ZENTRUM MATHEMATIK
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Lösung der 2D Wellengleichung mittels hierarchischer Matrizen

Michael Lintner

Lehrstuhl für Numerische Mathematik
und Wissenschaftliches Rechnen
Prof. Dr. Folkmar Bornemann

Lösung der 2D Wellengleichung mittels hierarchischer Matrizen

Michael Lintner

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. M. Brokate
Prüfer der Dissertation:

1. Univ.-Prof. Dr. F. Bornemann
2. Univ.-Prof. Dr. H. Faßbender
3. Univ.-Prof. Dr. Dr. h.c. W. Hackbusch
   Christian-Albrechts-Universität zu Kiel
   (schriftliche Beurteilung)

Die Dissertation wurde am 27.06.2002 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 28.11.2002 angenommen.

Abstract

The high-frequency 2D wave equation is solved by the exponential Gautschi method after projection onto a Finite Element space. Thereby, the discretization error is independent from the product of the time step size with the frequencies.

In the course of the computation of the discrete solution, we have to evaluate matrix-vector products with transcendental matrix functions of matrices of very large dimension. Approximating these matrix function-vector products by Krylov subspace methods yields a restriction of the time step size by the wave speed and the spatial mesh size. Therefore, we deploy a method to approximate the matrix-vector products by spectral decomposition of the matrix functions using the hierarchical matrices constructed by Hackbusch.

This leads to a method of almost linear complexity, i.e., linear up to logarithmic factors, for solving the high-frequency 2D wave equation under the assumption of smooth initial data. This method can also be applied to other linear PDEs like the heat equation or Schrödinger's equation.

Zusammenfassung

Die hochfrequente 2D Wellengleichung wird nach Projektion auf Finite Elemente im Ort durch das exponentielle Gautschi-Verfahren gelöst. Der Diskretisierungsfehler ist dabei unabhängig vom Produkt aus der Zeitschrittweite mit den Frequenzen.
Im Zuge der Berechnung der diskreten Lösung sind Matrix-Vektor-Produkte mit transzendenten Matrixfunktionen von Matrizen sehr großer Dimension auszuwerten. Nachdem die Approximation dieser Matrixfunktion-Vektor-Produkte mittels Krylovraummethoden die Beschränkung der Zeitschrittweite durch die Wellengeschwindigkeit und die Gitterbreite im Ort zur Folge hat, wird eine Methode entwickelt, die Matrix-Vektor-Produkte durch Spektralzerlegung der Matrixfunktionen mittels der von Hackbusch konstruierten hierarchischen Matrizen zu nähern.
Dies führt auf ein Verfahren fast linearer (d.h. bis auf logarithmische Terme linearer) Komplexität zur Lösung der hochfrequenten 2D Wellengleichung unter der Voraussetzung glatter Anfangsdaten im Ort. Dieses Verfahren lässt sich auch auf andere lineare PDEs wie die Wärmeleitungsgleichung oder die Schrödingergleichung anwenden.

Inhaltsverzeichnis

Einleitung

Die Modellierung des physikalischen Problems der hochfrequenten Wellenausbreitung in zwei Raumdimensionen führt auf eine hyperbolische Differentialgleichung mit bestimmten Anfangs- und Randbedingungen, die es nach erfolgter Diskretisierung in einem endlichdimensionalen Funktionenraum zu lösen gilt.

Wird die durch Semidiskretisierung im Ort mittels Finiter Elemente erhaltene ODE durch explizite Verfahren gelöst, unterliegt die Zeitschrittweite aus Stabilitätsgründen einer CFL-Bedingung. Ebenso hat die Verwendung impliziter Integrationsmethoden zur Vermeidung großer Phasenfehler die Beschränkung der Zeitschrittweite durch die Wellengeschwindigkeit und die Gitterbreite im Ort zur Folge.

Das primäre Ziel dieser Arbeit, die Entwicklung einer Methode zur Lösung der hochfrequenten 2D Wellengleichung ohne Beschränkung der Zeitschrittweite durch die Wellengeschwindigkeit und die Gitterbreite im Ort, konnte für genügend glatte Anfangsdaten im Ort unter Verwendung von exponentiellen Zeitintegratoren und Zuhilfenahme hierarchischer Matrizen erreicht werden.

Exponentielle Verfahren ermöglichen es, die Zeitschrittweite unabhängig von den Frequenzen zu wählen: Hochbruck und Lubich haben in [29] gezeigt, dass das auf eine Arbeit des Schweizer Mathematikers Gautschi aus dem Jahre 1961 zurückgehende exponentielle Gautschi-Verfahren Fehlerschranken 2. Ordnung besitzt, die unabhängig vom Produkt aus der Zeitschrittweite mit den Frequenzen sind.
Die hochfrequente 2D Wellengleichung wird durch Projektion auf Finite Elemente im Ort in ein Differentialgleichungssystem 2. Ordnung mit einer die hochfrequenten Oszillationen verursachenden symmetrisch positiv definiten Matrix übergeführt. Lösung dieses gewöhnlichen Differentialgleichungssystems mit dem Gautschi-Algorithmus erfordert die Auswertung von Matrixfunktion-Vektor-Produkten mit transzendenten Funktionen von Matrizen sehr großer Dimension. Die Argumente dieser transzendenten Matrixfunktionen sind stets skalare Vielfache der darstellenden Matrix des diskreten Laplace-Operators in einer bestimmten FE-Basis. Durch Anwendung eines Theorems aus [30] konnten wir zeigen, dass die Berechnung dieser Matrix-Vektor-Produkte mittels Krylovraummethoden (siehe [7], [28], [30]) wiederum einen von der Wellengeschwindigkeit und der Gitterbreite im Ort abhängigen Aufwand zum Erreichen einer vorgegebenen Genauigkeit besitzt.

Hackbusch hat in [19] und [20] Klassen sogenannter hierarchischer Matrizen ($\mathcal{H}$-Matrizen) konstruiert, mit deren Hilfe sich bestimmte Operatoren mit fast linearem (d.h. bis auf logarithmische Terme linearem) Aufwand approximieren lassen. Diese Matrizen bestehen nahe der Diagonalen aus vollbesetzten Matrixblöcken, während der Großteil der Außerdiagonalblöcke durch Niedrigrang-Matrizen approximiert wird. In [19] und [20] wurden approximierte

Grundoperationen für $\mathcal{H}$-Matrizen mit fast linearer Komplexität entwickelt. In der vorliegenden Arbeit wurden jetzt mehrere hierarchische Zerlegungen von $\mathcal{H}$-Matrizen mit fast linearer Komplexität samt Rückwärtsanalyse für die Approximationsfehler neu konstruiert. Eine davon, die LDL^T-Zerlegung symmetrisch indefiniter $\mathcal{H}$-Matrizen, erwies sich als überaus geeignet für die rückwärtsstabile Berechnung von Eigenpaaren des diskreten Laplace-Operators: Wir verwenden dazu ein auf Stewart ([43]) zurückgehendes simultanes Iterationsverfahren, das wir in der approximierten $\mathcal{H}$-Arithmetik mit fast linearem Aufwand abwickeln. Damit haben wir das Eigenwertproblem für den diskreten Laplace-Operator im Wesentlichen auf die hierarchische LDL^T-Zerlegung zurückgespielt. Analog zu den Eigenwerten und Eigenvektoren lassen sich auch die Singulärwerte und Singulärvektoren rückwärtsstabil in $\mathcal{H}$-Arithmetik berechnen.

Das in dieser Arbeit neu entwickelte Konzept zur Approximation der zur Lösung der 2D Wellengleichung mit dem Gautschi-Verfahren benötigten Matrixfunktion-Vektor-Produkte basiert auf Spektralzerlegung des diskreten Laplace-Operators: Nicht die Krylovräume, sondern die aus den Eigenräumen zu den kleinsten Eigenwerten des diskreten Laplace-Operators bestehenden Teilräume erwiesen sich als geeignete Kandidaten zur Approximation der Lösung der hochfrequenten Wellengleichung. Durch Diagonalisierung der die hochfrequenten Oszillationen verursachenden Matrix entkoppelt das Differentialgleichungssystem 2. Ordnung in lauter eindimensionale gewöhnliche Differentialgleichungen 2. Ordnung. So muss statt des exponentiellen Zeitintegrators im hochdimensionalen Koeffizientenraum der FE-Lösungen nur noch eine bestimmte Anzahl eindimensionaler gewöhnlicher Differentialgleichungen 2. Ordnung mit demselben Verfahren gelöst werden. Dies erleichtert zudem erheblich die durchzuführende Konditionsanalyse.

Die entwickelte Methode zur Lösung der hochfrequenten 2D Wellengleichung für glatte Anfangsdaten im Ort besitzt – wie alle approximierten Operationen mit $\mathcal{H}$-Matrizen – fast lineare (d.h. bis auf logarithmische Terme lineare) Komplexität. Aufwand und Approximationsgüte sind dabei unabhängig von der Wellengeschwindigkeit. Die Genauigkeit der ermittelten Lösung hängt nur von der Glattheit der Anfangsdaten ab: Je glatter eine Funktion, umso weniger Eigenfunktionen des Laplace-Operators werden für deren Approximation benötigt.

Neben der 2D Wellengleichung lassen sich auch die 2D Wärmeleitungsgleichung und Schrödingergleichung mittels desselben Verfahrens mit fast linearem Aufwand lösen: Für die Lösung der Schrödingergleichung gelten dieselben Voraussetzungen an die Glattheit der Anfangsdaten, während die Wärmeleitungsgleichung für genügend große Zeiten mit von der Glattheit der Anfangsdaten völlig unabhängiger Approximationsgüte gelöst werden kann.

Gavrilyuk, Hackbusch und Khoromskij haben in [14] bzw. [15] die Lösungsoperatoren einer parabolischen bzw. elliptischen Differentialgleichung durch

$\mathcal{H}$-Matrizen approximiert. Die Übertragung der Ideen aus [14] bzw. [15] auf die Lösungsoperatoren der (hyperbolischen!) Wellengleichung war nicht von Erfolg gekrönt. Was die darstellenden Matrizen der Lösungsoperatoren der homogenen Wellengleichung betrifft, gelang der heuristische Nachweis, dass sie für große Wellengeschwindigkeiten und feine Gitter im Ort gar nicht durch $\mathcal{H}$-Matrizen approximierbar sind. Das liegt im Wesentlichen daran, dass die den Singulärwerten innewohnende Information über die hochfrequenten Oszillationen in den Außerdiagonalblöcken nicht auf einige wenige Singulärwerte beschränkbar ist.

Mittels der in $\mathcal{H}$-Arithmetik berechneten Eigenpaare gelang jedoch die Konstruktion eines neuen Verfahrens zur Approximation einer bestimmten Klasse transzendenter Matrixfunktionen, deren Argumente skalare Vielfache der darstellenden Matrix des diskreten Laplace-Operators sind, durch Niedrigrang-Matrizen samt der Bereitstellung hervorragender Fehlerschätzer ohne zusätzlichen Aufwand. Dazu gehören sämtliche Matrizen, deren Singulärwerte ein genügend rasches Abklingverhalten aufweisen, wie beispielsweise die Exponentialfunktion des Laplace-Operators (Das ist die Lösung der Wärmeleitungsgleichung!). Dabei stellen diese Niedrigrang-Approximationen nichts anderes als eine approximierte Singulärwertzerlegung der jeweiligen Matrixfunktion dar.

Die vorliegende Arbeit gliedert sich im Wesentlichen in vier Teile:

Im **ersten Teil** (Kapitel 1 und 2) diskretisieren wir das kontinuierliche Problem zunächst im Ort mittels Finiter Elemente. Die daraus resultierende ODE in der Zeit wird dann mittels des exponentiellen Gautschi-Verfahrens in eine Drei-Term-Rekursion für die diskrete Lösung überführt. Die Galerkin-Konvergenz für die im Ort semidiskrete Wellengleichung konnte durch Übertragung der Beweisstrategie in [44] für dasselbe Vorhaben bei der Wärmeleitungsgleichung realisiert werden. Der Konvergenzbeweis für den Gautschi-Algorithmus in [29] konnte für die hier vorliegende, im Vergleich zu [29] spezielle Gestalt der rechten Seite stark vereinfacht werden. Für unser Verfahren zur Lösung der hochfrequenten Wellengleichung konnten wir somit Fehlerschranken 2. Ordnung sowohl für die Zeit- als auch für die Ortsdiskretisierung nachweisen.

Im **zweiten Teil** (Kapitel 3) wird zunächst der Nachweis erbracht, dass die Krylovunterräume die falsche Wahl zur Approximation der Lösung der hochfrequenten Wellengleichung sind. Im Anschluss daran wird eine unseren Bedürfnissen angepasste Klasse von 2D $\mathcal{H}$-Matrizen definiert, indem die Vorgehensweise aus [20] für das regelmäßige quadratische Gitter mit stückweise konstanten Funktionen auf das regelmäßige Dreiecksgitter mit stückweise linearen C^0-Funktionen übertragen wird. Dann werden die in [20, Kapitel 4.4.2 und 4.4.3] in ihrer Funktionsweise nur kurz geschilderten Matrix-Vektor-Multiplikation, Matrix-Addition, Matrix-Multiplikation und

Matrix-Invertierung für 2D $\mathcal{H}$-Matrizen detailliert ausgeführt sowie jeweils parallel zum hierarchischen Schema der approximierten Operationen mitlaufende a posteriori Fehlerschätzer konstruiert. Im Anschluss daran werden eine hierarchische Matrix-Vorwärts- sowie Matrix-Rückwärtssubstitution und damit eine hierarchische Cholesky-Zerlegung, LDL^T-Zerlegung, QR-Zerlegung und Polarzerlegung entwickelt sowie der Approximationsfehler jeweils durch Rückwärtsanalyse bestimmt. Die $\mathcal{H}$-QR-Zerlegung erfolgt durch eine $\mathcal{H}$-Matrix-Multiplikation, eine $\mathcal{H}$-Cholesky-Zerlegung und eine $\mathcal{H}$-Matrix-Vorwärtssubstitution und bietet anschließend die Möglichkeit, den orthogonalen Faktor in $\mathcal{H}$-Arithmetik zu reorthogonalisieren. Dabei wird – von der $\mathcal{H}$-Invertierung bis zum vorgestellten Algorithmus für die $\mathcal{H}$-QR-Zerlegung – die entscheidende Rolle der Kondition der Ausgangsmatrix für die Approximationsgüte anhand geeigneter numerischer Beispiele herausgearbeitet.

Nachdem uns nun ein Großteil der linearen Algebra in hierarchischer Arithmetik zur Verfügung steht, werden wir im **dritten Teil** in Kapitel 4 verschiedene Wege zur Darstellung bzw. Approximation transzendenter Matrixfunktionen mittels $\mathcal{H}$-Matrizen aufzeigen und analysieren, wobei die theoretischen Resultate stets durch anschauliche numerische Experimente in der Praxis erprobt werden. Zunächst zeigen wir, dass die Berechnung bestimmter Matrixfunktionen durch sukzessive Duplikation der Argumente ein schlecht konditioniertes Unterfangen darstellt. Dann folgt ein heuristischer Beweis für die Nichtapproximierbarkeit bestimmter oszillatorischer Matrixfunktionen durch $\mathcal{H}$-Matrizen und wie diese Matrizen durch Glättung in durch Niedrigrang-Matrizen approximierbare Matrixfunktionen überführt werden können. Die Tatsache der Nichtapproximierbarkeit eines Operators wird auch durch einen Blick auf die Nichtglattheit des zugehörigen Integralkerns deutlich gemacht. Die Niedrigrang-Approximationen an die geglätteten Matrixfunktionen werden schließlich aus der im Vorhergehenden entwickelten Spektralzerlegung in $\mathcal{H}$-Arithmetik der darstellenden Matrix des diskreten Laplace-Operators gewonnen.
In Kapitel 5 werden dann die 2D Wellengleichung, Schrödingergleichung bzw. Wärmeleitungsgleichung als entkoppelte Systeme einer bestimmten Anzahl eindimensionaler Wellen-, Schrödinger- bzw. Wärmeleitungsgleichungen gelöst: Diese Anzahl ist im Falle der Wellengleichung und Schrödingergleichung allein durch die Glattheit der Anfangsdaten vorgegeben, was durch geeignete numerische Experimente verdeutlicht wird. Für die Lösung der Wärmeleitungsgleichung ist sie für genügend große Zeiten auch davon unabhängig.

