Exponentielle Integratoren

—

Zeitintegrationsverfahren für Maxwell-Gleichungen und parabolische Systeme

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Georg Jansing

aus Düsseldorf

Düsseldorf, im März 2015

aus dem Mathematischen Institut
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Achim Schädle
Heinrich-Heine-Universität Düsseldorf

Korreferentin: Prof. Dr. Marlis Hochbruck
Karlsruher Institut für Technologie

Tag der mündlichen Prüfung: 30. März 2015

ZUSAMMENFASSUNG

Ziel dieser Arbeit ist es, bestimmte numerische Zeitintegrationsverfahren, so genannte exponentielle Integratoren, zu entwickeln, zu analysieren und zu verfeinern. Dabei soll der Fokus auf zwei Aspekten liegen. Zum einen wollen wir einfache Laser-Plasma-Interaktionen mit hochdichten Plasmen simulieren können. Für diese treten bei klassischen Zeitintegrationsverfahren durch den zum Plasma gehörigen großen Dichteparameter numerische Stabilitätsprobleme auf. Zum anderen betrachten wir Systeme gewöhnlicher Differentialgleichungen, die aus Ortsdiskretisierungen parabolischer Differentialgleichungen hervorgehen und bei welchen eine Matrix vor der Zeitableitung auftritt. Für solche Gleichungen wollen wir exponentielle Lösungsverfahren entwickeln, die ohne Anwendung von Inversen dieser Matrix auskommen.

Für die Simulation der Laser-Plasma-Interaktionen analysieren und modifizieren wir ein in der Literatur vorgeschlagenes symmetrisches Dreifach-Splitting-Verfahren, in dessen zentraler Komponente ein exponentieller Integrator enthalten ist. Es wurde auf die spezielle Form der zu lösenden Gleichung zugeschnitten. Von diesem Verfahren ist nicht bekannt, ob es tatsächlich die durch numerische Tests motivierten, mutmaßlichen Konvergenzeigenschaften besitzt. Wir werden eine etwas allgemeinere Version des Verfahrens analysieren und ermitteln, wie wir das ursprüngliche Verfahren modifizieren müssen, damit es tatsächlich die erwartete Konvergenzordnung erhält. Außerdem werden wir zeigen, dass der ursprüngliche Vorschlag ebenfalls konvergiert, wenn auch mit geringerer Ordnung.

Im zweiten Teil der Arbeit wenden wir uns der Zeitintegration ortsdikretisierter parabolischer Differentialgleichungen zu. Die Massematrix, die dabei vor der Zeitableitung auftritt, wird bei Verwendung von Krylov-Verfahren für die Approximation von Matrixfunktionen für exponentielle Integratoren oft formal durch Invertieren behandelt. Dies führt dazu, dass in jedem Krylov-Schritt ein lineares Gleichungssystem mit dieser Matrix gelöst werden muss. Wir kombinieren Krylov-Verfahren zur Lösung linearer Gleichungssysteme mit numerischen Konturintegrationsmethoden, wodurch wir auf explizites Lösen mit der Massematrix verzichten können. Im Vergleich zu klassischen Krylov-Verfahren für Matrixfunktionen müssen wir anstatt eines einzigen Krylov-Raumes nun mehrere aufbauen, jeder einzelne Krylov-Schritt ist jedoch günstiger. Der Aufbau der Krylov-Räume lässt sich auf triviale Weise parallelisieren.

SUMMARY

In this thesis we aim to develop, improve and analyze certain numerical time integration methods, so called exponential integrators. We focus on two different aspects: Firstly, we want to simulate simple laser-plasma interactions with overdense plasmas. The large density parameter belonging to the description of the plasma introduces numerical stability issues in classical time integration methods. Secondly, we consider systems of ordinary differential equations arising from the spatial discretizations of parabolic differential equations, where a matrix appears in front of the time derivative. We will develop exponential integrators that do not rely on solving linear systems with this matrix for this type of equations.

For the simulation of laser-plasma interactions we analyze and modify a triple splitting method that was proposed in literature. This method contains an exponential integrator in its central component and was specifically designed for the equations that describe those interactions. Besides numerical tests that indicate a certain order of convergence, rigorous error bounds are unknown. We will analyze a slightly generalized version of the beforementioned method to identify modifications to be able to prove the presumed convergence rate. Additionally, we show convergence of lesser order for the original method and actually see the reduced order in a numerical experiment.

In the second part of this thesis we move to the time integration of spatially discretized parabolic differential equations. The mass matrix arising in front of the time derivative is often handled formally by invsersion when Krylov methods for approximating the matrix functions in exponential integrators are used. This leads to the solution of a linear system in each Krylov step with this matrix. We combine Krylov methods for the solution of systems of linear equations with numerical contour integration methods, where these explicit solutions are not needed anymore. When compared to classical Krylov methods for matrix functions, we now have to build more Krylov subspaces, but each single Krylov step is cheaper. It is trivial to parallelize the construction of those subspaces.

Inhaltsverzeichnis

KAPITEL 1
EINFÜHRUNG

1.1 Problemstellung und Ziele

Ziel dieser Arbeit ist es, bestimmte numerische Zeitintegrationsverfahren, so genannte exponentielle Integratoren, zu entwickeln, zu analysieren und zu verfeinern. Dabei soll der Fokus auf zwei Aspekten liegen. Zum einen wollen wir einfache Laser-Plasma-Interaktionen mit hochdichten Plasmen simulieren können. Für diese treten bei klassischen Zeitintegrationsverfahren durch den zum Plasma gehörigen großen Dichteparameter numerische Stabilitätsprobleme auf. Zum anderen betrachten wir Systeme gewöhnlicher Differentialgleichungen, die aus Ortsdiskretisierungen parabolischer Differentialgleichungen hervorgehen und bei welchen eine Matrix vor der Zeitableitung auftritt. Für solche Gleichungen wollen wir exponentielle Lösungsverfahren entwickeln, die ohne Anwendung von Inversen dieser Matrix auskommen.

Für die Simulation der Laser-Plasma-Interaktionen analysieren und modifizieren wir ein in der Literatur vorgeschlagenes symmetrisches Dreifach-Splitting-Verfahren, in dessen zentraler Komponente ein exponentieller Integrator enthalten ist. Es wurde auf die spezielle Form der zu lösenden Gleichung zugeschnitten. Von diesem Verfahren ist nicht bekannt, ob es tatsächlich die durch numerische Tests motivierten, mutmaßlichen Konvergenzeigenschaften besitzt. Wir werden eine etwas allgemeinere Version des Verfahrens analysieren und ermitteln, wie wir das ursprüngliche Verfahren modifizieren müssen, damit es tatsächlich die erwartete Konvergenzordnung erhält. Außerdem werden wir zeigen, dass der ursprüngliche Vorschlag ebenfalls konvergiert, wenn auch mit geringerer Ordnung.

Bei der Analyse des Verfahrens werden wir die spezielle Struktur von auftretenden Matrizen ausnutzen, die durch Verwendung einer bestimmten Form von Dichteprofil und einer sehr einfachen Ortsdiskretisierung, dem Yee-Gitter, entsteht. Der Einsatz dieser Ortsdiskretisierung zur Lösung von Maxwell-Gleichungen ist sehr weit verbreitet, da sie sowohl konzeptionell als auch rechnerisch einfach ist und sich zudem wichtige topologische Eigenschaften der Gleichungen auf die diskretisierten Gleichungen übertragen. Da es sich dabei aber um einen Finite-Differenzen-Ansatz handelt, ist dieser nur dann gut einsetzbar, wenn das Rechengebiet eine sehr einfache Form hat. Auch bereiten filigrane Materialstrukturen größere Schwierigkeiten.

Eine Alternative zu diesem Ansatz bilden die so genannten Finite-Elemente-Methoden, von denen einige speziell auf die Maxwell-Gleichungen zugeschnitten sind. Diese sind sehr flexibel in Bezug auf die Struktur des Rechengebietes. Sie haben aber den Nachteil, dass bei Approximationen hoher Ordnung selbst für explizite numerische Zeitschrittverfahren üblicherweise viele Gleichungssysteme mit der Massematrix des diskretisierten Systems gelöst werden müssen. Vor allem bei der Verwendung exponentieller Integratoren entsteht dadurch ein beträchtlicher Mehraufwand.

Im zweiten Teil der Arbeit werden wir daher Verfahren vorstellen, die ohne die Lösung solcher Gleichungssysteme auskommen. Dabei kombinieren wir Krylov-Verfahren zur Lösung linearer Gleichungssysteme mit numerischen Konturintegrationsmethoden. Im Vergleich zu klassischen Krylov-Verfahren für Matrixfunktionen müssen wir anstatt eines einzigen Krylov-Raumes nun mehrere aufbauen, jeder einzelne Krylov-Schritt ist jedoch günstiger. Der Aufbau der Krylov-Räume lässt sich auf triviale Weise parallelisieren.

Die Arbeit ist wie folgt strukturiert: Im weiteren Verlauf des ersten Kapitels werden wir zunächst einige Vereinbarungen und Notationen festlegen, die dem Leser das Studieren des Textes erleichtern sollen.

Im zweiten Kapitel wenden wir uns physikalischen Grundlagen von Laser-Plasma-Interaktionen zu, die durch die Vlasov-Maxwell-Gleichungen beschrieben werden. Zuerst werden wir das physikalische Modell erläutern, um dann in den folgenden beiden Abschnitten die mathematischen Gleichungen, die den Laserpuls und das Plasma beschreiben, herzuleiten. Im Anschluss daran leiten wir einige Erhaltungsgrößen her, die uns später helfen werden, das numerische Verfahren zu analysieren.

In Kapitel 3 werden wir eine erste bekannte Version des exponentiellen Integrators vorstellen, um die Gleichungen im Kontext hochdichter Plasmen zu lösen. Nach ersten numerischen Tests werden wir feststellen, dass sich das Verfahren nicht so verhält, wie wir es auf den ersten Blick hätten erwarten können. Eine Erklärung dafür ist relativ schnell gefunden. Trotz vorhandenem Verständnis der Probleme führen erste Lösungsstrategien jedoch noch nicht zum Erfolg.

Dazu werden wir in Kapitel 4 Filterfunktionen in das numerische Schema einführen. Die Wahl derselben ist jedoch a priori nicht klar. Für die Analyse des gefilterten Verfahrens werden wir ein Resultat benutzen, welches mithilfe so genannter modulierter Fourier-Entwicklungen für ein exponentielles Mehrschrittverfahren Konvergenz zeigt. Um dieses Resultat verwenden zu können, bringen wir unser Verfahren in die passende Form und formulieren Voraussetzungen an die Ortsdiskretisierung. Damit erhalten wir eine Abschätzung für einen Teil des exponentiellen Integrators bzw. eine der Teilgleichungen. Um auch Konvergenz in den anderen Größen der Gleichungen zu erhalten, müssen wir auf Basis des ersten Resultats weitere Bedingungen an die Filterfunktionen und die Startwerte stellen und erhalten so eine Analyse für das vollständige Verfahren. Am Ende haben wir eine Zusammenstellung an Voraussetzungen sowohl an die Diskretisierung als auch an die Filterfunktionen zusammengetragen. Wir zeigen, dass erstere für die Ortsdiskretisierung mit dem Yee-Gitter erfüllt sind und für zweitere überhaupt eine Wahl von Filtern existiert, die allen Bedingungen gerecht wird. Dadurch wird sich die Verbesserung des Verfahrens ergeben.

In Kapitel 5 widmen wir uns den Gleichungen mit Massematrizen. Zu Beginn werden wir die zu lösenden abstrakten Gleichungen darstellen und erläutern, wie wir die Struktur der diskretisierten Gleichung erhalten. Dann stellen wir exponentielle Integratoren für diese Art von Gleichungen und deren Hauptwerkzeug – die Anwendung von Matrixfunktionen auf Vektoren – dar. Für die

Anwendungen auf Vektoren sind bei hochdimensionalen Systemen mit bestimmten Eigenschaften der Matrizen Krylov-Verfahren eine sehr beliebte Methode. Für diese diskutieren wir die klassische Einbindung der Massematrizen, mit der dort Gleichungssysteme gelöst werden müssen. Im Anschluss daran leiten wir eine Variante von Krylov-Verfahren her, die ohne diese Lösungen von Gleichungssystemen auskommt. Es stellt sich jedoch heraus, dass diese schlecht konvergiert. Es gelingt uns aber, auf Basis derselben Ideen und der Hinzunahme eines weiteren Hilfsmittels – numerischer inverser Laplace-Transformation – eine Verbesserung vorzuschlagen. Für diese erhält man auch theoretische Konvergenzaussagen. Mit der Diskussion einiger numerischer Beispiele und einem kurzen Ausblick schließen wir dieses Kapitel und damit den inhaltlichen Teil der Arbeit.

Im letzten Kapitel fassen wir die Ergebnisse zusammen. Im Anhang werden wir uns einigen mathematischen Grundlagen zuwenden, die allgemein und lange bekannt, aber für diese Arbeit von Bedeutung sind. Außerdem stellen wir einige Rechenregeln zusammen, die leicht zu beweisen sind, aber nicht unbedingt zum Standardstoff gehören.

Dies ist der richtige Zeitpunkt, um einige Worte an Personen zu richten, die auf eine oder andere Weise zu dieser Arbeit beigetragen haben.

Beginnen will ich natürlich mit meinem Betreuer Prof. Dr. Achim Schädle. Vielen Dank für unzählige Stunden bei der Jagd auf Formeln und Abschätzungen und bei der Fehlersuche in Programmen. All unsere Diskussionen sind von unschätzbarem Wert für mich. Außerdem bin ich für den Rückhalt in einer Zeit, als alles andere davon zu schwimmen schien, äußerst dankbar.

Mein Dank gilt auch meiner ursprünglichen Betreuerin Prof. Dr. Marlis Hochbruck, die mich für die numerische Mathematik im Allgemeinen und exponentielle Integratoren im speziellen begeisterte. Leider trennten sich unsere Wege nach kurzer Zeit. Umso mehr freue ich mich über die Übernahme der Zweitkorrektur.

Bendanken möchte ich mich weiterhin bei Prof. Dr. Florian Jarre, der mir auch schon lange vor der offiziellen Übernahme des Amtes als Mentor mit Rat und Tat zur Seite stand und mir half, auch in schwierigen Zeiten weiterzumachen und die Motivation wiederzufinden.

Bei Dr. Tobias Tückmantel, den ich mit meinen Fragen zu den physikalischen Fragestellungen bestimmt nicht nur einmal fast in den Wahnsinn getrieben habe. Bei Dr. Ludwig Gauckler. Im Zusammenhang mit seinem Besuch ist der sprichwörtliche Knoten geplatzt. Die Diskussionen rund um die modulierten Fourier-Entwicklungen haben die Wende gebracht. Bei PD Dr. Volker Grimm, der schon zu Studiumszeiten ein Mentor für mich war und dessen Begeisterung mich immer mitgerissen und dessen Geduld mir sowohl Anreiz als auch Vorbild war.

Bei den Arbeitsgruppen Angewandte Mathematik und Mathematische Optimierung am Mathematischen Institut der Heinrich-Heine Universität Düsseldorf. Für ein gutes, angenehmes und produktives Arbeitsklima. Legendäre Grillabende bei Schnee und Minusgraden eingeschlossen!

Bei unseren Lehrstuhlvertretern, besonders den Doktores Lothar Nannen, Daniel Ruprecht, Thomas Dickopf und Lars Röhe. Wegen der langen Vakanz gab es einige von Euch jungen und aufstrebenden Wissenschaftlern, die Ihr mich an Eurem Fachwissen teilhaben ließet und in Eure Freundschaft einschlosst.

Bei allen, die diese Arbeit – ob fachlich oder inhaltlich – Korrektur gelesen haben. Ihr alle tragt maßgeblich zur Qualität dieser Arbeit bei!

Zum Schluss, aber bestimmt nicht als letztes, möchte ich mich bei meiner Familie, allen voran natürlich meiner Partnerin Yumiko, aber auch Yumikos und meinen Eltern und meinen Geschwistern mit ihren Partnern bedanken. Ohne Eure Unterstützung hätte ich die oft sehr harte Zeit niemals durchgestanden! Ihr habt mich wieder aufgebaut, wenn ich völlig frustriert nach Hause kam, habt mich ausgehalten, wenn ich ungehalten über meine eigene Unzulänglichkeit war und wart der unverzichtbare Ausgleich, ohne den ich diese Aufgabe unmöglich hätte bewältigen können.

1.2 Notationen und Vereinbarungen

Einführungen von Namen und Definitionen werden *kursiv geschrieben*.

Die Raumdimension wird zumeist mit d notiert, das Rechengebiet wird mit $\Omega \subseteq \mathbb{R}^d$ bezeichnet. Wir gehen davon aus, dass das Gebiet einfach strukturiert ist, etwa polygonal oder mit stückweiser C^1-Berandung.

Vektoren werden mit Vektorpfeilen notiert, also ist $a \in \mathbb{R}$ eine skalare Größe und $\vec{x} \in \mathbb{R}^d$ ein Vektor. Entsprechend verfahren wir mit skalar- bzw. vektorwertigen Abbildungen, etwa $a : \Omega \longrightarrow \mathbb{R}$, $\vec{F} : \Omega \longrightarrow \mathbb{R}^d$. Die Komponenten eines Vektors werden ohne Pfeil, etwa $\vec{x} = (x_1, \dots, x_d)$, geschrieben. Matrizen werden fett gedruckt, z.B. $\mathbf{A} \in \mathbb{R}^{d \times d}$. Mit $\left\langle \cdot, \cdot \right\rangle$ bezeichnen wir, soweit nicht anders erwähnt, das euklidische Skalarprodukt im $\mathbb{R}^d$ und mit $\| \cdot \|$ die davon induzierte Norm. Die komplexe Einheit ist $\mathring{\mathrm{i}} = \sqrt{-1}$.

Für die Divergenz eines Vektorfeldes schreiben wir $\nabla \cdot \vec{F}$, um uns an die Standardnotation in der Physik zu halten. Das äußere Einheitsnormalenvektorfeld existiert unter obiger Voraussetzung an Ω fast überall und wir bezeichnen es mit $\nu = \nu(\vec{x})$. Der Integralsatz von Gauß,

$$\int_\Omega \nabla \cdot \vec{F} \, \mathrm{d}\vec{x} = \int_{\partial\Omega} \left\langle \vec{F}, \nu \right\rangle \mathrm{d}\vec{\sigma}, \tag{1.1}$$

ist dann immer erfüllt, solange für das Vektorfeld $\vec{F}$ gilt, dass $\nabla \cdot \vec{F} \in L^2(\Omega)$ existiert und eine Spur $\gamma(\vec{F}) \in L^2(\partial\Omega)$ besitzt.

In den Kapiteln über die Maxwell-Gleichungen befinden wir uns in Raumdimension $d = 3$. Die Variable d wird dort nicht verwendet. Die Komponenten des Ortsvektors $\vec{x} = (x, y, z)$ schreiben wir dann ohne Indizes. Dort benutzen wir x, y und z an Stelle eines Laufindexes auch zur Indizie-

rung. Für ortsabhängige Vektorfelder verwenden wir, falls nicht anders beschrieben

$$\vec{F} : \mathbb{R} \times \Omega \longrightarrow \mathbb{R}^3, \vec{F} = \vec{F}(t,x,y,z) = \vec{F}(t,\vec{x}) = \begin{bmatrix} F_x(t,x,y,z) \\ F_y(t,x,y,z) \\ F_z(t,x,y,z) \end{bmatrix}. \tag{1.2}$$

Die Rotation bezeichnen wir mit $\nabla \times \vec{F}$. Für Abbildungen von mehreren Argumenten schreiben wir für die Abbildung, die entsteht, wenn wir eine Komponente festhalten und den Rest variieren lassen wollen, einen Punkt für die noch freien Komponenten, etwa

$$a : U \times V \longrightarrow W, (u,v) \longmapsto a(u,v), \qquad a(u,\cdot) : V \longrightarrow W, v \longmapsto a(u,\cdot)(v) := a(u,v). \tag{1.3}$$

KAPITEL 2
PHYSIKALISCHES MODELL FÜR LASER-PLASMA-INTERAKTION

In diesem Kapitel beschreiben wir die physikalischen Hintergründe von Laser-Plasma-Interaktionen in groben Zügen, d.h. insoweit es für die Problemstellung und das weitere Verständnis dieser Arbeit erforderlich ist, und leiten das System von Gleichungen her, das wir numerisch lösen wollen.

Abschnitt 2.1 behandelt das physikalische Modell, welches den Kontext für die Gleichungen liefert. Abschnitt 2.2 führt die Maxwell-Gleichungen ein und 2.3 stellt die für diese Arbeit relevanten Teile der Gleichungen des Laser-Plasma-Modells vor. In Abschnitt 2.4 beschreiben wir ein spezielles Experiment, das uns als Testproblem dienen wird. Der letzte Abschnitt führt mit zwei Erhaltungsgrößen Hilfsmittel ein, die für die Analyse des Verfahrens nützlich sind.

2.1 Physikalisches Modell

Das vorgestellte Konzept zur physikalischen Beschreibung und Lösung entstammt [60, 59, 89, 88]. Den besten Überblick liefert dabei [88].

Bei Laser-Plasma-Interaktionen interessieren wir uns für Wechselwirkungen von Laserpulsen, die durch elektromagnetische Felder beschrieben werden, mit Materie, die durch Einwirkung des Lasers zu Plasma wird. Die elektromagnetischen Felder werden durch Maxwell-Gleichungen beschrieben. Die Teilchen des Plasmas werden durch Bewegungsgleichungen und durch Vlasov-Gleichungen beschrieben. Alles wird in einem großen physikalischen Modell zusammengefasst und soll numerisch simuliert werden. Die für diese Arbeit relevante Klasse von Problemen besteht aus solchen, bei denen das Plasma eine sehr hohe Dichte ρ aufweist. Diese macht die numerische Behandlung der Gleichungen besonders anspruchsvoll und wir wollen uns dieser Situation in den ersten Kapiteln widmen. Erwähnenswerte Beispiele für solche Probleme sind etwa die „target normal sheath acceleration“ (TNSA), vgl. [36, 57, 73], oder das Fast-Ignition-Konzept der Trägheitsfusion (FI of Inertial Confinement Fusion, ICF) [84]. Dabei treten Dichten bis zu $\rho = 1.000\rho_c$ bei TNSA und sogar bis $\rho = 10.000\rho_c$ im Fall von FI auf, wobei ρ_c die so genannte kritische Dichte ist, die von der Laser-Frequenz abhängt und ab dieser ein Laserpuls vom Plasma reflektiert wird. Für klassische explizite Verfahren erhält man aus Stabilitätsgründen dichteabhängige Schrittweiteneinschränkungen, vgl. etwa [60, 59], für die für uns relevanten Particle-In-Cell-Codes, kurz PIC-Codes.

Plasma ist Materie in einem Zustand, bei dem Elektronen und Atomkerne, die normalerweise zusammen ein Atom bilden, von einander gelöst sind und sich die Elektronen frei bewegen können.

Die Atomkerne – von den negativ geladenen Elektronen getrennt – sind positiv geladene Ionen.

Die Anzahl der auftretenden Teilchen ist extrem groß und es wird schwierig und kostspielig, das Verhalten jedes einzelnen Teilchens nachzuvollziehen. Interessante Informationen lassen sich allerdings auch schon aus deutlich gröberen Daten gewinnen. Wir betrachten daher zwei verschiedene Ansätze, die jeweils verschiedene Stärken und Schwächen haben, um die Datenmenge zu reduzieren und die Gleichungen handhabbar zu machen.

Zum einen werden so genannte PIC-Teilchen verwendet. Dabei handelt es sich um Makroteilchen, die eine feste Ausdehnung und Form haben und die eine feste Menge von echten Teilchen zusammenfassen. Die in einem Makroteilchen vereinten Teilchen bewegen sich dann alle gemeinsam, Wechselwirkungen untereinander werden vernachlässigt. Eigenschaften wie Masse oder Impuls resultieren aus der Summe der entsprechenden Größen der repräsentierten Teilchen. Um Wechselwirkungen von Teilchen dennoch simulieren zu können, werden entsprechend viele dieser Makroteilchen benötigt, deren Interaktion untereinander nicht vernachlässigt wird.

Eine andere Möglichkeit, die Teilchen zu behandeln, ist, sie durch kontinuierliche Teilchendichten darzustellen und durch Vlasov-Gleichungen zu beschreiben. Metadaten werden abhängig von der Zeit für alle Teilchen, die denselben Ort und dieselbe Geschwindigkeit haben, gesammelt. Verschiedene weitere Vereinfachungen reduzieren die Datenmenge. In diesem Fall spricht man von hydrodynamischer Behandlung.

Beide Beschreibungen haben ihre jeweiligen Vor- und Nachteile. Die Behandlung mit PIC-Teilchen leidet unter numerischer Dispersion und Diffusion. Die hydrodynamische Beschreibung ist stärker vereinfacht und kann nicht alle physikalischen Effekte beschreiben, ist dafür jedoch deutlich schneller und in den Situationen, in denen entsprechende Effekte nicht auftreten, genauer. Es werden dabei Verfahren für hyperbolische Differentialgleichungen, wie etwa Fluss-Korrektur-Verfahren [94] oder numerische Glätter [82], benötigt, um numerische Instabilitäten zu vermeiden. Die hydrodynamische Beschreibung ist die einzige praktische Möglichkeit, mit hohen Teilchendichten zu rechnen: Entweder wird eine zu große Menge an Makroteilchen benötigt, oder die PIC-Approximation wird dadurch zu ungenau, dass die Makroteilchen zu „fett“ werden. Das bedeutet dann, dass sie sehr viele echte Teilchen repräsentieren, deren Interaktionen untereinander vernachlässigt werden. Die Einleitung aus [88] stellt das Für und Wider beider Konzepte zusammen.

Neue Varianten [60, 89, 88] des Codes VLPL (virtual laser plasma laboratory) [72] verfolgen einen kombinierten Ansatz, bei dem das Plasma teils als PIC-Makroteilchen, teils kontinuierlich hydrodynamisch behandelt wird. Diese Varianten werden mit H-VLPL (hybrid virtual laser plasma laboratory) bezeichnet.

Um die Idee von hybriden Verfahren zu veranschaulichen, ist in Abbildung 2.1 das Bild einer Anfangsbedingung einer sehr einfachen PIC-Simulation dargestellt. Es zeigt ein Makroteilchen mit ansonsten kontinuierlich behandelten Elektronen. Dargestellt sind ein elektrisches Feld $\vec{E}$,

der Impuls $\vec{p}_h$ und die Dichte ρ_h der hydrodynamisch behandelten Elektronen, das PIC-Teilchen und die davon auf die Zellen verteilten Dichteanteile.

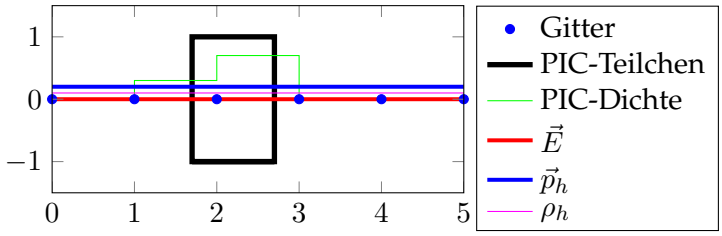


Abbildung 2.1: Einfache PIC-Simulation

In dieser Arbeit analysieren wir einen Teil des H-VLPL-Verfahrens, der speziell entwickelt wurde, um Simulationen mit hochdichten Plasmen durchführen zu können und der in [59, 89, 88] vorgestellt wird. Die PIC-Teilchen lassen wir außen vor.

2.2 Maxwell-Gleichungen

Wir betrachten die Maxwell-Gleichungen, die Wechselwirkungen zwischen elektrischen und magnetischen Feldern und Einflüsse von außen durch Materialien und Ströme beschreiben, vergleiche etwa die einführenden Kapitel aus [66, 6]. Sie sind gegeben durch

$$\partial_t \vec{D} = \nabla \times \vec{H} - \vec{j}, \tag{2.1a}$$
$$\partial_t \vec{B} = -\nabla \times \vec{E}, \tag{2.1b}$$
$$\vec{D} = \varepsilon \vec{E}, \tag{2.1c}$$
$$\vec{B} = \mu \vec{H} \tag{2.1d}$$

mit den zeitabhängigen Vektorfeldern $\vec{D}, \vec{E}, \vec{B}, \vec{H}, \vec{j}$. In dieser Formulierung sind nur nicht polarisierende, magnetisierungsfreie und lineare Materialien zugelassen und wir schränken uns zusätzlich auf isotropische Materialien, also skalare ε und μ, ein. Wir betrachten dabei die folgenden Feldgrößen:

- $\vec{E}$, die elektrische Feldstärke,
- $\vec{D}$, die elektrische Flussdichte,
- $\vec{H}$, die magnetische Feldstärke und
- $\vec{B}$, die magnetische Flussdichte

mit den ortsabhängigen und positiven Materialparametern

- ε, der elektrischen Permittivität und

- μ, der magnetischen Permeabilität,

sowie der Stromdichte

$$\vec{j} = \sigma \vec{E} + \vec{j}_a. \tag{2.2}$$

Dabei ist σ die Leitfähigkeit des Materials und $\vec{j}_a$ der von außen angelegte Strom. Mit

$$\vec{q} := \nabla \cdot \vec{D} \tag{2.3}$$

bezeichnen wir die elektrische Ladungsdichte. Aus (2.1a) ergibt sich dann

$$\partial_t \vec{q} = \nabla \cdot \partial_t \vec{D} = \nabla \cdot (\nabla \times \vec{H}) - \nabla \cdot \vec{j} \overset{(A.29)}{=} -\nabla \cdot \vec{j}. \tag{2.4}$$

Auf dieselbe Weise ergibt sich aus (2.1b)

$$\partial_t \nabla \cdot \vec{B} = \nabla \cdot (-\nabla \times \vec{E}) \overset{(A.29)}{=} 0. \tag{2.5}$$

Setzen wir nun noch $\nabla \cdot \vec{B}(0, \cdot) = 0$, also Divergenzfreiheit zum Startzeitpunkt, voraus, so ergibt sich daraus, dass das magnetische Feld zu allen Zeiten divergenzfrei ist:

$$\nabla \cdot \vec{B} = 0. \tag{2.6}$$

Die Gleichung (2.1a) wird als das *Maxwell-Ampèresche Gesetz*, (2.1c) als das *Gesetz von Faraday* und (2.3) und (2.6) als die *Gaußschen Gesetze* bezeichnet.

Wir eliminieren $\vec{H} = \mu^{-1} \vec{B}$ mit (2.1d) und $\vec{D} = \varepsilon \vec{E}$ mit (2.1c) und erhalten damit

$$\partial_t \varepsilon \vec{E} = \nabla \times \mu^{-1} \vec{B} - \vec{j} \tag{2.7a}$$

$$\partial_t \vec{B} = -\nabla \times \vec{E}. \tag{2.7b}$$

Wir lösen die Gleichung auf dem Gebiet Ω mit dem Rand $\partial\Omega$ und dem äußeren Normalenvektorfeld ν. Die Randbedingungen seien durch perfekt elektrische Leiter, engl. perfect electric conductors (PEC), mit

$$\nu \times \vec{E} = 0 \qquad \text{auf} \qquad \partial\Omega, \tag{2.8}$$

oder perfekt magnetische Leiter, engl. perfect magnetic conductors (PMC), mit

$$\nu \times \vec{H} = 0 \qquad \text{auf} \qquad \partial\Omega, \tag{2.9}$$

vgl. wieder [66] oder z.B. [93, Abschnitt 2.1.3], oder periodische Randbedingungen gegeben.

2.3 Gleichungen für das Laser-Plasma-Modell

Wir beschränken uns auf die Angabe der Gleichungen. Eine ausführlichere Herleitung lässt sich etwa in [88] finden.

Wir verwenden die Maxwell-Gleichungen für die elektromagnetischen Felder in CGS-Einheiten. Nach weiterer Umskalierung wie in [89] erhalten wir

$$\partial_t \vec{E} = \nabla \times \vec{B} - 2\pi \vec{j},$$
$$\partial_t \vec{B} = -\nabla \times \vec{E}.$$

Der Faktor 2π vor dem Strom entstand in persönlicher Kommunikation [87] durch eine etwas andere Skalierung der Zeit als in [89]. Die Teilchen seien durch ihre Dichten ρ_E für die Elektronen und ρ_I für die Ionen repräsentiert. Außer diesen Teilchen existiert kein weiteres Material, also $\varepsilon \equiv \varepsilon_0$, $\mu \equiv \mu_0$ und $\sigma = 0$ in (2.2). Wir verwenden die Annahme, dass sich die Ionen aufgrund ihrer großen Masse und damit großen Trägheit nicht bewegen. Dadurch erzeugen sie keine Ströme. Die Elektronen versehen wir zu Anfang mit demselben Verteilungsprofil wie die Ionen. Das elektrische Feld wirkt als Kraft auf die Elektronen, sodass sich deren Impuls ändert. Auch das magnetische Feld wirkt über die *Lorentz-Kraft* auf die Elektronen. Für niedrige Geschwindigkeiten ist diese aber klein und wird daher von uns vernachlässigt. Die Ionen ziehen die Elektronen zum größten Teil wieder auf ihren ursprünglichen Ort zurück. Dieser Effekt wird als *Plasma-Oszillation* bezeichnet. Wir vernachlässigen alle weitere Bewegung der Elektronen und halten ihren Ort als das Zentrum dieser lokalen Oszillationen fest, was zu einer zeitlich konstanten Dichte ρ_E führt. Durch ihre Impulse $\vec{p}_E$ erzeugen die Elektronen aber trotzdem Ströme, die auf das elektrische Feld wirken. Nach weiteren Umskalierungen und Vereinfachungen ergibt sich das vollständige System

$$\partial_t \vec{E} = \nabla \times \vec{B} - 2\pi e \rho \vec{v}, \tag{2.10a}$$
$$\partial_t \vec{B} = -\nabla \times \vec{E}, \tag{2.10b}$$
$$\partial_t \vec{p} = 2\pi e \vec{E}, \tag{2.10c}$$
$$\vec{v} = \frac{\vec{p}}{\gamma(\vec{p})}, \tag{2.10d}$$
$$\gamma(\vec{p}) = \sqrt{1 + \|\vec{p}\|^2}, \tag{2.10e}$$

wobei $\rho = \rho_E$, $\vec{p} = \vec{p}_E$. Die Dichte ρ ist die Teilchenzahldichte der Elektronen relativ zur kritischen Dichte ρ_c.

Ein Laserpuls ist durch

$$\begin{aligned} E_y = B_z &= a_0 \exp\left(-\frac{(2\pi[(x-x_0)-t])^2}{2\sigma_0^2}\right)\cos(2\pi[(x-x_0)-t]), \\ E_x &\equiv E_z \equiv B_x \equiv B_y \equiv 0 \end{aligned} \tag{2.11}$$

gegeben. Dieser löst das System (2.10) im Vakuum, falls also mit $\rho = 0$ kein Material vorhanden ist. Durch die Skalierung des Ortes und der Zeit wird beides in Laserperioden gemessen. Die Parameter a_0 und σ_0 beschreiben die Amplitude und Breite des Pulses. In Abbildung 2.2 sind die E_y- und die B_z-Komponente eines exemplarischen Pulses dargestellt.

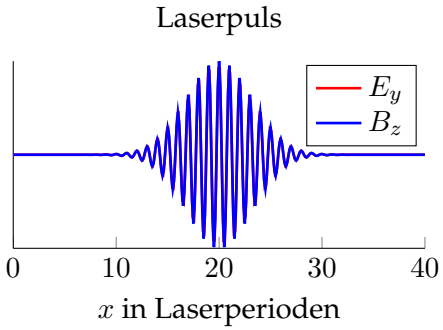


Abbildung 2.2: Darstellung eines Laserpulses (2.11) im Vakuum. Parameter: $x_0 = 50$, $\sigma_0 = 20$, $a_0 = 1$

Meistens nähern wir

$$\gamma \equiv 1, \tag{2.12}$$

was bedeutet, dass wir relativistische Effekte vernachlässigen. Wie das Vernachlässigen der Lorentz-Kraft bei kleinen Impulsen und damit niedrigen Geschwindigkeiten ist dies eine gute Approximation.

2.4 Modellanwendung: Laser-Reflexion an einer dünnen Folie

Wir verwenden die nun vorgestellte Anwendung als unser Testproblem für den Rest dieses Teiles der Arbeit. Dabei wird ein Laserpuls auf eine dünne Folie geschossen und dieser je nach Dichte des Materials entweder nur teilweise oder vollständig reflektiert. Wir gehen davon aus, dass die auftretenden Impulse des Plasmas klein bleiben und wir somit den relativistischen Faktor wie in

(2.12) vernachlässigen können. Es ergibt sich das System

$$\partial_t \vec{E} = \nabla \times \vec{B} - 2\pi e \rho \vec{p}, \tag{2.13a}$$

$$\partial_t \vec{B} = -\nabla \times \vec{E}, \tag{2.13b}$$

$$\partial_t \vec{p} = 2\pi e \vec{E}. \tag{2.13c}$$

Das Dichteprofil ρ sei von der Form

$$\rho(t, \vec{x}) = \rho(\vec{x}) = \begin{cases} \rho_F, & \text{falls } \vec{x} \in F, \\ 0, & \text{sonst}, \end{cases} \tag{2.14}$$

wobei F das örtliche Gebiet sei, an dem sich die Folie befindet, und ρ_F die hohe Teilchenzahldichte der Folie. In Abbildung 2.3 zeigen wir eindimensionale Darstellungen eines solchen Re-

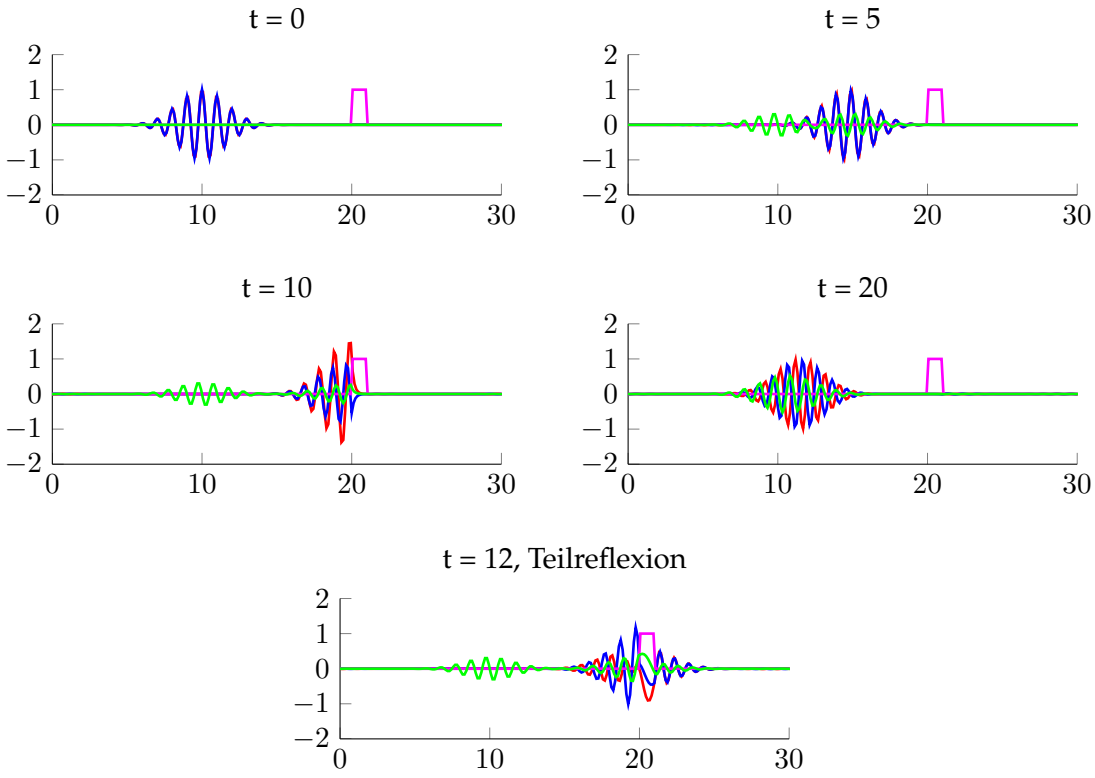


Abbildung 2.3: Lösung der Gleichungen (2.13) mit dem Laserpuls aus (2.11) als Startwert. Grün: p_y, rot: E_y, blau: B_z, schwarz: Hintergrunddichte ρ, der Maximalwert der Dichte ist auf zwei skaliert. Erste Grafik ($t = 0$): Startwert, zweite Grafik ($t = 5$): Propagation des Pulses im Vakuum, dritte Grafik ($t = 10$): Totalreflexion an der Dichtewand, vierte Grafik ($t = 20$): nach der Totalreflexion, fünfte Grafik ($t = 12$): Aufprall des Pulses bei einer Dichte $\rho < \rho_c$ unterhalb der kritischen Dichte, Teilreflexion

flexionsprozesses des Laser-Pulses zu verschiedenen Zeitpunkten, sowohl bei Teil- als auch bei vollständiger Reflexion. Die Impulse, die an der Stelle des Startwertes zurückbleiben, haben keine physikalische Bedeutung, da die Dichte in diesem Bereich verschwindet.

Für das den PIC-Codes zugrundeliegende Störmer-Verlet-Verfahren für die elektromagnetischen Felder ergibt sich in diesem Fall und falls $F = \Omega$, das Plasma also das gesamt Gebiet ausfüllt, die für die Stabilität des Verfahrens notwendige Bedingung

$$\tau \leq \frac{1}{\pi\sqrt{\rho}} \tag{2.15}$$

an die Schrittweite, vgl. etwa [33, Beispiel 3.4].

2.5 Erhaltungsgrößen

Unter einem Erhaltungsgesetz verstehen wir eine partielle Differentialgleichung der Form

$$\partial_t u(t,\vec{x}) - \nabla \cdot f(u(t,\vec{x})) = 0 \tag{2.16}$$

für eine geeignete Funktion f. Eines der bekanntesten Beispiele sind die Euler-Gleichungen für inkompressible Flüssigkeiten, vgl. etwa [69]. Für uns sind solche Gesetze wichtig, die sich aus der zu lösenden Differentialgleichung herleiten lassen. Dazu suchen wir eine Größe $E = E(\vec{p},\vec{E},\vec{B})$, für die

$$\partial_t E = 0, \quad \text{für alle } t \geq 0, \qquad \text{also} \qquad E(t) \equiv E(0), \quad \text{für alle } t \geq 0 \tag{2.17}$$

gilt. Dabei darf E vom Ort abhängen, muss es aber nicht. Ein solches Gesetz haben wir bereits in (2.5), der Erhaltung der Divergenz des magnetischen Flusses $\vec{B}$, gesehen. Wir wollen nun weitere Erhaltungsgrößen aus (2.10) herleiten.

Zunächst wollen wir die klassische Energie des Systems (2.10) ermitteln. Diese ist physikalisch von zentraler Bedeutung. Für den nicht-relativistischen Fall (2.12) ist sie gegeben durch

$$E(t) := \frac{1}{2}\int_\Omega \left\|\vec{E}\right\|^2 + \left\|\vec{B}\right\|^2 + \rho\left\|\vec{p}\right\|^2 \mathrm{d}\vec{x}, \tag{2.18}$$

denn für die zeitliche Ableitung dieses Ausdruckes ergibt sich

$$\partial_t E(t) = \frac{1}{2}\int_\Omega 2\langle\partial_t\vec{E},\vec{E}\rangle + 2\langle\partial_t\vec{B},\vec{B}\rangle + 2\rho\langle\partial_t\vec{p},\vec{p}\rangle \,\mathrm{d}\vec{x} \tag{2.19}$$

$$= \int_\Omega \langle\nabla\times\vec{B},\vec{E}\rangle - \langle\nabla\times\vec{E},\vec{B}\rangle \,\mathrm{d}\vec{x} + 2\pi e\rho\int_\Omega \langle\vec{p},\vec{E}\rangle - \langle\vec{E},\vec{p}\rangle \,\mathrm{d}\vec{x}. \tag{2.20}$$

Dabei verschwindet der zweite Summand wegen der Symmetrie des Skalarproduktes. Für den ersten müssen wir den Satz von Gauß (1.1) mit (A.30) auf $\vec{E}\times\vec{B}$ anwenden. Für alle drei Randbe-

dingungen aus Abschnitt 2.2 verschwindet das Randintegral

$$\int_{\partial\Omega} \langle \vec{E} \times \vec{B}, \vec{\nu} \rangle \, \mathrm{d}\vec{\sigma} = 0. \tag{2.21}$$

Der Vektor $\vec{E} \times \vec{B}$ wird als *Poynting-Vektor* bezeichnet. Er repräsentiert die Energieflussdichte der elektromagnetischen Felder: Durch seine Länge wird das Ausmaß des Energieflusses dargestellt und der Vektor zeigt in Richtung des Energietransports.

Wir wollen noch eine weitere Erhaltungsgröße herleiten, die für die spätere Analyse unseres Verfahrens von Bedeutung ist. Sie kommt eher aus der Betrachtungsweise, die Gleichungen für das elektrische Feld $\vec{E}$ und seine Zeitableitung, die wir für den Moment mit $\vec{F} := \partial_t \vec{E}$ bezeichnen wollen, als ein Hamilton-System zu verstehen. Wir beschränken uns auch hier auf das nichtrelativistische Modell (2.13) mit (2.12) und erhalten

$$\begin{aligned} \partial_{tt}\vec{E} &= \partial_t(\partial_t \vec{E}) = \partial_t \vec{F} \\ &= \partial_t(\nabla \times \vec{B} - 2\pi e \rho \vec{p}) \\ &= -\nabla \times \nabla \times \vec{E} - (2\pi e)^2 \rho \vec{E} \\ &=: -(2\pi e)^2 \omega^2 \vec{E} + g(\vec{E}) \\ &=: -\Omega^2 \vec{E} + g(\vec{E}) \end{aligned} \tag{2.22}$$

mit

$$\Omega(t, \vec{x}) = \Omega = 2\pi e \omega = 2\pi e \sqrt{\rho} \tag{2.23}$$

und

$$g(\vec{E}) = -\nabla \times \nabla \times \vec{E}. \tag{2.24}$$

Dabei ist g, wie auch die Anwendung von Ω, eine lineare Abbildung und wir könnten dies zu einem gemeinsamen linearen Operator zusammenfassen. Wir wollen jedoch den vermeintlich schwieriger zu behandelnden Teil, der vom Dichteterm ρ herrührt, gesondert hervorheben.

Um unsere Gleichungen als Hamilton-System auffassen zu können, müssen wir gewährleisten, dass $-\nabla \times \nabla \times$, aufgefasst als Operator $H^{\nabla\times\nabla\times}(\Omega) \longrightarrow (L^2(\Omega))^3$, selbstadjungiert ist. Wir fordern also

$$\int_\Omega \langle \vec{E}, -\nabla \times \nabla \times \vec{B} \rangle \, \mathrm{d}\vec{x} \overset{!}{=} \int_\Omega \langle -\nabla \times \nabla \times \vec{E}, \vec{B} \rangle \, \mathrm{d}\vec{x}. \tag{2.25}$$

Der Satz von Gauß (1.1) liefert uns den Randterm

$$0 \overset{!}{=} -\int_{\partial\Omega} \langle (\nabla \times \vec{E}) \times \vec{B} + \vec{E} \times (\nabla \times \vec{B}), \nu \rangle \, \mathrm{d}\sigma = \int_{\partial\Omega} \langle (\nabla \times \vec{E}) \times \nu, \vec{B} \rangle - \langle (\nabla \times \vec{B}) \times \nu, \vec{E} \rangle \, \mathrm{d}\sigma, \tag{2.26}$$

welcher im Falle periodischer Randbedingungen verschwindet. Nach [43] ist $-\nabla \times \nabla\times$ auch bei PEC- und PMC-Randbedingungen selbstadjungiert. Der Integrand in (2.26) hat keine dem

Poynting-Vektor $\vec{E} \times \vec{B}$ in (2.21) ähnliche physikalische Bedeutung.

Wir erhalten das Hamilton-System

$$\partial_t \vec{E} = \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\vec{F}}(\vec{E}, \partial_t \vec{E}), \quad \partial_t(\partial_t \vec{E}) = -\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\vec{E}}(\vec{E}, \partial_t \vec{E}) \tag{2.27}$$

mit der Hamilton-Funktion

$$\mathcal{H}(\vec{E}, \vec{F}) = \frac{1}{2}\|\vec{F}\|^2 + \frac{1}{2}\|\Omega\vec{E}\|^2 - \frac{1}{2}\left\langle g(\vec{E}), \vec{E}\right\rangle. \tag{2.28}$$

In Hamilton-Systemen bleibt die Hamilton-Funktion erhalten, also

$$\partial_t \mathcal{H}(\vec{E}(t), \partial_t \vec{E}(t)) = 0. \tag{2.29}$$

KAPITEL 3

SPLITTING-VERFAHREN FÜR LASER-PLASMA-INTERAKTIONEN MIT HOCHDICHTEN PLASMEN

In diesem Kapitel beschreiben wir das in [89, 59] vorgeschlagene – im Moment zunächst ungefilterte – numerische Verfahren, um die Gleichungen (2.10) im nichtrelativistischen Fall (2.12) zu lösen. Für dieses Verfahren ist es wesentlich, dass der relativistische Faktor vernachlässigt und die Dichte konstant gehalten wird. Wir rekapitulieren, dass klassische Fehleranalyse nicht das geeignete Mittel ist, um dieses Verfahren zu analysieren, da sie das Verhalten des Fehlers, das wir beobachten, nicht beschreiben kann. In der Literatur ist ähnliches Verhalten von verwandten Integratoren angewendet auf andere Gleichungen lange bekannt. Durch Einführung von Filterfunktionen werden solche Verfahren dahingehend verbessert, dass gleichmäßige Fehlerschranken möglich werden, vgl. etwa [46, 19, 34, 28]. In [89, 59] wurden solche Filter für das hier betrachtete Verfahren eingeführt, die Wahl dieser Funktionen ist dort jedoch heuristisch motiviert und es existiert bisher keine rigorose Fehleranalyse des gefilterten Verfahrens. Um das Konvergenzverhalten genauer zu verstehen, untersuchen wir, etwa nach dem Vorbild von [34, 28], die physikalische Energie des Systems und in wieweit das numerische Schema diese erhält. Nachdem wir gewisse Resonanzstellen identifizieren können, vereinfachen wir die Gleichungen und das Schema und vergleichen die Energie der numerischen Lösung in zwei aufeinanderfolgenden Zeitschritten in der Hoffnung, dass wir die Resonanzstellen wiederentdecken können. So wollen wir erkennen, an welchen Stellen im numerischen Schema wir gegebenenfalls welche Filterfunktionen ansetzen müssen.

3.1 Ortsdiskretisierung

Wir führen zunächst Bezeichungen für die ortsdiskretisierten Größen der Gleichungen ein. Dabei wollen wir uns nicht auf eine spezielle Ortsdiskretisierung festlegen. Um die Notation nicht zu überladen, bezeichnen wir die ortsdiskretisierten Vektorfelder weiterhin mit $\vec{p}$, $\vec{E}$ und $\vec{B}$. Aus (2.10) im nichtrelativistischen Fall (2.12) möge sich

$$\partial_t \vec{E} = \mathbf{C}_B \vec{B} - \tfrac{1}{c}\boldsymbol{\Omega}^2 \vec{p}, \tag{3.1a}$$

$$\partial_t \vec{B} = -\mathbf{C}_E \vec{E}, \tag{3.1b}$$

$$\partial_t \vec{p} = c\vec{E} \tag{3.1c}$$

ergeben, wobei $\mathbf{C}_B$ die ortsdiskretisierte Version des Rotationsoperators sei, der auf das magnetische Feld $\vec{B}$ angewendet wird, $\mathbf{C}_E$ entsprechend die für das elektrische Feld $\vec{E}$ und $\mathbf{\Omega}$ die Diskretisierung von $c\sqrt{\rho}$ mit $c := 2\pi e$. Die Randbedingungen der kontinuierlichen Gleichungen gehen dabei unter anderem in die Konstruktion von $\mathbf{C}_B$ und $\mathbf{C}_E$ ein. Die diskretisierten Startwerte seien als $\vec{p}(t_0) = \vec{p}_0$, $\vec{E}(t_0) = \vec{E}_0$ und $\vec{B}(t_0) = \vec{B}_0$ gegeben. Wir fassen $\vec{p}$, $\vec{E}$ und $\vec{B}$ dabei als Vektoren im $\mathbb{C}^N$ auf und setzen nur voraus, dass $\mathbf{C}_B$, $\mathbf{C}_E$ und $\mathbf{\Omega}$ zunächst beliebige Matrizen in $\mathbb{C}^{N\times N}$ sind.

