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))\right) = -\frac{1}{4\tau}\int_0^\tau\int_0^\xi\int_\sigma^\xi e^{\frac{\xi}{2}A}e^{\frac{\xi-\theta}{2}A}A\Big(b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_{j+1}))\right)\\ - b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\varphi_{A+\widetilde{b}}(\tau-\xi,\widetilde{u}(t_j))\right)\Big)\,d\theta\,d\sigma\,d\xi\end{aligned}$$

nach oben abgeschätzt werden. Mit Hilfe des Hauptsatzes der Differential- und Integralrechung folgt

$$\begin{aligned}&b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\widetilde{u}(t_{j+1}+\tau-\xi)\right) - b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\widetilde{u}(t_j+\tau-\xi)\right)\\ &\quad= \int_0^1 J_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\big(s\,\widetilde{u}(t_{j+1}+\tau-\xi) + (1-s)\widetilde{u}(t_j+\tau-\xi)\big)\right)\\ &\qquad\qquad\cdot\widetilde{\Phi}e^{\frac{\theta}{2}A}\big(\widetilde{u}(t_{j+1}+\tau-\xi) - \widetilde{u}(t_j+\tau-\xi)\big)\,ds.\end{aligned}$$

Die Variation-der-Konstanten-Formel angewandt auf die Differenz $\widetilde{u}(t_{j+1}+\tau-\xi) - \widetilde{u}(t_j+\tau-\xi)$ liefert insgesamt

$$\frac{1}{\tau}\left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))\right) = z_1^{(1)} + z_1^{(2)},$$

mit

$$\begin{aligned}z_1^{(1)} = -\frac{1}{4}\int_0^\tau\int_0^\xi\int_\sigma^\xi\int_0^1 e^{\left(\xi-\frac{\theta}{2}\right)A}AJ_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\big(s\,\widetilde{u}(t_{j+1}+\tau-\xi) + (1-s)\widetilde{u}(t_j+\tau-\xi)\big)\right)\\ \cdot\widetilde{\Phi}e^{\frac{\theta}{2}A}e^{(\tau-\xi)A}\frac{1}{\tau}\big(\widetilde{u}(t_{j+1}) - \widetilde{u}(t_j)\big)\,ds\,d\theta\,d\sigma\,d\xi,\\ z_1^{(2)} = -\frac{1}{4\tau}\int_0^\tau\int_0^\xi\int_\sigma^\xi\int_0^1 e^{\left(\xi-\frac{\theta}{2}\right)A}AJ_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}\big(s\,\widetilde{u}(t_{j+1}+\tau-\xi) + (1-s)\widetilde{u}(t_j+\tau-\xi)\big)\right)\\ \cdot\widetilde{\Phi}e^{\frac{\theta}{2}A}\int_0^{(\tau-\xi)} e^{(\tau-\xi-\nu)A}\big(\widetilde{b}(\widetilde{u}(t_{j+1}+\nu)) - \widetilde{b}(\widetilde{u}(t_j+\nu))\big)\,d\nu\,ds\,d\theta\,d\sigma\,d\xi.\end{aligned}$$

Das zweite Integral $z_1^{(2)}$ lässt sich mit (1.3), (4.1b), (3.5), (4.12) und (4.29) direkt abschätzen. Es gilt

$$\left\|z_1^{(2)}\right\| \leq \frac{1}{4\tau}\frac{\tau^4}{24}\widetilde{C_{g,1}}M_1 2C_{\mathsf{inv}}M_1\widetilde{C_{g,0}} \leq \frac{1}{48}M_1^2C_{\mathsf{inv}}\widetilde{C_{g,1}}\widetilde{C_{g,0}}\tau^3.$$

Um die Norm von $z_1^{(1)}$ nach oben zu beschränken, betrachten wir zunächst die Differenz

$$\begin{aligned}\frac{1}{\tau}\left(\widetilde{u}(t_{j+1}) - \widetilde{u}(t_j)\right) &= \frac{1}{\tau}(e^{\tau A} - I)\widetilde{u}(t_j) + \frac{1}{\tau}\int_0^\tau e^{(\tau-s)A}\widetilde{b}(\widetilde{u}(t_j+s))\,ds\\ &= (e^{\tau A} - I)(\tau A)^{-1}A\widetilde{u}(t_j) + \frac{1}{\tau}\int_0^\tau e^{(\tau-s)A}\widetilde{b}(\widetilde{u}(t_j+s))\,ds,\end{aligned}$$

wobei wir wieder die Variation-der-Konstanten-Formel (4.8b) verwendet haben. Da $\|(e^{\tau A} - I)(\tau A)^{-1}\| = \|\int_0^1 e^{\sigma\tau A}\, d\sigma\| \leq 1$ gilt, folgt mit (1.3) und den Radien aus (4.21), (3.5), (4.29), (4.1b) und (4.19)

$$\left\| \frac{1}{\tau} \left(\widetilde{u}(t_{j+1}) - \widetilde{u}(t_j) \right) \right\| \leq \widetilde{K} + C_{\text{inv}} M_1 \widetilde{C_{g,0}}. \tag{4.34}$$

Nun kann man $z_1^{(1)}$ mithilfe von (3.5), (4.1b), (4.12) und (4.29) nach oben beschränken durch

$$\left\| z_1^{(1)} \right\| \leq \frac{1}{4} \widetilde{C_{g,1}} M_1 \big(\widetilde{K} + C_{\text{inv}} M_1 \widetilde{C_{g,0}} \big) \frac{\tau^3}{6} = \frac{1}{24} M_1 \widetilde{C_{g,1}} \big(\widetilde{K} + C_{\text{inv}} M_1 \widetilde{C_{g,0}} \big) \tau^3.$$

Insgesamt folgt für z_1:

$$\begin{aligned} \left\| \frac{1}{\tau} \left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j)) \right) \right\| &\leq \left\| z_1^{(1)} \right\| + \left\| z_1^{(2)} \right\| \\ &\leq \frac{1}{24} M_1 \widetilde{C_{g,1}} \big(\frac{1}{2} M_1 C_{\text{inv}} \widetilde{C_{g,0}} + \widetilde{K} + C_{\text{inv}} M_1 \widetilde{C_{g,0}} \big) \tau^3. \end{aligned}$$

Nun folgen die Abschätzungen für Z_2. Zunächst lassen sich Z_2 und AZ_2 mit Hilfe von (4.1b), (3.5), (4.12), (4.15) und (4.29), nach oben durch

$$\begin{aligned} \|Z_2(\widetilde{u}(t_n))\| &\leq \frac{1}{4} M_1 C_{\text{inv}} \widetilde{C_{g,1}} \frac{\tau^3}{6} = \frac{1}{24} M_1 C_{\text{inv}} \widetilde{C_{g,1}} \tau^3, \\ \|AZ_2(\widetilde{u}(t_n))\| &\leq \frac{1}{4} M_1 \widetilde{C_{g,1}} \frac{\tau^3}{6} = \frac{1}{24} M_1 \widetilde{C_{g,1}} \tau^3, \end{aligned}$$

beschränken.

Zuletzt soll die folgende Differenz betrachtet werden:

$$\begin{aligned} \big(Z_2(\widetilde{u}(t_{j-1})) - Z_2(\widetilde{u}(t_j)) \big) = -\frac{1}{4} \int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi - \frac{\theta}{2})A} \widetilde{\Psi}_S \Big(J_b \left(\widetilde{\Phi} e^{\frac{\theta}{2} A} \varphi_{A+\widetilde{b}}(\tau - \xi, \widetilde{u}(t_{j-1}) \right) \\ - J_b \left(\widetilde{\Phi} e^{\frac{\theta}{2} A} \varphi_{A+\widetilde{b}}(\tau - \xi, \widetilde{u}(t_j)) \right) \Big) e^{\left(\frac{\theta}{2} + \tau - \xi\right) A} \, d\theta \, d\sigma \, d\xi. \end{aligned}$$

Wie bei der obigen Differenz zwischen $z_1(\widetilde{u}(t_j))$ und $z_1(\widetilde{u}(t_{j+1}))$ folgt mit Hilfe des Hauptsatzes der Differential- und Integralrechung für ein beliebiges $w \in \mathbb{R}^{2d}$

$$\begin{aligned} &\left(J_b \left(\widetilde{\Phi} e^{\frac{\theta}{2} A} \widetilde{u}(t_{j-1} + \tau - \xi) \right) - J_b \left(\widetilde{\Phi} e^{\frac{\theta}{2} A} \widetilde{u}(t_j + \tau - \xi) \right) \right)(w) \\ &\quad = \int_0^1 H_b \left(\widetilde{\Phi} e^{\frac{\theta}{2} A} (s \widetilde{u}(t_{j-1} + \tau - \xi) + (1 - s) \widetilde{u}(t_j + \tau - \xi)) \right) \\ &\qquad (\widetilde{\Phi} e^{\frac{\theta}{2} A} (\widetilde{u}(t_{j-1} + \tau - \xi) - \widetilde{u}(t_j + \tau - \xi)), w) \, ds. \end{aligned}$$

Die Variation-der-Konstanten-Formel angewandt auf die Differenz $\widetilde{u}(t_{j-1} + \tau - \xi) - \widetilde{u}(t_j + \tau - \xi)$ ergibt

$$\frac{1}{\tau} \left(Z_2(\widetilde{u}(t_{j-1})) - Z_2(\widetilde{u}(t_j)) \right)(w) = Z_2^{(1)}(w) + Z_2^{(2)}(w),$$

mit

$$Z_2^{(1)}(w) = -\frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi \int_0^1 e^{(\xi-\frac{\theta}{2})A}\widetilde{\Psi}_S H_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}(s\widetilde{u}(t_{j-1}+\tau-\xi)+(1-s)\widetilde{u}(t_j+\tau-\xi))\right)$$
$$\cdot\,(\widetilde{\Phi}e^{\frac{\theta}{2}A}e^{(\tau-\xi)A}\frac{1}{\tau}\big(\widetilde{u}(t_{j-1})-\widetilde{u}(t_j)\big), e^{(\frac{\theta}{2}+\tau-\xi)A}w)\,ds\,d\theta\,d\sigma\,d\xi,$$
$$Z_2^{(2)}(w) = -\frac{1}{4\tau}\int_0^\tau \int_0^\xi \int_\sigma^\xi \int_0^1 e^{(\xi-\frac{\theta}{2})A}\widetilde{\Psi}_S H_b\left(\widetilde{\Phi}e^{\frac{\theta}{2}A}(s\widetilde{u}(t_{j-1}+\tau-\xi)+(1-s)\widetilde{u}(t_j+\tau-\xi))\right)$$
$$\cdot\,(\widetilde{\Phi}e^{\frac{\theta}{2}A}\int_0^{\tau-\xi} e^{(\tau-\xi-\nu)A}\Big(\widetilde{b}(\widetilde{u}(t_{j-1}+\nu))-\widetilde{b}(\widetilde{u}(t_j+\nu))\Big)\,d\nu, e^{(\frac{\theta}{2}+\tau-\xi)A}w)\,ds\,d\theta\,d\sigma\,d\xi.$$

Beide Teile können mittels (1.3) mit den Radien aus (4.21),(3.5), (4.1b), (4.16), (4.29) und (4.34) nach oben abgeschätzt werden durch

$$\left\|Z_2^{(1)}\right\| \le \frac{1}{4}M_1 C_{\mathsf{inv}}\widetilde{C_{g,2}}M_1(\widetilde{K}+C_{\mathsf{inv}}M_1\widetilde{C_{g,0}})\frac{\tau^3}{6} = \frac{1}{24}M_1^2 C_{\mathsf{inv}}\widetilde{C_{g,2}}(\widetilde{K}+C_{\mathsf{inv}}M_1\widetilde{C_{g,0}})\tau^3,$$
$$\left\|Z_2^{(2)}\right\| \le \frac{1}{4}\frac{1}{\tau}M_1 C_{\mathsf{inv}}\widetilde{C_{g,2}}M_1 2M_1 C_{\mathsf{inv}}\widetilde{C_{g,0}}\frac{\tau^4}{24} = \frac{1}{48}M_1^3 C_{\mathsf{inv}}^2\widetilde{C_{g,2}}\widetilde{C_{g,0}}\tau^3.$$

Insgesamt gilt

$$\left\|\frac{1}{\tau}\left(Z_2(\widetilde{u}(t_{j-1}))-Z_2(\widetilde{u}(t_j))\right)\right\| \le \frac{1}{24}M_1^2 C_{\mathsf{inv}}\widetilde{C_{g,2}}\left(\widetilde{K}+\frac{3}{2}M_1 C_{\mathsf{inv}}\widetilde{C_{g,0}}\right)\tau^3.$$

□

4.1.3 Globaler Verfahrensfehler

Die Darstellung des lokalen Fehler aus Lemma 4.5 und 4.7 werden nun verwendet, um den globalen Fehler nach oben zu beschränken und damit zu zeigen, dass das Splittingverfahren (3.10) konvergent von der Ordnung zwei ist.

Satz 4.8. *(Globaler Fehler des modifizierten Problems) Es seien die Annahme 1.1 mit den Radien $\widetilde{r}_0, \widetilde{r}_1$ und $\widetilde{r}_2$ aus* (4.21)*, die Annahmen 3.1 und 4.1 erfüllt. Dann existiert eine Schrittweite τ_0, sodass für alle $\tau \le \tau_0$ der Fehler des Splittingverfahrens* (3.10) *als Näherung der Lösung des modifizierten Systems* (3.6) *der oberen Schranke*

$$\|u_n - \widetilde{u}(t_n)\| \le C\tau^2, \qquad 0 \le t_n = n\tau \le t_{\mathsf{end}}, \tag{4.35}$$

genügt, wobei die Konstanten C und τ_0 nur von C_{inv}, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$, $\widetilde{K}$, M_i, $i = 1,2,3,4$ und t_{end} abhängen.

Um die Beweisidee zu illustrieren, betrachtet man, wie auch schon beim lokalen Fehler, den linearen Fall (siehe Buchholz et al. (2018), Beweis von Theorem 4.1). Die Notation ist diesem Artikel entnommen (vergleiche Abschnitt 2.2.7). Der lokale Fehler $\delta_n = \widehat{\delta}_n + D_n$, mit $\|D_n\| \le C\tau^3$, konnte dabei im linearen Fall in die Anteile

$$\widehat{\delta}_n = \widehat{\delta}_n^{(1)} + \widehat{\delta}_n^{(2)}, \qquad \text{mit } \widehat{\delta}_n^{(1)} = A\widetilde{\Psi}Z_1\widetilde{u}(t_n) \quad \text{und} \quad \widehat{\delta}_n^{(2)} = Z_2\widetilde{\Phi}A^2\widetilde{u}(t_n),$$

zerlegt werden, wobei Z_1 und Z_2 nun Matrizen sind, deren Normen durch $C\tau^3$ beschränkt sind. Der globale Fehler $\widetilde{e}_n = \widetilde{u}_n - \widetilde{u}(t_n)$ kann mithilfe einer Teleskop-Summe geschrieben werden als

$$\begin{aligned}\widetilde{e}_n = (S_{\text{lin}}^n - e^{n\tau(A+\widetilde{B})})\widetilde{u}(t_0) &= \sum_{j=0}^{n-1} S_{\text{lin}}^{n-j-1}(S_{\text{lin}} - e^{\tau(A+\widetilde{B})})e^{j\tau(A+\widetilde{B})}\widetilde{u}(t_0) \\ &= \sum_{j=0}^{n-1} S_{\text{lin}}^{n-j-1}\delta_j.\end{aligned} \tag{4.36}$$

Die Analyse eines Splittingverfahrens für glatte Lösungen enden häufig an dieser Stelle, da die Terme der Größenordnung $\mathcal{O}(\tau^2)$ wie $\widehat{\delta}_n^{(1)}$ oder $\widehat{\delta}_n^{(2)}$ im lokalen Fehler nicht auftreten. Hat der Operator S Norm kleiner oder gleich eins, also ist das zugehörige Splittingverfahren stabil, so folgt für Terme der Ordnung $\mathcal{O}(\tau^3)$ wie D_n direkt

$$\left\| \sum_{j=0}^{n-1} S_{\text{lin}}^{n-j-1} D_j \right\| \leq C\tau^2.$$

Das ist das klassiche Argument des Fächers der Lady Windermeres. Im hochoszillatorischen Fall muss man mehr Aufwand betreiben, um zu garantieren, dass sich die Größenordnung von $\widehat{\delta}_n^{(1)}$ oder $\widehat{\delta}_n^{(2)}$ durch die Summation nicht ändert. Dafür betrachtet man

$$\widetilde{e}_n^{(1)} = \sum_{j=0}^{n-1} S_{\text{lin}}^{n-j-1}\widehat{\delta}_j^{(1)}.$$

In der Literatur (vergleiche Kapitel 2) wurde häufig partielle Summation verwendet, um solche Summen zu beschränken (eine Erläuterung zur partiellen Summation liefert Abschnitt B.2). Es folgt

$$\begin{aligned}\widetilde{e}_n^{(1)} &= \mathbb{S}_n\widehat{\delta}_0^{(1)} + \sum_{j=0}^{n-2} \mathbb{S}_{n-j-1}(\widehat{\delta}_{j+1}^{(1)} - \widehat{\delta}_j^{(1)}) \\ &= \mathbb{S}_n A\widetilde{\Psi}Z_1\widetilde{u}(t_0) + \sum_{j=0}^{n-2} \mathbb{S}_{n-j-1}A\widetilde{\Psi}Z_1(\widetilde{u}(t_{j+1}) - \widetilde{u}(t_j)),\end{aligned}$$

mit

$$\mathbb{S}_j = \sum_{k=0}^{j-1} S_{\text{lin}}^k.$$

Da man zeigen kann, dass die Differenz $\widetilde{u}(t_{j+1}) - \widetilde{u}(t_j)$ sich proportional zu τ beschränken lässt, ist die Summation über j dieser Differenz von der Größenordnung $\mathcal{O}(1)$. Es verbleibt zu beweisen, dass $\mathbb{S}_n A\widetilde{\Psi}Z_1$ die Größenordnung $\mathcal{O}(\tau^2)$ hat. Wegen $e^{\tau\widetilde{B}} = I + \tau\widetilde{B}$ (siehe (2.55)) und $\left\|\widetilde{B}\right\| \leq C$ gilt

$$S_{\text{lin}} = e^{\frac{\tau}{2}A}e^{\tau\widetilde{B}}e^{\frac{\tau}{2}A} = e^{\tau A} + \mathcal{O}(\tau).$$

Damit folgt zunächst

$$\mathbb{S}_n = \sum_{k=0}^{n-1} S_{\text{lin}}^k = \sum_{k=0}^{n-1} e^{k\tau A} + \mathcal{O}(1).$$

Intuitiv würde man nun gerne die geometrische Summenformel auf $\sum_{k=0}^{n-1} e^{k\tau A}$ anwenden. Nun ist $e^{\tau A} - I$ jedoch nicht invertierbar, wenn $\tau\Omega$ Eigenwerte bei Vielfachen von 2π hat. Man kann jedoch verwenden, dass $\mathbb{S}_j$ auf $A\widetilde{\Psi}Z_1$ angewandt wird. Denn $\tau A\widehat{\chi}$ mit $\widehat{\chi} = \widetilde{\Phi}$ oder $\widehat{\chi} = \widetilde{\Psi}$ lässt sich schreiben als

$$\tau A\widehat{\chi} = \tau\widehat{\chi}A = (e^{\tau A} - I)\widehat{\Theta}_\chi = \widehat{\Theta}_\chi(e^{\tau A} - I) \tag{4.37}$$

mit

$$\widehat{\Theta}_\chi = \widehat{\theta}_\chi(\tau\Omega), \qquad \widehat{\theta}_\chi(x) = \frac{1}{2}\begin{bmatrix} \cot(\frac{x}{2}) & -1 \\ 1 & \cot(\frac{x}{2}) \end{bmatrix}\begin{bmatrix} x\chi(x) & 0 \\ 0 & x\chi(x) \end{bmatrix}.$$

Dies folgt direkt mittels der Identitäten

$$\sin(x) = \cot(\tfrac{x}{2})\big(1 - \cos(x)\big) \qquad \text{und} \qquad \cos(x) + 1 = \cot(\tfrac{x}{2})\sin(x).$$

Die Filterbedingungen (4.1c) und (4.1d) liefern die Beschränktheit von $\widehat{\Theta}_\chi$:

$$\left\|\widehat{\Theta}_\chi\right\| \leq C. \tag{4.38}$$

Damit folgt für die Summe

$$\left\|\mathbb{S}_n A\widetilde{\Psi}Z_1\right\| = \left\|\mathbb{S}_n(e^{\tau A} - I)\widehat{\Theta}_\Psi\frac{1}{\tau}Z_1\right\| \leq C\tau^2,$$

denn $\left\|\mathbb{S}_n(e^{\tau A} - I)\right\| \leq C$ und man kann zusätzlich $\left\|\frac{1}{\tau}Z_1\right\| \leq C\tau^2$ nachweisen. Damit lässt sich also die Summe beschränken:

$$\left\|\sum_{j=0}^{n-1} S_{\text{lin}}^{n-j-1}\widehat{\delta}_j^{(1)}\right\| \leq C\tau^2.$$

Mit ähnlicher Technik lässt sich auch die zweite Summe über $\delta_j^{(2)}$ abschätzen.

Da man bereits für den lokalen Fehler eine vergleichbare Darstellung wie im linearen Fall finden konnte, wird zunächst versucht, diesen Beweisansatz auch auf Satz 4.8 zu übertragen. Die Darstellung des globalen Fehler erfolgt dabei nicht mehr über eine Teleskopsumme, da dies für eine nichtlineare Funktion g in dieser Form nicht möglich ist (hier geht wesentlich die Linearität von S_{lin} ein). Jedoch kann die Fehlerrekursion aus Grimm und Hochbruck (2006), siehe auch Abschnitt 2.2.5, verwendet werden. Daraus ergibt sich eine Darstellung des globalen Fehlers, die eine vergleichbare Analyse wie im linearen Fall erlaubt. Hier ist es ebenfalls möglich, die Terme der Größenordnung $\mathcal{O}(\tau^3)$ des lokalen Fehlers direkt abzuschätzen (vergleiche (4.41)). Für die verbleibenden Terme der Größenordnung $\mathcal{O}(\tau^2)$ kann partielle Summation verwendet werden. Auch für die Filterfunktionen benötigt man keine zusätzlichen Annahmen, da ähnliche Summen beschränkt werden müssen. Somit konnte ein großer Teil des Beweises des globalen Fehlers im linearen Fall für den nichtlinearen Fall adaptiert werden.

Gegenüber den bisherigen Konvergenzbeweisen aus García-Archilla et al. (1999), Hochbruck und Lubich (1999), Hairer et al. (2006) oder Grimm und Hochbruck (2006) unterscheidet sich dieser Beweis hauptsächlich darin, dass die Darstellung des lokalen Fehlers in völlig anderer Weise geschieht. Mithilfe dieser Darstellung und der Transformation der Geschwindigkeiten genügt es dann,

lediglich zwei partielle Summationen zur Abschätzung des globalen Fehlers zu verwenden. Auch die Filterbedingungen werden dadurch stark beeinflusst, siehe auch Abschnitt 4.3.