Im **vierten und letzten Teil** (Kapitel 6) werden schließlich die Grundzüge der Implementierung in der Programmiersprache C vom Speicherschema für die $\mathcal{H}$-Matrizen und der rekursiven Struktur deren approximierten Operationen bis hin zur L^2-orthogonalen Projektion von $L^2(\Omega)$ auf den FE-Raum V_h und zur Auswertung transzendenter Matrixfunktionen dargestellt.

Der Großteil der Kapitel folgt einem genesisorientierten Aufbau: Die einzelnen Resultate werden schrittweise entwickelt, wobei auch fehlgeschlagene Lösungsansätze und Sackgassen in der Entwicklung Erwähnung finden. Am Anfang eines jeden Kapitels werden die zentralen Ergebnisse zusammengefasst aufgelistet und deren Relevanz für die folgenden Abschnitte herausgestrichen, um dem Leser die jeweils gesteckten Ziele klarzumachen.

Danksagung

Ich möchte meinem Aufgabensteller und Betreuer Herrn Bornemann für die erhaltene Unterstützung danken, dass er mir nach dem Nichterreichen der gesteckten Ziele mit den Krylovraummethoden den Weg zu den $\mathcal{H}$-Matrizen gewiesen hat und mich den konstruktiven und unverzichtbaren Austausch von Theorie und Praxis auch und vor allem in der Entwicklung neuer Methoden gelehrt hat.
Mein Dank gilt außerdem Herrn Hackbusch sowie den Mitarbeitern seiner Arbeitsgruppe für wertvolle Anregungen und neue Impulse auf dem Gebiet der $\mathcal{H}$-Matrizen, die ich während meines Aufenthalts am Max-Planck-Institut für Mathematik in den Naturwissenschaften in Leipzig und am Mathematischen Forschungsinstitut Oberwolfach erhalten habe.
Schließlich sei noch allen Mitarbeitern am Lehrstuhl Prof. Bornemann für die große Hilfsbereitschaft vor allem im Umgang mit dem Rechner und in programmiertechnischen Fragen gedankt.

Kapitel 1

Das kontinuierliche Problem

In diesem Kapitel werden wir zunächst in Abschnitt 1.1 die zu behandelnde Aufgabe in einem geeigneten Funktionenraum formulieren sowie einige theoretische Resultate wiedergeben, die die Lösung dieser Aufgabe betreffen.

Im Anschluss daran werden wir in Abschnitt 1.2 bereits bestehende Lösungsmöglichkeiten skizzieren und deren Unzulänglichkeiten diskutieren.
Schließlich werden die Ziele genannt, die wir im Rahmen dieser Arbeit erreichen möchten.

1.1 Die Problemstellung

Im Folgenden sei $\Omega = (0,1)^2$ das offene Einheitsquadrat des $\mathbb{R}^2$. Wir betrachten die inhomogene 2D Wellengleichung auf dem Einheitsquadrat

$$u_{tt}(t,x) - c^2 \triangle u(t,x) = f(t,x), \quad x \in \Omega, t \in]0,T] \tag{1.1}$$

im hochfrequenten Bereich, d.h. für Wellengeschwindigkeiten $c \gg 1$, mit homogenen Dirichlet-Randbedingungen

$$u|_{\partial\Omega} \equiv 0$$

und den Anfangsbedingungen

$$u(0,\cdot) = u_0 \in H_0^1(\Omega), \quad u_t(0,\cdot) = \dot{u}_0 \in L^2(\Omega). \tag{1.2}$$

Für die Inhomogenität f gelte $f \in L^2(\Omega)$ und $T > 0$ sei eine feste Endzeit.

Wir haben uns aus Bequemlichkeitsgründen in dieser Arbeit auf das 2D Einheitsquadrat beschränkt. Man beachte jedoch, dass die im Folgenden erarbeitete Lösungsmethode auf beliebige beschränkte Gebiete $\Omega \subset \mathbb{R}^2$ verallgemeinerbar ist.

Bezeichne A den selbstadjungierten Friedrichs-Darstellungsoperator

$$A : D_A \subset H_0^1(\Omega) \cap H^2(\Omega) \to L^2(\Omega) \tag{1.3}$$

der Dirichlet-Form $(\nabla u, \nabla v)_{L^2}$, d.h.

$$(Au, v)_{L^2} = (\nabla u, \nabla v)_{L^2} = a(u, v), \quad u \in D_A, v \in H_0^1(\Omega)$$

(siehe [34, Seite 332 ff.]) auf dem 2D Einheitsquadrat Ω mit homogenen Dirichlet-Randbedingungen.

Das abstrakte Cauchy-Problem für die Wellengleichung in $L^2(\Omega)$

$$u_{tt} + c^2 Au = f, \quad u(0, \cdot) = u_0 \in H_0^1(\Omega),\ u_t(0, \cdot) = \dot{u}_0 \in L^2(\Omega) \tag{1.4}$$

mit $u(t, \cdot) \in H_0^1(\Omega) \cap H^2(\Omega)$ besitzt eine eindeutige Lösung (siehe z.B. [47, Seite 419 ff.] oder [32, Seite 96]), die formal durch

$$\begin{aligned} u(t,x) \quad = \quad & \cos(tc\sqrt{A})\, u_0(x) + (c\sqrt{A})^{-1} \sin(tc\sqrt{A})\, \dot{u}_0(x) + \\ & \int_0^t (c\sqrt{A})^{-1} \sin((t-s)c\sqrt{A}) f(s,x) ds \end{aligned} \tag{1.5}$$

gegeben ist.

1.2 Die Ziele dieser Arbeit

Projizieren wir die kontinuierliche Gleichung (1.4) im Ort auf Finite Elemente und lösen die resultierende gewöhnliche Differentialgleichung mittels eines expliziten Zeitintegrators, bekommen wir aus Stabilitätsgründen eine CFL-Bedingung, die die Zeitschrittweite durch die Gitterbreite h auf dem Einheitsquadrat sowie durch den Kehrwert der Wellengeschwindigkeit c^{-1} beschränkt.
Bei der Verwendung von impliziten Integrationsverfahren muss die Zeitschrittweite ebenso an die Gitterbreite im Ort und an die Wellengeschwindigkeit gekoppelt werden, um große Phasenfehler zu vermeiden: Die Lösungen $\cos(tc\sqrt{A})$ und $\sin(tc\sqrt{A})$ bzw. $\exp(itc\sqrt{A})$ der Wellengleichung haben oszillierenden Charakter, implizite Verfahren liefern jedoch eine rationale Approximation an die Exponentialfunktion! Um diese Oszillationen mittels rationaler Funktionen aufzulösen, muss die Zeitschrittweite entsprechend klein gewählt werden.

Nun stellt sich die folgende Frage: Können diese Beschränkungen an die Zeitschrittweite vermieden werden, und wenn ja, wie?

Würden wir statt der rationalen Approximationen die transzendenten Matrixfunktionen selbst in den Zeitintegrator einbauen, so würden wir keine Beschränkung der Zeitschrittweite mehr erhalten (siehe [29]). Damit hätte sich die Schwierigkeit der Lösung der Wellengleichung auf die Auswertung obiger transzendenter Matrixfunktionen bzw. von Matrixfunktion-Vektor-Produkten verlagert.

Hauptziel dieser Arbeit wird nun die Konstruktion eines numerischen Integrationsverfahrens sein, die hochfrequente Wellengleichung stabil und effizient unter Einbindung transzendenter Matrixfunktionen zu lösen und dadurch jegliche Beschränkung der Zeitschrittweite sowohl durch h als auch durch c^{-1} zu vermeiden.
Die Methode soll überdies adaptiv in Ort und Zeit sein, und die aus der feinen Ortsdiskretisierung resultierenden Matrizen großer Dimension sollen mit möglichst wenig Rechenaufwand unter Einhaltung einer vorgegebenen Genauigkeit behandelt werden.

Kapitel 2

Das diskrete Problem

In Abschnitt 2.1 wird das kontinuierliche Problem zunächst im Ort auf einen endlichdimensionalen Teilraum V_h von $H_0^1(\Omega)$ projiziert.

In Abschnitt 2.2 wird dann die Galerkin-Konvergenz für das semidiskrete Problem in V_h gezeigt. Dabei wird der in [44] für die Wärmeleitungsgleichung angegebene Konvergenzbeweis einfach auf die Wellengleichung übertragen.

In Abschnitt 2.3 wird die Lösung der aus obiger Projektion resultierenden gewöhnlichen Differentialgleichung auf ihre Sensitivität gegenüber Störungen bzgl. der Anfangswerte und der rechten Seite untersucht. Dabei stellt sich heraus, dass die Kondition der Lösung der hochfrequenten Wellengleichung direkt proportional zur Wellengeschwindigkeit und zur Zeit ist.

In Abschnitt 2.4 stellen wir das exponentielle Gautschi-Verfahren vor. Zunächst listen wir dessen wesentliche Eigenschaften und Vorzüge auf. Danach wird die Konvergenztheorie wegen der Abhängigkeit der Inhomogenität allein von der Zeit gegenüber der in [29] geführten stark vereinfacht dargelegt.

Schließlich wenden wir in Abschnitt 2.5 das exponentielle Gautschi-Verfahren auf die semidiskrete Wellengleichung in V_h an und zeigen die Konvergenz quadratischer Ordnung des diskreten Problems sowohl in der Zeit als auch im Ort.

2.1 Die FE-Semidiskretisierung im Ort

Wir projizieren die inhomogene 2D Wellengleichung zuerst im Ort auf den endlichdimensionalen Raum der stetigen stückweise linearen Dreieckselemente bzgl. der regelmäßigen Triangulierung des 2D Einheitsquadrates.

Problemstellung:
Ausgangspunkt ist die 2D Wellengleichung im hochfrequenten Bereich

$$u_{tt} - c^2 \triangle u = f, \;\; x \in \Omega, t \in]0, T] \tag{2.1}$$

mit homogenen Dirichlet-Randbedingungen

$$u|_{\partial\Omega} \equiv 0$$

und den Anfangsdaten

$$u(0,\cdot) = u_0 \in H_0^1(\Omega), \quad u_t(0,\cdot) = \dot{u}_0 \in L^2(\Omega). \tag{2.2}$$

Schwache Formulierung:
Die Variationsformulierung für diese hyperbolische Differentialgleichung lautet

$$(u_{tt}, v)_{L^2} + c^2(\nabla u, \nabla v)_{L^2} = (f, v)_{L^2} \ \ \forall v \in H_0^1(\Omega) \tag{2.3}$$

FE-Diskretisierung:
Sei $V_h = \text{span}(\{\psi_j^h\}_{j=1}^N)$ mit der Menge $\{\psi_j^h\}_{j=1}^N$ der zur betrachteten Triangulierung gehörigen nodalen Basisfunktionen.

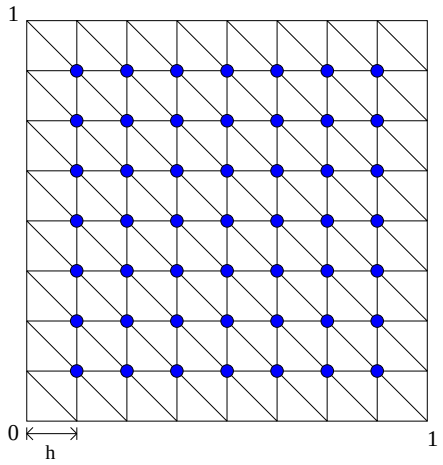

Abbildung 2.1: Regelmäßige Triangulierung des Einheitsquadrates Ω mit der Gitterbreite $h = \frac{1}{n+1}$

Wir projizieren nun Gleichung (2.3) von $H_0^1(\Omega)$ auf den Raum V_h der stetigen stückweise linearen Finiten Dreieckselemente auf dem regelmäßig triangulierten Einheitsquadrat, wobei $h = \frac{1}{n+1}$ dessen Gitterbreite bezeichnet bei n Freiheitsgraden in x- und in y-Richtung, also insgesamt $N = n^2$ Unbekannten:

$$u \in H_0^1(\Omega) \quad \longrightarrow \quad u_h = \sum_{j=1}^{N} u_j \psi_j^h \in V_h$$

Dadurch erhalten wir aus (2.3)

$$(\ddot{u}_h, \psi_j^h)_{L^2} + c^2(\nabla u_h, \nabla \psi_j^h)_{L^2} = (f, \psi_j^h)_{L^2}, \ \ j = 1, \ldots, N. \tag{2.4}$$

Bezeichne nun Q_h die L^2-orthogonale Projektion auf V_h, d.h.

$$(f, \psi_j^h)_{L^2} = (Q_h f, \psi_j^h)_{L^2} \quad \forall\, 1 \le j \le N, f \in L^2(\Omega),$$

und sei

$$(\nabla u_h, \nabla v_h)_{L^2} = (A_h^{op} u_h, v_h)_{L^2} = a(u_h, v_h), \;\; u_h, v_h \in V_h,$$

also $A_h^{op} = Q_h A Q_h : V_h \to V_h$ der Rieszsche Darstellungsoperator auf V_h.

Dann folgt die semidiskrete inhomogene Wellengleichung in V_h

$$\ddot{u}_h + c^2 A_h^{op} u_h = Q_h f, \;\; u_h :]0,T] \to V_h \tag{2.5}$$

mit den Anfangsdaten

$$u_{h0} = Q_h u_0, \quad \dot{u}_{h0} = Q_h \dot{u}_0,$$

eine gewöhnliche Differentialgleichung zweiter Ordnung mit dem bzgl. $(\cdot,\cdot)_{L^2}$ selbstadjungierten, unbeschränkten und positiv definiten diskreten Operator A_h^{op} auf V_h.

Die Konvergenztheorie werden wir für diese „basisfreie“ Version (2.5) führen, während sich für die algorithmische Realisierung die Darstellung in der nodalen FE-Basis empfiehlt:
Sei $u_h = \sum_{j=1}^N u_j \psi_j^h$ und $Q_h f = \sum_{j=1}^N f_j \psi_j^h$. Mit $\mathbf{u}_h = (u_1, \ldots, u_N)^T$ und $\mathbf{f}_h = (f_1, \ldots, f_N)^T$ ist (2.4) äquivalent zu

$$\sum_{i=1}^N \ddot{u}_i (\psi_i^h, \psi_j^h)_{L^2} + c^2 \sum_{i=1}^N u_i a(\psi_i^h, \psi_j^h) = \sum_{i=1}^N f_i (\psi_i^h, \psi_j^h)_{L^2}, \;\; j = 1, \ldots, N$$

$$\iff \qquad M_h \ddot{\mathbf{u}}_h + c^2 A_h \mathbf{u}_h = M_h \mathbf{f}_h, \tag{2.6}$$

wobei die symmetrisch positiv definiten $N \times N$-Matrizen

$$M_h = \Big((\psi_j^h, \psi_i^h)_{L^2} \Big)_{1 \le i,j \le N}$$

die zur nodalen FE-Basis $\{\psi_j^h\}_{j=1}^N$ gehörige Massenmatrix und

$$A_h = \Big(a(\psi_j^h, \psi_i^h) \Big)_{1 \le i,j \le N}$$

die zugehörige Steifigkeitsmatrix bezeichnen.