Wir wollen eine Fehlerabschätzung für diese ortsdiskretisierte Version der Gleichungen finden, daher verwenden wir zur Fehlermessung in der Theorie die $\|\cdot\| = \|\cdot\|_2$-Norm, benutzen gegebenenfalls die von dieser induzierte Matrixnorm und lassen den Index in diesem Kontext weg.

Hinweis zur Notation: Wir verwenden hier das Symbol $\mathbf{\Omega}$ für die Frequenzmatrix, um uns an die Standardliteratur anzupassen. Dieses ist im Gegensatz zum Gebiet Ω fett gedruckt und sollte nicht verwechselt werden.

3.2 Numerisches Verfahren

Um (3.1) zeitlich zu diskretisieren, benutzen wir ein so genanntes Splitting-Verfahren, wie es in [89, 59] vorgeschlagen wird. Dabei spalten wir die rechte Seite in drei Teile auf und schreiben

$$\partial_t \begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix} = f\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) = f_1\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) + f_2\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) + f_3\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) \tag{3.2}$$

mit

$$f_1\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \\ -\mathbf{C}_E\vec{E} \end{bmatrix}, \tag{3.3a}$$

$$f_2\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) = \begin{bmatrix} 0 \\ \mathbf{C}_B\vec{B} \\ 0 \end{bmatrix}, \tag{3.3b}$$

$$f_3\left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}\right) = \begin{bmatrix} 0 & c & \\ -\frac{1}{c}\mathbf{\Omega}^2 & 0 & \\ & & 0 \end{bmatrix} \begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix}. \tag{3.3c}$$

Da die Dichte ρ und damit die Matrix $\mathbf{\Omega}$ zeitlich konstant ist, können wir die Gleichungen

$$\partial_t \begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix} = f_i \left(\begin{bmatrix} \vec{p} \\ \vec{E} \\ \vec{B} \end{bmatrix} \right), \qquad \begin{aligned} \vec{p}(t_0) &= \vec{p}_0, \\ \vec{E}(t_0) &= \vec{E}_0, \\ \vec{B}(t_0) &= \vec{B}_0 \end{aligned} \tag{3.4}$$

für alle $i = 1, 2, 3$ sehr einfach analytisch lösen. Die Lösung der ersten beiden Gleichungen, mit f_i aus (3.3a) und (3.3b), ist affin und die Gleichung für $i = 3$ mit (3.3c) stellt einen harmonischen Oszillator dar:

$$\begin{aligned}
(3.4) + (3.3a)\ (i = 1) \Rightarrow\ & \vec{p}(t) \equiv \vec{p}_0, \vec{E}(t) \equiv \vec{E}_0, \\
& \vec{B}(t) = \vec{B}_0 - (t - t_0)\mathbf{C}_E \vec{E}_0, \\
(3.4) + (3.3b)\ (i = 2) \Rightarrow\ & \vec{p}(t) \equiv \vec{p}_0, \vec{B}(t) \equiv \vec{B}_0, \\
& \vec{E}(t) = \vec{E}_0 + (t - t_0)\mathbf{C}_B \vec{B}_0, \\
(3.4) + (3.3c)\ (i = 3) \Rightarrow\ & \vec{B}(t) \equiv \vec{B}_0, \\
& \begin{bmatrix} \vec{p} \\ \vec{E} \end{bmatrix}(t) = \exp\left((t - t_0) \begin{bmatrix} 0 & c \\ -\frac{1}{c}\mathbf{\Omega}^2 & 0 \end{bmatrix} \right) \begin{bmatrix} \vec{p} \\ \vec{E} \end{bmatrix}(t_0) \\
& \qquad = \begin{bmatrix} \cos((t - t_0)\mathbf{\Omega}) & (t - t_0)c\,\mathrm{sinc}((t - t_0)\mathbf{\Omega}) \\ -\frac{1}{c}\mathbf{\Omega}\sin((t - t_0)\mathbf{\Omega}) & \cos((t - t_0)\mathbf{\Omega}) \end{bmatrix} \begin{bmatrix} \vec{p}_0 \\ \vec{E}_0 \end{bmatrix}.
\end{aligned}$$

Die Matrix $\mathbf{\Omega}$ in der Lösung der dritten Gleichung spielt dabei die Rolle der Frequenz des harmonischen Oszillators.

Für die autonome gewöhnliche Differentialgleichung ist der *exakte Fluss* oder einfach nur *Fluss* φ_t definiert als die Abbildung, die den Startwert auf die Lösung zum Zeitpunkt t abbildet:

$$\partial_t y(t) = f(y(t)), \quad y(0) = y_0 \qquad \varphi_t(y_0) := y(t) = y(t; y_0). \tag{3.5}$$

Der Ansatz von Splitting-Verfahren ist es, meist exakte, Flüsse von aufgespalteten Gleichungen wie (3.4) ineinander zu verschachteln. Dazu bezeichnen wir die Flüsse von (3.4) mit ξ_t für $i = 1$, ζ_t für $i = 2$ und η_t für $i = 3$. Wir wählen gemäß [89, 59] die folgende, symmetrische Variante: Zuerst machen wir je einen halben Zeitschritt mit den zu (3.3a) und (3.3b) gehörigen Flüssen, dann einen vollen Zeitschritt mit dem zu (3.3c). Zum Schluss werden die beiden Halbschritte in umgekehrter Reihenfolge ausgeführt, also zuerst der zu (3.3b) und dann zu (3.3a). Damit ergibt sich der numerische Fluss des Verfahrens zu $\Psi_\tau = \xi_{\tau/2} \circ \zeta_{\tau/2} \circ \eta_\tau \circ \zeta_{\tau/2} \circ \xi_{\tau/2}$ und daraus das

numerische Schema

$$\vec{B}_{n+\frac{1}{2}} = \vec{B}_n - \tfrac{\tau}{2}\mathbf{C}_E\vec{E}_n \tag{3.6a}$$

$$\vec{E}_n^{+} = \vec{E}_n + \tfrac{\tau}{2}\mathbf{C}_B\vec{B}_{n+\frac{1}{2}} \tag{3.6b}$$

$$\begin{bmatrix} \vec{p}_{n+1} \\ \vec{E}_{n+1}^{-} \end{bmatrix} = \begin{bmatrix} \cos(\tau\mathbf{\Omega}) & \tau c\,\mathrm{sinc}(\tau\mathbf{\Omega}) \\ -\frac{1}{c}\mathbf{\Omega}\sin(\tau\mathbf{\Omega}) & \cos(\tau\mathbf{\Omega}) \end{bmatrix} \begin{bmatrix} \vec{p}_n \\ \vec{E}_n^{+} \end{bmatrix} \tag{3.6c}$$

$$\vec{E}_{n+1} = \vec{E}_{n+1}^{-} + \tfrac{\tau}{2}\mathbf{C}_B\vec{B}_{n+\frac{1}{2}} \tag{3.6d}$$

$$\vec{B}_{n+1} = \vec{B}_{n+\frac{1}{2}} - \tfrac{\tau}{2}\mathbf{C}_E\vec{E}_{n+1} \tag{3.6e}$$

zusammen mit den Startwerten $\vec{p}_0 = \vec{p}(t_0), \vec{E}_0 = \vec{E}(t_0), \vec{B}_0 = \vec{B}(t_0)$, den Zeitschritten $t_{n+1} = t_n + \tau$ und $\vec{p}_n \approx p(t_n)$, $\vec{E}_n \approx \vec{E}(t_n)$ und $\vec{B}_n \approx \vec{B}(t_n)$.

Für ein numerisches Verfahren $y_n \approx y(t_n)$ zur Lösung von (3.5) definieren wir den Fehler zur Zeit t_n als

$$e_n = y(t_n) - y_n. \tag{3.7}$$

Wir sagen, dass ein numerisches Verfahren *Konsistenzordnung* p [34, Definition II.1.2] besitzt, falls der lokale Fehler für eine hinreichend glatte rechte Seite f und alle Startwerte y_0 durch

$$\|e_1\| \leq C\tau^{p+1} \tag{3.8}$$

mit einer Konstante C, die nicht von τ abhängt, beschränkt ist. Wir sagen entsprechend, dass ein numerisches Verfahren *Konvergenzordnung* p besitzt, falls

$$\|e_n\| \leq C\tau^p, \qquad t_0 \leq t_n = t_n + n\tau \leq T \tag{3.9}$$

mit einer Konstante C gleichmäßig in $t_0 \leq t_n \leq T$, die also insbesondere nicht von n oder τ abhängt, erfüllt ist.

Wir haben das Verfahren (3.6) als symmetrisches Verfahren konstruiert. Nach klassischer Fehleranalyse wäre daher zu erwarten, dass dieses ein Schema mit gerader Konsistenzordnung liefert, vgl. etwa [34, Theorem II.3.2]. Da unser System linear ist, die rechte Seite insbesondere Lipschitzstetig, ist nach diesem Standpunkt auch die Fehlerfortpflanzung unproblematisch, vgl. z.B. [35, Lemma I.7.2] für Stabilität des expliziten Euler-Verfahrens in diesem Fall. Aus dieser Perspektive wäre für den Fehler zu erwarten, dass

$$\|\vec{p}(t_n) - \vec{p}_n\| \leq C\tau^2, \qquad \|\vec{E}(t_n) - \vec{E}_n\| \leq C\tau^2, \qquad \|\vec{B}(t_n) - \vec{B}_n\| \leq C\tau^2, \tag{3.10}$$

gilt. Dies wollen wir nun in einem numerischen Test, ähnlich wie in [59, 89], überprüfen.

3.3 Numerisches Experiment

Um das Verfahren (3.6) numerischen Tests zu unterziehen, verkleinern wir zunächst den Rechenaufwand, indem wir das Problem durch Wahl spezieller Startwerte auf eine Raumrichtung reduzieren. Wir werden das Verfahren mit extrem vielen Schrittweiten testen müssen, um eine möglichst gute Auflösung des Verhaltens des Fehlers zu gewährleisten. Daher sollte der Zeitaufwand für jeden einzelnen Durchlauf so gering wie möglich sein. Wir beobachteten bereits in dieser eindimensionalen Variante ein interessantes Verhalten.

Wir simulieren die Anwendung aus 2.4 und vereinfachen dazu die Gleichungen aus (2.13). Die Annahme, um diese auf eine Raumdimension zu reduzieren, ist es, dass alle auftretenden Funktionen und Vektorfelder nur in eine, hier die x-Richtung, variieren. Damit verschwinden alle Ortsableitungen in dazu orthogonalen Richtungen und es ergeben sich zwei entkoppelte Sätze von Gleichungen für p_y, E_y und B_z bzw. p_z, E_z und B_y:

$$\partial_t E_z = \partial_x B_y - 2\pi e \rho p_z, \tag{3.11a}$$

$$\partial_t B_y = \partial_x E_z, \tag{3.11b}$$

$$\partial_t p_z = 2\pi e E_z. \tag{3.11c}$$

Der zweite Satz von Gleichungen entsteht durch das Vertauschen der y- und z-Indizes und eine Vorzeichenumkehrung der beiden Ortsableitungen. Für die x-Komponenten ergibt sich

$$\partial_t B_x = 0, \qquad \partial_t E_x = -2\pi e \rho p_x, \qquad \partial_t p_x = 2\pi e E_x. \tag{3.12}$$

Wir wählen für die Startwerte $B_x(0,\cdot) \equiv E_x(0,\cdot) \equiv p_x(0,\cdot) \equiv 0$ und erhalten daher

$$B_x \equiv E_x \equiv p_x \equiv 0 \tag{3.13}$$

als Lösung für alle Zeiten.

Analog zur allgemeinen Version (2.18) erhalten wir die Energie

$$E(t) := \frac{1}{2}\int_\Omega E_z^2 + B_y^2 + \rho p_z^2 \,\mathrm{d}x + \frac{1}{2}\int_\Omega E_y^2 + B_z^2 + \rho p_y^2 \,\mathrm{d}x. \tag{3.14}$$

Wir wählen periodische Randbedingungen. Als Anfangswert wählen wir $\vec{E}(t_0,\cdot)$ und $\vec{B}(t_0,\cdot)$ als den Laserpuls aus (2.11) und die initialen Impulse als $\vec{p}(t_0,\cdot) \equiv 0$. Das Zentrum des Startpulses legen wir auf $x_0 = 10$ und die Breite auf $\sigma_0 = 10$ fest. Die Hintergrunddichte ρ wählen wir als dünne Wand, da sie die dünne Folie repräsentieren soll, auf die der Puls zufliegt und von der er reflektiert wird. Das Dichteprofil ist in (2.14) gegeben. Wir wählen die Simulationsbox $\Omega = [0, 30]$ und $F = [20, 21]$. Dies sind die Einstellungen, die wir bereits in Abbildung 2.3 verwendet haben.

Die Ortsdiskretisierung erfolgt mit finiten Differenzen der Ordnung zwei für die Rotationsoperatoren auf einem zwischen elektrischem und magnetischem Feld um einen halben Ortsschritt versetzten Gitter, der eindimensionalen Version des so genannten *Yee-Gitters* (vgl. [92]). Dabei handelt es sich um eine sehr weit verbreitete Diskretisierungsmethode für die Feldgrößen der Maxwell-Gleichungen, die unter anderem die Basis für die Feldlöser des VLPL-Codes darstellt und daher die für unsere Analyse kanonische Wahl darstellt. Um nicht interpolieren zu müssen, verwenden wir für die Impulse dasselbe Ortsgitter, wie für das elektrische Feld. Die Ortsschrittweite bezeichnen wir mit h und wählen sie als $h = \frac{1}{8}$. Die Hintergrunddichte wird punktweise auf dem Gitter ausgewertet. Die so entstandene ortsdiskretisierte Gleichung ist von der Form (3.1). Die diskrete Form der Energie (3.14) ist dann durch

$$E_h(t) = \frac{1}{2}\left(\|\vec{E}(t)\|^2 + \|\vec{B}(t)\|^2 + \frac{1}{c^2}\|\mathbf{\Omega}\vec{p}(t)\|^2\right) \tag{3.15}$$

gegeben, da durch die örtliche Versetzung der Gitter der Felder für die beiden Diskretisierungen der Rotationsoperatoren $\mathbf{C}_E = \mathbf{C}_B^T$ gilt.

Abbildung 3.1 auf Seite 28 stellt das Verhalten des Fehlers des numerischen Schemas (3.6) gegen die Zeitschrittweite τ dar. Der Fehler wurde dabei gegen eine Referenzlösung, berechnet mit kleiner Schrittweite, jeweils in der ∞-Norm und der L_2-Norm gemessen. Es werden beide Normen dargestellt, um zu zeigen, dass es sich bei dem dargestellten Verhalten nicht nur um Artefakte einer ungünstigen Wahl der Norm handelt.

In der ersten Zeile ist $\rho = 1.000$ und in der zweiten Zeile $\rho = 10^8$ gewählt. Die Plasmadichte ist also um den Faktor 1.000 beziehungsweise 10^8 größer als die kritische Dichte ρ_c, ab der der Puls reflektiert wird.

Nach klassischer Fehleranalyse wäre zu erwarten, dass die numerische Fehlerkurve gegen die Zeitschrittweite in logarithmischer Darstellung unter einer Geraden mit Steigung zwei liegt. Den Graphen aus Abbildung 3.1 nach zu urteilen, erhalten wir aber Ordnung Null, da es immer wieder Zeitschrittweiten gibt, bei denen der Fehler in einer festen Größenordnung ist. Dieses Verhalten des Fehlers ist bei hochoszillatorischen Differentialgleichungen bekannt und wird als die Entwicklung von *Resonanzen* bezeichnet. Im Fall der niedrigeren Dichte zeigen sich kleine Bereiche mit Ausreißern in der Fehlerkurve. Im Folgenden werden wir versuchen herauszufinden, welche Schrittweiten zu Resonanzen führen. Betrachten wir den Fall der hohen Dichte, ist das Verhalten des Fehlers komplizierter und scheint eher dem Zufall unterworfen zu sein. Hier ist es kaum vorhersehbar, wann eine Schrittweitenwahl einen kleinen Fehler liefert und wann einen etwa auf dem Niveau des Plateaus.

Zunächst werden wir erläutern, wieso die Erwartung an das Verhalten des Verfahrens nicht vollständig zutrifft. Dazu rekapitulieren wir, wie klassische Konvergenzresultate erzielt werden. Dann werden wir ähnliches Verhalten bei in der Literatur bekannten, verwandten Gleichungen und In-

tegratoren zum Vorbild nehmen, um das Resonanzverhalten zu erklären und detailierter zu verstehen.

3.4 Resonanzen

Der klassische Ordungsbegriff bei Einschnittverfahren (3.9) setzt voraus, dass die rechte Seite f hinreichend glatt ist. Die Konstante C aus (3.9) hängt dabei üblicherweise von Schranken von f und seinen Ableitungen ab.

Bei Betrachtung der ortsdiskretisierten Maxwell-Gleichungen aus (3.1) führen zwei der auftretenden Ausdrücke in der rechten Seite möglicherweise zu einer großen Konstante: die diskretisierten Rotationsoperatoren $\mathbf{C}_B$ und $\mathbf{C}_E$ und die Matrix $\mathbf{\Omega}^2$, die die Dichte ρ enthält. Der analytische Rotationsoperator ist unbeschränkt, weshalb eine Diskretisierung im Ort daher potentiell eine große Norm haben kann. Wir verwenden jedoch vergleichsweise grobe Ortsauflösungen, wie etwa im obigen Test mit $h = \frac{1}{8}$, sodass das in der Diskretisierung auftretende Quadrat der Inversen der Gitterweite nicht allzu groß ist. Die Abhängigkeit von der Gitterweite lässt sich nicht eliminieren, da das Verfahren (3.6) für $\rho \equiv 0$ dem Störmer-Verlet-Verfahren entspricht, welches in der klassischen Variante für die zeitliche Propagation in Kombination mit dem Yee-Gitter benutzt wird, vgl. [92]. Von diesem ist bekannt, dass der Zeitschritt Stabilitätsanforderderungen erfüllen muss, die abhängig von der Gitterweite sind, vgl. ebenfalls [92]. Kritisch ist dagegen der Dichteterm. Wir wollen hochdichte Plasmen simulieren können, bei denen die Dichte Werte von $\rho \approx 10^5$ in Teilen des Gitters annimmt, während die Felder und Impulse alle eine Größenordnung von eins haben. Konkret bedeutet dies, dass die rechte Seite f und seine Ableitungen alle die Dichte als Term enthalten. Damit findet sich dieser Faktor auch in Normen der Ableitungen bzw. der Lipschitz-Konstante von f und damit letztendlich auch die Fehlerkonstante C.

Damit widerspricht das Verhalten, welches wir im vorigen Abschnitt beobachtet haben, nicht der theoretischen Vorhersage. Wir erhalten zwar ein Verfahren der Ordnung zwei, die Fehlerkonstante ist jedoch so groß, dass die Gerade der theoretischen Fehlerschranke dadurch sehr weit oben und damit außerhalb des sichtbaren Bereiches der Graphen aus Abbildung 3.1 liegt. Für sehr viel kleinere Schrittweiten kann man erwarten, dass der Fehler dann tatsächlich klein wird. Damit ist der Vorteil des exponentiellen Integrators, den dieser vor klassischen expliziten Verfahren hat, zunächst wieder hinfällig. Diese unterliegen ebenfalls einer von der Dichte abhängigen Schrittweiteneinschränkung, z.B. (2.15) im Falle von Störmer-Verlet. Wenn wir mit dem Zeitschritt des exponentiellen Integrators in eine ähnliche Größenordnung hinunter gehen müssen, haben wir also nichts gewonnen.

Die Fehlerkurven in Abbildung 3.1 liegen jedoch außerhalb der Resonanzbereiche trotzdem unter einer Referenzkurve für Ordnung zwei, die keine große Verschiebung nach oben aufweist, was einer kleinen Fehlerkonstante entsprechen würde. Ein solches Verhalten ist in der Literatur bekannt

und es werden Filterfunktionen eingesetzt, um Resonanzen in den Fehlerkurven zu verhindern, vgl. etwa [46, 19, 34, 28]. Die Platzierung und Wahl dieser Funktionen ist jedoch in unserem Schema a priori nicht klar.

Im Folgenden werden wir versuchen ein Verständnis dafür zu entwickeln, wie sich die Bereiche von Schrittweiten verteilen, bei denen der Fehler des Verfahrens Resonanzen zeigt. Darauf aufbauend identifizieren wir diese Stellen mit entsprechendem Verhalten in verschieden Erhaltungsgrößen der Gleichungen und erkennen dabei ein passendes Muster in der Energie. Dann wollen mit einer Analyse des Verhaltens der Energie über einen Zeitschritt des numerischen Verfahrens eine passende Wahl und Platzierung der Filterfunktionen erkennen.

Eine der wichtigsten dieser Erhaltungsgrößen ist die Divergenz des magnetischen Feldes $\vec{B}$. Die Gleichung (2.6), $\nabla \cdot \vec{B} = 0$, ist fundamental für die Behandlung von elektromagnetischen Feldern. Diese ist jedoch in einer Raumdimension immer gewährleistet, da wir y- und z-Komponenten betrachten, die nur in x-Richtung variieren, und die x-Komponente Null gewählt wurde.

Die nächste wichtige Erhaltungsgröße ist die physikalische Energie E, die wir in (3.14) hergeleitet haben. Zu ihrer Untersuchung haben wir uns wieder der Situation des Beispiels aus Abschnitt 3.3 bedient. Dieses Mal haben wir allerdings nicht die Zeitschrittweite, sondern die maximale Hintergrunddichte ρ variiert. Eine solche Analyse ist in der Literatur üblich, vgl. etwa [34, XIII.2.5] oder [28], da sie Hinweise darauf gibt, für welche Bereiche von Werten der Dichte es zu Resonanzen kommen kann. Die Werte der Dichte hängen mit Frequenzen im Sinne des harmonischen Oszillators (3.4) für $i = 3$, mit denen die Lösung in der Zeit schwingt, zusammen, daher suchen wir auf diese Weise nach *Resonanzfrequenzen* von Algorithmus (3.22). Die Zeitschrittweite ist fest als $\tau = 0.02$ gewählt, die maximale Hintergrunddichte liegt einmal bei etwa 1.000 und einmal bei etwa 10^8. Dabei wurde die Dichte jeweils soweit variiert, dass sich das Produkt aus Zeitschrittweite und Frequenz, $\tau c\omega = \tau c\sqrt{\rho}$, in einem Intervall der Länge 4π bewegt. Zur Orientierung sind bei Vielfachen von 2π vertikale Linien eingezeichnet. Die Ergebnisse finden sich in Abbildung 3.2 auf Seite 29.

Die ersten beiden Zeilen von Abbildung 3.2 zeigen das Verhalten der diskreten Version (3.15) der Energie (3.14) und deren Fehler. Hier erkennen wir deutliche Ausschläge, bei denen dieser Fehler um mehrere Größenordnungen wächst.

In der letzten Zeile dieser Abbildung haben wir zum Vergleich den Fehler der ortsdiskretisierten Lösung aufgetragen. Dieses Mal wurde die Referenzlösung mit dem `expm` Befehl aus `MATLAB` bestimmt. Hier bestätigt sich, dass der Ortsfehler an genau den Resonanzstellen der Energie, bei den Vielfachen von 2π, ebenfalls Resonanzen entwickelt. Es erscheint also lohnenswert, das Verhalten des Verfahrens (3.6) im Bezug auf die Energie von (3.14) zu untersuchen und gegebenenfalls Filterfunktionen einzuführen, um die Resonanzen zu verhindern. Interessant ist, dass die Resonanzen nur bei Vielfachen von 2π, nicht etwa allen Vielfachen von π auftreten, wie es beispielsweise in [34, XIII.2.5] für das Fermi-Pasta-Ulam-Problem und die dortigen Integratoren der Fall ist.

3.5 Vereinfachte numerische Energieerhaltung

Wir untersuchen nun die numerische Erhaltung der Energie (3.14). Dazu wollen wir die Gleichungen zunächst etwas übersichtlicher gestalten. Wir betrachten nur das E_z/B_y-Paar und die p_z-Impulse und setzen $E_x \equiv B_x \equiv E_y \equiv B_z \equiv p_x \equiv p_y \equiv 0$. Des Weiteren nehmen wir an, dass die Dichte ρ räumlich konstant ist. Als Nächstes fordern wir periodische Randbedingungen und führen formal eine Fourier-Transformation der Feldgrößen durch. Sei etwa

$$E_z(t,x) = \sum_{k\in\mathbb{Z}} u_k(t) e^{\mathring{\imath}k\frac{2\pi}{|\Omega|}x}, \tag{3.16}$$

mit der Größe $|\Omega|$ des Gebietes in x-Richtung, also der Länge unseres Intervalls. Setzen wir weiterhin voraus, dass E_z hinreichend glatt ist, etwa $\sum_k |k u_k(t)|^2 < \infty$ für alle t, so ergibt sich für die x-Ableitung

$$\partial_x E_z = \sum_{k\in\mathbb{Z}} \mathring{\imath}k\frac{2\pi}{|\Omega|} u_k(t) e^{\mathring{\imath}k\frac{2\pi}{|\Omega|}x} = \sum_{k\in\mathbb{Z}} \mathrm{d}_k u_k(t) e^{\mathring{\imath}k\frac{2\pi}{|\Omega|}x}, \tag{3.17}$$

wobei die $\mathrm{d}_k := \mathring{\imath}k\frac{2\pi}{|\Omega|} \in \mathring{\imath}\mathbb{R}$ rein imaginär sind und den Ortsfrequenzen entsprechen. Die Fourier-Koeffizienten von B_y nennen wir v_k und die für p_z seien mit p_k bezeichnet. Dies ist eine kurze Notationsüberladung, die wir aber sogleich wieder auflösen: Wenn wir die Fourier-Reihen für E_z und B_y in (3.11b) einsetzen, ergibt sich

$$\sum_{k\in\mathbb{Z}} \partial_t v_k(t) e^{\mathring{\imath}k\frac{2\pi}{|\Omega|}x} = \sum_{k\in\mathbb{Z}} \mathrm{d}_k u_k(t) e^{\mathring{\imath}k\frac{2\pi}{|\Omega|}x} \qquad \Longleftrightarrow \qquad \partial_t v_k(t) = \mathrm{d}_k u_k(t) \quad \forall\, k \in \mathbb{Z}. \tag{3.18}$$

Dabei nutzen wir die Tatsache, dass die Wellenfunktionen $x \longmapsto e^{\mathring{\imath}k\frac{2\pi}{|\Omega|}x}$ ein vollständiges Orthogonalsystem von $L^2_{per}(\Omega)$ bilden, insbesondere also linear unabhängig sind. Damit entkoppeln die verschiedenen Summanden der Gleichung. Nach dem gleichen Prinzip behandeln wir die anderen Gleichungen (3.11). Betrachten wir nun nur eine einzige dieser Ortsfrequenzen, können wir den k-Index weglassen und erhalten das gekoppelte skalare System von Gleichungen

$$\partial_t u(t) = \mathrm{d} v(t) - c\omega^2 p(t), \tag{3.19a}$$

$$\partial_t v(t) = \mathrm{d} u(t), \tag{3.19b}$$

$$\partial_t p(t) = c u(t), \tag{3.19c}$$

wie gehabt mit $c = 2\pi e$. Die Ortsableitung geht also in eine Multiplikation mit der rein imaginären Zahl d über. Wie für den eindimensionalen Fall der Maxwell-Gleichungen haben wir keinen Vorzeichenwechsel vor diesem „Differentialoperator“. Schreiben wir das System in eine Differentialgleichung zweiter Ordnung für u um, so ergibt sich der Faktor $\mathrm{d}^2 \in \mathbb{R}_{\leq 0}$ passend zum negativ semidefiniten $-\nabla\times\nabla\times$-Operator aus der entsprechenden Umschrift von (2.13).

$\sqrt{\rho} = \omega^2$ ist wieder als groß angenommen und liefert die Zeitfrequenz. Analog zu (3.14) erhalten wir die Energie

$$E(t) = |u(t)|^2 + |v(t)|^2 + \omega^2 |p(t)|^2, \tag{3.20}$$

wobei wir wegen $\mathrm{d} \in \mathring{\mathrm{i}}\mathbb{R}$ die Identität

$$\mathrm{Re}(\mathring{\mathrm{i}}\,\mathrm{Im}(\mathrm{d})(\overline{u}v + u\overline{v})) = -\,\mathrm{Im}(\mathrm{d})(\mathrm{Im}(\overline{u}v) + \mathrm{Im}(u\overline{v})) = 0 \tag{3.21}$$

benutzen können.

Das Dreifach-Splitting (3.6) sieht dann wie folgt aus:

$$v_{n+\frac{1}{2}} = v_n + \tfrac{\tau}{2}\,\mathrm{d}_u u_n, \tag{3.22a}$$

$$u_n^+ = u_n + \tfrac{\tau}{2}\,\mathrm{d}_v v_{n+\frac{1}{2}}, \tag{3.22b}$$

$$\begin{bmatrix} p_{n+1} \\ u_{n+1}^- \end{bmatrix} = \begin{bmatrix} \cos(\tau c\omega) & \tau c\,\mathrm{sinc}(\tau c\omega) \\ -\omega\sin(\tau c\omega) & \cos(\tau c\omega) \end{bmatrix} \begin{bmatrix} p_n \\ u_n^+ \end{bmatrix}, \tag{3.22c}$$

$$u_{n+1} = u_{n+1}^- + \tfrac{\tau}{2}\,\mathrm{d}_v v_{n+\frac{1}{2}}, \tag{3.22d}$$

$$v_{n+1} = v_{n+\frac{1}{2}} + \tfrac{\tau}{2}\,\mathrm{d}_u u_{n+1}. \tag{3.22e}$$

Um später nachhalten zu können, welchen der Differentialoperatoren wir angewendet haben, unterscheiden wir diese durch den Index u bzw. v.

Als nächstes setzen wir dieses Schema in die Energie (3.20) ein. Dazu nennen wir die numerische Energie zum Zeitpunkt t_n:

$$E_n = |u_n|^2 + |v_n|^2 + \omega^2 |p_n|^2 \tag{3.23}$$

und wollen überprüfen, wie sich E_{n+1} im Verhältnis zu E_n verhält. Um die Berechnungen übersichtlich zu halten, lassen wir die Argumente der trigonometrischen Funktionen $\tau c\omega$ weg, da diese immer dieselben sind. Wir setzen jetzt sukzessive das Schema (3.22) von unten nach oben in die Energie E_{n+1} ein.

Wir haben nach (3.22e)

$$|v_{n+1}|^2 = \left(v_{n+\frac{1}{2}} + \frac{\tau\,\mathrm{d}_u}{2} u_{n+1}\right) \overline{\left(v_{n+\frac{1}{2}} + \frac{\tau\,\mathrm{d}_u}{2} u_{n+1}\right)} \tag{3.24}$$

$$= |v_{n+\frac{1}{2}}|^2 + \frac{\tau}{2} 2\,\mathrm{Re}(v_{n+\frac{1}{2}} \overline{\mathrm{d}_u\, u_{n+1}}) + \frac{\tau^2 |\mathrm{d}_u|^2}{4} |u_{n+1}|^2 \tag{3.25}$$

$$= |v_{n+\frac{1}{2}}|^2 - \tau\,\mathrm{Im}(\mathrm{d}_u)\,\mathrm{Im}(u_{n+1}\overline{v_{n+\frac{1}{2}}}) + \frac{\tau^2 |\mathrm{d}_u|^2}{4} |u_{n+1}|^2, \tag{3.26}$$

wobei wir wieder wegen $\mathrm{d}_u \in \mathring{\mathrm{i}}\mathbb{R}$ diesmal

$$\mathrm{Re}(v\overline{\mathrm{d}_u u}) = \mathrm{Re}(v(-\mathring{\mathrm{i}})\,\mathrm{Im}(\mathrm{d}_u)\overline{u}) = \mathrm{Im}(\mathrm{d}_u)\,\mathrm{Im}(v\overline{u}) = -\,\mathrm{Im}(\mathrm{d}_u)\,\mathrm{Im}(u\overline{v}) \tag{3.27}$$

benutzen. Weiterhin erhalten wir mit (3.22d)

$$\operatorname{Im}(u_{n+1}\overline{v_{n+\frac{1}{2}}}) = \operatorname{Im}(u^-_{n+1}\overline{v_{n+\frac{1}{2}}}) + \frac{\tau \operatorname{Im}(\mathrm{d}_v)}{2}|v_{n+\frac{1}{2}}|^2. \tag{3.28}$$

Als nächstes ergibt sich analog zur ersten Rechnung ebenfalls aus (3.22d)

$$|u_{n+1}|^2 = |u^-_{n+1}|^2 + \tau \operatorname{Im}(\mathrm{d}_v) \operatorname{Im}(u^-_{n+1}\overline{v_{n+\frac{1}{2}}}) + \frac{\tau^2|\mathrm{d}_v|^2}{4}|v_{n+\frac{1}{2}}|^2. \tag{3.29}$$

Die folgenden Terme ergeben sich aus der zweiten Zeile von (3.22c) zu

$$|u^-_{n+1}|^2 = \omega^2 \sin^2 |p_n|^2 - 2\omega \sin\cos \operatorname{Re}(u^+_n\overline{p_n}) + \cos^2 |u^+_n|^2 \tag{3.30a}$$

und

$$\operatorname{Im}(u^-_{n+1}\overline{v_{n+\frac{1}{2}}}) = \omega \sin \operatorname{Im}(v_{n+\frac{1}{2}}\overline{p_n}) + \cos \operatorname{Im}(u^+_n\overline{v_{n+\frac{1}{2}}}). \tag{3.30b}$$

Jetzt eliminieren wir p_{n+1} mithilfe der ersten Zeile von (3.22c) zu

$$|p_{n+1}|^2 = \cos^2 |p_n|^2 + 2\tau c \operatorname{sinc}\cos \operatorname{Re}(u^+_n\overline{p_n}) + \tau^2 c^2 \operatorname{sinc}^2 |u^+_n|^2. \tag{3.31}$$

Fahren wir mit (3.22b) fort, erhalten wir

$$|u^+_n|^2 = |u_n|^2 + \tau \operatorname{Im}(\mathrm{d}_v) \operatorname{Im}(u_n\overline{v_{n+\frac{1}{2}}}) + \frac{\tau^2|\mathrm{d}_v|^2}{4}|v_{n+\frac{1}{2}}|^2, \tag{3.32a}$$

$$\operatorname{Im}(u^+_n\overline{v_{n+\frac{1}{2}}}) = \operatorname{Im}(u_n\overline{v_{n+\frac{1}{2}}}) + \frac{\tau \operatorname{Im}(\mathrm{d}_v)}{2}|v_{n+\frac{1}{2}}|^2, \tag{3.32b}$$

$$\operatorname{Re}(u^+_n\overline{p_n}) = \operatorname{Re}(u_n\overline{p_n}) - \frac{\tau \operatorname{Im}(\mathrm{d}_v)}{2} \operatorname{Im}(v_{n+\frac{1}{2}}\overline{p_n}). \tag{3.32c}$$

Die erste Zeile des Schemas, (3.22a), liefert die restlichen Größen

$$|v_{n+\frac{1}{2}}|^2 = |v_n|^2 - \tau \operatorname{Im}(\mathrm{d}_u) \operatorname{Im}(u_n\overline{v_n}) + \frac{\tau^2|\mathrm{d}_u|^2}{4}|u_n|^2, \tag{3.33a}$$

$$\operatorname{Im}(u_n\overline{v_{n+\frac{1}{2}}}) = \operatorname{Im}(u_n\overline{v_n}) - \frac{\tau \operatorname{Im}(\mathrm{d}_u)}{2}|u_n|^2, \tag{3.33b}$$

$$\operatorname{Im}(v_{n+\frac{1}{2}}\overline{p_n}) = \operatorname{Im}(v_n\overline{p_n}) + \frac{\tau \operatorname{Im}(\mathrm{d}_u)}{2} \operatorname{Re}(u_n\overline{p_n}). \tag{3.33c}$$

Nun müssen wir diese Größen sukzessive in die numerische Energie zum Zeitpunkt $n+1$ einsetzen. Dazu verwenden wir die folgende Farbkodierung, um verschiedenes Auftreten von Koeffizienten zusammenzufassen: u, v, $u \cdot v$, p, $u \cdot p$ und $v \cdot p$. Die Größen können jeweils verschiedene Indizes haben, es tritt aber in jeder Zeile der Rechnung jeweils immer nur ein Index auf. Die Rechnung findet sich in (3.34) der Übersichtlichkeit halber auf einer eigenen Seite, Seite 31.

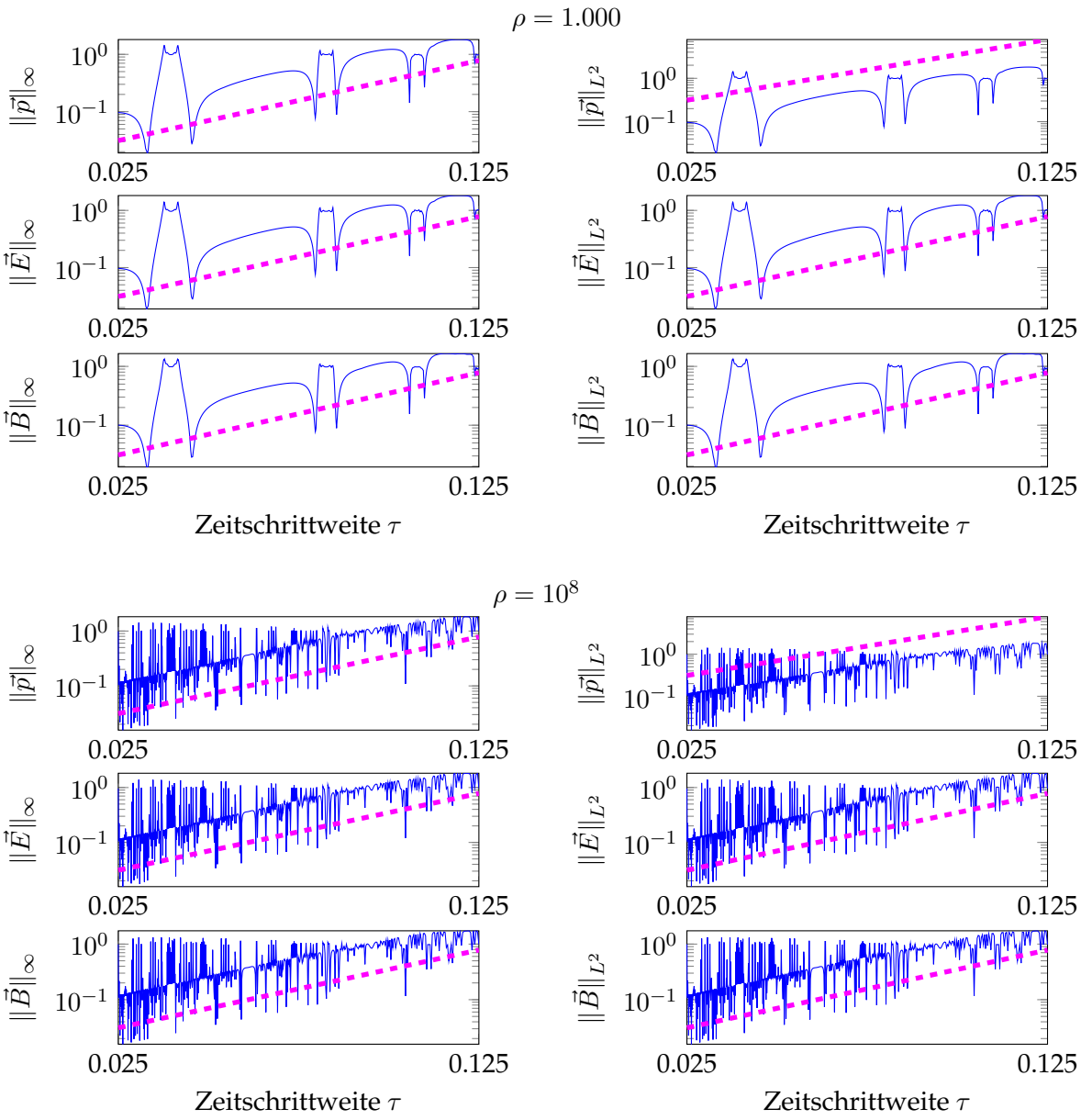


Abbildung 3.1: Darstellung des absoluten Fehlers der Impulse und der elektrischen und magnetischen Felder bei Anwendung von Verfahren (3.6) angewendet auf (3.11) mit dem Yee-Gitter zur Ortsdiskretisierung in Relation zur Zeitschrittweite τ. Die linken Abbildungen zeigen jeweils den Fehler in der ∞-Norm, die rechten entsprechend in der L_2-Norm. Als Referenz ist gestrichelt eine Kurve für Ordnung zwei dargestellt. Erste Zeile: Verhalten des Fehlers für eine Dichte von $\rho = 1.000$ in der Dichtewand, zweite Zeile: Verhalten des Fehlers für eine Dichte von $\rho = 10^8$ in der Dichtewand.

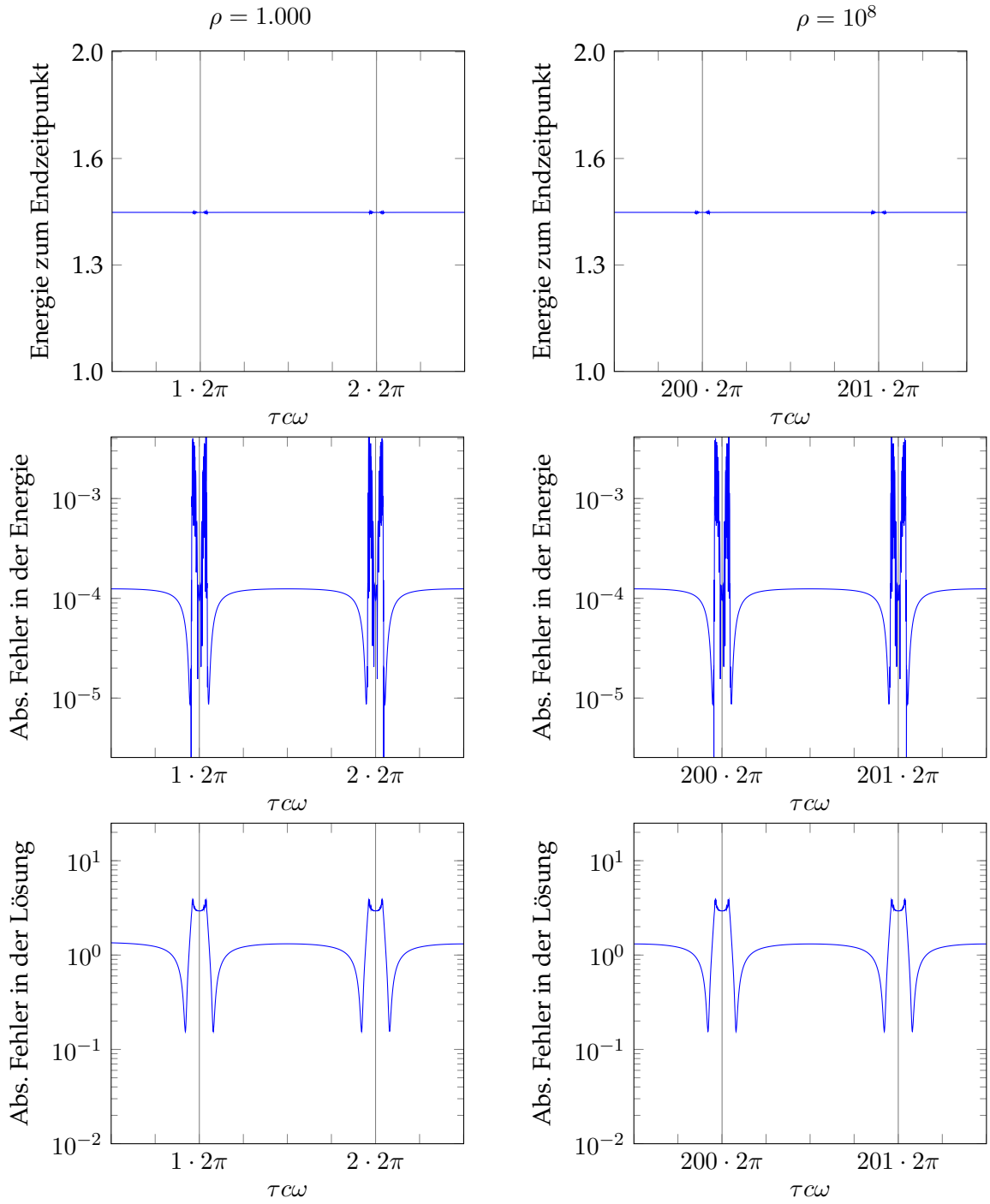


Abbildung 3.2: Darstellung verschiedener Größen bei Anwendung von Verfahren (3.6) angewendet auf (3.11) mit dem Yee-Gitter zur Ortsdiskretisierung für feste Zeitschrittweite τ in Relation zur Frequenz $c\omega$ und damit zur Dichte $\rho = \omega^2$. Die linken Abbildungen zeigen jeweils den Fehler in der ∞-Norm, die rechten jeweils in der L_2-Norm. Erste Spalte: Verhalten des Fehlers für eine Dichte von $\rho = 1000 n_c$ in der Dichtewand, zweite Spalte: Verhalten des Fehlers für eine Dichte von $\rho = 10^8 n_c$ in der Dichtewand.

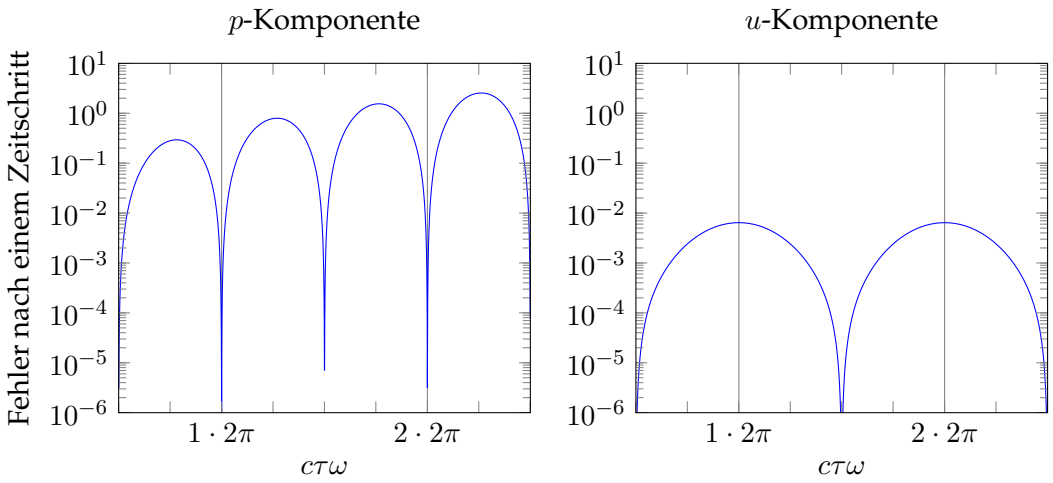


Abbildung 3.3: Koeffizienten des Fehlers der Energie (3.20) nach einem Schritt mit dem numerischen Verfahren (3.22), $\rho = 10^3$, $\mathrm{d} = 2\,\mathring{\mathrm{u}}$. Die Zeitschrittweite ist wieder bei $\tau = 0.02$ festgehalten und ω läuft, sodass $\tau c\omega$ im entsprechenden Intervall verläuft.

$$E_{n+1} = |u_{n+1}|^2 + |v_{n+1}|^2 + \omega^2 |p_{n+1}|^2 \tag{3.34a}$$

$$\overset{(3.26)}{=} |u_{n+1}|^2 + |v_{n+\frac{1}{2}}|^2 - \tau \operatorname{Im}(\mathrm{d}_u) \operatorname{Im}(u_{n+1}\overline{v_{n+\frac{1}{2}}}) + \frac{\tau^2|\mathrm{d}_u|^2}{4}|u_{n+1}|^2 + \omega^2|p_{n+1}|^2 \tag{3.34b}$$

$$\overset{(3.28)}{\overset{(3.29)}{=}} \left(1 + \frac{\tau^2|\mathrm{d}_u|^2}{4}\right)\left(|u_{n+1}^-|^2 + \tau \operatorname{Im}(\mathrm{d}_v)\operatorname{Im}(u_{n+1}^-\overline{v_{n+\frac{1}{2}}}) + \frac{\tau^2|\mathrm{d}_v|^2}{4}|v_{n+\frac{1}{2}}|^2\right) - \tau \operatorname{Im}(\mathrm{d}_u)\left(\operatorname{Im}(u_{n+1}^-\overline{v_{n+\frac{1}{2}}}) + \frac{\tau \operatorname{Im}(\mathrm{d}_v)}{2}|v_{n+\frac{1}{2}}|^2\right) + |v_{n+\frac{1}{2}}|^2 + \omega^2|p_{n+1}|^2 \tag{3.34c}$$

$$\begin{aligned}\overset{(3.30)}{\overset{(3.31)}{=}} &\left(1 + \frac{\tau^2|\mathrm{d}_u|^2}{4}\right)\left(\omega^2 \sin^2|p_n|^2 - 2\omega \sin\cos \operatorname{Re}(u_n^+\overline{p_n}) + \cos^2|u_n^+|^2\right) + \left(\tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u) + \frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}\right)\left(\omega \sin \operatorname{Im}(v_{n+\frac{1}{2}}\overline{p_n}) + \cos \operatorname{Im}(u_n^+\overline{v_{n+\frac{1}{2}}})\right) \\ &+ \left(1 - \frac{\tau^2 \operatorname{Im}(\mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{2} + \frac{\tau^2|\mathrm{d}_v|^2}{4} + \frac{\tau^4|\mathrm{d}_u|^2|\mathrm{d}_v|^2}{16}\right)|v_{n+\frac{1}{2}}|^2 + \omega^2\left(\cos^2|p_n|^2 + 2\tau c \operatorname{sinc}\cos \operatorname{Re}(u_n^+\overline{p_n}) + \tau^2c^2 \operatorname{sinc}^2|u_n^+|^2\right)\end{aligned} \tag{3.34d}$$

$$\begin{aligned}\overset{(3.32)}{=} &\left(\underbrace{\cos^2 + \overbrace{\omega^2\tau^2c^2 \operatorname{sinc}^2}^{\sin^2}}_{=1} + \frac{\tau^2|\mathrm{d}_u|^2}{4}\cos^2\right)\left(|u_n|^2 + \tau \operatorname{Im}(\mathrm{d}_v)\operatorname{Im}(u_n\overline{v_{n+\frac{1}{2}}}) + \frac{\tau^2|\mathrm{d}_v|^2}{4}|v_{n+\frac{1}{2}}|^2\right) + \left(\tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u) + \frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}\right)\cos\left(\operatorname{Im}(u_n\overline{v_{n+\frac{1}{2}}}) + \frac{\tau \operatorname{Im}(\mathrm{d}_v)}{2}|v_{n+\frac{1}{2}}|^2\right) \\ &+ \left(1 - \frac{\tau^2 \operatorname{Im}(\mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{2} + \frac{\tau^2|\mathrm{d}_v|^2}{4} + \frac{\tau^4|\mathrm{d}_u|^2|\mathrm{d}_v|^2}{16}\right)|v_{n+\frac{1}{2}}|^2 + \left(\underbrace{-2\omega \sin\cos + 2\omega\overbrace{\omega\tau c \operatorname{sinc}}^{=\sin}\cos}_{=0} - \frac{\tau^2|\mathrm{d}_u|^2}{2}\omega \sin\cos\right)\left(\operatorname{Re}(u_n\overline{p_n}) - \frac{\tau \operatorname{Im}(\mathrm{d}_v)}{2}\operatorname{Im}(v_{n+\frac{1}{2}}\overline{p_n})\right) \\ &+ \left(\tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u) + \frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}\right)\omega \sin \operatorname{Im}(v_{n+\frac{1}{2}}\overline{p_n}) + \left(\omega^2\underbrace{(\sin^2 + \cos^2)}_{=1} + \frac{\tau^2|\mathrm{d}_u|^2}{4}\omega^2 \sin^2\right)|p_n|^2\end{aligned} \tag{3.34e}$$

$$\begin{aligned}\overset{(3.33)}{=} &\left(1 + \frac{\tau^2|\mathrm{d}_u|^2}{4}\cos^2\right)|u_n|^2 + \left(\tau \operatorname{Im}(\mathrm{d}_v) + \frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}\cos(1+\cos) + \tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u)\cos\right)\left(\operatorname{Im}(u_n\overline{v_n}) - \frac{\tau \operatorname{Im}(\mathrm{d}_u)}{2}|u_n|^2\right) \\ &+ \left(1 - \frac{\tau^2 \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{2}(1+\cos) + \frac{\tau^4|\mathrm{d}_u|^2|\mathrm{d}_v|^2}{16}(1+\cos)^2\right)\left(|v_n|^2 - \tau \operatorname{Im}(\mathrm{d}_u)\operatorname{Im}(u_n\overline{v_n}) + \frac{\tau^2|\mathrm{d}_u|^2}{4}|u_n|^2\right) - \left(\frac{\tau^2|\mathrm{d}_u|^2}{2}\omega \sin\cos\right)\operatorname{Re}(u_n\overline{p_n}) \\ &+ \left(\frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}\omega \sin(1+\cos) + (\tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u))\,\omega \sin\right)\left(\operatorname{Im}(v_n\overline{p_n}) + \frac{\tau \operatorname{Im}(\mathrm{d}_u)}{2}\operatorname{Re}(u_n\overline{p_n})\right) + \left(1 + \frac{\tau^2|\mathrm{d}_u|^2}{4}\sin^2\right)\omega^2|p_n|^2\end{aligned} \tag{3.34f}$$

$$\begin{aligned}= &\left(1 - \frac{\tau^2 \operatorname{Im}(\mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{2}(1+\cos) + \frac{\tau^2|\mathrm{d}_u|^2}{4}\left((1+\cos)^2\left(1 - \frac{\tau^2 \operatorname{Im}(\mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{4}\right)^2 + \frac{\tau^2|\mathrm{d}_v|^2}{2}(1+\cos)\right)\right)|u_n|^2 + \left(\tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u)(1+\cos) + \frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}(\cos-2)(\cos+1)\right. \\ &\left.+ \frac{\tau^3 \operatorname{Im}(\mathrm{d}_u)|\mathrm{d}_v|^2}{2}(1+\cos)\right)\operatorname{Im}(u_n\overline{v_n}) + \left(1 - \frac{\tau^2 \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{2}(1+\cos) + \frac{\tau^4|\mathrm{d}_u|^2|\mathrm{d}_v|^2}{16}(1+\cos)^2\right)|v_n|^2 + \left(\frac{\tau^2 \operatorname{Im}(\mathrm{d}_u)\operatorname{Im}(\mathrm{d}_v)}{2} - \frac{\tau^2|\mathrm{d}_u|^2}{2}(1+\cos)\right)\omega \sin \operatorname{Re}(u_n\overline{p_n}) \\ &+ \left(\tau \operatorname{Im}(\mathrm{d}_v - \mathrm{d}_u) + \frac{\tau^3|\mathrm{d}_u|^2 \operatorname{Im}(\mathrm{d}_v)}{4}(1+\cos)\right)\omega \sin \operatorname{Im}(v_n\overline{p_n}) + \left(1 + \frac{\tau^2|\mathrm{d}_u|^2}{4}\sin^2\right)\omega^2|p_n|^2\end{aligned} \tag{3.34g}$$

Subtrahieren wir nun E_n von E_{n+1}, um den Fehler in der Energie über einen Zeitschritt zu bestimmen, verschwinden die Einsen in den Koeffizienten vor $|u_n|^2$, $|v_n|^2$ und $|p_n|^2$, grau markiert in (3.34g).