Beweis von Satz 4.8. Zuerst leitet man eine Fehlerrekursion analog zu Grimm und Hochbruck (2006) her. Dazu stellt man die exakte Lösung zunächst in Form des Splittingverfahrens (3.10) dar:

$$\widetilde{u}(t_{n+1}) = S(\widetilde{u}(t_n)) - \delta_n$$

Hierbei tritt der lokale Fehler $\delta_n = S(\widetilde{u}(t_n)) - \varphi_{A+\widetilde{b}}(\tau, \widetilde{u}(t_n))$ auf. Subtrahiert man nun die numerische Lösung $\widetilde{u}_{n+1}$, erhält man die folgende Rekursion für den globalen Fehler $\widetilde{e}_n = \widetilde{u}(t_n) - \widetilde{u}_n$:

$$\widetilde{e}_{n+1} = S(\widetilde{u}(t_n)) - S(\widetilde{u}_n) - \delta_n$$

Die Differenz zwischen $S(\widetilde{u}(t_n))$ und $S(\widetilde{u}_n)$ lässt sich schreiben als

$$\begin{aligned} S(\widetilde{u}(t_n)) - S(\widetilde{u}_n) &= e^{\frac{\tau}{2}A}\Big(\varphi_{\widetilde{b}}\left(\tau, e^{\frac{\tau}{2}A}\widetilde{u}(t_n)\right) - \varphi_{\widetilde{b}}\left(\tau, e^{\frac{\tau}{2}A}\widetilde{u}_n\right)\Big) \\ &= e^{\tau A}(\widetilde{u}(t_n) - \widetilde{u}_n) + \tau e^{\frac{\tau}{2}A}\big[\widetilde{b}(e^{\frac{\tau}{2}A}\widetilde{u}(t_n)) - \widetilde{b}(e^{\frac{\tau}{2}A}\widetilde{u}_n)\big] \\ &= e^{\tau A}\widetilde{e}_n + \tau e^{\frac{\tau}{2}A}\int_0^1 J_{\widetilde{b}}\left(e^{\frac{\tau}{2}A}(s\widetilde{u}(t_n) + (1-s)\widetilde{u}_n)\right)\widetilde{\Phi}e^{\frac{\tau}{2}A}\widetilde{e}_n\, ds \\ &= e^{\tau A}\widetilde{e}_n + \tau e^{\frac{\tau}{2}A}\mathcal{J}_n\widetilde{e}_n, \end{aligned}$$

mit

$$\mathcal{J}_n = \int_0^1 J_{\widetilde{b}}\left(e^{\frac{\tau}{2}A}(s\widetilde{u}(t_n) + (1-s)\widetilde{u}_n)\right)\widetilde{\Phi}e^{\frac{\tau}{2}A}\, ds.$$

Der globale Fehler genügt damit der Rekursionsformel

$$\widetilde{e}_{n+1} = e^{(n+1)\tau A}\widetilde{e}_0 + \tau e^{\frac{\tau}{2}A}\sum_{j=0}^{n} e^{(n-j)\tau A}\mathcal{J}_j\widetilde{e}_j - \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j. \tag{4.39}$$

Um den globalen Fehler nach oben zu beschränken, soll das diskrete Lemma von Gronwall B.3 verwendet werden. Dafür wird zunächst gezeigt, dass

$$\left\| \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j \right\| \leq C\tau^2$$

gilt. Nach Lemma 4.5 lässt sich der lokale Fehler aufteilen in

$$\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j = \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(1)} + \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)} + \sum_{j=0}^{n} e^{(n-j)\tau A}D_j. \tag{4.40}$$

Da $\|D_j\| \leq C\tau^3$ ist, folgt wegen $n\tau \leq t_{\text{end}}$ für die letzte Summe direkt

$$\left\| \sum_{j=0}^{n} e^{(n-j)\tau A}D_j \right\| \leq C\tau^2. \tag{4.41}$$

Im Folgenden werden nun die verbleibenden Summen nach oben beschränkt:

Man definiert zunächst

$$E_j = \sum_{k=0}^{j} e^{k\tau A}.$$

Partielle Summation angewandt auf die erste Summe auf der rechten Seite in (4.40) (siehe auch Abschnitt B.2) ergibt:

$$\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(1)} = E_n\delta_0^{(1)} + \sum_{j=0}^{n-1} E_{n-j-1}(\delta_{j+1}^{(1)} - \delta_j^{(1)})$$

Nun verwendet man die Darstellung aus Lemma 4.7. Daher folgt

$$\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(1)} = E_n A\widetilde{\Psi}_S z_1(\widetilde{u}(t_0)) + \sum_{j=0}^{n-1} E_{n-j-1}A\widetilde{\Psi}_S(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))). \tag{4.42}$$

Wie im linearen Fall verwendet man für $\widehat{\chi} = \widetilde{\Phi}$ oder $\widehat{\chi} = \widetilde{\Psi}$ die Matrix

$$\widehat{\Theta}_\chi = \widehat{\theta}_\chi(\tau\Omega), \qquad \widehat{\theta}_\chi(x) = \frac{1}{2}\begin{bmatrix} \cot(\frac{x}{2}) & -1 \\ 1 & \cot(\frac{x}{2}) \end{bmatrix}\begin{bmatrix} x\chi(x) & 0 \\ 0 & x\chi(x) \end{bmatrix}.$$

Damit ergeben sich die Identitäten

$$\tau A\widehat{\chi} = \tau\widehat{\chi}A = (e^{\tau A} - I)\widehat{\Theta}_\chi = \widehat{\Theta}_\chi(e^{\tau A} - I) \tag{4.43}$$

und wegen (4.1c) und (4.1d) ist

$$\left\|\widehat{\Theta}_\chi\right\| \leq C. \tag{4.44}$$

Damit lässt sich (4.42) schreiben als

$$\begin{aligned}\sum_{j=0}^{n} e^{(n-j)\tau A}\widehat{\delta}_j^{(1)} = {} & E_n(e^{\tau A} - I)\widehat{\Theta}_\Psi\left(\frac{1}{\tau}z_1(\widetilde{u}(t_0))\right) \\ & + \sum_{j=0}^{n-1} E_{n-j-1}(e^{\tau A} - I)\widehat{\Theta}_\Psi\frac{1}{\tau}\left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))\right).\end{aligned}$$

In Lemma 4.7 wurde bereits gezeigt, dass

$$\|z_1(\widetilde{u}(t_n))\| \leq C_1\tau^3 \qquad \text{und} \qquad \left\|\frac{1}{\tau}\left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))\right)\right\| \leq C_2\tau^3$$

gilt. Damit verbleibt $E_j(e^{\tau A} - I)$ zu beschränken. Es gilt

$$E_j(e^{\tau A} - I) = \sum_{k=0}^{j-1} e^{k\tau A}(e^{\tau A} - I) = e^{j\tau A} - I \quad \Rightarrow \quad \left\|E_j(e^{\tau A} - I)\right\| \leq 2.$$

Insgesamt kann man nun mit (4.38) und $n\tau \leq t_{\text{end}}$ die Summe abschätzen:

$$\begin{aligned}\left\|\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(1)}\right\| &\leq \left\|E_n(e^{\tau A}-I)\right\| \left\|\widehat{\Theta}_\Psi\right\| \left\|\frac{1}{\tau}z_1(\widetilde{u}(t_0))\right\| \\ &\quad + \sum_{j=0}^{n-1} \left\|E_{n-j-1}(e^{\tau A}-I)\right\| \left\|\widehat{\Theta}_\Psi\right\| \left\|\frac{1}{\tau}\left(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))\right)\right\| \\ &\leq C\tau^2.\end{aligned}$$

Es folgt die Berechnung einer oberen Schranke für $\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)}$. Definiere dafür

$$F_j = \sum_{k=0}^{j} \widetilde{u}(t_k).$$

Mit der Darstellung

$$\delta_n^{(2)} = Z_2(\widetilde{u}(t_n))\widetilde{\Phi}A^2\widetilde{u}(t_n)$$

aus Lemma 4.7 folgt mithilfe partieller Summation (siehe Appendix B.2)

$$\begin{aligned}\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)} &= \sum_{j=0}^{n} e^{(n-j)\tau A} Z_2(\widetilde{u}(t_j))\widetilde{\Phi}A^2\widetilde{u}(t_j) \\ &= Z_2(\widetilde{u}(t_n))\Phi A^2 F_n + \sum_{j=0}^{n-1} e^{(n-j)\tau A}\left(Z_2(\widetilde{u}(t_j)) - e^{-\tau A}Z_2(\widetilde{u}(t_{j+1}))\right)\widetilde{\Phi}A^2F_j.\end{aligned}$$

Mit $\widehat{\Theta}_\Phi$ und (4.37) folgt

$$\begin{aligned}&= \frac{1}{\tau}Z_2(\widetilde{u}(t_n))\Theta_\Phi(e^{\tau A}-I)AF_n \\ &\quad + \sum_{j=0}^{n-1} e^{(n-j)\tau A}\frac{1}{\tau}\left(Z_2(\widetilde{u}(t_j)) - e^{-\tau A}Z_2(\widetilde{u}(t_{j-1}))\right)\Theta_\Phi(e^{\tau A}-I)AF_j.\end{aligned}$$

Alle auftretenden Terme können einzeln beschränkt werden. Die Idee zur Umsetzung der partiellen Summation für $\delta_j^{(2)}$ entstand in Zusammenarbeit mit Ludwig Gauckler.

Betrachte zunächst die Differenz

$$\frac{1}{\tau}\left(Z_2(\widetilde{u}(t_j)) - e^{-\tau A}Z_2(\widetilde{u}(t_{j-1}))\right) = \frac{1}{\tau}e^{-\tau A}\left(Z_2(\widetilde{u}(t_j)) - Z_2(\widetilde{u}(t_{j-1}))\right) - \frac{1}{\tau}(e^{-\tau A}-I)Z_2(\widetilde{u}(t_j)).$$

Beide Terme können mit Hilfe des Lemmas 4.7 beschränkt werden durch

$$\begin{aligned}\left\|\frac{1}{\tau}e^{-\tau A}\left(Z_2(\widetilde{u}(t_j)) - Z_2(\widetilde{u}(t_{j-1}))\right)\right\| &\leq \left\|\frac{1}{\tau}\left(Z_2(\widetilde{u}(t_j)) - Z_2(\widetilde{u}(t_{j-1}))\right)\right\| \leq C\tau^3, \\ \left\|\frac{1}{\tau}(e^{-\tau A}-I)Z_2(\widetilde{u}(t_j))\right\| &\leq \left\|(e^{-\tau A}-I)(-\tau A)^{-1}\right\| \left\|AZ_2(\widetilde{u}(t_j))\right\| \leq C\tau^3,\end{aligned}$$

wobei $\|(e^{-\tau A} - I)(-\tau A)^{-1}\| = \| \int_0^1 e^{-\sigma\tau A} d\sigma\| \leq 1$ verwendet wurde.

Schlussendlich wird der Term $(e^{\tau A} - I)AF_j$ betrachtet. Zuerst wird die exakte Lösung künstlich eingefügt und schließlich die Variation-der-Konstanten-Formel angewendet. Es folgt

$$\begin{aligned}
(e^{\tau A} - I)AF_j &= A\sum_{k=0}^{j}(e^{\tau A} - I)\widetilde{u}(t_k) \\
&= A\sum_{k=0}^{j}(e^{\tau A}\widetilde{u}(t_k) - \varphi_{A+\widetilde{b}}(\tau, \widetilde{u}(t_k))) + A\sum_{k=0}^{j}(\varphi_{A+\widetilde{b}}(\tau, \widetilde{u}(t_k)) - \widetilde{u}(t_k)) \\
&= A\sum_{k=0}^{j}\int_0^\tau e^{(\tau-s)A}\widetilde{b}(\widetilde{u}(t_k + s))\, ds + A\sum_{k=0}^{j}(\widetilde{u}(t_{k+1}) - \widetilde{u}(t_k)) \\
&= \sum_{k=0}^{j}\int_0^\tau e^{(\tau-s)A}A\widetilde{b}(\widetilde{u}(t_k + s))\, ds + A(\widetilde{u}(t_{j+1}) - \widetilde{u}_0).
\end{aligned}$$

Damit lässt sich dieser Term mit Hilfe von (1.3), (3.5), (4.1b), und (4.19) beschränken durch

$$\left\|(e^{\tau A} - I)AF_j\right\| \leq j\tau M_1\widetilde{C_{g,0}} + 2\widetilde{K} \leq t_{\mathsf{end}}M_1\widetilde{C_{g,0}} + 2\widetilde{K}$$

für alle $j \leq n$ und $n\tau \leq t_{\mathsf{end}}$. Insgesamt folgt daraus

$$\begin{aligned}
\left\|\sum_{j=0}^{n} e^{(n-j)\tau A}\widehat{\delta}_j^{(2)}\right\| &\leq \frac{1}{\tau}\left\|Z_2(\widetilde{u}(t_n))\right\|\left\|\Theta_\Phi\right\|\left\|(e^{\tau A} - I)AF_n\right\| \\
&\quad + \sum_{j=0}^{n-1}\left\|e^{(n-j)\tau A}\right\|\left\|\frac{1}{\tau}\left(Z_2(\widetilde{u}(t_j)) - e^{\tau A}Z_2(\widetilde{u}(t_{j-1}))\right)\right\|\left\|\Theta_\Phi\right\|\left\|(e^{\tau A} - I)AF_j\right\| \\
&\leq C\tau^2.
\end{aligned}$$

Abschließend wurde damit

$$\left\|\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j\right\| \leq C_\delta\tau^2 \tag{4.45}$$

bewiesen.

Nun kann

$$\|\widetilde{e}_n\| \leq C\tau^2$$

gezeigt werden. Dafür seien folgende Konstanten definiert:

$$C_{\mathcal{J}} = M_1^2 C_{\mathsf{inv}}\widetilde{C_{g,1}} \qquad \text{und} \qquad \tau_0 = \left(\frac{C_{\mathsf{inv}}\widetilde{K}}{C_\delta(1 + t_{\mathsf{end}}C_{\mathcal{J}}e^{t_{\mathsf{end}}C_{\mathcal{J}}})}\right)^{\frac{1}{2}}.$$

Somit hängt τ_0 von $\widetilde{K}$, C_{inv}, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$, M_i, $i = 1, 2, 3, 4$ und t_{end} ab, aber nicht von n oder Ω. Damit ist die Schrittweiteneinschränkung weniger restriktiv wie eine CFL-Bedingung eines klassischen expliziten Verfahrens. Eine ähnliche obere Schranke findet sich auch bei Gauckler (2015), siehe auch Abschnitt 2.2.6. Es soll nun die Fehlerschranke

$$\|\widetilde{e}_n\| \leq C_\delta(1 + t_{\mathsf{end}}C_{\mathcal{J}}e^{t_{\mathsf{end}}C_{\mathcal{J}}})\tau^2 = C_E\tau^2, \qquad \text{für } \tau \leq \tau_0,$$

durch Induktion nach n gezeigt werden.

Der Induktionsanfang ist klar, denn $\widetilde{e}_0 = \widetilde{u}(0) - \widetilde{u}_0 = 0$.
Um den Induktionsschritt von $n \to n+1$ zu zeigen, wird zunächst die Stabilität des numerischen Verfahrens für alle $k \leq n$ und $\tau \leq \tau_0$ nachgewiesen:

$$\|\widetilde{u}_k\| \leq \|\widetilde{u}(t_k)\| + \|\widetilde{e}_k\| \leq C_{\mathsf{inv}}\widetilde{K} + C_\delta(1 + t_{\mathsf{end}} C_{\mathcal{J}} e^{t_{\mathsf{end}} C_{\mathcal{J}}})\tau^2 \leq 2C_{\mathsf{inv}}\widetilde{K}.$$

Hier erkennt man die Notwendigkeit einer oberen Schrittweiteneinschränkung τ_0. Ebenfalls ist klar, dass eine höhere Schranke angesetzt werden kann, welche sich dann in einer größeren Fehlerkonstante C_E widerspiegeln würde. Mit Hilfe der Stabilität, (3.5) und (4.29) folgt

$$\left\| e^{\frac{\tau}{2}A}(s\widetilde{u}(t_k) + (1-s)\widetilde{u}_k) \right\| \leq 2M_1 C_{\mathsf{inv}}\widetilde{K} < \widetilde{r}_1, \qquad \text{für } 0 \leq s \leq 1,$$

und damit mithilfe von (1.3), (4.1b)

$$\|\mathcal{J}_k\| = \left\| \int_0^1 J_{\widetilde{b}}\left(e^{\frac{\tau}{2}A}(s\widetilde{u}(t_k) + (1-s)\widetilde{u}_k) \right) \widetilde{\Phi} e^{\frac{\tau}{2}A}\, ds \right\| \leq C_{\mathcal{J}}, \qquad \text{für } k \leq n.$$

Mithilfe der Rekursionsformel für den Fehler $\widetilde{e}_{n+1}$ (4.39), der Abschätzung für die Summe über den lokalen Fehler (4.45) und $\widetilde{e}_0 = 0$ folgt

$$\begin{aligned} \|\widetilde{e}_{k+1}\| &= \left\| e^{(k+1)\tau A}\widetilde{e}_0 + \tau e^{\frac{\tau}{2}A} \sum_{j=0}^{k} e^{(k-j)\tau A} \mathcal{J}_j \widetilde{e}_j - \sum_{j=0}^{k} e^{(k-j)\tau A}\delta_j \right\| \\ &\leq C_\delta \tau^2 + \sum_{j=1}^{k} \tau C_{\mathcal{J}} \|\widetilde{e}_j\|, \qquad \text{für } k \leq n. \end{aligned}$$

Mithilfe des diskreten Lemmas von Gronwall B.3 folgt

$$\begin{aligned} \|\widetilde{e}_{n+1}\| &\leq C_\delta \tau^2 \left(1 + \sum_{j=1}^{n} \tau C_{\mathcal{J}} e^{t_{\mathsf{end}} C_{\mathcal{J}}} \right) \\ &\leq C_\delta \left(1 + t_{\mathsf{end}} C_{\mathcal{J}} e^{t_{\mathsf{end}} C_{\mathcal{J}}} \right) \tau^2. \end{aligned}$$

Damit ist die Induktionsbehauptung gezeigt und somit auch Satz 4.8. □

Bemerkung 4.9. *Satz 4.8 zeigt die Stabilität des Verfahrens* (3.10) *für* $0 \leq k \leq n$ *und* $0 \leq n\tau \leq t_{end}$.

Bemerkung 4.10. *(Schrittweiteneinschränkung)*
Im Fall, dass g *linear ist, ist keine Schrittweiteneinschränkung notwendig (siehe Buchholz et al. (2018)). Dies liegt vor allem an der Darstellung des globalen Fehlers über die Teleskopsumme (siehe* (4.36)*) anstelle der Rekursionsformel* (4.39)*. Der globale Fehler reduziert sich dabei auf eine Summe über den lokalen Fehler, der analog wie oben beschränkt werden kann. Die Stabilität des Verfahrens wird ebenfalls benötigt, kann jedoch für beliebige Schirttweiten* τ *garantiert werden (siehe Gleichung* (5.2) *aus Buchholz et al. (2018)).*

Auch kann man eine Schrittweiteneinschränkung vermeiden, wenn man annimmt, dass J_b gleichmäßig beschränkt ist, also $\|J_b(u)\| \leq \widetilde{C_{g,1}}$ unabhängig von $u \in \mathbb{R}^{2d}$. Gleiches war auch in Hochbruck und Lubich (1999) und Grimm und Hochbruck (2006) angenommen worden, weshalb man dort ebenfalls keine Schrittweiteneinschränkung benötigt.

Ist J_b gleichmäßig beschränkt, so ist

$$\|\mathcal{J}_j\| = \left\| \int_0^1 J_{\widetilde{b}}\left(e^{\frac{\tau}{2}A}(s\widetilde{u}(t_n) + (1-s)\widetilde{u}_j)\right) \widetilde{\Phi} e^{\frac{\tau}{2}A}\, ds \right\| \leq M_1^2 C_{inv} \widetilde{C_{g,1}} = C_{\mathcal{J}}.$$

Die Konstante $C_{\mathcal{J}}$ hängt dabei nicht von der Stabilität der numerischen Methode ab. Die Fehlerschranke folgt dann wie oben mit Hilfe des Lemmas von Gronwall.

Damit lässt sich schlussendlich der zentrale Konvergenzsatz der Arbeit formulieren:

Satz 4.11. *(Konvergenzresultat) Es gelten die Annahmen 1.1 mit den Radien $\widetilde{r}_0$, $\widetilde{r}_1$ und $\widetilde{r}_2$ aus* (4.21)*, die Annahmen 3.1 und 4.1. Dann existiert eine Schrittweite τ_0, sodass für alle $\tau \leq \tau_0$ der Fehler des Splittingverfahrens* (3.10) *angewandt auf* (3.2a) *beschränkt ist durch*

$$\|u_n - u(t_n)\| \leq C\tau^2, \qquad 0 \leq t_n = n\tau \leq t_{end},$$

wobei die Konstanten C und τ_0 nur von C_{inv}, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$, K, M_j, $j = 1, \ldots, 4$, und t_{end} abhängen, aber nicht von $\|\Omega\|$ oder τ.

Beweis von Satz 4.11. Sei $\widetilde{u}$ die exakte Lösung der modifizierten Gleichung (3.6). Kombiniert man die Ergebnisse aus Satz 4.2 (beachte, dass die Radien aus (4.7) kleiner sind als die Radien in (4.21) und somit der Satz auch mit obigen Radien verwendet werden darf) und Satz 4.8, so erhält man

$$\|u_n - u(t_n)\| \leq \|u_n - \widetilde{u}(t_n)\| + \|\widetilde{u}(t_n) - u(t_n)\| \leq C\tau^2, \qquad \text{für } \tau \leq \tau_0, 0 \leq t_n = n\tau \leq t_{\text{end}},$$

wobei die Konstanten C und τ_0 nur von C_{inv}, $\widetilde{C_{g,0}}$, $\widetilde{C_{g,1}}$, $\widetilde{C_{g,2}}$, K, L_g, M_j, $j = 1, 2, 3, 4$, und t_{end} abhängen. □

4.2 Vergleich mit der Fehleranalyse im linearen Fall

In diesem Abschnitt soll die Fehleranalyse aus dem vorangegangen Abschnitt für den semilinearen Fall mit der Fehleranalyse im linearen Fall verglichen werden. Speziell soll die Frage geklärt werden, ob die Ergebnisse der linearen mit der semilinearen Analyse reproduziert werden können. Ein Vergleich der Beweisideen fand teilweise bereits im vorherigen Kapitel als Motivation statt, jedoch ohne zu untersuchen, ob der lineare Fall in der semilinearen Analyse enthalten ist. Daraus ergeben sich hier einige Wiederholungen, die der Vollständigkeit dieses Abschnitts geschuldet sind. Zur Erläuterung der Notation siehe auch Abschnitt 2.2.7.

Folgend wird die semilineare Analyse für die lineare Funktion $g(q) = Gq$ mit einer Matrix G mit kleiner Norm betrachtet. Die Annahmen 1.1 an Ω und die Finite-Energie-Bedingung wurden auch im linearen Fall benötigt. Die Schranke $\widetilde{C_{g,0}}$ ist gegeben durch $\|Gq\| \leq \|G\|\, \widetilde{r}_0$ für alle $\|q\| \leq \widetilde{r}_0$. Die Beschränktheit der ersten und zweiten Ableitung von g wird im linearen Fall nicht explizit vorausgesetzt. Diese ist jedoch gegeben, da man die Norm von G als beschränkt annimmt. Weiterhin

wird hier g als Lipschitz-stetig vorausgesetzt, was für eine lineare Funktion mit kleiner Norm stets gegeben ist.

Wie in Kapitel 3 kann man die Differentialgleichung (1.1) in ein System erster Ordnung überführen, durch eine Multiplikation mit Ω^{-1} in der zweiten Komponente transformieren und schließlich mittels der Filterfunktionen Φ und Ψ_S die modifizierten Gleichungen (2.52) herleiten. Die Annahmen an die Filterfunktionen 4.1 müssen dabei auch im linearen Fall erfüllt sein, um die Fehlerschranke zu beweisen.

Um die Konsistenz zur vorliegenden Arbeit zu wahren, wird im Gegensatz zu Buchholz et al. (2018) hier auch ein Splitting in der vertauschter Reihenfolge verwendet (siehe Abschnitt 3.2: Betrachtet wird das Verfahren (3.10) mit dem Operator S anstelle T). Der Beweis kann jedoch analog zu Buchholz et al. (2018) geführt werden. Es sei weiterhin angemerkt, dass im linearen Fall

$$J_{\widetilde{b}}(u) = \widetilde{B} \qquad \text{und} \qquad H_{\widetilde{b}}(u)(y,z) = 0$$

gilt.

Die Hauptaussage aus Satz 4.11 im semilinearen Fall findet sich in Theorem 2.2 in Buchholz et al. (2018). Die Resultate unterscheiden sich durch die fehlende Schrittweiteneinschränkung im linearen Fall, die bereits in Bemerkung 4.10 thematisiert worden ist. Die Fehlerkonstante C aus der semilinearen Analyse hängt für $g(q) = Gq$ von $\widetilde{C_{g,0}} = \|G\|\,\widetilde{r}_0$, $\widetilde{C_{g,1}} = \|G\|$ und $\widetilde{C_{g,2}} = 0$ ab. Eine Oberschranke für den Radius $\widetilde{r}_0$ wurde in (4.21) angegeben. Dieser hängt nur von Größen ab, die auch in der Fehlerkonstante im linearen Fall auftauchen. Damit stimmen die Sätze im linearen Fall überein.

Satz 4.2 und Lemma 4.3 ergeben im linearen Fall genau die Aussagen aus Theorem 4.1 und Lemma 4.1 aus Buchholz et al. (2018) wieder, wobei hier wieder $C_{g,0} = \|G\|\,r_0$, $\widetilde{C_{g,1}} = \|G\|$ und $\widetilde{C_{g,2}} = 0$ mit r_0 aus (4.7) gewählt wird. Lediglich die Konstante $\widetilde{K}$ kann im linearen Fall unabhängig von C_{av} gewählt werden, denn der Beweis vereinfacht sich.