Wir wollen nun den Zusammenhang der beiden Darstellungen (2.5) und (2.6) genauer untersuchen.
Sei Φ_h der durch

$$\Phi_h : \mathbb{R}^N \to V_h; \mathbf{u}_h \mapsto u_h = \sum_{j=1}^N u_j \psi_j^h$$

definierte Isomorphismus zwischen $\mathbb{R}^N$ und V_h (vgl. [18, Kapitel 8]) sowie der (adjungierte oder) duale Operator Φ_h' durch

$$\Phi_h' : V_h^* \to (\mathbb{R}^N)^*; u_h^* \mapsto \Phi_h' u_h^* := u_h^* \circ \Phi_h$$

gegeben.
Sei X ein Hilbertraum. Dann ist die Abbildung $R_X : X \to X^*; x \mapsto x^*$ mit $x^*(y) = (x,y)_H \, \forall y \in X$ eine Isometrie zwischen dem Hilbertraum X und dem Dualraum X^*. Seien also die Isometrien

$$R_{\mathbb{R}^N} : \mathbb{R}^N \to (\mathbb{R}^N)^*; R_{\mathbb{R}^N}(\mathbf{v}_h)(\mathbf{u}_h) = \mathbf{v}_h^*(\mathbf{u}_h) = \langle \mathbf{u}_h, \mathbf{v}_h^* \rangle = \langle \mathbf{v}_h, \mathbf{u}_h \rangle_2$$

und

$$R_{V_h} : V_h \to V_h^*; R_{V_h}(v_h)(u_h) = v_h^*(u_h) = \langle u_h, v_h^* \rangle = (v_h, u_h)_{L^2}$$

gegeben, wobei $\langle \cdot, \cdot \rangle$ das jeweilige Dualitätsprodukt bezeichne.
Mit diesen Isometrien definieren wir die Hilbertraum-Adjungierte

$$\Phi_h^* := R_{\mathbb{R}^N}^{-1} \Phi_h' R_{V_h} : V_h \to \mathbb{R}^N.$$

Damit gilt für $v_h \in V_h$ und $\mathbf{u}_h \in \mathbb{R}^N$

$$\begin{aligned} (v_h, \Phi_h \mathbf{u}_h)_{L^2} &= v_h^*(\Phi_h \mathbf{u}_h) = \Phi_h' v_h^*(\mathbf{u}_h) = R_{\mathbb{R}^N} \Phi_h^* R_{V_h}^{-1} v_h^*(\mathbf{u}_h) \\ &= (\Phi_h^* v_h)^*(\mathbf{u}_h) = \langle \Phi_h^* v_h, \mathbf{u}_h \rangle_2 \end{aligned}$$

und somit $(\Phi_h^* v_h)_j = (v_h, \psi_j^h)_{L^2}$.

Lemma 2.1 (Massenmatrix und Steifigkeitsmatrix)
Es gelten folgende Beziehungen:

(a) $A_h = \Phi_h^* A_h^{op} \Phi_h$

(b) $M_h = \Phi_h^* \Phi_h$

(c) $M_h^{-1} A_h = \Phi_h^{-1} A_h^{op} \Phi_h$

Beweis: *(a)* Es ist

$$(A_h)_{ji} = (A_h^{op} \psi_i^h, \psi_j^h)_{L^2} = (A_h^{op} \Phi_h \mathbf{e}_i, \Phi_h \mathbf{e}_j)_{L^2} = \langle (\Phi_h^* A_h^{op} \Phi_h) \mathbf{e}_i, \mathbf{e}_j \rangle_2$$

sowie andererseits

$$(A_h)_{ji} = \langle A_h \mathbf{e}_i, \mathbf{e}_j \rangle_2$$

für alle $1 \le i, j \le N$. Daraus folgt sofort $A_h = \Phi_h^* A_h^{op} \Phi_h$.
(b) Weiter ist

$$(M_h)_{ji} = (\psi_i^h, \psi_j^h)_{L^2} = (\Phi_h \mathbf{e}_i, \Phi_h \mathbf{e}_j)_{L^2} = \langle (\Phi_h^* \Phi_h) \mathbf{e}_i, \mathbf{e}_j \rangle_2$$

sowie andererseits

$$(M_h)_{ji} = \langle M_h \mathbf{e}_i, \mathbf{e}_j \rangle_2$$

für alle $1 \le i, j \le N$. Daraus folgt sofort $M_h = \Phi_h^* \Phi_h$.
(c) Für die lineare Abbildung $M_h^{-1} A_h : \mathbb{R}^N \to \mathbb{R}^N$ folgt schließlich

$$M_h^{-1} A_h = \Phi_h^{-1} A_h^{op} \Phi_h$$

aus *(a)* und *(b)*. □

Zusammenfassung und Ausblick:

- Der Wechsel von der basisfreien Darstellung zur Darstellung in der nodalen FE-Basis erfolgt durch den Operator Φ_h. Um die Metrik zu erhalten, ist dabei das L^2-Skalarprodukt $(\cdot,\cdot)_{L^2}$ in V_h durch das mit M_h definierte Skalarprodukt $\langle\cdot,\cdot\rangle_{M_h} := \langle M_h\cdot,\cdot\rangle_2$ im $\mathbb{R}^N$ zu ersetzen.
- Während wir die nun folgende Konvergenztheorie für die Galerkin-Konvergenz in der basisfreien Darstellung führen werden, wird die algorithmische Realisierung in der Knotendarstellung erfolgen.

2.2 Die Galerkin-Konvergenz

Wir haben in Abschnitt 2.1 die kontinuierliche Wellengleichung von $H_0^1(\Omega)$ auf V_h projiziert:

$$\begin{array}{ccc} \text{PDE} & & \text{ODE} \\ \ddot{u} + c^2 A u = f & \longrightarrow & \ddot{u}_h + c^2 A_h^{op} u_h = f_h \\ \text{Lösung } u & & \text{Lösung } u_h \end{array}$$

Im Folgenden werden wir die Konvergenz $u_h \longrightarrow u$ der Lösung der semidiskreten Wellengleichung in V_h (ODE) gegen die Lösung der Wellengleichung in $H_0^1(\Omega)$ (PDE) für $h \to 0$ nachweisen, genauer $\|u - u_h\| \leq Ch^q$ in einer geeigneten Norm $\|\cdot\|$ mit einem $q \in \mathbb{N}$, wobei C – im Gegensatz zur Wellengeschwindigkeit c – hier und im weiteren Verlauf dieser Arbeit stets eine positive Konstante bezeichnet.

Wir übertragen im Folgenden die zweite Beweisstrategie von [44, Theorem 1.3] für die Wärmeleitungsgleichung auf die Wellengleichung.

Theorem 2.2 (Galerkin-Konvergenz für die Poisson-Gleichung)
Seien u_h und u die Lösungen der Gleichungen

$$A_h^{op} u_h = Q_h f \quad in\ V_h \tag{2.7}$$

und

$$Au = f \quad in\ H_0^1(\Omega). \tag{2.8}$$

Dann gilt für $1 \leq s \leq 2$

$$\|u_h - u\|_{L^2} \leq Ch^s \|u\|_{H^s} \quad und \quad \|\nabla(u_h - u)\|_{L^2} \leq Ch^{s-1}\|u\|_{H^s}.$$

Beweis: [44, Theorem 1.1] □

Bezeichne $R_h : H_0^1(\Omega) \to V_h$ die Ritz-Galerkin-Projektion von $H_0^1(\Omega)$ auf V_h, d.h.

$$a(R_h u, v_h) = a(u, v_h) \ \ \forall v_h \in V_h. \tag{2.9}$$

$R_h u$ ist die FE-Lösung u_h der diskreten Poisson-Gleichung (2.7).

Die folgende Fehlerschätzung ist eine direkte Konsequenz aus Theorem 2.2:

Lemma 2.3 (Direkte Konsequenz aus Theorem 2.2)
Mit der Ritz-Galerkin-Projektion R_h aus (2.9) gilt

$$\|R_h v - v\|_{L^2} + h\|\nabla(R_h v - v)\|_{L^2} \leq Ch^s \|v\|_{H^s}$$

für alle $v \in H^s(\Omega) \cap H_0^1(\Omega), 1 \leq s \leq 2$.

Unser Ziel ist nun eine L^2-Fehlerschätzung der Lösung des semidiskreten Cauchy-Problems

$$\ddot{u}_h + c^2 A_h^{op} u_h = Q_h f, \quad u_h(0) = Q_h u_0, \;\; \dot{u}_h(0) = Q_h \dot{u}_0. \tag{2.10}$$

Theorem 2.4 (Fehlerschätzung in der L^2-Norm)
Seien u und u_h die Lösungen der Anfangswertprobleme (AWP) (1.4) und (2.10). Dann gilt

$$\begin{aligned} \|u_h(t) - u(t)\|_{L^2} \quad \leq \quad & Ch^2 \Big[\|u_0\|_{H^2} + \textstyle\int_0^t \|u_t(s)\|_{H^2} ds + \\ & t \left(\|\dot{u}_0\|_{H^2} + \textstyle\int_0^t \|u_{tt}(s)\|_{H^2} ds \right) \Big] \end{aligned}$$

Beweis: Wir suchen nach einer oberen Schranke für $\|u_h - u\|_{L^2}$. Dazu schreiben wir $u_h - u = \theta + \rho$ mit $\theta = u_h - R_h u$, $\rho = R_h u - u$, und beachten, dass der Operator R_h mit der Zeitableitung $\dot{}$ vertauscht.
Die Abschätzung für $\rho(t)$ in der L^2-Norm

$$\begin{aligned} \|\rho(t)\|_{L^2} \quad &= \quad \|R_h u(t) - u(t)\|_{L^2} \leq Ch^2 \|u(t)\|_{H^2} \\ &\leq \quad Ch^2 \left(\|u_0\|_{H^2} + \textstyle\int_0^t \|u_t(s)\|_{H^2} ds \right) \end{aligned}$$

folgt sofort mit Lemma 2.3.

Im Hauptteil des Beweises vergleichen wir nun die Lösung des semidiskreten Problems mit der Ritz-Galerkin-Projektion der exakten Lösung.
Die Lösung der homogenen Wellengleichung

$$u_{tt} + c^2 A u = 0, \;\; u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0 \tag{2.11}$$

ist formal gegeben durch

$$u_{hom}(t) = \cos(tc\sqrt{A})\, u_0 + (c\sqrt{A})^{-1} \sin(tc\sqrt{A})\, \dot{u}_0, \tag{2.12}$$

die Lösung der inhomogenen Wellengleichung

$$u_{tt} + c^2 A u = f, \;\; u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0 \tag{2.13}$$

ergibt sich mittels Variation der Konstanten zu

$$u_{inhom}(t) = u_{hom}(t) + \int_0^t (c\sqrt{A})^{-1} \sin((t-s)c\sqrt{A}) f(s) ds. \tag{2.14}$$

Der diskrete Laplace-Operator $A_h^{op} : V_h \to V_h$ steht mit unseren übrigen Operatoren durch die Gleichung

$$A_h^{op} R_h = Q_h A \tag{2.15}$$

in Beziehung, denn

$$\begin{aligned}(A_h^{op} R_h u, v_h)_{L^2} &= a(R_h u, v_h) = a(u, v_h) \\ &= (Au, v_h)_{L^2} = (Q_h Au, v_h)_{L^2} \quad \forall v_h \in V_h.\end{aligned}$$

Unter Verwendung von (1.4), (2.10) und (2.15) folgt nun

$$\begin{aligned}\theta_{tt} + c^2 A_h^{op}\theta &= (u_{h,tt} + c^2 A_h^{op} u_h) - (R_h u_{tt} + c^2 A_h^{op} R_h u) \\ &= Q_h f + (Q_h - R_h)u_{tt} - Q_h(u_{tt} + c^2 Au) \\ &= Q_h(I - R_h)u_{tt} = -Q_h \rho_{tt}\end{aligned} \tag{2.16}$$

für $t > 0$.
Mit der Lösung

$$u_{h,hom}(t) = \cos(tc\sqrt{A_h^{op}})\, u_{0h} + (c\sqrt{A_h^{op}})^{-1} \sin(tc\sqrt{A_h^{op}})\, \dot{u}_{0h} \tag{2.17}$$

der homogenen semidiskreten Wellengleichung

$$u_{h,tt} + c^2 A_h^{op} u_h = 0, \;\; u_h(0) = u_{0h}, \quad \dot{u}_h(0) = \dot{u}_{0h} \tag{2.18}$$

und der Lösung

$$u_{h,inhom}(t) = u_{h,hom}(t) + \int_0^t (c\sqrt{A_h^{op}})^{-1} \sin((t-s)c\sqrt{A_h^{op}}) f_h(s) ds \tag{2.19}$$

der inhomogenen semidisktreten Wellengleichung

$$u_{h,tt} + c^2 A_h^{op} u_h = f_h, \;\; u_h(0) = u_{0h}, \quad \dot{u}_h(0) = \dot{u}_{0h} \tag{2.20}$$

folgt aus (2.16)

$$\begin{aligned}\theta(t) &= \cos(tc\sqrt{A_h^{op}})\, \theta(0) + (c\sqrt{A_h^{op}})^{-1} \sin(tc\sqrt{A_h^{op}})\, \dot{\theta}(0) - \\ &\quad \int_0^t (c\sqrt{A_h^{op}})^{-1} \sin((t-s)c\sqrt{A_h^{op}}) Q_h \rho_{tt}(s) ds\end{aligned} \tag{2.21}$$

Der Operator $\cos(tc\sqrt{A_h^{op}})$ ist „stabil in $L^2(\Omega)$“, d.h. es gilt

$$\|\cos(tc\sqrt{A_h^{op}})\, u_{0h}\|_{L^2} \leq \|u_{0h}\|_{L^2}$$

für alle $u_{0h} \in V_h, t \geq 0$: Aufgrund der Selbstadjungiertheit von A_h^{op} besitzt der Operator $\cos(tc\sqrt{A_h^{op}})$ nämlich ein Orthonormalsystem von Eigenfunktionen

bzgl. $(\cdot,\cdot)_{L^2}$ mit allen Eigenwerten betragsmäßig kleiner gleich 1.
Analog dazu ist

$$\|(c\sqrt{A_h^{op}})^{-1}\sin(tc\sqrt{A_h^{op}})\,\dot{u}_{0h}\|_{L^2} \le t\,\|\dot{u}_{0h}\|_{L^2}$$

für alle $\dot{u}_{0h} \in V_h, t \ge 0$.
Da Q_h Norm 1 in $L^2(\Omega)$ hat, impliziert (2.21)

$$\|\theta(t)\|_{L^2} \le \|\theta(0)\|_{L^2} + t\left(\|\dot{\theta}(0)\|_{L^2} + \int_0^t \|\rho_{tt}(s)\|_{L^2} ds\right). \tag{2.22}$$

Daher gilt mit

$$\begin{aligned}\|\theta(0)\|_{L^2} &= \|u_{0h} - R_h u_0\|_{L^2} \le \|u_{0h} - u_0\|_{L^2} + \|R_h u_0 - u_0\|_{L^2}\\ &\le Ch^2\|u_0\|_{H^2},\end{aligned}$$

$$\begin{aligned}\|\dot{\theta}(0)\|_{L^2} &= \|\dot{u}_{0h} - R_h \dot{u}_0\|_{L^2} \le \|\dot{u}_{0h} - \dot{u}_0\|_{L^2} + \|R_h \dot{u}_0 - \dot{u}_0\|_{L^2}\\ &\le Ch^2\|\dot{u}_0\|_{H^2},\end{aligned}$$

und

$$\|\rho_{tt}\|_{L^2} = \|R_h u_{tt} - u_{tt}\|_{L^2} \le Ch^2\|u_{tt}\|_{H^2}$$

insgesamt

$$\|\theta(t)\|_{L^2} \le Ch^2\left(\|u_0\|_{H^2} + t\|\dot{u}_0\|_{H^2} + t\int_0^t \|u_{tt}(s)\|_{H^2} ds\right)$$

und die Behauptung von Theorem 2.4

$$\begin{aligned}\|u_h(t) - u(t)\|_{L^2} &\le \|\rho(t)\|_{L^2} + \|\theta(t)\|_{L^2}\\ &\le Ch^2\Big[\|u_0\|_{H^2} + \textstyle\int_0^t \|u_t(s)\|_{H^2} ds +\\ &\qquad t\left(\|\dot{u}_0\|_{H^2} + \textstyle\int_0^t \|u_{tt}(s)\|_{H^2} ds\right)\Big]\end{aligned}$$

□

Die gewünschte Abschätzung für θ war im obigen Beweis eine Konsequenz der Stabilität der Lösungsoperatoren der homogenen semidiskreten Wellengleichung in Kombination mit der Fehlerschätzung für $\rho_{tt} = (R_h - I)u_{tt}$ in $\|\cdot\|_{L^2}$.

2.3 Sensitivitätsanalyse der Matrixexponentialfunktion

Es gibt diverse Artikel zur Konditionsanalyse der Matrixexponentiellen (siehe z.B. [1], [39], [45], [46]). Da wir es hier mit der Exponentialfunktion einer Matrix zu tun haben, deren sämtliche Eigenwerte auf der imaginären Achse liegen, empfiehlt sich ein Vorgehen nach dem Vorbild von Gröbner/Aleksejew

(siehe z.B. [6, Seite 58 ff.]).