Zur Verifikation wurden MAPLE-Plots der kontinuierlichen Formel mit den MATLAB-Codes für den Zeitintegrator verglichen. So ist zumindest sichergestellt, dass die dominierenden Terme übereinstimmen.

Der auf den ersten Blick größte Fehlerterm ist dabei der vor $|p_n|^2$, nämlich $\frac{\tau^2|\mathrm{d}_u|^2}{4}\omega^2\sin^2(\tau c\omega)$. Der Faktor ω^2 kann dabei nicht von τ^2 ausgeglichen werden. In der Simulation in 3.3 mit $\rho = 10^8$ haben wir etwa $\omega \sim \tau^{-2}$, sodass ein Faktor der Größenordnung ω übrig bleibt. Allerdings hat die Sinusfunktion ihre Maxima bei $\tau c\omega = \pi/2 + k\pi$, $k \in \mathbb{Z}$ und sogar Nullstellen bei $k2\pi$, $k \in \mathbb{N}_{>0}$, bei denen wir die Resonanzen in Abbildung 3.1 festgestellt haben. Im der linken Grafik von Abbildung 3.3 ist der Verlauf dieses Koeffizienten des Fehlers aufgezeichnet. Der Fehler wächst mit der Frequenz, hat aber Nullstellen bei den vermuteten Resonanzstellen, die durch einen senkrechten grauen Strich markiert sind.

Ähnliches gilt auch für die Koeffizienten der Mischterme $\mathrm{Re}(u_n\overline{p_n})$ und $\mathrm{Im}(v_n\overline{p_n})$. Diese sind ebenfalls potenziell noch groß, haben aber Nullstellen an den beobachteten Resonanzstellen.

Die Koeffizienten vor $|u_n|^2$ und $|v_n|^2$ und dem letzten Mischterm $\mathrm{Im}(u_n\overline{v_n})$ zeigen ein interessanteres Verhalten. Das rechte Bild von Abbildung 3.3 zeigt exemplarisch den Verlauf des $|u_n|^2$-Koeffizienten. Die Führungskoeffizienten enthalten alle einen Faktor $(1+\cos(\tau c\omega))$, dessen Betrag sein Maximum bei Vielfachen von 2π hat, also genau an den Resonanzstellen. Allerdings wachsen diese nicht mit ω. Dies legt den Verdacht nahe, dass die Resonanzen durch die Feldkomponenten entstehen und diese dann über die Zeit durch die Kopplung mit den Impulsen durch die Hintergrunddichte verstärkt werden.

Insgesamt zeigt diese lokale Betrachtung des Fehlers in der Energie über einen Zeitschritt also ein anderes Verhalten als die globale über die gesamte Simulationszeit. Dies gibt uns einerseits den Hinweis, dass lokale Betrachtungen allein hier nicht ausreichend sind, andererseits, dass die Interaktion zwischen den Feldkomponenten $\vec{E}$ und $\vec{B}$, bzw. hier u und v, mit den Impulsen $\vec{p}$ bzw. p von großer Bedeutung ist.

Bei dieser Analyse ist uns bewusst geworden, wie nahe unser Integrator mit den Verfahren aus [34, XIII.2.5] und [28] verwandt ist. Dadurch wurde der Fokus auf das Umschreiben der Gleichungen (2.13) in eine Wellengleichung (2.22) und die Hamilton-Funktion (2.28) gelenkt. Es wird sich im weiteren Verlauf der Arbeit herausstellen, dass dies ein deutlich besserer Zugang zur Analyse des Zeitintegrators ist. Die Verwandtschaft geht sogar so weit, dass wir unser Verfahren in die Form aus [34, XIII.2.5] umschreiben können. Die physikalische Laserenergie (2.18), die wir bisher betrachtet haben, scheint bei der Analyse hingegen nicht hilfreich zu sein.

KAPITEL 4
FEHLERABSCHÄTZUNG FÜR DAS SPLITTING-VERFAHREN

In diesem Kapitel werden wir eine andere Herangehensweise verwenden, um das Verhalten des Splitting-Verfahrens (3.6) für die Gleichungen (2.13) zu verstehen. Die erfolgreiche Vorgehensweise ist es, dass wir nicht versuchen, die Fehler für alle Komponenten gleichzeitig zu analysieren. Vielmehr werden wir uns zuerst auf das elektrische Feld $\vec{E}$ konzentrieren. Dieses hat durch die direkte Interaktion mit den beiden anderen Komponenten eine herausgehobene Stellung. Das magnetische Feld $\vec{B}$ und die Impulse $\vec{p}$ wechselwirken nur indirekt über $\vec{E}$ miteinander.

Wir werden die $\nabla \times \nabla \times$-Formulierung aus (2.22), (2.24) verwenden, welche als Hamilton-System mit der Hamilton-Funktion (2.28) zusätzliche Struktur erhält. Entsprechend werden wir durch gegenseitiges Einsetzen das magnetische Feld und die Impulse aus dem numerischen Verfahren eliminieren und eine Mehrschrittformulierung für die numerische Lösung des elektrischen Feldes finden.

Es stellt sich heraus, dass in [34, Abschnitt XIII.4.1] für eine solche Formulierung bereits eine Fehleranalyse existiert, die auf so genannten *modulierten Fourier-Entwicklungen* basiert. Die Grundidee ist dabei, ähnlich wie bei klassischen Fourier-Entwicklungen, oszillatorische Lösungen von Differentialgleichungen in ihre Frequenzen zu zerlegen. Dabei ist es aber erlaubt, dass die Koeffizienten in der Zeit auf eine glatte, nicht-oszillatorische Weise variieren. Solche Entwicklungen können sowohl für die analytische als auch die numerische Lösung hergeleitet werden. Abgesehen von hier verwendeten Abschätzungen für den Fehler numerischer Verfahren, kann man diese unter anderem auch verwenden, um das Langzeitverhalten von analytischen [12, 32, 21, 14] oder numerischen [30, 13] Lösungen von Hamilton-Systemen und Wellengleichungen zu verstehen. Vor Kurzem wurde in [20] ein Konvergenzresultat für die Wellengleichungen mit polynomialer Nichtlinearität gezeigt. Dort stammen die hohen Frequenzen aber anders als hier aus der Ortsdiskretisierung des Differentialoperators. Für die nichtlineare Schrödinger-Gleichung können Erhaltungsresultate für die analytischen und im Ort diskretisierten Lösungen [22] sowie für die vollständig diskretisierten Lösungen [23, 18] bewiesen werden. Einen guten Überblick über die vorhandenen Ergebnisse liefert [31].

Bei der Anwendung modulierter Fourier-Entwicklungen oder allgemeiner im Kontext hochoszillatorischer Differentialgleichungen wie etwa in [28] ist es üblich, so genannte Filterfunktionen zu benutzen. Diese werden dazu eingesetzt um Resonanzstellen, wie wir sie z.B. in Abbildung 3.2 finden, zu beseitigen. Einen expliziten Einsatz dieser Eigenschaft werden wir später unter anderem in (4.7), (4.81) oder zur Abschätzung von (4.104) sehen.

Wir beginnen damit, in unser Schema (3.6) nach dem Vorbild von [89] bzw. [59] Filterfunktionen

einzuführen. Wir legen uns aber zunächst nicht auf eine spezielle Wahl der Filter fest. Dann zitieren wir den Konvergenzsatz aus [34, XIII.4.1] und bringen in den folgenden Abschnitten unsere Gleichungen und das Verfahren Stück für Stück in die richtige Form, um diesen Satz anwenden zu können. Dazu müssen wir die Größen $\vec{B}$ und $\vec{p}$ sowohl aus der analytischen Gleichung als auch aus dem numerischen Verfahren eliminieren. Wir verwenden tatsächlich die einfachste Form der modulierten Fourier-Entwicklungen, bei denen nur eine hohe Frequenz und keine Nebenfrequenzen $k\omega$, $k \in \mathbb{Z} \setminus \{-1, 0, 1\}$, oder gar mehrere verschiedene, nahe beieinander liegende Frequenzen auftreten.

Ein Problem, das sich bei der Anwendung des Konvergenzresultates aus [34, XIII.4.1] ergibt, ist, dass die Startwerte, die dort gefordert sind, nicht zu unseren Gleichungen passen. Wir werden dies dadurch umgehen, dass wir zuerst die Fehlerabschätzung des numerischen Verfahrens gegen die analytische Lösung mit den für unseren Fall falschen Startwerten verwenden. Der Fehler in den Startwerten ist aber kontrollierbar klein. Mithilfe von Stabilitätseigenschaften der Lösung zeigen wir dann, dass der Fehler über die Zeit auch nicht größer wird, und erhalten so eine Abschätzung gegen die „richtige“ analytische Lösung, also jene mit den ursprünglichen Anfangswerten.

Auf Basis der Konvergenzaussage für das elektrische Feld können wir daraufhin eine Abschätzung zuerst für das magnetische Feld aufbauen und danach auch die Impulse abschätzen. Hier werden wir die Filterungseigenschaften direkt nutzen und dann unter Verwendung einiger trigonometrischer Identitäten das Aufsummieren von Fehlern und damit die Entwicklung von Resonanzen, wie wir sie im vorigen Abschnitt gesehen haben, unterbinden.

Während dieser Analyse werden wir immer wieder Bedingungen an die Filterfunktionen aufstellen. Dabei ist es jedoch zunächst unklar, ob es überhaupt solche gibt, die allen Bedingungen gerecht werden. Daher werden wir im Anschluss eine Wahl vorstellen, die tatsächlich die gewünschten Eigenschaften hat. Außerdem werden wir auch die ursprüngliche Wahl aus [89] diskutieren.

Zu guter Letzt werden wir unsere Ergebnisse in einem numerischen Test illustrieren.

4.1 Einführung von Filterfunktionen

Der erste Schritt ist die Einführung von Filterfunktionen in das numerische Schema. Außerdem ist es zweckdienlich, den Parameter c ganz zum Dichteparameter zu verschieben. Wenden wir uns zunächst den Filtern zu: Diese fügen wir nach dem Vorbild von [89] bzw. [59] in das numerische

Schema an allen Stellen ein, die sinnvoll erscheinen. Es ergibt sich das Verfahren

$$\vec{B}_{n+\frac{1}{2}} = \vec{B}_n - \tfrac{\tau}{2}\psi_B\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\mathbf{C}_E\,\phi_E\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\vec{E}_n, \tag{4.1a}$$

$$\vec{E}_n^{+} = \vec{E}_n + \tfrac{\tau}{2}\psi_E\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\mathbf{C}_B\,\phi_B\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\vec{B}_{n+\frac{1}{2}}, \tag{4.1b}$$

$$\begin{bmatrix}\vec{p}_{n+1}\\ \vec{E}_{n+1}^{-}\end{bmatrix} = \begin{bmatrix}\cos(\tau\mathbf{\Omega}) & \tau c\,\mathrm{sinc}(\tau\mathbf{\Omega})\\ -\frac{1}{c}\mathbf{\Omega}\sin(\tau\mathbf{\Omega}) & \cos(\tau\mathbf{\Omega})\end{bmatrix}\begin{bmatrix}\vec{p}_n\\ \vec{E}_n^{+}\end{bmatrix}, \tag{4.1c}$$

$$\vec{E}_{n+1} = \vec{E}_{n+1}^{-} + \tfrac{\tau}{2}\psi_E\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\mathbf{C}_B\,\phi_B\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\vec{B}_{n+\frac{1}{2}}, \tag{4.1d}$$

$$\vec{B}_{n+1} = \vec{B}_{n+\frac{1}{2}} - \tfrac{\tau}{2}\psi_B\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\mathbf{C}_E\,\phi_E\left(\tfrac{\tau}{2}\mathbf{\Omega}\right)\vec{E}_{n+1}. \tag{4.1e}$$

Die Filterfunktionen $\psi_i, \phi_i : \mathbb{C} \longrightarrow \mathbb{C}$ seien zunächst skalare, ganze Funktionen, die $\psi_i(\mathbb{R}_{\geq 0}) \subset \mathbb{R}$, $\phi_i(\mathbb{R}_{\geq 0}) \subset \mathbb{R}$, $i \in \{E, B\}$ erfüllen.

Wir gehen zu der Größe $\widetilde{\vec{p}} := \frac{1}{c}\vec{p}$ über, um c und $\mathbf{\Omega}$ besser zusammenfassen zu können. Um die Notation einfach zu halten, dividieren wir die erste Zeile von (4.1c) durch c, lassen danach die neu entstandene Tilde wieder weg und verbergen auch die Argumente der trigonometrischen und Filterfunkionen. Letztere sind immer $\tau\mathbf{\Omega}$ bzw. $\frac{\tau}{2}\mathbf{\Omega}$. Aus der ortskontinuierlichen Gleichung (2.10) im nichtrelativistischen Fall (2.12) ergibt sich die Differentialgleichung

$$\partial_t\vec{E} = \nabla\times\vec{B} - c^2\omega^2\vec{p}, \tag{4.2a}$$

$$\partial_t\vec{B} = -\nabla\times\vec{E}, \tag{4.2b}$$

$$\partial_t\vec{p} = \vec{E}. \tag{4.2c}$$

Entsprechend stellen wir auch die ortsdiskretisierte Version (3.1) um

$$\partial_t\vec{E} = \mathbf{C}_B\vec{B} - \mathbf{\Omega}^2\vec{p}, \tag{4.3a}$$

$$\partial_t\vec{B} = -\mathbf{C}_E\vec{E}, \tag{4.3b}$$

$$\partial_t\vec{p} = \vec{E}, \tag{4.3c}$$

wodurch die explizite Abhängigkeit der rechten Seite von c eliminiert wird. Obiges Verfahren hat dann die Form

$$\vec{B}_{n+\frac{1}{2}} = \vec{B}_n - \tfrac{\tau}{2}\psi_B\cdot\mathbf{C}_E\phi_E\cdot\vec{E}_n, \tag{4.4a}$$

$$\vec{E}_n^{+} = \vec{E}_n + \tfrac{\tau}{2}\psi_E\cdot\mathbf{C}_B\phi_B\cdot\vec{B}_{n+\frac{1}{2}}, \tag{4.4b}$$

$$\begin{bmatrix}\vec{p}_{n+1}\\ \vec{E}_{n+1}^{-}\end{bmatrix} = \begin{bmatrix}\cos & \tau\,\mathrm{sinc}\\ -\mathbf{\Omega}\sin & \cos\end{bmatrix}\begin{bmatrix}\vec{p}_n\\ \vec{E}_n^{+}\end{bmatrix}, \tag{4.4c}$$

$$\vec{E}_{n+1} = \vec{E}_{n+1}^{-} + \tfrac{\tau}{2}\psi_E\cdot\mathbf{C}_B\phi_B\cdot\vec{B}_{n+\frac{1}{2}}, \tag{4.4d}$$

$$\vec{B}_{n+1} = \vec{B}_{n+\frac{1}{2}} - \tfrac{\tau}{2}\psi_B\cdot\mathbf{C}_E\phi_E\cdot\vec{E}_{n+1}. \tag{4.4e}$$

Um die Multiplikationen nicht mit den Argumenten der Filter zu verwechseln, haben wir einen Malpunkt eingefügt – dieser bedeutet *kein* Skalarprodukt. Wir erhalten wieder ein Splitting-Verfahren der Form

$$\Psi_\tau = \xi_{\tau/2} \circ \zeta_{\tau/2} \circ \eta_\tau \circ \zeta_{\tau/2} \circ \xi_{\tau/2},$$

wobei ξ der numerische Fluss in (4.4a) bzw. (4.4e), ζ derjenige aus (4.4b) bzw. (4.4d) und η der des harmonischen Oszillators aus (4.4c) sei. Dabei sind im Gegensatz zu Abschnitt 3.2 ξ und ζ diesmal keine exakten Flüsse.

Um insgesamt ein konsistentes Verfahren zu erhalten, müssen auch die numerischen Lösungen der Teilgleichungen (4.3), insbesondere die beiden gestörten Flüsse ξ und ζ, konsistent sein. Dazu benötigen wir die Eigenschaften

$$\psi_i(z) \longrightarrow 1, \quad \phi_i(z) \longrightarrow 1 \quad \text{für} \quad z \longrightarrow 0, \quad i \in \{E, B\}\,. \tag{4.5}$$

Wir wollen nach wie vor ein symmetrisches Verfahren erhalten. Ein Splitting-Verfahren obiger Form ist sofort symmetrisch, sobald dies für die Teilflüsse ξ_τ, ζ_τ und η_τ gilt. Als exakter Fluss von (3.4) für $i = 3$ bzw. (4.3a) ist der mittlere Term η_τ automatisch symmetrisch. Die beiden anderen numerischen Flüsse sind durch die Einführung der Filter nun aber keine exakten Lösungen von Teilgleichungen mehr, wodurch diese Symmetrie nicht mehr direkt folgt. Wählen wir die Filter als gerade Funktionen, d.h.

$$\psi_i(z) = \psi_i(-z), \qquad \phi_i(z) = \phi_i(-z) \qquad \text{für alle} \qquad z \in \mathbb{C}, \quad i \in \{E, B\}\,, \tag{4.6}$$

so erhalten wir $\xi_\tau^{-1} = \xi_{-\tau}$ bzw. $\zeta_\tau^{-1} = \zeta_{-\tau}$ und damit die Symmetrie dieser Flüsse und damit auch die des Verfahrens (4.4).

Eine erste Wahl für eine der Filterfunktionen erhalten wir wie in [89] aus der Forderung, die Divergenz des magnetischen Feldes $\vec{B}$ zu erhalten (2.5). Diese ist eine physikalisch wichtige Eigenschaft und soll auch von der Zeitdiskretisierung respektiert werden. Aus dem Verfahren ergibt sich für die ortskontinuierlichen Gleichungen, also wenn in (4.4a) und (4.4e) der diskrete Rotationsoperator $\mathbf{C}_E$ durch den kontinuierlichen $\nabla\times$ ersetzt wird und wir $\vec{E}$ und $\vec{B}$ wieder als Vektorfelder auffassen:

$$\begin{aligned} \nabla \cdot \vec{B}_n \overset{!}{=} \nabla \cdot \vec{B}_{n+1} \quad &= \quad \nabla \cdot \vec{B}_n + \frac{\tau}{2} \nabla \cdot \left[\psi_B \cdot \left(-\nabla \times \left(\phi_E \cdot (\vec{E}_n + \vec{E}_{n+1}) \right) \right) \right] \\ &\overset{(A.31)}{=} \nabla \cdot \vec{B}_n + \frac{\tau}{2} \left\langle \nabla \psi_B, -\nabla \times \left(\phi_E \cdot (\vec{E}_n + \vec{E}_{n+1}) \right) \right\rangle . \end{aligned} \tag{4.7}$$

Das Argument von ψ_B ist im kontinuierlichen Fall $\tau c \omega$. Da ω und die elektrischen Felder beide ortsabhängig sind und der zweite Summand für beliebige ω und $\vec{E}$ verschwinden soll, muss also

$\psi_B \equiv$ const gelten. Mit (4.5) ergibt sich insgesamt

$$\psi_B \equiv 1. \tag{4.8}$$

Auf dieselbe Art und Weise zu verhindern, dass das magnetische Feld die Divergenz des elektrischen Feldes beeinflusst, vgl. (2.4), lässt sich nicht mit einer so einfachen Bedingung gewährleisten. Wir werden später sehen, dass eine entsprechende Wahl von ψ_E unattraktiv ist.

Um unser weiteres Vorgehen besser motivieren zu können, geben wir als Nächstes das Resultat an, mit dem wir arbeiten wollen, um den Fehler des Verfahrens zu kontollieren.

4.2 Zuhilfenahme modulierter Fourier-Entwicklungen

Unser Ziel ist es, für die Abschätzung des elektrischen Feldes ein Resultat zu benutzen, das mithilfe modulierter Fourier-Entwicklungen bewiesen wird. Zunächst zitieren wir dieses aus [34] und diskutieren dann Stück für Stück die Voraussetzungen:

Theorem 4.1 (Hairer, Lubich, Wanner, [34, XIII.4.1, Theorem 4.1])**.** *Wir betrachten die numerische Lösung des Systems*

$$\partial_{tt} x(t) = -\mathbf{\Omega}^2 x(t) + g(x(t)), \qquad x(t_0) = x_0, \quad \partial_t x(t_0) = \dot{x}_0, \tag{4.9}$$

wobei

$$g(x) = -\nabla U(x) \qquad \textit{und} \qquad \mathbf{\Omega} = \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{\omega}\,\mathbf{Id} \end{bmatrix}, \qquad \widetilde{\omega} \gg 1 \tag{4.10}$$

und g eine glatte Gradienten-Nichtlinearität sei. Die Anfangsbedingungen mögen

$$\frac{1}{2}\big\langle \partial_t x(0), \partial_t x(0) \big\rangle + \frac{1}{2}\big\langle \mathbf{\Omega} x(0), \mathbf{\Omega} x(0) \big\rangle \leq H_0, \tag{4.11}$$

für H_0 unabhängig von $\widetilde{\omega}$, erfüllen. Diese Gleichung möge mit der Methode

$$x_{n+1} - 2\cos(\tau\mathbf{\Omega}) x_n + x_{n-1} = \tau^2 \psi(\tau\mathbf{\Omega}) g(\phi(\tau\mathbf{\Omega}) x_n) \tag{4.12}$$

mit einer Schrittweite $\tau \leq \tau_0$ mit einem hinreichend kleinen τ_0, das von $\widetilde{\omega}$ unabhängig ist, gelöst werden, für das $\tau\widetilde{\omega} \geq c_0 > 0$ gilt. Die Filterfunktionen ψ und ϕ seien dabei gerade, reellwertige Funktionen mit $\psi(0) = \phi(0) = 1$. Die erste Iterierte x_1 werde mittels

$$x_1 = \cos(\tau\mathbf{\Omega}) x_0 + \tau\,\mathrm{sinc}(\tau\mathbf{\Omega})\dot{x}_0 + \frac{1}{2}\tau^2 \psi(\tau\mathbf{\Omega}) g(\phi(\tau\mathbf{\Omega}) x_0) \tag{4.13}$$

bestimmt. Wir benötigen die folgenden Bedingungen an die Filterfunktionen:

$$|\psi(z)| \leq C_1 \operatorname{sinc}^2(\tfrac{1}{2}z), \tag{4.14a}$$
$$|\phi(z)| \leq C_2 |\operatorname{sinc}(\tfrac{1}{2}z)|, \tag{4.14b}$$
$$|\psi(z)\phi(z)| \leq C_3 |\operatorname{sinc}(z)|. \tag{4.14c}$$

für $z \geq 0$. Dann erhalten wir eine Abschätzung der Ordnung zwei an den Fehler des Verfahrens:

$$\|x_n - x(t_n)\| \leq C\tau^2 \qquad \textit{für} \qquad t_n := n\tau \leq T. \tag{4.15}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von T, H_0, Schranken an die Ableitungen von g sowie den Konstanten C_1, C_2 und C_3.

Falls anstatt von (4.14a) die abgeschwächte Bedingung

$$|\psi(z)| \leq C_0 |\operatorname{sinc}(\tfrac{1}{2}z)| \tag{4.16}$$

erfüllt ist, reduziert sich die Konvergenzordnung auf eins,

$$\|x_n - x(t_n)\| \leq C\tau \qquad \textit{für} \qquad t_n := n\tau \leq T. \tag{4.17}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen.

Wir werden nun Schritt für Schritt zeigen, dass bzw. unter welchen Voraussetzungen die Gleichungen (2.13) und das Verfahren (4.4) die Voraussetzungen von Theorem 4.1 erfüllen.

4.3 $\nabla \times \nabla\times$-Formulierung

Der erste Schritt ist es, die Form (4.9) nachzuweisen. Dies haben wir rein formal für die ursprüngliche Form der ortskontinuierlichen Gleichungen in der $\nabla \times \nabla\times$-Formulierung aus (2.22)-(2.24) bereits erreicht. Der Übergang $\vec{p}$ zu $\frac{1}{c}\vec{p}$ ändert dabei nichts. Entsprechend erhalten wir aus dem ortsdiskretisierten System (4.3)

$$\begin{cases} \partial_{tt}\vec{E}(t) = -\mathbf{\Omega}^2\vec{E}(t) + \mathbf{G}\vec{E}(t), & \text{(4.18a)} \\ \vec{E}(t_0) = \vec{E}_0, \quad \partial_t\vec{E}(t_0) = \mathbf{C}_B\vec{B}(t_0) - \mathbf{\Omega}^2\vec{p}(t_0) := \dot{\vec{E}}_0, & \text{(4.18b)} \end{cases}$$

wobei $\mathbf{G} := -\mathbf{C}_B\mathbf{C}_E$ und erhalten mit $g(\vec{E}) := \mathbf{G}\vec{E}$ die Form (4.9).

Die nötigen Voraussetzungen an die Ortsdiskretisierung fassen wir nun zusammen.

Voraussetzung 4.2. *Die Ortsdiskretisierung und das Dichteprofil ρ seien so gewählt, dass die Matrix $\mathbf{\Omega}$,*

die aus der Diskretisierung von $c\sqrt{\rho}$ hervorgeht, die Gestalt

$$\mathbf{\Omega} = \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{\omega}\,\mathbf{Id} \end{bmatrix}, \qquad \widetilde{\omega} := c\sqrt{\rho_F} \qquad \rho_F \gg 1, \quad c \in \mathbb{R}_{>0} \tag{4.19}$$

habe. Die beiden Diskretisierungen des Rotationsoperators, $\mathbf{C}_B$ und $\mathbf{C}_E$, seien zueinander transponiert, also

$$\mathbf{C}_E = \mathbf{C}_B^T. \tag{4.20}$$

Die Matrix

$$\mathbf{G} = -\mathbf{C}_B\mathbf{C}_E \preccurlyeq 0 \tag{4.21}$$

sei negativ semidefinit. Die Normen der Matrizen $\mathbf{C}_B$ und $\mathbf{C}_E$ seien durch eine gemeinsame Konstante C_c, die unabhängig von $\widetilde{\omega}$ ist, etwa

$$\|\mathbf{C}_B\| \leq C_c, \qquad \|\mathbf{C}_E\| \leq C_c, \tag{4.22}$$

beschränkt.

Etwa bei periodischen Randbedingungen erhalten wir $\langle -\nabla \times \nabla \times \vec{E}, \vec{E}\rangle = \langle \vec{E}, -\nabla \times \nabla \times \vec{E}\rangle = -\|\nabla \times \vec{E}\|^2$, also ist der analytische $\nabla \times \nabla\times$-Operator symmetrisch negativ semidefinit. Außerdem ist er linear. Es ist daher natürlich, dies auch von der ortsdiskretisierten Version $\mathbf{G}$ zu fordern. Mit den Schranken (4.22) ist auch

$$\|\mathbf{G}\| = \| - \mathbf{C}_B\mathbf{C}_E\| \leq C_c^2 =: C_g \tag{4.23}$$

unabhängig von $\widetilde{\omega}$ beschränkt. Es sei darauf hingewiesen, das die Voraussetzung der Existenz der Schranken C_c bzw. C_g eine wesentliche Einschränkung in dem Sinne liefert, dass wir keine von der Ortsschrittweite h unabhängigen Fehlerabschätzungen erhalten können, solange diese Konstanten in die Fehlerabschätzung eingehen. Die Einschränkung an die Struktur von $\mathbf{\Omega}$ werden wir zu einem späteren Zeitpunkt diskutieren.

Die Frequenz aus [34] und den verwandten Artikeln, die auch in Theorem 4.1 auftritt, ist hier durch $\widetilde{\omega} = c\sqrt{\rho_F}$ gegeben. Die trigonometrischen und Filterfunktionen werden auf $\tau\mathbf{\Omega}$, also den gleichen Argumenten wie in diesem Kontext ausgewertet.

Da $g(\vec{E}) = \mathbf{G}\vec{E}$ linear und $\mathbf{G} = -\mathbf{C}_B\mathbf{C}_E = -\mathbf{C}_B\mathbf{C}_B^T$ mit (4.20) symmetrisch ist, können wir

$$U(\vec{E}) := -\tfrac{1}{2}\vec{E}^T\mathbf{G}\vec{E} \quad \text{definieren, sodass} \quad g(\vec{E}) = -\nabla U(\vec{E}), \tag{4.24}$$

womit wir die Potentialstruktur nachgewiesen haben. Nach persönlicher Korrespondenz mit einem der Autoren [61] wird bestätigt, dass dieses g die Glattheitsannahmen von Theorem 4.1 erfüllt, also insbesondere nicht selbst beschränkt sein muss.

Die entsprechende ortsdiskretisierte Version der über die Zeit erhaltenen Hamilton-Funktion er-

gibt sich zu

$$\mathcal{H}(\vec{E},\vec{F}) = \tfrac{1}{2}\|\vec{F}\|^2 + \tfrac{1}{2}\|\mathbf{\Omega}\vec{E}\|^2 - \tfrac{1}{2}\left\langle \vec{E}, \mathbf{G}\vec{E}\right\rangle = \tfrac{1}{2}\|\vec{F}\|^2 + \tfrac{1}{2}\|\mathbf{\Omega}\vec{E}\|^2 + U(\vec{E}). \tag{4.25}$$

Diese bezeichnen wie wieder mit $\mathcal{H}$, da wir uns von hier ab nur noch auf diese ortsdiskretisierte Version beziehen.

Mithilfe der Variation-der-Konstanten-Formel erhalten wir die Darstellungen

$$\begin{aligned}\vec{E}(t) = \cos((t-t')\mathbf{\Omega})\vec{E}(t') + (t-t')\,\mathrm{sinc}((t-t')\mathbf{\Omega})\partial_t\vec{E}(t') \\ + (t-t')\int_0^1 (t-t')(1-\xi)\,\mathrm{sinc}((t-t')(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t'(1-\xi)+t\xi)\,\mathrm{d}\xi\end{aligned} \tag{4.26a}$$

und

$$\begin{aligned}\partial_t\vec{E}(t) = -\mathbf{\Omega}\sin((t-t')\mathbf{\Omega})\vec{E}(t') + \cos((t-t')\mathbf{\Omega})\partial_t\vec{E}(t') \\ + (t-t')\int_0^1 \cos((t-t')(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t'(1-\xi)+t\xi)\,\mathrm{d}\xi\end{aligned} \tag{4.26b}$$

der Lösungen von (4.18).

4.4 Mehrschrittformulierung des numerischen Verfahrens

Wir wollen nun auf (4.12) hinarbeiten. Auf dem Weg dorthin erhalten wir eine weitere Bedingung an die Filterfunktionen. Wir beginnen damit, die Verfahrensschritte aus (4.4) ineinander einzusetzen, dabei (4.8) auszunutzen und erhalten

$$\begin{aligned}\vec{p}_{n+1} &= \cos\cdot\vec{p}_n + \tau\,\mathrm{sinc}\cdot\vec{E}_n + \tau^2\tfrac{1}{2}\,\mathrm{sinc}\cdot\psi_E\cdot\mathbf{C}_B\phi_B\cdot\vec{B}_n \\ &\quad - \tau^3\tfrac{1}{4}\,\mathrm{sinc}\cdot\psi_E\cdot\mathbf{C}_B\phi_B\cdot\mathbf{C}_E\phi_E\cdot\vec{E}_n, \qquad (4.27)\\ \vec{E}_{n+1} &= -\mathbf{\Omega}\sin\cdot\vec{p}_n + \cos\cdot\vec{E}_n + \tau\tfrac{1}{2}(\cos+1)\cdot\psi_E\cdot\mathbf{C}_B\phi_B\cdot\vec{B}_n \\ &\quad - \tau^2\tfrac{1}{4}(\cos+1)\cdot\psi_E\cdot\mathbf{C}_B\phi_B\cdot\mathbf{C}_E\phi_E\cdot\vec{E}_n \qquad (4.28)\end{aligned}$$

und

$$\vec{B}_{n+1} = \vec{B}_n - \tau\tfrac{1}{2}\mathbf{C}_E\phi_E\cdot\left(\vec{E}_n + \vec{E}_{n+1}\right). \tag{4.29}$$

Wir erkennen nun in (4.27) und (4.28), dass wir tatsächlich ein Verfahren erhalten, welches den zu $\mathbf{\Omega}$ gehörigen Anteil von (4.18) exakt behandelt und den zur Nichtlinearität g aus (4.9) bzw. deren

Darstellung $\mathbf{G}$ gehörigen filtert. Allerdings haben wir einen zusätzlichen Filter ϕ_B zwischen den beiden Rotationsoperatoren, den Faktoren von $\mathbf{G}$, eingeführt, was wir nun beheben, indem wir

$$\phi_B \equiv 1 \tag{4.30}$$

fordern. Damit erhalten wir für $\vec{p}$ und $\vec{E}$ die Einschrittformulierung

$$\vec{p}_{n+1} = \cos\cdot\vec{p}_n + \tau\,\mathrm{sinc}\cdot\vec{E}_n + \tau^2\tfrac{1}{2}\,\mathrm{sinc}\cdot\psi_E\cdot\mathbf{C}_B\vec{B}_n + \tau^3\tfrac{1}{4}\,\mathrm{sinc}\cdot\psi_E\cdot\mathbf{G}\phi_E\cdot\vec{E}_n, \tag{4.31}$$

$$\vec{E}_{n+1} = -\mathbf{\Omega}\sin\cdot\vec{p}_n + \cos\cdot\vec{E}_n + \tau\tfrac{1}{2}(\cos+1)\cdot\psi_E\cdot\mathbf{C}_B\vec{B}_n + \tau^2\tfrac{1}{4}(\cos+1)\cdot\psi_E\cdot\mathbf{G}\phi_E\cdot\vec{E}_n. \tag{4.32}$$

Unser Splitting-Ansatz führt also auf natürliche Weise bereits den Filter sinc bei den Impulsen und $\frac{1}{2}(1+\cos)$ für das elektrische Feld ein, selbst wenn wir $\phi_E \equiv \psi_E \equiv 1$ wählen würden. Um die Mehrschrittformulierung (4.12) zu erhalten, müssen wir nun wie dort auf der linken Seite einen Vorwärts- und einen Rückwärtsschritt des Verfahrens mit Schrittweiten τ bzw. $-\tau$ addieren und $2\cos\vec{E}_n$ subtrahieren. Mit den Symmetrieeigenschaften der Filter und den trigonometrischen Funktionen erhalten wir daraus die korrekte Form

$$\vec{E}_{n-1} - 2\cos\cdot\vec{E}_n + \vec{E}_{n+1} = \tau^2\tfrac{1}{2}(\cos+1)\cdot\psi_E\cdot\mathbf{G}\phi_E\cdot\vec{E}_n. \tag{4.33}$$

Wir erhalten also (4.12) mit

$$\psi(z) = \tfrac{1}{2}(\cos(z)+1)\psi_E(\tfrac{1}{2}z), \qquad \phi(z) = \phi_E(\tfrac{1}{2}z) \qquad \text{für alle} \qquad z\in\mathbb{C}. \tag{4.34}$$

Für Theorem 4.1 benötigen wir, dass ψ und ϕ gerade, reellwertige Funktionen mit $\psi(0) = \phi(0) = 1$ sind. Mit den Bedingungen an die Filter ψ_E, ϕ_E am Anfang von Abschnitt 4.1 sind diese erfüllt.

Die Bedingungen (4.14) lesen sich dann wie folgt:

$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)| \le C_1\,\mathrm{sinc}^2(\tfrac{1}{2}z), \tag{4.35a}$$

$$|\phi_E(\tfrac{1}{2}z)| \le C_2|\,\mathrm{sinc}(\tfrac{1}{2}z)|, \tag{4.35b}$$

$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)\phi_E(\tfrac{1}{2}z)| \le C_3|\,\mathrm{sinc}(z)| \tag{4.35c}$$

für $z \ge 0$, die abgeschwächte Bedingung (4.16) wird zu

$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)| \le C_0|\,\mathrm{sinc}(\tfrac{1}{2}z)| \tag{4.36}$$

für $z \ge 0$, mit Konstanten C_0, C_1, C_2 und C_3, die unabhängig von $\widetilde{\omega} = c\sqrt{\rho_F}$ sind.

Betrachten wir nun ortskontinuierlich das Update der Divergenz des elektrischen Feldes, so ergibt sich mit (4.32) nach Ersetzen von $\mathbf{C}_B$ und $\mathbf{C}_E$ durch $\nabla\times$ wie in (4.7)

$$\nabla\cdot\vec{E}_{n+1} = \nabla\cdot(-c\omega\sin\cdot\vec{p}_n + \cos\cdot\vec{E}_n) + \tfrac{\tau}{2}\left\langle\nabla((\cos+1)\cdot\psi_E), \nabla\times(\vec{B}_n - \tfrac{\tau}{2}\nabla\times\vec{E}_n)\right\rangle. \tag{4.37}$$

Wollen wir passend zu (2.4) verhindern, dass das magnetische Feld Einfluss nimmt, müssen wir $(\cos(z)+1)\psi_E(z) \equiv \text{const}$ oder umgekehrt $\psi_E(z) = \text{const}/(\cos(z)+1)$ wählen. Das hätte aber zwei ungünstige Folgen: Zum einem erhalten wir in ψ_E Singularitäten in $(1+2k)\pi$, $k \in \mathbb{Z}$. Im Filter ψ wären diese zwar hebbar, doch müssen wir für die Splitting-Formulierung (4.1) ψ_E direkt auswerten. Zum anderen würden wir mit dieser Wahl den auf natürliche Weise aus dem Splitting-Ansatz entstandenen Filter wieder aufheben und den exponentiellen Integrator zu einem klassischen Verfahren zurückentwickeln.

4.5 Anfangswerte und Energie

Die nächste Voraussetzung, die wir für die Anwendbarkeit von Theorem 4.1 benötigen, ist die Forderung an die Startwerte (4.13). Wir subtrahieren diesen gewünschten Startwert von (4.28) und erhalten mit (4.32) und der Filterfunktion ψ aus (4.34) den Defekt

$$\begin{aligned}
&\vec{E}_1 - \cos\cdot\vec{E}_0 - \tau\,\mathrm{sinc}\cdot\dot{\vec{E}}_0 - \tau^2\tfrac{1}{2}\psi\cdot\mathbf{G}\phi\cdot\vec{E}_0\\
=&-\mathbf{\Omega}\sin\cdot\vec{p}_0 + \cos\cdot\vec{E}_0 + \tau\tfrac{1}{2}(\cos+1)\cdot\psi_E\cdot\mathbf{C}_B\vec{B}_0 + \tau^2\tfrac{1}{4}(\cos+1)\cdot\psi_E\cdot\mathbf{G}\phi_E\cdot\vec{E}_0\\
&-\cos\cdot\vec{E}_0 - \tau\,\mathrm{sinc}\cdot(\mathbf{C}_B\vec{B}_0 - \mathbf{\Omega}^2\vec{p}_0) - \tau^2\tfrac{1}{4}(\cos+1)\cdot\psi_E\cdot\mathbf{G}\phi_E\cdot\vec{E}_0\\
=&\tau(\tfrac{1}{2}(\cos+1)\cdot\psi_E - \mathrm{sinc})\cdot\mathbf{C}_B\vec{B}_0,
\end{aligned} \tag{4.38}$$

wobei wir für die zweite Gleichung die Identität $z\,\mathrm{sinc}(z) = \sin(z)$ verwenden. Auf diese Weise lässt sich (4.13) also nicht erfüllen. Daher gehen wir zu den modifizierten Startwerten

$$\vec{E}_0{}' := \vec{E}_0 \qquad \dot{\vec{E}}_0{}' := \chi(\tau\mathbf{\Omega})\mathbf{C}_B\vec{B}_0 - \mathbf{\Omega}^2\vec{p}_0 \tag{4.39}$$

über. Für diese erhalten wir

$$\vec{E}_1{}' - \cos\cdot\vec{E}_0{}' - \tau\,\mathrm{sinc}\cdot\dot{\vec{E}}_0{}' - \tau^2\tfrac{1}{2}\psi\cdot\mathbf{G}\phi\cdot\vec{E}_0{}' = \tau(\tfrac{1}{2}(\cos+1)\cdot\psi_E - \mathrm{sinc}\cdot\chi)\cdot\mathbf{C}_B\vec{B}_0 = 0, \tag{4.40}$$

die zweite Gleichung für die Wahl

$$\chi(z) = \tfrac{1}{2}\frac{\cos(z)+1}{\mathrm{sinc}(z)}\psi_E(\tfrac{1}{2}z) = \frac{\psi(z)}{\mathrm{sinc}(z)}. \tag{4.41}$$

Mit der Forderung

$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)| \le C_4|\,\mathrm{sinc}(z)| \tag{4.42}$$

für $z \ge 0$ mit einer Konstante $C_4 > 0$, die unabhängig von $\widetilde{\omega}$ ist, erreichen wir, dass χ, insbesondere auch an den Nullstellen seines Nenners, beschränkt ist. Dies ist eine Verschärfung von (4.35c), denn wir fordern die Abschätzung ohne Zuhilfenahme des Filters ϕ_E und nach (4.35b)

muss ψ_E beschränkt sein und damit folgt dann (4.35c) aus (4.42). In Abschnitt 4.7 werden wir weitere Bedingungen diskutieren, damit $\vec{E}_0$ und $\vec{E}_0{}'$ nahe genug beieinander liegen.

Die Bedingungen an die Filterfunktionen (4.35) sollen vorerst als Voraussetzungen bestehen bleiben. Konkrete Wahlen der Filter werden wir in Abschnitt 4.12 diskutieren.

Zum Schluss formulieren wir Forderungen an die Startwerte, die die initiale-Energie-Bedingung (4.11) gewährleisten. Unter der Bedingung, dass

$$\|\mathbf{\Omega}\vec{E}_0\|^2 \leq \tfrac{2}{3}H_0, \qquad \|\mathbf{C}_B\vec{B}_0\|^2 \leq \tfrac{1}{3}(\tfrac{2}{C_4})^2 H_0, \qquad \|\mathbf{\Omega}^2\vec{p}_0\|^2 \leq \tfrac{1}{3}H_0 \tag{4.43}$$

mit einem H_0, das unabhängig von $\widetilde{\omega}$ sei, erhalten wir sowohl

$$\tfrac{1}{2}\langle \mathbf{\Omega}\vec{E}_0{}', \mathbf{\Omega}\vec{E}_0{}'\rangle + \tfrac{1}{2}\langle \dot{\vec{E}}_0{}', \dot{\vec{E}}_0{}'\rangle \leq H_0, \tag{4.44}$$

als auch

$$\tfrac{1}{2}\langle \mathbf{\Omega}\vec{E}_0, \mathbf{\Omega}\vec{E}_0\rangle + \tfrac{1}{2}\langle \dot{\vec{E}}_0, \dot{\vec{E}}_0\rangle \leq H_0, \tag{4.45}$$

wobei o.B.d.A. $2/C_4 \leq 1$. Für die später gewählten Filterfunktionen erhalten wir mit $C_4 = 2$ sogar Gleichheit. Damit ist die initiale-Energie-Bedingung (4.11) also für beide Paare von Startwerten erfüllt.

Für die weitere Fehleranalyse benötigen wir eine Abschätzung der Hamilton-Funktion $\mathcal{H}$, einschließlich des Potentials, und eine Abschätzung des elektrischen Feldes $\vec{E}$ zum Startzeitpunkt – und damit zu allen Zeiten – unabhängig von $\widetilde{\omega}$. Dazu fordern wir

$$-\langle \mathbf{G}\vec{E}_0, \vec{E}_0\rangle = \|\mathbf{C}_E\vec{E}_0\|^2 \leq 2H_0, \qquad \|\vec{E}_0\|^2 \leq H_0 \tag{4.46}$$

mit H_0 aus (4.43), unabhängig von $\widetilde{\omega}$. Mit der Hamilton-Funktion (4.25) erhalten wir mit der Schranke an den Startwert des elektrischen Feldes zusätzlich noch eine Aussage darüber, wie sich die Lösung überall dort verhält, wo die Diagonale von $\mathbf{\Omega}$ nicht besetzt ist. Die mittlere Annahme aus (4.43) und die erste aus (4.46) sind als Glattheitsannahmen an die Felder zu verstehen, da $\mathbf{C}_B$ und $\mathbf{C}_E$ die Diskretisierungen des Rotationsoperators darstellen. Schließlich fordern wir für den Startwert des magnetischen Feldes

$$\|\vec{B}_0\|^2 \leq H_0 \tag{4.47}$$

und erhalten insgesamt die folgende Erhaltungsaussage.

Lemma 4.3 (Stabilität von Lösungen)**.** *Betrachte die analytische Lösung der ortsdiskretisierten Maxwell-Gleichungen in der $-\nabla\times\nabla\times$-Formulierung* (4.18)*. Das magnetische Feld sei via* (4.3b) *gegeben. Es gebe eine Konstante H_0, sodass die Startwerte* (4.43) *und* (4.46) *erfüllen.*

Dann gilt für die Hamilton-Funktion aus (4.25)

$$\mathcal{H}(\vec{E}(t), \partial_t \vec{E}(t)) \leq 2H_0 \qquad \textit{und} \qquad \|\vec{E}(t)\| \leq (1 + 2(T - t_0))\sqrt{H_0} \qquad \textit{für} \quad t_0 \leq t \leq T. \tag{4.48}$$

Insbesondere ergeben sich

$$\begin{aligned} \|\partial_t \vec{E}(t)\| &\leq 2\sqrt{H_0}, \\ \|\mathbf{\Omega}\vec{E}(t)\| &\leq 2\sqrt{H_0} \qquad \textit{für} \quad t_0 \leq t \leq T. \\ \textit{und } \|\mathbf{C}_E \vec{E}(t)\| &\leq 2\sqrt{H_0} \end{aligned} \tag{4.49}$$

Falls zusätzlich (4.47) *erfüllt ist, erhalten wir weiterhin*

$$\|\vec{B}(t)\| \leq (1 + 2(T - t_0))\sqrt{H_0} \qquad \textit{für} \quad t_0 \leq t \leq T. \tag{4.50}$$

Beweis. Mit (4.43) erhalten wir nach (4.45), dass $\frac{1}{2}\langle \mathbf{\Omega}\vec{E}_0, \mathbf{\Omega}\vec{E}_0 \rangle + \frac{1}{2}\langle \dot{\vec{E}}_0, \dot{\vec{E}}_0 \rangle \leq H_0$. Der dritte Summand der Hamilton-Funktion (4.25) ist nach der 1. Bedingung aus (4.46) ebenfalls durch H_0 beschränkt. Damit folgt die Abschätzung an die Hamilton-Funktion zum Startzeitpunkt und damit auch für alle Zeiten. Die Ungleichungen (4.49) sind eine direkte Folgerung daraus, denn die linken Seiten sind – mit einem Faktor $\frac{1}{2}$ und einem Quadrat versehen – ein Teil der Hamiltonfunktion.

Die Abschätzung für $\|\vec{E}(t)\|$ folgt mit dem Hauptsatz der Differential- und Integralrechnung via

$$\begin{aligned} \|\vec{E}(t)\| = \left\| \vec{E}_0 + \int_{t_0}^{t} \partial_t \vec{E}(\xi)\, \mathrm{d}\xi \right\| &\overset{(4.46),(4.48)}{\leq} \sqrt{H_0} + (t - t_0) \max_{\xi \in [t_0, t]} \sqrt{2\mathcal{H}(\vec{E}(\xi), \partial_t \vec{E}(\xi))} \\ &\leq (1 + 2(T - t_0))\sqrt{H_0}. \end{aligned}$$

Analog erhalten wir die Abschätzung für das magnetische Feld, indem wir den Hauptsatz auf (4.3b) anwenden und die Abschätzung für den Startwert des magnetischen Feldes (4.47) und die gerade hergeleitete Abschätzung für das elektrische Feld, die erste aus (4.49), benutzen. □

Bemerkung 4.4. Das oben angegebene Lemma lässt zu, dass die Lösung über die Zeit wächst. Wir zeigen, dass diese Abschätzung auch scharf ist. Der Grund für dieses Wachstum ist der nichttriviale Kern der Frequenzmatrix $\mathbf{\Omega}$. Dies führt zu dem vorhergesagten Wachstum, wie das folgende einfache Beispiel illustriert. Wir betrachten dazu etwa die Gleichung

$$u'(t) = \mathbf{A}u(t), \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad u(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

welche die Umformulierung von $v''(t) = 0$, $v(0) = 0$, $v'(0) = 1$ in ein System erster Ordnung

darstellt. Die Matrix $\mathbf{A}$ ist nilpotent mit $\mathbf{A}^2 = 0$ und wir erhalten

$$\exp(t\mathbf{A}) = \mathbf{Id} + t\mathbf{A} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \quad \Rightarrow \quad u(t) = \begin{bmatrix} t \\ 1 \end{bmatrix}.$$

Die Lösung wächst also linear.

Falls die Randbedingungen so gewählt sind, dass zusätzlich zur Hamilton-Funktion (4.25) auch die klassische Energie (2.18) erhalten bleibt und sich dies auf die Ortsdiskretisierung überträgt, wird dieses Wachstum sowohl für das elektrische als auch das magnetische Feld verhindert. Nach der Umskalierung $\widetilde{\vec{p}} = \frac{1}{c}\vec{p}$ ergibt sich die diskrete Form der Energie (3.14) zu

$$E_h(t) := \frac{1}{2}(\|\vec{E}(t)\|^2 + \|\vec{B}(t)\|^2 + \|\mathbf{\Omega}\vec{p}(t)\|^2). \tag{4.51}$$

Falls diese ebenfalls erhalten ist, also $E_h(t) \equiv E_h(t)$ für alle t, sind sowohl $\|\vec{E}(t)\|^2$ als auch $\|\vec{B}(t)\|^2$ durch $E_h(t_0)$ beschränkt. Wir wollen uns allerdings nicht auf diesen Fall einschränken.