Es folgt die Darstellung des lokalen Fehlers, hier in Lemma 4.5. Die Annahmen sind identisch mit Lemma 5.1 aus Buchholz et al. (2018), jedoch muss man zuerst die Darstellung vereinfachen, um die Aussagen vergleichen zu können. Aus Lemma 4.5 erhält man für $g(q) = Gq$:

$$\begin{aligned}
\delta_n^{(1)} + \delta_n^{(2)} &= -\frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A} A^2 \widetilde{B} e^{\frac{\theta}{2}A} e^{(\tau-\xi)(A+\widetilde{B})} \widetilde{u}(t_n)\, d\theta\, d\sigma\, d\xi \\
&\quad - \frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A} \widetilde{B} A^2 e^{(\frac{\theta}{2}+\tau-\xi)A} \widetilde{u}(t_n)\, d\theta\, d\sigma\, d\xi \\
&= -\frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A} \big(A^2\widetilde{B} + \widetilde{B}A^2\big) e^{(\frac{\theta}{2}+\tau-\xi)A} \widetilde{u}(t_n)\, d\theta\, d\sigma\, d\xi \\
&\quad + \frac{1}{4}\int_0^\tau \int_0^\xi \int_\sigma^\xi e^{(\xi-\frac{\theta}{2})A} A^2 \widetilde{B} e^{\frac{\theta}{2}A} \int_0^{\tau-\xi} e^{(\tau-\xi-\nu)A} \widetilde{B} \widetilde{u}(t_n+\nu)\, d\nu\, d\theta\, d\sigma\, d\xi,
\end{aligned}$$

Hierbei wird die Variation-der-Konstanten-Formel für $e^{(\tau-\xi)(A+\widetilde{B})}$ in $\delta_n^{(1)}$ verwendet. Das zweite Integral hat die Größenordnung $\mathcal{O}(\tau^3)$ und kann deswegen in D_n gefasst werden. Der erste Term hat bis auf den Faktor $\frac{1}{4}$ anstelle von $\frac{1}{2}$ Faktoren gleicher Größenordnung wie $\hat{\delta}_n$ aus Buchholz et al.

(2018). Insbesondere der kritische Anteil

$$L = A^2\widetilde{B} + \widetilde{B}A^2$$

tritt ebenfalls auf. Der Faktor $\frac{1}{2}$ anstatt $\frac{1}{4}$ liefert lediglich eine größere Fehlerschranke. Er lässt sich folgendermaßen erklären: Im semilinearen Fall wird

$$\left[A, e^{\xi\widetilde{B}}\right] w = \left[A, \varphi_{\widetilde{b}}(\xi, \cdot)\right]_L (w) = \xi \left[A, \widetilde{b}\right]_L (w) = \xi \left[A, \widetilde{B}\right] (w).$$

verwendet. Im linearen Fall dagegen wird der Kommutator mithilfe des Hauptsatzes der Differential- und Integralrechung geschrieben als

$$\left[A, e^{\xi\widetilde{B}}\right] = \int_0^\xi e^{(\xi-\theta)\widetilde{B}} \left[A, \widetilde{B}\right] e^{\theta\widetilde{B}}\, d\theta.$$

Dadurch bedingt wird im linearen Fall der Term $e^{\xi\widetilde{B}} \left[A, \widetilde{B}\right] e^{\frac{\xi}{2}A}$ künstlich addiert und subtrahiert. Dies ist in der vorliegenden semilinearen Analyse nicht nötig. Die Vereinfachung wäre bereits im linearen Fall möglich gewesen.

In Lemma 4.7 wird der lokale Fehler schließlich in eine geeignete Form gebracht um den globalen Fehler abzuschätzen. Die geschieht auch im linearen Fall in Lemma 5.2. Abgesehen von den Änderungen in der Darstellung des lokalen Fehlers ist die Form wie eben gesehen gleich. Die beschränkten Anteile z_1 und Z_2 sind hier jedoch Matrizen Z_1 und Z_2, die nicht von der Lösung $\widetilde{u}(t_n)$ abhängen. Zusätzlich werden im semilinearen Fall die Abschätzungen

$$\left\| \frac{1}{\tau}(z_1(\widetilde{u}(t_{j+1})) - z_1(\widetilde{u}(t_j))) \right\| \leq C\tau^3 \tag{4.46}$$

$$\left\| \frac{1}{\tau}(Z_2(\widetilde{u}(t_{j-1})) - Z_2(\widetilde{u}(t_j))) \right\| \leq C\tau^3 \tag{4.47}$$

gezeigt. Die Abschätzungen sind deutlich einfacher, wenn z_1 und Z_2 nicht mehr von $\widetilde{u}(t_n)$ abhängen. Sie werden bei der linearen Analyse deshalb im Beweis von Theorem 5.1 direkt bewiesen.

Satz 4.8 findet sich im linearen Fall in Theorem 5.1. Die Aussage ist identisch bis auf die Schrittweiteneinschränkung (siehe hierzu Bemerkung 4.10). Der Beweis unterscheidet sich jedoch vor allem in der Darstellung des globalen Fehlers, vergleiche dafür mit dem Abschnitt nach Satz 4.8, der die Motivation der Beweisidee von Satz 4.8 darstellt.

Insgesamt erkennt man im Vergleich mit Buchholz et al. (2018), dass die Aussage für den linearen Fall in der vorliegenden Arbeit enthalten ist. Die Darstellungen des lokalen Fehlers unterscheiden sich unwesentlich, die Beweisideen sind häufig identisch, wobei die Nichtlinearität von g an einigen Stellen zusätzliche Arbeit erzeugte. Nur für den globalen Fehler musste eine neue Darstellung gewählt werden. Jedoch konnte die Technik der partiellen Summation ebenfalls angewandt werden.

4.3 Einordnung der Analyse

Die folgende Analyse wurde bereits im Rahmen der Veröffentlichung des linearen Problems in Buchholz et al. (2018) durchgeführt und findet sich in dem Artikel in Abschnitt 6. Sie ist zur Vollständigkeit hier ebenfalls aufgeführt.

Die Entwicklung der trigonometrischen Integratoren wurde bereits in Kapitel 2.2 vorgestellt. Die in der vorliegenden Arbeit durchgeführt Fehleranalyse unterscheidet sich von diesen Arbeiten in ihrer Aussage, den Bedingungen an die Filterfunktionen und natürlich in der Beweistechnik, ausgenommen natürlich der Analyse des linearen Falls in Abschnitt 2.2.7. Die Aussage in Satz 4.11 unterscheidet sich darin, dass in den transformierten Geschwindigkeiten $v = \Omega^{-1}q'$ ebenfalls Konvergenz zweiter Ordnung in τ erzielt werden konnte, während für die Geschwindigkeiten q' bisher nur Konvergenz erster Ordnung gezeigt wurde (vergleiche Hairer und Lubich (2000) oder Grimm und Hochbruck (2006)). Für kleine Schrittweiten mit $\tau \|\Omega\| = O(1)$, lässt sich mit Satz 4.11 Konvergenz erster Ordnung in q' zeigen.

Weiterhin unterscheiden sich die Analysen in den Bedingungen, die an die Filterfunktionen gestellt werden. Der hier vorgestellte Integrator (3.10) entspricht bis auf Vertauschung der Reihenfolge des Splittings dem trigonometrischen Integrator aus Grimm und Hochbruck (2006), wenn man die Filterfunktionen ϕ, ψ, ψ_1, and ψ_0 in Grimm und Hochbruck (2006) oder Hairer et al. (2006) folgender Maßen wählt:

$$\phi = \phi, \qquad \psi_1 = \psi_S, \qquad \psi = \operatorname{sinc}(\cdot)\psi_S, \qquad \psi_0 = \cos(\cdot)\psi_S.$$

Man kann nun die Filterbedingungen $(11) - (16)$ in Grimm und Hochbruck (2006) (Bedingungen (2.33)-(2.39) aus Abschnitt 2.2.5) und (XIII.4.1) und (XIII.4.8) in Hairer et al. (2006) mit denen aus Annahme 4.1 vergleichen. Die Beschränktheit der Filterfunktionen (4.1b) findet sich ebenso in Bedingung (11) in Grimm und Hochbruck (2006). Aus den Bedingungen $(13) - (16)$ in Grimm und Hochbruck (2006) folgt sofort, dass die ϕ und $\psi_1 = \psi_S$ verschwinden müssen, wenn $x = 2k\pi$ mit $k \in \mathbb{Z}$ ist. Diese Bedingung findet man auch in (4.1d) der vorliegenden Arbeit.

Jedoch findet man in Grimm und Hochbruck (2006) keine Bedingung, die ein Abklingverhalten wie x^{-1} für $x \to \infty$ fordert, wie in Bedingung (4.1c). Bedingung (15) fordert vergleichbares für den Filter ψ. Geht man jedoch von Filterfunktionen aus, die ein symmetrisches Verfahren garantieren, so stellt dies keine zusätzliche Annahme an $\psi_1 = \psi_S$ dar.

Die Bedingungen (XIII.4.1) und (XIII.4.8) aus Hairer et al. (2006) implizieren die Bedingungen (4.1b)–(4.1d) der vorliegenden Arbeit. Es sei noch angemerkt, dass die gezeigten Bedingungen an die Filterfunktionen lediglich hinreichend aber nicht zwangsläufig notwendig sind.

Der größte Unterschied liegt jedoch in der Beweistechnik. Die Darstellung des lokalen Fehlers in Terme der Größenordnung $\mathcal{O}(\tau^2)$ und $\mathcal{O}(\tau^3)$ durch den Hauptsatz der Differential- und Integralrechung und den Kommutatoren verbindet die trigonometrischen Integratoren mit der Analyse von Splittingverfahren, siehe Kapitel 2.1. Weiterhin etabliert sich der Einsatz der partiellen Summation um den globalen Fehler abzuschätzen. Interessant ist auch die Zuhilfenahme der modifizierten Gleichung. Sie kann als ein τ^2-Störung der Gleichung (3.2a) verursacht durch die Filterfunktionen verstanden werden. Diese Störung erklärt auch, weshalb die Fehlerkonstante wächst, sobald die Schrittweiten den nichtsteifen Bereich ($\tau\omega_{\max} < 2$) erreicht haben, vergleiche dazu die Abbildungen der Simulationsfehler und ihre Beschreibungen in Kapitel 5.

KAPITEL 5

SIMULATION EINER SCHWINGERKETTE

Im folgenden Kapitel wird ein bekanntes Testproblem verwendet, um die Effizienz des in der vorliegenden Arbeit untersuchten Splittingverfahrens und des verwendten trigonometrischen Integrators im Beispiel zu demonstrieren. Alle Simulationen in der vorliegenden Arbeit wurden mit MATLAB (The Mathworks, Version R2018a) durchgeführt.

5.1 Lineare Schwingungskette

Folgendes Beispielproblem stammt aus García-Archilla et al. (1999) und wurde dort verwendet, um eine Stabilitätsanalyse des Impulsverfahrens durchzuführen.

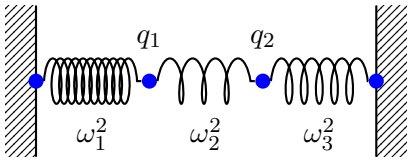

Abbildung 5.1: Kette aus linearen Federn mit verschiedenen Steifigkeiten

Man betrachtet eine Schwingerkette mit zwei Massepunkten, die an den Rändern fest eingespannt ist, siehe Abbildung 5.1. Die äußeren Federn sind steif mit Federsteifigkeit $\omega_1^2 = 1000$ und $\omega_3^2 = 2$. Die verbleibende, mittlere Feder ist weich mit Federsteifigkeit $\omega_2^2 = \frac{1}{2}$. Alle drei Federn haben die Länge eins und die Massepunkte haben Masse eins. Es wirken weiterhin keine zusätzlichen Kräfte auf das System.

Bezeichnet man mit q_1 bzw. q_2 die Verschiebung des ersten bzw. zweiten Masspunktes aus der Ruhelage, so werden die Bewegungen durch die Lösung $q(t) = [q_1(t), q_2(t)]^T$ der Differentialgleichung (1.1) mit $t_{\text{end}} = 1$,

$$\Omega^2 = \begin{bmatrix} \omega_1^2 & 0 \\ 0 & \omega_3^2 \end{bmatrix}, \qquad \text{und} \qquad g(q) = Gq, \qquad \text{mit } G = \omega_2^2 \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix},$$

und Anfangswerten

$$q(0) = \begin{bmatrix} \omega_1^{-1} \\ \omega_2^{-1} \end{bmatrix} \qquad\qquad q'(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

beschrieben. Die Hamiltonfunktion des Systems (1.1) ist gegeben durch

$$\mathcal{H}(q,p) = \frac{1}{2}\left\|\Omega q\right\|^2 + \frac{1}{2}\left\|p\right\|^2 - \frac{1}{2}q^T G q.$$

Diese entspricht physikalisch der Gesamtenergie des Systems. Die Lösung von (1.1) erhält $\mathcal{H}$. Es gilt also

$$\mathcal{H}(q(t), q'(t)) \equiv \mathcal{H}_0 = \mathcal{H}(q(0), q'(0)), \qquad \text{für alle } t \geq 0.$$

Da die Matrix G negativ semidefinit ist, folgt daraus direkt die Finite-Energie-Bedingung

$$\left\|\Omega q(t)\right\|^2 + \left\|q'(t)\right\|^2 \leq \mathcal{H}_0.$$

Damit sind alle Bedingungen aus der Annahme 1.1 erfüllt.

5.1.1 Numerische Simulation

Zur Simulation der Lösung wird das System zuerst, wie in Kapitel 3 beschrieben, transformiert und schließlich ein Strang-Splittingverfahren (3.10) auf die modifizierte Gleichung angewandt. Es werden hier die Filterfunktionen

$$\phi(\xi) = \operatorname{sinc}(\xi) \qquad \text{und} \qquad \psi_S(\xi) = \operatorname{sinc}(\xi)$$

verwendet, welche die Annahme 4.1 erfüllen. Es wird nun die numerische Lösung mit 1000 verschiedenen Schrittweiten zwischen $[10^{-3}, 10^{-1}]$ berechnet. Im Gegensatz zu glatten Problemen genügt es hier nicht, den Fehler am Endzeitpunkt zu berechnen, da sich der numerische Fehler ebenfalls oszillatorisch verhält. Daher wird hier der Fehler nach jedem Zeitschritt berechnet und das Maximum über alle Zeitschritte bei gleicher Schrittweite gebildet.

Die Abbildungen 5.2 und 5.3 zeigen den Fehler des Splittingverfahrens, einmal in den Positionen q (oben) und einmal für die transformierten Geschwindigkeiten $v = \Omega^{-1}q'$ (unten). Die gestrichelte blaue Linie beschreibt dabei den Fehler für das Splittingverfahren ohne Filterfunktionen. Der Fehler des Splittingverfahrens mit Filterfunktionen ist durch die orangefarbene, durchgezogene Linie dargestellt. Die lilafarbene, gestrichpunktete Linie hat Steigung zwei und dient somit als Referenz, um die Ordnung des Verfahrens zu veranschaulichen.

Sowohl in den Positionen q, als auch in den transformierten Geschwindigkeiten v erkennt man deutlich die Ordnungsreduktion des Verfahrens ohne Filterfunktionen auf Ordnung Null für gewisse große Schrittweiten. Man sieht auch die Konvergenz zweiter Ordnung des Verfahrens ohne Filterfunktionen im nichtsteifen Bereich (also für kleine Schrittweiten). Das Verfahren mit Filterfunktionen zeigt demgegenüber Konvergenz zweiter Ordnung für jegliche Schrittweiten. Kleinere harmlosere Oszillationen des Fehlers verbleiben von numerischen Resonanzen höherer Ordnung. Sie sind aber beschränkt und führen so zu keiner Ordnungsreduktion. Solche Oszillationen traten ebenfalls bei der Simulation mit dem trigonometrischen Integrator aus Grimm und Hochbruck (2006) auf.

Weiterhin steigt die Fehlerkonstante für das Verfahren mit Filterfunktionen an, wenn die Schrittweiten den nichtsteifen Bereich erreichen. Dies lässt sich dadurch erklären, dass man die Filter als eine Störung eines Verfahrens der Ordnung zwei auffassen kann (vergleiche die modifizierte Gleichung aus Abschnitt 3.1). Dass eine solche Störung den Fehler vergrößert und dieser deshalb für kleine

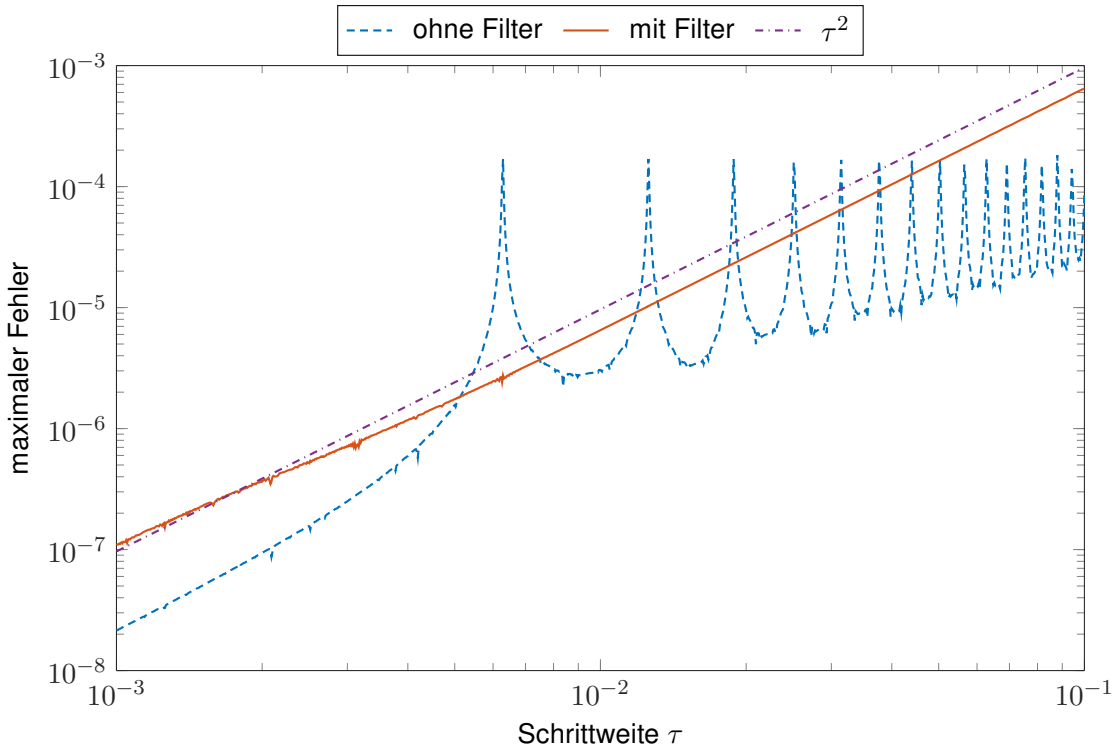

Abbildung 5.2: Fehler des Strang-Splittingverfahrens in den Positionen q

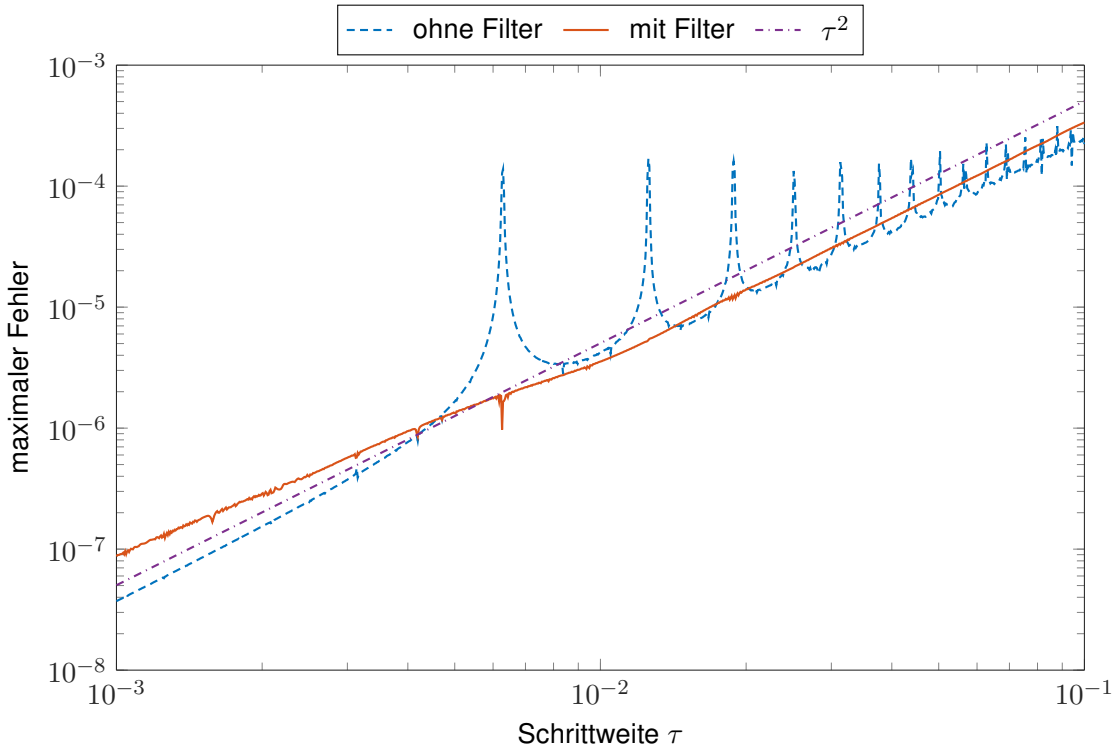

Abbildung 5.3: Fehler des Strang-Splittingverfahrens in den transformierten Geschwindigkeiten $v = \Omega^{-1} q'$

Schrittweiten über der Fehlerkurve des Verfahrens ohne Filter liegt, ist nicht verwunderlich, denn durch die Störung muss ein zusätzlicher Fehlerterm berücksichtigt werden.

Die Abbildung 5.4 zeigt die Hamiltonfunktion und die Abbildung 5.5 die Finite-Energie der numerischen Approximation zu jedem berechneten Zeitschritt. Das Verfahren ohne Filterfunktionen wird dabei mit der blauen gestrichelten und das Verfahren mit Filterfunktionen mit der orangefarbene durchgezogenen Linie dargestellt.

Wie oben beschrieben ist die Hamiltonfunktion der exakten Lösung konstant und hat einen Wert von $\mathcal{H}_0 = 1.0628$. Das Splittingverfahren mit Filterfunktionen approximiert diesen Wert bis auf kleine Schwingungen sehr gut. Wird jedoch das Verfahren ohne Filterfunktionen verwendet, so fällt die Energie für große Schrittweiten ab. Die Energieerhaltung wird erst für kleinere Schrittweiten gewährleistet.

Wie erwartet, ist die Finite-Energie nicht konstant, jedoch ist sie stets kleiner als $\mathcal{H}_0$. Auch hier erkennt man, dass die Finite-Energie durch das Verfahren ohne Filterfunktionen für große, resonante Schrittweiten nicht gut approximiert wird.