Betrachten wir die gewöhnliche Differentialgleichung 1. Ordnung

$$\dot{x} = f(x) = ic\sqrt{A}x \tag{2.23}$$

mit dem Anfangswert $x(0) = x_0$ und einer symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{N\times N}$.
Die Lösung von (2.23) lautet

$$x(t) = \Phi^t x_0 = e^{ic\sqrt{A}t}x_0,$$

wobei die Einparametergruppe von Diffeomorphismen Φ den Phasenfluß der Differentialgleichung (2.23) bezeichnet.
Die sogenannte Propagationsmatrix

$$W(t-s) := D_x\Phi^{t-s}x|_{x=\Phi^s x_0} \in \mathbb{R}^{N\times N},$$

welche die Variationsgleichung

$$\frac{d}{dt}W(t-s) = f_x(\Phi^t x_0)\cdot W(t-s),\ \ W(0) = I$$

löst, gibt an, wie stark sich eine **Störung δx_s der Lösung $x(s)$ zum Zeitpunkt** s für $t > s$ verstärkt. Es gilt nämlich linearisiert $\delta x(t) \doteq W(t-s)\delta x_s$.
Für die Differentialgleichung (2.23) ergibt sich die Variationsgleichung zu

$$\frac{d}{dt}W(t-s) = ic\sqrt{A}\cdot W(t-s),\ \ W(0) = I$$

und somit die unitäre Propagationsmatrix $W(t-s) = e^{ic\sqrt{A}(t-s)}$.

Wir lassen nun **parametrische Störungen der rechten Seite** $f(x) = f(x;\lambda)$ zu (Wir stören die symmetrisch positiv definite Matrix A.) und untersuchen, wie sich Störungen in λ auf die Lösung $x(t)$ des AWP auswirken.
Dabei erweitern wir die Differentialgleichung durch Ankoppeln der Parameter als neue Variablen, so dass sich die Untersuchung von Störungen bzgl. der Parameter auf die soeben behandelte Untersuchung der Störungen bzgl. der Anfangswerte zurückführen lässt. Sei also

$$\begin{aligned} \dot{x}(t) &= f(x(t);\lambda(t)) = ic\sqrt{\lambda(t)}x(t) \quad , \quad x(0) = x_0 \\ \dot{\lambda}(t) &= 0 \qquad\qquad\qquad\qquad\quad , \quad \lambda(0) = \lambda_0 = A \end{aligned}$$

Die Sensitivitätsmatrix $P(t;\lambda_0) = \frac{d}{d\lambda}\Phi^t x_0|_{\lambda=\lambda_0}$, die die inhomogene Variationsgleichung

$$\frac{d}{d\lambda}P(t;\lambda_0) = f_x(\Phi^t x_0;\lambda_0)P(t;\lambda_0) + f_\lambda(\Phi^t x_0;\lambda_0)$$

mit homogenen Anfangswerten $P(0;\lambda_0) = 0$ löst, gibt die Änderung der Lösung $x(t)$ mit λ im Punkt λ_0 an: $\delta x(t) \doteq P(t;\lambda_0)\delta\lambda$.

Aus der Lösung $W(t-s)$ der homogenen Variationsgleichung erhalten wir mittels Variation der Konstanten

$$\begin{aligned} P(t;\lambda_0) &= \int_0^t W(t-s) f_\lambda(\Phi^s x_0;\lambda_0) ds \\ &= \int_0^t e^{ic\sqrt{A}(t-s)} ic\tfrac{1}{2}\lambda_0^{-\frac{1}{2}} \cdot \Phi^s x_0 ds, \end{aligned}$$

wobei hier $\cdot$ für das Argument von $P(t;\lambda_0)$ steht, und für die Norm der Sensitivitätsmatrix

$$\begin{aligned} \|P(t;\lambda_0)\| &= \max_{\|E\|=1} \|P(t;\lambda_0)E\| \\ &= \max_{\|E\|=1} \| \int_0^t e^{ic\sqrt{A}(t-s)} ic\tfrac{1}{2} A^{-\frac{1}{2}} E\Phi^s x_0 ds\| \\ &\leq \tfrac{c}{2} \int_0^t \|A^{-\frac{1}{2}}\| \cdot \|E\| \cdot \|\Phi^s x_0\| ds \leq Cc \int_0^t \|x(s)\| ds \\ &\leq Ctc \max_{s\in[0,t]} \|x(s)\| \end{aligned}$$

wegen der Unitarität des Operators $e^{ic\sqrt{A}(t-s)}$ und der Beschränkheit des Operators $A^{-\frac{1}{2}}$.

Für **nicht-parametrische Störungen $\delta f(x)$ der rechten Seite** $f(x)$ gilt nach Aleksejew (1961) und Gröbner (1960):
Das AWP $\dot{x} = f(x),\ x(0) = x_0$ habe die Lösung x, das gestörte Problem $\dot{x} = f(x) + \delta f(x),\ x(0) = x_0$ die Lösung $\bar{x} = x + \delta x$.
Dann ist

$$\delta x(t) = \int_0^t M(t-s)\delta f(\bar{x}(s)) ds$$

für t hinreichend nahe bei 0, wobei M eine stetige matrixwertige Abbildung ist („nichtlineare Variation der Konstanten“, siehe [6, Seite 59]).
Für kleine Störungen δf ergibt sich durch Linearisierung sowohl $\delta f(\bar{x}(s)) \doteq \delta f(x(s))$ als auch $M(t-s) \doteq W(t-s)$ und somit $\delta x(t)$ zu

$$\delta x(t) \doteq \int_0^t W(t-s)\delta f(x(s)) ds$$

für t hinreichend nahe bei 0 mit der zu x gehörigen Propagationsmatrix $W(t-s)$ (siehe [6, Seite 60]). Damit ergibt sich

$$\delta x(t) \doteq \int_0^t e^{ic\sqrt{A}(t-s)} \delta f(x(s)) ds$$

und

$$\|\delta x(t)\| \doteq \int_0^t \|\delta f(x(s))\| ds \leq t \max_{s\in[0,t]} \|\delta f(x(s))\|.$$

Für Störungen der Form $\delta f(x) = icEx$ ergibt sich das zu vorhin analoge Resultat

$$\|\delta x(t)\| \stackrel{\cdot}{\leq} tc \max_{s\in[0,t]} \|x(s)\| \cdot \|E\|.$$

Fazit: Stören wir also die darstellende Matrix des diskreten Laplace-Operators, so ist in der Lösung der Wellengleichung mit einer „linearen“ Verstärkung des Eingabefehlers um einen zu tc proportionalen Faktor zu rechnen.

2.4 Der Gautschi-Algorithmus

Wir haben bisher die kontinuierliche Wellengleichung auf den FE-Raum V_h projiziert. Löst man die daraus resultierende ODE 2. Ordnung in der Zeit mittels expliziter Verfahren, so erhält man aus Stabilitätsgründen eine CFL-Bedingung der Form $\tau = Chc^{-1}$, wobei τ die Zeitschrittweite, h die Gitterbreite im Ort und c die Wellengeschwindigkeit bezeichnen.
Die Anwendung impliziter Verfahren liefert aus Genauigkeitsgründen der rationalen Approximation an die Exponentialfunktion eine analoge Beschränkung von τ durch h und c^{-1}.
Um diese Zeitschrittweitenbeschränkung zu umgehen, verwenden wir statt einer polynomialen bzw. rationalen Approximation an die Exponentialfunktion diese Funktion selbst, also statt einem expliziten bzw. impliziten einen exponentiellen Zeitintegrator.

2.4.1 Das Gautschi-Verfahren

Wir betrachten das gewöhnliche lineare Differentialgleichungssystem 2. Ordnung der Dimension N

$$\ddot{y}(t) + Ay(t) = f(t), \;\; y(0) = y_0, \dot{y}(0) = \dot{y}_0 \tag{2.24}$$

mit einer symmetrisch positiv semidefiniten Matrix $A \in \mathbb{R}^{N\times N}$, in dem hochfrequente Oszillationen vom linearen Teil Ay herrühren.
Das numerische Schema basiert auf der Forderung, dass das System (2.24) mit konstanter Inhomogenität f exakt integriert werde. Wir beweisen, dass das Integrationsschema Fehlerschranken 2. Ordnung besitzt, die unabhängig vom Produkt aus der Zeitschrittweite mit den Frequenzen sind: Weder aus Stabilitäts- noch Genauigkeitsgründen muss die Zeitschrittweite durch die Frequenzen von A beschränkt werden (siehe [29]).

Als Gautschi-Algorithmus wird nun die folgende Drei-Term-Rekursion zur Lösung des Differentialgleichungssystems (2.24) bezeichnet:

$$y_{n+1} - 2y_n + y_{n-1} = \tau^2\sigma(\tau^2 A)(-Ay_n + f_n), \tag{2.25}$$

wobei y_n eine Näherung von $y(t_n)$, $t_n = n\tau$ und $f_n = f(t_n)$ ist sowie

$$\sigma(x^2) = \left(\frac{\sin\frac{1}{2}x}{\frac{1}{2}x}\right)^2 = 2\,\frac{1-\cos x}{x^2}.$$

Das Drei-Schritt-Verfahren (2.25) hat die Startwerte y_0 und

$$y_1 = \cos(\tau\sqrt{A})\, y_0 + (\sqrt{A})^{-1} \sin(\tau\sqrt{A})\, \dot{y}_0 + \frac{1}{2}\tau^2 \sigma(\tau^2 A)\, f_0.$$

Man beachte, dass für eine symmetrisch positiv semidefinite Matrix A eine eindeutig definierte Matrix B mit $B^2 = A$ existiert, d.h. A hat eine eindeutige Quadratwurzel $\sqrt{A}$.

2.4.2 Eigenschaften des Gautschi-Verfahrens

1) Struktur des Gautschi-Algorithmus:
$\sigma \equiv 1$ würde ein diskretes Schema liefern, das durch Diskretisierung der 2. Zeitableitung mittels des zentralen Differenzenquotienten 2. Ordnung entsteht (Störmer-Methode, siehe [26, Seite 462]). Einfügen obiger transzendenter Funktion σ macht daraus einen exponentiellen Zeitintegrator.

2) Ein Algorithmus für die Ableitung:
Näherungen für die Ableitung der Lösung lassen sich durch die folgende Drei-Term-Rekursion berechnen:

$$\dot{y}_{n+1} = \dot{y}_{n-1} + 2\tau\psi(\tau^2 A)(-Ay_n + f_n) \tag{2.26}$$

mit $\psi(x^2) = \frac{\sin x}{x}$ und den Startwerten $\dot{y}_0$ und

$$\dot{y}_1 = -\sqrt{A}\sin(\tau\sqrt{A})\, y_0 + \cos(\tau\sqrt{A})\, \dot{y}_0 + \tau\psi(\tau^2 A)\, f_0.$$

3) Exaktheit des diskreten Verfahrens für konstante Inhomogenität:
Das Gautschi-Verfahren liefert die exakte Lösung von (2.24) für $f \equiv \text{const}$:

$$\begin{aligned}
& y(t+\tau) - 2y(t) + y(t-\tau) \\
= \;& \cos(\tau\sqrt{A})\, y(t) + (\sqrt{A})^{-1}\sin(\tau\sqrt{A})\, \dot{y}(t) + \\
& \textstyle\int_0^\tau (\sqrt{A})^{-1}\sin((\tau-s)\sqrt{A}) f\, ds - 2y(t) + \\
& \cos(\tau\sqrt{A})\, y(t) - (\sqrt{A})^{-1}\sin(\tau\sqrt{A})\, \dot{y}(t) - \\
& \textstyle\int_0^{-\tau} (\sqrt{A})^{-1}\sin((\tau+s)\sqrt{A}) f\, ds \\
= \;& 2\cos(\tau\sqrt{A})\, y(t) - 2y(t) + 2A^{-1} f - 2A^{-1}\cos(\tau\sqrt{A})\, f \\
= \;& -2(I - \cos(\tau\sqrt{A}))y(t) + 2(I - \cos(\tau\sqrt{A}))A^{-1} f \\
= \;& 2\tau^2 \tfrac{I - \cos(\tau\sqrt{A})}{\tau^2} A^{-1}(-Ay(t) + f) \\
= \;& \tau^2 \sigma(\tau^2 A)(-Ay(t) + f)
\end{aligned}$$

sowie

$$\begin{aligned}
& \dot{y}(t+\tau) - \dot{y}(t-\tau) \\
= \;& -\sqrt{A}\sin(\tau\sqrt{A})\, y(t) + \cos(\tau\sqrt{A})\, \dot{y}(t) + \textstyle\int_0^\tau \cos((\tau-s)\sqrt{A}) f\, ds \\
& -\sqrt{A}\sin(\tau\sqrt{A})\, y(t) - \cos(\tau\sqrt{A})\, \dot{y}(t) - \textstyle\int_0^{-\tau} \cos((\tau+s)\sqrt{A}) f\, ds \\
= \;& -2\sqrt{A}\sin(\tau\sqrt{A})\, y(t) + 2(\sqrt{A})^{-1}\sin(\tau\sqrt{A})\, f \\
= \;& 2\tau(\tau\sqrt{A})^{-1}\sin(\tau\sqrt{A})(-Ay(t) + f) \\
= \;& 2\tau\psi(\tau^2 A)(-Ay(t) + f)
\end{aligned}$$

4) Energieerhaltung:
Die Gautschi-Methode ist energieerhaltend, da die Lösungen y_n und $\dot{y}_n$ der homogenen Drei-Term-Rekursionen wegen Eigenschaft **3)** der exakten Lösung $y(t_n)$ und deren Ableitung $\dot{y}(t_n)$ entsprechen und die Energie für die kontinuierliche homogene Gleichung erhalten ist: Mit

$$E_{ges}(t) = E_{kin}(t) + E_{pot}(t) = \frac{1}{2}\langle \dot{y}(t), \dot{y}(t)\rangle_2 + \frac{1}{2}\langle Ay(t), y(t)\rangle_2$$

gilt nämlich

$$\dot{E}_{ges}(t) = \langle \ddot{y}(t), \dot{y}(t)\rangle_2 + \langle Ay(t), \dot{y}(t)\rangle_2 = 0.$$

2.4.3 Konvergenz des Gautschi-Verfahrens

Der im Folgenden geführte Beweis lehnt sich an die Darstellung in [29] an, ist aber wegen der Abhängigkeit der Inhomogenität f allein von der Zeit stark vereinfacht worden.

Theorem 2.5 (Konvergenz in der euklidischen Norm)
Für den Fehler $e_n = y_n - y(t_n)$ des exponentiellen Gautschi-Verfahrens (2.25) gilt

$$\|y_n - y(t_n)\|_2 \leq \tau^2 \left(\frac{M_1}{6} t_n + \frac{M_2}{12} t_n^2 \right),$$

wobei $\|f^{(k)}(t)\|_2 \leq M_k \ \ \forall t \in [0,T], k = 1,2,3.$

Man beachte, dass die Fehlerschranke gleichmäßig in den Frequenzen ist, d.h. hohe Frequenzen haben keine Einschränkung von τ zur Folge.

Für den Beweis von Theorem 2.5 benötigen wir die folgenden drei Lemmata:

Lemma 2.6 (Abbrechfehler des Gautschi-Verfahrens)
Der Abbrechfehler

$$d_n = y(t_{n+1}) - 2y(t_n) + y(t_{n-1}) - \tau^2 \sigma(\tau^2 A)(-Ay(t_n) + f(t_n))$$

ist von der Form

$$d_n = \tau^4 L_n + \tau^5 z_n$$

mit $\|L_n\|_2 \leq M_2$ und $\|z_n\|_2 \leq M_3$.