4.6 Abschätzung für das elektrische Feld mit modifizierten Anfangsbedingungen

Die Überlegungen aus den Abschnitten 4.3 bis 4.5 zeigen, unter welchen Bedingungen wir Theorem 4.1 auf die Gleichung mit den modifizierten Startwerten anwenden können. Das folgende Lemma fasst die Ergebnisse dieser Abschnitte zusammen:

Lemma 4.5. *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen in der $\nabla \times \nabla\times$-Formulierung (4.18a) mit dem Verfahren (4.33). Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ gelte $\tau \leq \tau_0$, τ_0 unabhängig von $\widetilde{\omega}$, und $\tau\widetilde{\omega} \geq c_0 > 0$. Für gegebene $\vec{E}_0$, $\vec{B}_0$ und $\vec{p}_0$, die (4.43) erfüllen, seien die Anfangswerte durch (4.39) mit (4.41) gegeben. Der erste Zeitschritt werde als*

$$\vec{E}_1{}' = \cos\cdot\vec{E}_0{}' + \tau\,\mathrm{sinc}\cdot\dot{\vec{E}}_0{}' + \tfrac{1}{4}\tau^2(\cos+1)\cdot\psi_E\cdot g\left(\phi_E\cdot\vec{E}_0{}'\right) \tag{4.52}$$

bestimmt. Die numerische Lösung werde mit $\vec{E}_n{}'$ bezeichnet. Die Filterfunktionen $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ seien gerade, reellwertige Funktionen mit $\psi_E(0) = \psi_E(0) = 1$ und mögen den Abschätzungen (4.35) und (4.42) mit Konstanten $C_1, \ldots, C_4$ unabhängig von $\widetilde{\omega}$ genügen.

Die analytische Lösung werde mit $\vec{E}'(t)$ bezeichnet. Dann erhalten wir die Abschätzung

$$\|\vec{E}_n{}' - \vec{E}'(t_n)\| \leq C\tau^2 \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.53}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge

des Zeitintervalls $(T - t_0)$ *sowie den Konstanten* H_0, C_g *aus* (4.23) *und* $C_1, ..., C_3$.

Falls anstatt von (4.35a) *die abgeschwächte Bedingung* (4.36) *erfüllt ist, reduziert sich die Konvergenzordnung auf eins,*

$$\|\vec{E}_n{}' - \vec{E}'(t_n)\| \leq C\tau \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.54}$$

In diesem Fall ist in den Abhängigkeiten von C *die Konstante* C_1 *durch* C_0 *zu ersetzen.*

Beweis. Wir vollziehen die Voraussetzungen von Theorem 4.1 noch einmal in Kurzform nach. Die $\nabla \times \nabla\times$-Formulierung (4.18a) sichert (4.9) mit $g(\vec{E}) := \mathbf{G}\vec{E}$ und (4.19) und liefert die Struktur der Matrix $\mathbf{\Omega}$ in (4.10) mit $\widetilde{\omega} := c\sqrt{\rho_F}$. Nach (4.24) sichert Voraussetzung 4.2 die Potentialstruktur von g in (4.10). Da das System autonom ist, die Nichtlinearität also nicht von der Zeit abhängt, können wir die Zeit um t_0 verschieben. Aus (4.43) erhalten wir nach (4.44) mit (4.42) die Energiebedingung (4.11) an die modifizierten Anfangswerte. Die Forderungen aus (4.22) an die Ortsdiskretisierung sichern zudem die Schranke an $\mathbf{G}$ (4.23) und damit die Glattheit der Nichtlinearität g für Theorem 4.1. Auch die Zeitschrittweite ist passend zu den Voraussetzungen dieser Aussage gewählt. Durch die Wahl von ψ_E und ϕ_E sind nach (4.34) und den Abschätzungen (4.35) die Voraussetzungen an die Filter ψ und ϕ erfüllt. Das Mehrschrittverfahren (4.33) und der erste Zeitschritt (4.52) entsprechen dem Zeitschrittverfahren (4.12) und (4.13). Damit sind alle Voraussetzungen von Theorem 4.1 erfüllt und dessen Anwendung liefert die gewünschte Fehlerabschätzung, wobei die Konstante C dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung abhängt, dafür aber von der Länge des Zeitintervalls $(T - t_0)$, H_0, Schranken an die Ableitungen von g sowie den Konstanten C_1 bzw. C_0, C_2 und C_3.

Die erste Ableitung des oben gewählten g ist gegeben als $\mathrm{D}g = \mathbf{G}$, also $\|\,\mathrm{D}g\,\| = \|\mathbf{G}\| \leq C_g$. Alle weiteren Ableitungen von g verschwinden, sind also durch Null beschränkt.

N.B.: Die Konstante C hängt nicht von C_4 ab, denn die einzige Stelle, für die wir die Abschätzung (4.42) explizit benötigen, ist die Abschätzung an die initiale oszillatorische Energie (4.43). Man kann also sagen, dass C_4 in H_0 enthalten ist. □

Wir formulieren diese Aussage nun für die ursprünglichen Gleichungen und das ursprüngliche Verfahren.

Korollar 4.6. *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen* (4.3) *mit dem numerischen Schema* (4.4). *Die Ortsdiskretisierung, Zeitschrittweite und Filterfunktionen* ψ_E *und* ϕ_E *mögen wie in Lemma* 4.5 *gewählt sein, die beiden übrigen Filterfunktionen seien* (4.8) *und* (4.30) *entsprechend als* $\psi_B \equiv \phi_B \equiv 1$. *Die Startwerte seien* $\vec{E}_0$, $\vec{B}_0$ *und* $\vec{p}_0$ *und mögen* (4.43) *erfüllen. Mit* $\vec{E}'(t)$ *bezeichnen wir die analytische Lösung von* (4.18a) *mit den Startwerten*

$$\vec{E}'(t_0) = \vec{E}_0{}', \qquad \partial_t \vec{E}'(t_0) = \dot{\vec{E}}_0{}',$$

aus (4.39) *mit* (4.41).

Dann erhalten wir die Abschätzung

$$\|\vec{E}_n - \vec{E}'(t_n)\| \leq C\tau^2 \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.55}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$ aus (4.19), τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T - t_0)$ sowie den Konstanten H_0, C_g aus (4.23) und $C_1, ..., C_3$.

Falls anstatt von (4.35a) die abgeschwächte Bedingung (4.36) erfüllt ist, reduziert sich die Konvergenzordnung auf eins,

$$\|\vec{E}_n - \vec{E}'(t_n)\| \leq C\tau \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.56}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen.

Beweis. Wir haben gezeigt, dass das numerische Schema (4.4) mit (4.8) und (4.30) der Mehrschrittformulierung (4.33) genügt. Durch die Wahl der Startwerte mit (4.39) erhalten wir aus (4.40) mit der passenden Wahl von χ aus (4.41), dass der $\vec{E}$-Anteil des Verfahrens (4.4) und der zweite Startwert des Mehrschrittverfahrens,

$$\vec{E}_1{}' = \vec{E}_1$$

übereinstimmen. Da wir auf der exakten Lösung starten und damit beide Startwerte für die Mehrschrittformulierung übereinstimmen, sind dann auch alle weiteren Iterierten gleich:

$$\vec{E}_n{}' = \vec{E}_n, \qquad n \geq 2.$$

Die Ortsdiskretisierung, Zeitschrittweite und Filterfunktionen sind wie in Lemma 4.5 gewählt. Damit dürfen wir dieses anwenden und wir erhalten die Aussage für $n \geq 1$. Für $n = 0$ ist sie sowieso erfüllt, da wir auf der exakten Lösung starten. □

Bemerkung 4.7. Wir erhalten hier zunächst eine Fehlerabschätzung gegen die analytische Lösung mit gestörten Anfangswerten. Für eine Abschätzung gegen die analytische Lösung mit den ursprünglichen Anfangswerten müssen wir erreichen, dass die Anfangswerte nahe genug beieinander liegen, und benötigen die Stabilität von analytischen Lösungen. Diesem Thema wenden wir uns im nächsten Abschnitt zu.

Darüber hinaus müssen wir den Schritt von der $\nabla\times\nabla\times$-Formulierung (4.18) zurück zur ursprünglichen Maxwell-Formulierung (4.3) gehen. Auch dies müssen wir an dieser Stelle noch zurückstellen, da die Störung der Anfangswerte in (4.39) inkompatibel mit der Maxwell-Formulierung ist. Wir können nur den Anfangswert $\vec{B}_0$ direkt mit χ multiplizieren, können aber den Rotationsoperator $\mathbf{C}_B$ und diese matrixwertige Funktion nicht vertauschen. Wir können diesen Schritt also erst auf den ungestörten Anfangswerten vollziehen.

4.7 Fehler in den Anfangswerten und Stabilität bei gestörten Anfangswerten

Wir wollen nun den Fehler des numerischen Verfahrens gegen die „richtige“ analytische Lösung, also diejenige mit ungestörten Anfangswerten, finden. Wir nutzen dazu die bereits gewonnenen Erkenntnisse der vorigen Abschnitte wie folgt aus:

$$\begin{aligned}\|\vec{E}_n - \vec{E}(t_n)\| &\leq \|\vec{E}_n - \vec{E}'(t_n)\| + \|\vec{E}'(t_n) - \vec{E}(t_n)\| \\ &\leq \widetilde{C}\tau^2 + \|\vec{E}'(t_n) - \vec{E}(t_n)\| \qquad \text{(Korollar 4.6)} \\ &\overset{!}{\leq} C\tau^2.\end{aligned}$$

Wir suchen also eine von $\widetilde{\omega}$ unabhängige $\mathcal{O}(\tau^2)$-Abschätzung für die analytischen Lösungen. Um dies zu erreichen, müssen wir zunächst den Fehler im Startwert kontrollieren und werden dazu die weitere Bedingung aus (4.58) an die Startwerte und die zusätzliche Bedingung (4.57) an die Filter benötigen. Im Anschluss daran werden wir mithilfe von Stabilitätseigenschaften zeigen, dass sich der Fehler nicht zu sehr verstärkt.

Lemma 4.8 (Fehler in den den Startwerten)**.** *Für gegebene $\vec{p}_0$, $\vec{E}_0$ und $\vec{B}_0$ bezeichnen wir die Startwerte für (4.18a) aus (4.18b) mit $(\vec{E}_0, \dot{\vec{E}}_0)$ und die aus (4.39) mit $(\vec{E}'_0, \dot{\vec{E}}'_0)$. Die Störung χ aus (4.41) möge (4.42) genügen. Außerdem fordern wir, dass die zusätzliche Abschätzung*

$$|\operatorname{sinc}(z) - \psi(z)| = \left|\operatorname{sinc}(z) - \tfrac{1}{2}(\cos(z)+1)\psi_E(\tfrac{1}{2}z)\right| \leq C_5 z^2 |\operatorname{sinc}(z)| \tag{4.57}$$

für $z \geq 0$ erfüllt ist. Der Startwert für das magnetische Feld möge zusätzlich zu (4.43) der Bedingung

$$\|\boldsymbol{\Omega}^2 \mathbf{C}_B \vec{B}_0\|^2 \leq H_0 \tag{4.58}$$

genügen. Die Konstanten H_0 aus (4.43), C_4 aus (4.42) und C_5 seien dabei unabhängig von $\widetilde{\omega}$ aus (4.19). Dann gilt

$$\left\| \begin{bmatrix} \vec{E}_0 \\ \dot{\vec{E}}_0 \end{bmatrix} - \begin{bmatrix} \vec{E}'_0 \\ \dot{\vec{E}}'_0 \end{bmatrix} \right\| = \left\| \dot{\vec{E}}_0 - \dot{\vec{E}}'_0 \right\| \leq C\tau^2 \tag{4.59}$$

mit einer Konstante C, die nicht von $\widetilde{\omega}$ abhängt, dafür aber von H_0 und C_5.

Beweis. Da wir $\vec{E}'_0 = \vec{E}_0$ gewählt haben, also $\Delta\vec{E}_0 := \vec{E}_0 - \vec{E}'_0 = 0$, brauchen wir nur die zweiten Komponenten

$$\begin{aligned}\Delta\dot{\vec{E}}_0 := \dot{\vec{E}}_0 - \dot{\vec{E}}'_0 &= \mathbf{C}_B\vec{B}_0 - \boldsymbol{\Omega}^2\vec{p}_0 - \chi(\tau\boldsymbol{\Omega})\mathbf{C}_B\vec{B}_0 + \boldsymbol{\Omega}^2\vec{p}_0 \\ &= (1 - \chi(\tau\boldsymbol{\Omega}))\mathbf{C}_B\vec{B}_0\end{aligned}$$

zu betrachten. Es sei darauf hingewiesen, dass sich hinter der Zahl 1 die Identitätsmatrix verbirgt. Um den Faktor τ^2 aus $\Delta\dot{\vec{E}}_0$ zu gewinnen, benutzen wir (4.57). Dadurch entsteht der Faktor $\mathbf{\Omega}^2$, den wir dann in das magnetische Feld verschieben und mit (4.58) kontrollieren. Wir erhalten mit den beiden Filterabschätzungen (4.42) und (4.57)

$$|1 - \chi(z)| \leq C_5 z^2.$$

Definieren wir $f(z) := (1 - \chi(z))/z^2$, so ergibt sich daraus $|f(z)| \leq C_5$ für $z \geq 0$ und $f(z)z^2 = 1 - \chi(z)$. Also erhalten wir

$$\|\Delta\dot{\vec{E}}_0\| = \|f(\tau\mathbf{\Omega})\tau^2\mathbf{\Omega}^2\nabla \times \vec{B}_0\| \leq C_5\tau^2\sqrt{H_0} \tag{4.60}$$

nach der obigen Bedingung (4.58) an den Startwert. Insgesamt erhalten wir also

$$\left\| \begin{bmatrix} \Delta\vec{E}_0 \\ \Delta\dot{\vec{E}}_0 \end{bmatrix} \right\| = \sqrt{\|\Delta\vec{E}_0\|^2 + \|\Delta\dot{\vec{E}}_0\|^2} \leq C_5\sqrt{H_0}\tau^2, \tag{4.61}$$

wobei nach Voraussetzung keine der beiden Konstanten von $\widetilde{\omega}$ abhängt. □

Als nächstes stellen wir sicher, dass der Fehler in den Startwerten über die Zeit nicht zu sehr verstärkt wird.

Lemma 4.9 (Stabilität von Lösungen bei gestörten Anfangswerten)**.** *Wir betrachten die analytische Lösung der ortsdiskretisierten Maxwell-Gleichungen in der $\nabla \times \nabla\times$-Formulierung (4.18a) mit den Startwerten*

$$\vec{E}(t_0) = \Delta\vec{E}_0 = 0, \qquad \partial_t\vec{E}(t_0) = \Delta\dot{\vec{E}}_0. \tag{4.62}$$

Die Ortsdiskretisierung erfülle Voraussetzung 4.2.

Dann gilt

$$\left\| \begin{bmatrix} \vec{E} \\ \partial_t\vec{E} \end{bmatrix}(t) \right\| \leq \sqrt{1 + (T - t_0)^2}\|\Delta\dot{\vec{E}}_0\| \quad \text{für alle} \quad t_0 \leq t \leq T. \tag{4.63}$$

Beweis. Nach (4.21), (4.20) und (4.19) ist die Matrix $(\mathbf{G} - \mathbf{\Omega}^2) =: -\mathbf{B}^2$ symmetrisch negativ semidefinit. Damit ist die Matrix $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ symmetrisch positiv semidefinit mit orthogonalem $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{Id}$ und diagonalem $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, $\lambda_i \geq 0$ für alle $1 \leq i \leq N$.

Wir schreiben (4.18a) in ein System erster Ordnung mit $\dot{\vec{E}}(t) := \partial_t\vec{E}(t)$ um und erhalten

$$\partial_t \begin{bmatrix} \vec{E} \\ \dot{\vec{E}} \end{bmatrix}(t) = \mathbf{A} \begin{bmatrix} \vec{E} \\ \dot{\vec{E}} \end{bmatrix}(t), \qquad \begin{bmatrix} \vec{E} \\ \dot{\vec{E}} \end{bmatrix}(t_0) = \begin{bmatrix} \Delta\vec{E}_0 \\ \Delta\dot{\vec{E}}_0 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & \mathbf{Id} \\ -\mathbf{B}^2 & 0 \end{bmatrix}. \tag{4.64}$$

Die Lösung dieses linearen Systems ist gegeben als

$$\begin{bmatrix} \vec{E} \\ \dot{\vec{E}} \end{bmatrix}(t) = \exp((t-t_0)\mathbf{A}) \begin{bmatrix} \vec{E} \\ \dot{\vec{E}} \end{bmatrix}(t_0) \overset{\Delta\vec{E}_0=0}{=} \begin{bmatrix} (t-t_0)\operatorname{sinc}((t-t_0)\mathbf{B})\Delta\dot{\vec{E}}_0 \\ \cos((t-t_0)\mathbf{B})\Delta\dot{\vec{E}}_0 \end{bmatrix}.$$

Da wir bei othogonalen Zerlegungen die Matrixfunktion auf der Diagonalmatrix auswerten dürfen, vgl. (A.24), erhalten wir mit der Eigenwertzerlegung von $\mathbf{B}$

$$\operatorname{sinc}(t\mathbf{B}) = \mathbf{U}\operatorname{sinc}(t\mathbf{D})\mathbf{U}^T, \qquad \cos(t\mathbf{B}) = \mathbf{U}\cos(t\mathbf{D})\mathbf{U}^T. \tag{4.65}$$

Daraus ergibt sich dann für die Matrixnormen mit der Orthogonalität von $\mathbf{U}$

$$\|\operatorname{sinc}(t\mathbf{B})\| \le \underbrace{\|\mathbf{U}\|}_{=1}\|\operatorname{sinc}(t\mathbf{D})\|\underbrace{\|\mathbf{U}^T\|}_{=1} \le \max_{1\le i\le N}|\operatorname{sinc}(t\lambda_i)| \le 1, \quad \|\cos(t\mathbf{B})\| \le \max_{1\le i\le N}|\cos(t\lambda_i)| \le 1, \tag{4.66}$$

indem wir ausnutzen, dass die λ_i reell sind. Wir erhalten dann wegen $t \le T$

$$\begin{aligned} \left\| \begin{bmatrix} \vec{E} \\ \dot{\vec{E}} \end{bmatrix}(t) \right\| &\le \sqrt{((t-t_0)\|\operatorname{sinc}(t\mathbf{B})\|\|\Delta\dot{\vec{E}}_0\|)^2 + (\|\cos(t\mathbf{B})\|\|\Delta\dot{\vec{E}}_0\|)^2} \\ &\le \sqrt{1+(T-t_0)^2}\|\Delta\dot{\vec{E}}_0\|. \end{aligned}$$

□

4.8 Abschätzung für das elektrische Feld

Mit den Ergebnissen aus dem letzten Abschnitt können wir nun das Hauptresultat für das elektrische Feld beweisen.

Satz 4.10 (Fehlerabschätzung für das elektrische Feld)**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen* (4.3) *mit dem numerischen Schema* (4.4)*. Die analytische Lösung werde mit* $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ *bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ *gelte* $\tau \le \tau_0$*,* τ_0 *unabhängig von* $\widetilde{\omega}$*, wobei* $\tau\widetilde{\omega} \ge c_0 > 0$*. Die Filterfunktionen* $\psi_E, \phi_E : \mathbb{R}_{\ge 0} \longrightarrow \mathbb{R}$ *seien gerade, reellwertige Funktionen mit* $\psi_E(0) = \phi_E(0) = 1$ *und mögen den Abschätzungen* (4.35)*,* (4.42) *und* (4.57) *genügen. Die beiden übrigen Filterfunktionen seien als* $\psi_B \equiv \phi_B \equiv 1$ *gewählt. Die Startwerte seien* $\vec{E}_0$*,* $\vec{B}_0$ *und* $\vec{p}_0$ *und mögen* (4.43) *und* (4.58) *erfüllen. Die auftretenden Konstanten* $C_1, \ldots, C_5$ *und* H_0 *seien unabhängig von* $\widetilde{\omega}$*.*

Dann erhalten wir für das elektrische Feld $\vec{E}$ *in der Lösung von* (4.3) *bzw.* (4.4) *die Abschätzung*

$$\|\vec{E}_n - \vec{E}(t_n)\| \le C\tau^2 \qquad \textit{für} \qquad t_n := t_0 + n\tau \le T. \tag{4.67}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T - t_0)$ sowie den Konstanten H_0, C_g aus (4.23) und C_1, C_2, C_3 und C_5.

Falls anstatt von (4.35a) die abgeschwächte Bedingung (4.36) erfüllt ist, reduziert sich die Konvergenzordnung auf eins,

$$\|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.68}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen.

Beweis. Wir behandeln zuerst den Fall, dass die starken Bedingungen an die Filter erfüllt sind. Wir teilen den Fehler, wie im letzten Abschnitt, in die beiden Anteile

$$\|\vec{E}_n - \vec{E}(t_n)\| \leq \|\vec{E}_n - \vec{E}'(t_n)\| + \|\vec{E}'(t_n) - \vec{E}(t_n)\|$$

auf, wobei $\vec{E}'(t)$ die analytische Lösung von (4.18a) mit den gestörten Startwerten aus (4.39) und χ aus (4.41) bezeichnet. Die Voraussetzungen von Korollar 4.6 sind erfüllt und wir können den ersten Summanden mit

$$\|\vec{E}_n - \vec{E}'(t_n)\| \leq \widehat{C}\tau^2$$

kontrollieren, wobei die Konstante $\widehat{C}$ nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung abhängt, dafür aber von $(T - t_0)$, H_0, C_g, C_1, C_2 und C_3.

Weiterhin haben wir gezeigt, dass die $\vec{E}$-Komponente der Lösung von (4.4) der $\nabla\times\nabla\times$-Formulierung (4.18a) genügt. Also sind wir in der Situation der beiden Lemmata des vorigen Abschnitts. Bezeichnen wir wie vorher mit $\Delta\vec{E}_0 := \vec{E}_0 - \vec{E}'_0 = 0$ und $\Delta\dot{\vec{E}}_0 := \dot{\vec{E}}_0 - \dot{\vec{E}}'_0$ die Fehler in den Startwerten zwischen den beiden analytischen Lösungen, so ist aufgrund der Linearität des Flusses der Differentialgleichung

$$\begin{bmatrix} (\vec{E} - \vec{E}')(t) \\ (\dot{\vec{E}} - \dot{\vec{E}}')(t) \end{bmatrix} = \begin{bmatrix} \Delta\vec{E} \\ \Delta\dot{\vec{E}} \end{bmatrix}(t),$$

wobei $\Delta\vec{E}(t)$,$\Delta\dot{\vec{E}}(t)$ die Lösung mit den entsprechenden Startwerten sei. Weiterhin haben wir in den Voraussetzungen des Satzes diejenigen für die beiden Lemmata eingebaut und dürfen diese folgerichtig anwenden. Sie zeigen insgesamt

$$\begin{aligned} \|\vec{E}'(t) - \vec{E}(t)\| \leq \left\| \begin{bmatrix} \Delta\vec{E} \\ \Delta\dot{\vec{E}} \end{bmatrix}(t) \right\| &\overset{\text{L. 4.9}}{\leq} \sqrt{1 + (T - t_0)^2}\,\|\Delta\dot{\vec{E}}_0\| \\ &\overset{\text{L. 4.8}}{\leq} \sqrt{1 + (T - t_0)^2}\,\widetilde{C}\tau^2 \quad \text{für} \quad t_0 \leq t \leq T \end{aligned}$$

für eine von $\widetilde{\omega}$ unabhängige Konstante $\widetilde{C}$, die aber von H_0 und C_5 abhängt.

Wir erhalten die gewünschte Aussage

$$\|\vec{E}_n - \vec{E}(t_n)\| \leq (\widehat{C} + \sqrt{1 + (T - t_0)^2}\,\widetilde{C})\tau^2 \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T \tag{4.69}$$

mit $C := \widehat{C} + \sqrt{1+(T-t_0)^2}\widetilde{C}$, welche nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung abhängt, dafür aber von $(T-t_0)$, sowie den Konstanten H_0, C_g, C_1, C_2, C_3 und C_5.

Falls nur die schwache Version der Filterbedingung (4.36) erfüllt ist, benutzen wir den Zusatz von Korollar 4.6 und erhalten

$$\|\vec{E}_n - \vec{E}'(t_n)\| \leq \widehat{C}\tau.$$

Schätzen wir dann

$$\|\vec{E}'(t) - \vec{E}(t)\| \leq \sqrt{1+(T-t_0)^2}\widetilde{C}\tau^2 \leq \sqrt{1+(T-t_0)^2}\widetilde{C}\tau_0\tau$$

ab, erhalten wir die Aussage erster Ordnung. □

4.9 Abschätzung für das magnetische Feld

Wir wenden uns nun dem magnetischen Feld zu. Mithilfe des Konvergenzresultats für das elektrische können wir auch für das magnetische Feld die Konvergenz beweisen. Wir stellen dazu die folgende weitere Bedingung

$$|\operatorname{sinc}(z) - \phi_E(\tfrac{1}{2}z)| \leq C_6|z\sin(\tfrac{1}{2}z)| \tag{4.70}$$

für $z \geq 0$ an die Filterfunktionen mit einer Konstante C_6, die nicht von $\widetilde{\omega}$ aus (4.19) abhängt, auf.

Satz 4.11 (Abschätzung für das magnetische Feld)**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen (4.3) mit dem numerischen Schema (4.4). Die analytische Lösung werde mit $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ gelte $\tau \leq \tau_0$, τ_0 unabhängig von $\widetilde{\omega}$, wobei $\tau\widetilde{\omega} \geq c_0 > 0$. Die Filterfunktionen $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ seien gerade, reellwertige Funktionen mit $\psi_E(0) = \phi_E(0) = 1$ und mögen den Abschätzungen (4.35), (4.42), (4.57) und (4.70) genügen. Die beiden übrigen Filterfunktionen seien als $\psi_B \equiv \phi_B \equiv 1$ gewählt. Die Startwerte seien $\vec{E}_0$, $\vec{B}_0$ und $\vec{p}_0$ und mögen (4.43), (4.46) und (4.58) erfüllen. Die auftretenden Konstanten C_c, $C_1, \ldots, C_6$ und H_0 seien unabhängig von $\widetilde{\omega}$.

Dann erhalten wir für das magnetische Feld $\vec{B}$ in der Lösung von (4.3) bzw. (4.4) die Abschätzung

$$\|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau^2 \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.71}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T-t_0)$, der Fehlerkonstante C_E aus der Fehlerabschätzung des elektrischen Feldes sowie den Konstanten C_c aus (4.22) und H_0, C_1, C_2, C_3, C_5 und C_6.

Falls anstatt von (4.35a) die abgeschwächte Bedingung (4.36) erfüllt ist, reduziert sich die Konvergenzordnung auf eins,

$$\|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.72}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen.

Beweis. Wir behandeln zuerst den Fall, dass die starken Bedingungen an die Filter erfüllt sind. Da die Voraussetzungen von Satz 4.10 erfüllt sind, dürfen wir diesen anwenden und bezeichnen die Fehlerkonstante mit C_E. Weiterhin sind auch die Voraussetzungen von Lemma 4.3 erfüllt und dessen Stabilitätsergebnisse gültig.

Zur Abkürzung der Notation definieren wir wieder $\dot{\vec{E}}_0 := \mathbf{C}_B\vec{B}_0 - \mathbf{\Omega}^2\vec{p}_0 = \partial_t\vec{E}(t_0)$.

Aus (4.3b) und (4.29) ergibt sich die Fehlerrekursion

$$\vec{B}(t_{n+1}) - \vec{B}_{n+1} = \vec{B}(t_n) - \vec{B}_n - \frac{\tau}{2}\mathbf{C}_E\left(2\int_0^1 \vec{E}(t_n+\tau s)\,\mathrm{d}s - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})(\vec{E}_n + \vec{E}_{n+1})\right).$$

Wegen $t_{n+1} = t_n + \tau$ ist $t_n + \tau s = t_{n+1} + \tau(s-1)$ und wir erhalten aus der Darstellung der Lösung der $\vec{E}$-Komponente mit der Variation-der-Konstanten-Formel (4.26a)

$$\begin{aligned}
&2\int_0^1 \vec{E}(t_n+\tau s)\,\mathrm{d}s - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})(\vec{E}_n + \vec{E}_{n+1}) \\
&= \int_0^1 \vec{E}(t_n+\tau s)\,\mathrm{d}s + \int_0^1 \vec{E}(t_{n+1}+\tau(s-1))\,\mathrm{d}s - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})(\vec{E}_n + \vec{E}_{n+1}) \\
&= \left(\int_0^1 \cos(\tau s\mathbf{\Omega})\,\mathrm{d}s - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right)\vec{E}(t_n) + \left(\int_0^1 \cos(\tau(s-1)\mathbf{\Omega})\,\mathrm{d}s - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right)\vec{E}(t_{n+1}) \\
&\qquad + \tau\left(\int_0^1 s\,\mathrm{sinc}(\tau s\mathbf{\Omega})\,\mathrm{d}s\;\partial_t\vec{E}(t_n) + \int_0^1 (s-1)\,\mathrm{sinc}(\tau(s-1)\mathbf{\Omega})\,\mathrm{d}s\;\partial_t\vec{E}(t_{n+1})\right) \\
&\qquad + \tau^2\left(\int_0^1 s^2 I_n^+(\tau,s)\,\mathrm{d}s + \int_0^1 (1-s)^2 I_{n+1}^-(\tau,s)\,\mathrm{d}s\right) \\
&\qquad + \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\left((\vec{E}(t_n) - \vec{E}_n) + (\vec{E}(t_{n+1}) - \vec{E}_{n+1})\right),
\end{aligned}$$

wobei

$$I_n^+(\tau,s) := \int_0^1 (1-\xi)\,\mathrm{sinc}(\tau s(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t_n+\tau s\xi)\,\mathrm{d}\xi, \tag{4.73a}$$

$$I_{n+1}^-(\tau,s) := \int_0^1 (1-\xi)\,\mathrm{sinc}(\tau(s-1)(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t_{n+1}+\tau(s-1)\xi)\,\mathrm{d}\xi. \tag{4.73b}$$

Die Integrale vor $\vec{E}(t_n)$, $\vec{E}(t_{n+1})$, $\partial_t\vec{E}(t_n)$ und $\partial_t\vec{E}(t_{n+1})$ können wir mithilfe von (A.18) und (A.19) bestimmen. Die cosc-Funktion in der nächsten Gleichung ist in (A.10) definiert. Durch Einsetzen

der Fehlerrekursion erhalten wir, da wir auf exakten Startwerten starten,

$$\vec{B}(t_n) - \vec{B}_n = -\frac{\tau}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \left[\left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right) \vec{E}(t_l) + \left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right) \vec{E}(t_{l+1}) \right] \tag{4.74a}$$

$$- \frac{\tau^2}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \left[\operatorname{cosc}(\tau\mathbf{\Omega})\partial_t \vec{E}(t_l) - \operatorname{cosc}(\tau\mathbf{\Omega})\partial_t \vec{E}(t_{l+1}) \right] \tag{4.74b}$$

$$- \frac{\tau^3}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \left[\int_0^1 s^2 I_l^+(\tau, s)\,\mathrm{d}s + \int_0^1 (1-s)^2 I_{l+1}^-(\tau, s)\,\mathrm{d}s \right] \tag{4.74c}$$

$$- \frac{\tau}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \phi_E(\tfrac{\tau}{2}\mathbf{\Omega}) \left[(\vec{E}(t_l) - \vec{E}_l) + (\vec{E}(t_{l+1}) - \vec{E}_{l+1}) \right]. \tag{4.74d}$$

Wir werden jetzt die Terme aus dieser Gleichung zeilenweise abschätzen. Dies führen wir aufgrund der Komplexität der notwendigen Argumente in umgekehrter Reihenfolge durch. Nach (4.35b) ist ϕ_E insbesondere durch C_2 und nach (4.22) ist $\mathbf{C}_E$ durch C_c beschränkt. Satz 4.10 liefert uns eine Schranke für die Differenzen von numerischer und analytischer Lösung der $\vec{E}$-Komponente. Insgesamt erhalten wir für (4.74d)

$$\begin{aligned} \left\| \frac{\tau}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \phi_E(\tfrac{\tau}{2}\mathbf{\Omega}) \left[(\vec{E}(t_l) - \vec{E}_l) + (\vec{E}(t_{l+1}) - \vec{E}_{l+1}) \right] \right\| &\leq \tfrac{\tau}{2} C_c C_2 C_E \tau^2 \sum_{l=0}^{n} 2 \\ &\leq C_c C_2 C_E (T - t_0)\tau^2. \end{aligned} \tag{4.75}$$

Fahren wir mit (4.74c) fort. Wegen (A.20) gilt aufgrund der Diagonalstruktur von $\mathbf{\Omega}$ auch, dass $\|\operatorname{sinc}(\tau\mathbf{\Omega})\| \leq 1$ beschränkt ist. Damit und mit (4.48) aus Lemma 4.3 und der Beschränktheit von $\mathbf{G}$ aus (4.23) erhalten wir für die beiden Integrale aus (4.73)

$$\max\left\{ \|I_l^+(\tau, s)\|, \|I_{l+1}^-(\tau, s)\| \right\} \leq \tfrac{1}{2} C_g (1 + 2(T - t_0))\sqrt{H_0}.$$

Dadurch können wir in (4.74c) die Integrale und damit auch deren Summe abschätzen. Insgesamt erhalten wir

$$\begin{aligned} &\left\| \frac{\tau^3}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \left[\int_0^1 s^2 I_l^+(\tau, s)\,\mathrm{d}s + \int_0^1 (1-s)^2 I_{l+1}^-(\tau, s)\,\mathrm{d}s \right] \right\| \\ &\qquad \leq \frac{1}{6} C_c C_g (1 + 2(T - t_0))\sqrt{H_0}(T - t_0)\tau^2, \end{aligned} \tag{4.76}$$

wobei wir durch die Summation über alle Zeitschritte wieder eine τ-Potenz verlieren.

Durch die geschickte Aufteilung am Anfang und den Vorzeichenwechsel in (A.19) wird (4.74b) zu einer Teleskopsumme. Diese fällt bis auf das erste und letzte Glied zusammen. Wegen (A.20),

(4.49) und den bereits verwendeten Schranken erhalten wir dann

$$\begin{aligned}\left\|\frac{\tau^2}{2}\mathbf{C}_E\sum_{l=0}^{n-1}\left[\operatorname{cosc}(\tau\boldsymbol{\Omega})\partial_t\vec{E}(t_l)-\operatorname{cosc}(\tau\boldsymbol{\Omega})\partial_t\vec{E}(t_{l+1})\right]\right\| \\ = \left\|\frac{\tau^2}{2}\mathbf{C}_E\operatorname{cosc}(\tau\boldsymbol{\Omega})(\dot{\vec{E}}_0-\partial_t\vec{E}(t_n))\right\| \\ \leq \frac{1}{2}C_c\frac{1}{2}\cdot 2\cdot 2\sqrt{H_0}\tau^2 = C_c\sqrt{H_0}\tau^2. \end{aligned} \tag{4.77}$$

Kommen wir zum ersten Summanden (4.74a). Die Idee der Abschätzung wurde durch [28, Lemma 3] inspiriert, wobei wir anstelle der partiellen Summation hier wieder eine Teleskopsumme erhalten werden. Wir benutzen zunächst noch einmal die Variation-der-Konstanten-Formel (4.26a), um das $\vec{E}$-Feld als

$$\vec{E}(t_l) = \cos(l\tau\boldsymbol{\Omega})\vec{E}_0 + l\tau\operatorname{sinc}(l\tau\boldsymbol{\Omega})\dot{\vec{E}}_0 + l^2\tau^2\int_0^1(1-\xi)\operatorname{sinc}(l\tau(1-\xi)\boldsymbol{\Omega})\mathbf{G}\vec{E}(t_0+\tau l\xi)\,\mathrm{d}\xi \tag{4.78}$$

auszudrücken. So erhalten wir Summen von trigonometrischen Funktionen. Jetzt benutzen wir trigonometrische Identitäten, um diese kollabieren zu lassen. Wir rechnen zunächst rein formal. Es gelten

$$\cos(lz) = \frac{\sin((l+\frac{1}{2})z)-\sin((l-\frac{1}{2})z)}{2\sin(\frac{1}{2}z)},$$

und

$$l\operatorname{sinc}(lz) = -\frac{\cos((l+\frac{1}{2})z)-\cos((l-\frac{1}{2})z)}{2z\sin(\frac{1}{2}z)}.$$

Die Singularitäten, die entstehen, werden nachher durch die Filter geglättet. Durch Summation erhalten wir

$$\begin{aligned}\left(\operatorname{sinc}(z)-\phi_E(\tfrac{1}{2}z)\right)\left(\sum_{l=0}^{n-1}\cos(lz)+\sum_{l=1}^{n}\cos(lz)\right) \\ = \sin(nz)\cos(\tfrac{1}{2}z)\frac{\operatorname{sinc}(z)-\phi_E(\frac{1}{2}z)}{\sin(\frac{1}{2}z)}, \\ \left(\operatorname{sinc}(z)-\phi_E(\tfrac{1}{2}z)\right)\left(\sum_{l=0}^{n-1}l\operatorname{sinc}(lz)+\sum_{l=1}^{n}l\operatorname{sinc}(lz)\right) \\ = -\left((\cos(nz)-1)\cos(\tfrac{1}{2}z)\right)\frac{\operatorname{sinc}(z)-\phi_E(\frac{1}{2}z)}{z\sin(\frac{1}{2}z)}.\end{aligned} \tag{4.79}$$

Die trigonometrischen Funktionen vor den Brüchen auf den rechten Seiten sind jeweils beschränkt.

Definieren wir

$$\chi_i(z) := \frac{\operatorname{sinc}(z) - \phi_E(\frac{1}{2}z)}{z^i \sin(\frac{1}{2}z)},$$

so erhalten wir aus Voraussetzung (4.70) an den Filter die Abschätzungen

$$|\chi_0(z)| \leq C_6|z| \qquad \text{und} \qquad |\chi_1(z)| \leq C_6, \tag{4.80}$$

die die Singularitäten kontrollieren. Es fehlt nun noch ein letztes Teil für die Abschätzung der ersten beiden Summanden auf der rechten Seite von (4.78). Dem ersten dieser Summanden fehlt ein Faktor τ. Diesen erhalten wir mit derselben Idee, die wir auch in (4.60) aus dem Beweis zu Lemma 4.8 benutzt haben. Diesmal müssen wir allerdings nur eine τ-Potenz gewinnen. Nach Definition der χ-Funktionen erhalten wir $\chi_0(z) = z\chi_1(z)$. Damit können wir feiner, als in der ersten Ungleichung aus (4.80), abschätzen. Es ergibt sich

$$\|\chi_0(\tau\mathbf{\Omega})\vec{E}_0\| = \|\chi_1(\tau\mathbf{\Omega})\tau\mathbf{\Omega}\vec{E}_0\| \leq C_6\tau\sqrt{\tfrac{2}{3}H_0}, \tag{4.81}$$

wobei $\mathbf{\Omega}\vec{E}_0$ mit den Bedingungen (4.43) an die Startwerte kontrolliert wird. Den letzten Summanden in (4.78) lassen wir zunächst stehen, den Rest von (4.74a) können wir nun abschätzen:

$$\begin{aligned}
&\left\| \frac{\tau}{2}\mathbf{C}_E \sum_{l=0}^{n-1} \left[\left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right)\vec{E}(t_l) + \left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right)\vec{E}(t_{l+1}) \right] \right\| \\
&\quad \overset{(4.78)}{\leq} C_c\frac{\tau}{2}\left(\left\| \left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right) \left(\sum_{l=0}^{n-1} \cos(l\tau\mathbf{\Omega}) + \sum_{l=1}^{n} \cos(l\tau\mathbf{\Omega}) \right) \vec{E}_0 \right\| \right. \\
&\qquad\quad + \left\| \tau \left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right) \left(\sum_{l=0}^{n-1} l\operatorname{sinc}(l\tau\mathbf{\Omega}) + \sum_{l=1}^{n} l\operatorname{sinc}(l\tau\mathbf{\Omega}) \right) \dot{\vec{E}}_0 \right\| \\
&\qquad\quad \left. + \left\| \tau^2\mathbf{C}_E \sum_{l=0}^{n}{}' l^2 \left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right) \int_0^1 (1-\xi)\operatorname{sinc}(l\tau(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t_0+\tau l\xi)\,\mathrm{d}\xi \right\| \right) \\
&\quad \overset{(4.79)}{\leq} C_c\frac{\tau}{2}\Big(\|\sin(n\tau\mathbf{\Omega})\cos(\tfrac{\tau}{2}\mathbf{\Omega})\|\|\chi_0(\tau\mathbf{\Omega})\vec{E}_0\| \\
&\qquad\qquad + \tau\|((\cos(n\tau\mathbf{\Omega}) - 1)\cos(\tfrac{\tau}{2}\mathbf{\Omega}))\|\|\chi_1(\tau\mathbf{\Omega})\|\|\dot{\vec{E}}_0\| \\
&\qquad\qquad + 2\,\|J_n(\tau)\|\Big) \\
&\quad \underset{(4.80)}{\overset{(4.81)}{\leq}} \frac{1}{2}C_c\left(C_6\sqrt{\tfrac{2}{3}H_0} + 2C_6\sqrt{2H_0} + 2\|J_n(\tau)\| \right)\tau^2.
\end{aligned} \tag{4.82}$$

Dabei ist

$$J_n(\tau) := \tau \sum_{l=0}^{n}{}' l^2 \left(\operatorname{sinc}(\tau\mathbf{\Omega}) - \phi_E(\tfrac{\tau}{2}\mathbf{\Omega})\right) \int_0^1 (1-\xi)\operatorname{sinc}(l\tau(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t_0+\tau l\xi)\,\mathrm{d}\xi, \tag{4.83}$$

wobei der an das Summenzeichen angebrachte Strich bedeutet, dass der erste und letzte Summand mit einem Gewicht von $\frac{1}{2}$ zu versehen sind.

Für die Funktionenen

$$\vartheta_i(z) := \frac{\operatorname{sinc}(z) - \phi_E(\frac{1}{2}z)}{z^i}, \qquad i \in \{0,1,2\}\,, \tag{4.84}$$

erhalten wir die Beziehungen

$$z\vartheta_i(z) = \vartheta_{i-1}(z) \qquad i \in \{1,2\}\,, \tag{4.85}$$

und mit Voraussetzung (4.70) die Abschätzung

$$|\vartheta_1(z)| \leq C_6 |\sin(\tfrac{1}{2}z)| \leq C_6. \tag{4.86}$$

Weiterhin erhalten wir mit (A.9) die Identität

$$\vartheta_0(z) l(1-\xi)\operatorname{sinc}(l(1-\xi)z) = \vartheta_1(z) z l(1-\xi)\operatorname{sinc}(l(1-\xi)z) = \vartheta_1(z)\sin(l(1-\xi)z). \tag{4.87}$$

Damit lässt sich in J_n einer der beiden l-Faktoren via

$$J_n(\tau) = \tau \sum_{l=0}^{n}{}' l\vartheta_1(\tau\mathbf{\Omega}) \int_0^1 \sin(l\tau(1-\xi)\mathbf{\Omega})\mathbf{G}\vec{E}(t_0 + \tau l\xi)\,\mathrm{d}\xi \tag{4.88}$$

eliminieren. Von dem verbleibenden Faktor l abgesehen, ist der Term unter der Summe beschränkt. Damit könnten wir eine $\mathcal{O}(n)$-Abschätzung für J_n erhalten und damit eine Fehlerschranke der Ordnung eins für das magnetische Feld. Für Konvergenz zweiter Ordnung schätzen wir jedoch J_n noch einmal mithilfe von (4.70) schärfer ab. Mit dieser Voraussetzung und (A.9) erhalten wir

$$|\operatorname{sinc}(z) - \phi_E(\tfrac{1}{2}z)| \leq C_6 |z\sin(\tfrac{1}{2}z)| = C_6 |z(\tfrac{1}{2}z)\operatorname{sinc}(\tfrac{1}{2}z)| \overset{(A.20)}{\leq} \tfrac{1}{2}C_6 z^2 \tag{4.89}$$

und damit für die in (4.84) definierte Funktion ϑ_2

$$|\vartheta_2(z)| \leq \tfrac{1}{2}C_6. \tag{4.90}$$

Es gilt

$$\frac{\partial}{\partial\xi}\vartheta_2(z)\cos(l(1-\xi)z) = \vartheta_2(z) l z \sin(l(1-\xi)z) \overset{(4.85)}{=} \vartheta_1(z) l \sin(l(1-\xi)z).$$

Damit können wir für differenzierbares f mit Ableitung f' die partielle Integration

$$\begin{aligned}\int_0^1 \vartheta_1(z) l \sin(l(1-\xi)z) f(t_0 + l\tau\xi)\,\mathrm{d}\xi =& \left[\vartheta_2(z)\cos(l(1-\xi)z) f(t_0 + l\tau\xi)\right|_{\xi=0}^{\xi=1} \\ &- \int_0^1 \vartheta_2(z)\cos(l(1-\xi)z) f'(t_0 + l\tau\xi) l\tau\,\mathrm{d}\xi\end{aligned}$$

durchführen. Angewandt auf J_n in der Form (4.88) erhalten wir

$$\begin{aligned} J_n(\tau) =& \tau \sum_{l=0}^{n}{}' \left[\vartheta_2(\tau\mathbf{\Omega}) \cos(l\tau(1-\xi)\mathbf{\Omega}) \mathbf{G} \vec{E}(t_0 + l\tau\xi) \right|_{\xi=0}^{\xi=1} \\ & - \tau \sum_{l=0}^{n}{}' \int_0^1 \vartheta_2(\tau\mathbf{\Omega}) \cos(l\tau(1-\xi)\mathbf{\Omega}) \mathbf{G} \partial_t \vec{E}(t_0 + l\tau\xi) l\tau \, \mathrm{d}\xi, \end{aligned}$$

wodurch der Randterm keinen Faktor l mehr enthält und das neue Integral den Faktor τ hinzugewonnen hat. Die Terme im ersten Summenzeichen sind alle beschränkt, diejenigen im zweiten wegen $l\tau \leq (T - t_0)$ ebenfalls. Damit können wir J_n mit (4.90), der Beschränktheit des Kosinus, den Schranken an $\mathbf{G}$ und $\vec{E}_0$ und Lemma 4.3 gegen

$$\|J_n(\tau)\| \leq (T-t_0)\tfrac{1}{2} C_6 C_g (1 + 1 + 2(T-t_0) + 2(T-t_0))\sqrt{H_0}$$

abschätzen. Zusammen mit (4.82) ergibt sich

$$\|(4.74a)\| \leq \frac{1}{2} C_c \left(C_6 \sqrt{\tfrac{2}{3} H_0} + 2 C_6 \sqrt{2 H_0} + (T-t_0)\tfrac{1}{2} C_6 C_g 2 (1 + 2(T-t_0))\sqrt{H_0} \right) \tau^2$$

und mit (4.74) und den Abschätzungen für die Terme (4.82), (4.77) und (4.76)

$$\begin{aligned} \|\vec{B}(t_n) - \vec{B}_n\| \leq & \left[\frac{1}{2} C_c \left(C_6 \sqrt{\tfrac{2}{3} H_0} + 2 C_6 \sqrt{2 H_0} + \tau (T-t_0)\tfrac{1}{2} C_6 C_g 2 (1 + 2(T-t_0))\sqrt{H_0} \right) \tau^2 \right. \\ & \left. + C_c \sqrt{H_0} \tau^2 + \frac{1}{6} C_c C_g (1 + 2(T-t_0))(T-t_0)\sqrt{H_0} + C_c C_2 C_E (T-t_0) \right] \tau^2. \end{aligned} \tag{4.91}$$

Alle auftretenden Konstanten sind nach Voraussetzung von $\widetilde{\omega}$ unabhängig.

Falls nur die schwache Version der Filterbedingung (4.36) erfüllt ist, verwenden wir den Zusatz aus Satz 4.10 und der Fehlerterm (4.74d) verliert eine Ordnung. Dann müssen wir in den Abschätzungen für die anderen drei Summanden (4.74a), (4.74b) und (4.74c) je einmal $\tau \leq \tau_0 \leq (T - t_0)$ abschätzen und erhalten globale Konvergenz der Ordnung eins. □

Bemerkung 4.12. Für die Wahl $\phi_E(z) = \mathrm{sinc}(2z)$ verschwindet die linke Seite der Bedingung (4.70) und wir können $C_6 = 0$ wählen. Außerdem verschwindet auch der erste Summand in (4.74), wodurch sich obiger Beweis drastisch vereinfacht. Allerdings muss dann noch gezeigt werden, dass dieser Filter alle vorherigen Annahmen an die Filterfunktionen und eine weitere, die im nächsten Abschnitt für die Impulse aufgestellt wird, erfüllt. Dies ist ohne weitere Probleme mit den Techniken aus Abschnitt 4.12 möglich. Wir haben diese längere Variante gewählt, da diese uns mehr Freiheiten für die Wahl des Filters lässt. Wir stellen eine Wahl von Filterfunktionen vor, bei der ϕ_E nur mit der halben Frequenz schwingen muss, wir aber trotzdem Ordnung zwei erhalten.

Bemerkung 4.13. In (4.79) können wir erkennen, wie die Namen „Resonanz“ und „Filterfunktion“ entstanden sind. Die Summen auf den linken Seiten können sich um ungünstigsten Fall, d.h., falls die Argumente der trigonometrischen Funktionen etwa alle Maximalstellen dieser Funktionen sind, zu n aufsummieren, wodurch Resonanzen im Fehler enstehen. Die rechte Seite dieser Formeln zeigt jeweils, wann dies passiert, nämlich an den Nullstellen von $\sin(\frac{1}{2}z)$, also bei geraden Vielfachen von π – genau dort, wo wir sie im vorigen Kapitel auch numerisch erkannt haben. Falls der Filter $\phi_E \equiv 1$ gewählt wird, wie es im ungefilterten Verfahren der Fall ist, kann dies nicht verhindert werden, da $\operatorname{sinc}(z) - 1$ nicht an den richtigen Stellen klein wird. Falls die Filterfunktion jedoch so gewählt ist, dass die beiden Brüche auf den rechten Seiten von (4.79) beschränkt bleiben, wird dieses Aufsummieren und damit die Resonanz verhindert. Das Filtern der Singularitäten bewirken wir also dadurch, dass wir sie zu der Differenz zwischen sinc und der Filterfunktion verschieben.

4.10 Abschätzung für die Impulse

Wenden wir uns nun den Impulsen zu. Mithilfe des Konvergenzresultates für das elektrische und das magnetische Feld können wir auch für diese Konvergenz beweisen. Wir benötigen dazu die folgenden weiteren Bedingungen

$$|\psi_E(z)| \leq C_7 \tag{4.92}$$

und

$$|\operatorname{sinc}^2(\tfrac{1}{2}z) - \operatorname{sinc}(z)\phi_E(\tfrac{1}{2}z)| \leq C_8 \sin^2(\tfrac{1}{2}z) \tag{4.93}$$

für $z \geq 0$ an die Filterfunktionen mit Konstanten C_7 und C_8, die nicht von $\widetilde{\omega}$ aus (4.19) abhängen.

Satz 4.14 (Abschätzung für die Impulse)**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen (4.3) mit dem numerischen Schema (4.4). Die analytische Lösung werde mit $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ gelte $\tau \leq \tau_0$, τ_0 unabhängig von $\widetilde{\omega}$, wobei $\tau\widetilde{\omega} \geq c_0 > 0$. Die Filterfunktionen $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ seien gerade, reellwertige Funktionen mit $\psi_E(0) = \phi_E(0) = 1$ und mögen den Abschätzungen (4.35), (4.42), (4.57), (4.70), (4.92) und (4.93) genügen. Die beiden übrigen Filterfunktionen seien als $\psi_B \equiv \phi_B \equiv 1$ gewählt. Die Startwerte seien $\vec{E}_0$, $\vec{B}_0$ und $\vec{p}_0$ und mögen (4.43), (4.46), (4.47) und (4.58) erfüllen. Die auftretenden Konstanten C_c, $C_1, \ldots, C_8$ und H_0 seien unabhängig von $\widetilde{\omega}$.