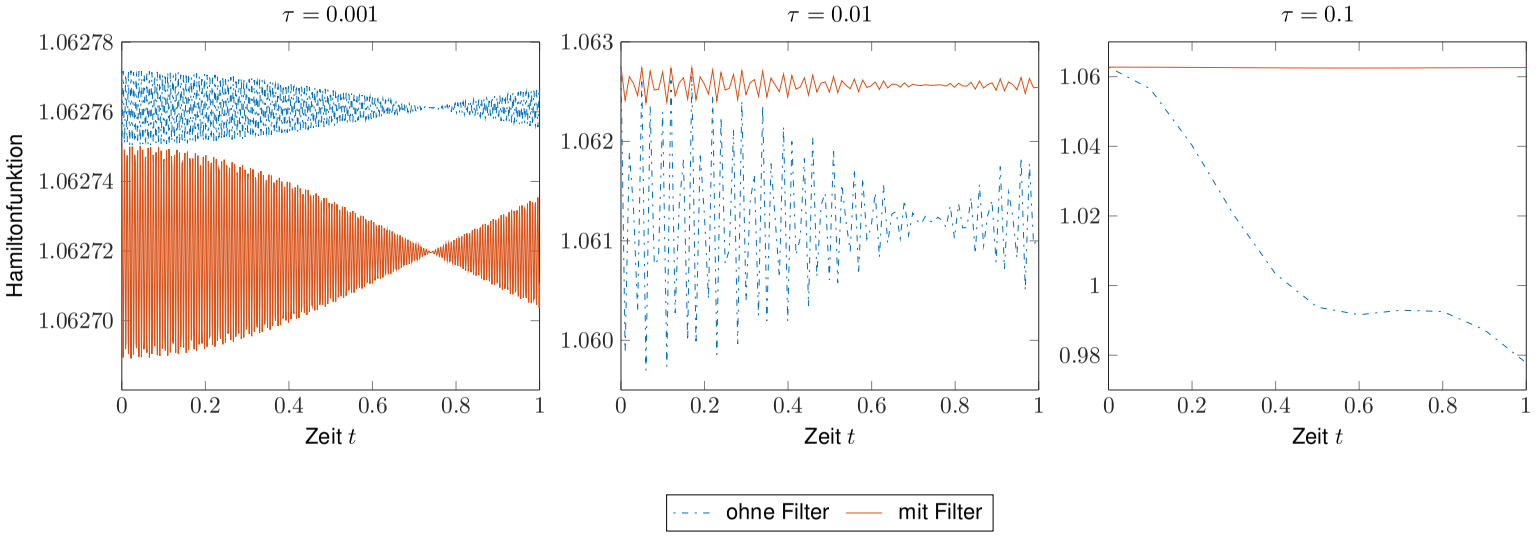

Abbildung 5.4: Abbildung der Hamiltonfunktion $\mathcal{H}(q_n, q'_n)$ für die Iterierten q_n und q'_n des Strang-Splittingverfahrens

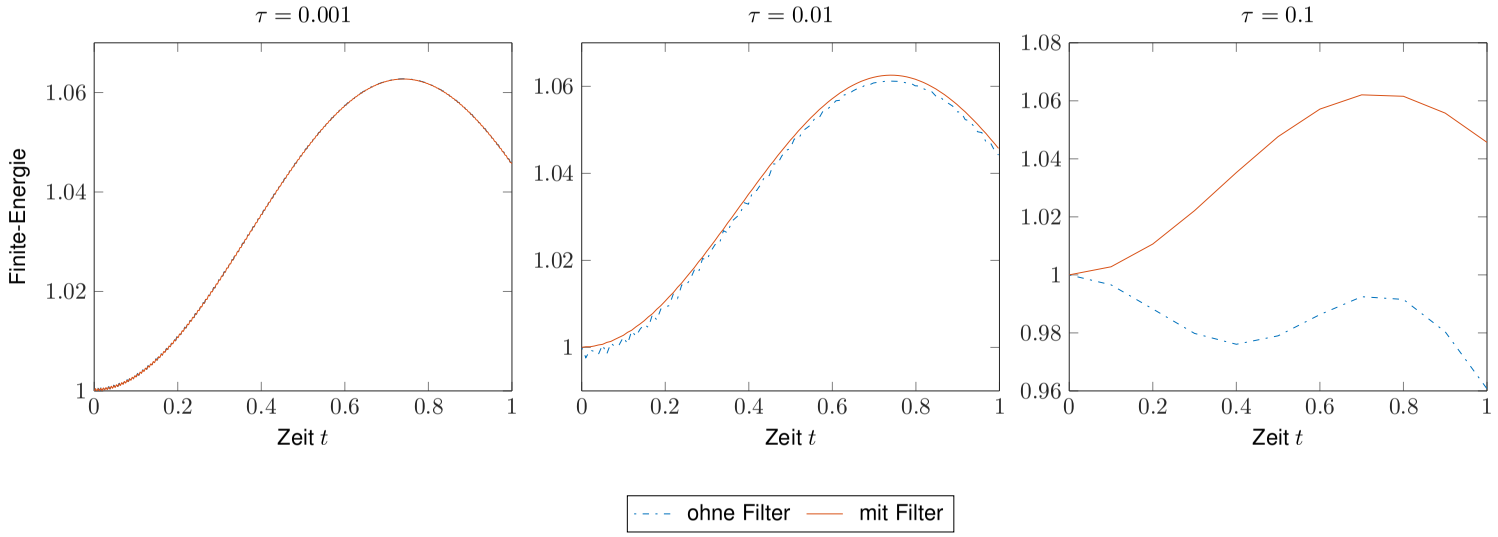

Abbildung 5.5: Abbildung der Finite-Energie $\|\Omega q_n\|^2 + \|q'_n\|^2$ für die Iterierten q_n und q'_n des Strang-Splittingverfahrens

5.2 Das Fermi-Pasta-Ulam-Tsingou-Problem

Das Fermi-Pasta-Ulam-Tsingou-Problem (FPUT) wurde 1955 in Fermi, Ulam und Pasta (1955) veröffentlicht. Es gilt heute als Meilenstein in der Entwicklung der Computersimulationen und ist der Standardtest für Probleme mit mehreren Zeitskalen. So ist es auch Beispiel in fast allen in Kapitel 2.2 erwähnten Artikeln. Es lassen sich mithilfe des FPUT-Problems sowohl algorithmische Instabilitäten als auch numerische Resonanzen erkennen, vergleiche Kapitel 1.2. Häufig wird es auch unter dem Namen Fermi-Pasta-Ulam-Problem referenziert und damit die Arbeit der Mathematikerin Mary Tsingou unterschlagen. Ihre Verdienste würdigte Thierry Dauxois in Dauxois (2008).

Das Modell stammt aus der Mechanik und demonstriert das Schwingungsverhalten einer Kette mit $2m$ Massepunkten. Diese sind mit alternierend weichen, kubisch nichtlinearen und steifen, linearen Federn verbunden. Lineare Federn sind dabei solche, bei denen die Federkraft linear von der Auslenkung der Feder abhängt. Die Federkraft nichtlinearer Federn hängt im Gegensatz dazu nichtlinear, in diesem Fall kubisch, von der Auslenkung der Feder ab. An den Endpunkten ist die Kette fest eingespannt. Das Modell ist in Abbildung 5.6 dargestellt.

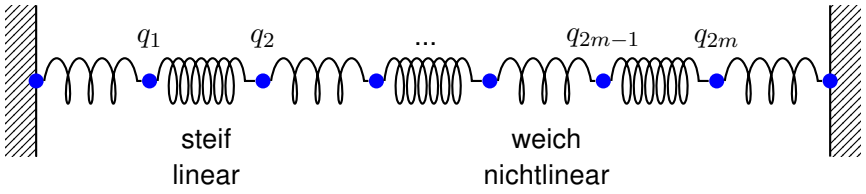

Abbildung 5.6: Kette aus alternierenden weichen nichtlinearen und steifen linearen Federn

Das FPUT-Problem ist eins der ersten dokumentieren Computerexperimente und wurde erstmals auf dem MANIAC I durchgeführt. Dieses einfache Modell zeigte in der Simulation ein höchst unerwartetes dynamisches Verhalten. Während sich vergleichbare, lineare Systeme periodisch verhalten, erwarteten die Autoren hier, dass man durch die Nichtlinearität der Federn ein ergodisches System erhält (das heißt, das System erreicht alle möglichen Zustände). Die Bewegungen waren jedoch quasi-periodisch. Die Simulation liefert damit einen wichtigen Beitrag für die Chaosforschung und gilt heute als Pionierarbeit. Sie zeigt auch den Wert der Computersimulation bei der Analyse von physikalischen Systemen.

5.3 Modellbildung

In diesem Kapitel wird nun das FPUT-Problem mit kubischer Nichtlinearität untersucht. Die Modellierung stammt aus Hairer et al. (2006), Kapitel I.5. Bezeichnet man mit $q_1, \ldots, q_{2m}$ die Verschiebung der Massepunkte aus ihrer Ruhelage und mit $p_i = q_i'$ deren Geschwindigkeiten, so wird die Bewegung mithilfe der Hamiltonfunktion

$$\widetilde{\mathcal{H}}(p,q) = \frac{1}{2}\sum_{i=1}^{m}(p_{2i-1}^2 + p_{2i}^2) + \frac{\omega^2}{4}\sum_{i=1}^{m}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{m}(q_{2i+1} - q_{2i})^4,$$

beschrieben, wobei ω der Federsteifigkeit der steifen, linearen Federn entspricht und dementsprechend groß sein soll. Für die hier vorgestellte Simulation wird $\omega = 1000$ gewählt. Es wird nun eine

Variablentransformation durchgeführt, sodass man später eine Diagonalmatrix für Ω erhält. Es sei

$$x_{0,i} = (q_{2i} + q_{2i-1})/\sqrt{2}, \qquad x_{1,i} = (q_{2i} - q_{2i-1})/\sqrt{2},$$
$$y_{0,i} = (p_{2i} + p_{2i-1})/\sqrt{2}, \qquad y_{1,i} = (p_{2i} - p_{2i-1})/\sqrt{2}.$$

Dabei stellt $x_{0,i}$ die skalierte Verschiebung und $x_{1,i}$ die skalierte Dehnung der i-ten steifen Feder und $y_{0,i}$, $y_{1,i}$ ihre jeweiligen Geschwindigkeiten dar. Man erhält in den neuen Variablen die folgende Hamiltonfunktion

$$\mathcal{H}(y,x) = \frac{1}{2}\sum_{i=1}^{m}(y_{0,i}^2 + y_{1,i}^2) + \frac{\omega^2}{2}\sum_{i=1}^{m} x_{1,i}^2 + \frac{1}{4}\Big((x_{0,1} - x_{1,1})^4 + \sum_{i=1}^{m-1}(x_{0,i+1} - x_{1,i+1} - x_{0,i} - x_{1,i})^4 + (x_{0,m} + x_{1,m})^4\Big).$$

Wie beim linearen Problem beschreibt die Hamiltonfunktion physikalisch die Gesamtenergie des Systems und wird von der exakten Lösung des Hamiltonsystems erhalten.

Aus der Hamiltonfunktion lassen sich die Bewegungsgleichungen ableiten. Es ergibt sich dabei zunächst eine Differentialgleichung der Form

$$x'' = -\widetilde{\Omega}^2 x + \tilde{g}(x), \qquad x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

mit Vektoren $x_j = [x_{j,1}, \ldots, x_{j,m}]^T,\ j \in 0, 1$, einer kubischen Nichtlinearität $\widetilde{g}$ und einer Matrix $\widetilde{\Omega}$ mit

$$\widetilde{\Omega} = \begin{bmatrix} 0 & 0 \\ 0 & \omega I \end{bmatrix},$$

mit Eigenwerten ω und Null. Um Annahme 3.1 zu erfüllen, muss Ω positiv definit sein. Hier wird nun ein Shift verwendet, um eine äquivalente Differentialgleichung mit positiv definiter Matrix Ω zu erhalten:

$$\Omega = \tilde{\Omega} + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \qquad g(x) = \tilde{g}(x) + \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \tag{5.1}$$

Die Matrix Ω erfüllt nun die Annahme 1.1 und 3.1, jedoch nicht die Funktion g. Sie selbst und ihre Ableitungen sind zwar auf Kugeln beschränkt, nur ist sie nicht Lipschitz-stetig. Weiterhin ist die Finite-Energie-Bedingung im ursprünglichen System beschränkt,

$$\frac{1}{2}\left\|\widetilde{\Omega}\begin{bmatrix} x_0 \\ x_1 \end{bmatrix}\right\|^2 + \frac{1}{2}\left\|\begin{bmatrix} y_0 \\ y_1 \end{bmatrix}\right\|^2 \leq \mathcal{H}(y(0), x(0)),$$

da alle auftretenden Größen auch in der Hamiltonfunktion $\mathcal{H}$ enthalten sind, die verbleibende Größe nichtnegativ ist, und die Hamiltonfunktion konstant ist. Betrachtet man jedoch die Finite-Energie im geshifteten System

$$\mathcal{F}(y,x) = \frac{1}{2}\left\|\Omega\begin{bmatrix} x_0 \\ x_1 \end{bmatrix}\right\|^2 + \frac{1}{2}\left\|\begin{bmatrix} y_0 \\ y_1 \end{bmatrix}\right\|^2,$$

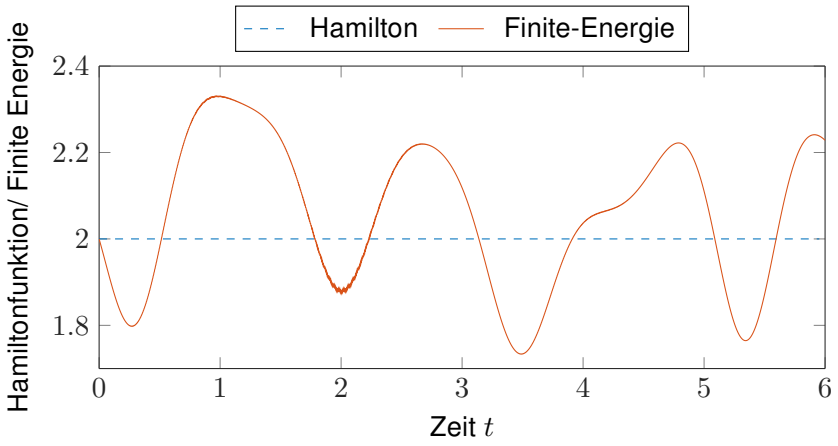

Abbildung 5.7: Hamiltonfunktion $\mathcal{H}(y_n, x_n)$ und der Finiten-Energie $\mathcal{F}(y_n, x_n)$ für die Iterierten der Referenzlösung

so tritt wegen des Shiftes zusätzlich $\|x_0\|^2$ auf. Im geshifteten System kann die Finite-Energie nicht mehr durch die Erhaltung der Hamiltonfunktion beschränkt werden. Jedoch lässt sich analytisch die Wohlgestelltheit der Gleichung (1.1) mit Ω und g aus (5.1) und damit die Beschränktheit der Lösung x und ihrer Ableitung y im endlichen Existenzintervall $[0, t_{\text{end}}]$ nachweisen. Verwendet man nun die Variation-der-Konstanten-Formel ergibt sich die Darstellung

$$x(t) = \cos(t\Omega)x(0) + \Omega^{-1}\sin(t\Omega)x'(0) + \int_0^t \Omega^{-1}\sin((t-s)\Omega)g(x(s))\,ds.$$

Mithilfe dieser Gleichung lässt sich nun die Beschränktheit der Finiten-Energie im geshifteten System für endliche Zeitintervalle $[0, t_{\text{end}}]$ ableiten. Um diese Aussage zu bekräftigen, enthält die Abbildung 5.7 eine Darstellung der Hamiltonfunktion $\mathcal{H}(y_n, x_n)$ und der Finiten-Energie $\mathcal{F}(y_n, x_n)$ für die Iterierten der Referenzlösung. Hier erkennt man den Erhalt der Hamiltonfunktion über das gesamte Zeitintervall. Die Finite-Energie ist nicht konstant, jedoch beschränkt.

Es wird im Folgenden ein System mit sechs Massepunkten simuliert ($m = 3$). Als Anfangswerte werden

$$x_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad x_1 = \begin{bmatrix} \omega^{-1} \\ 0 \\ 0 \end{bmatrix}, \qquad y_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad y_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

gewählt. Das zeitliche Simulationsintervall sei $[0, 6]$.

5.3.1 Numerische Simulation

Wie bereits im linearen Fall wendet man das Strang-Splittingverfahren (3.10) auf die modifizierte Gleichung mit den Filterfunktionen

$$\phi(\xi) = \operatorname{sinc}(\xi) \qquad \text{und} \qquad \psi_S(\xi) = \operatorname{sinc}(\xi)$$

an. Konnte man für das lineare Problem aus Abschnitt 5.1 die exakte Lösung durch die Matrixexponentialfunktion ausdrücken und damit die Referenzlösung in MATLAB mittels der Matlabfunktion *expm* beschaffen, so wird zur Fehlerschätzung im nichtlinearen Fall eine Referenzlösung mit dem

Störmer-Verlet-Verfahren berechnet, vergleiche Kapitel 1.1.1. Wie in Abschnitt 1.2 gesehen, muss die Schrittweite dieser Referenzlösung sehr klein gewählt werden, um die Stabilitätsbedingung des Verfahrens einzuhalten. Weiterhin soll trotzdem der maximale Fehler des Strang-Splittingverfahrens über alle Zeitpunkte berechnet werden. Wählt man also eine Schrittweite τ aus, so muss die Referenzlösung an allen Zeitpunkten $t_k = k\tau$ mit $t_k \leq t_{\text{end}}$ vorliegen. Es sollen zusätzlich möglichst viele verschiedene Schrittweiten τ betrachtet werden, um eine gute Auflösung möglicher Resonanzen zu erreichen, während möglichst wenig Aufwand in die Berechnung der Referenzlösungen investiert werden sollen.

Man kann dies gewährleisten, indem man sich zuerst eine hochzusammengesetzte Zahl N_{max} („highly composite number“) wählt, also eine Zahl, die mehr Teiler besitzt als alle kleineren Zahlen. Alle Teiler N dieser Zahl kann man als Schrittweiten t_{end}/N verwenden. Für die Schrittweite τ_{ref} der Referenzlösung wählt man sich einen geeigneten Skalierungsfaktor c und setzt

$$\tau_{\text{ref}} = \frac{t_{\text{end}}}{cN_{\text{max}}},$$

so dass τ_{ref} der Stabilitätsbedingung des Störmer-Verlet-Verfahrens genügt. Der Fehler der Referenzlösung ist für c hinreichend groß deutlich kleiner, als der Fehler des Verfahrens. Dies schränkt die Auswahl der Schrittweiten τ trotzdem stark ein. Man kann die Anzahl an möglichen Schrittweiten vergrößern, indem man weitere Referenzlösungen zu leicht modifizierten Endzeiten $\tilde{t}_{\text{end}}$ berechnet, wobei $\tilde{t}_{\text{end}} - t_{\text{end}}$ klein sein sollte.

Für diese Simulation wurden fünf verschiedene Referenzlösungen mit $t_{\text{end}} = 5.99, 5.999, 6, 6.001, 6.01$ mit einer Referenzschrittweite $\tau_{\text{ref}} < 2.3 \cdot 10^{-7}$ berechnet. Während für jede einzelne Simulation jeweils 96 oder 97 mögliche Schrittweiten in $[10^{-3}, 10^{-1}]$ berechnet wurden, so konnten in Kombination der Ergebnisse 483 verschiedene Schrittweiten genutzt werden.

Die Abbildungen 5.8 und 5.9 zeigen den Fehler des Strang-Splittingverfahrens (3.10) in x und den transformierten Variablen $v = \Omega^{-1}y$. Die blaue gestrichelte Linie entspricht dem Fehler des Verfahrens ohne Filterfunktionen, die orangefarbene, durchgezogene Linie dem Fehler des Verfahrens mit Filterfunktionen. Auch hier lassen sich deutliche Resonanzen im Fehler ohne Filterfunktionen ausmachen. Der Integrator mit Filterfunktionen zeigt, wie erwartet, Konvergenz der Ordnung zwei für alle Schrittweiten. Wie auch im linearen Beispiel vergrößert sich die Fehlerkonstante, sobald die Schrittweiten den nicht-steifen Bereich erreichen. Interessanterweise ist der Fehler des Verfahrens mit Filterfunktionen in den transformierten Geschwindigkeiten auch im nicht-steifen Bereich kleiner als der Fehler ohne Filterfunktionen.

Die Finite-Energie $\mathcal{F}(y_n, x_n)$ und die Hamiltonfunktion $\mathcal{H}(y_n, x_n)$ der Iterierten x_n und y_n des Verfahrens sind in den Abbildungen 5.10 und 5.11 dargestellt. Das Verfahren mit Filterfunktionen zeigt dabei für alle Schrittweiten eine gute Approximation an die Referenzlösung, vergleiche mit der Darstellung der Hamiltonfunktion und der Finiten-Energie der Iterierten der Referenzlösung in Abbildung 5.7, während das Verfahren ohne Filterfunktionen an den Resonanzstellen deutliche Abweichungen erkennen lässt.

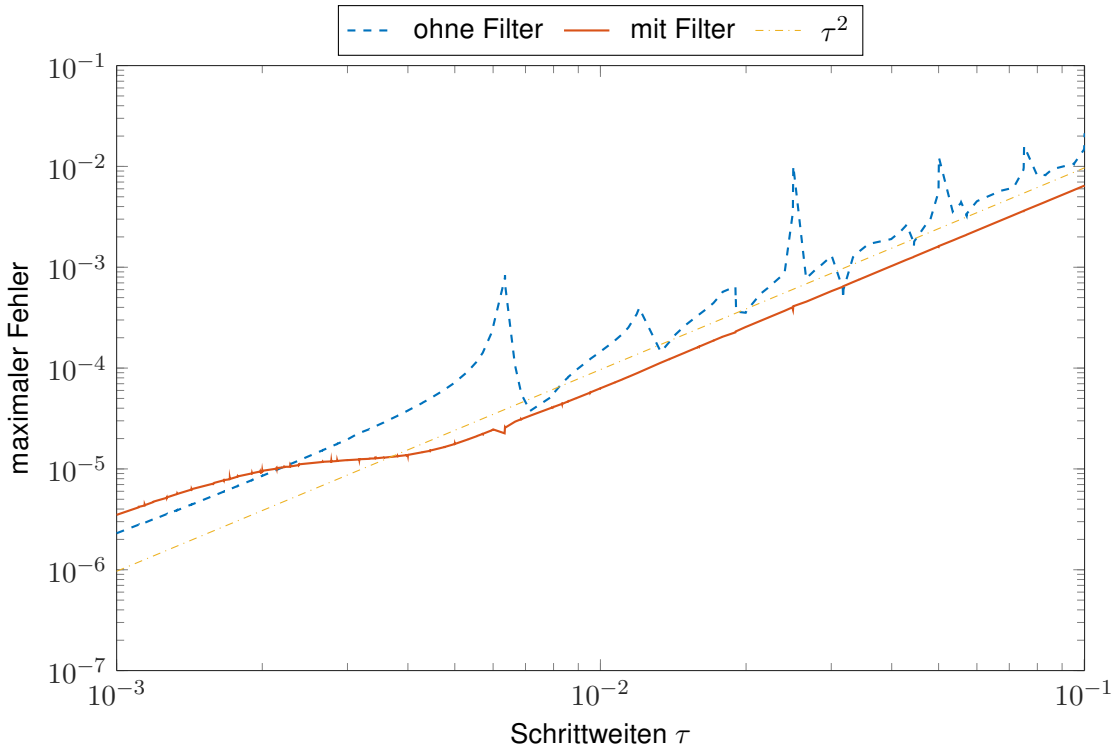

Abbildung 5.8: Fehler des Strang-Splittingverfahrens in x

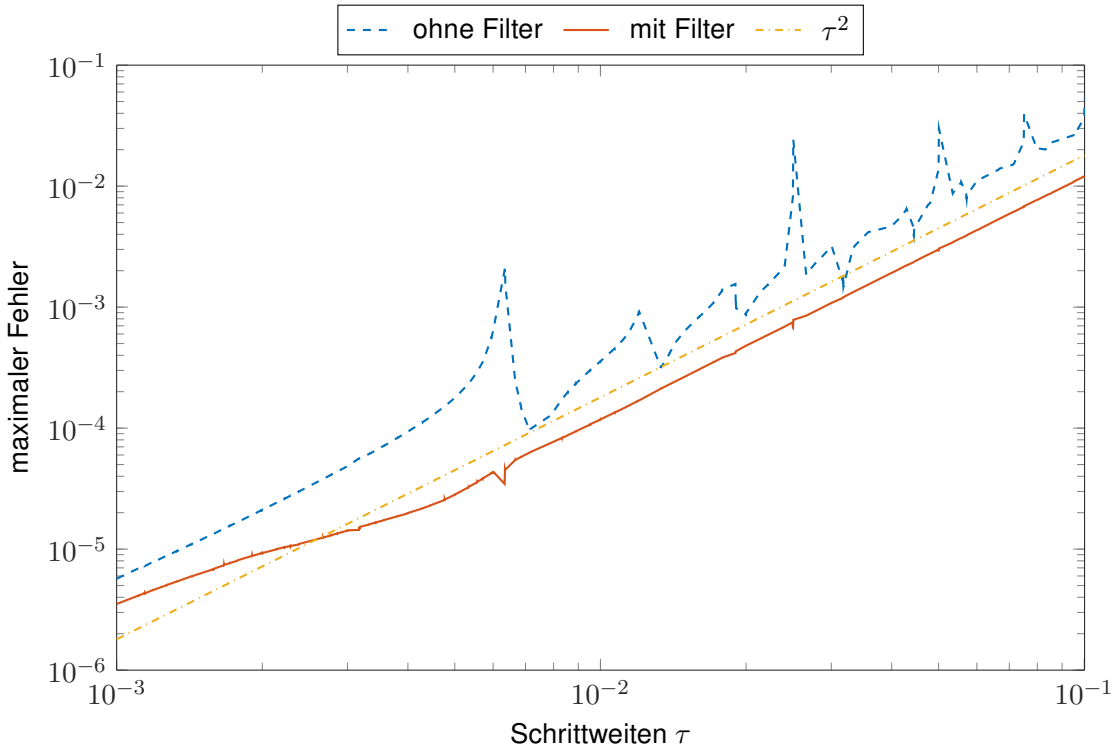

Abbildung 5.9: Fehler des Strang-Splittingverfahrens in den transformierten Variablen $v = \Omega^{-1} y$