Beweis: Mit der exakten Lösungsformel

$$\begin{aligned} y(t+\tau) \quad = \quad & \cos(\tau\sqrt{A})\, y(t) + (\sqrt{A})^{-1} \sin(\tau\sqrt{A})\, \dot{y}(t) + \\ & \textstyle\int_0^\tau (\sqrt{A})^{-1} \sin((\tau - s)\sqrt{A}) f(t+s) ds \end{aligned}$$

gilt

$$d_n = \int_0^\tau (\sqrt{A})^{-1} \sin((\tau - s)\sqrt{A}) \left[f(t_n + s) - 2f(t_n) + f(t_n - s) \right] ds$$

und mit

$$f(t_n \pm s) - f(t_n) = \pm \dot{f}(t_n)s + \ddot{f}(t_n)\frac{s^2}{2} \pm f^{(3)}(t_n^{\pm})\frac{s^3}{6},$$

wobei $t_n^{\pm} = t_n^{\pm}(s)$ zwischen t_n und $t_n \pm s$ liegt, folgt

$$\begin{aligned}
d_n &= \int_0^\tau (\sqrt{A})^{-1} \sin((\tau - s)\sqrt{A}) \left[2\ddot{f}(t_n)\tfrac{s^2}{2} + (f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\tfrac{s^3}{6} \right] ds \\
&= \tau^2 \int_0^1 (\tau\sqrt{A})^{-1} \sin((1-\theta)\tau\sqrt{A}) \\
&\quad \left[2\ddot{f}(t_n)\tfrac{(\theta\tau)^2}{2} + (f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\tfrac{(\theta\tau)^3}{6} \right] d\theta \\
&= \underbrace{\tau^4 \int_0^1 (\tau\sqrt{A})^{-1} \sin((1-\theta)\tau\sqrt{A})\ddot{f}(t_n)\theta^2 d\theta}_{=:L_n} + \\
&\quad \underbrace{\tau^5 \int_0^1 (\tau\sqrt{A})^{-1} \sin((1-\theta)\tau\sqrt{A})(f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\frac{\theta^3}{6} d\theta}_{=:z_n},
\end{aligned}$$

wobei

$$\begin{aligned}
\|L_n\|_2 &\leq \int_0^1 \|(\tau\sqrt{A})^{-1} \sin((1-\theta)\tau\sqrt{A})\|_2 \, \|\ddot{f}(t_n)\|_2 \theta^2 d\theta \\
&\leq \int_0^1 (1-\theta)\theta^2 d\theta \cdot M_2 = \tfrac{1}{12} M_2
\end{aligned}$$

und

$$\begin{aligned}
\|z_n\|_2 &\leq \int_0^1 \|(\tau\sqrt{A})^{-1} \sin((1-\theta)\tau\sqrt{A})\|_2 \, \|f^{(3)}(t_n^+) - f^{(3)}(t_n^-)\|_2 \tfrac{\theta^3}{6} d\theta \\
&\leq \int_0^1 (1-\theta)\tfrac{\theta^3}{6} d\theta \cdot 2M_3 = \tfrac{1}{60} M_3
\end{aligned}$$

ist. □

Lemma 2.7 (Drei-Term-Rekursion für die Fehler e_n)
Die Fehler $e_n = y_n - y(t_n)$ genügen der Gleichung

$$e_{n+1} = W_n e_1 - \sum_{j=1}^{n} W_{n-j} d_j$$

mit $W_n = (\sin(n+1)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1}$.

Beweis: Es ist

$$\begin{aligned}
& e_{n+1} - 2e_n + e_{n-1} = \tau^2 \sigma(\tau^2 A)(-Ae_n) - d_n \\
\iff \quad & e_{n+1} - 2\cos(\tau\sqrt{A})\, e_n + e_{n-1} = -d_n \qquad (2.27)
\end{aligned}$$

Für die Anfangswerte der symmetrischen inhomogenen Drei-Term-Rekursion (2.27) gilt $e_0 = 0$ und $e_1 = -d_0$.
Die homogene Drei-Term-Rekursion

$$e_{n+1} - 2\cos(\tau\sqrt{A})\, e_n + e_{n-1} = 0, \;\; n \geq 1$$

besitzt die linear unabhängigen Lösungen $e_n = \cos(n\tau\sqrt{A})$ und $e_n = \sin(n\tau\sqrt{A})$ (siehe z.B. [5]).
Die inhomogene Drei-Term-Rekursion

$$e_n - 2\cos(\tau\sqrt{A})\, e_{n-1} + e_{n-2} = \delta_{jn}, \;\; n \geq j$$

mit der Inhomogenität $\delta_{jn} = \begin{cases} I & \text{falls } j = n \\ 0 & \text{sonst} \end{cases}$, $j \in \mathbb{N}_0$ und den Anfangswerten $e_{j-2} = e_{j-1} = 0$ wird gelöst durch

$$e_n = W_{n-j} = (\sin(n-j+1)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1}, \;\; n \geq j.$$

Die Lösung der inhomogenen Drei-Term-Rekursion (2.27) ergibt sich schließlich durch Superposition der diskreten Greenschen Funktionen W_{n-j} zu

$$e_{n+1} = \sum_{j=0}^{n} W_{n-j}(-d_j) = W_n e_1 - \sum_{j=1}^{n} W_{n-j} d_j$$

□

Lemma 2.8 (Hilfsabschätzung)
Es gilt

$$\| \sum_{j=1}^{n} W_{n-j} d_j \|_2 \leq \tau^2 M_2 t_n^2 + \tau^3 M_3 t_n^2.$$

Beweis:

$$\begin{aligned} \| \textstyle\sum_{j=1}^{n} W_{n-j} d_j \|_2 &\leq \textstyle\sum_{j=1}^{n} \|W_{n-j}\|_2 \, \|d_j\|_2 \\ &\leq \textstyle\sum_{j=1}^{n} (n-j+1) \left(\frac{M_2}{12}\tau^4 + \frac{M_3}{60}\tau^5 \right) \\ &\dot{\leq} \;\; \frac{n^2}{2} \left(\frac{M_2}{12}\tau^4 + \frac{M_3}{60}\tau^5 \right) = \tau^2 \frac{M_2}{24} t_n^2 + \tau^3 \frac{M_3}{120} t_n^2 \end{aligned}$$

□

Beweis von Theorem 2.5:

$$\begin{aligned} \|W_n e_1\|_2 &= \|(\sin(n+1)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1} \\ &\quad \left[\tfrac{1}{2}\tau^2 \sigma(\tau^2 A) f_0 - \textstyle\int_0^\tau (\sqrt{A})^{-1} \sin((\tau - s)\sqrt{A}) f(s) ds \right] \|_2 \\ &= \|(\sin(n+1)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1} \\ &\quad \textstyle\int_0^\tau (\sqrt{A})^{-1} \sin((\tau - s)\sqrt{A})(f_0 - f(s)) ds \|_2 \\ &\leq (n+1) \; \textstyle\int_0^\tau (\tau - s) s \, \|\dot{f}(\xi(s))\|_2 \, ds \;\; \text{mit} \;\; \xi(s) \in [0, s] \\ &\leq (n+1) \, \frac{\tau^3}{6} \max_{0 \leq t \leq \tau} \|\dot{f}(t)\|_2 \dot{\leq} \frac{\tau^2}{6} M_1 t_n \end{aligned}$$

Mit Lemma 2.7 und Lemma 2.8 folgt nun in führender Ordnung

$$\begin{aligned}\|e_{n+1}\|_2 &\leq \|W_n e_1\|_2 + \|\sum_{j=1}^{n} W_{n-j} d_j\|_2 \\ &\leq \tau^2 \left(\tfrac{M_1}{6} t_n + \tfrac{M_2}{24} t_n^2\right)\end{aligned}$$

□

Wir werden nun analog zur vorherigen Vorgehensweise den Fehler der Ableitung $\dot{e}_n = \dot{y}_n - \dot{y}(t_n)$ in der euklidischen Norm abschätzen, um daraufhin diesen Abschnitt über die Konvergenz des Gautschi-Verfahrens mit einer Fehlerschätzung in der Energienorm zu beschließen.
Analog zu Lemma 2.6 gilt

Lemma 2.9 (Abbrechfehler für die Ableitung)
Der Abbrechfehler

$$\dot{d}_n = \dot{y}(t_{n+1}) - \dot{y}(t_{n-1}) - 2\tau\psi(\tau^2 A)(-Ay(t_n) + f(t_n))$$

ist von der Form

$$\dot{d}_n = \tau^3 \dot{L}_n + \tau^4 \dot{z}_n$$

mit $\|\dot{L}_n\|_2 \leq M_2$ *und* $\|\dot{z}_n\|_2 \leq M_3$.

Beweis: Mit der Ableitung nach τ der exakten Lösungsformel

$$\begin{aligned}\dot{y}(t+\tau) &= -\sqrt{A}\sin(\tau\sqrt{A})\, y(t) + \cos(\tau\sqrt{A})\, \dot{y}(t) + \\ &\quad \textstyle\int_0^\tau \cos((\tau - s)\sqrt{A}) f(t+s) ds\end{aligned}$$

gilt

$$\begin{aligned}\dot{d}_n &= \textstyle\int_0^\tau \cos((\tau-s)\sqrt{A}) \left[f(t_n+s) - 2f(t_n) + f(t_n - s)\right] ds \\ &= \textstyle\int_0^\tau \cos((\tau-s)\sqrt{A}) \left[2\ddot{f}(t_n)\frac{s^2}{2} + (f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\frac{s^3}{6}\right] ds \\ &= \tau \textstyle\int_0^1 \cos((1-\theta)\tau\sqrt{A}) \\ &\quad \left[2\ddot{f}(t_n)\frac{(\theta\tau)^2}{2} + (f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\frac{(\theta\tau)^3}{6}\right] d\theta \\ &= \tau^3 \underbrace{\int_0^1 \cos((1-\theta)\tau\sqrt{A})\ddot{f}(t_n)\theta^2 d\theta}_{=:\dot{L}_n} + \\ &\quad \tau^4 \underbrace{\int_0^1 \cos((1-\theta)\tau\sqrt{A})(f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\frac{\theta^3}{6} d\theta}_{=:\dot{z}_n},\end{aligned}$$

wobei

$$\begin{aligned}\|\dot{L}_n\|_2 &\leq \textstyle\int_0^1 \|\cos((1-\theta)\tau\sqrt{A})\|_2 \, \|\ddot{f}(t_n)\|_2 \, \theta^2 d\theta \\ &\leq \textstyle\int_0^1 \theta^2 d\theta \cdot M_2 = \frac{1}{3} M_2\end{aligned}$$

und

$$\begin{aligned}\|\dot{z}_n\|_2 &\leq \textstyle\int_0^1 \|\cos((1-\theta)\tau\sqrt{A})\|_2 \, \|f^{(3)}(t_n^+) - f^{(3)}(t_n^-)\|_2 \frac{\theta^3}{6} d\theta \\ &\leq \textstyle\int_0^1 \frac{\theta^3}{6} d\theta \cdot 2M_3 = \frac{1}{12} M_3\end{aligned}$$

ist. □

Lemma 2.10 (Drei-Term-Rekursion für die Fehler $\dot{e}_n$)
Die Fehler $\dot{e}_n = \dot{y}_n - \dot{y}(t_n)$ genügen der Gleichung

$$\dot{e}_{n+1} = \dot{W}_n \dot{e}_1 + \sum_{j=1}^{n} \dot{W}_{n-j} \dot{h}_j$$

mit $\dot{W}_n = \begin{cases} I & \textit{falls } n = 0, 2, 4, \ldots \\ 0 & \textit{falls } n = 1, 3, 5, \ldots \end{cases}$ *und* $\dot{h}_j = -2\sqrt{A}\sin(\tau\sqrt{A})\, e_j - \dot{d}_j$.

Beweis: Es ist

$$\begin{aligned} &\dot{e}_{n+1} - \dot{e}_{n-1} = 2\tau\psi(\tau^2 A)(-Ae_n) - \dot{d}_n \quad \Longleftrightarrow \\ &\dot{e}_{n+1} - \dot{e}_{n-1} = -2\sqrt{A}\sin(\tau\sqrt{A})\, e_n - \dot{d}_n =: \dot{h}_n \end{aligned} \tag{2.28}$$

Für die Anfangswerte der inhomogenen Drei-Term-Rekursion (2.28) gilt $\dot{e}_0 = 0$ und $\dot{e}_1 = -\dot{d}_0$.
Die homogene Drei-Term-Rekursion

$$\dot{e}_{n+1} - \dot{e}_{n-1} = 0, \;\; n \geq 1$$

besitzt die Fundamentallösungen $\dot{e}_n = I$ und $\dot{e}_n = (-1)^n I$.
Die inhomogene Drei-Term-Rekursion

$$\dot{e}_n - \dot{e}_{n-2} = \delta_{jn}, \;\; n \geq j$$

mit der Inhomogenität $\delta_{jn} = \begin{cases} I & \text{falls } j = n \\ 0 & \text{sonst} \end{cases}$, $j \in \mathbb{N}_0$ und den Anfangswerten $\dot{e}_{j-2} = \dot{e}_{j-1} = 0$ wird gelöst durch

$$\dot{e}_n = \dot{W}_{n-j} = \begin{cases} I & \text{falls } n = j, j+2, \ldots \\ 0 & \text{falls } n = j+1, j+3, \ldots \end{cases}$$

Die Lösung der inhomogenen Drei-Term-Rekursion (2.28) ergibt sich somit durch Superposition der diskreten Greenschen Funktionen $\dot{W}_{n-j}$ zu

$$\begin{aligned} \dot{e}_{n+1} &= \textstyle\sum_{j=0}^{n} \dot{W}_{n-j}\dot{h}_j = \dot{W}_n\dot{e}_1 + \sum_{j=1}^{n} \dot{W}_{n-j}\dot{h}_j \\ &= \begin{cases} \dot{e}_1 + \displaystyle\sum_{\substack{j=1 \\ j\text{gerade}}}^{n} \dot{h}_j & \text{falls } n \text{ gerade} \\ \displaystyle\sum_{\substack{j=1 \\ j\text{ungerade}}}^{n} \dot{h}_j & \text{falls } n \text{ ungerade} \end{cases} \end{aligned}$$

□

Lemma 2.11 (Hilfsabschätzung)
Es gilt

$$\| \sum_{\substack{j=1 \\ j\,ger./ung.}}^{n} \dot{h}_j \|_2 \leq \tau \left(\frac{M_1}{2} t_n + \frac{M_2}{6} t_n^2 \right).$$

Dabei bedeutet „j ger./ung.“ entweder gerade oder ungerade.

Beweis:

$$
\begin{aligned}
& \|\sqrt{A}\sin(\tau\sqrt{A})\, e_j\|_2 \\
= \;& \|\sqrt{A}\sin(\tau\sqrt{A})\ \textstyle\sum_{k=0}^{j-1} W_{j-1-k}(-d_k)\|_2 \\
\leq \;& \|\sqrt{A}\sin(\tau\sqrt{A}) W_{j-1}\, e_1\|_2 + \|\sqrt{A}\sin(\tau\sqrt{A})\ \textstyle\sum_{k=1}^{j-1} W_{j-1-k} d_k\|_2 \\
= \;& \|\sqrt{A}\sin(\tau\sqrt{A})\sin(j\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1} \\
& \textstyle\int_0^\tau (\sqrt{A})^{-1}\sin((\tau-s)\sqrt{A})(f_0 - f(s))ds\|_2 + \\
& \|\textstyle\sum_{k=1}^{j-1}\sqrt{A}\sin(\tau\sqrt{A})\sin((j-k)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1} \\
& \Big(\tau^4 \textstyle\int_0^1 (\tau\sqrt{A})^{-1}\sin((1-\theta)\tau\sqrt{A})\ddot{f}(t_n)\theta^2 d\theta + \\
& \tau^5 \textstyle\int_0^1 (\tau\sqrt{A})^{-1}\sin((1-\theta)\tau\sqrt{A})(f^{(3)}(t_n^+) - f^{(3)}(t_n^-))\frac{\theta^3}{6} d\theta\Big)\|_2 \\
\leq \;& \textstyle\int_0^\tau s\,\|\dot{f}(\xi(s))\|_2\, ds \qquad\qquad \text{mit } \xi(s)\in[0,s] \\
& + \textstyle\sum_{k=1}^{j-1}\Big(\tau^3\int_0^1 \theta^2 d\theta\cdot M_2 + \tau^4\int_0^1 \frac{\theta^3}{6}d\theta\cdot 2M_3\Big) \\
= \;& \tau^2\tfrac{M_1}{2} + \textstyle\sum_{k=1}^{j-1}\big(\tau^3\frac{M_2}{3} + \tau^4\frac{M_3}{12}\big) \\
= \;& \tau^2\tfrac{M_1}{2} + (j-1)\tau^3\tfrac{M_2}{3} + (j-1)\tau^4\tfrac{M_3}{12}
\end{aligned}
$$

Damit ergibt sich nun

$$
\begin{aligned}
\|\sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n} \dot{h}_j\|_2 &= \|\sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n} (2\sqrt{A}\sin(\tau\sqrt{A})\, e_j + \dot{d}_j)\|_2 \\
&\leq \sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n} 2\,\|\sqrt{A}\sin(\tau\sqrt{A})\, e_j\|_2 + \sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n} \|\dot{d}_j\|_2 \\
&\leq \sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n} 2\left(\tau^2\tfrac{M_1}{2} + (j-1)\tau^3\tfrac{M_2}{3} + (j-1)\tau^4\tfrac{M_3}{12}\right) + \\
&\qquad \sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n} \left(\tau^3\tfrac{M_2}{3} + \tau^4\tfrac{M_3}{12}\right) \\
&\dot{\leq}\; n\tau^2\tfrac{M_1}{2} + \tfrac{n^2}{2}\tau^3\tfrac{M_2}{3} = \tau\left(\tfrac{M_1}{2}t_n + \tfrac{M_2}{6}t_n^2\right)
\end{aligned}
$$

□

Theorem 2.12 (Konvergenz in der Energienorm)
Für den Fehler $e_n = y_n - y(t_n)$ des exponentiellen Gautschi-Verfahrens (2.25) gilt

$$\|y_n - y(t_n)\|_a \leq \tau\left(M_1 t_n + \frac{M_2}{3}t_n^2\right),$$

wobei $\|e_n\|_a^2 = \langle Ae_n, e_n\rangle_2 + \|\dot{e}_n\|_2^2$ ist.