Dann erhalten wir für den Anteil der Impulse $\vec{p}$ der Lösung von (4.3) bzw. (4.4) die Abschätzung

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau^2 \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.94}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge

des Zeitintervalls $(T-t_0)$, den Fehlerkonstanten C_E und C_B aus den Fehlerabschätzungen des elektrischen und des magnetischen Feldes sowie den Konstanten C_c aus (4.22) und H_0, C_1 ,C_2, C_3 und $C_5, ..., C_8$.

Falls anstatt von (4.35a) die abgeschwächte Bedingung (4.36) erfüllt ist, reduziert sich die Konvergenzordnung auf eins,

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.95}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen.

Beweis. Wir behandeln zuerst den Fall, dass die starken Bedingungen an die Filter erfüllt sind. Da wir die Voraussetzungen der Sätze 4.10 und 4.11 erfüllen, dürfen wir diese anwenden. Wir bezeichnen die Fehlerkonstanten mit C_E und C_B. Weiterhin erfüllen wir auch die Voraussetzungen von Lemma 4.3 und können auf dessen Stabilitätsergebnisse zurückgreifen.

Wir benutzen die Darstellung (4.31) der numerischen Lösung und den Hauptsatz der Differential- und Integralrechnung, angewandt auf (4.3c), sowie die Variation-der-Konstanten-Formel (4.26a) für die Darstellung der analytischen Lösung und erhalten

$$\begin{aligned}
\vec{p}(t_{n+1}) - \vec{p}_{n+1} &= \vec{p}(t_n) + \tau \left(\int_0^1 \cos(\tau s \boldsymbol{\Omega}) \, \mathrm{d}s \vec{E}(t_n) - \operatorname{sinc}(\tau \boldsymbol{\Omega}) \vec{E}_n \right) \\
&\quad + \tau \int_0^1 \tau s \operatorname{sinc}(\tau s \boldsymbol{\Omega}) \, \mathrm{d}s \, \partial_t \vec{E}(t_n) - \cos(\tau \boldsymbol{\Omega}) \vec{p}_n \\
&\quad - \tfrac{\tau^2}{2} \operatorname{sinc}(\tau \boldsymbol{\Omega}) \psi_E(\tfrac{\tau}{2}\boldsymbol{\Omega}) \mathbf{C}_B \vec{B}_n + \tau^3 I_n \\
&= \vec{p}(t_n) + \tau \operatorname{sinc}(\tau \boldsymbol{\Omega})(\vec{E}(t_n) - \vec{E}_n) + \tau^3 I_n \\
&\quad + \tau^2 \left(\operatorname{cosc}(\tau \boldsymbol{\Omega}) \mathbf{C}_B \vec{B}(t_n) - \tfrac{1}{2} \operatorname{sinc}(\tau \boldsymbol{\Omega}) \psi_E(\tfrac{\tau}{2}\boldsymbol{\Omega}) \mathbf{C}_B \vec{B}_n \right) \qquad (4.96a) \\
&\quad - \tau^2 \boldsymbol{\Omega}^2 \operatorname{cosc}(\tau \boldsymbol{\Omega}) \vec{p}(t_n) - \cos(\tau \boldsymbol{\Omega}) \vec{p}_n \qquad (4.96b)
\end{aligned}$$

mit (A.18), (A.19) und (4.3a), wobei

$$I_n := \int_0^1 s^2 \int_0^1 (1-\xi) \operatorname{sinc}(\tau s (1-\xi) \boldsymbol{\Omega}) \mathbf{G} \vec{E}(t_n + \tau s \xi) \, \mathrm{d}\xi \, \mathrm{d}s - \tfrac{1}{4} \operatorname{sinc}(\tau \boldsymbol{\Omega}) \psi_E(\tfrac{\tau}{2}\boldsymbol{\Omega}) \mathbf{G} \phi_E(\tfrac{\tau}{2}\boldsymbol{\Omega}) \vec{E}_n \tag{4.97}$$

ist. Für (4.96b) benutzen wir (A.15) und erhalten

$$-\tau^2 \boldsymbol{\Omega}^2 \operatorname{cosc}(\tau \boldsymbol{\Omega}) \vec{p}(t_n) - \cos(\tau \boldsymbol{\Omega}) \vec{p}_n = \cos(\tau \boldsymbol{\Omega})(\vec{p}(t_n) - \vec{p}_n) - \vec{p}(t_n), \tag{4.98}$$

für (4.96a) analog mit (A.17)

$$\begin{aligned}
&\tau^2 \left(\operatorname{cosc}(\tau \boldsymbol{\Omega}) \mathbf{C}_B \vec{B}(t_n) - \tfrac{1}{2} \operatorname{sinc}(\tau \boldsymbol{\Omega}) \psi_E(\tfrac{\tau}{2}\boldsymbol{\Omega}) \mathbf{C}_B \vec{B}_n \right) \\
&\qquad = \frac{\tau^2}{2} \Big((\operatorname{sinc}^2(\tfrac{\tau}{2}\boldsymbol{\Omega}) - \operatorname{sinc}(\tau \boldsymbol{\Omega}) \psi_E(\tfrac{\tau}{2}\boldsymbol{\Omega})) \mathbf{C}_B \vec{B}(t_n) \\
&\qquad\qquad + \operatorname{sinc}(\tau \boldsymbol{\Omega}) \psi_E(\tfrac{\tau}{2}\boldsymbol{\Omega}) \mathbf{C}_B (\vec{B}(t_n) - \vec{B}_n) \Big) . \qquad (4.99)
\end{aligned}$$

Wir definieren die weitere Hilfsgröße

$$J_n := \tau \left(\operatorname{sinc}(\tau\mathbf{\Omega})(\vec{E}(t_n) - \vec{E}_n) + \tau \operatorname{sinc}(\tau\mathbf{\Omega})\psi_E(\tfrac{\tau}{2}\mathbf{\Omega})\mathbf{C}_B(\vec{B}(t_n) - \vec{B}_n) + \tau^2 I_n \right). \tag{4.100}$$

Mit Beschränktheit von ψ_E (4.92), mit $\|\vec{E}_n\| \leq \|\vec{E}(t_n)\| + \|\vec{E}_n - \vec{E}(t_n)\|$, mit der E-Feld-Abschätzung in Satz 4.10 und Stabilitätslemma 4.3 erhalten wir zunächst

$$\|I_n\| \leq \left(\tfrac{1}{3}\tfrac{1}{2} + \tfrac{1}{4}C_7C_2\right) C_g(1 + 2(T - t_0))\sqrt{H_0} + \tfrac{1}{4}C_7C_gC_2C_E\tau^2 =: C_I, \tag{4.101}$$

wobei wir den Faktor $\tau^2 \leq \tau_0^2$ in der Konstante eliminieren können. Damit, mit der Beschränktheit von sinc, ψ_E und $\mathbf{C}_B$ und den vorigen beiden Sätzen können wir dann

$$\|J_n\| \leq C_E\tau^2 + \tau C_7C_cC_B\tau^2 + C_I\tau^2 =: C_J\tau^2 \tag{4.102}$$

abschätzen. Wir erhalten dann aus (4.96) mit (4.98) und (4.99) die Fehlerrekursion

$$\vec{p}(t_n) - \vec{p}_n = \tau \sum_{l=0}^{n} \cos^l(\tau\mathbf{\Omega})J_{n-l-1} + \tfrac{\tau^2}{2} \sum_{l=0}^{n} \cos^l(\tau\mathbf{\Omega})(\operatorname{sinc}^2(\tfrac{\tau}{2}\mathbf{\Omega}) - \operatorname{sinc}(\tau\mathbf{\Omega})\psi_E(\tfrac{\tau}{2}\mathbf{\Omega}))\mathbf{C}_B\vec{B}(t_{n-l-1}). \tag{4.103}$$

Der erste Summand hat wegen der Abschätzung von J_n und dem führenden τ die richtige Ordnung. Die Abhängigkeit von $\tau_0 \leq T - t_0$ kann man eliminieren. Für den zweiten Summanden wenden wir noch einmal trigonometrische Identitäten an, denn es gilt

$$\cos^n(z) = \frac{\cos^{n+1}(z) - \cos^n(z)}{-2\sin^2(\frac{1}{2}z)},$$

und wir erhalten durch die partielle Summation (A.32) mit $f_l := \frac{\cos^l(z)}{-2\sin^2(\frac{1}{2}z)}$ und $g_l := \mathbf{C}_B\vec{B}(t_{n-l-1})$

$$\begin{aligned}
&(\operatorname{sinc}^2(\tfrac{1}{2}z) - \operatorname{sinc}(z)\psi_E(\tfrac{1}{2}z)) \sum_{l=0}^{n-1} \cos^l(z)\mathbf{C}_B\vec{B}(t_{n-l-1}) \\
&\quad = \frac{\operatorname{sinc}^2(\tfrac{1}{2}z) - \operatorname{sinc}(z)\psi_E(\tfrac{1}{2}z)}{-2\sin^2(\tfrac{1}{2}z)} \\
&\qquad \cdot \left(\sum_{l=1}^{n} \cos^l(z)\mathbf{C}_B(\vec{B}(t_{n-l}) - \vec{B}(t_{n-l-1})) + \cos^n(z)\mathbf{C}_B\vec{B}_0 - \mathbf{C}_B\vec{B}(t_n) \right).
\end{aligned} \tag{4.104}$$

Hier kommt wieder unsere Filterfunktion ins Spiel, denn mit (4.93) erhalten wir

$$\left| \frac{\operatorname{sinc}^2(\tfrac{1}{2}z) - \operatorname{sinc}(z)\psi_E(\tfrac{1}{2}z)}{-2\sin^2(\tfrac{1}{2}z)} \right| \leq \tfrac{1}{2}C_8$$

und wir dürfen mit Lemma 4.3

$$\|(4.104)\| \leq \tfrac{1}{2}C_8 \left(\sum_{l=0}^{n-1} \|\mathbf{C}_B(\vec{B}(t_{n-l}) - \vec{B}(t_{n-l-1}))\| + C_c\,[1+1+2(T-t_0)]\,\sqrt{H_0} \right)$$

abschätzen. Die letzte nötige τ-Potenz erhalten wir ein weiteres Mal mit dem Hauptsatz der Differential- und Integralrechnung via

$$\begin{aligned} \|\mathbf{C}_B(\vec{B}(t_{n-l}) - \vec{B}(t_{n-l-1}))\| &= \left\| \mathbf{C}_B \left(\vec{B}(t_{n-l-1}) - \tau \int_0^1 \mathbf{C}_E E(t_{n-l-1} + \tau\xi)\,\mathrm{d}\xi - \vec{B}(t_{n-l-1}) \right) \right\| \\ &\leq \tau C_g (1 + 2(T-t_0))\sqrt{H_0}. \end{aligned}$$

Zusammenfassend erhalten wir

$$\|\vec{p}(t_n) - \vec{p}_n\| \leq (T-t_0)C_J\tau^2 + \tfrac{1}{4}C_8(((T-t_0)C_g + C_c)(1+2(T-t_0)) + C_c)\sqrt{H_0}\tau^2 \qquad (4.105)$$

Alle auftretenden Konstanten sind nach Voraussetzung von $\widetilde{\omega}$ unabhängig.

Falls nur die schwache Version der Filterbedingung (4.36) erfüllt ist, verwenden wir den Zusatz aus den Sätzen 4.10 und 4.11. Dann reduzieren sich in der Norm (4.102) von J_n (4.100) und in der Norm (4.101) von I_n (4.97) die τ-Potenzen der Beiträge der Fehler aus dem elektrischen und dem magnetischen Feld um je eins. Die Ordnungsreduktion des magnetischen Feldes ist nicht gravierend, da wir vor dem mittleren Summanden von J_n noch einen Vorfaktor τ ausnutzen können. Die Reduktion im elektrischen Feld reduziert die τ-Potenz vor dem ersten Summanden aus (4.102) und im zweiten Summanden aus (4.101). I_n muss nur gegen eine τ-unabhängige Konstante abgeschätzt werden, der erste Summand aus (4.102) reduziert jedoch die Abschätzung von J_n in (4.102) auf erste Ordnung. □

4.11 Vollständige Fehlerabschätzung für das Splitting-Verfahren mit Voraussetzungen an die Filterfunktionen

Korollar 4.15 (Fehlerabschätzung mit skalierten Impulsen)**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen* (4.3) *mit dem numerischen Schema* (4.4)*. Die analytische Lösung werde mit* $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ *bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ *gelte* $\tau \leq \tau_0$, τ_0 *unabhängig von* $\widetilde{\omega}$*, wobei* $\tau\widetilde{\omega} \geq c_0 > 0$*. Die Filterfunktionen* $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ *seien gerade, reellwertige Funktionen mit* $\psi_E(0) = \phi_E(0) = 1$ *und mögen den Abschätzungen* (4.35), (4.42), (4.57), (4.70), (4.92) *und* (4.93) *genügen. Die beiden übrigen Filterfunktionen seien als* $\psi_B \equiv \phi_B \equiv 1$ *gewählt. Die Startwerte seien* $\vec{E}_0$, $\vec{B}_0$ *und* $\vec{p}_0$ *und mögen* (4.43), (4.46), (4.47) *und* (4.58) *erfüllen. Die auftretenden Konstanten* $C_c, C_1, \ldots, C_8$ *und* H_0 *seien unabhängig von* $\widetilde{\omega}$*.*

Dann erhalten wir die Abschätzung

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau^2, \quad \|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau^2, \quad \|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau^2, \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.106}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T - t_0)$sowie den Konstanten C_c aus (4.22) und H_0, C_1, C_2, C_3 und $C_5, ..., C_8$.

Falls anstatt von (4.35a) die abgeschwächte Bedingung (4.36) erfüllt ist, reduziert sich die Konvergenzordnung auf eins,

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau, \quad \|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau, \quad \|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau, \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.107}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen. □

Da wir obige Fehlerabschätzung eigentlich für das ursprüngliche Verfahren (4.1) und die ursprünglichen Gleichungen (2.13) gesucht haben, formulieren wir das Resultat auch für diese Variante. Um zu der besser analysierbaren Formulierung zu gelangen, hatten wir die Impulse $\widetilde{\vec{p}} = \frac{1}{c}\vec{p}$ umskaliert und den Beweis für $\widetilde{\vec{p}}$geführt, wobei wir aus Notationsgründen auf die Tilde verzichtet haben. Unter Verwendung des vorigen Korollars und unter Umkehrung dieser Operation erhalten wir das folgende Korollar.

Korollar 4.16 (Fehlerabschätzung für das Splitting-Verfahren)**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen (3.1) mit dem numerischen Schema (4.1). Die analytische Lösung werde mit $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ gelte $\tau \leq \tau_0$, τ_0 unabhängig von $\widetilde{\omega}$, wobei $\tau\widetilde{\omega} \geq c_0 > 0$. Die Filterfunktionen $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ seien gerade, reellwertige Funktionen mit $\psi_E(0) = \phi_E(0) = 1$ und mögen den Abschätzungen (4.35), (4.42), (4.57), (4.70), (4.92) und (4.93) genügen. Die beiden übrigen Filterfunktionen seien als $\psi_B \equiv \phi_B \equiv 1$ gewählt. Die Startwerte seien $\vec{E}_0$, $\vec{B}_0$ und $\vec{p}_0$ und mögen (4.43), (4.46), (4.47) und (4.58) erfüllen. Die auftretenden Konstanten $C_c, C_1, \ldots, C_8$ und H_0 seien unabhängig von $\widetilde{\omega}$.

Dann erhalten wir die Abschätzung

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau^2, \quad \|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau^2, \quad \|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau^2, \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.108}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T - t_0)$ sowie den Konstanten C_c aus (4.22) und H_0, C_1, C_2, C_3 und $C_5, ..., C_8$.

Falls anstatt von (4.35a) die abgeschwächte Bedingung (4.36) erfüllt ist, reduziert sich die Konvergenzord-

nung auf eins,

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau, \quad \|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau, \quad \|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau, \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.109}$$

In diesem Fall ist in den Abhängigkeiten von C die Konstante C_1 durch C_0 zu ersetzen. □

Bemerkung 4.17. Die Fehlerkonstanten lassen sich in den Sätzen 4.10 in (4.69), 4.11 in (4.91) und 4.14 in (4.105) bis auf diejenigen, die aus Theorem 4.1 stammen, nachvollziehen.

Wir erhalten ein quadratisches Wachstum mit der Länge des Zeitintervalls $(T - t_0)$, welches mit dem linearen Wachstum des elektrischen Feldes (4.48) zusammen mit Aufsummierungsargumenten und dem quadratischen Wachstum des magnetischen Feldes (4.50) aus Lemma 4.3 zusammenhängt. Wenn die Differentialgleichung auch (4.51) erfüllt, lässt sich dieses Wachstum auf lineares Wachstum mit $(T - t_0)$ reduzieren.

4.12 Wahl der Filterfunktionen und Konvergenzresultate

Das Konvergenzresultat des vorigen Abschnittes liefert die erhoffte Fehlerschranke unter diversen Bedingungen an die Filterfunktionen. A priori ist jedoch nicht klar, dass diese überhaupt simultan erfüllbar sind.

In [89] wurde, allerdings ohne Kenntnis der hier gezeigten Theorie und nur auf Basis numerischer Experimente, die Wahl

$$\psi_E(z) = \phi_E(z) = \operatorname{sinc}(z) \qquad \text{und} \qquad \psi_B \equiv \phi_B \equiv 1 \tag{4.110}$$

vorgeschlagen. Es zeigt sich, dass diese Wahl nicht ganz ausreicht, um alle Bedingungen zu erfüllen. Stattdessen verwenden wir alternativ die Wahl

$$\psi_E(z) = \operatorname{sinc}^2(z), \qquad \phi_E(z) = \operatorname{sinc}(z) \qquad \text{und} \qquad \psi_B \equiv \phi_B \equiv 1. \tag{4.111}$$

Im Folgenden zeigen wir, dass mit dieser Wahl alle Voraussetzungen erfüllt werden. Zur Übersichtlichkeit fassen wir alle Filtervoraussetzungen aus Korollar 4.16 hier noch einmal zusammen. Dort wird gefordert, dass die Filterfunktionen $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ gerade, reellwertige Funktio-

nen mit $\psi_E(0) = \phi_E(0) = 1$ seien, die die Bedingungen

$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)| \leq C_1 \operatorname{sinc}^2(\tfrac{1}{2}z), \tag{4.35a}$$
$$|\phi_E(\tfrac{1}{2}z)| \leq C_2 |\operatorname{sinc}(\tfrac{1}{2}z)|, \tag{4.35b}$$
$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)\phi_E(\tfrac{1}{2}z)| \leq C_3 |\operatorname{sinc}(z)|, \tag{4.35c}$$
$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)| \leq C_4 |\operatorname{sinc}(z)|, \tag{4.42}$$
$$\left|\operatorname{sinc}(z) - \tfrac{1}{2}(\cos(z)+1)\psi_E(\tfrac{1}{2}z)\right| \leq C_5 z^2 |\operatorname{sinc}(z)|, \tag{4.57}$$
$$|\operatorname{sinc}(z) - \phi_E(\tfrac{1}{2}z)| \leq C_6 |z \sin(\tfrac{1}{2}z)|, \tag{4.70}$$
$$|\psi_E(z)| \leq C_7 \tag{4.92}$$

und

$$|\operatorname{sinc}^2(\tfrac{1}{2}z) - \operatorname{sinc}(z)\phi_E(\tfrac{1}{2}z)| \leq C_8 \sin^2(\tfrac{1}{2}z) \tag{4.93}$$

erfüllen mögen. Falls (4.35a) nicht erfüllbar ist, kann diese durch

$$|(\cos(z)+1)\psi_E(\tfrac{1}{2}z)| \leq C_0 |\operatorname{sinc}(\tfrac{1}{2}z)| \tag{4.36}$$

für eine Konvergenz von erster Ordnung ersetzt werden. Die beiden übrigen Filterfunktionen seien als $\psi_B \equiv \phi_B \equiv 1$ gewählt.

Offensichtlich sind die beiden Filterfunktionen ψ_E, ϕ_E sowohl aus (4.110) als auch aus (4.111) gerade, reellwertig und mit dem richtigen Wert im Ursprung ausgestattet. Auch ψ_B und ϕ_B sind korrekt gewählt. Wir müssen also die Ungleichungen verifizieren.

In Abschnitt 3.4 haben wir festgestellt, dass die Resonanzen im Fehler schon für die Wahl ohne Filterfunktionen, also $\psi_E \equiv \phi_E \equiv 1$, immer bei Vielfachen von 2π auftreten. Für andere Verfahren der Bauart (4.12) treten aber auch bei ungeraden Vielfachen von π Resonanzen auf, vgl. etwa [34, Kapitel XIII] und [28, 30]. Die Anforderungen an die Filter, die sich auf eine Filterung bei ungeraden Vielfachen von π beziehen, sind (4.35c), (4.42) und (4.57), alle anderen haben durch den Faktor $\frac{1}{2}$ im Argument eine niedrigere Frequenz. Für diese Bedingungen liefert jedoch bereits der natürlich auftretende Faktor $\cos(z) + 1$ die Filterung bei den ungeraden Vielfachen, sodass Resonanzen auch im ungefilterten Verfahren nur bei geraden Vielfachen auftreten.

Wir diskutieren nun das Verhalten der beiden Filterauswahlen (4.110) und (4.111) im Bezug auf obige Filterabschätzungen.

Gleich die erste Abschätzung ist für (4.110) verletzt, denn die rechte Seite fordert doppelte Nullstellen an Vielfachen von 2π und ein quadratisches Abklingverhalten für $z \longrightarrow \infty$. Um dies zu veranschaulichen, haben wir in Abbildung 4.1 beide Seiten der Abschätzung nach einer Multiplikation mit $(z/2)^2$ dargestellt. Andererseits ist wegen $|\cos(z) + 1| \leq 2$ die Bedingung (4.36) mit

$C_0 = 2$ erfüllt. Im Gegensatz dazu erfüllt (4.111) die Bedingung (4.35a) für die bessere Konvergenz mit $C_1 = 2$. Dieses Verhalten ist, wieder nach Multiplikation mit $(z/2)^2$, in Abbildung 4.2 illustriert.

Die beiden Wahlen (4.110) und (4.111) erfüllen offensichtlich die zweite Bedingung (4.35b) mit $C_2 = 1$, sogar mit Gleichheit.

Aus der vierten und zweiten Bedingung zusammen erhalten wir die dritte mit $C_3 = C_2C_4$, sodass wir diese nicht zu diskutieren brauchen.

Aus

$$\cos(z) + 1 = 2\cos^2(\tfrac{1}{2}z) \quad \text{und} \quad \operatorname{sinc}(\tfrac{1}{2}z)\cos(\tfrac{1}{2}z) = \operatorname{sinc}(z) \tag{4.112}$$

erhalten wir wegen der Beschränktheit von $\cos$ und sinc, dass die Bedingung (4.42) von beiden Wahlen (4.110) und (4.111) der Filterfunktionen mit Konstante $C_4 = 2$ erfüllt ist. Für (4.45) hatten wir angenommen, dass $C_4 \geq 2$ gelte, wir müssen diese Schranke also nicht weiter vergrößern. In den Abbildungen 4.1 bzw. 4.2 sind beide Seiten der Bedingung nach Multiplikation mit z gezeichnet.

Ebenfalls aus (4.112) folgen

$$|\operatorname{sinc}(z) - \tfrac{1}{2}(\cos(z)+1)\operatorname{sinc}(\tfrac{1}{2}z)| = |\operatorname{sinc}(z)(1-\cos(\tfrac{1}{2}z))| \overset{(A.15)}{=} |\operatorname{sinc}(z)\tfrac{1}{4}z^2\operatorname{cosc}(\tfrac{1}{2}z)| \overset{(A.20)}{\leq} \tfrac{1}{8}z^2|\operatorname{sinc}(z)|$$

und mit c_3 aus (A.14)

$$|\operatorname{sinc}(z) - \tfrac{1}{2}(\cos(z)+1)\operatorname{sinc}^2(\tfrac{1}{2}z)| = |\operatorname{sinc}(z)(1-\operatorname{sinc}(z))| \overset{(A.16)}{=} |\operatorname{sinc}(z)z^2\,\mathrm{c}_3(z)| \overset{(A.20)}{\leq} |\tfrac{1}{6}\operatorname{sinc}(z)z^2|.$$

Damit erfüllen (4.110) und (4.111) die Abschätzung (4.57) mit den Konstanten $C_5 = \frac{1}{8}$ bzw. $C_5 = \frac{1}{6}$. In den Abbildungen 4.1 bzw. 4.2 sind beide Seiten der Bedingung nach Multiplikation mit z gezeichnet.

Für $z \neq 0$ folgt aus

$$\operatorname{sinc}(z) - \operatorname{sinc}(\tfrac{1}{2}z) = \frac{(\cos(\frac{1}{2}z)-1)}{(\frac{1}{2}z)^2}(\tfrac{1}{2}z)^2\operatorname{sinc}(\tfrac{1}{2}z) \overset{(A.15),(A.9)}{=} \tfrac{1}{2}\operatorname{cosc}(\tfrac{1}{2}z)z\sin(\tfrac{1}{2}z),$$

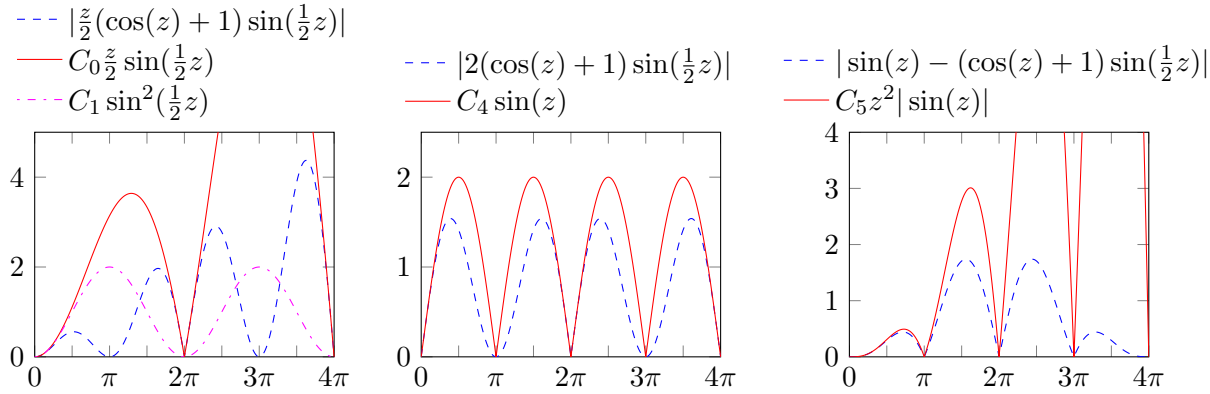
dass (4.70) mit $C_6 = \frac{1}{4}$ erfüllt ist, da $|\operatorname{cosc}| \leq \frac{1}{2}$. Im Ursprung steht auf beiden Seiten der Ungleichung Null. In der Abbildung 4.3 sind beide Seiten der Bedingung nach Multiplikation mit z gezeichnet.

Die Bedingung (4.92) ist aufgrund der Beschränktheit von sinc (A.20) mit $C_7 = 1$ erfüllt.

Für $z \neq 0$ folgt aus

$$\operatorname{sinc}^2(\tfrac{1}{2}z) - \operatorname{sinc}(z)\operatorname{sinc}(\tfrac{1}{2}z) = \operatorname{sinc}^2(\tfrac{1}{2}z)(1-\cos(\tfrac{1}{2}z)) = \sin^2(\tfrac{1}{2}z)\operatorname{cosc}(\tfrac{1}{2}z),$$

dass (4.93) mit $C_8 = \frac{1}{2}$ erfüllt ist, da $|\operatorname{cosc}| \leq \frac{1}{2}$. Im Ursprung steht auf beiden Seiten der Ungleichung Null. In der Abbildung 4.3 sind beide Seiten der Bedingung gezeichnet.



(4.36), erf. bzw (4.35a), verl. (4.42) (4.57)

Abbildung 4.1: Plots der Abschätzungen, die den ψ_E-Filter betreffen, für die Wahl $\psi_E = \operatorname{sinc}$ aus (4.110). Blau-gestrichelt ist jeweils die rechte Seite der Abschätzung, rot-durchgezogen die linke Seite der jeweiligen Ungleichung dargestellt. In der ersten Grafik ist zusätzlich noch die rechte Seite der verletzten starken Bedingung in magenta-strich-punkt eingezeichnet. Die Konstanten sind als $C_0 = 2$, $C_1 = 2$, $C_4 = 2$ und $C_5 = \frac{1}{8}$ gewählt.

Damit haben wir gezeigt:

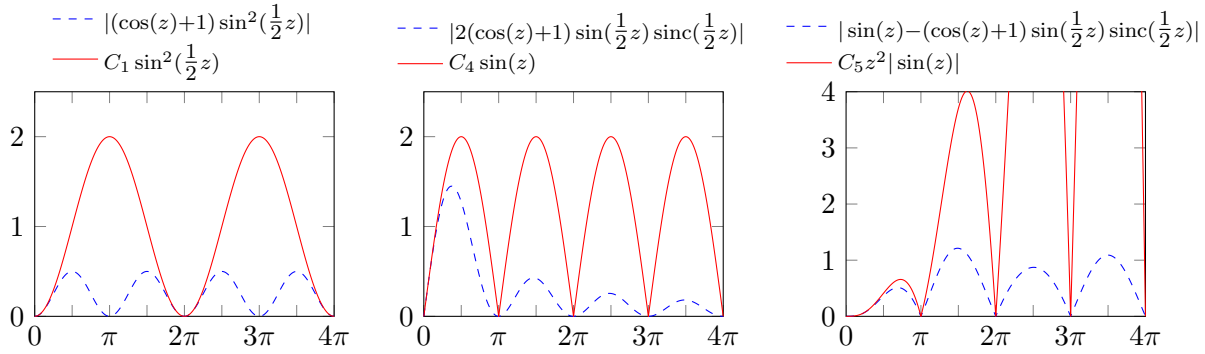
Satz 4.18 (Wahl der Filterfunktionen)**.** *Für die Wahl* (4.110) *gilt, dass die Filterfunktionen* $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ *gerade, reellwertige Funktionen mit* $\psi_E(0) = \phi_E(0) = 1$ *sind, die die Bedingungen* (4.36) *mit* $C_0 = 2$, (4.35b) *mit* $C_2 = 1$, (4.35c) *mit* $C_3 = 2$, (4.42) *mit* $C_4 = 2$, (4.57) *mit* $C_5 = \frac{1}{6}$, (4.70) *mit* $C_6 = \frac{1}{4}$, (4.92) *mit* $C_7 = 1$ *und* (4.93) *mit* $C_8 = \frac{1}{2}$ *erfüllen. Die beiden übrigen Filterfunktionen sind als* $\psi_B \equiv \phi_B \equiv 1$ *gewählt.*

Für die Wahl (4.111) *gilt, dass die Filterfunktionen* $\psi_E, \phi_E : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ *gerade, reellwertige Funktionen mit* $\psi_E(0) = \phi_E(0) = 1$ *sind, die die Bedingungen* (4.35a) *mit* $C_1 = 2$, (4.35b) *mit* $C_2 = 1$, (4.35c) *mit* $C_3 = 2$, (4.42) *mit* $C_4 = 2$, (4.57) *mit* $C_5 = \frac{1}{6}$, (4.70) *mit* $C_6 = \frac{1}{4}$, (4.92) *mit* $C_7 = 1$ *und* (4.93) *mit* $C_8 = \frac{1}{2}$ *erfüllen. Die beiden übrigen Filterfunktionen sind als* $\psi_B \equiv \phi_B \equiv 1$ *gewählt.* □

Nun können wir die Hauptresultate dieses Kapitels formulieren. Zuerst formulieren wir die Konvergenzaussage für das Verfahren aus [89], welches die Wahl (4.110) verwendet.

Theorem 4.19 (Fehlerabschätzung für das Splitting-Verfahren mit ursprünglicher Wahl der Filterfunktionen)**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen* (3.1) *mit dem numerischen Schema* (4.1)*. Die analytische Lösung werde mit* $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ *bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2*. Für die Zeitschrittweite* τ *gelte* $\tau \leq \tau_0$, τ_0 *unabhängig von* $\widetilde{\omega}$, *wobei* $\tau\widetilde{\omega} \geq c_0 > 0$*. Die Filterfunktionen seien als* (4.110) *gewählt. Die Startwerte seien* $\vec{E}_0$, $\vec{B}_0$



(4.35a) (4.42) (4.57)

Abbildung 4.2: Plots der Abschätzungen, die den ψ_E-Filter betreffen, für die Wahl $\psi_E = \operatorname{sinc}^2$ aus (4.111). Blau-gestrichelt ist jeweils die rechte Seite der Abschätzung, rot-durchgezogen die linke Seite der jeweiligen Ungleichung dargestellt. Die Konstanten sind als $C_1 = 2$, $C_4 = 2$ und $C_5 = \frac{1}{6}$ gewählt.

und $\vec{p}_0$ und mögen (4.43), (4.46), (4.47) und (4.58) erfüllen. Die auftretenden Konstanten C_c und H_0 seien unabhängig von $\widetilde{\omega}$.

Dann erhalten wir die Abschätzung

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau, \quad \|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau, \quad \|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau, \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.113}$$

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T - t_0)$ sowie den Konstanten C_c aus (4.22) und H_0. □

Mit der neuen Wahl der Filterfunktionen aus (4.111) erhalten wir das folgende verbesserte Konvergenzresultat.

Theorem 4.20 (Fehlerabschätzung für das Splitting-Verfahren mit Filterfunktionen aus (4.111))**.** *Wir betrachten die numerische Lösung der ortsdiskretisierten Maxwell-Gleichungen (3.1) mit dem numerischen Schema (4.1). Die analytische Lösung werde mit $\vec{p}(t), \vec{E}(t), \vec{B}(t)$ bezeichnet.*

Die Ortsdiskretisierung erfülle Voraussetzung 4.2. Für die Zeitschrittweite τ gelte $\tau \leq \tau_0$, τ_0 unabhängig von $\widetilde{\omega}$, wobei $\tau\widetilde{\omega} \geq c_0 > 0$. Die Filterfunktionen seien als (4.111) gewählt. Die Startwerte seien $\vec{E}_0$, $\vec{B}_0$ und $\vec{p}_0$ und mögen (4.43), (4.46), (4.47) und (4.58) erfüllen. Die auftretenden Konstanten C_c und H_0 seien unabhängig von $\widetilde{\omega}$.

Dann erhalten wir die Abschätzung

$$\|\vec{p}_n - \vec{p}(t_n)\| \leq C\tau^2, \quad \|\vec{E}_n - \vec{E}(t_n)\| \leq C\tau^2, \quad \|\vec{B}_n - \vec{B}(t_n)\| \leq C\tau^2, \qquad \text{für} \qquad t_n := t_0 + n\tau \leq T. \tag{4.114}$$

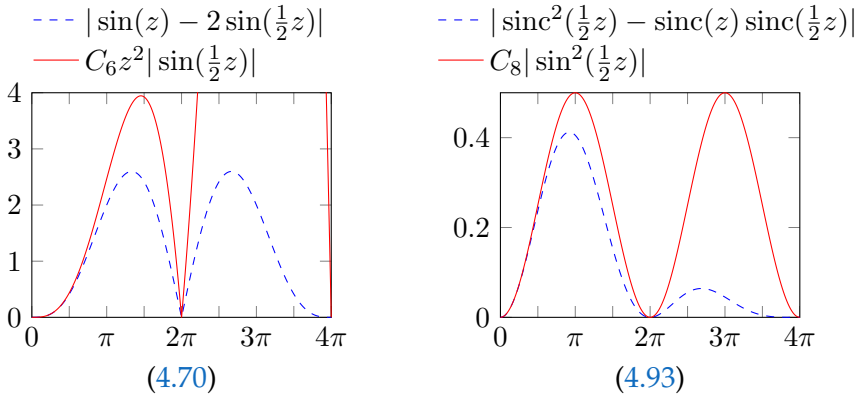


Abbildung 4.3: Plots der Abschätzungen, die den ϕ_E-Filter betreffen, für die Wahl $\psi_E = \operatorname{sinc}$. Blau-gestrichelt ist jeweils die rechte Seite der Abschätzung, rot-durchgezogen die linke Seite der jeweiligen Ungleichung dargestellt. Die Konstanten sind als $C_6 = \frac{1}{4}$ und $C_8 = \frac{1}{2}$ gewählt.

Die Konstante C hängt dabei nicht von $\widetilde{\omega}$, τ, n oder Ableitungen der Lösung ab, dafür aber von der Länge des Zeitintervalls $(T - t_0)$ sowie den Konstanten C_c aus (4.22) und H_0. □

Bemerkung 4.21. In unserer Analyse stellen wir zwei Voraussetzungen an die Frequenzmatrix $\mathbf{\Omega}$. Die eine ist die Diagonalstruktur, die, wie auch schon in [34] erwähnt, nicht wesentlich ist, solange $\mathbf{\Omega}$ unitär diagonalisierbar ist. Dann kann man ohne Einschränkungen zu einer Basis aus Eigenvektoren übergehen, ohne die Struktur der Nichtlinearität g zu beeinflussen.

Wesentlich ist hingegen die Blockstruktur. Diese gibt vor, dass $\mathbf{\Omega}$ nur die beiden Eigenwerte Null und $\widetilde{\omega}$ besitzen darf, wobei die letztere die Oszillationsfrequenz der Lösung beschreibt. Wenn die Matrix mehrere (hohe) Frequenzen besitzt, verkompliziert sich die Theorie, vgl. etwa [34, XIII.9]. Eine Konvergenztheorie für diesen Fall ist bisher unbekannt.

4.13 Numerische Tests

Wir haben jetzt eine theoretische Vorhersage über die Konvergenz des Splitting-Verfahrens (4.1) zur Verfügung. Im Folgenden illustrieren wir diese Vorhersage durch ein numerisches Beispiel. Dazu untersuchen wir wieder die Situation aus Abschnitt 3.3.

Durch Diskretisierung werden aus den unbeschränkten Differentialoperatoren Matrizen mit endlicher Norm der Form

$$\mathbf{C}_B = \tfrac{1}{h}(\operatorname{diag}(0,-1,1) + \vec{e_n}\vec{e_1}^T), \qquad \mathbf{C}_E = \tfrac{1}{h}(\operatorname{diag}(-1,1,0) - \vec{e_1}\vec{e_n}^T).$$

Dabei ist $\mathbf{C}_B$ die Diskretisierung des ∂_x-Operators in der Gleichung für die Änderung des $\vec{E}$-

Feldes, $\mathbf{C}_E$ diejenige aus der Gleichung für die Änderung des $\vec{B}$-Feldes. Sie erfüllen

$$\|\mathbf{C}_B\vec{B}\| \leq \tfrac{2}{h}\|\vec{B}\|, \qquad \|\mathbf{C}_E\vec{E}\| \leq \tfrac{2}{h}\|\vec{E}\|,$$

womit wir (4.22) verifiziert haben. Weiterhin ergibt sich

$$\mathbf{G} = \frac{1}{h^2}(\mathrm{diag}(1,-2,1) + \vec{e_n}\vec{e_1}^T + \vec{e_1}\vec{e_n}^T),$$

womit $\mathbf{G}$ eine symmetrische und aufgrund der Diagonaldominanz und der negativen Diagonaleinträge eine negativ semidefinite Matrix ist, die also die Voraussetzung (4.21) erfüllt. Dadurch, dass wir den Dichteterm punktweise auf dem Gitter auswerten, wird $\mathbf{\Omega}$ zu einer Diagonalmatrix. Nach Umsortierung der Komponenten erhalten wir ohne Einschränkung die Form

$$\mathbf{\Omega} = \begin{bmatrix} 0 & 0 \\ 0 & c\sqrt{\rho_F}\,\mathbf{Id} \end{bmatrix},$$

wie sie in (4.19) gefordert ist.

Mit diesen Daten sind die ortsdiskretisierten Gleichungen durch (3.1) gegeben, wobei die Ortsdiskretisierung die Voraussetzung 4.2 erfüllt. Damit befinden wir uns in derselben Situaltion wie schon in Abschnitt 3.3.

Die Bedingungen an die Startwerte, die für die Konvergenz gefordert sind, sind durch

$$\|\mathbf{\Omega}\vec{E}_0\|^2 \leq \tfrac{2}{3}H_0, \qquad \|\mathbf{C}_B\vec{B}_0\|^2 \leq \tfrac{1}{3}(\tfrac{2}{C_4})^2 H_0, \qquad \|\mathbf{\Omega}^2\vec{p}_0\|^2 \leq \tfrac{1}{3}H_0, \tag{4.43}$$

$$-\langle \mathbf{G}\vec{E}_0, \vec{E}_0\rangle = \|\mathbf{C}_E\vec{E}_0\|^2 \leq 2H_0, \qquad \|\vec{E}_0\|^2 \leq H_0 \tag{4.46}$$

$$\|\vec{B}_0\|^2 \leq H_0 \tag{4.47}$$

und

$$\|\mathbf{\Omega}^2\mathbf{C}_B\vec{B}_0\|^2 \leq H_0 \tag{4.58}$$

gegeben, wobei H_0 unabhängig von $\widetilde{\omega} = \sqrt{\rho_F}$ sein muss. Die mittlere Bedingung aus (4.43) und die Bedingungen (4.46) und (4.47) sind dabei unkritisch, da die Startwerte unabhängig von $\widetilde{\omega}$ sind. Die Forderung, dass die Größen in den anderen Bedingungen nach den Multiplikationen mit $\mathbf{\Omega}$ noch beschränkt sein sollten, bedeutet, dass die Felder in den Bereichen, die zum unteren Block der Matrix $\mathbf{\Omega}$ gehören, entsprechend der Potenz von $\mathbf{\Omega}$ klein sein müssen.

Analytisch ist dies für den Laserpuls aus (2.11) nicht erfüllbar, da der exp-Term zwar sehr schnell, aber nicht von $\widetilde{\omega}$ abhängig klein wird. Für unsere praktischen Experimente wählen wir das Zentrum x_0 und die Breite σ_0 des Pulses, sodass die Werte sehr klein sind, und setzen diese auf Null. Dabei ist der entstandene Fehler so klein, dass er unterhalb der Rechengenauigkeit liegt, also im

numerischen Rauschen untergeht. Obige Wahlen für x_0 und σ_0 ermöglichen dies.

Damit sind alle Bedingungen der Konvergenzaussagen aus den Theoremen 4.19 und 4.20 erfüllt. Wir führen die Experimente aus 3.3 noch einmal, diesmal mit den beiden gefilterten Verfahren (4.110) und (4.111), durch. Das Ergebnis ist in Abbildung 4.4 dargestellt.

Wir können erkennen, dass beide gefilterten Varianten des Verfahrens die Situation schon deutlich verbessern. Im Gegensatz zur ungefilterten Version, die große Bereiche hat, in denen sie oberhalb der Ordnung-2-Kurve liegt, ist die Gefahr, eine ungünstige Schrittweite zu treffen, bei den beiden anderen sehr klein bzw. gar nicht vorhanden. Für die hohe Dichte scheint auch für (4.110) eine perfekte Ordnung-2-Kurve zu entstehen. Dies war vermutlich der Grund, dass das eigentlich schlechtere Konvergenzverhalten des Verfahrens in [89] übersehen wurde. Im Bild der niedrigeren Dichte sind die Spitzen im $\vec{E}$-Feld aber deutlich zu sehen.

Um erkennen zu können, dass auch die anderen beiden Felder – wenn auch kleinere – Ausschläge haben, und auch, um noch einmal etwas genauer zu überprüfen, dass das neue Verfahren keine Ausreißer aufweist, haben wir den Bereich um den größten Ausreißer des Verfahrens mit der Filterwahl aus (4.110) in Abbildung 4.5 noch einmal vergrößert dargestellt. Hier sind tatsächlich alle möglichen Schrittweitenwahlen in diesem Bereich, die man überhaupt treffen kann, aufgetragen, wenn man für den letzten Zeitschritt die Schrittweite nicht anpassen will. In dieser Situation ergibt sich, $n\tau = T - t_0$, also $\tau = \frac{T-t_0}{n}$, $n \in \mathbb{N}$. Für $\tau_0 \leq \tau \leq \tau_1$ erfüllten dies also nur endlich viele τ.

Diesmal haben wir die Kurve für das Verfahren (4.110) als dünne Kurve dargestellt, um die Oszillationen der Fehlerkurve besser erkennen zu können. Der Fehler dieses Verfahrens scheint sogar nur unterhalb einer Konstante zu liegen und nicht, wie die Theorie vorhersagt, unterhalb einer Kurve mit Steigung eins. Dies ist vermutlich ein numerisches Artefakt, bei dem die Schrittweiten das eigentliche Maximum der Oszillationen nicht treffen.

Unsere Theorie beschränkt sich durch die Strukturannahme an die Matrix $\mathbf{\Omega}$ sehr, was die Freiheit der Wahl für das Dichteprofil betrifft. In der Matrix ist nur eine einzige hohe Frequenz zugelassen. Der Hauptgrund dafür ist, dass wir die Abschätzung des Fehlers im elektrischen Feld auf Theorem 4.1 stützen. Der Rest der hier gezeigten Theorie lässt sich mit kleinen Änderungen auch etwa für symmetrisches, positiv-semidefinites $\mathbf{\Omega}$ zeigen.

In [46, 26] gibt es eine Fehlerabschätzung für ein Mehrschrittverfahren der Form (4.12), die genau diese Eigenschaft der Matrix $\mathbf{\Omega}$ fordert. Allerdings ist die Wahl der Filterfunktion ψ in diesem Fall auf $\operatorname{sinc}^2(\frac{1}{2}z)$ eingeschränkt, was wegen des zusätzlichen $\frac{1}{2}(\cos(z)+1)$-Faktors in (4.33) für unser Verfahren nicht passt. Außerdem wird von der Nichtlinearität g globale Beschränktheit gefordert, was in unserem Fall aufgrund der Linearität ebenfalls nicht gewährleistet werden kann. Um das erste dieser beiden Probleme zu beheben, müsste der Beweis in [46] genauer untersucht und angepasst werden. Dies ist vermutlich möglich. Das zweite Problem lässt sich augenscheinlich ohne Änderung des Beweises in „g muss innerhalb eines Streifens entlang der exakten Lösung

beschränkt bleiben“ abschwächen. Auch dies müsste genauer untersucht werden. Diese Analyse wäre sehr interessant, geht aber über den Rahmen dieser Arbeit hinaus. Weiterhin wird von der Filterfunktion ϕ gefordert, dass sie Nullstellen bei allen ganzzahligen Vielfachen von π außer dem Ursprung hat, wodurch wir die Frequenz des Filters $\phi_E = \operatorname{sinc}(2z)$ im Vergleich zu (4.111) verdoppeln. Der zweite Startwert muss, wie ebenfalls in Theorem 4.1, auch (4.13) genügen, was wiederum ein Vorgehen wie in den Abschnitten 4.5, 4.6 und 4.7 erforderlich macht.

Wir führen einen einfachen numerischen Test mit der Situation wie am Anfang dieses Abschnittes, aber einem Dichteprofil der Form

$$\rho(x) = \begin{cases} \rho_F, & \text{falls } x \in F, \\ \frac{\rho_F}{2}(1+\cos((1+\frac{x-x_l}{|G_l|})\pi), & \text{falls } x \in F_l, \\ \frac{\rho_F}{2}(1+\cos((1+\frac{x_r-x}{|G_r|})\pi), & \text{falls } x \in F_r, \\ 0, & \text{sonst} \end{cases} \tag{4.115}$$

durch, wobei wir die Dichte nun glatter in den Randstücken F_l und F_r steigen und abfallen lassen. Eine Grafik dazu findet sich in Abbildung 4.6. Die Fehlerkurven, die sich in diesem Fall ergeben, werden in Abbildung 4.7 dargestellt. Wir haben dabei sowohl die Filter aus (4.111) als auch die durch [46, 26] motivierte Wahl

$$\psi_E(z) = \operatorname{sinc}^2(z), \qquad \phi_E(z) = \operatorname{sinc}(2z) \qquad \text{und} \qquad \psi_B \equiv \phi_B \equiv 1 \tag{4.116}$$

getestet.

Das Verfahren zeigt für beide Wahlen der Filterfunktionen numerisch die Ordnung zwei. Die durch die Theorie aus [46, 26] motivierte stärkere Filterung scheint also nicht nötig zu sein oder unser spezielles Problem hat nicht die Eigenschaft, den entsprechenden Resonanzfall auszulösen oder die Auflösung der Schrittweiten reicht nicht aus. In jedem Fall wäre es in zukünftigen Forschungen interessant zu ermitteln, um welchen dieser Fälle es sich handelt.

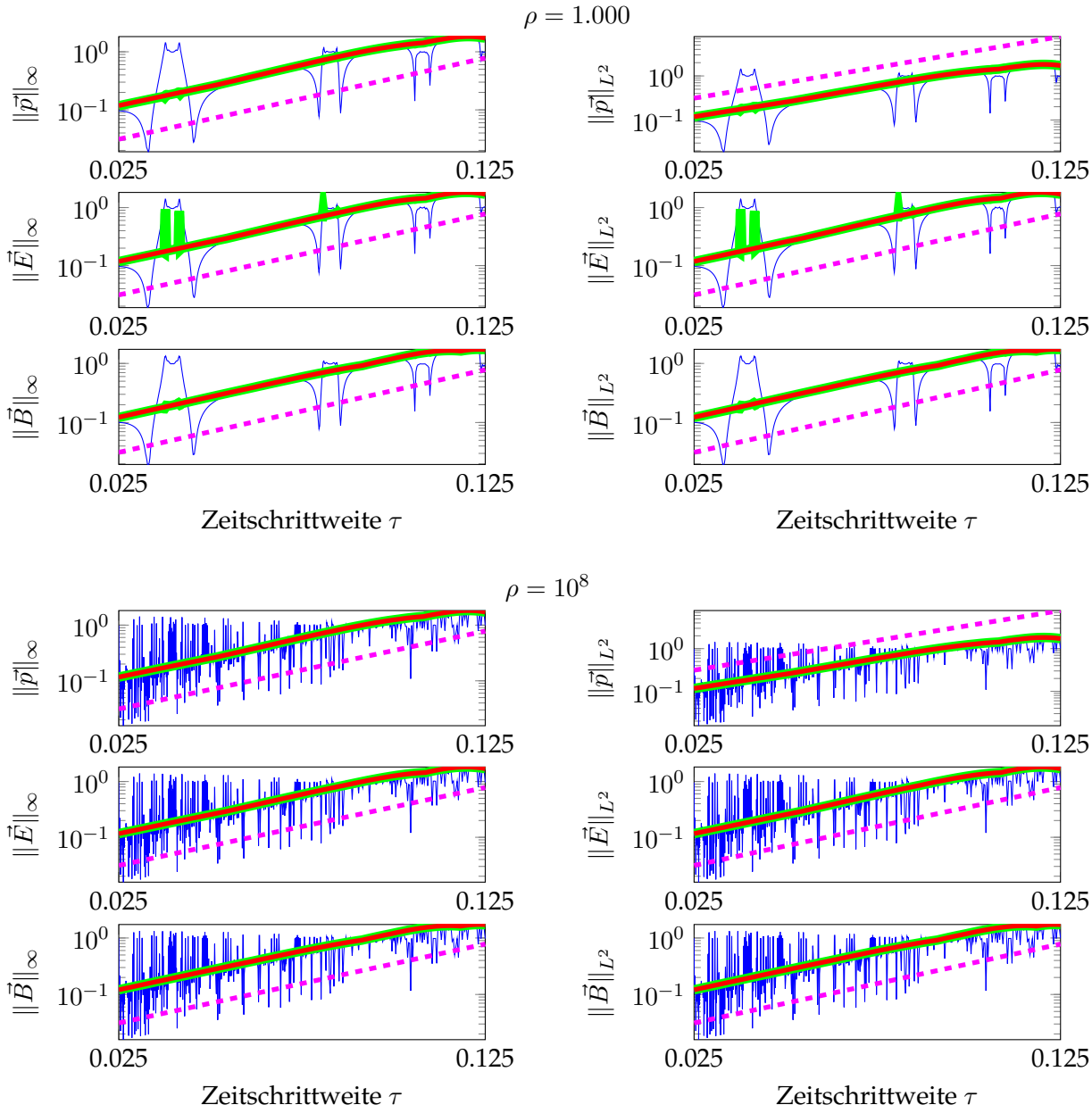


Abbildung 4.4: Darstellung des absoluten Fehlers der Impulse und der elektrischen und magnetischen Felder in Relation zur Zeitschrittweite τ. Die linken Abbildungen zeigen jeweils den Fehler in der ∞-Norm, die rechten jeweils in der L_2-Norm. Als Referenz ist gestrichelt eine Ordnung-zwei-Kurve dargestellt. Blau, dünn: das ungefilterte Verfahren (3.6); grün, dick: das mit (4.110) gefilterte Verfahren; rot, mittel: das mit (4.111) gefilterte Verfahren. Erste Zeile: Verhalten des Fehlers für eine Dichte 1.000 in der Dichtewand, zweite Zeile: Verhalten des Fehlers für eine Dichte 10^8 in der Dichtewand.