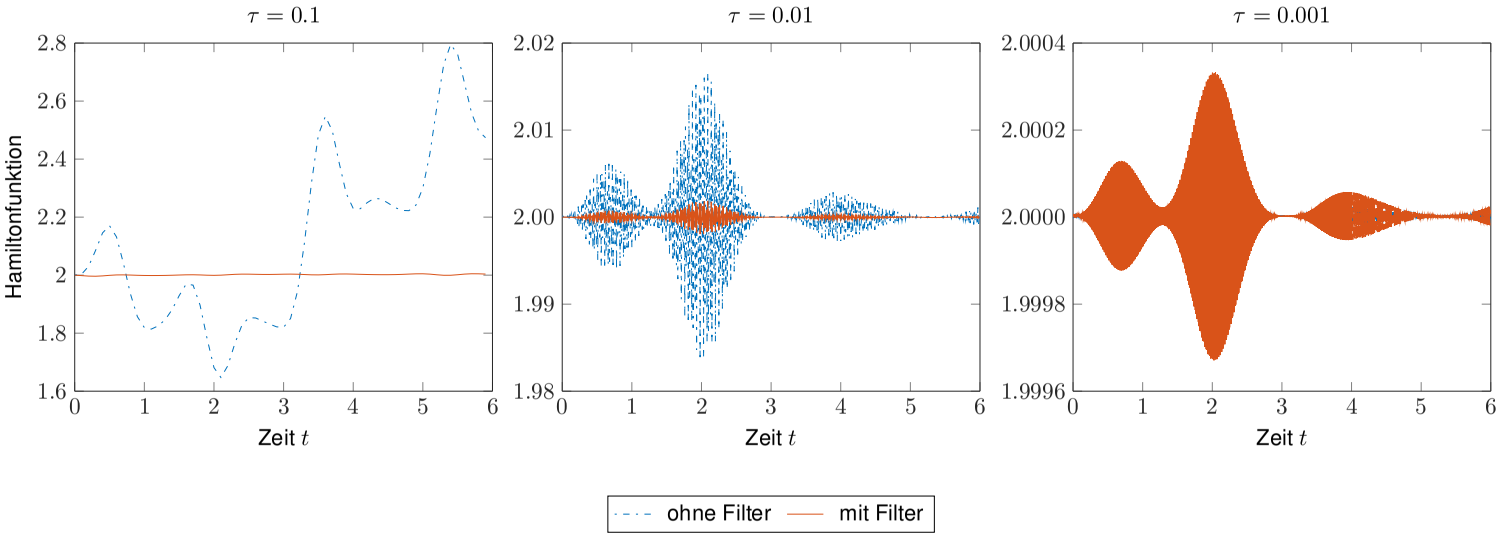

Abbildung 5.10: Abbildung der Hamiltonfunktion $\mathcal{H}(y_n, x_n)$ der Iterierten x_n und y_n des Strang-Splittingverfahrens

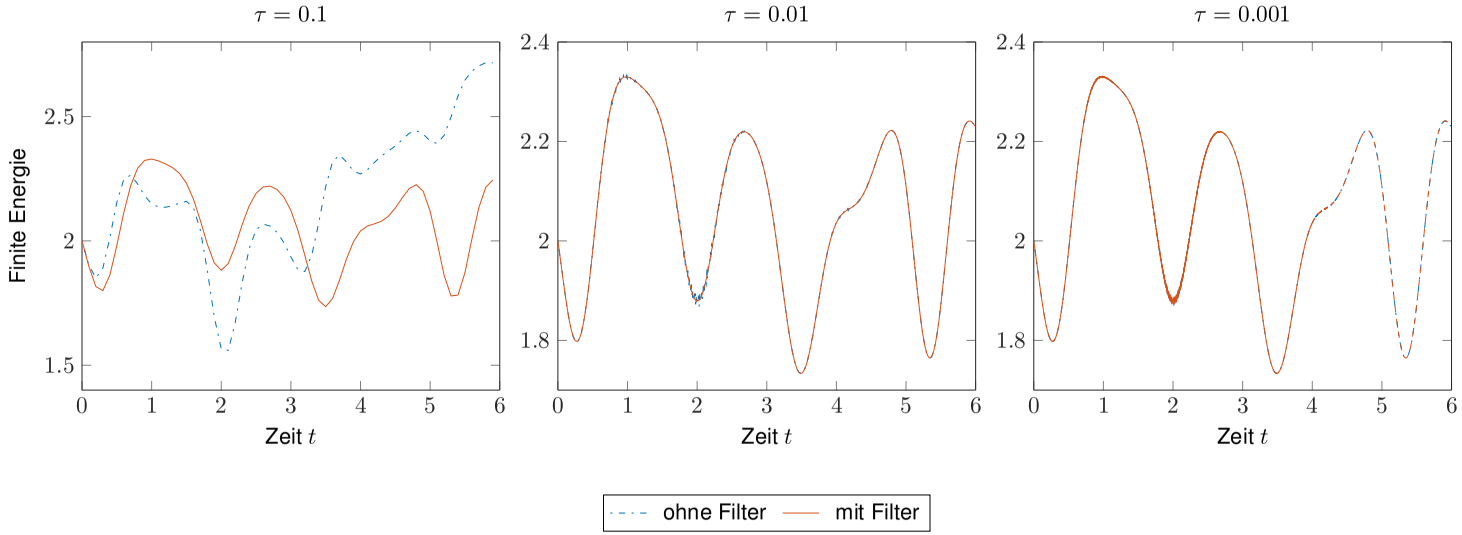

Abbildung 5.11: Abbildung der Finiten-Energie $\mathcal{F}(y_n, x_n)$ der Iterierten x_n und y_n des Strang-Splittingverfahrens

KAPITEL 6

SIMULATION EINER LASER-PLASMA-INTERAKTION

Laser-Plasma-Interaktionen beschreiben die Wechselwirkungen zwischen einem Laserpuls und Materie. Die Materie wird dabei stark erhitzt und verwandelt sich in ein Plasma. Man bezeichnet mit Plasma einen Zustand, in dem sich Elektronen aus den ansonsten neutralen Atomen lösen. Deshalb enthält ein Plasma freie Ladungsträger wie Ionen und Elektronen. Die Technologie zur starken Beschleunigung von Ionen mittels hochintensiver Laserpulse mit stark reduzierter Energieverteilung ist noch recht neu und wurde erstmals in Hegelich et al. (2006) vorgestellt.

Solche Interaktionen haben vielerlei Anwendungen. Beispielsweise können sie den klassischen Teilchenbescheuniger ersetzen, der in Naturwissenschaften und Medizin zahlreich verwendet wird. In der Medizin lassen sich damit Tumore gezielt entfernen, ohne umliegendes Gewebe zu beschädigen, siehe dazu Linz und Alonso (2016). Außerdem ermöglichen Laser-Plasma-Beschleuniger neue Ansätze in der Fusionsenergie, siehe beispielsweise Kirkwood et al. (2013).

Hier sollen nun Experimente betrachtet werden, bei denen die Plasmadichte sehr groß wird. Dies tritt beispielsweise auf, wenn Elektronen durch feste Objekte wie dünne Folien geschossen werden. Dies wird TNSA (Target Normal Sheath Acceleration) genannt (siehe Hegelich et al. (2002), Robson et al. (2006)). Hierbei treten Plasmadichten von $\rho_E = 100 - 1000\rho_c$ auf, die von der Frequenz des Lasers ω_p abhängen. Dabei bezeichnet ρ_c die kritische Plasmadichte, bei welcher der Laserpuls vom Plasma reflektiert wird. Eine weitere wichtige Anwendung ist das Konzept der schnellen Entzündung FI (Fast Ignition) in der Trägheitsfusion (siehe Tabak et al. (1994)). Hierbei werden Dichten von bis zu $\rho_E = 10.000\rho_c$ erreicht.

In Düsseldorf, Jena und München startete 2004 der Sonderforschungsbereich TR18 „Relativistische Laser-Plasma-Dynamik“, der die Physik ultra-intensiver Laserinteraktionen mit Materie erforschen sollte. Das Teilprojekt B5 „Fortgeschrittene numerische Methoden zur Simulation der Wechselwirkungen relativistischer Kurzpuls-Laser mit hochdichten Plasmen“ von Prof. Dr. A. Pukhov und Prof. Dr. M. Hochbruck beschäftigte sich insbesondere damit, die Laser-Plasma-Interaktionen numerisch zu simulieren, um aufwendige experimentelle Tests zu vermeiden. Hier wurde der C++ Code H-VLPL, Hybrid Virtual Laser Plasma Laboratory, entwickelt, dessen Verhalten im Folgenden analysiert werden soll.

Ein Laser wird durch ein elektromagnetisches Feld simuliert, welches durch die Maxwell-Gleichungen modelliert wird. Die Plasmateilchen werden meist kinetisch durch Bewegungsgleichungen beschrieben. Jedoch enthält ein Plasma eine große Anzahl von Teilchen, weshalb man dazu überging, eine Wolke aus Teilchen als sogenanntes Makroteilchen zu simulieren. Die Bewegung aller Teilchen im Makroteilchen werden als eine Bewegung des Makroteilchens simuliert und die Wechselwirkungen unter den Teilchen in einem Makroteilchen vernachlässigt. Solche Simulationen bezeichnet man als PIC (Particle-in-Cell), siehe Birdsall und Langdon (1991) oder Hockney und Eastwood (1992). Eine Vielzahl Codes für dreidimensionale, parallele Simulationen mittels PIC sind bereits vorhanden, darunter unter anderen OSIRIS, VORPAL und OOPIC. In diesem Beispiel wird der Code VLPL, Virtual Laser Plasma Laboratory, aus Pukhov (1999) betrachtet.

Generell gilt PIC als sehr detailreich, jedoch wird die Simulation sehr aufwendig, wenn man große Plasmadichten betrachtet. Hier müssen entweder die Makroteilchen sehr groß gewählt werden, wodurch die Simulation ungenau wird, oder es müssen sehr viele kleine Makroteilchen simuliert werden, was letztlich wie beim Code VLPL auf eine Schrittweiteneinschränkung des Zeitschritts τ im Sinne von

$$\tau \leq \frac{2}{\omega_p}$$

führt, um die Stabilität des Verfahrens zu gewährleisten. Neben PIC gibt es jedoch eine weitere Möglichkeit, Plasma zu beschreiben: die hydrodynamische Modellierung. Die Teilchen werden dabei wie eine Flüssigkeit behandelt. Mathematisch modelliert man das Plasma mithilfe der Boltzmann-Vlasov-Gleichungen. In dem oben beschriebenen Fall ist es sinnvoll, beide Modelle zu einem hybriden Modell zu kombinieren. Dabei wird heißes Plasma mit niedrigen Dichten über PIC simuliert, während kaltes Plasma mit hohen Dichten über das hydrodynamische Modell beschrieben wird. Solche hybriden Modelle findet man beispielsweise in Mason (1980) oder Gremillet, Bonnaud und Amiranoff (2002).

Wissenschaftlern im Teilprojekt B5 am SFB TR18 gelang die Umsetzung eines solchen hybriden Ansatzes erstmals in Liljo, Karmakar, Pukhov und Hochbruck (2008) in einer Raumdimension. Der entstandene Code wurde H-VLPL genannt. Der Code ist unabhängig von der Plasmadichte stabil, da hier ein implizites Verfahren im Umgang mit der hohen Plasmadichte eingesetzt wurde, um Stabilitätseinschränkungen zu vermeiden.

In einem weiteren Schritt versuchte man, einen Code für Simulationen in drei Raumdimensionen zu entwickeln. Dabei ließ sich der Ansatz aus dem 1D-Code nicht direkt übertragen, da die implizite Berechnung in 3D zu aufwändig ist, so dass sie keine Vorteile gegenüber einer expliziten Simulation mit PIC-Codes liefert. Daher wurde schlussendlich ein neues Verfahren für die Zeitintegration entwickelt. Es handelt sich dabei um ein explizites Dreifach-Splittingverfahren, bei dem analog zur gemittelten Impulsmethode von García-Archilla et al. (1999), siehe Abschnitt 2.2.3, oder dem in dieser Arbeit vorgestelltem Splittingverfahren Filterfunktionen verwendet werden, um auftretende Resonanzen zu vermeiden, wie sie in Abschnitt 1.2 der vorliegenden Arbeit beschrieben werden. Das entstandene Verfahren wurde in Tückmantel, Pukhov, Liljo und Hochbruck (2010) vorgestellt. Beide Simulationen sind außerdem Thema in den Dissertationen von Liljo (2010) und Tückmantel (2012).

Obwohl das Dreifach-Splitting mit den eingeführten Filterfunktionen in Experimenten gute Ergebnisse erzielte und Konvergenz der Ordnung zwei zeigte, war eine Konvergenzanalyse lange unbekannt.

Diese gelang erstmals in Jansing (2015) und wurde später in Jansing und Schädle (2017) publiziert. Dafür wurde die Differentialgleichung und das Verfahren zunächst transformiert und anschließend bekannte Konvergenzresultate aus Hairer et al. (2006) verwendet. Für den Konvergenzbeweis in Jansing und Schädle (2017) ist jedoch die spezielle Form der Semidiskretisierung entscheidend. In diesem Kapitel wird ein alternativer Beweis unter Verwendung des Satz 4.11 unter allgemeineren Voraussetzungen geführt.

6.1 Modellierung des Problems

In diesem Kapitel wird ein vereinfachtes Modell aus Tückmantel et al. (2010) betrachtet. Das Interesse liegt auf hochozillatorischen Problemen, weshalb die folgenden Gleichungen nur die hybriden (hydrodynamisch behandelten) Teilchen des Plasmas über die Impulse der Elektronen $\boldsymbol{p}$ beschreiben. Der Laser wird durch das elektrische Feld $\boldsymbol{E}$ und den magnetischen Fluss $\boldsymbol{B}$ beschrieben. Zunächst sei $D \subset \mathbb{R}^3$ offen, beschränkt und besitze einen Lipschitz-Rand ∂D und dem äußerem Normalenvektor ν. Man erhält in dimensionslosen Variablen $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{p} : \mathbb{R} \times D \to \mathbb{R}^3$, $\omega : D \to \mathbb{R}$

$$\frac{d\boldsymbol{p}}{dt} = \boldsymbol{E}, \tag{6.1a}$$

$$\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} - \omega^2 \boldsymbol{p}, \tag{6.1b}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{6.1c}$$

vergleiche Jansing und Schädle (2017), Gleichung (2.1). Dabei sei $\omega(x) \geq \delta > 0$ für alle $x \in D$. Auf dem Rand werden perfekte elektrische Leiter (PEC - perfectly electric conductors) vorgegeben, also

$$\nu \times \boldsymbol{E} = 0 \qquad \text{auf } \partial D.$$

Zusätzlich sind die folgenden Anfangswerte gegeben:

$$\boldsymbol{p}(x,0) = \boldsymbol{p}_0, \qquad \boldsymbol{E}(x,0) = \boldsymbol{E}_0, \qquad \boldsymbol{B}(x,0) = \boldsymbol{B}_0 \qquad \text{für } x \in D.$$

6.2 Diskretisierung in Raum und Zeit

Die obige Konvergenzanalyse aus Abschnitt 4 umfasst lediglich den Fehler in der Zeit. Daher wird das System (6.1) zuerst örtlich diskretisiert. Dies kann beispielsweise mit zentralen finiten Differenzen (Yee-Schema) auf dem Yee-Gitter geschehen. Die verwendete Gitterschrittweite wird als h bezeichnet. Es ergeben sich die gewöhnlichen Differentialgleichungen

$$\boldsymbol{p}_h' = \boldsymbol{E}_h, \tag{6.2a}$$

$$\boldsymbol{E}_h' = G_{\text{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h - \Omega_h^2 \boldsymbol{p}_h, \tag{6.2b}$$

$$\boldsymbol{B}_h' = -G_{\text{curl}}^{\boldsymbol{E}} \boldsymbol{E}_h, \tag{6.2c}$$

$$\tag{6.2d}$$

mit diskreten Anfangswerten

$$\boldsymbol{p}_h(0) = \boldsymbol{p}_h^0, \qquad \boldsymbol{E}_h(0) = \boldsymbol{E}_h^0, \qquad \boldsymbol{B}_h(0) = \boldsymbol{B}_h^0. \tag{6.3}$$

Dabei bezeichnen $G_{\text{curl}}^{\boldsymbol{B}}$ und $G_{\text{curl}}^{\boldsymbol{E}}$ die diskretisierten $\nabla\times$-Operatoren und Ω_h stellt die Diskretisierung des Parameters ω dar. An die Differentialgleichung werden die folgenden Annahmen gestellt:

Annahme 6.1. • *Die Matrix Ω_h ist gegeben durch $\Omega_h = \operatorname{diag}(\omega(x_j))$. Weiter gilt*

$$\left\|\Omega_h^{-1}\right\| \leq \frac{1}{\delta}.$$

- *Die Matrixnormen von $G_{curl}^{\boldsymbol{E}}$ und $G_{curl}^{\boldsymbol{B}}$ hängen wesentlich von der Schrittweite der Ortsdiskretisierung h ab. Jedoch wird hier angenommen, dass die Matrixnorm beschränkt werden kann, das heißt es soll*

$$\left\|\nabla_h^{\boldsymbol{F}} \times \boldsymbol{F}\right\| \leq C_{curl} \|\boldsymbol{F}\| \qquad \text{für } \boldsymbol{F} = \boldsymbol{E}, \boldsymbol{B} \tag{6.4}$$

gelten. Denn es soll der Fall

$$C_{curl} \sim h^{-1} \ll \|\Omega_h\| = \omega_{max}$$

betrachtet werden, wobei ω_{max} die größte auftretende Frequenz des Systems beschreibt. Die Oszillationen in der Lösungen stammen somit von Ω_h und nicht von der Ortsdiskretisierung der $\nabla\times$-Operatoren.

Bemerkung 6.2. • *Mit $\|\cdot\| = \|\cdot\|_{L^2(D),h}$ wird hier die diskrete L^2-Norm bezeichnet, die von der konkreten Diskretisierung abhängt.*

- *Die Matrix Ω_h ist als Diagonalmatrix symmetrisch.*
- *Die beiden Matrizen $G_{curl}^{\boldsymbol{E}}$ und $G_{curl}^{\boldsymbol{B}}$ unterscheiden sich aufgrund der Randbedingungen und gegebenfalls aufgrund der Gitterstruktur der zugrunde liegenden Ortsdiskretisierung. Dies kann man sich beispielsweise in 1D in Liljo (2010), Gleichung (5.6), klar machen. Man beachte, dass die hergeleiteten Fehlerschranken nur unabhängig von der Zeitschrittweite τ sind, jedoch nicht von der Ortsschrittweite h.*

Im nächsten Schritt soll ein Splittingverfahren zur Zeitintegration hergeleitet werden. Das System lässt sich in drei Teilprobleme zerlegen, die man jeweils einfach lösen kann. Es ist

$$\boldsymbol{u}' = \begin{bmatrix} \boldsymbol{p}_h' \\ \boldsymbol{E}_h' \\ \boldsymbol{B}_h' \end{bmatrix} = \begin{bmatrix} \boldsymbol{E}_h \\ G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h - \Omega_h^2 \boldsymbol{p}_h \\ -G_{\mathsf{curl}}^{\boldsymbol{E}} \boldsymbol{E}_h \end{bmatrix} = A_M \boldsymbol{u} + B_M \boldsymbol{u} + C_M \boldsymbol{u}$$

mit der Lösung $\boldsymbol{u} = [\boldsymbol{p}_h^T, \boldsymbol{E}_h^T, \boldsymbol{B}_h^T]^T$ und den Matrizen

$$A_M = \begin{bmatrix} 0 & I & 0 \\ -\Omega_h^2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B_M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -G_{\mathsf{curl}}^{\boldsymbol{E}} & 0 \end{bmatrix}, \qquad C_M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & G_{\mathsf{curl}}^{\boldsymbol{B}} \\ 0 & 0 & 0 \end{bmatrix},$$

und den Anfangswerten aus (6.3). Der Index M steht hier für „Maxwell“. Das erste Teilproblem reduziert sich auf den entkoppelten harmonischen Oszillator $\boldsymbol{p}'' = -\Omega_h^2 \boldsymbol{p}$ in jedem Gitterpunkt, da $\boldsymbol{E}_h' = p$ ist. Die exakte Lösung der Gleichung $u_{A_M}' = A_M u_{A_M}$, $u_{A_M}(0) = u_{A_M}^0$ ist damit gegeben durch

$$u_{A_M}(t) = \varphi_t^{\mathcal{A}}\left(u_{A_M}^0\right) = \begin{bmatrix} \cos(t\Omega_h) & t\operatorname{sinc}(t\Omega_h) & 0 \\ -\Omega_h \sin(t\Omega_h) & \cos(t\Omega_h) & 0 \\ 0 & 0 & I \end{bmatrix} u_{A_M}^0.$$

Man kann die Lösungen der Teilprobleme $u'_{B_M} = B_M u_{B_M}$, $u_{B_M}(0) = u^0_{B_M}$ und $u'_{C_M} = C_M u_{C_M}$, $u_{C_M}(0) = u^0_{C_M}$ direkt angeben, diese lauten

$$u_{B_M}(t) = \varphi_t^{\mathcal{B}}\left(u^0_{B_M}\right) = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -tG^{\boldsymbol{E}}_{\mathsf{curl}} & I \end{bmatrix} u^0_{B_M} = (I + tB_M)u^0_{B_M},$$

$$u_{C_M}(t) = \varphi_t^{\mathcal{C}}\left(u^0_{C_M}\right) = \begin{bmatrix} I & 0 & 0 \\ 0 & I & tG^{\boldsymbol{B}}_{\mathsf{curl}} \\ 0 & 0 & I \end{bmatrix} u^0_{C_M} = (I + tC_M)u^0_{C_M}.$$

Nun kann man eine numerische Approximation an die Lösung des Systems (6.2) berechnen, indem man ein symmetrisches Dreifach-Splittingverfahren verwendet:

$$\Phi_\tau = \varphi^{C_M}_{\frac{\tau}{2}} \circ \varphi^{B_M}_{\frac{\tau}{2}} \circ \varphi^{A_M}_{\tau} \circ \varphi^{B_M}_{\frac{\tau}{2}} \circ \varphi^{C_M}_{\frac{\tau}{2}}.$$

Das entstehende Integrationsverfahren kann in Form eines Impulsverfahrens geschrieben werden (vergleiche Verlet-I in Grubmüller et al. (1991) oder r-RESPA in Tuckerman et al. (1992)). Es lautet:

Kick:
$$\boldsymbol{B}^{n+\frac{1}{2}} = \boldsymbol{B}^n - \frac{\tau}{2} G^{\boldsymbol{E}}_{\mathsf{curl}} \boldsymbol{E}^n \tag{6.5a}$$
$$(\boldsymbol{E}^+)^n = \boldsymbol{E}^n + \frac{\tau}{2} G^{\boldsymbol{B}}_{\mathsf{curl}} \boldsymbol{B}^{n+\frac{1}{2}} \tag{6.5b}$$

Oszillation:
$$\begin{bmatrix} \boldsymbol{p}^{n+1} \\ (\boldsymbol{E}^-)^{n+1} \end{bmatrix} = \begin{bmatrix} \cos(\tau\Omega_h) & \tau\,\mathrm{sinc}(\tau\Omega_h) \\ -\Omega_h \sin(\tau\Omega_h) & \cos(\tau\Omega_h) \end{bmatrix} \begin{bmatrix} \boldsymbol{p}^n \\ (\boldsymbol{E}^+)^n \end{bmatrix} \tag{6.5c}$$

Kick:
$$\boldsymbol{E}^{n+1} = (\boldsymbol{E}^-)^{n+1} + \frac{\tau}{2} G^{\boldsymbol{B}}_{\mathsf{curl}} \boldsymbol{B}^{n+\frac{1}{2}} \tag{6.5d}$$
$$\boldsymbol{B}^{n+1} = \boldsymbol{B}^{n+\frac{1}{2}} - \frac{\tau}{2} G^{\boldsymbol{E}}_{\mathsf{curl}} \boldsymbol{E}^{n+1}. \tag{6.5e}$$

Bei diesem Dreifach-Splittingverfahren treten Resonanzen im Fehler ähnlich wie in Abschnitt 1.2 auf, siehe dafür Tückmantel et al. (2010), Fig. 1. Wie in dem Artikel vorgeschlagen, fügt man nun Filterfunktionen hinzu, um ein Verfahren zweiter Ordnung zu erhalten. Dazu seien die Funktionen ϕ und ψ_M reellwertige, gerade, analytische Funktionen und

$$\Phi = \phi(\tau\Omega_h), \qquad \Psi_M = \psi_M(\tau\Omega_h).$$

Die Filter werden dabei nur auf den Maxwell-Teil der Gleichung angewandt. Bemerkung 6.3 liefert eine Begründung für den Einsatz der Filterfunktionen an den gewählten Positionen. Man erhält das folgende modifizierte Splittingverfahren:

Kick:
$$\boldsymbol{B}^{n+\frac{1}{2}} = \boldsymbol{B}^n - \frac{\tau}{2} G^{\boldsymbol{E}}_{\mathsf{curl}} \Phi \boldsymbol{E}^n \tag{6.6a}$$
$$(\boldsymbol{E}^+)^n = \boldsymbol{E}^n + \frac{\tau}{2} \Psi_M G^{\boldsymbol{B}}_{\mathsf{curl}} \boldsymbol{B}^{n+\frac{1}{2}} \tag{6.6b}$$