Beweis: Wir behandeln zunächst den Term $\langle Ae_{n+1}, e_{n+1}\rangle_2 = \|\sqrt{A}e_{n+1}\|_2^2$.
Es ist $\|\sqrt{A}e_{n+1}\|_2 = \|W_n\sqrt{A}e_1 - \sum_{j=1}^n W_{n-j}\sqrt{A}d_j\|_2$ mit

$$
\begin{aligned}
\|W_n\sqrt{A}e_1\|_2 &\leq \|(\sin(n+1)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1}\sqrt{A} \\
&\qquad \textstyle\int_0^\tau(\sqrt{A})^{-1}\sin((\tau-s)\sqrt{A})(f_0 - f(s))ds\|_2 \\
&\leq (n+1)\textstyle\int_0^\tau s ds\cdot M_1 \doteq \tau\frac{M_1}{2}t_n
\end{aligned}
$$

und

$$\begin{aligned}\|\sum_{j=1}^{n} W_{n-j}\sqrt{A}d_j\|_2 &\dot{=} \|\sum_{j=1}^{n}(\sin(n-j+1)\tau\sqrt{A})(\sin\tau\sqrt{A})^{-1}\sqrt{A}\\ &\quad \tau^4\int_0^1(\tau\sqrt{A})^{-1}\sin((1-\theta)\tau\sqrt{A})\ddot{f}(t_n)\theta^2 d\theta\|_2\\ &\leq \sum_{j=1}^{n}(n-j+1)\tau^3\int_0^1\theta^2 d\theta\cdot M_2\\ &\dot{\leq} \frac{n^2}{2}\tau^3\frac{M_2}{3} = \tau\frac{M_2}{6}t_n^2.\end{aligned}$$

Mit der Ableitung der exakten Lösungsformel gilt

$$\begin{aligned}\|\dot{e}_1\|_2 &= \|\dot{y}_1 - \dot{y}(t_1)\|_2\\ &= \|(\sqrt{A})^{-1}\sin(\tau\sqrt{A})\,f_0 - \int_0^\tau\cos((\tau-s)\sqrt{A})f(s)ds\|_2\\ &= \|\int_0^\tau\cos((\tau-s)\sqrt{A})(f_0-f(s))ds\|_2\\ &\leq \int_0^\tau s\,\|\dot{f}(\xi(s))\|_2\,ds \leq \tau^2\frac{M_1}{2}\end{aligned}$$

mit $\xi(s)\in[0,s]$.
Nach Lemma 2.10 und Lemma 2.11 ist somit

$$\begin{aligned}\|\dot{e}_{n+1}\|_2 &\leq \|\dot{e}_1\|_2 + \|\sum_{\substack{j=1\\ j\text{ger./ung.}}}^{n}\dot{h}_j\|_2\\ &\leq \tau^2\frac{M_1}{2} + \tau\left(\frac{M_1}{2}t_n + \frac{M_2}{6}t_n^2\right)\\ &\dot{=} \tau\left(\frac{M_1}{2}t_n + \frac{M_2}{6}t_n^2\right)\end{aligned}$$

Also gilt für den Fehler e_n in der Energienorm

$$\begin{aligned}\|e_n\|_a &= \sqrt{\|\sqrt{A}e_n\|_2^2 + \|\dot{e}_n\|_2^2}\\ &\leq \sqrt{\tau^2\left(\frac{M_1}{2}t_n + \frac{M_2}{6}t_n^2\right)^2 + \tau^2\left(\frac{M_1}{2}t_n + \frac{M_2}{6}t_n^2\right)^2}\\ &= \tau\sqrt{2}\left(\frac{M_1}{2}t_n + \frac{M_2}{6}t_n^2\right)\end{aligned}$$

□

2.4.4 Lineare Stabilität des Gautschi-Verfahrens

Wir untersuchen nun das lineare System

$$\ddot{y} = -Ay$$

mit symmetrisch positiv semidefinitem A auf lineare Stabilität.

Theorem 2.13 (Lineare Stabilität des Gautschi-Algorithmus)
Die Drei-Term-Rekursion

$$\begin{aligned}&y_{n+1} - 2y_n + y_{n-1} = \tau^2\sigma(\tau^2 A)(-Ay_n)\\ \iff\quad &y_{n+1} - 2\cos(\tau\sqrt{A})\,y_n + y_{n-1} = 0\end{aligned}$$

ist stabil im Sinne dass

$$\|y_n\|_2 \leq n(\|y_0\|_2 + \|y_1\|_2),\ \ n>1$$

gilt.

Beweis: [29, Kap. 5, Theorem 2 für den Spezialfall $B=0$ und die nachfolgende Betrachtung] □

2.5 Anwendung des Gautschi-Verfahrens auf die semidiskrete Wellengleichung

Gegeben sei die semidiskrete Wellengleichung

$$\ddot{u}_h + c^2 A_h^{op} u_h = f_h, \;\; u_h(0) = u_{0h}, \quad \dot{u}_h(0) = \dot{u}_{0h} \tag{2.29}$$

mit der Lösung $u_h \in V_h$. Auf diese gewöhnliche Differentialgleichung 2. Ordnung wenden wir nun den Gautschi-Algorithmus (2.25) an. Dies liefert die folgende Drei-Term-Rekursion

$$u_{h,n+1} = 2u_{h,n} - u_{h,n-1} + \tau^2 \sigma(\tau^2 c^2 A_h^{op})(-c^2 A_h^{op} u_{h,n} + f_{h,n}) \tag{2.30}$$

mit den Startwerten $u_{h,0} = u_{0h}$ und

$$u_{h,1} = \cos(\tau c\sqrt{A_h^{op}})\, u_{0h} + (c\sqrt{A_h^{op}})^{-1} \sin(\tau c\sqrt{A_h^{op}})\, \dot{u}_{0h} + \frac{1}{2}\tau^2 \sigma(\tau^2 c^2 A_h^{op})\, f_{h,0}$$

für die in Ort und Zeit diskrete Lösung der Wellengleichung $u_{h,\tau}$.

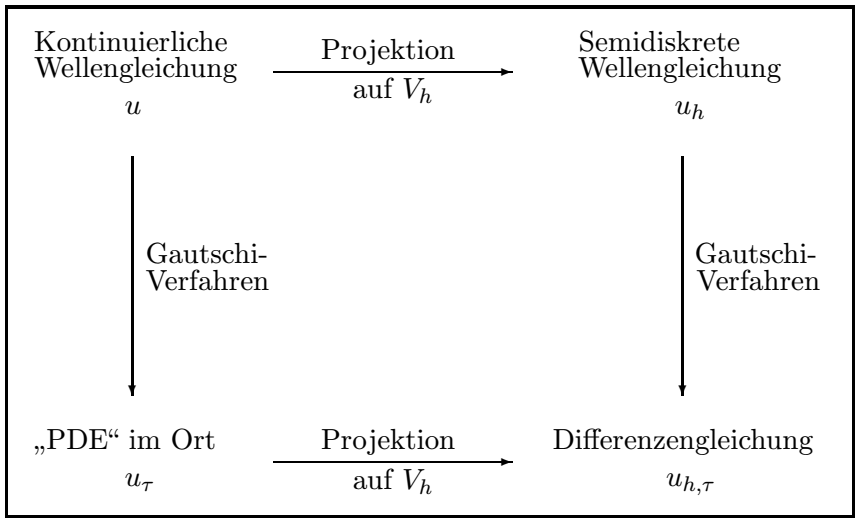

Abbildung 2.2: Konvergenzdiagramm für die Wellengleichung

Unser Ziel ist es nun, eine Abschätzung der Form

$$\|u - u_{h,\tau}\|_{L^2} \leq C(h^2 + \tau^2)$$

zu beweisen.

Es gilt $\|u - u_{h,\tau}\|_{L^2} \leq \|u - u_h\|_{L^2} + \|u_h - u_{h,\tau}\|_{L^2}$.
Die Galerkin-Konvergenz für zeitabhängige Funktionen $u_h \to u$ für $h \to 0$ wurde bereits in Abschnitt 2.2 gezeigt: $\|u - u_h\|_{L^2} \leq Ch^2$.

Die Konvergenz $u_{h,\tau} \to u_h$ für $\tau \to 0$ mit der damit verbundenen Abschätzung in der L^2-Norm verhält sich nun wie folgt:
Die Knotendarstellung der semidiskreten Wellengleichung

$$\ddot{\mathbf{u}}_h + c^2 M_h^{-1} A_h \mathbf{u}_h = \mathbf{f}_h \tag{2.31}$$

ist äquivalent zu

$$\ddot{\mathbf{v}}_h + c^2 M_h^{-1/2} A_h M_h^{-1/2} \mathbf{v}_h = M_h^{1/2} \mathbf{f}_h \tag{2.32}$$

mit $\mathbf{v}_h = M_h^{1/2} \mathbf{u}_h$.
Die Matrix $M_h^{-1/2} A_h M_h^{-1/2}$ ist mit M_h und A_h ebenfalls symmetrisch positiv definit, was die Anwendung des Gautschi-Algorithmus in der in Kapitel 2.4 vorgestellten Version auf (2.32) mit der in Abschnitt 2.4.3 bewiesenen Fehlerschätzung in der euklidischen Norm ermöglicht.
Nun ist jedoch

$$\|\mathbf{v}_{h,n} - \mathbf{v}_h(t_n)\|_2 = \|M_h^{1/2}(\mathbf{u}_{h,n} - \mathbf{u}_h(t_n))\|_2 = \|u_{h,n} - u_h(t_n)\|_{L^2},$$

womit der Übergang von $(\mathbb{R}^N, \langle\cdot,\cdot\rangle_2)$ nach $(\mathbb{R}^N, \langle\cdot,\cdot\rangle_{M_h})$ sowie nach $(V_h, (\cdot,\cdot)_{L^2})$ hergestellt ist.
Der Gautschi-Algorithmus kann also unter Verwendung von $\langle\cdot,\cdot\rangle_{M_h}$ auch direkt auf (2.31) angewendet werden – Die Matrix $M_h^{-1} A_h$ ist nämlich bzgl. $\langle\cdot,\cdot\rangle_{M_h}$ symmetrisch positiv definit! – wie auch unter Verwendung von $(\cdot,\cdot)_{L^2}$ auf die basisfreie Darstellung (2.29).

Für festes $h > 0$ gilt also nach Abschnitt 2.4.3 $\|u_h - u_{h,\tau}\|_{L^2} \leq C_h \tau^2$, womit noch die Gleichmäßigkeit in h der Konvergenz in der Zeit zu zeigen bleibt:
Für die diskrete Lösung $u_{h,n}$ der Differentialgleichung

$$\ddot{u}_h + c^2 A_h^{op} u_h = f_h$$

gilt nach Theorem 2.5

$$\|u_{h,n} - u_h(t_n)\|_{L^2} \leq \tau^2 \left(\frac{M_{h,1}}{6} t_n + \frac{M_{h,2}}{12} t_n^2 \right)$$

mit

$$\max_{0 \leq t \leq T} \|f_h^{(k)}(t)\|_{L^2} \leq M_{h,k}, \;\; k = 1, \ldots, 3.$$

Nun gilt unter der Voraussetzung

$$\max_{0 \leq t \leq T} \|f^{(k)}(t,x)\|_{L^2} \leq M_k, \;\; k = 1, \ldots, 3$$

mit $f_h = Q_h f$ wegen der Vertauschbarkeit von $\frac{\partial}{\partial t}$ und Q_h

$$\begin{aligned} \|f_h^{(k)}(t)\|_{L^2} &= \|\tfrac{\partial^k}{\partial t^k}(Q_h f(t,x))\|_{L^2} = \|Q_h \tfrac{\partial^k}{\partial t^k} f(t,x)\|_{L^2} \\ &\leq \|\tfrac{\partial^k}{\partial t^k} f(t,x)\|_{L^2} \leq M_k \text{ unabhängig von } h. \end{aligned}$$

Aufgrund der Unabhängigkeit der Konstanten M_k von h gilt also

$$\|u_h - u_{h,\tau}\|_{L^2} \leq C\tau^2$$

und somit insgesamt unsere angestrebte Abschätzung für die diskrete Lösung

$$\|u - u_{h,\tau}\|_{L^2} \leq C(h^2 + \tau^2).$$

□

Zur Kommutativität des Diagramms in Abbildung 2.2:

Besteht die Projektion von $H_0^1(\Omega)$ auf den FE-Raum V_h in der Ersetzung

$$u \to u_h, \;\; A \to A_h^{op} = Q_h A \text{ sowie } \; f \to f_h = Q_h f,$$

dann kommutiert das Diagramm in Abbildung 2.2:

1. Weg: Projiziere zunächst in obigem Sinne auf V_h und wende dann den Gautschi-Algorithmus auf die im Ort semidiskrete Wellengleichung an:

$$\begin{aligned}
&\ddot{u} + c^2 A u = f \to \\
&\ddot{u}_h + c^2 A_h^{op} u_h = f_h \to \\
&u_{h,n+1} - 2u_{h,n} + u_{h,n-1} = \tau^2 \sigma(\tau^2 c^2 A_h^{op})(-c^2 A_h^{op} u_{h,n} + f_{h,n}) \text{ mit} \\
&u_{h,1} = \cos(\tau c\sqrt{A_h^{op}})\, u_{h,0} + (c\sqrt{A_h^{op}})^{-1} \sin(\tau c\sqrt{A_h^{op}})\, \dot{u}_{h,0} + \\
&\qquad \tfrac{1}{2}\tau^2 \sigma(\tau^2 c^2 A_h^{op})\, f_{h,0}
\end{aligned}$$

2. Weg: Wende zunächst den Gautschi-Algorithmus auf die kontinuierliche Wellengleichung an und projiziere dann in obigem Sinne auf V_h:

$$\begin{aligned}
&\ddot{u} + c^2 A u = f \to \\
&u_{n+1} - 2u_n + u_{n-1} = \tau^2 \sigma(\tau^2 c^2 A)(-c^2 A u_n + f_n) \text{ mit} \\
&u_1 = \cos(\tau c\sqrt{A})\, u_0 + (c\sqrt{A})^{-1} \sin(\tau c\sqrt{A})\, \dot{u}_0 + \\
&\qquad \tfrac{1}{2}\tau^2 \sigma(\tau^2 c^2 A)\, f_0 \to \\
&u_{n+1,h} - 2u_{n,h} + u_{n-1,h} = \tau^2 \sigma(\tau^2 c^2 A_h^{op})(-c^2 A_h^{op} u_{n,h} + f_{n,h}) \text{ mit} \\
&u_{1,h} = \cos(\tau c\sqrt{A_h^{op}})\, u_{0,h} + (c\sqrt{A_h^{op}})^{-1} \sin(\tau c\sqrt{A_h^{op}})\, \dot{u}_{0,h} + \\
&\qquad \tfrac{1}{2}\tau^2 \sigma(\tau^2 c^2 A_h^{op})\, f_{0,h}
\end{aligned}$$

Nachdem

$$f_{h,n} = f_h(t_n) = Q_h f(t_n, x) = Q_h f_n(x) = f_{n,h} \;\; \forall \;\; n \in \mathbb{N}_0$$

sowie

$$\begin{aligned}
u_{0,h} &= Q_h u_0(x) = u_{h,0} \text{ und} \\
\dot{u}_{0,h} &= Q_h \dot{u}_0(x) = \dot{u}_{h,0}
\end{aligned}$$

gilt, folgt die Gleichheit $u_{h,n} = u_{n,h}$ für alle $n \in \mathbb{N}_0$.