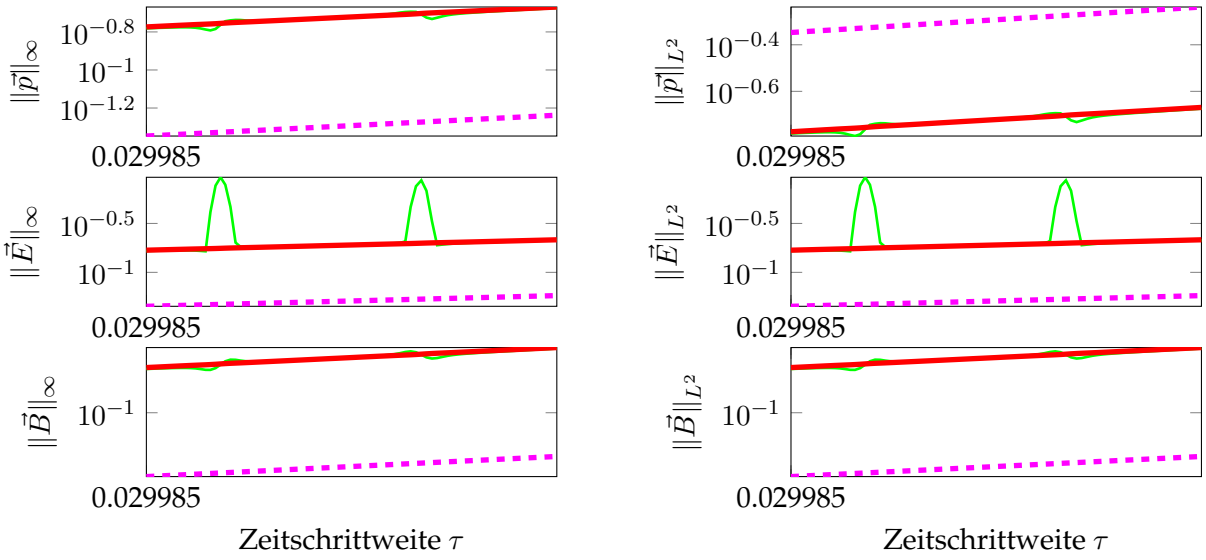


Abbildung 4.5: Vergrößerter Ausschnitt von Abbildung 4.4 ohne das ungefilterte Verfahren (3.6) für $\tau \in [0.12, 0.135]$ und nur für die niedrigere Dichte $\rho = 1.000$. Grün, dünn: das mit (4.110) gefilterte Verfahren; rot, mittel: das mit (4.111) gefilterte Verfahren.

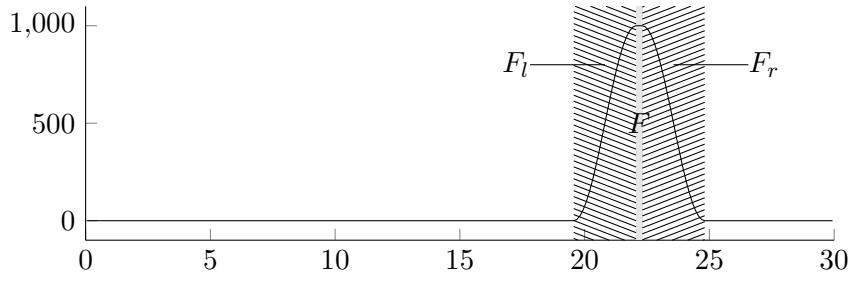


Abbildung 4.6: Dichteprofil der Form (4.115).

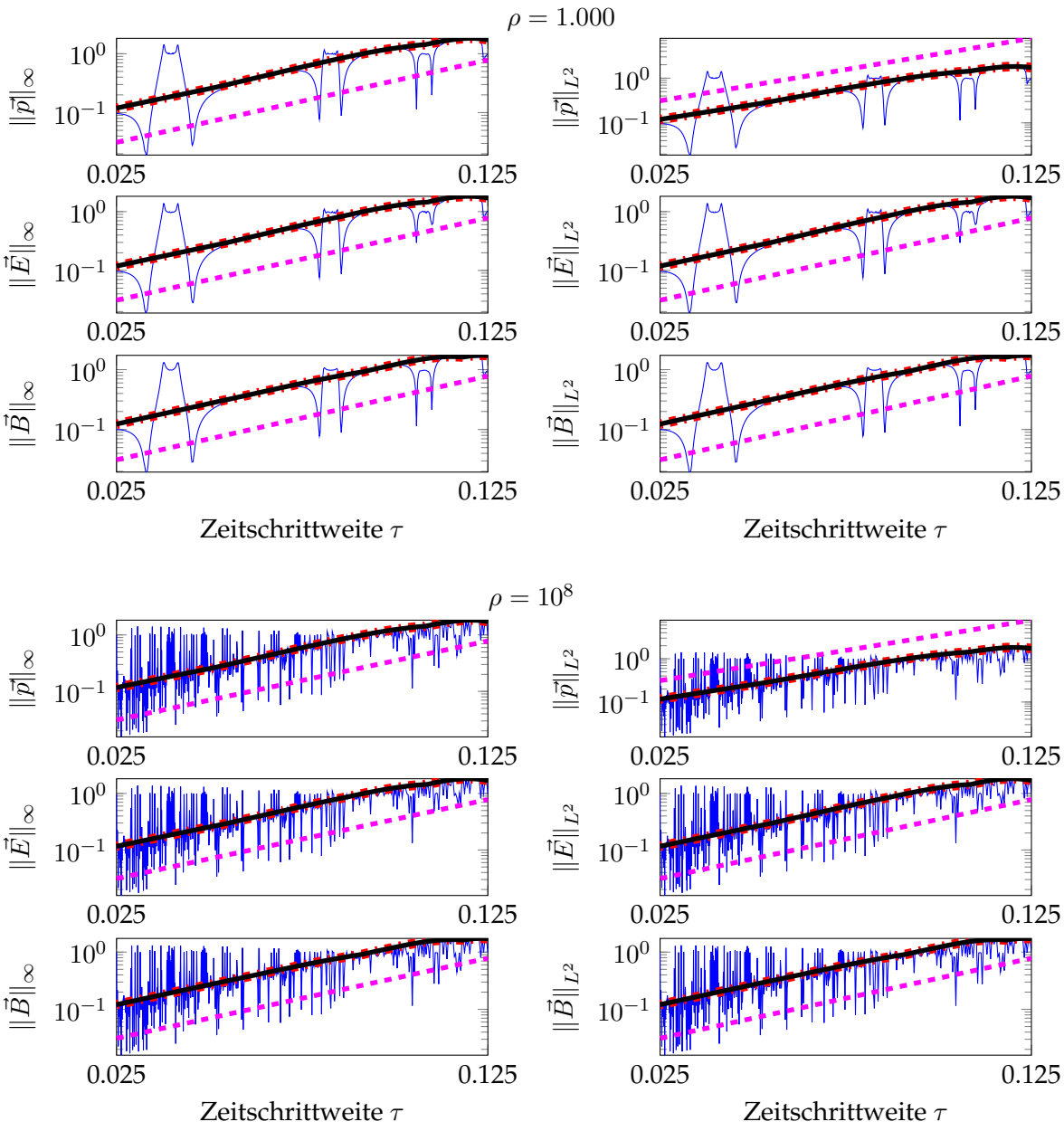


Abbildung 4.7: Darstellung des absoluten Fehlers der Impulse und der elektrischen und magnetischen Felder wie in Abbildung 4.4 für ein Dichteprofil der Form (4.115). Blau durchgezogen: ungefiltertes Verfahren; rot gestrichelt, dick: das mit (4.111) gefilterte Verfahren; schwarz, dünn: das mit (4.116) gefilterte Verfahren.

KAPITEL 5

MASSE-KRYLOV-VERFAHREN FÜR SEKTORIELLE MATRIXFUNKTIONEN

In diesem Kapitel wenden wir unseren Fokus von den hochdichten Plasmen ab und starten mit einem allgemeineren Ansatz. Bisher haben wir unser numerisches Verfahren speziell für die gegebene Differentialgleichung konstruiert. Der exponentielle Integrator kam durch die exakte Lösung der Gleichung (3.4) mit der rechten Seite (3.3c) ins Spiel.

Jetzt suchen wir approximative Lösungen von allgemeinen abstrakten gewöhnlichen Differentialgleichungen der Form:

$$\text{Finde } u \in C^1([t_0, T], \mathcal{X}), \text{ sodass} \qquad \begin{cases} \partial_t u(t) = Au(t) + g(t, u(t)), & t_0 \le t \le T, \\ u(t_0) = u_0 \end{cases} \tag{5.1}$$

in einem Hilbert-Raum $(\mathcal{X}, \langle \cdot, \cdot \rangle)$. Dabei sei $A : D(A) \longrightarrow \mathcal{X}$ ein linearer, abgeschlossener, dicht definierter Operator auf $\mathcal{X}$, der eine analytische Halbgruppe $\exp(\cdot A) : \mathbb{R}_{\ge 0} \longrightarrow \mathcal{X}$ auf $\mathcal{X}$ erzeugt und $g : \mathbb{R}_{\ge 0} \times \mathcal{X} \longrightarrow \mathcal{X}$ hinreichend glatt und integrierbar. Theorie und Voraussetzungen zur Lösbarkeit solcher Gleichungen für den linearen Fall finden sich zum Beispiel in [71, 16]. Vom Operator A werden wir verlangen, dass er zu einer passenden Bilinearform a auf einem Unterraum $D(A) \subset V \subset \mathcal{X}$ assoziiert sei.

Zunächst wollen wir eine allgemeine Lösungsformel für (5.1) angeben, die als Basis für den numerischen Ansatz zur Zeitdiskretisierung dient. Im Weiteren gehen wir dann auf eine Diskretisierungsmöglichkeit für den Operator A, also die Ortsdiskretisierung, ein. Die daraus hervorgehende gewöhnliche Differentialgleichung ist semilinear, wobei vor der Zeitableitung eine Matrix, die so genannte Massematrix auftritt. Exponentielle Integratoren müssen in jedem Zeitschritt mehrere lineare Gleichungssysteme mit dieser Matrix lösen. Ziel dieses Kapitels ist es, Verfahren zu entwickeln, bei denen auf ebendieses Lösen verzichtet werden kann.

Wir werden zwei Möglichkeiten vorstellen, die das gesetzte Ziel erreichen können. Dabei löst der erste Ansatz zwar das Problem, es zeigt sich jedoch in numerischen Experimenten, dass das Verfahren schlechte Konvergenzeigenschaften besitzt. Ein kurzer Blick auf die Analyse verwandter Verfahren lässt auch den Grund vermuten. Eine Behebung des Problems liefert der zweite Ansatz. Von diesem können wir auch theoretisch zeigen, dass er konvergiert.

Der Aufbau dieses Kapitels ist wie folgt: Nach der Einführung der allgemeinen Lösungsformel im ersten Abschnitt stellen wir darauf aufbauende exponentielle Runge-Kutta-Verfahren als Repräsentanten allgemeiner exponentieller Integratoren in kurzer Form vor. Der wichtigste Baustein exponentieller Verfahren ist die Anwendung der auftretenden Auswertungen von Matrixfunktionen auf Vektoren. Im zweiten Abschnitt gehen wir auf eine Ortsdiskretisierung mit finiten Ele-

menten ein und stellen Eigenschaften der entstehenden Matrizen dar. Die Matrixfunktionen der Integratoren müssen auf diesen Matrizen ausgewertet werden.

Die Abschnitte 5.3 bis 5.6 sind der Herleitung von Krylov-Verfahren für Matrixfunktionen gewidmet. Der erste dieser Abschnitte stellt die ursprüngliche Variante für die Lösung linearer Gleichungssysteme dar. Um Krylov-Verfahren für Matrixfunktionen herleiten zu können, benötigen wir für letztere noch eine Darstellungsformel über ein Kurvenintegral in der komplexen Ebene. Üblicherweise wird hier die Cauchy-Integralformel verwendet, wir benutzen in Abschnitt 5.4 jedoch die Laplace-Transformation, die für spätere Zwecke dienlicher ist. Mithilfe dieser approximieren wir im folgenden Abschnitt die Anwendung von Matrixfunktionen auf Vektoren und im letzten erklären wir, wie man dabei die Struktur der ursprünglichen Gleichung in das Verfahren integrieren kann.

Das erste Krylov-Verfahren für Matrixfunktionen, welches ohne die Lösung von Gleichungssystemen mit der Massematrix auskommt, kurz masselösungsfreies Krylov-Verfahren, folgt in Abschnitt 5.7. Außerdem diskutieren wir ein einfaches Beispiel, das uns die Schwächen dieser Variante aufzeigt. Eine kurze Betrachtung des Fehlers gibt uns Indizien für die Ursache. Um diese zu umgehen, stellen wir in den Abschnitten 5.8 und 5.9 die Approximation der Matrixfunktionen mit numerischer inverser Laplace-Transformation vor. Diese nutzen wir im letzten Abschnitt in Kombination mit Krylov-Verfahren, um unser vollständiges Verfahren zu entwickeln. Wir schließen damit ab, auch für diesen Ansatz numerische Versuche durchzuführen.

5.1 Variation-der-Konstanten-Formel und exponentielle Integratoren

Die allgemeine Lösung von (5.1) liefert die *Variation-der-Konstanten-Formel*

$$u(t) = \exp((t-t_0)A)u_0 + \int_{t_0}^{t} \exp((t-\xi)A)g(\xi, u(\xi))\,\mathrm{d}\xi, \qquad t_0 \le t \le T. \tag{5.2}$$

Das Integral ist als Bochner-Integral zu verstehen. Dabei sei darauf hingewiesen, dass diese Formel einen allgemeineren Lösungsbegriff liefert als (5.1), denn jede Lösung von (5.1) lässt sich durch die Variation-der-Konstanten-Formel ausdrücken, aber letztere liefert auch Lösungen, falls $u_0 \notin D(A)$. In diesem Fall redet man von *milden Lösungen*, vgl. etwa [16, VI.7] für den Fall, dass g nur von der Zeit t abhängt.

Diese Formel ist in verschiedenen Versionen unter vielen Namen bekannt, etwa die *Variation-der-Parameter-Formel*, das *Duhamel-Prinzip* oder die *Alekseev-Gröbner-Formel*. Dabei werden unterschiedliche Voraussetzungen gemacht. Gemeinsam haben alle offensichtlich die eindeutige Lösbarkeit der linearen Gleichung mit $g = 0$. Varianten sind etwa $g(t,u) = Bu$, B linear beschränkt, vgl. [16, III.3], oder $g(t,u) = g(t)$ nur zeitabhängig, vgl. [16, IV.7]. Für den endlichdimensionalen

Fall lässt sich die Formel einfach verifizieren, indem man für

$$\varphi(\xi) := \exp((t-\xi)A)u(\xi) \tag{5.3}$$

den Hauptsatz der Differential- und Integralrechnung anwendet, vgl. etwa [54].

Numerische Verfahren, die auf dieser Formel aufbauen, versuchen nun einerseits die Anwendung der Halbgruppe auf Elemente von $\mathcal{X}$, andererseits das Integral zu approximieren. Die gängigste Variante für zweiteres ist die Näherung mit einer Quadraturformel. Dieser Ansatz liefert so genannte *exponentielle Runge-Kutta-Verfahren* [49, 48]. Wir wählen die Form aus [56]:

$$U_{ni} = u_n + c_i\tau_n\varphi_1(c_i\tau_n A)F(t_n,u_n) + \tau_n\sum_{j=1}^{s} a_{ij}(\tau_n A)D_{nj}, \qquad 1 \le i \le s \tag{5.4a}$$

$$u_{n+1} = u_n + \tau_n\varphi_1(\tau_n A)F(t_n,u_n) + \tau_n\sum_{i=1}^{s} b_i(\tau_n A)D_{ni}, \tag{5.4b}$$

wobei $t_{n+1} = t_n + \tau_n$ und

$$D_{ni} = g(t_n + c_i\tau_n, U_{ni}) - g(t_n,u_n), \quad 1 \le i \le s, \qquad F(t,u) = Au + g(t,u). \tag{5.5}$$

Dabei sind die Koeffizientenfunktionen b_j Linearkombinationen der zu Anfang in (A.1) definierten ganzen Funktionen φ_k und a_{ij} von $\varphi_k(c_i\,\cdot\,)$. Diese Funktionen, und damit auch b_i und a_{ij}, können genauso, wie die Halbgruppe selbst, auf Elemente von $\mathcal{X}$ angewendet werden.

Es gibt noch viele weitere Ansätze, die alle diese gemeinsame Basis haben, auf die wir hier aber nicht näher eingehen wollen. Einen sehr guten Überblick liefert [50]. Die obige Wahl von exponentiellen Runge-Kutta-Verfahren ist für uns nicht wesentlich, da sich unsere Konstruktionen direkt auf andere exponentielle Integratoren verallgemeinern lassen.

Die Auswertung der Exponential- und φ_k-Funktionen auf Operatoren bezeichnen wir als *Operatorfunktionen*. Da wir numerisch rechnen wollen, werden wir in Abschnitt 5.2 die Gleichung (5.1) auch im Ort diskretisieren und erhalten dann aus den Operatoren Matrizen und sprechen dann von *Matrixfunktionen*. Für genau diese Anwendung der Matrixfunktionen auf Vektoren werden wir in diesem Kapitel neue Varianten vorschlagen.

Grundsätzlich lassen sich die Auswertungsmethoden von Matrixfunktionen in zwei Kategorien einteilen. Die eine bilden so genannte direkte Methoden, bei denen die Matrixfunktion selbst ausgewertet wird. Die Anwendung auf einen Vektor ist dann ein einfaches Matrix-Vektor-Produkt. Zu diesen direkten Methoden zählen etwa abgeschnittene Taylor-Entwicklungen, Padé-Approximationen, vgl. etwa [90] für solche an die Exponentialfunktion, Scaling & Squaring-Methoden [39], der Schur-Parlett-Algorithmus [40, Kapitel 9 und 10], Kontur-Integrationsmethoden [80, 64], feste polynomielle Approximationen wie Tschebycheff-Polynome [62, Kapitel III.2] oder Interpo-

lationsverfahren wie Leja-Punkt-Methoden [10]. Wir haben hier nur einige der möglichen Verfahren aufgezählt und davon auch nur wenige Repräsentanten referenziert. Eine vollständige Liste überschreitet den Rahmen dieser Arbeit. Einen guten Überblick über moderne direkte Methoden liefern [40, 41].

Die Matrizen aus typischen Diskretisierungen für partielle Differentialgleichungen sind, wie im nächsten Abschnitt motiviert, groß und dünn besetzt. Beim Einsatz obiger direkter Methoden treten dabei zwei Probleme auf. Einerseits sind Matrixfunktionen dünnbesetzter Matrizen in der Regel voll besetzt. Der Speicherbedarf für die Matrixfunktionen ist also um ein Vielfaches größer als derjenige für die diskretisierten Operatoren. Andererseits müssen viele der genannten Algorithmen etwa auch auf Matrix-Matrix-Produkte zurückgreifen, lineare Gleichungssysteme der vollen Dimension lösen oder andere Operationen auf den vollen Matrizen ausführen. Deren Laufzeit ist quadratisch oder kubisch in der Dimension der Matrizen, was für hochdimensionale Systeme äußerst teuer ist.

Die zweite Kategorie von Verfahren nutzt aus, dass die Matrixfunktion jeweils auf Vektoren angewendet wird, und vermeidet dabei, dass volle Matrizen aufgestellt werden müssen. Der Hauptaufwand wird dabei zumeist durch vielfache Matrix-Vektor-Produkte verursacht. Bei dünn besetzten Matrizen haben diese meist keine quadratische Laufzeit in der Dimension der Matrix und es müssen keine vollen Matrizen abspeichert werden. Es gibt in [2] etwa eine Variante der Scaling & Squaring-Methode, die auf das explizite Auswerten der vollen Matrix verzichtet. Den bekanntesten und mit Abstand am weitest verbreiteten Vertreter dieser Kategorie bilden die Krylov-Verfahren, vgl. etwa [76, 45]. Der erste in der Praxis einsetzbare exponentielle Integrator, der auf der zweitgenannten Arbeit aufbaut, ist der so genannte `exp4` aus [47]. Unsere Ansätze bauen dabei auf der speziellen Struktur der entstehenden gewöhnlichen Differentialgleichung auf, die aus der örtlichen Diskretisierung der abstrakten Formulierung (5.1) hervorgeht.

5.2 Schwache Formulierung, Methode der finiten Elemente

Wir betrachten den Fall, dass der Raum $\mathcal{X}$ ein Funktionenraum und A ein Differentialoperator auf diesem ist. Darunter fallen auch die Maxwell-Gleichungen (2.10). Für unseren Ansatz benötigen wir jedoch parabolische Gleichungen, bei denen die Anwendung der Halbgruppe die Lösung glättet.

Wir überführen (5.1) zunächst in ihre schwache Formulierung. Dazu bilden wir das Skalarprodukt mit einer passenden Testfunktion. Wir suchen $u \in C^1([t_0, T], V)$, sodass

$$\begin{cases} \langle \partial_t u(t), v \rangle = a(u, v) + \langle g(t, u(t)), v \rangle & \forall\, t_0 \leq t \leq T, \\ \langle u(t_0), v \rangle = \langle u_0, v \rangle \end{cases} \quad \forall\, v \in V, \tag{5.6}$$

mit einem passenden Unterraum $D(A) \subseteq V \subseteq \mathcal{X}$ gilt. Sei $j : V \longrightarrow \mathcal{X}$ eine lineare, beschränkte Abbildung mit dichtem Bild und die Sesquilinearform a dabei j-elliptisch und beschränkt. Wir setzen voraus, das A im Sinne von [5, Korollar 5.11], [4, Theorem 4.2] der zu (a, j) assoziierte Operator ist.

Für die Ortsdiskretisierung wählen wir einen geeigneten endlichdimensionalen Unterraum $V_h = \operatorname{span}(\phi_1, \ldots, \phi_N)$ von V, der über seine Basis $\{\phi_1, \ldots, \phi_N\}$ gegeben ist. Die Elemente der Basis bezeichnen wir als *Ansatzfunktionen* oder *shape functions*.

Wir verwenden für Elemente des großen Funktionenraumes $v \in V$ keinen Index, für Elemente des Unterraumes $v_h \in V_h$ den Index h und für die Koeffizientenvektoren $\vec{v}_h \in \mathbb{C}^N$ einen Vektorpfeil.

Mögliche Wahlen für V_h sind zum Beispiel die Funktionenräume zu finiten Elementen für parabolische Probleme [8, 7], zu unstetigen finiten Elementen [38] oder etwa zu Nédélec-Elementen für die Maxwell-Gleichungen [67, 68, 93]. Nédélec-Elemente sind darauf zugeschnitten, solche Unstetigkeiten zuzulassen.

Wichtig ist hier nur, dass die Basis $\{\phi_1, \ldots, \phi_N\}$ des diskreten Raumes der Ansatzfunktionen explizit zur Verfügung steht, bezüglich derer die Lösung u_h in diesem Unterraum dargestellt ist, etwa

$$u_h(t) = \sum_{i=1}^{N} u_i(t)\phi_i. \tag{5.7}$$

Um nicht zu viele Indizes zu verwenden, wird bei den Einträgen von $\vec{v}_h = (v_i)_{i=1}^N$ auf den Index h verzichtet. Wir identifizieren

$$v_h = \mathrm{id}(\vec{v}_h) := \sum_{i=1}^{N} v_i\phi_i \tag{5.8}$$

miteinander.

Wir verwenden also einen Ansatz der Trennung der Zeit- und Ortsvariablen, welcher in der Literatur als *Linienmethode*, engl. *method of lines* bekannt ist.

Indem wir den Such- und den Testraum in (5.6) von V auf V_h einschränken, erhalten wir das folgende $N = \dim(V_h)$-dimensionale System gewöhnlicher Differentialgleichungen erster Ordnung

$$\mathbf{M}\partial_t\vec{u}_h(t) = \mathbf{A}\vec{u}_h(t) + \vec{g}(t, \vec{u}_h(t)), \qquad t \geq t_0, \qquad \mathbf{M}\vec{u}_h(t_0) = \vec{l}(u_0). \tag{5.9}$$

Die beiden Matrizen

$$\mathbf{A} = (a(\phi_i, \phi_j))_{i,j=1}^N \qquad \text{und} \qquad \mathbf{M} = (\langle \phi_i, \phi_j \rangle)_{i,j=1}^N \tag{5.10}$$

sind die aus dem Differentialoperator A bzw. aus seiner zugehörigen Sesquilinearform a hervorgegangene *Steifigkeitsmatrix* $\mathbf{A}$ und die Gram-Matrix der Basis zum $\mathcal{X}$-Skalarprodukt, genannt

Massematrix $\mathbf{M}$. In

$$\vec{g} : [t_0, T] \times \mathbb{C}^N \longrightarrow \mathbb{C}^N, (t, \vec{v}_h) \longmapsto \vec{l}(g(t, \mathrm{id}(\vec{v}_h)))$$

gehen die Nichtlinearität g der kontinuierlichen Gleichung ein. Die lineare Abbildung $\vec{l}$ ist dabei gegeben als

$$\vec{l} : X \longrightarrow \mathbb{C}^N, v \longmapsto (\langle v, \phi_i \rangle)_{i=1}^N,$$

Es gilt $\vec{l}|_{V_h} = \mathrm{id}^{-1}$ mit id aus (5.8).

Das natürliche Skalarprodukt auf $\mathbb{C}^N$ ist dabei gegeben als

$$\langle \vec{u}_h, \vec{v}_h \rangle := \langle \vec{u}_h, \vec{v}_h \rangle_{\mathbf{M}}, \tag{5.11}$$

denn dann ist $\langle \vec{u}_h, \vec{v}_h \rangle = \langle \mathrm{id}(\vec{u}_h), \mathrm{id}(\vec{v}_h) \rangle_{\mathcal{X}}$. Bezüglich dieses Skalarproduktes wird id aus (5.8) zu einer Isometrie, wenn V_h mit der $\mathcal{X}$-Norm versehen wird.

Die Matrix $\mathbf{M}$ ist als Gram-Matrix zu einer Basis insbesondere invertierbar. Daher lässt sich (5.9) zu

$$\partial_t \vec{u}_h(t) = \mathbf{M}^{-1}\mathbf{A}\vec{u}_h(t) + \mathbf{M}^{-1}\vec{g}(t, \vec{u}_h(t)), \qquad t \geq t_0, \qquad \vec{u}_h(t_0) = \mathbf{M}^{-1}\vec{l}(u_0) \tag{5.12}$$

umformen.

Damit haben wir wieder eine Gleichung der Form (5.1), diesmal auf dem endlichdimensionalen Hilbert-Raum $\mathbb{C}^N$. Exponentielle Integratoren, wie etwa die exponentiellen Runge-Kutta-Verfahren (5.4), lassen sich also anwenden. Dabei ist zu beachten, dass das Auswerten der rechten Seite oder das Anwenden des Operators der Gleichung jeweils das Lösen eines linearen Gleichungssystems mit der Massematrix erfordert.

Bei Finite-Elemente-Methoden beginnen wir zur Konstruktion eines Ansatzraumes damit, das zum Differentialoperator A gehörige örtliche Gebiet in einfache Teilgebiete zu zerlegen, etwa Linienstücke in einer Raumdimension, Drei- oder Vierecke im zweidimensionalen Fall, Tetraeder oder verformte Quader in drei Raumdimensionen. Dann definieren wir die Ansatzfunktionen, z.B. als Polynome, stückweise auf diesen Teilgebieten und setzen sie dann zu einer globalen Funktion so zusammen, dass sie hinreichend glatt sind, um in V zu liegen. Dabei werden die lokalen Ansatzfunktionen auf einzelnen Elementen so gewählt, dass sie auf einem möglichst großen Teil des Gebietes verschwinden. Dadurch umfassen die Träger der globalen Ansatzfunktionen jeweils immer nur wenige oder sogar nur einzelne Teilgebiete. Alle Skalarprodukte aus (5.10), bei denen die beiden Ansatzfunktionen ϕ_i und ϕ_j keinen gemeinsamen Träger besitzen, verschwinden, wodurch die Steifigkeits- und Massematrizen $\mathbf{A}$ und $\mathbf{M}$ eine sehr dünne Besetzungsstruktur besitzen.

Um eine gute örtliche Auflösung zu erhalten, wollen wir eine feine Zerlegung des Gebietes, einen hochdimensionalen Ansatzraum auf den Teilgebieten oder eine Kombination aus beidem erlauben. Dadurch wird die Dimension N des globalen Ansatzraumes V_h sehr groß.

Je feiner die Ortsdiskretisierung wird, desto besser werden auch oszillatorische Eigenkomponen-

ten des Operators in der Matrix $\mathbf{A}$ aufgelöst. Dadurch wächst die Konditionszahl von $\mathbf{A}$ stark. Die Massematrix $\mathbf{M}$ hat als Darstellung der Norm des Hilbert-Raumes eine deutlich bessere Kondition. Diese kann bei hohem Polynomgrad oder unstrukturierten Gittern aber durchaus auch in der Größenordnung von einigen Zehnerpotenzen sein. Dies wirkt sich entsprechend auf die Lösung von Gleichungssystemen mit den beiden Matrizen aus.

Insgesamt werden (5.9) bzw. (5.12) zu einem großen semilinearen System gewöhnlicher Differentialgleichungen mit dünnbesetzten Matrizen $\mathbf{A}$ und $\mathbf{M}$.

5.3 Krylov-Verfahren für lineare Gleichungssysteme

Die letzte fehlende Komponente, um ein vollständiges numerisches Verfahren zu formulieren und (5.1) numerisch lösen zu können, ist die Approximation der Anwendung der Matrixfunktionen auf Vektoren, ausgehenf von (5.12) etwa $\varphi_1(\tau_n\mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}F(t_n,u_n)$ oder $a_{ij}(\tau_n\mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}D_{nj}$. Dies wollen wir erreichen, ohne explizit Gleichungssysteme mit der Massematrix $\mathbf{M}$ lösen zu müssen. Unser Ansatz basiert dabei auf Krylov-Verfahren, welche wir zunächst erläutern. Den Zusammenhang mit Matrixfunktionen klären wir in Abschnitt 5.5. Da Krylov-Verfahren zur Lösung linearer Gleichungssysteme lange bekannt und weit verbreitet sind, verweisen wir zur Übersicht auf [77] und verzichten in diesem Abschnitt ansonsten weitgehend auf Referenzen.

Wir betrachten zunächst Krylov-Verfahren zur Approximation von Lösungen von linearen Gleichungssystemen $\mathbf{A}\vec{x} = \vec{v}$. Diese konstruieren eine Matrix $\mathbf{V}_m = \left[\vec{v}_1,\ldots,\vec{v}_m\right] \in \mathbb{C}^{N\times m}$, deren Spalten eine Orthonormalbasis bezüglich des Euklidischen Skalarproduktes des *Krylov-Raumes*

$$\mathcal{K}_m(\mathbf{A},\vec{v}) = \operatorname{span}\left\{\vec{v},\mathbf{A}\vec{v},\ldots,\mathbf{A}^{m-1}\vec{v}\right\}$$

bilden. Die Konstruktion erfolgt mittels eines modifizierten Gram-Schmidt-Verfahrens. Wir sprechen bei dieser Konstruktion dieser Basis vom *Arnoldi-Algorithmus*. In letzterem Fall erhält man kurze 3-Term-Rekursionen.

Wir nehmen an, dass die Dimension des Krylov-Raumes m im Vergleich zur vollen Dimension des Systems gering ist, d.h. $m \ll N$. Beginnend mit $\vec{v}_1 = \vec{v}/\beta$, $\beta = \|\vec{v}\|$, wird diese in jedem Schritt um einen Vektor erweitert. Durch die Orthogonalisierung entsteht eine zweite Matrix,

$$\widetilde{\mathbf{H}}_m = \begin{bmatrix} & \mathbf{H}_m & \\ 0 & \ldots & 0\ h_{m+1,m} \end{bmatrix} \in \mathbb{C}^{(m+1)\times m}, \qquad \mathbf{H}_m \in \mathbb{C}^{m\times m},$$

die Hessenberg-Gestalt besitzt – also eine obere Dreiecksmatrix, bei der auch die erste untere Nebendiagonale noch besetzt ist, sodass die Krylov-Relation

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_{m+1}\widetilde{\mathbf{H}}_m = \mathbf{V}_m\mathbf{H}_m + h_{m+1,m}\vec{v}_{m+1}\vec{e}_m^H \qquad \text{mit} \qquad \mathbf{V}_m^H\mathbf{V}_m = \mathbf{Id}_m \tag{5.13}$$

erfüllt ist. Damit gilt

$$\vec{v} = \beta \mathbf{V}_m \vec{e}_1 \qquad \text{und} \qquad \mathbf{H}_m = \mathbf{V}_m^H \mathbf{A} \mathbf{V}_m \tag{5.14}$$

ist die Projektion von $\mathbf{A}$ auf den Krylov-Raum $\mathcal{K}_m(\mathbf{A}, \vec{v})$. Als Approximation der Lösung des Gleichungssystems $\mathbf{A}\vec{x} = \vec{v}$ wählen wir einen Vektor $\vec{x}_m = \mathbf{V}_m \vec{y}_m$, $\vec{y}_m \in \mathbb{C}^m$ aus dem Krylov-Raum. Das entsprechende Residuum ist $\vec{r}_m = \mathbf{A}\vec{x}_m - \vec{v}$.

Für die Wahl von $\vec{y}_m$ gibt es zwei unterschiedliche Ansätze. Der eine fordert die Minimierung des Residuums über dem Krylov-Raum, was auf das GMRES-Verfahren [78], oder im Fall einer symmetrischen Matrix $\mathbf{A}$ auf das MINRES-Verfahren [70] hinführt. Der andere Ansatz, dem auch wir folgen wollen, fordert, dass das Residuum orthogonal auf dem Krylov-Raum steht. Damit erhalten wir das FOM-Verfahren [75] oder im Fall einer symmetrischen, positiv-definiten Matrix $\mathbf{A}$ das cg-Verfahren [37]. In diesem Falle erhalten wir

$$\vec{x}_m = \beta \mathbf{V}_m \mathbf{H}_m^{-1} \vec{e}_1. \tag{5.15}$$

Da wir $m \ll N$ annehmen, ist das Lösen des Gleichungssystems mit der Hessenberg-Matrix $\mathbf{H}_m$ günstig.

Es gibt viele Kriterien, ab wann eine Approximation hinreichend gut ist. Auf diese wollen wir im Kontext dieser Arbeit nicht weiter eingehen. Wir stellen nun vor, wie sich Krylov-Verfahren nutzen lassen, um Matrixfunktionen zu approximieren. Die Herleitung basiert üblicherweise auf der Cauchy-Integralformel (A.22). Für unsere Zwecke ist es jedoch sinnvoller, die Laplace-Transformation (A.25) zugrunde zu legen.

5.4 Darstellung sektorieller Operatorfunktionen über die Laplace-Transformation

Eine analytische Abbildung $\Psi : D \subseteq \mathbb{C} \longrightarrow \mathbb{C}$ heißt *sektoriell*, falls es Konstanten $\delta \in (0, \pi/2)$, $M > 0$, $\gamma \in \mathbb{R}$ und $\nu \geq 1$ gibt, sodass die Abschätzung

$$|\Psi(\lambda)| \leq \frac{M}{|\lambda - \gamma|^\nu} \qquad \text{für alle} \qquad \lambda - \gamma \in \Sigma_\delta \tag{5.16}$$

erfüllt ist, wobei $\Sigma_\delta + \gamma \subseteq D$ und

$$\Sigma_\delta := \{ z \in \mathbb{C} \mid |\arg(z)| < \pi - \delta \},$$

vgl. [64]. Ist für eine Funktion ψ ihre Laplace-Transformierte Ψ mit

$$\psi(t) = \mathcal{L}^{-1}[\Psi](t) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(t\lambda) \Psi(\lambda) \, \mathrm{d}\lambda$$

sektoriell, so dürfen wir die Kurve Γ in dieser Formel so wählen, dass sie positiv ortientiert ist und das Komplement des Sektors $\Sigma_\delta + \gamma$ umläuft, vgl. [85, 64]. Genauer kann Γ als eine einfache Kurve, parametrisiert durch eine glatte Abbildung $T : \mathbb{R} \longrightarrow \mathbb{C}$ so, dass

$$\lim_{x \longrightarrow \pm\infty} \operatorname{Im} T(x) = \pm\infty, \qquad \lim_{x \longrightarrow \pm\infty} \frac{\operatorname{Re} T(x)}{|x|} < 0 \tag{5.17}$$

erfüllt ist, gewählt werden, für die dann $\operatorname{Re}(z) \leq -b|z|$, für $z \longrightarrow \infty$ und $b > 0$ gilt, vgl. [65]. In Abschnitt 5.8 werden wir auf gute Wahlen für diese Kurve eingehen.

Der Operator A heißt *sektoriell*, falls es Konstanten $\delta \in (0, \pi/2)$, $M > 0$ und $\gamma \in \mathbb{R}$ gibt, sodass der Sektor $\Sigma_\delta + \gamma$ in der Resolventenmenge von A enthalten ist und

$$\|(\lambda \operatorname{Id} - A)^{-1}\| \leq \frac{M}{|\lambda - \gamma|} \quad \text{für alle} \quad \lambda - \gamma \in \Sigma_\delta \tag{5.18}$$

erfüllt ist.

Ist A der zu einer Sesquilinearform a assoziierte Operator und ist a j-elliptisch und sektoriell im Sinne von [5, Lektion 5], [4, Kapitel 4], so überträgt sich die Sektorialität mit demselben Winkel auf A, vgl. [5, Proposition 5.5] bzw. [4, Theorem 4.3].

Für die skalare Version der φ_k-Funktionen nutzen wir die Darstellung über die Laplace-Transformation aus (A.27) und setzen für das Argument nun einen sektoriellen Operator A ein. Es ergibt sich

$$\varphi_k(\tau A) = \mathcal{L}^{-1}\left[\Phi_k(\,\cdot\,, \tau A)\right](1) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(\lambda) \Phi_k(\lambda, \tau A)\, \mathrm{d}\lambda, \qquad \Phi_k(\lambda, \tau A) = \frac{1}{\lambda^k} (\lambda \operatorname{Id} - \tau A)^{-1}. \tag{5.19}$$

Für festes $\tau \geq 0$ und sektorielle Operatoren A erhalten wir die Abschätzung

$$\|\Phi_k(\lambda, \tau A)\| \leq \frac{M}{|\lambda - \gamma|^\nu} \quad \text{für ein} \quad \nu \geq 1 \quad \text{und alle} \quad \lambda - \gamma \in \Sigma_\delta \tag{5.20}$$

mit möglicherweise neuen und von τ, A und $R_k := \lambda \longmapsto \frac{1}{\lambda^k}$ abhängigen Konstanten $\delta \in (0, \pi/2)$, $M > 0$ und $\gamma \in \mathbb{R}$, vgl. wieder [64]. Insbesondere enthält das Komplement des Sektors $\Sigma_\delta + \gamma$ neben dem Spektrum von A alle Pole von R_k.

Andere Operatorfunktionen, etwa solche, die man für exponentielle Mehrschrittverfahren [51] benötigt, können als Linearkombinationen der φ_k-Funktionen ausgedrückt werden. Darstellungen für die in [11] hergeleiteten alternativen Mehrschrittverfahren finden sich in [64]. Allgemein

wollen wir Operatorfunktionen auswerten, die die Gestalt

$$\varphi(\tau A) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(\lambda)\Phi(\lambda, \tau A)\,\mathrm{d}\lambda = \mathcal{L}^{-1}[\Phi(\bullet, \tau A)](1), \qquad \Phi(\lambda, \tau A) = R(\lambda)(\lambda\,\mathrm{Id} - \tau A)^{-1}, \tag{5.21}$$

mit einer rationalen Funktion $R : \mathbb{C} \longrightarrow \mathbb{C}$ besitzen.

Solche Paare aus Funktion φ und Operator A, bei denen die transformierte Funktion $\Phi(\,\cdot\,, \tau A)$ sektoriell im Sinne von (5.16) ist, bezeichnen wir als *sektorielle Operatorfunktion*.

Für Matrizen ist eine Abschätzung der Form (5.18) immer möglich, da ihr Wertebereich beschränkt ist. Für Finite-Elemente-Diskretisierungen hängen diese Konstanten jedoch zunächst von der Gitterweite ab. Um für diese Art von Matrizen eine Gitterunabhängige Abschätzung zu erhalten verwenden wir zunächst den auf den Raum V_h eingeschränkte Operator A_h, welcher ebenfalls sektoriell ist, mit denselben Konstanten γ, δ und M, wie A, vgl. etwa [63]. Mit der Sektorialität von A_h erhalten wir

$$\|(\lambda\mathbf{M} - \mathbf{A})^{-1}\|_{\mathbf{M}} \leq \frac{M}{|\lambda - \gamma|} \qquad \text{für alle} \qquad \lambda - \gamma \in \Sigma_\delta \tag{5.22}$$

mit denselben – also gitterunabhängigen – Konstanten γ, δ und M für die Sektorialitätsabschätzung des kontinuierlichen Operators A. Paare aus solchen Matrizen zusammen mit einer Funktion φ, so dass

$$\|R(\lambda)(\lambda\mathbf{M} - \tau\mathbf{A})^{-1}\|_M \leq \frac{M}{|\lambda - \gamma|^\nu} \quad \text{für ein} \quad \nu \geq 1 \quad \text{und alle} \quad \lambda - \gamma \in \Sigma_\delta, \tag{5.23}$$

wobei R die rationale Funktion aus der Laplace-Transformierten von φ ist, bezeichnen wir als *sektorielle Matrixfunktion*.

5.5 Standard Krylov-Verfahren für Matrixfunktionen

Im Folgenden verwenden wir Krylov-Verfahren, um die Anwendungen von Matrixfunktionen auf Vektoren auszuwerten, ohne die Matrixfunktionen explizit auszuwerten. Zunächst benutzen wir das euklidische Skalarprodukt. Dazu verwenden wir die Darstellung aus (5.19) für die φ_k Funktionen und erhalten eine (gitterabhängige) Abschätzung der Form (5.18) für $\mathbf{M}^{-1}\mathbf{A}$ in der euklidischen Norm. Die üblichere Herleitung über die Cauchy-Integralformel (A.22) ist vollkommen analog.

Wir setzen $\widetilde{\mathbf{A}} := \mathbf{M}^{-1}\mathbf{A}$, $\vec{w} := \mathbf{M}^{-1}\vec{v}$ und verstecken vorübergehend die Anwendung der Inversen der Massematrix. Aus (5.21) erhalten wir für den diskretisierten Differentialoperator aus (5.12)

$$\varphi(\tau\mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}\vec{v} = \varphi(\tau\widetilde{\mathbf{A}})\vec{w} = \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(\lambda)R(\lambda)(\lambda\,\mathbf{Id} - \tau\widetilde{\mathbf{A}})^{-1}\vec{w}\,\mathrm{d}\lambda. \tag{5.24}$$

Unter dem Integral steht nun die Lösung eines linearen Gleichungssystems, für welches wir ein Krylov-Verfahren benutzen können. Allerdings erhalten wir für jedes λ und τ eine andere Lösung. Wie man leicht nachvollziehen kann, gilt

$$\mathcal{K}_m(\lambda \mathbf{Id} - \tau \widetilde{\mathbf{A}}, \vec{w}) \subseteq \mathcal{K}_m(\widetilde{\mathbf{A}}, \vec{w}) \tag{5.25}$$

für alle $\lambda \in \mathbb{C}$ und wir können für alle λ denselben Krylov-Raum verwenden. Offensichtlich ergibt sich dabei aus (5.13) $\mathbf{V}_m^H(\lambda \mathbf{Id}_N - \tau \widetilde{\mathbf{A}})\mathbf{V}_m = \lambda \mathbf{Id}_m - \tau \mathbf{H}_m$. Bezeichnen wir mit $\vec{x}_m(\lambda, \tau) = \beta \mathbf{V}_m(\lambda \mathbf{Id}_m - \tau \mathbf{H}_m)^{-1}\vec{e}_1$ die Krylov-Näherung, die sich aus der Orthogonalisierung des Residuums ergibt, können wir rückwärts einsetzen:

$$\varphi(\tau \widetilde{\mathbf{A}})\vec{w} \approx \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(\lambda) R(\lambda) \beta \mathbf{V}_m (\lambda \mathbf{Id} - \tau \mathbf{H}_m)^{-1} \vec{e}_1 \, \mathrm{d}\lambda = \beta \mathbf{V}_m \varphi(\tau \mathbf{H}_m)\vec{e}_1. \tag{5.26}$$

Der Fehler dieses Verfahrens ist gegeben durch

$$\begin{aligned} \epsilon_m &= \varphi(\tau \widetilde{\mathbf{A}})\vec{w} - \beta \mathbf{V}_m \varphi(\tau \mathbf{H}_m)\vec{e}_1 \\ &= \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(\lambda) R(\lambda) \left[(\lambda \mathbf{Id} - \tau \widetilde{\mathbf{A}})^{-1} \vec{w} - \beta \mathbf{V}_m (\lambda \mathbf{Id} - \tau \mathbf{H}_m)^{-1} \vec{e}_1 \right] \mathrm{d}\lambda, \end{aligned}$$

wobei $\left[(\lambda \mathbf{Id} - \tau \widetilde{\mathbf{A}})^{-1} \vec{w} - \beta \mathbf{V}_m (\lambda \mathbf{Id} - \tau \mathbf{H}_m)^{-1} \vec{e}_1 \right]$ der Fehler beim Lösen der linearen Gleichungssysteme ist. Da wir diesen Fehler nicht ausrechnen können, ohne bereits die exakte Lösung zu kennen, ersetzen wir ihn durch das Residuum $\vec{w} - \beta(\lambda \mathbf{Id} - \tau \widetilde{\mathbf{A}})\mathbf{V}_m(\lambda \mathbf{Id} - \tau \mathbf{H}_m)^{-1}\vec{e}_1$ und erhalten das verallgemeinerte Residuum

$$\begin{aligned} \rho_m &= \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(\lambda) R(\lambda) \left[\vec{v} - \beta(\lambda \mathbf{Id} - \tau \widetilde{\mathbf{A}}) \mathbf{V}_m (\lambda \mathbf{Id} - \tau \mathbf{H}_m)^{-1} \vec{e}_1 \right] \mathrm{d}\lambda \\ &= \beta \tau h_{m+1,m} \vec{v}_{m+1} \vec{e}_m^H \varphi(\tau \mathbf{H}_m)\vec{e}_1 \end{aligned} \tag{5.27}$$

unter Ausnutzung der Krylov-Relation (5.13) und der ersten Gleichung aus (5.14).

In Algorithmus 1 ist der vollständige Arnoldi-Algorithmus für Matrixfunktionen angegeben. Das Auswerten der Matrixfunktion auf der kleinen $m \times m$-Matrix $\mathbf{H}_m$ kann dann mit einem der direkten Verfahren erledigt werden, da der Aufwand aufgrund der geringen Dimension vertretbar ist.

Wir müssen bei jeder Anwendung von $\widetilde{\mathbf{A}}$ in Zeile 5, also in jedem Krylov-Schritt, einmal ein lineares Gleichungssystem mit der Massematrix lösen. Für symmetrische oder schiefsymmetrische Matrizen $\widetilde{\mathbf{A}}$ überträgt sich die entsprechende Struktur auf $\mathbf{H}_m$, wodurch letztere Matrix Tridiagonalgestalt erhält. Dadurch reduzieren sich die Zeilen 6 bis 9 auf kurze 3-Term-Rekursionen für die Bestimmung des Vektors $\vec{v}_{m+1}$. Es muss also nicht die vollständige Krylov-Basis gespeichert werden und die Anzahl benötigter Skalarprodukte wird deutlich reduziert. Das Anwenden von $\mathbf{V}_m$ am Ende kann dann mit Rückwärtsrechnen ausgeführt werden, vgl. etwa den `exp4`-Code zum Ver-

Algorithmus 1 Standard Arnoldi-Algorithmus für Matrixfunktionen

1: **procedure** MATRIXFUNCTIONARNOLDI($\widetilde{A}$, w, τ, φ)
2: $\beta \leftarrow \|w\|$
3: $v_1 \leftarrow w/\beta$
4: **for** $m = 1, 2, \ldots$ **do**
5: $\widetilde{v}_{m+1} \leftarrow \widetilde{A} v_m$
6: **for** $j = 1, 2, \ldots, m$ **do**
7: $h_{j,m} \leftarrow \langle v_j, \widetilde{v}_{m+1} \rangle$
8: $\widetilde{v}_{m+1} \leftarrow \widetilde{v}_{m+1} - h_{j,m} v_j$
9: **end for**
10: $h_{m+1,m} \leftarrow \|\widetilde{v}_{m+1}\|$
11: $v_{m+1} \leftarrow \widetilde{v}_{m+1} / h_{m+1,m}$
12: **end for**
13: **return** $\beta V_m \varphi(\tau H_m) e_1$
14: **end procedure**

fahren aus [47]. Allerdings ist $\widetilde{\mathbf{A}}$ selbst bei symmetrischer Matrix $\mathbf{M}$ und (schief-)symmetrischem $\mathbf{A}$ nur dann (schief-)symmetrisch, wenn die beiden Matrizen kommutieren. Dies ist im Allgemeinen und nahezu in allen konkreten Anwendungen nicht der Fall.

5.6 Verwendung des Masseskalarproduktes

Es bleibt zu klären, wie wir die Massematrix „richtig“ in das Krylov-Verfahren integrieren. Der erste und vielleicht offensichtlichste Ansatz ist, das Skalarprodukt und die Norm in Algorithmus 1 durch die natürlichen, zum Funktionenraum V bzw. V_h passenden zu ersetzen. Mit (5.11) ersetzen wir also die Norm in den Zeilen 2 und 10 durch $\| \cdot \|_{\mathbf{M}}$ und das Skalarprodukt in Zeile 7 durch $\langle \cdot, \cdot \rangle_{\mathbf{M}}$.

Der Nachteil der Vektor-2-Norm gegenüber der L^2-Norm ist es, dass alle Fehlerkomponenten gleich behandelt werden. Ansatzfunktionen hoher Ordnung oder solche, die auf im Verhältnis zum restlichen Gitter sehr kleinen Teilgebieten definiert sind, leisten jedoch in der problembezogenen L^2-Norm nur einen verhältnismäßig kleinen Beitrag zum Gesamtfehler.

Außerdem gewinnt man durch das passende Skalarprodukt wegen $\langle \vec{v}, \widetilde{\mathbf{A}} \vec{w} \rangle_{\mathbf{M}} = \langle \vec{v}, \mathbf{A} \vec{w} \rangle$ die Symmetrieeigenschaften von $\mathbf{A}$ in der Hessenberg-Matrix $\mathbf{H}_m$ zurück, wie wir es von Gleichungen ohne Massematrix kennen. Wie am Ende des vorigen Abschnittes diskutiert, hatten wir diese zunächst verloren. Da sich in $\mathbf{A}$ die Symmetrieeigenschaften des ursprünglichen Differentialoperators widerspiegeln, die sich so auf $\mathbf{H}_m$ übertragen, ist dies vom theoretischen Standpunkt aus gesehen eine schöne Eigenschaft. Weiterhin erhalten wir auch den rechnerischen Vorteil der Ersparnis von Rechenzeit und Speicherplatz in Form kurzer 3-Term-Rekursionen wieder zurück.

Dieses Vorgehen ist mittlerweile in der Literatur verbreitet, vgl. etwa [27, 53, 42] mit Anwendun-

gen auf rationale Krylov-Verfahren, Maxwell-Gleichungen mit räumlicher Discontinuous-Galerkin-Approximation und für Magnus-Integratoren.

Auch in [79] und [58] wurde dieser Ansatz verfolgt. Die klassische Fehleranalyse für Krylov-Verfahren bei Matrixfunktionen aus [76] und [45] kann angewendet werden, die Fehlernorm wird dabei durch die zum Problem passende L^2-Norm ersetzt. [58] integriert die Methode in das exponentielle Integratorenpaket EXPODE [55, 56] und stellt auch die übrige Fehlermessung für die Fehlerschätzer auf diese Norm ein.