Oszillation:
$$\begin{bmatrix} \boldsymbol{p}^{n+1} \\ (\boldsymbol{E}^-)^{n+1} \end{bmatrix} = \begin{bmatrix} \cos(\tau\Omega_h) & \tau\,\mathrm{sinc}(\tau\Omega_h) \\ -\Omega_h \sin(\tau\Omega_h) & \cos(\tau\Omega_h) \end{bmatrix} \begin{bmatrix} \boldsymbol{p}^n \\ (\boldsymbol{E}^+)^n \end{bmatrix} \tag{6.6c}$$

Kick:
$$\boldsymbol{E}^{n+1} = (\boldsymbol{E}^-)^{n+1} + \frac{\tau}{2} \Psi_M G^{\boldsymbol{B}}_{\mathsf{curl}} \boldsymbol{B}^{n+\frac{1}{2}} \tag{6.6d}$$
$$\boldsymbol{B}^{n+1} = \boldsymbol{B}^{n+\frac{1}{2}} - \frac{\tau}{2} G^{\boldsymbol{E}}_{\mathsf{curl}} \Phi \boldsymbol{E}^{n+1}. \tag{6.6e}$$

Bemerkung 6.3. *Zu Beginn ist nicht klar, an welcher Stelle die Filterfunktionen sinnvoller Weise stehen. Deswegen führt man zunächst Filter vor und nach der Anwendung der diskreten Curl-Matrizen $G^{\boldsymbol{E}}_{curl}$ und $G^{\boldsymbol{B}}_{curl}$ ein, also*

$$
\begin{aligned}
\textbf{\textit{Kick:}} \qquad & \boldsymbol{B}^{n+\frac{1}{2}} = \boldsymbol{B}^n - \frac{\tau}{2}\Psi_B G^{\boldsymbol{E}}_{curl}\Phi_E \boldsymbol{E}^n \\
& (\boldsymbol{E}^+)^n = \boldsymbol{E}^n + \frac{\tau}{2}\Psi_E G^{\boldsymbol{B}}_{curl}\Phi_B \boldsymbol{B}^{n+\frac{1}{2}} \\
\textbf{\textit{Oszillation:}} \qquad & \begin{bmatrix} \boldsymbol{p}^{n+1} \\ (\boldsymbol{E}^-)^{n+1} \end{bmatrix} = \begin{bmatrix} \cos(\tau\Omega_h) & \tau\,\mathrm{sinc}(\tau\Omega_h) \\ -\Omega_h \sin(\tau\Omega_h) & \cos(\tau\Omega_h) \end{bmatrix} \begin{bmatrix} \boldsymbol{p}^n \\ (\boldsymbol{E}^+)^n \end{bmatrix} \\
\textbf{\textit{Kick:}} \qquad & \boldsymbol{E}^{n+1} = (\boldsymbol{E}^-)^{n+1} + \frac{\tau}{2}\Psi_E G^{\boldsymbol{B}}_{curl}\Phi_B \boldsymbol{B}^{n+\frac{1}{2}} \\
& \boldsymbol{B}^{n+1} = \boldsymbol{B}^{n+\frac{1}{2}} - \frac{\tau}{2}\Psi_B G^{\boldsymbol{E}}_{curl}\Phi_E \boldsymbol{E}^{n+1}.
\end{aligned}
$$

Bereits in Liljo (2010) in Abschnitt 5.2.1 wird erklärt, dass man $\Psi_B \equiv 1$ wählt, um weiterhin ein divergenzfreies Magnetfeld $\boldsymbol{B}$ zu garantieren. Die Wahl $\Phi_B \equiv 1$ wird jedoch ausschließlich anhand numerischer Beispiele motiviert. In Jansing (2015), Abschnitt 4.4 und Gleichung (4.30), wird diese Wahl aus der Zweischritt-Formulierung (Transformation zur Wellengleichung, siehe dazu Abschnitt 6.3) motiviert. Die zusätzliche Filterung an dieser Stelle ist nicht nötig.

6.2.1 Symmetrie des Verfahrens

Um das Splittingverfahren in eine Zweischritt-Formulierung zu bringen, wird verwendet, dass das Verfahren symmetrisch und damit in der Zeit umkehrbar ist (folgt aus Definition A.1). Dies soll hier kurz begründet werden.

Nach Konstruktion ist das Ausgangsverfahren Φ_τ symmetrisch: Die exakten Flüsse autonomer Differentialgleichungen sind stets umkehrbar (siehe (A.3), kann hier auch elementar nachgerechnet werden). Schreibt man

$$\phi_\tau = \varphi_\tau^{C_M} \circ \varphi_\tau^{B_M} \circ \varphi_\tau^{A_M},$$

so ist

$$\Phi_\tau = \phi_{\frac{\tau}{2}} \circ \phi^*_{\frac{\tau}{2}}$$

und damit nach Definition A.1 symmetrisch.

Nun werden zusätzlich Filterfunktionen eingeführt, um das gewünschte Splittingverfahren (6.6) zu erreichen. Es kann jedoch elementar nachgerechnet werden, dass die Umkehrbarkeit des Flusses mit symmetrischen (geraden) Filterfunktionen, wie sie oben gewählt werden, erhalten bleibt. Genauer gesagt, mit

$$
\begin{aligned}
\varphi_\tau^{\widetilde{B_M}} &= I + \tau\widetilde{B_M}, & \widetilde{B_M} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -G^{\boldsymbol{E}}_{\mathsf{curl}}\Phi & 0 \end{bmatrix}, \\
\varphi_\tau^{\widetilde{C_M}} &= I + \tau\widetilde{C_M}, & \widetilde{C_M} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \Psi_M G^{\boldsymbol{B}}_{\mathsf{curl}} \\ 0 & 0 & 0 \end{bmatrix},
\end{aligned}
$$

gelten die Identitäten

$$(\varphi_\tau^{\widetilde{B_M}})^{-1} = \varphi_{-\tau}^{\widetilde{B_M}}, \qquad (\varphi_\tau^{\widetilde{C_M}})^{-1} = \varphi_{-\tau}^{\widetilde{C_M}}$$

Damit erhält man analog wie oben beim Ausgangsverfahren Φ_τ die Symmetrie.

6.3 Transformation zur semilinearen Wellengleichung

Es wird nun die Idee aus Jansing und Schädle (2017) aufgegriffen und das System erster Ordnung (6.2) auf eine Differentialgleichung zweiter Ordnung reduziert. Dafür ist es zunächst notwendig, höhere Regularität in der Zeit für $\boldsymbol{E}_h$ zu fordern. Man erhält

$$\frac{\partial \boldsymbol{E}_h}{\partial t^2} = G_{\mathsf{curl}}^{\boldsymbol{B}} \frac{\partial \boldsymbol{B}_h}{\partial t} - \Omega_h^2 \frac{\partial \boldsymbol{p}_h}{\partial t} = -G_{\mathsf{curl}}^{\boldsymbol{B}} G_{\mathsf{curl}}^{\boldsymbol{E}} \boldsymbol{E}_h - \Omega_h^2 \boldsymbol{E}_h, \tag{6.7a}$$

mit Anfangswerten

$$\boldsymbol{E}_h(0) = \boldsymbol{E}_h^0, \tag{6.7b}$$

$$\boldsymbol{E}_h'(0) = G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 - \Omega_h^2 \boldsymbol{p}_h^0. \tag{6.7c}$$

Die Gleichung (6.7a) wurde somit in die Form (1.1) mit Matrix $\Omega = \Omega_h$ und der linearen Funktion $g(\boldsymbol{E}_h) = G_M \boldsymbol{E}_h = -G_{\mathsf{curl}}^{\boldsymbol{B}} G_{\mathsf{curl}}^{\boldsymbol{E}} \boldsymbol{E}_h$ gebracht. Im nächsten Abschnitt soll nun gezeigt werden, dass auch der verwendete Integrator (6.6) dem Splittingverfahren aus Buchholz et al. (2018) entspricht. Für den Zusammenhang zwischen diesem Splittingverfahren und dem Splittingverfahren (3.10) aus der vorliegenden Arbeit sei auf Abschnitte 2.2.7, 3.2 und 4.2 verwiesen.

6.3.1 Transformation des Integrators

Die Darstellung des Splittingverfahrens (6.6) als Zweischritt-Formulierung stammt aus Jansing und Schädle (2017) (Abschnitt 6.1. Reformulation). Zunächst setzt man die einzelnen Schritte des Verfahrens ineinander ein und erhält die folgenden drei Gleichungen:

$$\begin{aligned}
\boldsymbol{p}_{n+1} &= \cos(\tau\Omega_h)\boldsymbol{p}_n + \left(\tau \operatorname{sinc}(\tau\Omega_h) - \frac{\tau^3}{4}\operatorname{sinc}(\tau\Omega_h)\Psi_M G_{\mathsf{curl}}^{\boldsymbol{B}} G_{\mathsf{curl}}^{\boldsymbol{E}} \Phi\right) \boldsymbol{E}_n \\
&\quad + \frac{\tau^2}{2}\operatorname{sinc}(\tau\Omega_h)\Psi_M G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_n, \\
\boldsymbol{E}_{n+1} &= -\Omega_h \sin(\tau\Omega_h)\boldsymbol{p}_n + \left(\cos(\tau\Omega_h) - \frac{\tau^2}{4}(1+\cos(\tau\Omega_h))\Psi_M G_{\mathsf{curl}}^{\boldsymbol{B}} G_{\mathsf{curl}}^{\boldsymbol{E}} \Phi\right) \boldsymbol{E}_n \\
&\quad + \frac{\tau}{2}(1+\cos(\tau\Omega_h))\Psi_M G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_n, \\
\boldsymbol{B}_{n+1} &= \boldsymbol{B}_n - \frac{\tau}{2} G_{\mathsf{curl}}^{\boldsymbol{E}} \Phi(\boldsymbol{E}_n + \boldsymbol{E}_{n+1}).
\end{aligned}$$

Da das Splittingverfahren (6.6) symmetrisch ist (siehe Abschnitt 6.2.1), kann man $\boldsymbol{E}_{n-1}$ direkt berechnen, indem man einen Rückwärtsschritt ausführt (siehe dazu Abschnitt 6.2.1), das heißt man ersetzt $\boldsymbol{E}_{n+1}$ durch $\boldsymbol{E}_{n-1}$ und τ durch $-\tau$. Mithilfe der trigonometrischen Identität

$$\tfrac{1}{2}\big(1+\cos(\tau\Omega_h)\big) = \cos^2(\tfrac{1}{2}\tau\Omega_h)$$

und $G_M\boldsymbol{E} = -G^{\boldsymbol{B}}_{\mathsf{curl}}G^{\boldsymbol{E}}_{\mathsf{curl}}\boldsymbol{E}$ erhält man die Zweischritt-Formulierung

$$\boldsymbol{E}_{n+1} - 2\cos(\tau\Omega_h)\boldsymbol{E}_n + \boldsymbol{E}_{n-1} = \tau^2\cos^2(\tfrac{1}{2}\tau\Omega_h)\Psi_M G_M\Phi\boldsymbol{E}_n \tag{6.8a}$$

mit Startschritt

$$\boldsymbol{E}_1^{\mathsf{M}} = -\Omega_h\sin(\tau\Omega_h)\boldsymbol{p}_0 + \cos(\tau\Omega_h)\boldsymbol{E}_0 + \tau\cos^2(\tfrac{1}{2}\tau\Omega_h)\Psi_M\Big(G^{\boldsymbol{B}}_{\mathsf{curl}}\boldsymbol{B}_0 + \frac{\tau}{2}G_M\Phi\boldsymbol{E}_0\Big). \tag{6.8b}$$

Da im Folgenden ein weiterer Startvektor auftreten wird, bezeichnet man den Startvektor $\boldsymbol{E}_1^{\mathsf{M}}$ des Verfahrens (6.6) mit dem Index M.

Man kann nun zeigen, dass die hergeleitete Zweischritt-Formulierung äquivalent zu dem trigonometrischen Integrator aus Buchholz et al. (2018) ist, dessen Konvergenz zweiter Ordnung mit ähnlichen Techniken wie in Abschnitt 4 nachgewiesen worden ist. Man setzt dafür die Variablen

$$q(t) = \boldsymbol{E}_h(t), \qquad q_n = \boldsymbol{E}_n, \tag{6.9a}$$

für die Matrizen

$$\Omega = \Omega_h, \qquad G_M = -G^{\boldsymbol{B}}_{\mathsf{curl}}G^{\boldsymbol{E}}_{\mathsf{curl}}, \tag{6.9b}$$

und für die Filterfunktion (siehe Bemerkung 6.4)

$$\operatorname{sinc}(\tau\Omega_h)\Psi_S = \cos^2\left(\frac{1}{2}\tau\Omega_h\right)\Psi_M, \tag{6.9c}$$

wobei Ψ_S die Filterfunktion des Splittingverfahrens und Ψ_M die hier verwendete Filterfunktion des Maxwell-Integrators ist. Damit hat sowohl die Gleichung (6.7a) die Form der Gleichung (1.1) mit linearer Funktion g, als auch die Zweischritt-Formulierung (6.8a) die Darstellung (2.57a). Lediglich der Startschritt stimmt nicht vollständig überein. Verwendet man das Verfahren aus Buchholz et al. (2018), so erhält man mit $\boldsymbol{E}_h'(0) = G^{\boldsymbol{B}}_{\mathsf{curl}}\boldsymbol{B}_h^0 - \Omega_h^2\boldsymbol{p}_h^0$ den Startwert

$$\begin{aligned}\boldsymbol{E}_1^{\mathsf{S}} &= \cos(\tau\Omega_h)\boldsymbol{E}_0 + \sin(\tau\Omega_h)\Omega_h^{-1}(G^{\boldsymbol{B}}_{\mathsf{curl}}\boldsymbol{B}_0 - \Omega_h^2\boldsymbol{p}_0) + \frac{\tau}{2}\sin(\tau\Omega_h)\Omega_h^{-1}\Psi_S G_M\Phi\boldsymbol{E}_0\\ &= -\sin(\tau\Omega_h)\Omega_h\boldsymbol{p}_0 + \cos(\tau\Omega_h)\boldsymbol{E}_0 + \tau\cos^2(\tfrac{1}{2}\tau\Omega_h)\left(G^{\boldsymbol{B}}_{\mathsf{curl}}\boldsymbol{B}_0 + \frac{\tau}{2}\Psi_M G_M\Phi\boldsymbol{E}_0\right).\end{aligned}$$

Die Differenz beider Startwerte ist somit

$$\boldsymbol{E}_1^{\mathsf{M}} - \boldsymbol{E}_1^{\mathsf{S}} = \tau\cos^2(\tfrac{1}{2}\tau\Omega_h)(\Psi_M - I)G^{\boldsymbol{B}}_{\mathsf{curl}}\boldsymbol{B}_0.$$

Der Unterschied besteht darin, dass das Splittingverfahren (2.57) für eine Differentialgleichung 2. Ordnung hergeleitet wurde und damit den gegebenen Anfangswert $q'(0)$ verwendet. Das hier vorgestellte Dreifach-Splittingverfahren (6.6) ist jedoch für drei gekoppelte Differentialgleichungen 1. Ordnung ausgelegt. Der Startwert $\boldsymbol{E}_1^{\mathsf{M}}$ verwendet somit die Filterfunktion Ψ_M auf $G^{\boldsymbol{B}}_{\mathsf{curl}}\boldsymbol{B}_0$, die für $\boldsymbol{E}_1^{\mathsf{S}}$ nicht verwendet wird. Den Fehler im Startwert gilt es nachfolgend zu untersuchen.

Bemerkung 6.4. *Die Bedingung* (6.9c) *wird beispielsweise von*

$$\Psi_S = \cos^2\left(\frac{1}{2}\tau\Omega_h\right), \qquad\qquad \Psi_M = \operatorname{sinc}(\tau\Omega_h)$$

erfüllt. Jedoch soll die Filterfunktion Ψ_S letzlich geeignet gewählt werden, um Konvergenz des Splittingverfahrens (6.8a) *zu gewährleisten. Wie dies zu erfüllen ist, wird Annahme 6.8 zeigen.*

6.4 Konvergenznachweis

Im nächsten Abschnitt soll die Konvergenz des Splittingverfahrens mit bereits bekannten Aussagen gezeigt werden. Dafür betrachtet man die modifizierte Differentialgleichung

$$\widetilde{\boldsymbol{E}_h}'' = -\Omega_h^2 \widetilde{\boldsymbol{E}_h} + G_M \widetilde{\boldsymbol{E}_h} \tag{6.10a}$$

mit Anfangswerten

$$\widetilde{\boldsymbol{E}_h}(0) = \boldsymbol{E}_h^0, \tag{6.10b}$$

$$\widetilde{\boldsymbol{E}_h}'(0) = \Psi_M G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 - \Omega_h^2 \boldsymbol{p}_h^0, . \tag{6.10c}$$

Wendet man darauf das Splittingverfahren (2.57) an, so erhält man das Verfahren (6.8a) mit Startwert $\boldsymbol{E}_1^{\mathsf{S}}$ aus (6.8b). Diese kleine Störung im Anfangswert $\widetilde{\boldsymbol{E}_h}'$ durch die Filterfunktion Ψ_M bewirkt, dass die exakte Lösung $\widetilde{\boldsymbol{E}_h}$ wie $\widetilde{u}$ aus Abschnitt 3.1 vom Zeitschritt τ abhängt. Es soll nun gezeigt werden, dass sich die Differenz der analytischen Lösung $\boldsymbol{E}_h$ von (6.7a) mit Anfangswerten $\boldsymbol{E}_h(0)$ und $\boldsymbol{E}_h'(0)$ von der analytsichen Lösung $\widetilde{\boldsymbol{E}_h}$ mit Anfangswerten $\widetilde{\boldsymbol{E}_h}(0)$ und $\widetilde{\boldsymbol{E}_h}'(0)$ proportional zu τ^2 beschränken lässt. Dafür wird die folgende zusätzliche Annahme benötigt:

Annahme 6.5. *Für den Anfangswert $\boldsymbol{B}_h^0$ gilt*

$$\left\|\Omega_h G_{curl}^{\boldsymbol{B}} \boldsymbol{B}_h^0\right\| \leq C_{\boldsymbol{B}_h} \tag{6.11}$$

mit einer Konstante $C_{\boldsymbol{B}_h}$, die nicht von Ω_h abhängt.

Bemerkung 6.6. *Eine ähnliche Bedingung wie* (6.11) *findet sich in stärkerer Form auch in Jansing und Schädle (2017), vergleiche* (4.5).

Mit dieser Annahme lässt sich das folgende Lemma beweisen:

Lemma 6.7. *(vergleiche Lemma 6.4 aus Jansing und Schädle (2017)) Es gelte die Annahme 6.1. Die gerade Filterfunktion ψ_M erfülle die Bedingungen* (4.1a) *und* (4.1b). *Der Anfangswert $\boldsymbol{B}_h^0$ genügt der Annahme 6.5. Dann folgt für die Differenz der Lösung $\boldsymbol{E}_h$ von* (6.7) *und der Lösung $\widetilde{\boldsymbol{E}_h}$ aus* (6.10)

$$\left\|\boldsymbol{E}_h(t) - \widetilde{\boldsymbol{E}_h}(t)\right\| \leq C\tau^2, \qquad 0 \leq t \leq t_{end},$$

mit einer Konstante C, die von M_4, $C_{\boldsymbol{B}_h}$, C_{curl} und t_{end} abhängt, aber nicht von Ω_h oder τ.

Beweis. Schreibt man die Differentialgleichung (6.7a) in ein System erster Ordnung, so erhält man

$$\begin{bmatrix} \boldsymbol{E}_h(t) \\ \boldsymbol{E}_h'(t) \end{bmatrix}' = (A_h + B_h) \begin{bmatrix} \boldsymbol{E}_h(t) \\ \boldsymbol{E}_h'(t) \end{bmatrix},$$

mit

$$A_h = \begin{bmatrix} 0 & I \\ -\Omega_h^2 & 0 \end{bmatrix}, \qquad B_h = \begin{bmatrix} 0 & 0 \\ G_M & 0 \end{bmatrix}.$$

Diese Differentialgleichung wird mit den zwei verschiedenen Anfangswerten

$$\begin{bmatrix} \boldsymbol{E}_h(0) \\ \boldsymbol{E}_h'(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{E}_h^0 \\ G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 - \Omega_h^2 \boldsymbol{p}_h^0 \end{bmatrix}, \qquad \begin{bmatrix} \widetilde{\boldsymbol{E}_h}(0) \\ \widetilde{\boldsymbol{E}_h}'(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{E}_h^0 \\ \Psi_M G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 - \Omega_h^2 \boldsymbol{p}_h^0 \end{bmatrix},$$

versehen, insbesondere gilt $\boldsymbol{E}_h(0) = \widetilde{\boldsymbol{E}_h}(0)$. Weiterhin sei

$$R_h(t) = e^{tA_h} = \begin{bmatrix} \cos(t\Omega_h) & t\,\mathrm{sinc}(t\Omega_h) \\ -\Omega_h \sin(t\Omega_h) & \cos(t\Omega_h) \end{bmatrix}.$$

Mit Hilfe der Variation-der-Konstanten-Formel ergibt sich für die Differenz

$$\begin{bmatrix} \boldsymbol{E}_h(t) \\ \boldsymbol{E}_h'(t) \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{E}_h}(t) \\ \widetilde{\boldsymbol{E}_h}'(t) \end{bmatrix} = R_h(t) \left(\begin{bmatrix} \boldsymbol{E}_h(0) \\ \boldsymbol{E}_h'(0) \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{E}_h}(0) \\ \widetilde{\boldsymbol{E}_h}'(0) \end{bmatrix} \right) + \int_0^t R_h(t-s) B_h \left(\begin{bmatrix} \boldsymbol{E}_h(s) \\ \boldsymbol{E}_h'(s) \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{E}_h}(s) \\ \widetilde{\boldsymbol{E}_h}'(s) \end{bmatrix} \right) ds.$$

Die erste Differenz ist

$$\begin{aligned} R_h(t) \left(\begin{bmatrix} \boldsymbol{E}_h(0) \\ \boldsymbol{E}_h'(0) \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{E}_h}(0) \\ \widetilde{\boldsymbol{E}_h}'(0) \end{bmatrix} \right) &= R_h(t) \begin{bmatrix} 0 \\ (I - \Psi_M) G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 \end{bmatrix} \\ &= \begin{bmatrix} t\,\mathrm{sinc}(t\Omega_h)(I - \Psi_M) G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 \\ \cos(t\Omega_h)(I - \Psi_M) G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 \end{bmatrix} \\ &= \begin{bmatrix} \tau^2 \sin(t\Omega_h)(I - \Psi_M)(\tau\Omega_h)^{-2} \Omega_h G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 \\ \tau \cos(t\Omega_h)(I - \Psi_M)(\tau\Omega_h)^{-1} \Omega_h G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 \end{bmatrix}. \end{aligned}$$

Der Term ist somit mit Hilfe von (4.2) und (6.11) beschränkt durch

$$\left\| \tau^2 \sin(t\Omega_h)(I - \Psi_M)(\tau\Omega_h)^{-2} \Omega_h G_{\mathsf{curl}}^{\boldsymbol{B}} \boldsymbol{B}_h^0 \right\| \leq M_4 C_{\boldsymbol{B}_h} \tau^2.$$

Außerdem ist

$$R_h(t-s) B_h = \begin{bmatrix} (t-s)\,\mathrm{sinc}\big((t-s)\Omega_h\big) G_M & 0 \\ \cos((t-s)\Omega_h) G_M & 0 \end{bmatrix}.$$

Man erhält für die Differenz

$$\left\| \boldsymbol{E}_h(t) - \widetilde{\boldsymbol{E}_h}(t) \right\| \leq M_4 C_{\boldsymbol{B}_h} \tau^2 + \int_0^t (t-s) C_{\mathsf{curl}}^2 \left\| \boldsymbol{E}_h(s) - \widetilde{\boldsymbol{E}_h}(s) \right\| ds,$$

und damit mit dem Lemma von Gronwall B.2 die Aussage $\left\| \boldsymbol{E}_h(t) - \widetilde{\boldsymbol{E}_h}(t) \right\| \leq C\tau^2$. □

Im Folgenden soll nun das obige Lemma 6.7 mit der Aussage aus Satz 2.5 kombiniert werden, um die Konvergenz des Verfahrens (6.8) angewandt auf die Lösung $\boldsymbol{E}_h$ aus (6.7a) zu zeigen. Dazu werden zunächst die Voraussetzungen des Satzes aufgeführt:

Annahme 6.8.

- *Es gelte die Finite-Energie-Bedingung*

$$\|\Omega_h \boldsymbol{E}_h(t)\|^2 + \left\|\boldsymbol{E}_h'(t)\right\|^2 \leq K_M, \qquad 0 \leq t \leq t_{\mathit{end}}. \tag{6.12}$$

  Es genügt, wenn die Finite-Energie-Bedingung zum Zeitpunkt $t = 0$ erfüllt ist, was man im Beweis von Lemma B.1 sieht.