Kapitel 3

Hierarchische Matrizen

Wir haben bisher in den Kapiteln 1 und 2 die inhomogene 2D Wellengleichung betrachtet und diese im Ort auf Finite Elemente projiziert sowie in der Zeit mittels eines exponentiellen Integrators diskretisiert.
Der Algorithmus zur Bestimmung der diskreten Lösung enthält dabei mehrere Matrix-Vektor-Produkte mit transzendenten Matrixfunktionen, wobei die Dimension dieser Matrizen gerade der Anzahl der Unbekannten auf dem regelmäßig triangulierten 2D Einheitsquadrat entspricht:

Nach Anwendung des Gautschi-Verfahrens auf die aus (2.6) durch Multiplikation von links mit der Inversen der Massenmatrix M_h^{-1} resultierende Differentialgleichung 2. Ordnung

$$\ddot{\mathbf{u}}_h + c^2 M_h^{-1} A_h \mathbf{u}_h = \mathbf{f}_h \tag{3.1}$$

müssen die drei Matrixfunktion-Vektor-Produkte $\cos(\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{v}_h$, $\psi(\tau^2 c^2 M_h^{-1}A_h)\,\mathbf{v}_h$ und $\sigma(\tau^2 c^2 M_h^{-1}A_h)\,\mathbf{v}_h$ mit den transzendenten Matrixfunktionen $\cos(x)$, $\psi(x^2) = \frac{\sin x}{x}$ und $\sigma(x^2) = \left(\frac{\sin \frac{1}{2}x}{\frac{1}{2}x}\right)^2 = 2\,\frac{1-\cos x}{x^2}$ sowie Vektoren $\mathbf{v}_h$ der Dimension $N = n^2$ berechnet werden.

Das Umgehen der Stabilitätsprobleme bei expliziten und der Phasenfehler bei impliziten durch Verwendung transzendentaler Verfahren hat somit zu einer Verlagerung der Schwierigkeiten von der Diskretisierung auf die algebraische Ebene geführt.

Wir werden zunächst in Abschnitt 3.1 zeigen, dass obige Matrixfunktion-Vektor-Produkte für große Wellengeschwindigkeiten c und kleine Gitterbreite h mittels Krylovraummethoden nicht approximiert werden können.

In Abschnitt 3.2 führen wir die von Hackbusch in [19] und [20] entwickelten $\mathcal{H}$-Matrizen anhand des Beispiels einer Rang-k-Matrix ein.

In Abschnitt 3.3 werden wir ein für unsere Zwecke passendes hierarchisches Format in Anlehnung an die in [20] aufgeführte Konstruktion angeben. Der einzige wesentliche Unterschied zu [20, Kapitel 4] besteht dabei in der Verwendung stückweise linearer C^0-Elemente auf dem regelmäßig triangulierten

Einheitsquadrat statt stückweise konstanter Funktionen bzgl. des regelmäßigen Rechtecksgitters auf dem Einheitsquadrat.

In Abschnitt 3.4 wird die gesamte approximierte $\mathcal{H}$-Arithmetik von der einfachsten Operation, der Matrix-Vektor-Multiplikation, bis hin zur QR-Zerlegung und Polarzerlegung behandelt. Für die späteren Anwendungen sind dabei alle Operationen mit Ausnahme der QR-Zerlegung und Polarzerlegung relevant. Eine detaillierte Übersicht zur $\mathcal{H}$-Arithmetik wird am Anfang des betreffenden Abschnitts gegeben.

3.1 Krylovraummethoden zur Berechnung von Matrixfunktion-Vektor-Produkten

Die Betrachtung von beispielsweise $\sigma(\tau^2 c^2 M_h^{-1} A_h)\,\mathbf{v}_h$ führt mehrere Probleme vor Augen:

- $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ ist eine transzendente Matrixfunktion.
- Die Matrizen M_h und A_h haben die „große“ Dimension $N = n^2$.
- Die Berechnung der Inversen von M_h ist daher viel zu aufwendig und würde überdies die schwache Besetztheit von M_h zerstören.

Abhilfe hierfür bieten Krylovraummethoden zur Approximation des Produktes einer Matrixfunktion mit einem Vektor (siehe z.B. [7], [28], [30]). Diese Technik besteht in der orthogonalen Projektion der Matrix vom $\mathbb{R}^N$ auf einen Krylovunterraum $K_m = K_m(M_h^{-1} A_h, \mathbf{v}_h)$ der Dimension $m \ll N$ und der Berechnung des Matrixfunktion-Vektor-Produktes in K_m, genauer: $\sigma(\tau^2 c^2 M_h^{-1} A_h)\,\mathbf{v}_h \approx V_m\,\sigma(\tau^2 c^2 H_m)\,V_m^T\,\mathbf{v}_h$, wobei die Matrix $V_m = [\mathbf{v}_1, \ldots, \mathbf{v}_m] \in \mathbb{R}^{N\times m}$ die Arnoldi-Orthonormalbasis von $K_m(M_h^{-1} A_h, \mathbf{v}_h)$ enthält und H_m die orthogonale Projektion von $M_h^{-1} A_h$ auf K_m ist (H_m ist eine obere Hessenbergmatrix.) (siehe z.B. [30]).

Die Berechnung der Approximation über die Krylovraumprojektion gliedert sich wie folgt:

1. Berechnung von V_m und H_m mittels des Arnoldi-Verfahrens (siehe z.B. [40])
2. Skalierung von $\tau^2 c^2 H_m$, so dass $\|4^{-k}\tau^2 c^2 H_m\|_2 \leq \frac{1}{2}$
3. Berechnung von $\sigma(4^{-k}\tau^2 c^2 H_m)$ mittels Taylorreihe bis zur gewünschten Genauigkeit und daraus rekursive Ermittlung von $\sigma(\tau^2 c^2 H_m)$ über den Duplikationsalgorithmus

$$\sigma(4x^2) = \sigma(x^2)\left(1 - \frac{1}{4}\,x^2\sigma(x^2)\right)$$

4. Berechnung von $V_m\,\sigma(\tau^2 c^2 H_m)\,\mathbf{e}_1\,\|\mathbf{v}_h\|_2$.

Bemerkung: Die Attraktivität der Krylovraummethoden liegt in deren superlinearen Konvergenz begründet, die i. Allg. schon ab kleinen Iterationszahlen m auftritt.

Untersuchung der Konvergenz der Krylovraumiteration für oszillierende Funktionen wie $\cos(x)$, $\psi(x^2)$ oder $\sigma(x^2)$:

Nach [30, Theorem 4] ist der Fehler der Arnoldi-Approximation an $e^{\tau A}\mathbf{v}$ für antihermitesche Matrizen $A \in \mathbb{C}^{N\times N}$ mit Eigenwerten in einem Intervall der Länge 4ρ auf der imaginären Achse und normiertem $\mathbf{v} \in \mathbb{C}^N$ durch

$$\epsilon_m := \|e^{\tau A}\mathbf{v} - V_m e^{\tau H_m}\mathbf{e}_1\|_2 \leq 12e^{-\frac{(\rho\tau)^2}{m}}\left(\frac{e\rho\tau}{m}\right)^m, \;\; m \geq 2\rho\tau$$

beschränkt und für $m < \rho\tau$ i. Allg. keine substantielle Fehlerverringerung gegeben.

Wir kommen nun zu den Krylovraumapproximationen an $\cos(\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{v}_h$, $\psi(\tau^2c^2M_h^{-1}A_h)\,\mathbf{v}_h$ und $\sigma(\tau^2c^2M_h^{-1}A_h)\,\mathbf{v}_h$.
Für den größten Eigenwert von $M_h^{-1}A_h$ gilt

$$\begin{aligned}\lambda_{\max}(M_h^{-1}A_h) &= \max_{\mathbf{v}_h\neq 0}\frac{\langle M_h^{-1}A_h\mathbf{v}_h,\mathbf{v}_h\rangle_2}{\langle \mathbf{v}_h,\mathbf{v}_h\rangle_2} = \max_{\mathbf{v}_h\neq 0}\frac{\langle M_h^{-\frac{1}{2}}A_hM_h^{-\frac{1}{2}}\mathbf{v}_h,\mathbf{v}_h\rangle_2}{\langle \mathbf{v}_h,\mathbf{v}_h\rangle_2}\\ &= \max_{\mathbf{u}_h\neq 0}\frac{\langle A_h\mathbf{u}_h,\mathbf{u}_h\rangle_2}{\langle M_h\mathbf{u}_h,\mathbf{u}_h\rangle_2} = \max_{u_h\in V_h\setminus\{0\}}\frac{|u_h|_{H^1}^2}{\|u_h\|_{L^2}^2}\\ &\leq Ch^{-2}\end{aligned}$$

unter Ausnutzung der inversen Ungleichung

$$|u_h|_{H^1}^2 \leq Ch^{-2}\|u_h\|_{L^2}^2$$

im letzten Schritt und damit

$$\|M_h^{-1}A_h\|_2 \geq \lambda_{\max}(M_h^{-1}A_h) = \mathcal{O}(h^{-2}) = \mathcal{O}(N).$$

Der kleinste Eigenwert von $M_h^{-1}A_h$ ist wegen

$$\lambda_{\min}(M_h^{-1}A_h) = \min_{\mathbf{u}_h\neq 0}\frac{\langle A_h\mathbf{u}_h,\mathbf{u}_h\rangle_2}{\langle M_h\mathbf{u}_h,\mathbf{u}_h\rangle_2} = \min_{u_h\in V_h\setminus\{0\}}\frac{|u_h|_{H^1}^2}{\|u_h\|_{L^2}^2} \geq \frac{1}{C} > 0$$

mit C unabhängig von h, wobei im letzten Schritt die Poincaré-Friedrichs-Ungleichung

$$\|u_h\|_{L^2} \leq C|u_h|_{H^1}$$

benutzt wurde.
Sämtliche Eigenwerte der Matrix $i\tau c\sqrt{M_h^{-1}A_h}$ liegen also auf der imaginären Achse in einem Intervall der Länge $C\tau ch^{-1}$. Weiter ist $i\tau c\sqrt{M_h^{-1}A_h}$ bzgl. $\langle\cdot,\cdot\rangle_{M_h}$ antihermitesch: Es existiert nämlich eine eindeutige bzgl. $\langle\cdot,\cdot\rangle_{M_h}$ symmetrisch positiv definite Quadratwurzel $\sqrt{M_h^{-1}A_h}$ aus $M_h^{-1}A_h$, womit sofort

$$\langle i\tau c\sqrt{M_h^{-1}A_h}\mathbf{u}_h,\mathbf{v}_h\rangle_{M_h} = -\langle \mathbf{u}_h, i\tau c\sqrt{M_h^{-1}A_h}\mathbf{v}_h\rangle_{M_h}$$

für alle $\mathbf{u}_h, \mathbf{v}_h \in \mathbb{R}^N$ folgt.
Somit ist für die Krylovraumapproximation an $\cos(\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{v}_h = \Re\left(\exp(i\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{v}_h\right)$ nach [30] superlineare Konvergenz erst ab einer Iterationszahl $m_0 \eqsim \tau c h^{-1}$ zu erwarten.

Numerische Experimente für die drei obengenannten Matrixfunktion-Vektor-Produkte haben dieses Konvergenzverhalten bestätigt (siehe Abbildung 3.1 für das Matrixfunktion-Vektor-Produkt $\cos(\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{v}_h$).

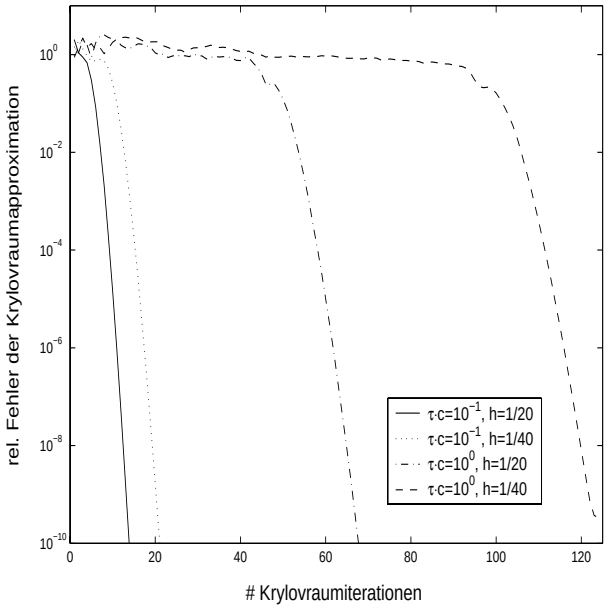

Abbildung 3.1: Relativer Fehler der Krylovraumapproximation an $\cos(\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{v}_h$, wobei $\mathbf{v}_h$ ein Zufallsvektor der Länge 1 ist, für $\tau c = 10^{-1}$ und 10^0 sowie $h = \frac{1}{20}$ und $\frac{1}{40}$.

Dies liefert also eine Schrittweitenbeschränkung der Form $\tau \leq Chc^{-1}$ (vom Typ der CFL-Bedingung bei expliziten Verfahren).

Nachdem unser Hauptziel gerade darin besteht, die Zeitschrittweite τ unabhängig von h und c wählbar zu machen, müssen wir nach anderen Möglichkeiten suchen, um obige Matrixfunktion-Vektor-Produkte zu berechnen.
Diese Suche führte schließlich zu den hierarchischen Matrizen und deren approximierter Arithmetik.

3.2 Einführung in $\mathcal{H}$-Matrizen

Die hierarchischen Matrizen ($\mathcal{H}$-Matrizen) wurden von Hackbusch in [19] und [20] eingeführt (siehe weiter [21], [22], [23], [24], ...).
$\mathcal{H}$-Matrizen haben folgende Eigenschaften:

1. Sie sind schwach besetzt im Sinne dass nur wenige Daten zu deren Darstellung notwendig sind.

2. I. Allg. sind exakte Summen, Produkte und Inverse von $\mathcal{H}$-Matrizen nicht länger im $\mathcal{H}$-Format, jedoch deren Realisierungen in einer approximierten Arithmetik ($\mathcal{H}$-Arithmetik).

3. Die Matrix-Vektor-Multiplikation und die approximierte Addition, Multiplikation und Invertierung von $\mathcal{H}$-Matrizen sind von fast linearer Komplexität.

Ein einführendes Beispiel:
Einfachstes Beispiel für eine $\mathcal{H}$-Matrix ist eine Rang-k-Matrix (Rk-Matrix), das ist eine Matrix mit höchstens Rang k.
Eine Rk-Matrix $R \in \mathbb{R}^{n\times m}$ besitzt die Darstellung

$$R = \sum_{i=1}^{k} \mathbf{a}_i \mathbf{b}_i^T = \mathbf{a}\mathbf{b}^{\mathbf{T}} = [\mathbf{a}, \mathbf{b}]$$

mit $\mathbf{a} = [\mathbf{a_1}, \ldots, \mathbf{a_k}] \in \mathbb{R}^{\mathbf{n}\times\mathbf{k}}$ und $\mathbf{b} = [\mathbf{b_1}, \ldots, \mathbf{b_k}] \in \mathbb{R}^{\mathbf{m}\times\mathbf{k}}$ als Summe von k Rang-1-Matrizen $\mathbf{a}_i\mathbf{b}_i^T,\ i = 1, \ldots, k$.

Eigenschaften von Rk-Matrizen:

1. Zur Darstellung bzw. Speicherung einer Rk-Matrix R genügen $k(n + m)$ Elemente.

2. Die Matrix-Vektor-Multiplikation $R\mathbf{v}$ einer Rk-Matrix R mit einem Vektor $\mathbf{v}$ benötigt nur $\mathcal{O}(n + m)$ Operationen.
   Die Multiplikation einer Rk-Matrix R von rechts oder von links mit einer beliebigen vollbesetzten Matrix M liefert wieder eine Rk-Matrix, wobei der Aufwand für letztere Operation gerade dem von k Matrix-Vektor-Multiplikationen mit der Matrix M entspricht.