In [17] wird eine andere Gewichtungsvariante für Krylov-Verfahren wie FOM oder GMRES vorgeschlagen. [29] stellen allerdings dessen Effektivität im Vergleich zu vorkonditionierten Krylov-Varianten in Zweifel. Die dort verwendete Skalierungsmethode stammt jedoch nicht wie hier aus dem Problem, sondern wird adaptiv aus dem Residuum bestimmt und ändert sich in jedem Schritt. Die Aussagen aus [29] treffen nicht auf L^2-Norm und -Skalarprodukt zu.

Das Verwenden von Massenorm und -skalarprodukt sollte also als die richtige Vorgehensweise bei Standard Krylov-Verfahren mit Matrixfunktionen angesehen werden. Dieses Vorgehen erreicht allerdings nicht das Ziel dieses Kapitels, nämlich das Vermeiden der Lösung von linearen Gleichungssystemen mit der Massematrix. Diesem Problem widmen wir uns im nächsten Abschnitt.

5.7 Vermeidung von Lösungen von Massesystemen

In diesem Abschnitt werden wir eine Variante des in Algorithmus 1 vorgestellten Krylov-Verfahrens für Matrixfunktionen herleiten, die ohne explizite Lösungen von linearen Gleichungssystemen mit der Massematrix auskommt. Wenn wir in (5.24) die Inverse von $\mathbf{M}$ nicht in $\widetilde{\mathbf{A}}$ und $\vec{w}$ verstecken, können wir

$$\begin{aligned}\varphi(\tau\mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}\vec{v} &= \frac{1}{2\pi\,\mathbb{i}}\int_\Gamma \exp(\lambda)R(\lambda)(\lambda\,\mathbf{Id}-\tau\mathbf{M}^{-1}\mathbf{A})^{-1}\mathbf{M}^{-1}\vec{v}\,\mathrm{d}\lambda \\ &= \frac{1}{2\pi\,\mathbb{i}}\int_\Gamma \exp(\lambda)R(\lambda)(\lambda\mathbf{M}-\tau\mathbf{A})^{-1}\vec{v}\,\mathrm{d}\lambda \end{aligned} \tag{5.28}$$

umstellen. Es ist nun wieder möglich, ein Krylov-Verfahren zu verwenden, um die linearen Gleichungssysteme, dieses Mal mit den verallgemeinerten Resolventen $(\lambda\mathbf{M}-\tau\mathbf{A})^{-1}$, zu lösen. Allerdings sind im Gegensatz zu (5.25) die Krylov-Räume $\mathcal{K}_m(\lambda\mathbf{M}-\tau\mathbf{A},\vec{v})$ und $\mathcal{K}_m(\mathbf{A},\vec{v})$ im Allgemeinen nicht ineinander enthalten. Dies ignorieren wir zunächst und benutzen für die Projektion den fest gewählten Raum $\mathcal{K}_m(\mathbf{A},\vec{v})$. Für alle $\lambda \neq 0$ erhalten wir keine Krylov-Approximation an die Lösung des Gleichungssystems im klassischen Sinne, sondern lediglich die Lösung nach der Projektion auf einen festen Unterraum. In diesem Kontext ist das Verwenden des Masseskalarproduktes nicht sinnvoll, da wir jetzt $\mathbf{A}$ anstelle von $\mathbf{M}^{-1}\mathbf{A}$ projizieren. Mit dem Standardskalarprodukt überträgt sich die Struktur von $\mathbf{A}$ auf die von $\mathbf{H}_m$.

Wir fordern, dass das Residuum orthogonal auf dem Krylov-Raum steht, und erhalten die Näherung

$$\vec{x}_m(\lambda,\tau) = \beta \mathbf{V}_m(\lambda \mathbf{G}_m - \tau \mathbf{H}_m)^{-1}\vec{e}_1 \tag{5.29}$$

an die Lösung der linearen Gleichungssysteme mit der Orthogonalprojektion

$$\mathbf{G}_m = \mathbf{V}_m^H \mathbf{M} \mathbf{V}_m \tag{5.30}$$

von $\mathbf{M}$ auf $\mathcal{K}_m(\mathbf{A},\vec{v})$. Rückwärts Einsetzen liefert die Näherung

$$\varphi(\tau \mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}\vec{v} \approx \beta \mathbf{V}_m \varphi(\tau \mathbf{G}_m^{-1}\mathbf{H}_m)\mathbf{G}_m^{-1}\vec{e}_1 \tag{5.31}$$

für die Matrixfunktion mit derselben Idee wie in (5.28). Wir tauschen also das Lösen der $m+1$ linearen Gleichungssysteme mit der Matrix $\mathbf{M}$, für die Bestimmung des Startwertes und bei jeder Anwendung von $\widetilde{\mathbf{A}}$, gegen das deutlich günstigere Lösen von $m+1$ linearen Gleichungssystemen mit der Matrix $\mathbf{G}_m$, und zwar m-mal mit den Spalten von $\mathbf{H}_m$ und einmal mit $\vec{e}_1$ als rechte Seiten. Unter der realistischen Voraussetzung, dass $\mathbf{M}$ nicht nur invertierbar ist, sondern die Null auch nicht im Wertebereich liegt, gilt dies auch für $\mathbf{G}_m$ und die Matrix ist damit ebenso invertierbar.

Für ein Paar $(\mathbf{H}_m, \mathbf{G}_m)$, für das eine verallgemeinerte Eigenwertzerlegung

$$\mathbf{H}_m \mathbf{W_m} = \mathbf{G}_m \mathbf{W_m} \mathbf{D_m} \tag{5.32}$$

mit einer Diagonalmatrix $\mathbf{D_m}$ und einer – ggf. orthogonalen oder unitären, aber vor allem invertierbaren – Matrix $\mathbf{W_m}$ aus Eigenvektoren existiert, lässt sich die Matrixfunktion aus (5.31) mithilfe der Cauchy-Integralformel (A.22) schreiben als

$$\begin{aligned}
\varphi(\tau \mathbf{G}_m^{-1}\mathbf{H}_m)\mathbf{G}_m^{-1} &= \frac{1}{2\pi \mathring{\mathbb{1}}} \int_\Gamma \varphi(\lambda)(\lambda \mathbf{G}_m - \tau \mathbf{H}_m)^{-1}\,\mathrm{d}\lambda \\
&= \frac{1}{2\pi \mathring{\mathbb{1}}} \int_\Gamma \varphi(\lambda)\mathbf{W_m}(\lambda\,\mathbf{Id} - \tau \mathbf{D_m})^{-1}\mathbf{W_m}^{-1}\mathbf{G}_m^{-1}\,\mathrm{d}\lambda \\
&= \mathbf{W_m}\varphi(\tau \mathbf{D_m})\mathbf{W_m}^{-1}\mathbf{G}_m^{-1}.
\end{aligned} \tag{5.33}$$

Falls man also die Matrixfunktionen auf der kleinen Matrix mit einer Eigenwertzerlegung berechnet, kann der QR-Algorithmus für die Eigenwertbestimmung durch den QZ-Algorithmus [25] ersetzt werden und es muss zusätzlich nur eine Lösung eines Gleichungssystems mit der Matrix $\mathbf{G}_m$ berechnet werden.

Der Fehler des Verfahrens ist gegeben als

$$\begin{aligned}
\epsilon_m &= \varphi(\tau \mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}\vec{v} - \beta \mathbf{V}_m \varphi(\tau \mathbf{G}_m^{-1}\mathbf{H}_m)\mathbf{G}_m^{-1}\vec{e}_1 \\
&= \frac{1}{2\pi \mathring{\mathbb{1}}} \int_\Gamma \varphi(\lambda)\left[(\lambda \mathbf{M} - \tau \mathbf{A})^{-1}\vec{v} - \beta \mathbf{V}_m(\lambda \mathbf{G}_m - \tau \mathbf{H}_m)^{-1}\vec{e}_1\right]\mathrm{d}\lambda,
\end{aligned}$$

wobei $\left[(\lambda\mathbf{M}-\tau\mathbf{A})^{-1}\vec{v}-\beta\mathbf{V}_m(\lambda\mathbf{G}_m-\tau\mathbf{H}_m)^{-1}\vec{e}_1\right]$ der Fehler ist, der beim Lösen des linearen Gleichungssystems für ein beliebiges festes λ entsteht. Dabei haben wir auch hier die Cauchy-Integralformel für die Darstellung verwendet. Wir ersetzen diesen unbekannten Fehler wieder durch das Residuum und erhalten

$$\rho_m = \beta\tau h_{m+1,m}\vec{e}_m^H\varphi(\tau\mathbf{G}_m^{-1}\mathbf{H}_m)\mathbf{G}_m^{-1}\vec{e}_1\vec{v}_{m+1} + \beta\tau(\mathbf{V}_m\mathbf{G}_m-\mathbf{M}\mathbf{V}_m)\varphi(\tau\mathbf{G}_m^{-1}\mathbf{H}_m)\mathbf{G}_m^{-1}\mathbf{H}_m\mathbf{G}_m^{-1}\vec{e}_1. \tag{5.34}$$

Dabei entsteht der zweite Summand dadurch, dass die Matrix $\mathbf{M}$ künstlich mit $\mathbf{V}_m$ vertauscht werden muss. Wir haben dabei zwischen $\mathbf{M}$, $\mathbf{V}_m$ und $\mathbf{G}_m$ keine Relation wie (5.13), sondern nur (5.30), welches der Folgerung (5.14) aus dieser Relation entspricht. Den Faktor

$$\mathbf{V}_m\mathbf{G}_m-\mathbf{M}\mathbf{V}_m \tag{5.35}$$

können wir als eine Art von Kommutator zwischen der Krylov-Basis und der Massematrix interpretieren und wir können ihn als „Strafe“ dafür ansehen, dass wir die „falschen“ Krylov-Räume benutzen, um die verschobenen Gleichungssysteme zu lösen.

Algorithmus 2 Arnoldi-Algorithmus für Matrixfunktionen ohne explizites Lösen mit der Massematrix

procedure MATRIXFUNCTIONARNOLDI(A, M, v, τ, φ)
  $\beta \leftarrow \|v\|$
  $v_1 \leftarrow v/\beta$
  $g_{1,1} \leftarrow \langle v_1, Mv_1\rangle$
  for $m = 1, 2, \ldots$ **do**
    $\widetilde{v}_{m+1} \leftarrow Av_m$
    for $j = 1, 2, \ldots, m$ **do**
      $h_{j,m} \leftarrow \langle v_j, \widetilde{v}_{m+1}\rangle$
      $\widetilde{v}_{m+1} \leftarrow \widetilde{v}_{m+1} - h_{j,m}v_j$
    end for
    $h_{m+1,m} \leftarrow \|\widetilde{v}_{m+1}\|$
    $v_{m+1} \leftarrow \widetilde{v}_{m+1}/h_{m+1,m}$
    $\widehat{v}_{m+1} \leftarrow Mv_{m+1}$
    for $j = 1, 2, \ldots, m+1$ **do**
      $g_{j,m+1} \leftarrow g_{m+1,j} \leftarrow \langle v_j, \widehat{v}_{m+1}\rangle$
    end for
  end for
  return $\beta V_m\varphi(\tau G_m^{-1}H_m)G_m^{-1}e_1$, z.B. mit (5.33)
end procedure

Das entstandene Verfahren ist in Algorithmus 2 zusammengefasst. Für symmetrische oder schiefsymmetrische Matrizen $\mathbf{A}$ überträgt sich die entsprechende Struktur auf $\mathbf{H}_m$, wodurch letztere Matrix Tridiagonalgestalt erhält. Die Matrix $\mathbf{G}_m$ erhält entsprechend die Struktur von $\mathbf{M}$, ist aber *nicht* von Hessenberg-Gestalt. Insbesondere muss auch im symmetrischen Fall die Matrix $\mathbf{V}_m$

(oder $\mathbf{MV}_m$) vollständig im Speicher gehalten werden.

Falls das Residuum (5.34) in jedem Schritt, oder wie üblich in immer größer werdenden Intervallen, ausgerechnet werden soll, kann zusätzlich zu $\mathbf{V}_m$ die Matrix $\mathbf{MV}_m$ gespeichert und entsprechend erweitert werden.

Testen wir nun das neue Verfahren mit einem einfachen numerischen Experiment. Als Testgleichung wählen wir die Wärmeleitungsgleichung

$$\begin{cases} \partial_t u(t,\vec{x}) = a^2 \Delta u(t,\vec{x}) + g(t,u(t,\vec{x})) & \text{in } \Omega, \\ \partial_\nu u(t,\vec{x}) = 0 & \text{auf } \partial\Omega, \\ u(t_0,\vec{x}) = u_0(\vec{x}) \end{cases} \tag{5.36}$$

zunächst mit $\Omega = [0,1]$ in einer Raumdimension und $g(t,u(t)) \equiv g_0$. Als Startwert benutzen wir $u_0(\vec{x}) = u_0(x) = \max\left\{-|x-\frac{1}{2}|+\frac{1}{4},0\right\}$. Die exakte Lösung ist als die Anwendung der φ_1-Operatorfunktion des Δ-Operators auf den Startwert gegeben. Zur Ortsdiskretisierung verwenden wir lineare finite Elemente auf einer regelmäßigen Unterteilung und erhalten eine gewöhnliche Differentialgleichung der Form (5.9) mit tridiagonalen Matrizen $\mathbf{A}$ und $\mathbf{M}$. Die Ortsschrittweite wählen wir als $h = \frac{1}{150}$ und erhalten somit ein System der Dimension $N = 151$. Die exakte Lösung der ortsdiskretisierten Gleichung für zeitlich konstantes g ist durch

$$\vec{u}(t) = \mathbf{M}^{-1}\vec{l}(u_0) + (t-t_0)\varphi_1((t-t_0)\mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}(\mathbf{A}\mathbf{M}^{-1}\vec{l}(u_0) + \vec{l}(g_0))$$

gegeben. Wir müssen also im Wesentlichen die Matrixfunktion $\varphi_1((t-t_0)\mathbf{M}^{-1}\mathbf{A})$ auf einen Vektor der Form $\mathbf{M}^{-1}\vec{v}$ anwenden. Da wir uns für die Güte der Approximation der Auswertung der Matrixfunktionen interessieren, berechnen wir nur einen Zeitschritt der Länge $\tau = 0,1$.

In Abbildung 5.1 sind Kurven für die Fehler der verschiedenen Varianten der Krylov-Verfahren aufgetragen. Dabei haben wir das Standard Krylov-Verfahren für Matrixfunktionen, Algorithmus 1, sowohl mit Verwendung des Standard- als auch des Masseskalarproduktes und die neue Variante ohne Lösungen von Massesystemen, Algorithmus 2, verwendet. Der Fehler des letzten Verfahrens liegt dabei – trotz des geringeren Aufwandes – unterhalb des Fehlers der Standardverfahren, die ihrerseits fast dieselben Ergebnisse liefern. Auch liegt das verallgemeinerte Residuum näher am Approximationsfehler. Erwartungsgemäß ist bei den Standardvarianten das in der L_2-Norm gemessene Residuum der Version mit dem Masseskalarprodukt näher am in der L_2-Norm gemessenen Fehler. Das sehr ähnliche Fehlerverhalten der beiden Standardvarianten ist könnte dadurch zu erklären sein, dass die Massematrix nur zwei besetzte Nebendiagonalen besitzt und die Einträge entlang der Diagonalen jeweils fast konstant sind. Masse- und Standardskalarprodukt unterscheiden sich daher hauptsächlich durch den Faktor der Ortsschrittweite Δx. Inbesondere liefert das Orthogonalisieren fast dieselben Vektoren.

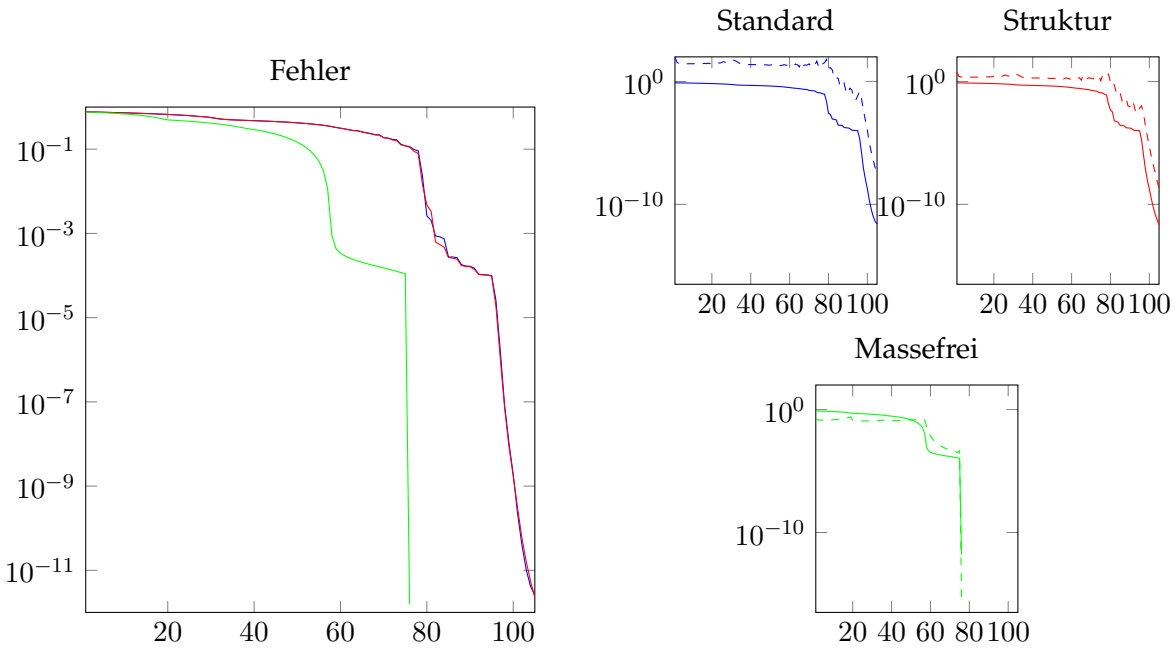


Abbildung 5.1: Testproblem (5.36) mit 1D linearen finiten Elementen, $\varphi \equiv \varphi_1$, $\tau = 0.1$. Fehlermessung in der L_2-Norm. x-Achse: Anzahl der Krylov-Schritte, y-Achse: relativer Fehler. Links: Approximation (5.26) mit Standard- (Standard, blau) und Masseskalarprodukt (Struktur, rot) und masselösungsfreie Approximation (5.31) (Massefrei, grün). Rechts: Fehler (durchgezogen) gegen verallgemeinertes Residuum (gestrichelt), (5.27) bzw. (5.34) für alle drei Varianten.

Nach diesem vielversprechenden Ergebnis wollen wir dasselbe Experiment in zwei Raumdimensionen durchführen. In (5.36) wird dabei das Gebiet durch das Einheitsquadrat $\Omega = [0,1]^2$ ersetzt, das Ortsgitter bleibt regelmäßig mit einer Gitterweite von $h = \frac{1}{40}$ und damit $N = 1681$, der Startwert ist gewählt als $u_0(\vec{x}) = \max\left\{-\|\vec{x} - [\frac{1}{2}, \frac{1}{2}]^T\| + \frac{1}{4}, 0\right\}$.

Die Ergebnisse sind in Abbildung 5.2 im gleichen Format wie in der eindimensionalen Variante dargestellt. Dieses Mal verhält sich das neue Verfahren jedoch deutlich schlechter, als im eindimensionalen Fall. Das Standard-Krylov-Verfahren konvergiert sehr schnell, wohingegen Algorithmus 2 zunächst sogar noch etwas schneller, aber ab einer Größenordnung von etwa 10^{-2} sehr langsam den Fehler verkleinert. Bei Tests mit gröberer Ortsauflösung zeigte sich ähnliches Verhalten. Der Fehler wird erst klein, wenn die Krylov-Raum-Dimension m schon nahe bei der Dimension des ursprünglichen Gleichungssystems N liegt, und erreicht dann in den wenigen restlichen Schritten die Maschinengenauigkeit.

Um dieses Verhalten zu bestätigen, haben wir in Abbildung 5.3 die Fehler der Krylov-Verfahren bei zwei weiteren Ortsdiskretisierungen dargestellt. Zunächst verwenden wir finite Elemente der Ordnung vier, weiterhin auf dem strukturierten Gitter, dann die selben Elemente auf einem unstrukturierten Gitter. Die verwendeten Gitter sind ebenfalls dargestellt und verwenden Gitter-

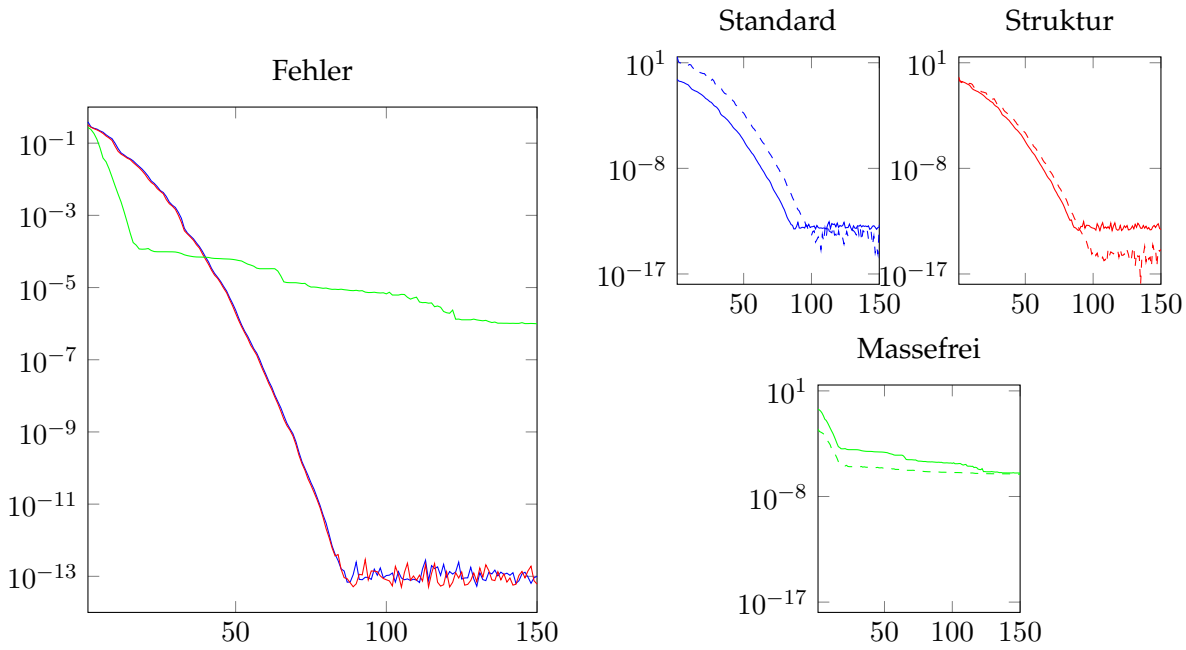


Abbildung 5.2: Testproblem (5.36) mit 2D linearen finiten Elementen, $\varphi \equiv \varphi_1$, $\tau = 0.1$. Fehlermessung in der L_2-Norm. x-Achse: Anzahl der Krylov-Schritte, y-Achse: relativer Fehler. Links: Approximation (5.26) mit Standard- (Standard, blau) und Masseskalarprodukt (Struktur, rot) und masselösungsfreie Approximation (5.31) (Massefrei, grün). Rechts: Fehler (durchgezogen) gegen verallgemeinertes Residuum (gestrichelt), (5.27) bzw. (5.34) für alle drei Varianten.

weiten von etwa $h = 0.2$, sind also gröber als das Gitter des letzten Beispieles. Die Ergebnisse sind sogar noch schlechter.

Hier erkennen wir aber einen – wenn auch kleinen – Unterschied zwischen den Standard Krylov-Varianten mit den beiden verschiedenen Skalarprodukten. Je höher der Polynomgrad und je ungleichmäßiger die Triangulierung, desto stärker weicht die Massematrix von einer Diagonalmatrix mit konstanten Diagonalen ab und damit die Gewichtung der Freiheitsgrade untereinander im Masseskalarprodukt.

Für die Fehleranalyse von Saad [76] ist die Eigenschaft

$$p(\mathbf{M}^{-1}\mathbf{A})\vec{v}_1 = \mathbf{V}_m p(\mathbf{H}_m)\vec{e}_1 \qquad \text{für alle} \qquad p \in \mathcal{P}_{m-1} \tag{5.37}$$

wesentlich. Wenn wir diese in unserem Fall nachzuweisen versuchen, treten in der Differenz beider Seiten ähnliche Kommutatoren wie in (5.35), diesmal zwischen $\mathbf{M}^{-1}$ bzw. $\mathbf{G}_m^{-1}$ und $\mathbf{V}_m$, auf. Dies ist auf den Verlust der Invarianzeigenschaft (5.25) zurückzuführen. Auch eine analoge Aussage zu Lemma 1 aus der Analyse von Hochbruck und Lubich [45], auf der die dortigen Fehlerabschätzungen beruhen, wird ohne diese Invarianz schwierig herzuleiten sein. Die oben gezeigten numerischen Ergebnisse legen nahe, dass wir auch keinen Ersatz für (5.37) bzw. dieses Lemma

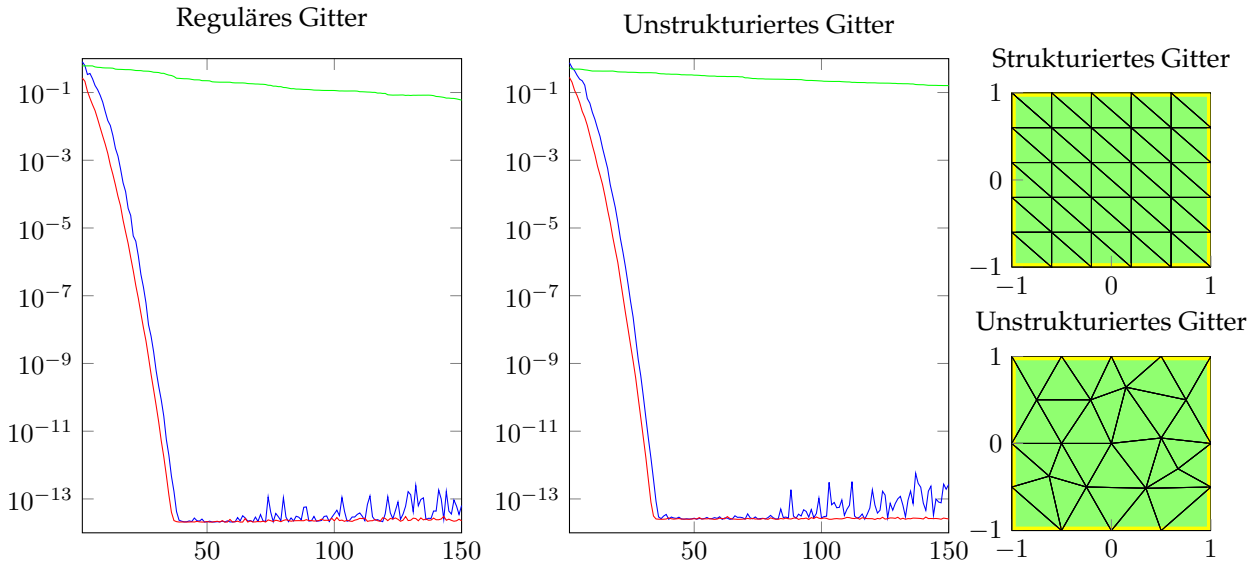


Abbildung 5.3: Testproblem (5.36) mit 2D finiten Elementen der Ordnung vier, $\varphi \equiv \varphi_1$, $\tau = 0.1$. x-Achse: Anzahl der Krylov-Schritte. y-Achse: relativer Fehler in der Approximation (5.26) mit Standard- (Standard) und Masseskalarprodukt (Struktur) und der masselösungsfreien Approximation (5.31) (Massefrei), gemessen in der L_2-Norm. Links: strukturiertes Gitter, Mitte: unstrukturiertes Gitter, rechts: Plot der beiden Gitter.

werden finden können.

Im weiteren Verlauf dieses Kapitels werden wir versuchen, das Problem der schlechten Konvergenz durch eine Verbesserung des Verfahrens zu beheben. Um zu verhindern, dass wir die linearen Gleichungssysteme in den „falschen“ Krylov-Räumen lösen, steigen wir auf die „richtigen“ um, also auf $\mathcal{K}_m(\lambda\mathbf{M} - \tau\mathbf{A}, \vec{v})$. Dies ist für die überabzählbar vielen $\lambda \in \Gamma$ aber nicht möglich. Dies führt uns zu der Idee, die Kurve Γ zu diskretisieren und damit zu Verfahren für eine numerische – nicht wie bisher verwendet analytische – inverse Laplace-Transformation, wie sie unter anderem in [64, 65, 91] vorgeschlagen wird, überzugehen. Diese wird in den nächsten beiden Abschnitten eingeführt. Danach kombinieren wir sie im finalen Abschnitt mit unseren Krylov-Verfahren und diskutieren einige Beispiele.

5.8 Numerische inverse Laplace-Transformation

Der klassische Ansatz, Krylov-Verfahren für Matrixfunktionen zu verwenden, benutzt die Darstellung (5.21) zweimal und die Auswertung der Matrixfunktion auf einer projizierten Version der Matrix bzw. des Operators. Wie wir im letzten Abschnitt herausgearbeitet haben, ist jedoch nach der Umstellung (5.28) die direkte Auswertung der Matrixfunktionen auf der Projektion nicht ideal. Als Alternative verwenden wir das Krylov-Verfahren nicht direkt nach der Laplace-Trans-

formation, sondern kehren diese zuerst numerisch um. Dann können die dort entstehenden Gleichungssysteme mit einem Krylov-Verfahren gelöst werden. Dazu führen wir nun eine numerische Methode zur Invertierung der Laplace-Transformation ein.

Wir verwenden dabei das Vorgehen aus den Arbeiten [64, 65, 80]. Zunächst wird eine spezielle Kontur Γ gewählt, deren Verlauf in der komplexen Halbebene sich auf die Konvergenz der Diskretisierungsmethode auswirkt. [64, 65] wählen die Kontur als eine Hyperbel, die durch die Abbildung

$$T : \mathbb{R} \longrightarrow \Gamma, \qquad x \longmapsto \mu(1 - \sin(\alpha + \mathring{\imath} x)) + \gamma \tag{5.38}$$

gegeben ist. Dabei ist μ ein Skalierungsparameter und $0 < \alpha < \frac{\pi}{2} - \delta$ sowie γ und δ aus der Abschätzung (5.20) abhängig vom Operator und der rationalen Funktion $R(z)$. Damit gilt $T(\mathbb{R}) \subseteq \Sigma_\delta$. In Abbildung 5.4 ist eine solche Kurve zusammen mit den entsprechenden Sektoren für den Operator und die rationale Funktion dargestellt.

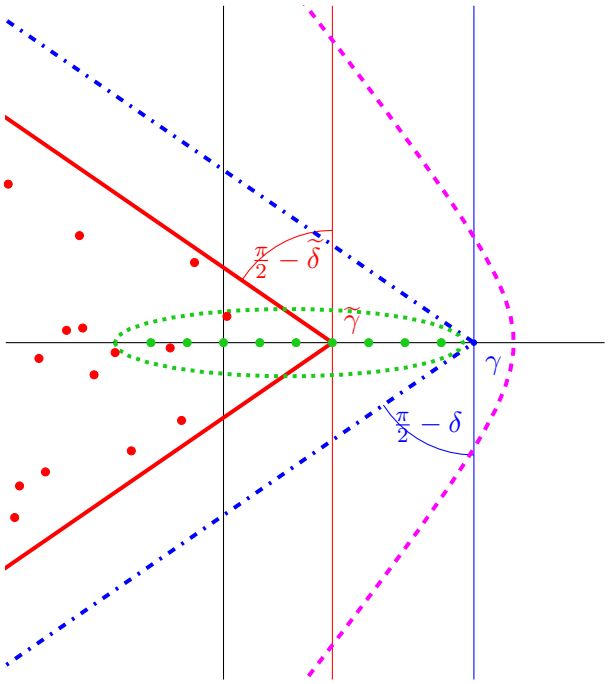


Abbildung 5.4: Sektor, in dem das Spektrum von τA (rot-durchgezogen) enthalten ist, und ein zweiter Sektor, der zusätzlich die Pole enthält (blaue Strichpunkte). Hyperbel (5.38) in pink-gestrichelt. Rote Punkte symbolisieren das Spektrum von τA, grüne die Pole von R.

Alternativen zu dieser Wahl sind etwa Parabeln [24] oder Talbot-Konturen [85], falls $\delta = 0$. [86] identifizieren die Wahl der Kurve mit einer rationalen Approximation und verwenden eine Best-Approximation.

Sobald wir eine geeignete Kurve Γ durch eine Parametrisierung wie (5.38) zur Hand haben, schneiden wir die Kurve ab und benutzen die Trapezregel, um das parametrisierte Integral zu approxi-

mieren. Dies liefert für eine Funktion ψ und ihre Laplace-Transformierte Ψ mit $\psi(t) = \mathcal{L}^{-1}[\Psi](t)$ die Näherung

$$\psi(t) \approx \sum_{l=-K}^{K} w_l \exp(t\lambda_l)\Psi(\lambda_l) \tag{5.39}$$

mit der Schrittweite h und

$$w_l = -\frac{h}{2\pi \mathbb{i}} T'(lh), \qquad \lambda_l = T(lh), \qquad -K \leq l \leq K. \tag{5.40}$$

Die Konvergenz dieser Näherung und die Wahl der Parameter liefert uns Theorem 2.1 aus [64] als Zusammenfassung der Aussagen aus [65].

Theorem 5.1 (López-Fernández, Palencia, Schädle, [64, Theorem 2.1]). *Die Laplace-Transformierte Ψ möge sektoriell mit einem Öffnungswinkel δ und Exponenten ν aus (5.16) sein, und α und d seien so gewählt, dass*

$$0 < \alpha - d < \alpha + d < \frac{\pi}{2} - \delta \tag{5.41}$$

gilt. Für $t_0 > 0$, $\Lambda \geq 1$ und $K \geq 1$ seien die Parameter

$$h = \frac{a(\theta^*)}{K}, \qquad \mu = \frac{2\pi d K(1-\theta^*)}{\Lambda t_0 a(\theta^*)} \tag{5.42}$$

mit der Abbildung

$$a : (0,1) \longrightarrow \mathbb{R}, \theta \longmapsto \operatorname{arccosh}\left(\frac{\Lambda}{(1-\theta)\sin(\alpha)}\right)$$

und der Minimalstelle

$$\theta^* = \operatorname*{arg\,min}_{\theta\in(0,1)} \left(\epsilon \exp\left(\tfrac{2\pi K(1-\theta)}{a(\theta)}\right) + \exp\left(\tfrac{-2\pi K\theta}{a(\theta)}\right)\right) \tag{5.43}$$

definiert. Dabei sei ϵ die Genauigkeit der Auswertungen der Laplace-Transformierten Ψ und der elementaren Rechenoperationen in (5.39).

Dann gibt es positive Konstanten c und C, sodass der Fehler der Approximation (5.39)

$$E_K(t) = \psi(t) - \sum_{l=-K}^{K} w_l \exp(t\lambda_l)\Psi(\lambda_l) \tag{5.44}$$

mit den Quadraturknoten und -gewichten aus (5.40) durch

$$\|E_K(t)\| \leq C t^{\nu-1}(\epsilon + \exp(-cK)) \tag{5.45}$$

gleichmäßig in $t \in [t_0, \Lambda t_0]$ abgeschätzt werden kann.

Außerdem ist der folgende Zusatz verfügbar, vgl. [64, Zusatz zu Theorem 2.1]: Falls keine zu-

verlässigen Informationen über die Genauigkeit ϵ vorliegen, können wir die Wahl (5.43) nicht verwenden. In diesem Fall garantiert die Wahl von $\theta^* = 1 - 1/K$ immer noch die Kontrolle von Rundungsfehlern, aber wir erhalten nur eine etwas schwächere Fehlerabschätzung wie $\mathcal{O}(\epsilon + \exp(-cK/\ln(K))$.

5.9 Approximation sektorieller Matrixfunktionen

Wir verwenden nun (5.39), um sektorielle Matrixfunktionen zu approximieren. Dazu sei für die Funktionen φ die transformierte Funktion Φ aus (5.21) gegeben. Für einen sektoriellen Operator A und eine vorgegebene Zeitschrittweite τ können wir dann

$$\varphi(\tau A)v \approx \sum_{l=-K}^{K} w_l \exp(\lambda_l)\Phi(\lambda_l, \tau A)v \tag{5.46}$$

nähern. Für die Matrizen $(\mathbf{A}, \mathbf{M})$ einer Finite-Elemente-Diskretisierung erhalten wir mit (5.22) Sektorialität für $\varphi(\tau\mathbf{M}^{-1}\mathbf{A})\mathbf{M}^{-1}\vec{v}$ mit Konstanten unabhängig von der Ortsdiskretisierung. Mit Theorem 5.1 erhalten wir exponentielle Konvergenz in der Anzahl K von Quadraturpunkten.

Für eine feste Schrittweite τ ist damit alles getan. Theorem 5.1 gibt uns aber die Freiheit, das Argument t der Funktion ψ um die Größenordnung Λ zu variieren. Dies nutzen wir bisher nicht aus, da wir ψ an nur einer festen Stelle – für $t = 1$ – auswerten.

Für exponentielle Runge-Kutta-Verfahren (5.4) oder auch exponentielle Rosenbrock-Verfahren [52] oder etwa den `exp4` aus [47] müssen $\varphi_k(c_i\tau A)$ für verschiedene $c_i \in (0, 1]$ ausgewertet werden. Außerdem ist im Kontext variabler-Schrittweiten-Implementierungen mit Fehlerschätzern damit zu rechnen, dass wir nach einem verworfenen Schritt die Schrittweite reduzieren müssen. Dann müssen dieselben Matrixfunktionen auf dieselben Vektoren, aber mit neuer Schrittweite $\widetilde{\tau}$, die nicht allzuweit von τ entfernt ist, angewendet werden.

Es wäre daher erstrebenswert, den größten Aufwand der Approximation (5.39), die Resolventenapproximation, wiederverwenden zu können.

Dabei müssen wir zwei Arten von Matrixfunktionen unterscheiden. Die einfachere von beiden besteht aus denjenigen für die Mehrschrittverfahren aus [11]:

$$\psi(t) = \varphi(t; \tau A) = \varphi(t \cdot \tau A) = \mathcal{L}^{-1}[\Phi(\bullet, \tau A)](t). \tag{5.47}$$

Diese sind, anders als die φ_k-Funktionen (A.1), nicht über eine Faltung in einem Intervall der Länge eins, sondern der Länge t definiert. In diesem Fall ist nichts weiter zu unternehmen. Die

zweite Art von Matrixfunktionen, die φ_k-Funktionen (A.1), skalieren sich jedoch nach (A.28) als

$$\psi(t) = \varphi_k(t\tau A) = \frac{1}{t^{k-1}}\mathcal{L}^{-1}[\Phi(\bullet, \tau A)](t). \tag{5.48}$$

Hier müssen wir also noch durch den entsprechenden Faktor dividieren und verlieren in der Fehlerschranke einen Faktor von $\Lambda^{k-1} = \Lambda^{\nu-1}$.

Allgemeinere Funktionen φ, bei denen die rationale Funktion R komplizierter als ein einfaches inverses Monom ist, kann man nicht mehr direkt skalieren, denn hier ergibt sich aus (5.21) mit einer Integraltransformation

$$\varphi(t\tau A) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma \exp(t\lambda) R(t\lambda)(\lambda - \tau A)^{-1} \,\mathrm{d}\lambda = „\mathcal{L}^{-1}\left[R(t\cdot\bullet)(\bullet - \tau A)^{-1}\right](t)“. \tag{5.49}$$

Dadurch kann man diese nicht mehr als Laplace-Transformierte von Φ darstellen, da diese wieder explizit t-abhängig wäre. Bei den meisten Funktionen φ handelt es sich aber, wie z.B. bei den γ und $\widetilde{\gamma}$-Funktionen der Mehrschrittverfahren aus [51] und auch den meisten anderen Matrixfunktionen im Kontext von exponentiellen Integratoren, um eine Faltung der Exponentialfunktion mit einem Polynom. Dann lässt sich φ als Linearkombination der φ_k-Funktionen – entweder auf dem Intervall $[0, 1]$ oder auf $[0, t]$ wie in [11] – schreiben. Deshalb kann die Auswertung solcher Matrixfunktionen auf die Auswertung der φ_k-Funktionen zurückgeführt werden.

Um verschiedene Matrixfunktionen mit denselben Resolventen auswerten zu können, müssen wir lediglich die rationale Funktion austauschen und auf den Quadraturpunkten neu auswerten. Dabei ist zu beachten, dass wir den Sektor Σ_δ, um den die durch T aus (5.38) parametrisierte Kurve Γ herumläuft, so groß wählen, dass die Pole aller auftretenden rationalen Funktionen darin enthalten sind. Dadurch verschlechtern wir möglicherweise geringfügig die Fehlerkonstanten einiger der Funktionen, die mit einem kleineren Sektor auskommen. Es ist aber zu erwarten, dass der Fehlerterm, der zum größten Sektor gehört, den Gesamtfehler dominiert.

Diese Eigenschaft wird uns sehr nützlich sein, denn wir werden wieder alle Produkte aus Matrixfunktionen mit demselben Vektor simultan ausrechnen können.

5.10 Kontur-Krylov-Verfahren für sektorielle Matrixfunktionen

Wir benutzen nun die Approximation (5.46), um die Anwendung sektorieller Matrixfunktionen auf Vektoren zu approximieren. Anwendungen der verallgemeinerten Resolventen $(\lambda\mathbf{M} - \tau\mathbf{A})^{-1}$ werden mit einem Krylov-Verfahren, etwa FOM oder GMRES, approximiert. In Algorithmus 3 ist das Verfahren für Linearkombinationen von φ_k-Funktionen zusammengefasst.

Der teuerste Anteil von Algorithmus 3, die Lösung der Gleichungssysteme in den Krylov-Räumen, lässt sich auf triviale Weise, ohne Kommunikation zwischen den Prozessen, parallelisieren. Falls

Algorithmus 3 Kontur-Krylov-Verfahren für sektorielle Matrixfunktionen

procedure MATRIXFUNCTIONCONTOURKRYLOV(A, M, v, τ, c_i, a_{ij}, K, δ, γ, α, d)
  // Approximiere $r_i \approx \sum_{j=0}^{p} a_{ij}\varphi_j(c_i\tau M^{-1}A)M^{-1}v$, $1 \leq i \leq s$
  // K Anzahl der Quadraturpunkte
  // δ, γ aus (5.18)
  // $\gamma > 0$, falls $a_{ij} \neq 0$ für ein $j > 0$, damit der Pol von $1/\lambda^k$ im Sektor liegt
  // α, d mit (5.41)
  $t_0 \leftarrow \min_i c_i$, $\Lambda \leftarrow \max_i c_i/t_0$
  $\theta^* \leftarrow 1 - \frac{1}{K}$
  Bestimme h und μ aus (5.42)
  Berechne Quadraturknoten λ_l und -gewichte w_l aus (5.40)
  if A und M reell **then**
    $w_l \leftarrow 2w_l$, $-K \leq l \leq -1$
    entferne λ_l, w_l, $l \geq 1$
  end if
  $r_i \leftarrow 0$, $t_i \leftarrow c_i/t_0$, $1 \leq i \leq s$
  for all (λ_l, w_l) **do** // parallel
    Approximiere $x_l \approx (\lambda_l M - \tau A)^{-1}v$ mit einem Krylov-Verfahren
    for $j = 1, ..., p$ **do**
      $r_i \leftarrow r_i + a_{ij}w_l\frac{\exp(t_i\lambda_l)}{(t_i\lambda_l)^j}x_l$
    end for
  end for
  return r_i, $1 \leq i \leq s$
end procedure

jeder der Krylov-Prozesse auf einem eigenen Prozessor läuft, ist der zeitliche Aufwand dominiert durch das Aufstellen eines einzigen Krylov-Raumes und wir sind somit in der gleichen Situation, wie bei klassischen Krylov-Verfahren für Matrixfunktionen. Dabei ist im Fall des neuen Algorithmus 3 jeder einzelne Krylov-Schritt deutlich günstiger, da wir kein lineares Gleichungssystem lösen müssen.

Die φ_k-Funktionen gehen nur über die rationale Funktion $R(\lambda) = 1/\lambda^k$ aus (5.19) in Zeile 19 ein. Um Algorithmus 3 für andere Matrixfunktionen zu verwenden, kann diese Funktion durch die entsprechende aus (5.21) ersetzt werden. Dabei ist auf die korrekte Skalierung mit t_i zu achten.

Für reelle Matrizen $\mathbf{A}$ und $\mathbf{M}$ wird in der Abfrage in Zeile 11 die Anzahl der Krylov-Räume auf etwa die Häflte reduziert, vgl. [64]. Für reell-symmetrische Matrizen $\mathbf{A}$ und $\mathbf{M}$ sind die Resolventen komplex-symmetrisch, hierfür lassen sich ebenfalls Lanczos-artige 3-Term-Rekursionen konstruieren, vgl. z.B. [9]. Für hermitesche oder schief-hermitesche Matrizen geht die Struktur verloren.

Wir verwenden die sichere Parameterwahl aus dem Zusatz von Theorem 5.1, da wir die Genauigkeit der Krylov-Approximationen nicht kennen.

Algorithmus 3 verzichtet nicht nur auf die Lösung mit linearen Gleichungssystemen mit der Massematrix, sondern wir können ihn insbesondere auch für nicht-invertierbare Matrizen $\mathbf{M}$ anwenden. Auch Eigenschaften wie positive Definitheit oder Symmetrie werden nicht benötigt.

Bei der Wahl der verwendeten Krylov-Verfahren ist man bis auf die Voraussetzung, dass sie mit gegebenenfalls durch die Verschiebung mit einem komplexen Vielfachen der Massematrix geänderten Symmetrieeigenschaften der verallgemeinerten Resolventen umgehen können müssen, frei. Außerdem können Vorkonditionierungsmethoden, z.B. [77, Kapitel 9], eingesetzt werden, was in den klassischen Krylov-Varianten für Matrixfunktionen nicht so ohne Weiteres möglich ist.

Zum Abschluss wollen wir einige numerische Beispiele vorstellen. Bei allen getesteten Beispielen haben wir die Wahl $\gamma = 0$, $\alpha = 0,7$ und $d = 0,6$ verwendet. Alle anderen Parameter ergeben sich aus Theorem 5.1 unter Verwendung des Zusatzes für die sichere Wahl von θ^*.

5.10.1 2D-Wärmeleitungsgleichung

Beginnen wollen wir mit dem Testbeispiel für das Krylov-Verfahren ohne die Lösung von Masse-Systemen, Gleichung (5.36). Wir wählen dieselbe Ortsdiskretisierung wie in der letzten Variante von Abschnitt 5.7: finite Elemente vierter Ordnung auf einem unstrukturierten Gitter auf dem Einheitsquadrat, vgl. Abbildung 5.3. Als Krylov-Verfahren verwenden wir FOM. Die Ergebnisse finden sich in Abbildung 5.5.

Wir stellen fest, dass ähnlich, wie im Fall des masselösungsfreien Krylov-Verfahrens, Algorithmus 2, der Fehler des neuen Verfahrens ebenfalls zu stagnieren scheint. In diesem Fall ist dies jedoch darauf zurückzuführen, dass wir die ebenfalls eingezeichnete Genauigkeit der Kontur-

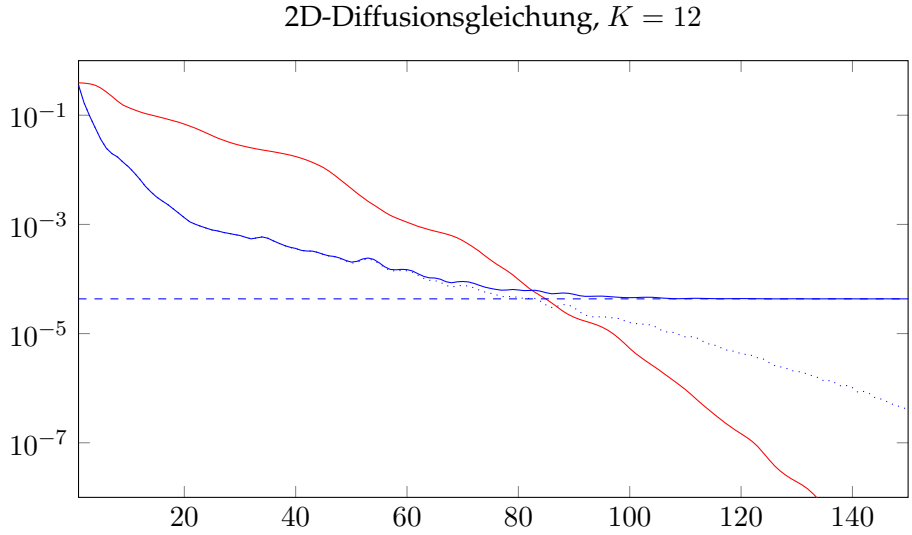


Abbildung 5.5: Testproblem (5.36) mit 2D finiten Elementen der Ordnung vier, $\varphi \equiv \varphi_1$, $\tau = 0,1$, $\gamma = 0$, $\alpha = 0,7$ und $d = 0,6$. x-Achse: Anzahl der Krylov-Schritte. y-Achse: relativer Fehler in der L_2-Norm. Rot, durchgezogen: Fehler der klassischen Approximation (5.26) mit Masseskalarprodukt gemessen gegen die exakte Lösung des ortsdiskretisierten Systems; blau, gestrichelt: Fehler der Kontur gegen selbige; blau, durchgezogen: Fehler des Kontur-Krylov-Verfahrens 3 gegen selbige; blau, gepunktet: Fehler des Kontur-Krylov-Verfahrens gegen die Konturlösung.

Approximation (5.39) erreicht haben. Eine Vergrößerung des Parameters K, der Anzahl von Quadraturpunkten auf der Kurve Γ, verschiebt das Plateau nach unten. Dies illustrieren wir im nächsten Beispiel.

5.10.2 2D-Wärmeleitungsgleichung mit Advektion und transparenten Randbedingungen

Um unser Testproblem etwas interessanter zu gestalten, führen wir einen Advektionsterm ein. Außerdem wollen wir die Gleichung nun auf dem ganzen $\mathbb{R}^2$ lösen.

$$\begin{cases} \partial_t u(t,\vec{x}) = a^2 \Delta u(t,\vec{x}) + \vec{b}^T \nabla u(t,\vec{x}), & \text{in } \mathbb{R}^2, t \geq t_0, \\ u(t,\vec{x}) = 0, & \|\vec{x}\| \longrightarrow \infty, \\ u(t_0,\vec{x}) = u_0(\vec{x}) \end{cases} \tag{5.50}$$

Die Nichtlinearität setzen wir dieses Mal auf Null. Wir müssen also die Exponentialfunktion anstatt von φ_1 auswerten und anwenden. Als Startwert benutzen wir eine Gauß-Verteilung. Die

analytische, also räumlich und zeitlich exakte, Lösung ist dann gegeben als

$$u(t,\vec{x}) = \frac{1}{t}\exp\left(-\frac{1}{4ta^2}\left[\left(x_1 - tb_1\right)^2 + \left(x_2 - tb_2\right)^2\right]\right), \tag{5.51}$$

vgl. [74]. Zur Ortsdiskretisierung wählen wir eine Finite-Elemente-Approximation der Ordnung vier auf $[-4,4]^2$ zusammen mit transparenten Randbedingungen [74]. Die Parameter für die Randbedingungen sind auf $s_0 = -5$ und $N_\xi = 3$ gesetzt. Das verwendete Gitter ist unstrukturiert und hat eine Maschenweite von etwa $h \approx 0,1$.

Mithilfe der analytischen Lösung können wir die Genauigkeit der örtlichen Approximation bestimmen.

Wir wählen $a = 0.5$, $\vec{b} = [1,5;1,5]^T$. Als Krylov-Verfahren verwenden wir FOM. Die Ergebnisse für drei verschiedene Werte der Anzahl der Quadraturpunkte K finden sich in Abbildung 5.6.