- *Seien ϕ und ψ_{MS} gerade, reellwertige, analytische Funktionen. Für $\chi = \phi, \psi_{MS}$ gelten die Filterbedingungen aus Annahme 4.1.*

- *Die Filterfunktionen ψ_S und ψ_M seien gegeben durch*

$$\psi_M(x) = \operatorname{sinc}(x)\psi_{MS}(x),$$
$$\psi_S(x) = \cos^2\left(\frac{x}{2}\right)\psi_{MS}(x).$$

Bemerkung 6.9. *In der zweiten Annahme wird eine zusätzliche Funktion ψ_{MS} eingeführt. Mithilfe der dritten Annahme ist garantiert, dass einerseits die Äquivalenz der Verfahren gewährleistet ist, also*

$$\operatorname{sinc}(\tau\Omega_h)\psi_S(\tau\Omega_h) = \cos^2\left(\frac{\tau\Omega_h}{2}\right)\psi_M(\tau\Omega_h)$$

gilt, und andererseits Ψ_S den Voraussetzungen von Satz 2.5 genügt.

Satz 6.10. *Es gelten die Annahmen 6.1, 6.5 und 6.8. Dann konvergiert die numerische Lösung $\boldsymbol{E}_n$ definiert in* (6.6) *in zweiter Ordnung gegen die exakte Lösung $\boldsymbol{E}_h(t_n)$ der Differentialgleichung* (6.2) *mit der Fehlerschranke*

$$\|\boldsymbol{E}_h(t_n) - \boldsymbol{E}_n\| \leq C\tau^2, \qquad 0 \leq t_n = n\tau \leq t_{end},$$

mit einer Konstante C, die von C_{curl}, $C_{\boldsymbol{B}_h}$, K_M, δ^{-1}, M_j, $j = 1, \dots 4$, und t_{end} abhängt, jedoch nicht von Ω_h oder τ.

Beweis. Zunächst entspricht die Iterierte $\boldsymbol{E}_n$ aus (6.6) der Iterierten $\boldsymbol{E}_n$ aus (6.8). Mit den Annahmen 6.1 und 6.8 und den Variablen und Matrizen aus (6.9) sind alle Voraussetzungen von Satz 2.5 erfüllt. Da sich der Integrator (6.8) ergibt, wenn man das Splittingverfahren (2.57) auf die modifizierte Differentialgleichung (6.10) anwendet, folgt somit

$$\left\|\widetilde{\boldsymbol{E}_h}(t_n) - \boldsymbol{E}_n\right\| \leq C\tau^2, \qquad 0 \leq t_n = n\tau \leq t_{\text{end}},$$

mit einer Konstante C, die nur von δ^{-1}, C_{curl}, K_M, M_j, $j = 1, \dots, 4$, und t_{end} abhängt. Diese Aussage wird nun mit der Aussage aus Lemma 6.7 kombiniert. Da die für Lemma 6.7 gestellten Voraussetzungen ebenfalls erfüllt sind, folgt

$$\|\boldsymbol{E}_h(t_n) - \boldsymbol{E}_n\| \leq \left\|\boldsymbol{E}_h(t_n) - \widetilde{\boldsymbol{E}_h}(t_n)\right\| + \left\|\widetilde{\boldsymbol{E}_h}(t_n) - \boldsymbol{E}_n\right\| \leq C\tau^2$$

für alle $0 \leq t_n = n\tau \leq t_{\text{end}}$. Die Konstante C hängt dabei nur von C_{curl}, $C_{\boldsymbol{B}_h}$, K_M, δ^{-1}, M_j, $j = 1, \dots 4$, und t_{end} ab, jedoch nicht von Ω_h oder τ. □

Mit Satz 6.10 wurde die Konvergenz zweiter Ordnung des Verfahrens (6.6) angewandt auf (6.2) nachgewiesen. Die Konvergenz zweiter Ordnung kann auch auf $\boldsymbol{B}_n$ und $\boldsymbol{p}_n$ übertragen werden, siehe dazu Jansing und Schädle (2017), Theorem 6.6 und Theorem 6.8. Es ist lediglich zu erwarten, dass wie in Jansing und Schädle (2017) weitere Filterbedingungen notwendig sind.

Numerische Experimente finden sich in Jansing und Schädle (2017), Abschnitt 8, wobei hier eine Laserfrequenz der Form

$$\omega(x) = \begin{cases} \omega_0, & x \in D_1 \\ 0 & \text{sonst} \end{cases}$$

für $D_1 \subseteq D$ betrachtet wird. Das Resultat aus Satz 6.10 ist interessant, weil es beliebige (zeitunabhängige) Frequenzen in der Matrix Ω_h zulässt, solange Annahme 6.1 erfüllt ist. Weiterhin konnte die Bedingung (4.5) aus Jansing und Schädle (2017) abgeschwächt werden zu (6.11).

6.4.1 Vergleich der Bedingungen an die Filterfunktionen

In den verschiedenen Konvergenzanalysen treten verschiedene, jeweils hinreichende Bedingungen an die Filterfunktionen auf. In Jansing und Schädle (2017) sind ϕ und ψ_M ebenfalls gerade, reellwertige Funktionen, die ebenfalls $\phi(0) = \psi_M(0) = 1$ erfüllen (siehe (5.1)). Zusätzlich werden die folgenden Voraussetzungen an die Filterfunktionen gestellt:

$$|(\cos(2\xi)+1)\psi_M(\xi)| \leq C_1 \operatorname{sinc}^2(\xi), \tag{6.13a}$$

$$|\phi(\xi)| \leq C_2 \left|\operatorname{sinc}(\xi)\right|, \tag{6.13b}$$

$$|(\cos(2\xi)+1)\psi_M(\xi)\phi(\xi)| \leq C_3 \left|\operatorname{sinc}(2\xi)\right|, \tag{6.13c}$$

$$|(\cos(2\xi)+1)\psi_M(\xi)| \leq C_4 \left|\operatorname{sinc}(2\xi)\right|, \tag{6.13d}$$

$$\left|\operatorname{sinc}(2\xi) - \frac{1}{2}(\cos(2\xi)+1)\psi_M(\xi)\right| \leq C_5\xi^2 \left|\operatorname{sinc}(2\xi)\right|, \tag{6.13e}$$

siehe auch $(5.2a) - e))$ mit $\xi = 2z$. Die dritte Bedingung ergibt sich dabei aus der zweiten und der vierten Bedingung mit $C_3 = C_2C_4$ und kann deshalb vernachlässigt werden.

G. Jansing und A. Schädle benötigen nur eine Bedingung an den Filter ϕ. Diese enthält jedoch die hier geforderten Bedingung (4.1b) und fordert ebenfalls Nullstellen der Funktion ϕ an Vielfachen von π. Für die vorliegende Analyse sind jedoch nur Nullstellen für alle geraden Vielfachen von π nötig (vergleiche Bedingung (4.1d)). Allerdings wird in der vorliegenden Analyse zusätzlich gefordert, dass $\phi(\xi)$ sich wie ξ^{-1} verhält für $\xi \to \infty$ (vergleiche Bedingung (4.1c) und (4.1d)).

Betrachtet man die verbleibenden Bedingungen für ψ_M, so ergibt sich ebenfalls die Bedingung (4.1b) und auch die Nullstellen bei Vielfachen von 2π werden gefordert. Die Forderung von Nullstellen bei ungeraden Vielfachen von π wird hier jedoch durch den Faktor $\cos(2z)+1$ erfüllt. Dies ist also analog zu den Voraussetzungen in der vorliegenden Arbeit. Wie auch für ϕ wird in Jansing und Schädle (2017) keine Bedingung $|\xi\psi_M(\xi)| \leq M_2$ wie in (4.1c) gefordert.

In der vorliegenden Arbeit wird zusätzlich die Analytizität der Filterfunktionen gefordert. Ohne diese muss man Bedingung (4.2) explizit fordern. G.Jansing und A. Schädle brauchen in Lemma 6.4 eine vergleichbare Bedingung, die sie jedoch durch die Filterbedingung (6.13e) garantieren.

Anhänge

ANHANG A

GRUNDLAGENWISSEN

In den meisten Kapiteln der vorliegenden Arbeit wird das Landau-Symbol $\mathcal{O}$ verwenden, um eine asymptotische obere Schranke anzugeben. Es ist für Funktionen $f, g : \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$ allgemein definiert durch

$$f(x) = \mathcal{O}(g(x)), \;\; x \to a \qquad \Leftrightarrow \qquad \limsup_{x \to a} \left| \frac{f(x)}{g(x)} \right| < \infty, \qquad \text{für } a \in \mathbb{R} \cup \{-\infty, \infty\}.$$

A.1 Fluss einer Differentialgleichung

Die vorliegenden Arbeit verwendet das fundamentale Konzept eines Flusses einer Differentialgleichung zur Zeit t. Dabei bezeichnet der Fluss eine Abbildung, die jeden Punkt y_0 im Phasenraum auf einen Punkt $y(t)$ abbildet, der die Lösung der Differentialgleichung

$$\dot{y} = f(y) \tag{A.1}$$

mit Anfangswert $y(0) = y_0$ zur Zeit t darstellt. Die Abbildung ist damit definiert durch:

$$\varphi_t(y_0) = y(t), \qquad \text{für } y(0) = y_0$$

(vergleiche Hairer et al. (2006), Seite 2). Aus der Definition und unter der Annahme, dass die Differentialgleichung eindeutig lösbar ist, erhält man unmittelbar, dass der Fluss φ_t die folgenden Halbgruppeneigenschaften erfüllen muss:

$$\begin{aligned} \varphi_0(y_0) &= y_0 \\ \varphi_t(\varphi_s(y_0)) &= \varphi_{t+s}(y_0) \end{aligned} \tag{A.2}$$

Ein numerisches Einschrittverfahrenzu approximativen Lösung von (A.1) kann man durch den numerischen Fluss ϕ_τ darstellen als

$$y_{n+1} = \phi_\tau(y_n), \qquad n = 0, 1, \ldots$$

wobei $y_n \approx y(t_n)$ mit $t_n = n\tau$ für $n = 0, 1, \ldots$.

A.2 Symmetrie eines numerischen Verfahrens

Der Fluss φ_t einer autonomen Differentialgleichung

$$\dot{y} = f(y), \qquad y(t_0) = y_0,$$

genügt wegen den Halbgruppeneingeschaften (A.2) der Identität

$$\varphi_{-t}^{-1} = \varphi_t \tag{A.3}$$

denn

$$\varphi_{-t}(\varphi_t(y_0)) = \varphi_{-t+t}(y_0) = \varphi_0(y_0) = y_0. \tag{A.4}$$

Diese Eigenschaft besitzt jedoch nicht unbedingt jeder numerische Fluss ϕ_h. Dies wird sehr anschaulich in Hairer et al. (2006), Kapitel II.3, Abbildung 3.1, illustiert. Jedoch gibt es numerische Verfahren, die diese Eigenschaft bewahren. Man nennt sie symmetrisch oder in der Zeit umkehrbar. Zur Definition benötigt man die Adjungierten eines numerischen Verfahrens:

Definition A.1 (vergleiche Definition 3.1 aus Hairer et al. (2006))**.** *Das adjungierte Verfahren ϕ_t^* eines Verfahrens ϕ_t ist die inverse Abbildung des ursprünglichen Verfahrens mit umgekehrtem Zeitschritt $-t$, das heißt*

$$\phi_t^* = \phi_{-t}^{-1}$$

Mit anderen Worten, $y_1 = \phi_t^*(y_0)$ ist implizit über $\phi_{-t}(y_1) = y_0$ definiert. Ein Verfahren, das $\phi_t^* = \phi_t$ erfüllt, heißt symmetrisch.

Aus der obigen Definition folgt unmittelbar, dass ein Symmetrie und Zeit-Umkehrbarkeit eines numerischen Verfahrens äquivalente Eigenschaften sind.

A.3 Klassischer Ordnungsbegriff

Gegeben sei das Anfangswertproblem

$$y'(t) = f(y(t)), \qquad t \in [t_0, T], \tag{A.5}$$
$$y(t_0) = y_0 \tag{A.6}$$

mit $f : \mathbb{R}^d \to \mathbb{R}^d$ und ein numerisches Einschrittverfahren $y_1 = \phi_\tau(y_0)$, welches Approximationen $y_1 \approx y(t_0 + \tau)$ an die exakte Lösung liefert.

Unter Konvergenz der Ordnung p des numerischen Verfahrens mit Fluss ϕ_τ versteht man allgemein, dass der globale Fehler proportional zu τ^p beschränkt ist, also

$$y_n - y(t_0 + n\tau) = \mathcal{O}(\tau^p).$$

Um dies zu erreichen, müssen meist einige Voraussetzungen an das Anfangswertproblem und an das Verfahren erfüllt sein. Betrachtet man beispielsweise das klassische Euler-Verfahren

$$y_1 = \phi_\tau^E(y_0) = y_0 + \tau f(y_0),$$

so ist das Euler-Verfahren ϕ_τ^E konvergent der Ordnung eins, wenn $f \in C^2(\mathbb{R}^d)$ ist. Das Störmer-Verlet-Verfahren für ein Anfangswertproblem 2. Ordnung mit rechter Seite f ist konvergent der Ordnung zwei für $f \in C^1(\mathbb{R}^d)$ und geeigneten Startwerten (zum Störmer-Verlet-Verfahren siehe Abschnitt 1.1.1). Dabei geht die Größe der ersten Ableitung von f in die Fehlerkonstante ein. Um dies zu beweisen, verwendet man den Technik des *Lady Windermeres Fächer*. Eine Beschreibung dieser Beweistechnik findet sich in Hairer, Nø rsett und Wanner (1993), Abschnit I.7.

Bei einer klassischen Fehleranalyse geht sowohl in die Abschätzung des lokalen Fehlers (meist mithilfe der Taylorentwicklung) als auch in der Abschätzung der Fehlerfortpflanzung (meist mit Lipschitz-Stetigkeit der rechten Seite) die Annahme $\tau \|f'(y)\| \ll 1$ ein. Für hochoszillatorischen Probleme ist dies jedoch eine äußerst restriktive Schrittweiteneinschränkung, die man nicht in Kauf nehmen möchte. Auch steife Probleme erfüllen diese Annahme nicht. Steife Probleme werden umfangreich in Hairer und Wanner (1996) diskutiert. Der Unterschied zwischen einem steifen und einem hochoszillatorischem Problem besteht in der Regularität der Lösung. Steife Probleme können durchaus eine hohe Regularität der Lösung fordern, die für hochoszillaorische Probleme nicht angenommen werden kann. Umgangssprachlich wird häufig auch vom „steifen (Schrittweiten-)Bereich“ gesprochen, wenn Schrittweiten gewählt werden, für die $\tau \|f'(y)\| > 1$ ist.

A.4 Matrixfunktionen

Unter einer Matrixfunktion versteht man die Anwendung einer skalaren Funktion f auf eine Matrix $A \in \mathbb{C}^{d\times d}$, sodass $f(A)$ eine Matrix gleicher Dimension wie A ist. Matrixfunktionen treten in zahlreichen Feldern der Mathematik und ihrer Anwendungen auf. Beispielsweise ist bereits die Inverse A^{-1} eine Matrixfunktion, die zur Berechnung der Lösung linearer Gleichungssysteme benötigt wird. In der vorliegenden Arbeit werden Matrixfunktionen benötigt, um die exakte Lösung von Differentialgleichung darzustellen. Beispielsweise ist die exakte Lösung des harmonischen Oszillators

$$u'' = -\Omega^2 u, \qquad u(0) = u_0, \quad u'(0) = u_0'$$

durch die Matrixexponentialfunktion

$$\begin{bmatrix} u(t) \\ u'(t) \end{bmatrix} = \exp\left(t \begin{bmatrix} 0 & I \\ -\Omega^2 & 0 \end{bmatrix}\right) = \begin{bmatrix} \cos(t\Omega) & t\,\mathrm{sinc}(t\Omega) \\ -\Omega \sin(t\Omega) & \cos(t\Omega) \end{bmatrix} \begin{bmatrix} u_0 \\ u_0' \end{bmatrix}.$$

gegeben. Hier treten durch die Darstellung der Matrixexponentialfunktion weitere trigonometrische Matrixfunktionen auf. Gleiches geschieht auch bei der Darstellung der Lösung der Gleichung (3.2a) durch die Variation-der-Konstanten-Formel (4.8a). Weiterhin werden Matrixfunktionen benötigt, um exponentielle Integratoren zur Approximation der Lösung von Differentialgleichung zu konstruieren, vergleiche Hochbruck und Ostermann (2010). Die Klasse der trigonometrischen Integratoren sind ebenfalls exponentielle Integratoren, sie werden in Abschnitt 1.1.2 vorgestellt. Zur Berechnung der Approximation mithilfe eines trigonometrischen Integrators ist die Berechnung von trigonometrische Matrixfunktionen sowie Filterfunktionen notwendig. Neben den Differentialgleichungen und ihrer numerischen Approximation gibt es noch weitere Anwendungen von Matrixfunktionen wie beispielsweise das Feld der inversen Probleme und ihre Regularisierung.

Dabei gibt es zahlreiche Möglichkeiten eine Matrixfunktion zu definieren und auch numerisch zu approximieren. Eine umfassende Einführung bietet Higham (2008). Das Rechnen mit Matrizen und eine Vielzahl an numerischen Approximationen von Matrizen ist in Golub und Van Loan (2013) beschrieben.

Ist $f(\lambda)$ ein Polynom oder eine rationale Funktion, so kann man die Matrixfunktion direkt angeben, indem man λ durch A ersetzt. So gilt

$$p(\lambda) = \lambda^2 + \lambda - 1, \qquad \Rightarrow \qquad p(A) = A^2 + A - I,$$
$$f(\lambda) = \frac{1+t^2}{1-t}, \qquad \Rightarrow \qquad f(A) = (I-A)^{-1}(I+A^2), \qquad \text{falls } 1 \notin \sigma(A),$$

wobei $\sigma(A)$ die Menge aller Eigenwerte von A ist (Spektrum von A). Neben dieser einfachen Definition für Polynome und rationale Funktionen, wird für diese Arbeit Matrixfunktionen für analytische Funktionen f und einer symmetrisch, positiv-semidefiniten Matrix A benötigt. Aus der Symmetrie der Matrix folgt direkt, dass diese diagonalisierbar ist

$$A = X^{-1}\Lambda X, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d),$$

wobei $\lambda_i \in \mathbb{R}$ die Eigenwerte der Matrix A sind. Für eine analytische Funktion f definiert man

$$f(A) = X^{-1} f(\Lambda) X^{-1}, \qquad f(\Lambda) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_d)),$$

siehe Higham (2008), Definition 1.2.

A.4.1 Numerische Approximation von Matrixfunktionen

Kennt man die Eigenwerte und Eigenvektoren einer Matrix, so kann man die Matrixfunktion wie oben gesehen explizit ausrechnen. Es gibt aber auch eine Vielzahl an numerischen Approximationsverfahren. Numerische Verfahren sind vorallem für Matrizen $A \in \mathbb{R}^{d\times d}$ mit großer Dimension d sinnvoll. Für solche Matrizen ist die Bestimmung aller Eigenwerte nicht mehr (effizient) möglich.

In Higham (2008) werden ab Kapitel 4 zahlreiche Verfahren angegeben, um eine Matrixfunktion zu approximieren. Eines der einfachsten Verfahren für allgemeine analytische Matrixfunktionen ist das Abschneiden der Taylorreihe. Hier hängt allerdings der Abschneidefehler von der Norm der Matrix ab, vergleiche Higham (2008) Theorem 4.8. Dies ist für Matrixfunktionen von Matrizen A mit großer Norm keine sinnvolle Approximation. Der Fehler ist zu groß. Es gibt jedoch andere Verfahren, die mit Matrizen mit großer Norm umgehen können, beispielsweise rationale Krylov-Verfahren, vergleiche Grimm und Hochbruck (2008), oder die „Scaling and Squaring“-Verfahren (Skalieren und Quadrieren), die in Higham (2008), Kapitel 10.3 und den dort erwähnten Referenzen vorgestellt werden.

ANHANG B

HILFREICHE LEMMATA

Lemma B.1. *Gegeben sei die Differentialgleichung* (1.1) *mit den Bedingungen aus der Annahme 1.1. Ist die Finite-Energie-Bedingung für* $t = 0$ *erfüllt, so erfüllt die Lösung die Finite-Energie--Bedingung zu allen Zeiten* $0 \leq t \leq t_{end}$.

Beweis. Es sei zunächst

$$v = \begin{bmatrix} q \\ q' \end{bmatrix}, \qquad \widehat{\Omega} = \begin{bmatrix} \Omega & 0 \\ 0 & I \end{bmatrix},$$
$$A_* = \begin{bmatrix} 0 & I \\ -\Omega^2 & 0 \end{bmatrix} \qquad b_*(v) = \begin{bmatrix} 0 & 0 \\ g(v) & 0 \end{bmatrix}.$$

Damit ist (1.1) äquivalent zu dem folgenden System erster Ordnung

$$v' = A_* v + b_*(v), \qquad v_0 = \begin{bmatrix} q_0 \\ q_0' \end{bmatrix}.$$

Es ist dann

$$R(t) = e^{tA_*} = \begin{bmatrix} \cos(t\Omega) & t\,\mathrm{sinc}(t\Omega) \\ -\Omega \sin(t\Omega) & \cos(t\Omega) \end{bmatrix}$$

Es soll gezeigt werden, dass

$$\left\| \widehat{\Omega} v_0 \right\| \leq K_0 \qquad \Rightarrow \qquad \left\| \widehat{\Omega} v(t) \right\| \leq K_0 + t\widetilde{C_{g,0}}$$

folgt. Mit Hilfe der Variation-der-Konstanten-Formel erhält man

$$v(t) = R(t)v_0 + \int_0^t R(t-s) b_*(v(s))\, ds. \tag{B.1}$$

Sei $\widetilde{R}(t)$ definiert durch

$$\widehat{\Omega} R(t) = \widetilde{R}(t)\widehat{\Omega} \qquad \text{mit } \widetilde{R}(t) = \begin{bmatrix} \cos(t\Omega) & \sin(t\Omega) \\ -\sin(t\Omega) & \cos(t\Omega) \end{bmatrix}.$$

Durch Multiplikation der Gleichung (B.1) mit $\widehat{\Omega}$ folgt

$$\widehat{\Omega}v(t) = \widetilde{R}(t)\widehat{\Omega}v_0 + \int_0^t \widetilde{R}(t-s)\widehat{b}_*(v(s))\,ds.$$

Damit folgt für $\|v(s)\| \leq \widetilde{r}_0$ und $\left\|\widehat{\Omega}v_0\right\| \leq K_0$

$$\begin{aligned}\left\|\widehat{\Omega}v(t)\right\| &\leq \left\|\widehat{\Omega}v_0\right\| + \int_0^t \|g(v(s))\|\,ds \\ &\leq K_0 + t\widetilde{C_{g,0}} \\ &\leq K_0 + t_{\text{end}}\widetilde{C_{g,0}}\end{aligned}$$

für alle $0 < t \leq t_{\text{end}}$. □

B.1 Zwei Varianten des Lemmas von Gronwall

Lemma B.2. *(Lemma von Gronwall)*
Seien $y(t)$, $f(t)$ und $g(t)$ nicht negative Funktionen auf dem Intervall $[0, t_{end}]$, die einseitige Grenzwerte für alle Punkte $t \in [0, t_{end}]$ haben. Weiter gelte für $0 \leq t \leq t_{end}$

$$y(t) \leq f(t) + \int_0^t g(s)y(s)\,ds.$$

Dann folgt für $0 \leq t \leq t_{end}$

$$y(t) \leq f(t) + \int_0^t g(s)f(s)\exp\left(\int_s^t g(u)\,du\right)\,ds.$$

Beweis. Siehe Holte (2009). □

Lemma B.3. *(Diskretes Lemma von Gronwall)*
Es seien $(y_n)_{n\in\mathbb{N}}$, $(f_n)_{n\in\mathbb{N}}$ und $(g_n)_{n\in\mathbb{N}}$ nicht negative Folgen und es gelte

$$y_n \leq f_n + \sum_{k=0}^{n-1} g_k y_k, \qquad \text{für } n \geq 1.$$

Dann ist

$$y_n \leq f_n + \sum_{k=0}^{n-1} f_k g_k \exp\left(\sum_{j=k+1}^{n-1} g_j\right), \qquad \text{für } n \geq 1.$$

Beweis. Siehe Holte (2009). □

B.2 Partielle Summation

Die (Abelsche) partielle Summation (siehe auch Heuser (1991)) ist zunächst für eine natürliche Zahl n und reelle Zahlen $a_0, a_1, \ldots, a_n, b_0, b_1, \ldots, b_n$ definiert

$$\sum_{k=0}^{n} a_k b_k = A_n b_n + \sum_{k=0}^{n-1} A_k (b_k - b_{k+1}) \qquad \text{mit} \qquad A_k = \sum_{j=0}^{k} a_j.$$

Oder analog

$$\sum_{k=0}^{n} a_k b_k = a_n B_n + \sum_{k=0}^{n-1} (a_k - a_{k+1}) B_k, \qquad \text{mit} \qquad B_k = \sum_{j=0}^{k} b_j.$$

Im Beweis des globalen Fehlers in Satz 4.8 wird diese Umformung nun verwendet um die Summen

$$\sum_{j=0}^{n} e^{(n-j)\tau A} \delta_j^{(1)} \qquad \text{und} \qquad \sum_{j=0}^{n} e^{(n-j)\tau A} \delta_j^{(2)}$$

zu transformieren. Nun handelt es sich hierbei nicht um reele Zahlen, sondern um Matrizen und Vektoren. Nun ist partielle Summation immer möglich, wenn a_k auf b_k linear wirkt, d.h. man Summen wie A_k bilden beziehungsweise auseinander ziehen kann. Die Kommutivität von a_k und b_k spielt keine Rolle.