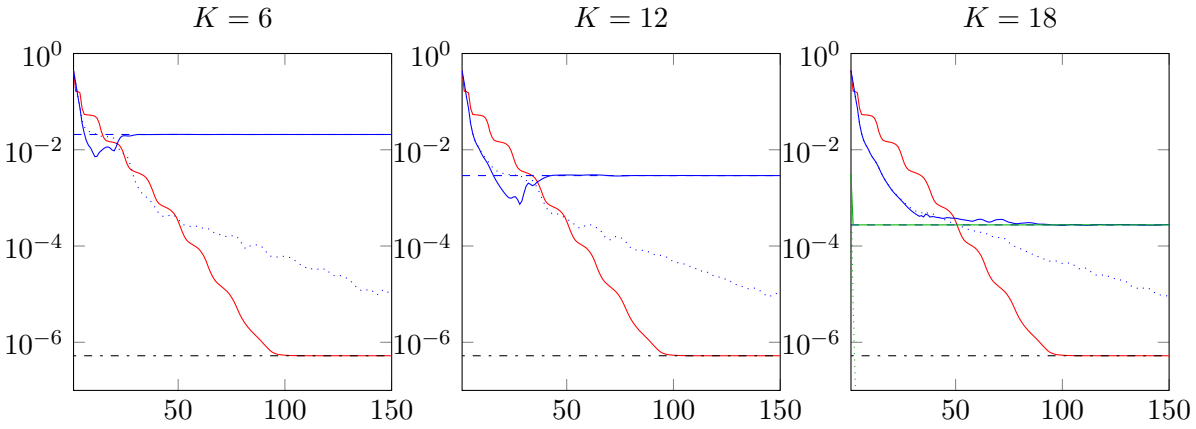


Abbildung 5.6: Testproblem (5.50) mit 2D finiten Elementen der Ordnung vier und transparenten Randbedingungen, $\varphi \equiv \exp$, $\tau = 0,1$, $\gamma = 0$, $\alpha = 0,7$ und $d = 0,6$. x-Achse: Anzahl der Krylov-Schritte. y-Achse: relativer Fehler in der L_2-Norm. Rot, durchgezogen: Fehler der klassischen Approximation (5.26) mit Masseskalarprodukt gemessen gegen die exakte Lösung; blau, gestrichelt: Fehler der Kontur gegen selbige; blau, durchgezogen: Fehler des Kontur-Krylov-Verfahrens, Algorithmus 3, gegen selbige; blau, gepunktet: Fehler des Kontur-Krylov-Verfahrens gegen die Konturlösung; grün, durchgezogen: Fehler des Kontur-Krylov-Verfahrens mit ILU-Vorkonditionierung gegen die exakte Lösung (nur rechts); schwarz, Strichpunkte: Fehler der Ortsdiskretisierung.

Algorithmus 3 konvergiert jeweils bis zur möglichen Genauigkeit der Konturapproximation und jeweis in weniger Schritten, als das klassische Krylov-Verfahren zur entsprechenden Genauigkeit. Allerdings ist zu vermuten, dass dies bei weiterer Erhöhung von K nicht mehr der Fall sein wird, da sich der Fehler zwischen Kontur und Kontur-Krylov für alle K in etwa gleich verhält und abflacht. Die Fehler von Quadratur- und Krylov-Verfahren können hier als mehr oder weniger unabhängig voneinander angesehen werden und einzeln unter gewünschte Toleranzen gedrückt

werden.

Soll der Fehler der Krylov-Approximation schneller reduziert werden, können z.B. Vorkonditionierungstechniken eingesetzt werden. Bei einer einfachen unvollständigen LR-Zerlegung (ILU) mit einer Abschneidetoleranz von 10^{-4} sind die Krylov-Approximationen nach ein bis zwei Schritten schon kleiner, als der Quadraturfehler. Wir haben dies nur für $K = 18$ in die Grafik eingezeichnet, um diese übersichtlich zu halten. Für die beiden anderen Werte ist das Verhalten qualitativ dasselbe. Im nächsten Beispiel werden wir uns diesem Thema etwas ausführlicher widmen.

Die exponentielle Konvergenz der Kontur-Approximation ist gut zu erkennen: Das Hinzufügen einer festen Anzahl von Quadraturpunkten verkleinert den Fehler jeweils um etwa eine Zehnerpotenz.

5.10.3 Eddy-Currents-Modell der Maxwell-Gleichungen in 3D und Vorkonditionierung

Als nächstes wollen wir uns den Maxwell-Gleichungen widmen, die im ersten Teil der Arbeit eine zentrale Rolle spielen. Da diese jedoch hyperbolisch sind, verwenden wir ein Wirbelstrommodell, auch Eddy-Currents-Modell genannt, [3], bei dem Verschiebungsströme vernachlässigt werden und welches, passend zu diesem Kontext, parabolisch ist. In der $\nabla \times \nabla\times$-Formulierung erhalten wir

$$\begin{cases} \partial_t \vec{H}(t,\vec{x}) = -\nabla \times \nabla \times \vec{H}(t,\vec{x}) + \nabla \times \vec{j}(t,\vec{x}) & \text{in } \Omega, \\ (\nabla \times \vec{H}) \times \vec{\nu} = 0 & \text{auf } \partial\Omega, \\ \vec{H}(t_0,\vec{x}) = \vec{H}_0(\vec{x}) \end{cases} \tag{5.52}$$

mit $\Omega = [-1,1]^3$, einem äußeren Strom $\vec{j}$, den wir auf Null setzen. Die Materialparameter sind als $\epsilon = \mu = 1$ gewählt. Als Startwert wählen wir

$$\vec{H}_0(\vec{x}) = \frac{2\alpha}{w_0} \exp\left(-\frac{x_1^2 + x_2^2 + x_3^2}{w_0}\right) \begin{bmatrix} x_2 \\ -x_1 - x_3 \\ x_2 \end{bmatrix} \tag{5.53}$$

mit $\alpha = 5$ und $w_0 = 0.05$, vgl. [44].

Zur Ortsdiskretisierung verwenden wir $H^{\nabla\times}$-reguläre finite Elemente [93] der Ordnung vier auf einem Tetraedergitter, das von Netgen [81] mit einer maximalen Gitterweite von $h = 0.2$ generiert wurde. Das resultierende System gewöhnlicher Differentialgleichungen hat eine Dimension von $N = 147.640$. Als Krylov-Verfahren verwenden wir FOM. Zur Vorkonditionierung verwenden wir eine unvollständige LR-Zerlegung mit einer absoluten Abschneidetoleranz von 10^{-2}. Die Faktoren verbrauchen dabei zwischen 40% und 52% des Speichers der jeweiligen verallgemeinerten Resolventen. Die Ergebnisse für die Anzahl der Quadraturpunkte $K = [4, 6, 8]$ finden sich in

Abbildung 5.7.

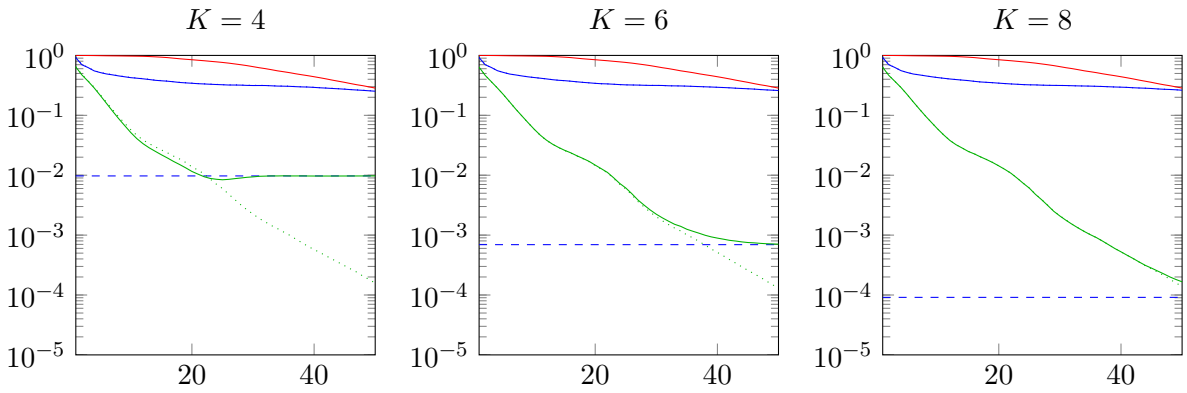


Abbildung 5.7: Testproblem (5.52) mit 3D $H^{\nabla\times}$-konformen finiten Elementen der Ordnung vier und transparenten Randbedingungen, $\varphi \equiv \exp$, $\tau = 0,1$, $\gamma = 0$, $\alpha = 0,7$ und $d = 0,6$. x-Achse: Anzahl der Krylov-Schritte. y-Achse: relativer Fehler in der L_2-Norm. Rot, durchgezogen: Fehler der klassischen Approximation (5.26) mit Masseskalarprodukt gemessen gegen die exakte Lösung des ortsdiskretisierten Systems; blau, gestrichelt: Fehler der Kontur gegen selbige; blau, durchgezogen: Fehler des Kontur-Krylov-Verfahrens, Algorithmus 3, gegen selbige; blau, gepunktet: Fehler des Kontur-Krylov-Verfahrens gegen die Konturlösung; grün, durchgezogen: Fehler des Kontur-Krylov-Verfahrens mit ILU-Vorkonditionierung gegen die exakte Lösung.

Außerdem haben wir in Abbildung 5.8 für $K = 6$ verschiedene einfache Vorkonditionierer miteinander verglichen.

Abbildung 5.7 zeigt uns, dass sowohl das klassische Krylov-Verfahren, als auch das nicht vorkonditionierte Kontur-Verfahren in den ersten 50 Schritten den Fehler nur um eine knappe Zehnerpotenz reduziert. Der Bereich, in dem das klassische Krylov-Verfahren schnell konvergiert, beginnt erst deutlich nach Dimension 50. Das vorkonditionierte Verfahren zeigt – bei recht geringem Mehraufwand an Speicher – jedoch eine deutlich verbesserte Konvergenz. Hier gewinnen wir alle 10 bis 15 Schritte eine weitere Zehnerpotenz und erhalten damit ähnlich gute Konvergenzeigenschaften, wie die reine Konturapproximation.

In Abbildung 5.8 zeigt der Vergleich der Vorkonditionierungsmethoden, dass es sich auch durchaus lohnt, mit sehr einfachen Methoden zu arbeiten. Die Verwendung der Diagonale zur Vorkonditionierung verdoppelt die Konvergenzgeschwindigkeit im Bezug auf das unvorkonditionierte Verfahren, die Verwendung des unteren Dreiecksblocks konvergiert bereits etwa halb so schnell, wie die ILU-Variante.

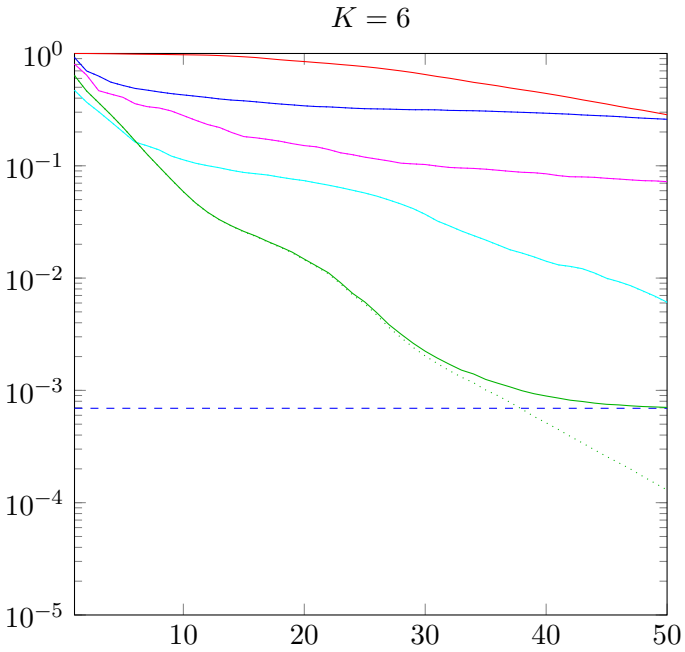


Abbildung 5.8: Gleiche Daten, wie in Abbildung 5.7. Zusätzlich: violett: Fehler des Kontur-Krylov-Verfahrens mit diagonaler Vorkonditionierung gegen exakte Lösung; türkis: Fehler des Kontur-Krylov-Verfahrens mit unterem Dreiecksblock als Vorkonditionierung gegen selbige.

5.10.4 Ausblick: Sobolev-Gleichungen und DAEs

Zum Abschluss wollen wir einen kurzen Ausblick geben, in welche möglichen Richtungen sich das neue Verfahren weiterentwickeln lassen könnte. Wir können es ohne algorithmische Abänderungen auf Gleichungen der Form

$$\text{finde } u \in C^1([t_0, T], \mathcal{X}), \text{ sodass } \quad \begin{cases} M\partial_t u(t) = Au(t) + g(t, u(t)), & t_0 \le t \le T, \\ u(t_0) = u_0, \end{cases}$$

bzw. deren Diskretisierungen anwenden, wobei M ebenfalls ein Differentialoperator ist, solange das Problem parabolisch bleibt. Die verallgemeinerten Resolventen müssen außerhalb eines Sektors in der linken Halbebene überall definiert sein und entsprechend abfallen. In diesem Fall würde die klassische Krylov-Variante, Algorithmus 1, in jedem Krylov-Schritt ein lineares Gleichungssystem mit dem zweiten Differentialoperator lösen müssen, was sehr teuer ist. Hier können wir nicht damit rechnen, mit wenigen Schritten des cg-Verfahrens auszukommen. In Algorithmus 3 müssen dafür Gleichungssysteme von um einen anderen Differentialoperator verschobenen Differentialoperator gelöst werden. Es bleibt zu untersuchen, welche der Varianten, gegebenenfalls auch unter welchen Bedingungen, die bessere ist.

Auch wenn wir für die Herleitung eine invertierbare Matrix $\mathbf{M}$ benötigt haben, wird diese Eigen-

schaft in Algorithmus 3 nicht mehr verwendet. Wir können das Verfahren also rein formal auch auf solche Gleichungen anwenden, bei denen die Massematrix singulär ist. Damit sind wir der Lösung von differential-algebraischen Gleichungen (DAEs) einen Schritt näher gekommen. Von der theoretischen Seite her haben wir allerdings im Moment keine Konvergenzaussagen für diese zur Verfügung.

In [83] findet sich Theorie für Sobolev-Gleichungen. Diese beinhaltet auch inhomogene Gleichungen der Form

$$M\partial_t u(t) = Au(t) + f(t),$$

bei der M ein nicht-invertierbarer Operator sein darf. Beispiel 5.1.1 aus diesem Werk behandelt den Matrixfall und das dort folgende Theorem 5.1.2 liefert eine der Cauchy-Integralformel, kombiniert mit der Variation-der-Konstanten-Formel, ähnliche Darstellung der Lösung. Immer wieder werden dort Darstellungen mit verallgemeinerten Resolventen verwendet. Im Operatorfall könnte es eine wichtige Voraussetzung sein, dass A (M, p)-sektoriell ist.

Solche Darstellungen könnten als Ansatz dienen, um Algorithmus 3 im Fall von Operatoren oder nicht-invertierbaren Matrizen zu analysieren.

KAPITEL 6
FAZIT UND AUSBLICK

Wir haben erfolgreich das Verfahren aus [89] analysiert und verbessert. Dabei konnten wir mit einer rigorosen Fehleranalyse des ursprünglichen Verfahrens zeigen, dass es mit erster Ordnung konvergiert, und eine Verbesserung vorschlagen, um die eigentlich erwartete Konvergenz zweiter Ordnung zu gewährleisten.

Wir haben das numerische Schema dazu als Mehrschrittverfahren für die Komponente, die zum elektrischen Feld gehört, umformuliert und benutzen eine Aussage, die über modulierte Fourier-Entwicklungen bewiesen wird. Dadurch wird eine Anpassung der Startwerte notwendig und die Störung mit einfacher Stabilitätsanalyse kontrolliert. Die Fehlerabschätzung für die anderen Komponenten gelingt unter anderem mit Techniken, wie sie z.B. in [28] zu finden sind. Numerische Tests illustrieren die Resultate.

Bei der Analyse nutzten wir aus, dass im Wesentlichen nur eine hohe zeitliche Frequenz in der Lösung auftritt, die sich aus einer speziellen Wahl des Dichteprofils ergibt. Dies war notwendig, um ein Konvergenzresultat aus [34] nutzen zu können. Offen ist im Moment jedoch, ob nicht auch mehrere hohe Frequenzen zugelassen werden könnten. Der nächste Schritt wäre, die Analyse aus [46] anzupassen, für welche nur allgemeine symmetrische, positv-semidefinite Matrizen $\mathbf{\Omega}$ vorausgesetzt werden. Das Verfahren dort passt allerdings nicht ganz, da der feste Filter $\psi \equiv \mathrm{sinc}^2$ gefordert ist, bei uns jedoch $\psi \equiv \frac{1}{2}(1+\cos)\cdot \mathrm{sinc}^2$ auftritt. Ein numerischer Test lässt vermuten, dass es sich lohnt, in Richtung [46] weiterzuarbeiten.

Im zweiten Teil der Arbeit haben wir ein neues Verfahren vorgeschlagen, mit dem wir die Anwendung von Matrixfunktionen auf Vektoren approximieren können. Dabei haben wir Ideen aus Krylov-Verfahren mit denen von Konturintegrationsmethoden vereint. Die Kombination aus der Fehleranalyse für die auf der Konturintegration basierende inverse Laplace-Transformation mit solcher für klassische Krylov-Verfahren liefert Konvergenz für das gesamte Verfahren. Offen ist im Moment noch eine Fehleranalyse für die Konvergenz der Krylov-Verfahren, die gleichmäßig in der Anzahl der Quadraturpunkte ist. Die numerischen Resultate geben Anlass zur Annahme, dass diese gelingen könnte.

Die Verwendung von Vorkonditionierung der für die Lösung auftretenden linearen Gleichungssysteme wurde untersucht und zeigte eine deutliche Verbesserung der Ergebnisse. Somit können z.B. auch geometrische oder algebraische Mehrgitterverfahren zur Konvergenzbeschleunigung verwendet werden. Generell können nun auch andere bekannte Techniken, wie etwa die Verwendung des letzten Zeitschrittes als Startnäherung, für die Beschleunigung der Krylov-Verfahren eingesetzt werden, was bei klassischen Varianten für Matrixfunktionen oft schwierig ist. Eine Kopplung der Zielgenauigkeit ϵ für die inverse Laplace-Transformation mit Fehlerschätzern für Krylov-Verfahren wäre ebenfalls interessant. Dies könnte sich jedoch als schwierig herausstellen,

da die Wahl der Quadraturpunkte von ϵ abhängt.

Vom rein technischen Standpunkt aus gesehen ist auch die Behandlung von differential-algebraischen Systemen und Sobolev-Gleichungen möglich. Eine Analyse für einen solchen Fall ist bisher nicht vorhanden und könnte ein interessantes Ziel weiterer Untersuchungen darstellen.

APPENDIX A
MATHEMATISCHE GRUNDLAGEN

In Anhang stellen wir einige Grundlagen vor. Die Aussagen sind lange bekannt oder einfach nachzuvollziehen, daher verzichten wir auf Beweise.

A.1 Exponentialfunktion, φ- und trigonometrische Funktionen

Im Zusammenhang mit exponentiellen Integratoren treten die so genannten φ-Funktionen auf, die eng mit der Exponentialfunktion zusammenhängen. Wir wollen diese hier einführen und einige Eigenschaften nachweisen. Wir definieren diese als Faltung skalierter Monome mit der Exponentialfunktion:

$$\varphi_k(z) = \int_0^1 \exp((1-\xi)z)\frac{\xi^{k-1}}{(k-1)!}\,\mathrm{d}\xi, \quad k \geq 1, \qquad \varphi_0(z) := \exp(z). \tag{A.1}$$

Um zu klären, wann und wie sich diese Funktionen auf Operatoren anwenden lassen, müssen wir einige Grundlagen der Halbgruppentheorie zur Verfügung haben. Darauf gehen wir in Abschnitt 5.1 ein. Im Matrixfall können wir die Potenzreihenentwicklung oder Cauchy-Integralformel (A.22) verwenden, um das Matrixexponential zu erklären, und können dann das Integral auswerten. Es gilt die Rekursionsformel

$$\varphi_{k+1}(z) = \frac{\varphi_k(z) - \varphi_k(0)}{z}, \qquad k \geq 0, z \neq 0, \tag{A.2}$$

die zu

$$\varphi_k(z) = \frac{\varphi_0(z) - \sum_{l=0}^{k-1}\varphi_l(0)z^l}{z^{k-1}} = \frac{\exp(z) - \sum_{l=0}^{k-1}\frac{z^l}{l!}}{z^{k-1}}, \qquad k \geq 0, z \neq 0 \tag{A.3}$$

führt. Die φ_k-Funktionen haben die Potenzreihenentwicklung

$$\varphi_k(z) := \sum_{l\geq 0}\frac{z^l}{(k+l)!}, \qquad \text{also} \qquad \varphi_k(0) = \frac{1}{k!}. \tag{A.4}$$

Der Konvergenzradius ist ∞ und damit sind die φ_k-Funktionen ganze Funktionen. Die Unstetigkeit von (A.2) in $z = 0$ ist also hebbar. Die Rekursionsformel geben wir auch zusätzlich in der für alle z anwendbaren Form

$$z\varphi_{k+1}(z) = \varphi_k(z) - \varphi_k(0), \qquad k \geq 0 \tag{A.5}$$

an. Diese hat den Vorteil, dass sie, im Gegensatz zu (A.2), auch mit singulären Matrizen oder Operatoren als Argument anwendbar ist.

Für die trigonometrischen Funktionen $\sin$ und $\cos$, die eng mit der Exponentialfunktion zusam-

menhängen, können wir ähnliche Funktionen definieren. Wir beschränken uns hier auf einige Spezialfälle. Allgemeinere Definitionen für beliebige Monome wie in (A.1) sind analog möglich. Die bekannteste und für uns wichtigste dieser Funktionen ist die sinc-Funktion, manchmal auch si-Funktion genannt. Sie ist gegeben als

$$\mathrm{sinc}(z) := \int_0^1 \cos((1-\xi)z)\,\mathrm{d}\xi \tag{A.6}$$

$$= \begin{cases} \frac{\sin(z)}{z}, & \text{falls } z \neq 0, \\ 1, & \text{sonst.} \end{cases} \tag{A.7}$$

Es ist zu beachten, dass diese Funktion auch oft mit einem um π skalierten Argument definiert wird. Die Standard-Implementierung aus `MATLAB` etwa verwendet diese. Da $\sin$ und $1/z$ ungerade sind, ist sinc gerade. Aus der Potenzreihe der Sinusfunktion ergibt sich direkt

$$\mathrm{sinc}(z) = \sum_{l\geq 0} (-1)^l \frac{z^{2l}}{(2l+1)!} \tag{A.8}$$

mit Konvergenzradius ∞. Auch diese Funktion ist eine ganze Funktion und die Unstetigkeit (A.7) ist hebbar. Darüber hinaus geben wir auch die für alle z – und damit auch für allgemeine Matrizen – verwendbare Version

$$z\,\mathrm{sinc}(z) = \sin(z) \qquad \forall\, z \in \mathbb{C} \tag{A.9}$$

von (A.7) an. Die nächsthöhere, ebenfalls gerade ganze Funktion bezeichnen wir als

$$\mathrm{cosc}(z) = \int_0^1 \cos((1-\xi)z)\xi\,\mathrm{d}\xi \tag{A.10}$$

$$= \begin{cases} \frac{1-\cos(z)}{z^2}, & \text{falls } z \neq 0, \\ \frac{1}{2}, & \text{sonst} \end{cases} \tag{A.11}$$

mit der Potenzreihe

$$\mathrm{cosc}(z) = \sum_{l\geq 0} (-1)^l \frac{z^{2l}}{(2l+2)!}. \tag{A.12}$$

Wir gehen noch einen Schritt weiter und definieren die nächsthöhere Funktion

$$\mathrm{c}_3(z) = \begin{cases} \frac{1-\mathrm{sinc}(z)}{z^2}, & \text{falls } z \neq 0, \\ \frac{1}{6}, & \text{sonst} \end{cases} \tag{A.13}$$

mit der Potenzreihe

$$\mathrm{c}_3(z) = \sum_{l\geq 0} (-1)^l \frac{z^{2l}}{(2l+3)!}. \tag{A.14}$$

Für alle $z \in \mathbb{C}$ gelten die Identitäten

$$z^2 \operatorname{cosc}(z) = 1 - \cos(z), \tag{A.15}$$

$$z^2 \operatorname{c}_3(z) = 1 - \operatorname{sinc}(z), \tag{A.16}$$

$$\operatorname{cosc}(2z) = \frac{1}{2}\operatorname{sinc}^2(z), \tag{A.17}$$

$$\int_0^1 \cos(\xi z)\,\mathrm{d}\xi = \int_0^1 \cos((\xi-1)z)\,\mathrm{d}\xi = \operatorname{sinc}(z), \tag{A.18}$$

und

$$\int_0^1 \xi \operatorname{sinc}(\xi z)\,\mathrm{d}\xi = -\int_0^1 (\xi-1)\operatorname{sinc}((\xi-1)z)\,\mathrm{d}\xi = \int_0^1 (1-\xi)\operatorname{sinc}((1-\xi)z)\,\mathrm{d}\xi = \operatorname{cosc}(z) \tag{A.19}$$

und die Abschätzungen

$$|\operatorname{sinc}(z)| \le 1, \qquad |\operatorname{cosc}(z)| \le \tfrac{1}{2} \qquad |\operatorname{c}_3(z)| \le \tfrac{1}{6} \qquad \forall\, z \in \mathbb{R}. \tag{A.20}$$

A.2 Cauchy-Integralformel und Matrixfunktionen

Exponentielle Integratoren zeichnen sich dadurch aus, dass wir die Exponentialfunktion, die im vorigen Abschnitt definierten φ_k-Funktionen und die damit verwandten trigonometrischen Funktionen auf Operatoren bzw. deren endlichdimensionale Diskretisierungen auswerten, um gewisse lineare Anteile der zu lösenden Differentialgleichung exakt zu bestimmen. Über Halbgruppentheorie werden wir im Operatorfall herausfinden, was dies zu bedeuten hat. Im Matrix-Fall sprechen wir von einer *Matrixfunktion*. Für diesen gibt es eine alternative Definition, die mit dem halbgruppentheoretischen Ansatz zusammenfällt. Er basiert auf der Cauchy-Integralformel, welche wir hier angeben, vgl etwa [1]:

Sei $D \subseteq \mathbb{C}$ ein Gebiet und $\varphi : D \longrightarrow \mathbb{C}$ holomorph. Sei Γ eine geschlossene Kurve in D und $n(\Gamma, z)$ die Windungszahl von Γ um $z \in D$. Dann gilt

$$n(\Gamma, z)\varphi(z) = \frac{1}{2\pi\,\mathring{\mathrm{i}}} \int_\Gamma \frac{\varphi(\lambda)}{\lambda - z}\,\mathrm{d}\lambda \qquad \forall\, z \in D \setminus \Gamma. \tag{A.21}$$

Wir benutzen diese Formel, um zu erklären, was wir unter Anwendung einer Matrixfunktion auf einen Vektor verstehen wollen, vgl. etwa [15]. Es gilt

$$\varphi(\mathbf{A})v = \frac{1}{2\pi\,\mathring{\mathrm{i}}} \int_\Gamma \varphi(\lambda)(\lambda\,\mathbf{Id} - \mathbf{A})^{-1} v\,\mathrm{d}\lambda = \frac{1}{2\pi\,\mathring{\mathrm{i}}} \int_\Gamma \varphi(\lambda) R(\lambda, \mathbf{A}) v\,\mathrm{d}\lambda \tag{A.22}$$

für eine positiv orientierte Kurve Γ, die den *Wertebereich* der Matrix

$$\mathcal{F}(\mathbf{A}) := \left\{ \frac{x^H \mathbf{A} x}{x^H x} \middle| \, x \neq 0 \right\} \tag{A.23}$$

so umschließt, dass $n(\Gamma, \rho) = 1$ für alle $\rho \in \mathcal{F}(\mathbf{A})$ ist. Jeder Punkt des Wertebereichs wird also genau einmal von Γ umlaufen. Da $\sigma(\mathbf{A})$[1]$\subseteq \mathcal{F}(\mathbf{A})$, trifft $\Gamma \subseteq \rho(\mathbf{A})$[2] nirgends einen Eigenwert und die Resolvente ist für alle $\lambda \in \Gamma$ definiert.

Falls $\varphi(z) = \sum_{l \geq 0} a_l z^l$ analytisch mit Konvergenzradius r ist, kann man alternativ auch

$$\varphi(\mathbf{A}) = \sum_{l \geq 0} a_l \mathbf{A}^l = \sum_{l \geq 0} a_l \mathbf{X}^{-1} \mathbf{D}^l \mathbf{X} = \mathbf{X}^{-1} \varphi(\mathbf{D}) \mathbf{X} \tag{A.24}$$

definieren. Wegen $\|\mathbf{A}^l\| \leq \|\mathbf{A}\|^l$ für eine submultiplikative Matrixnorm konvergiert diese Reihe, solange $\|\mathbf{A}\| < r$ in einer beliebigen submultiplikativen Matrixnorm, da im Endlichdimensionalen alle Normen äquivalent sind. Die mittlere Gleichung gilt dabei, falls $\mathbf{A} = \mathbf{X}^{-1}\mathbf{D}\mathbf{X}$ eine Zerlegung der Matrix ist. Die letzte Gleichung gilt nur, solange auch $\|\mathbf{D}\| < r$.

A.3 Laplace-Transformation

Die Laplace-Transformation $\mathcal{L}$ ist ein linearer Operator, der gegeben ist als

$$\mathcal{L} := f \longmapsto \mathcal{L}[f] = F = \left(\lambda \longmapsto \int_0^\infty \exp(-\lambda z) f(z) \, \mathrm{d}z \right), \tag{A.25}$$

wobei λ so gewählt sei, dass das Integral existiert. Die Menge aller solcher Punkte λ heißt Konvergenzbereich von f bzw. F. Die Inverse $\mathcal{L}^{-1}$ dazu ist gegeben über das so genannte *Bromwich-Integral* als

$$\mathcal{L}^{-1} := F \longmapsto \mathcal{L}^{-1}[F] = f = \left(z \longmapsto \frac{1}{2\pi \mathring{\mathrm{i}}} \int_\Gamma \exp(z\lambda) F(\lambda) \, \mathrm{d}\lambda \right) \tag{A.26}$$

mit einer Kurve Γ, die eine Parallele zur imaginären Achse ist und durch den Konvergenzbereich von f bzw. F verläuft. Für die φ_k-Funktionen (A.1) erhalten wir die Darstellung

$$\varphi_k(z) = \mathcal{L}^{-1} \left[\mathcal{L}[\exp(z \cdot \bullet)] \mathcal{L} \left[\frac{(\bullet)^{k-1}}{(k-1)!} \right] \right] (1) = \mathcal{L}^{-1} \left[\frac{1}{(\bullet)^k (\bullet - z)} \right] (1), \tag{A.27}$$

für $z \in \mathbb{C}$, vgl. [64, (3.10),(3.11)], wobei die Faltung in (A.1) durch Laplace-Transformation zu einer Multiplikation wird. Die Darstellung (A.27) gilt, ohne den Zwischenschritt, auch für die Exponentialfunktion φ_0. Wir können das Argument skalieren, indem wir eine einfache Umskalierung der

[1] $\sigma(\mathbf{A}) = \{\lambda \in \mathbb{C} \mid \lambda \, \mathbf{Id} - A \text{ ist nicht Invertierbar}\}$ bezeichnet das Spektrum von $\mathbf{A}$
[2] $\rho(\mathbf{A}) = \mathbb{C} \setminus \sigma(\mathbf{A})$ bezeichnet die Resolventenmenge von $\mathbf{A}$

Kurve vornehmen, und erhalten

$$\varphi_k(tz) = \frac{1}{t^{k-1}} \mathcal{L}^{-1} \left[\frac{1}{(\bullet)^k} \frac{1}{(\bullet - z)} \right] (t). \tag{A.28}$$

A.4 Allgemeine Rechenregeln

Zum Abschluss tragen wir einige einfache Rechenregeln zusammen. Für $\vec{F} \in C^2(\mathbb{R}^3, \mathbb{R}^3)$ gilt

$$\nabla \cdot \nabla \times \vec{F} = 0. \tag{A.29}$$

Für $\vec{F} = \vec{E} \times \vec{B}$ gilt

$$\nabla \cdot \vec{F} = \left\langle \nabla \times \vec{B}, \vec{E} \right\rangle - \left\langle \nabla \times \vec{E}, \vec{B} \right\rangle. \tag{A.30}$$

Für $\Psi \in C^1(\mathbb{R}^3, \mathbb{R})$ und $V \in C^2(\mathbb{R}^3, \mathbb{R}^3)$ erhalten wir dann

$$\nabla \cdot (\Psi \, \nabla \times V) = \left\langle \nabla \Psi, \nabla \times V \right\rangle + \Psi \underbrace{(\nabla \cdot \nabla \times V)}_{=0} = \left\langle \nabla \Psi, \nabla \times V \right\rangle. \tag{A.31}$$

Partielle Summation

$$\sum_{l=0}^{n-1} (f_{l+1} - f_l) g_l = \sum_{l=0}^{n-1} f_l (g_{l-1} - g_l) + f_n g_{n-1} - f_0 g_{-1}. \tag{A.32}$$

LITERATURVERZEICHNIS

[1] Lars Valerian Ahlfors. *Complex analysis: an introduction to the theory of analytic functions of one complex variable*. International series in pure and applied mathematics. McGraw-Hill, 1979.
(Referenced in: A.2.)

[2] Awad H. Al-Mohy and Nicholas John Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM Journal on Scientific Computing*, 33(2):488–511, 2011.
(Referenced in: 5.1.)

[3] Habib Ammari, Anbalisa Buffa, and Jean-Claude Nédélec. A justification of eddy currents model for the maxwell equations. *SIAM Journal on Applied Mathematics*, 60(5):1805–1823, May 2000.
(Referenced in: 5.10.3.)

[4] Wolfgang Arendt and Tom A.F.M. ter Elst. From forms to semigroups. In Wolfgang Arendt, Joseph A. Ball, Jussi Behrndt, Karl-Heinz Förster, Volker Mehrmann, and Carsten Trunk, editors, *Spectral Theory, Mathematical System Theory, Evolution Equations, Differential and Difference Equations*, volume 221 of *Operator Theory: Advances and Applications*, pages 47–69. Springer Basel, 2012.
(Referenced in: 5.2 and 5.4.)

[5] Wolfgang Arendt and Jürgen Voigt. *Form methods for evolution equations, and applications*. Internet Seminar on Evolution Equations. 2014.
(Referenced in: 5.2 and 5.4.)

[6] Alain Bossavit. *Computational electromagnetism: variational formulations, complementarity, edge elements*. Academic Press, 1998.
(Referenced in: 2.2.)

[7] Dietrich Braess. *Finite Elemente - Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie, 5.* Springer, 2013.
(Referenced in: 5.2.)

[8] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, 2002.
(Referenced in: 5.2.)

[9] Angelika Bunse-Gerstner and Ronald Stöver. On a conjugate gradient-type method for solving complex symmetric linear systems. *Linear Algebra and its Applications*, 287(1–3):105 – 123,

1999.
(Referenced in: 5.10.)

[10] Marco Caliari, Marco Vianello, and Luca Bergamaschi. Interpolating discrete advection-diffusion propagators at leja sequences. *Journal of Computational and Applied Mathematics*, 172(1):79 – 99, 2004.
(Referenced in: 5.1.)

[11] Mari Paz Calvo and César Palencia. A class of explicit multistep exponential integrators for semilinear problems. *Numerische Mathematik*, 2006.
(Referenced in: 5.4, 5.9, and 5.9.)

[12] David Cohen, Ernst Hairer, and Christian Lubich. Modulated fourier expansions of highly oscillatory differential equations. *Foundations of Computational Mathematics*, 3(4):327–345, 2003.
(Referenced in: 4.)

[13] David Cohen, Ernst Hairer, and Christian Lubich. Conservation of energy, momentum and actions in numerical discretizations of nonlinear wave equations. *Numerische Mathematik*, 110(2):113–143, 2008.
(Referenced in: 4.)

[14] David Cohen, Ernst Hairer, and Christian Lubich. Long-time analysis of nonlinearly perturbed wave equations via modulated Fourier expansions. *Archive for Rational Mechanics and Analysis*, 187(2):341–368, 2008.
(Referenced in: 4.)

[15] Nelson Dunford and Jacob Theodore Schwartz. *Linear Operators. Part I: General Theory. Reprint of the 1958 original.* John Wiley & Sons, New York, 1988.
(Referenced in: A.2.)

[16] Klaus-Jochen Engel and Rainer Nagel. *One-Parameter Semigroups for Linear Evolution Equations*. Springer, 2000.
(Referenced in: 5 and 5.1.)

[17] Azeddine Essai. Weighted FOM and GMRES for solving nonsymmetric linear systems. *Numerical Algorithms*, 18:277–292, 1998.
(Referenced in: 5.6.)

[18] Erwan Faou, Ludwig Gauckler, and Christian Lubich. Plane wave stability of the split-step fourier method for the nonlinear Schrödinger equation. *Forum of Mathematics, Sigma*, 2, 8 2014.
(Referenced in: 4.)

[19] B. García-Archilla, J. Sanz-Serna, and R. Skeel. Long-time-step methods for oscillatory differential equations. *SIAM Journal on Scientific Computing*, 20(3):930–963, 1998.
(Referenced in: 3 and 3.4.)

[20] Ludwig Gauckler. Error analysis of trigonometric integrators for semilinear wave equations. *ArXiv e-prints*, jul 2014.
(Referenced in: 4.)

[21] Ludwig Gauckler, Ernst Hairer, Christian Lubich, and Daniel Weiss. Metastable energy strata in weakly nonlinear wave equations. *Communications in Partial Differential Equations*, 37(8):1391–1413, 2012.
(Referenced in: 4.)

[22] Ludwig Gauckler and Christian Lubich. Nonlinear Schrödinger equations and their spectral semi-discretizations over long times. *Foundations of Computational Mathematics*, 10(2):141–169, 2010.
(Referenced in: 4.)

[23] Ludwig Gauckler and Christian Lubich. Splitting integrators for nonlinear Schrödinger equations over long times. *Foundations of Computational Mathematics*, 10(3):275–302, 2010.
(Referenced in: 4.)

[24] Ivan P. Gavrilyuk and Vologymyr Makarov. Exponentially convergent algorithms for the operator exponential with applications to inhomogeneous problems in banach spaces. *SIAM Journal on Numerical Analysis*, 43(5):2144–2171, 2005.
(Referenced in: 5.8.)

[25] Gene Howard Golub and Charles Francis Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996.
(Referenced in: 5.7.)

[26] Volker Grimm. A note on the Gautschi-type method for oscillatory second-order differential equations. *Numerische Mathematik*, 102(1):61–66, 2005.
(Referenced in: 4.13 and 4.13.)

[27] Volker Grimm. Resolvent Krylov subspace approximation to operator functions. *BIT Numerical Mathematics*, 52(3):639–659, 2012.
(Referenced in: 5.6.)

[28] Volker Grimm and Marlis Hochbruck. Error analysis of exponential integrators for oscillatory second-order differential equations. *Journal of Physics A: Mathematical and General*, 39:5495–5507, 2006.
(Referenced in: 3, 3.4, 3.5, 4, 4.9, 4.12, and 6.)

[29] Stefan Güttel and Jennifer Pestana. Some observations on weighted GMRES. *Numerical Algorithms*, pages 1–20, 2014.
(Referenced in: 5.6.)

[30] Ernst Hairer and Christian Lubich. Long-time energy conservation of numerical methods for oscillatory differential equations. *SIAM Journal on Numerical Analysis*, 38(2):414–441, jul 2000.
(Referenced in: 4 and 4.12.)

[31] Ernst Hairer and Christian Lubich. Modulated Fourier expansions for continuous and discrete oscillatory systems. In *Foundations of Computational Mathematics, Budapest 2011*, pages 113–128. Cambridge University Press, 2012. Cambridge Books Online.
(Referenced in: 4.)

[32] Ernst Hairer and Christian Lubich. On the energy distribution in Fermi-Pasta-Ulam lattices. *Archive for Rational Mechanics and Analysis*, 205(3):993–1029, 2012.
(Referenced in: 4.)

[33] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the störmer-verlet method. *Acta Numerica*, 12:399–450, 5 2003.
(Referenced in: 2.4.)

[34] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration - Structure-preserving algorithms for ordinary differential equations*. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, Heidelberg, 2nd ed. edition, 2006.
(Referenced in: 3, 3.2, 3.4, 3.5, 4, 4, 4.2, 4.1, 4.3, 4.12, 4.21, and 6.)

[35] Ernst Hairer, Syvert Paul Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I (2Nd Revised. Ed.): Nonstiff Problems*. Springer-Verlag New York, Inc., New York, NY, USA, 1993.
(Referenced in: 3.2.)

[36] M. Hegelich, S. Karsch, G. Pretzler, D. Habs, K. Witte, W. Guenther, M. Allen, A. Blazevic, J. Fuchs, J. C. Gauthier, M. Geissel, P. Audebert, T. Cowan, and M. Roth. Mev ion jets from short-pulse-laser interaction with thin foils. *Physical Review Letters*, 89:085002, Aug 2002.
(Referenced in: 2.1.)

[37] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.
(Referenced in: 5.3.)

[38] Jan Hesthaven and Timothy Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Texts in Applied Mathematics. Springer, New York, 2008.
(Referenced in: 5.2.)

[39] Nicholas John Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26:1179–1193, 2005.
(Referenced in: 5.1.)

[40] Nicholas John Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.
(Referenced in: 5.1 and 5.1.)

[41] Nicholas John Higham and Awad H. Al-Mohy. Computing matrix functions. *Acta Numerica*, 19:159–208, 2010.
(Referenced in: 5.1.)

[42] David Hipp, Marlis Hochbruck, and Alexander Ostermann. An exponential integrator for non-autonomous parabolic problems. *ETNA*, 2014.
(Referenced in: 5.6.)

[43] Ralf Hiptmair, Peter Robert Kotiuga, and Sébastien Tordeux. Self-adjoint curl operators. *Annali di Matematica Pura ed Applicata*, 191(3):431–457, 2012.
(Referenced in: 2.5.)

[44] Ralf Hiptmair and Achim Schädle. Non-reflecting boundary conditions for maxwell's equations. *Computing*, 71(3):265–292, 2003.
(Referenced in: 5.10.3.)

[45] Marlis Hochbruck and Christian Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 34(5):1911–1925, 1997.
(Referenced in: 5.1, 5.6, and 5.7.)

[46] Marlis Hochbruck and Christian Lubich. A Gautschi-type method for oscillatory second-order differential equations. *Numerische Mathematik*, 83(3):403–426, 1999.
(Referenced in: 3, 3.4, 4.13, 4.13, and 6.)

[47] Marlis Hochbruck, Christian Lubich, and Hubert Selhofer. Exponential integrators for large systems of differential equations. *SIAM Journal on Scientific Computing*, 19(5):1552–1574, 1998.
(Referenced in: 5.1, 5.5, and 5.9.)

[48] Marlis Hochbruck and Alexander Ostermann. Explicit exponential Runge–Kutta methods for semilinear parabolic problems. *SIAM Journal on Numerical Analysis*, 43(3):1069–1090, 2005.
(Referenced in: 5.1.)

[49] Marlis Hochbruck and Alexander Ostermann. Exponential Runge–Kutta methods for parabolic problems. *Applied Numerical Mathematics*, 53(2–4):323–339, 2005.
(Referenced in: 5.1.)

[50] Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 2010.
(Referenced in: 5.1.)

[51] Marlis Hochbruck and Alexander Ostermann. Exponential multistep methods of Adams-type. *BIT Numerical Mathematics*, 51(4):889–908, 2011.
(Referenced in: 5.4 and 5.9.)

[52] Marlis Hochbruck, Alexander Ostermann, and Julia Schweitzer. Exponential Rosenbrock-type methods. *SIAM Journal on Numerical Analysis*, 47:786–803, 2009.
(Referenced in: 5.9.)

[53] Marlis Hochbruck, Tomislav Pažur, Andreas Schulz, Ekkachai Thawinan, and Christian Wieners. Efficient time integration for discontinuous Galerkin approximations of linear wave equations. *Zeitschrift für Angewandte Mathematik und Mechanik*, 2014.
(Referenced in: 5.6.)

[54] Helge Holden, Christian Lubich, and Nils Henrik Risebro. Operator splitting for partial differential equations with Burgers nonlinearity. *Mathematics of Computation*, 82(281), 2013.
(Referenced in: 5.1.)

[55] Georg Jansing. EXPODE - advanced exponential time integration toolbox for MATLAB, code documentation. *arXiv:1108.2655*, 2011.
(Referenced in: 5.6.)

[56] Georg Jansing. EXPODE - advanced exponential time integration toolbox for MATLAB. *arXiv:1404.4580*, 2014.
(Referenced in: 5.1 and 5.6.)

[57] M. Kaluza, J. Schreiber, M. I. K. Santala, G. D. Tsakiris, K. Eidmann, J. Meyer-ter Vehn, and K. J. Witte. Influence of the laser prepulse on proton acceleration in thin-foil experiments. *Physical Review Letters*, 93:045003, Jul 2004.
(Referenced in: 2.1.)

[58] Florian Kleen. Discontinuous Galerkin Verfahren für Maxwellgleichungen. Masterarbeit, Heinrich-Heine Universität Düsseldorf, 2013.
(Referenced in: 5.6.)

[59] Jalo Liljo. *Hybride Verfahren zur Simulation der Wechselwirkung relativistischer Kurzpuls-Laser mit hochdichten Plasmen*. Dissertation, Heinrich-Heine Universität Düsseldorf, 2010.
(Referenced in: 2.1, 2.1, 3, 3.2, 3.2, 3.2, 4, and 4.1.)

[60] Jalo Liljo, Anupam Karmakar, Alexander Pukhov, and Marlis Hochbruck. One-dimensional electromagnetic relativistic PIC-hydrodynamic hybrid simulation code H-VLPL (hybrid virtual laser plasma lab). *Computer Physics Communications*, 179(6):371 – 379, 2008.
(Referenced in: 2.1 and 2.1.)

[61] Christian Lubich. Persönliche Kommunikation.
(Referenced in: 4.3.)

[62] Christian Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. European Mathematical Society, Zürich, Switzerland, 2008.
(Referenced in: 5.1.)

[63] Christian Lubich and Alexander Ostermann. Linearly implicit time discretization of non-linear parabolic equations. *IMA Journal of Numerical Analysis*, 15(4):555–583, 1995.
(Referenced in: 5.4.)

[64] María López-Fernández. A quadrature based method for evaluating exponential-type functions for exponential methods. *BIT Numerical Mathematics*, 50(3):631–655, 2010.
(Referenced in: 5.1, 5.4, 5.4, 5.7, 5.8, 5.8, 5.1, 5.10, and A.3.)

[65] María López-Fernández, César Palencia, and Achim Schädle. A spectral order method for inverting sectorial laplace transforms. *SIAM Journal on Numerical Analysis*, 44(3):1332–1350, 2006.
(Referenced in: 5.4, 5.7, 5.8, and 5.8.)

[66] Peter Monk. *Finite element methods for Maxwell's equations*. Clarendon Press, Oxford, 2003.
(Referenced in: 2.2 and 2.2.)

[67] Jean-Claude Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numerische Mathematik*, 35:315–341, 1980.
(Referenced in: 5.2.)

[68] Jean-Claude Nédélec. A new family of mixed finite elements in $\mathbb{R}^3$. *Numerische Mathematik*, 50(1):57–81, November 1986.
(Referenced in: 5.2.)

[69] Pradip Niyogi, Sunil Kumar Chakrabartty, and Manas Kumar Laha. *Introduction to Computational Fluid Dynamics*. Pearson, 2002.
(Referenced in: 2.5.)

[70] Christopher C. Paige and Michael A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.
(Referenced in: 5.3.)

[71] Amnon Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer, 1983.
(Referenced in: 5.)

[72] Alexander Pukhov. Three-dimensional electromagnetic relativistic particle-in-cell code VLPL (virtual laser plasma lab). *Journal of Plasma Physics*, 61:425–433, April 1999.
(Referenced in: 2.1.)

[73] Lynne Robson, P. T. Simpson, Robert J. Clarke, Kenneth W. D. Ledingham, Filip Lindau, Olle Lundh, Tom McCanny, Patrick Mora, David Neely, Claes-Göran Wahlstrom, Matthew Zepf, and Paul McKenna. Scaling of proton acceleration driven by petawatt-laser-plasma interactions. *Nature Physics*, 3(1):58–62, January 2007.
(Referenced in: 2.1.)

[74] Daniel Ruprecht, Achim Schädle, and Frank Schmidt. Transparent boundary conditions based on the pole condition for time-dependent, two-dimensional problems. *Numerical Methods for Partial Differential Equations*, 29(4):1367–1390, 2013.
(Referenced in: 5.10.2.)

[75] Yousef Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of Computation*, 37(155):105–126, 1981.
(Referenced in: 5.3.)

[76] Yousef Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 29(1):209–228, 1992.
(Referenced in: 5.1, 5.6, and 5.7.)

[77] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2nd edition, 2003.
(Referenced in: 5.3 and 5.10.)

[78] Yousef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.
(Referenced in: 5.3.)

[79] Maximilian Schneiders. Krylovraumerfahren zur Approximation der Exponentialfunktion. Bachelorarbeit, Heinrich-Heine Universität Düsseldorf, September 2013.
(Referenced in: 5.6.)

[80] Achim Schädle, María López-Fernández, and Christian Lubich. Fast and oblivious convolution quadrature. *SIAM Journal on Scientific Computing*, 28(2):421–438, 2006.
(Referenced in: 5.1 and 5.8.)

[81] Joachim Schöberl. Netgen – an advancing front 2d/3d-mesh generator based on abstract rules. *Computing and Visualization in Science*, 1(1):41–52, 1997.
(Referenced in: 5.10.3.)

[82] Chi-Wang Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservatio n laws. In *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, volume 1697 of *Lecture Notes in Mathematics*, chapter 4, pages 325–432. Springer Berlin / Heidelberg, 1998. 10.1007/BFb0096355.
(Referenced in: 2.1.)

[83] Georgy A. Sviridyuk and Vladimir E. Fedorov. *Linear Sobolev Type Equations and Degenerate Semigroups of Operators*. De Gruyter, Berlin, Boston, 2003.
(Referenced in: 5.10.4.)

[84] Max Tabak, James Hammer, Michael E. Glinsky, William L. Kruer, Scott C. Wilks, John Woodworth, E. Michael Campbell, Michael D. Perry, and Rodney J. Mason. Ignition and high gain with ultrapowerful lasers. *Physics of Plasmas*, 1(5):1626–1634, 1994.
(Referenced in: 2.1.)

[85] A. Talbot. The accurate numerical inversion of Laplace transforms. *IMA Journal of Applied Mathematics*, 23(1):97–120, 1979.
(Referenced in: 5.4 and 5.8.)

[86] Lloyd Nicholas Trefethen, J. André C. Weideman, and Thomas Schmelzer. Talbot quadratures and rational approximations. *BIT Numerical Mathematics*, 46(3):653–670, 2006.
(Referenced in: 5.8.)

[87] Tobias Tückmantel. Persönliche Kommunikation.
(Referenced in: 2.3.)

[88] Tobias Tückmantel. *Hybrid particle-in-cell simulations of relativistic plasmas*. Dissertation, Heinrich-Heine Universität Düsseldorf, 2013.
(Referenced in: 2.1, 2.1, 2.1, and 2.3.)

[89] Tobias Tückmantel, Alexander Pukhov, Jalo Liljo, and Marlis Hochbruck. Three-dimensional relativistic particle-in-cell hybrid code based on an exponential integrator. *IEEE Transactions on Plasma Science*, 38(9):2383–2389, 2010.
(Referenced in: 2.1, 2.1, 2.1, 2.3, 3, 3.2, 3.2, 3.2, 4, 4, 4.1, 4.1, 4.12, 4.12, 4.13, and 6.)

[90] Gerhard Wanner, Ernst Hairer, and Syvert Paul Nørsett. Order stars and stability theorems. *BIT Numerical Mathematics*, 18(4):475–489, 1978.
(Referenced in: 5.1.)

[91] J. André C. Weideman and Lloyd Nicholas Trefethen. Parabolic and hyperbolic contours for computing the bromwich integral. *Mathematics of Computation*, pages 1341–1356, 2007.
(Referenced in: 5.7.)

[92] Kane S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas and Propagation*, pages 302–307, 1966.
(Referenced in: 3.3 and 3.4.)

[93] Sabine Zaglmayr. *High Order Finite Element Methods for Electromagnetic Field Computation*. Dissertation, Johannes Kepler Universität, Linz, 2006.
(Referenced in: 2.2, 5.2, and 5.10.3.)

[94] Steven T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, 1979.

(Referenced in: 2.1.)