Im Fall der ersten Summe ist $a_j = e^{j\tau A}$ eine Matrix, die linear auf den Vektor $b_j = \delta_{n-j}^{(1)}$ wirkt. Mit

$$E_j = \sum_{k=0}^{j} e^{k\tau A}$$

folgt somit

$$\begin{aligned}
\sum_{j=0}^{n} e^{(n-j)\tau A} \delta_j^{(1)} &= \sum_{j=0}^{n} e^{j\tau A} \delta_{n-j}^{(1)} \\
&= E_n \delta_0^{(1)} + \sum_{j=0}^{n-1} E_j (\delta_{n-j}^{(1)} - \delta_{n-j-1}^{(1)}) \\
&= E_n \delta_0^{(1)} + \sum_{j=0}^{n-1} E_{n-j-1} (\delta_{j+1}^{(1)} - \delta_j^{(1)}).
\end{aligned}$$

Im zweiten Fall folgt mit Lemma 4.7 $\delta_n^{(2)} = Z_2(\widetilde{u}(t_n))\widetilde{\Phi} A^2 \widetilde{u}(t_n)$. Definiert man nun

$$a_j = e^{(n-j)\tau A} Z_2(\widetilde{u}(t_j))\widetilde{\Phi} A^2, \qquad b_j = \widetilde{u}(t_j),$$

so ist zwar Z_2 nichtlinear abhängig vom Argument $\widetilde{u}(t_j)$, dies wird allerdings in a_j mitgeführt. Die Matrix a_j wirkt dann linear auf den Vektor $b_j = \widetilde{u}(t_j)$, so ergibt sich mit

$$F_j = \sum_{k=0}^{j} \widetilde{u}(t_k)$$

die zweiten Variante der partiellen Summation

$$\begin{aligned}\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)} &= \sum_{j=0}^{n} e^{(n-j)\tau A} Z_2(\widetilde{u}(t_j))\widetilde{\Phi}A^2\widetilde{u}(t_j) \\ &= Z_2(\widetilde{u}(t_n))\Phi A^2 F_n + \sum_{j=0}^{n-1} e^{(n-j)\tau A}\left(Z_2(\widetilde{u}(t_j)) - e^{-\tau A} Z_2(\widetilde{u}(t_{j+1}))\right)\widetilde{\Phi}A^2 F_j.\end{aligned}$$

ABBILDUNGSVERZEICHNIS

LITERATURVERZEICHNIS

Adams, R. A. & Fournier, J. J. F. (2003). *Sobolev spaces* (2. Aufl., Bd. 140). Elsevier/Academic Press, Amsterdam.

Aulbach, B. (2004). *Gewöhnliche Differenzialgleichungen* (2. Aufl. Aufl.). München: Elsevier, Spektrum Akadem. Verl.

Auzinger, W., Kassebacher, T., Koch, O. & Thalhammer, M. (2017). Convergence of a Strang splitting finite element discretization for the Schrödinger-Poisson equation. *ESAIM Math. Model. Numer. Anal.*, *51* (4), 1245–1278. Zugriff auf https://doi.org/10.1051/m2an/2016059 doi: 10.1051/m2an/2016059

Baumstark, S., Faou, E. & Schratz, K. (2018). Uniformly accurate exponential-type integrators for Klein-Gordon equations with asymptotic convergence to the classical NLS splitting. *Math. Comp.*, *87* (311), 1227–1254. Zugriff auf https://doi.org/10.1090/mcom/3263 doi: 10.1090/mcom/3263

Birdsall, C. K. & Langdon, A. B. (1991). *Plasma physics via computer simulation*. Bristol [u.a.]: Hilger.

Blanes, S. & Casas, F. (2016). *A concise introduction to geometric numerical integration*. Boca Raton: CRC Press.

Brigham, E. O. (1982). *FFT: schnelle Fourier-Transformation*. R. Oldenbourg Verlag, München.

Buchholz, S., Gauckler, L., Grimm, V., Hochbruck, M. & Jahnke, T. (2016). Two different approaches to highly oscillatory problems. In E. Faou, E. Hairer, M. Hochbruck & C. Lubich (Hrsg.), *Oberwolfach reports: Geometric numerical integration* (Bd. 13, S. 887-889). European Mathematical Society Publishing House.

Buchholz, S., Gauckler, L., Grimm, V., Hochbruck, M. & Jahnke, T. (2018). Closing the gap between trigonometric integrators and splitting methods for highly oscillatory differential equations. *IMA J. Numer. Anal.*, *38* (1), 57–74. Zugriff auf https://doi.org/10.1093/imanum/drx007 doi: 10.1093/imanum/drx007

Cano, B. & Moreta, M. J. (2010). Multistep cosine methods for second-order partial differential systems. *IMA J. Numer. Anal.*, *30* (2), 431–461. Zugriff auf https://doi.org/10.1093/imanum/drn043 doi: 10.1093/imanum/drn043

Cano, B. & Moreta, M. J. (2013). High-order symmetric multistep cosine methods. *Appl. Numer. Math.*, *66*, 30–44. Zugriff auf https://doi.org/10.1016/j.apnum.2012.11.005 doi: 10.1016/j.apnum.2012.11.005

Cohen, D. (2006). Conservation properties of numerical integrators for highly oscillatory Hamiltonian systems. *IMA J. Numer. Anal.*, *26* (1), 34–59. Zugriff auf https://doi.org/10.1093/imanum/dri020 doi: 10.1093/imanum/dri020

Cohen, D., Gauckler, L., Hairer, E. & Lubich, C. (2015). Long-term analysis of numerical integrators for oscillatory Hamiltonian systems under minimal non-resonance conditions. *BIT*, *55* (3), 705–732. Zugriff auf http://dx.doi.org/10.1007/s10543-014-0527-8 doi: 10.1007/s10543-014-0527-8

Cohen, D., Hairer, E. & Lubich, C. (2005). Numerical energy conservation for multi-frequency oscillatory differential equations. *BIT*, *45* (2), 287–305. Zugriff auf https://doi.org/10.1007/s10543-005-7121-z doi: 10.1007/s10543-005-7121-z

Cohen, D., Larsson, S. & Sigg, M. (2013). A trigonometric method for the linear stochastic wave equation. *SIAM J. Numer. Anal.*, *51* (1), 204–222. Zugriff auf https://doi.org/10.1137/12087030X doi: 10.1137/12087030X

Dauxois, T. (2008). Fermi, Pasta, Ulam and a mysterious lady. *arXiv preprint arXiv:0801.1590*.

Deuflhard, P. (1979). A study of extrapolation methods based on multistep schemes without parasitic solutions. *Z. Angew. Math. Phys.*, *30* (2), 177–189. Zugriff auf https://doi.org/10.1007/BF01601932

Dong, X. (2014). Stability and convergence of trigonometric integrator pseudospectral discretization for N-coupled nonlinear Klein-Gordon equations. *Appl. Math. Comput.*, *232*, 752–765. Zugriff auf https://doi.org/10.1016/j.amc.2014.01.144 doi: 10.1016/j.amc.2014.01.144

Einkemmer, L. & Ostermann, A. (2014). Convergence analysis of Strang splitting for Vlasov-type equations. *SIAM J. Numer. Anal.*, *52* (1), 140–155. Zugriff auf http://dx.doi.org/10.1137/130918599 doi: 10.1137/130918599

Einkemmer, L. & Ostermann, A. (2015). A splitting approach for the Kadomtsev-Petviashvili equation. *J. Comput. Phys.*, *299*, 716–730. Zugriff auf http://dx.doi.org/10.1016/j.jcp.2015.07.024 doi: 10.1016/j.jcp.2015.07.024

Einkemmer, L. & Ostermann, A. (2018). A comparison of boundary correction methods for Strang splitting. *Discrete Contin. Dyn. Syst. Ser. B*, *23* (7), 2641–2660. Zugriff auf https://doi.org/10.3934/dcdsb.2018081 doi: 10.3934/dcdsb.2018081

Engquist, B., Fokas, A., Hairer, E. & Iserles, A. H. (Hrsg.). (2009). *Highly oscillatory problems* (1. publ. Aufl.). Cambridge [u.a.]: Cambridge University Press.

Faou, E., Ostermann, A. & Schratz, K. (2015). Analysis of exponential splitting methods for inhomogeneous parabolic equations. *IMA J. Numer. Anal.*, *35* (1), 161–178. Zugriff auf http://dx.doi.org/10.1093/imanum/dru002 doi: 10.1093/imanum/dru002

Faou, E. & Schratz, K. (2014). Asymptotic preserving schemes for the Klein-Gordon equation in the non-relativistic limit regime. *Numer. Math.*, *126* (3), 441–469. Zugriff auf https://doi.org/10.1007/s00211-013-0567-z doi: 10.1007/s00211-013-0567-z

Fermi, E., Ulam, S. & Pasta, J. (1955). *Studies of nonlinear problems* (Bericht). University of California.

García-Archilla, B., Sanz-Serna, J. M. & Skeel, R. D. (1999). Long-time-step methods for oscillatory differential equations. *SIAM J. Sci. Comput.*, *20* (3), 930–963. Zugriff auf http://dx.doi.org/10.1137/S1064827596313851 doi: 10.1137/S1064827596313851

Gauckler, L. (2015). Error analysis of trigonometric integrators for semilinear wave equations. *SIAM J. Numer. Anal.*, *53* (2), 1082–1106. Zugriff auf http://dx.doi.org/10.1137/140977217 doi: 10.1137/140977217

Gautschi, W. (1961). Numerical integration of ordinary differential equations based on trigonometric polynomials. *Numer. Math.*, *3*, 381–397. Zugriff auf https://doi.org/10.1007/BF01386037

Golub, G. H. & Van Loan, C. F. (2013). *Matrix computations* (4. ed. Aufl.). Baltimore, Md.: Johns Hopkins University Pr.

Gremillet, L., Bonnaud, G. & Amiranoff, F. (2002). Filamented transport of laser-generated relativistic electrons penetrating a solid target. *Physics of Plasmas*, *9* (3), 941-948. Zugriff auf https://doi.org/10.1063/1.1432994 doi: 10.1063/1.1432994

Grimm, V. (2002). *Exponentielle Integratoren als Lange-Zeitschritt-Verfahren für oszillatorische Differentialgleichungen zweiter Ordnung* (Dissertation, Heinrich-Heine-Universität Düsseldorf). Zugriff auf https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=2164

Grimm, V. (2005a). A note on the Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.*, *102* (1), 61–66. Zugriff auf http://dx.doi.org/10.1007/s00211-005-0639-9

Grimm, V. (2005b). On error bounds for the Gautschi-type exponential integrator applied to oscillatory second-order differential equations. *Numer. Math.*, *100* (1), 71–89. Zugriff auf https://doi.org/10.1007/s00211-005-0583-8 doi: 10.1007/s00211-005-0583-8

Grimm, V. & Hochbruck, M. (2006). Error analysis of exponential integrators for oscillatory second-order differential equations. *J. Phys. A*, *39* (19), 5495–5507. Zugriff auf http://dx.doi.org/10.1088/0305-4470/39/19/S10 doi: 10.1088/0305-4470/39/19/S10

Grimm, V. & Hochbruck, M. (2008). Rational approximation to trigonometric operators. *BIT*, *48* (2), 215–229. Zugriff auf http://dx.doi.org/10.1007/s10543-008-0185-9 doi: 10.1007/s10543-008-0185-9

Grubmüller, H., Heller, H., Windemuth, A. & Schulten, K. (1991). Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Molecular Simulation*, *6* (1-3), 121–142. doi: 10.1080/08927029108022142

Hairer, E. & Lubich, C. (2000). Long-time energy conservation of numerical methods for oscillatory differential equations. *SIAM J. Numer. Anal.*, *38* (2), 414–441. Zugriff auf https://doi.org/10.1137/S0036142999353594

Hairer, E., Lubich, C. & Wanner, G. (2003). Geometric numerical integration illustrated by the Störmer/Verlet method. *Acta Numerica*, *12*, 399–450.

Hairer, E., Lubich, C. & Wanner, G. (2006). *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations* (Second Aufl., Bd. 31). Springer-Verlag, Berlin.

Hairer, E., Nø rsett, S. P. & Wanner, G. (1993). *Solving ordinary differential equations I: Nonstiff problems* (Second Aufl., Bd. 8). Springer-Verlag, Berlin.

Hairer, E. & Wanner, G. (1996). *Solving ordinary differential equations II: Stiff and differential-algebraic problems* (Second Aufl., Bd. 14). Springer-Verlag, Berlin. Zugriff auf https://doi.org/10.1007/978-3-642-05221-7 doi: 10.1007/978-3-642-05221-7

Hansen, E. & Ostermann, A. (2009). Exponential splitting for unbounded operators. *Math. Comp.*, *78* (267), 1485–1496. Zugriff auf http://dx.doi.org/10.1090/S0025-5718-09-02213-3 doi: 10.1090/S0025-5718-09-02213-3

Hansen, E. & Ostermann, A. (2016). High-order splitting schemes for semilinear evolution equations. *BIT*, *56* (4), 1303–1316. Zugriff auf http://dx.doi.org/10.1007/s10543-016-0604-2 doi: 10.1007/s10543-016-0604-2

Hansen, E., Ostermann, A. & Schratz, K. (2016). The error structure of the Douglas-Rachford splitting method for stiff linear problems. *J. Comput. Appl. Math.*, *303*, 140–145. Zugriff auf https://doi.org/10.1016/j.cam.2016.02.037 doi: 10.1016/j.cam.2016.02.037

Hegelich, M., Albright, B. J., Cobble, J., Flippo, K., Letzring, S., Paffett, M., ... Fernández, J. C. (2006, 26. 01). Laser acceleration of quasi-monoenergetic MeV ion beams. *Nature*, *439*, 441 EP -. Zugriff auf http://dx.doi.org/10.1038/nature04400

Hegelich, M., Karsch, S., Pretzler, G., Habs, D., Witte, K., Guenther, W., ... Roth, M. (2002, Aug). MeV ion jets from short-pulse-laser interaction with thin foils. *Phys. Rev. Lett.*, *89*, 085002. Zugriff auf https://link.aps.org/doi/10.1103/PhysRevLett.89.085002 doi: 10.1103/PhysRevLett.89.085002

Heuser, H. (1991). *Lehrbuch der Analysis. Teil 2* (Sechste Aufl.). B. G. Teubner, Stuttgart.

Higham, N. J. (2008). *Functions of matrices : Theory and computation*. Philadelphia: Society for Industrial and Applied Mathematics.

Hochbruck, M., Jahnke, T. & Schnaubelt, R. (2015). Convergence of an ADI splitting for Maxwell's equations. *Numer. Math.*, *129* (3), 535–561. Zugriff auf http://dx.doi.org/10.1007/s00211-014-0642-0 doi: 10.1007/s00211-014-0642-0

Hochbruck, M. & Lubich, C. (1999). A Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.*, *83* (3), 403–426. Zugriff auf http://dx.doi.org/10.1007/s002110050456 doi: 10.1007/s002110050456

Hochbruck, M. & Lubich, C. (2003). On Magnus integrators for time-dependent Schrödinger equations. *SIAM J. Numer. Anal.*, *41* (3), 945–963. Zugriff auf http://link.aip.org/link/?SNA/41/945/1 doi: 10.1137/S0036142902403875

Hochbruck, M. & Ostermann, A. (2010). Exponential integrators. *Acta Numer.*, *19*, 209–286. Zugriff auf http://dx.doi.org/10.1017/S0962492910000048 doi: 10.1017/S0962492910000048

Hockney, R. W. & Eastwood, J. W. (1992). *Computer simulation using particles* (Repr. Aufl.). Bristol [u.a.]: Hilger.

Holden, H., Karlsen, K. H., Lie, K.-A. & Risebro, N. H. (2010). *Splitting methods for partial differential equations with rough solutions*. Zürich: European Mathematical Society (EMS). Zugriff auf http://dx.doi.org/10.4171/078 (Analysis and MATLAB programs) doi: 10.4171/078

Holden, H., Lubich, C. & Risebro, N. H. (2013). Operator splitting for partial differential equations with Burgers nonlinearity. *Math. Comp.*, *82* (281), 173–185. Zugriff auf http://dx.doi.org/10.1090/S0025-5718-2012-02624-X doi: 10.1090/S0025-5718-2012-02624-X

Holte, J. M. (2009). *Discrete Gronwall lemma and applications.* Zugriff auf http://homepages.gac.edu/~holte/publications/GronwallLemma.pdf (MAA north central section meeting at University of North Dakota)

Hundsdorfer, W. H. & Verwer, J. G. (2007). *Numerical solution of time dependent advection diffusion reaction equations* (Corr. 2. print. Aufl.). Berlin: Springer. Zugriff auf http://swbplus.bsz-bw.de/bsz266173926cov.htm

Jahnke, T. & Lubich, C. (2000). Error bounds for exponential operator splittings. *BIT*, *40* (4), 735–744. Zugriff auf http://dx.doi.org/10.1023/A:1022396519656 doi: 10.1023/A:1022396519656

Jahnke, T. & Mikl, M. (2018). Adiabatic midpoint rule for the dispersion-managed nonlinear Schrödinger equation. *Numer. Math.*, *138* (4), 975–1009. Zugriff auf https://doi.org/10.1007/s00211-017-0926-2 doi: 10.1007/s00211-017-0926-2

Jansing, G. (2015). *Exponentielle Integratoren - Zeitintegration für Maxwell-Gleichungen und parabolische Systeme* (Dissertation, Heinrich-Heine-Universität Düsseldorf). Zugriff auf https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=34954

Jansing, G. & Schädle, A. (2017, 02). Convergence analysis of an explicit splitting method for laser plasma interaction simulations. *ETNA*, *48*.

Kirkwood, R. K., Moody, J. D., Kline, J., Dewald, E., Glenzer, S., Divol, L., ... Lindl, J. (2013). A review of laser–plasma interaction physics of indirect-drive fusion. *Plasma Physics and Controlled Fusion*, *55* (10), 103001. Zugriff auf http://stacks.iop.org/0741-3335/55/i=10/a=103001

Koch, O. & Lubich, C. (2011). Variational-splitting time integration of the multi-configuration time-dependent Hartree-Fock equations in electron dynamics. *IMA J. Numer. Anal.*, *31* (2), 379–395. Zugriff auf http://dx.doi.org/10.1093/imanum/drp040 doi: 10.1093/imanum/drp040

Krämer, P. & Schratz, K. (2017). Efficient time integration of the Maxwell-Klein-Gordon equation in the non-relativistic limit regime. *J. Comput. Appl. Math.*, *316*, 247–259. Zugriff auf https://doi.org/10.1016/j.cam.2016.07.007 doi: 10.1016/j.cam.2016.07.007

Liljo, J. (2010). *Hybride Verfahren zur Simulation der Wechselwirkung relativistischer Kurzpuls-Laser mit hochdichten Plasmen* (Dissertation, Heinrich-Heine-Universität Düsseldorf). Zugriff auf https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=15909

Liljo, J., Karmakar, A., Pukhov, A. & Hochbruck, M. (2008). One-dimensional electromagnetic relativistic PIC-hydrodynamic hybrid simulation code H-VLPL (hybrid virtual laser plasma lab). *Comput. Phys. Comm.*, *179* (6), 371–379. Zugriff auf https://doi.org/10.1016/j.cpc.2008.03.008 doi: 10.1016/j.cpc.2008.03.008

Linz, U. & Alonso, J. (2016, Dec). Laser-driven ion accelerators for tumor therapy revisited. *Phys. Rev. Accel. Beams*, *19*, 124802. Zugriff auf https://link.aps.org/doi/10.1103/PhysRevAccelBeams.19.124802 doi: 10.1103/PhysRevAccelBeams.19.124802

Lubich, C. (2008). On splitting methods for Schrödinger-Poisson and cubic nonlinear Schrödinger equations. *Math. Comp.*, *77* (264), 2141–2153. Zugriff auf http://dx.doi.org/10.1090/S0025-5718-08-02101-7 doi: 10.1090/S0025-5718-08-02101-7

Magnus, W. (1954). On the exponential solution of differential equations for a linear operator. *Comm. Pure Appl. Math.*, *7*, 649–673. Zugriff auf https://doi.org/10.1002/cpa.3160070404 doi: 10.1002/cpa.3160070404

Mason, R. J. (1980). Monte Carlo hybrid modeling of electron transport in laser produced plasmas. *The Physics of Fluids*, *23* (11), 2204-2215. Zugriff auf https://aip.scitation.org/doi/abs/10.1063/1.862903 doi: 10.1063/1.862903

McLachlan, R. I. & Quispel, G. R. W. (2002, 1). Splitting methods. *Acta Numerica*, *11*, 341–434. Zugriff auf http://journals.cambridge.org/article_S0962492902000053 doi: 10.1017/S0962492902000053

Pazy, A. (1983). *Semigroups of linear operators and applications to partial differential equations* (Bd. 44). New York: Springer. Zugriff auf https://doi.org/10.1007/978-1-4612-5561-1

Pukhov, A. (1999). Three-dimensional electromagnetic relativistic particle-in-cell code VLPL (Virtual Laser Plasma Lab). *Journal of Plasma Physics*, *61* (3), 425–433.

Robson, L., Simpson, P. T., Clarke, R. J., Ledingham, K. W. D., Lindau, F., Lundh, O., ... McKenna, P. (2006, 10. 12). Scaling of proton acceleration driven by petawatt-laser–plasma interactions. *Nature Physics*, *3*, 58 EP -. Zugriff auf http://dx.doi.org/10.1038/nphys476

Sanz-Serna, J. M. (2008). Mollified impulse methods for highly oscillatory differential equations. *SIAM J. Numer. Anal.*, *46* (2), 1040–1059. Zugriff auf https://doi.org/10.1137/070681636 doi: 10.1137/070681636

Störmer, C. (1907). Sur les trajectoires des corpuscules électrisés dans l'espace. Applications à l'aurore boréale et aux perturbations magnétiques. *Radium (Paris)*, *4* (1), 2-5. Zugriff auf https://hal.archives-ouvertes.fr/jpa-00242218 doi: 10.1051/radium:01907004010201

Strang, G. (1968). On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, *5* (3), 506-517. Zugriff auf http://www.jstor.org/stable/2949700

Tabak, M., Hammer, J., Glinsky, M. E., Kruer, W. L., Wilks, S. C., Woodworth, J., ... Mason, R. J. (1994). Ignition and high gain with ultrapowerful lasers. *Physics of Plasmas*, *1* (5), 1626-1634. Zugriff auf https://doi.org/10.1063/1.870664 doi: 10.1063/1.870664

Thalhammer, M., Caliari, M. & Neuhauser, C. (2009). High-order time-splitting Hermite and Fourier spectral methods. *J. Comput. Phys.*, *228* (3), 822–832. Zugriff auf http://dx.doi.org/10.1016/j.jcp.2008.10.008 doi: 10.1016/j.jcp.2008.10.008

Trotter, H. F. (1959). On the product of semi-groups of operators. *Proc. Amer. Math. Soc.*, *10*, 545–551. Zugriff auf https://doi.org/10.2307/2033649 doi: 10.2307/2033649

Tuckerman, M., Berne, B. J. & Martyna, G. J. (1992). Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, *97* (3), 1990-2001. Zugriff auf https://doi.org/10.1063/1.463137 doi: 10.1063/1.463137

Tückmantel, T. (2012). *Hybrid particle-in-cell simulations of relativistic plasmas* (Dissertation, Heinrich-Heine-Universität Düsseldorf). Zugriff auf https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=23952

Tückmantel, T., Pukhov, A., Liljo, J. & Hochbruck, M. (2010). Three-dimensional relativistic particle-in-cell hybrid code based on an exponential integrator. *IEEE Transactions on Plasma Science*, *38* (9), 2383-2389. Zugriff auf http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5535138 doi: 10.1109/TPS.2010.2056706

Verlet, L. (1967, Jul). Computer “experiments“ on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, *159*, 98–103. Zugriff auf https://link.aps.org/doi/10.1103/PhysRev.159.98 doi: 10.1103/PhysRev.159.98