3. Die Summe zweier Rk-Matrizen R_1 und R_2 ist i. Allg. keine Rk-Matrix mehr, sondern eine Rang-$2k$-Matrix. Für eine geeignete Rk-Approximation an die exakte Summe $R_1 + R_2$ betrachte man die Singulärwertzerlegung

   $$R_1 + R_2 = U\Sigma V^T \ \text{ mit } \ \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{2k}, 0, \ldots, 0),$$

   wobei $\sigma_1 \geq \ldots \geq \sigma_{2k} \geq 0$, und setze die k kleineren Singulärwerte $\sigma_{k+1}, \ldots, \sigma_{2k}$ auf Null. Die resultierende Rk-Matrix

   $$R_1 +_{Rk} R_2 = U\Sigma' V^T \ \text{ mit } \ \Sigma' = \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$$

ist die Bestapproximation an $R_1 + R_2$ in der Frobeniusnorm sowie in der Spektralnorm. Der Approximationsfehler in der Frobeniusnorm entspricht dabei $\sqrt{\sum_{i=k+1}^{2k} \sigma_i^2}$, der in der Spektralnorm ist gleich σ_{k+1} (siehe [19]).

Berechnung der Rang-k-Bestapproximation an $R_1 + R_2$:
Um die Rk-Approximation an die exakte Summe zweier Rk-Matrizen zu berechnen, muss nicht die (viel zu aufwendige) Singulärwertzerlegung der vollbesetzten Matrix $R_1 + R_2$ berechnet werden, sondern es genügt die Lösung eines Eigenwertproblems der Dimension $2k$ (siehe [19]) bzw. die Singulärwertzerlegung einer Matrix in $\mathbb{R}^{2k \times 2k}$ (siehe [17]).
Erfolgt die Lösung über das $2k$-dimensionale Eigenwertproblem, müssen die daraus erhaltenen Eigenvektoren der symmetrischen (!) Matrix $R^T R$, $R = R_1 + R_2$, mittels orthogonaler Iteration nachverbessert werden, da die Matrix des $2k$-dimensionalen Eigenwertproblems unsymmetrisch ist.

Bemerkungen:

1. Der Aufwand für die approximierte Addition zweier Rk-Matrizen liegt bei $\mathcal{O}(n+m) + \mathcal{O}(1)$, wobei das $2k$-dimensionale Eigenwertproblem bzw. die $2k$-dimensionale Singulärwertzerlegung im $\mathcal{O}(1)$-Term enthalten ist.

2. Für die Fehlerschätzung ist kein zusätzlicher Aufwand nötig, da die dafür notwendigen abgeschnittenen Singulärwerte schon im Zuge des $2k$-dimensionalen Eigenwertproblems bzw. der $2k$-dimensionalen Singulärwertzerlegung berechnet wurden.

$\mathcal{H}$-Matrizen sind nun im Wesentlichen hierarchisch strukturierte Blockmatrizen, deren Blöcke aus vollbesetzten Matrizen (hauptsächlich nahe der Diagonalen) und aus Rang-k-Matrizen niedrigen Ranges bestehen. Die Blockgestalt einer solchen $\mathcal{H}$-Matrix, die Lage der vollbesetzten und der Rk-Blöcke und der Rang der Rk-Blöcke hängen dabei vom jeweiligen Problem und der gewünschten Approximationsgüte ab.

3.3 FE-Matrizen als $\mathcal{H}$-Matrizen

Wir werden im Folgenden eine geeignete Klasse von $\mathcal{H}$-Matrizen für die Darstellung der Steifigkeitsmatrix A_h und der Massenmatrix M_h bzgl. des regelmäßig triangulierten 2D Einheitsquadrates samt ihrer Inversen definieren.
Mit A_h und M_h als $\mathcal{H}$-Matrizen können dann sämtliche Matrixoperationen in der approximierten Arithmetik durchgeführt werden.
In [20] wurden geeignete Block-Partitionierungen für den Fall stückweise konstanter Finiter Elemente zunächst auf dem regelmäßigen Rechtecksgitter in $\Omega = (0,1)^2$ konstruiert (siehe [20, Kapitel 4]). Allgemeine 2D Gitter wurden dann einfach auf das regelmäßige quadratische Gitter projiziert und dessen bereits konstruierte Hierarchie auf das unregelmäßige Gitter „zurückgespielt“ (siehe [20, Kapitel 5]).

Auf diese Weise werden wir nun eine geeignete Block-Partitionierung für das

unser Problem betreffende regelmäßig triangulierte Einheitsquadrat mit stückweise linearen Finiten Elementen gewinnen.

3.3.1 Definition einer geeigneten Klasse von $\mathcal{H}$-Matrizen

Sei $I = \{(i,j) : 1 \leq i,j \leq n\}$ die n^2-elementige Indexmenge aller inneren Knoten von Ω (siehe Abbildung 3.2).

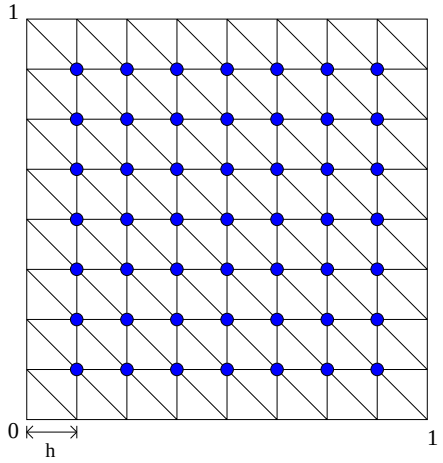

Abbildung 3.2: Regelmäßig trianguliertes Einheitsquadrat mit $h = \frac{1}{n+1}$

Definition 3.1 ($\mathcal{H}$-Baum)
$T = T(I)$ ist ein $\mathcal{H}$-Baum über I $:\Longleftrightarrow$

1. *$I \in T$ Wurzel*
2. *Alle Scheitel $t \in T$ sind Teilmengen von I*
3. *Falls t kein Blatt ist, ist t disjunkte Vereinigung seiner Söhne $S(t)$:*

$$t = \dot{\bigcup_{s \in S(t)}} s$$

Definition 3.2 (T-Block-Partitionierung)
$P = \{I_j : 1 \leq j \leq k\}$ heißt T-Block-Partitionierung von I $:\Longleftrightarrow$

$$I = \dot{\bigcup_{j=1}^{k}} I_j \ \text{ und } \ P \subset T.$$

Bemerkung: Jede T-Block-Partitionierung P entspricht eineindeutig den Blättern eines $\mathcal{H}$-Teilbaums T' von T: $P = \mathcal{L}(T')$.

Wir konstruieren den zum regelmäßigen Dreiecksgitter gehörigen $\mathcal{H}$-Baum

$T_1 = T(I)$, indem wir zunächst das Einheitsquadrat, dem die gesamte Indexmenge I als Wurzel (Scheitel vom Level 0) entspricht, in 4 gleich große Quadrate unterteilen. Die jeweils darin enthaltenen Gitterpunkte bilden nun die 4 Söhne von I (die 4 Scheitel vom Level 1), wobei die auf einer Trennlinie liegenden Knoten im Folgenden stets zum Quadrat links bzw. unterhalb des Knotens geschlagen werden. Dieses Verfahren wird nun rekursiv auf die 4 soeben erhaltenen Quadrate solange angewendet, bis durch eine weitere Unterteilung leere Quadrate entstehen würden.

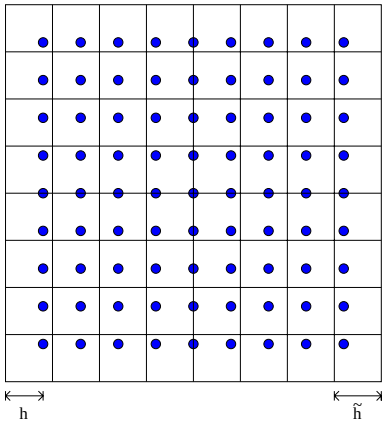

Abbildung 3.3: $n^2 = 9^2$ innere Knoten sowie $\tilde{h}^{-2} = 8^2$ Teilquadrate vom Level $p = 3$ des 2D Einheitsquadrates Ω. Letztere entsprechen den Blättern des $\mathcal{H}$-Baums T_1.

Wird auf diese Weise dem regulären Dreiecksgitter das reguläre quadratische Gitter mit der Gitterbreite

$$\tilde{h} = 2^{-p} \text{ mit } p \text{ definiert durch } 2^p \leq n < 2^{p+1}$$

zugrunde gelegt, dann sind alle Blöcke bis zum Level p nichtleer.
So ist beispielsweise für $n = 9$ in Abbildung 3.3 $p = 3$ und alle quadratischen Blöcke bis zum Level 3 – das sind all jene bis zu einer Seitenlänge von 2^{-3} – sind nichtleer.

Konstruktion von $T_2 = T(I \times I)$ mittels T_1:

Wir starten diesmal mit $I \times I$ anstelle von I und definieren die 16 Söhne b_{ij} von $I \times I$ durch $b_{ij} = (t_i, t_j)$, t_i, t_j Sohn von $I, 1 \leq i, j \leq 4$. Die Konstruktion wird nun bis zum höchsten Level p von T_1 weitergeführt. Der damit erhaltene 2D $\mathcal{H}$-Baum $T_2 = T(I \times I)$ besitzt natürlich dieselbe Tiefe wie der 1D $\mathcal{H}$-Baum $T_1 = T(I)$.

Wir konstruieren nun eine unserer Problemstellung angepasste Block-Partitionierung P_2 von $I \times I$ als disjunkte Vereinigung von Teilmengen von

$I \times I$. Wir verwenden dabei das Zulässigkeitskriterium

$$\min\{\operatorname{diam}(t_1), \operatorname{diam}(t_2)\} \leq 2\eta \operatorname{dist}(t_1, t_2) \tag{3.2}$$

für einen Block $b = (t_1, t_2) \in T_2$ mit $\eta = \frac{\sqrt{2}}{2}$ (siehe [20]). Während $\operatorname{diam}(t_i)$ den Durchmesser des zu $t_i \in T_1$ gehörigen Teilquadrates bezeichnet, steht $\operatorname{dist}(t_1, t_2)$ für den minimalen Abstand der zu t_1 und t_2 gehörigen Teilquadrate.

Definition 3.3 (Zulässigkeit von Blöcken)
Ein Block $b = (t_1, t_2) \in T_2$ heißt zulässig, falls er entweder ein Blatt ist oder die Zulässigkeitsbedingung (3.2) erfüllt.

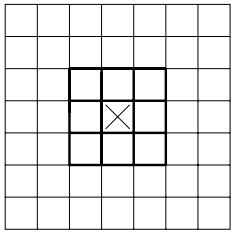

Abbildung 3.4: Unzulässige Blöcke für einen gegebenen Block „×“ für $\eta = \frac{\sqrt{2}}{2}$

Abbildung 3.4 zeigt die unzulässigen Blöcke für einen gegebenen Block für $\eta = \frac{\sqrt{2}}{2}$.

Nachdem in jedem Teilquadrat des zugrundegelegten quadratischen Gitters mindestens ein Knoten liegt und die Träger der nodalen Basisfunktionen ψ_j^h gerade aus den am Knoten j angrenzenden Dreiecken bestehen (siehe Abbildung 3.5), garantiert die Wahl von $\eta = \frac{\sqrt{2}}{2}$ die exakte Darstellbarkeit der FE-Matrizen A_h und M_h als hierarchische Matrizen (siehe Abbildung 3.4): Die Matrixelemente $(M_h)_{ij}$ und $(A_h)_{ij}$ sind nämlich gleich 0, falls die Träger der nodalen Basisfunktionen ψ_i^h und ψ_j^h nur Randpunkte gemeinsam haben.

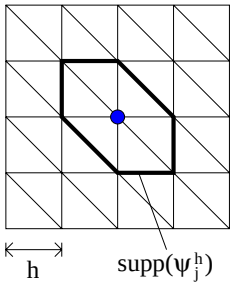

Abbildung 3.5: Träger der nodalen Basisfunktion ψ_j^h

Wir suchen nun nach der minimalen zulässigen Block-Partitionierung P_2 von $I \times I$, d.h. nach jener zulässigen Partitionierung mit der minimalen Anzahl von Blöcken bzw. mit den größtmöglichen Blöcken.
Wir starten beim Block $I \times I$. Da dieser nicht zulässig ist, betrachten wir dessen 16 Söhne. Nachdem diese auch nicht zulässig sind, gehen wir zu deren je 16 Söhnen über. Die zulässigen darunter werden in P_2 archiviert, die nicht zulässigen weiter durch deren 16 Söhne ersetzt, bis wir an den Blättern von T_2 angelangt sind.
Blöcke aus P_2, die die Zulässigkeitsbedingung (3.2) erfüllen, werden nun durch Rk-Matrizen dargestellt, alle übrigen Blöcke (sie sind Teil der Blätter von T_2) durch vollbesetzte Matrizen. Dies führt uns zur Definition einer für unser Problem geeigneten Klasse von $\mathcal{H}$-Matrizen:

Definition 3.4 ($\mathcal{H}$-Matrizen)
Sei P_2 eine Block-Partitionierung von $I \times I$. Dann definiert

$$\mathcal{M}_{\mathcal{H},k}(I \times I, P_2) := \{M \in \mathbb{R}^{I \times I} : \mathit{Rang}(M^b) \leq k(b) \text{ } \mathit{für\ jeden\ Block}\ M^b \\ \mathit{mit}\ b \in P_2\ \mathit{erfüllt\ die\ Zulässigkeitsbedingung\ (3.2)}\}$$

die Menge der von P_2 induzierten $\mathcal{H}$-Matrizen.

Bemerkungen:

1. Was die Anordnung der Indizes in einer $\mathcal{H}$-Matrix M betrifft, wird immer mit dem linken oberen Viertel des Einheitsquadrates begonnen und im Uhrzeigersinn bis zum linken unteren Viertel fortgefahren. Dabei werden ihrerseits geviertelte Teilquadrate auf dieselbe Weise abgearbeitet (siehe Abbildung 3.6). Die Indizes in den einzelnen Matrixblöcken einer $\mathcal{H}$-Matrix werden dann im jeweiligen Teilquadrat von $(0,1)^2$ zeilenweise von links unten nach rechts oben angeordnet, wie wir es vom 5-Punkte-Stern der Diskretisierung des Laplace-Operators auf dem regelmäßigen Dreiecksgitter kennen.

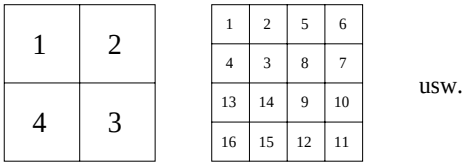

Abbildung 3.6: Indizierung der Blöcke in einer $\mathcal{H}$-Matrix

2. Mit der Festlegung auf eine Block-Partitionierung P_2 ist die Menge $\mathcal{M}_{\mathcal{H},k}(I \times I, P_2)$ bereits eindeutig definiert.

3. Der Rang k der Rk-Blöcke M^b ist variabel ($k = k(b)$).

4. FE-Matrizen als $\mathcal{H}$-Matrizen: Die altbekannten zeilenweise von links unten nach rechts oben durchnumerierten FE-Matrizen A_h und M_h wurden durch geeignetes Umordnen der Zeilen und Spalten, d.h. durch Konjugation mit einer geeigneten Permutationsmatrix, auf hierarchische Struktur gebracht:
   Für eine hierarchische Matrix $C_{\mathcal{H}} \in \mathcal{M}_{\mathcal{H},k}(I \times I, P_2)$ gilt nämlich gerade $C_{\mathcal{H}} = PCP^T$ mit einer geeigneten Permutationsmatrix P, wobei C die Matrix mit der üblichen zeilenweisen Anordnung von links unten nach rechts oben bezeichnet. Wir werden nun im Folgenden mit A_h und M_h stets deren hierarchisierte Versionen bezeichnen.

Verallgemeinerung:

Wir können nun statt des quadratischen Gitters mit $\tilde{h} = 2^{-p}$ allgemeiner die Gitterbreite $\tilde{h} = 2^{-p+\delta p}$ mit $\delta p \geq 0$ zugrundelegen. Das führt dazu, dass wir mit der rekursiven Viertelung des quadratischen Referenzgitters bereits um δp Levels früher aufhören und somit die Tiefe der resultierenden $\mathcal{H}$-Bäume um δp verringern.
Dadurch wird ebenso die Tiefe der rekursiven Struktur der $\mathcal{H}$-Matrizen um δp verringert; dafür haben aber die kleinsten Matrixblöcke (sie entsprechen den Blättern von P_2) größere Dimension: Für $\delta p \geq 0$ gilt für die Dimension $\dim(M^b)$ der Matrixblöcke M^b vom größten Level $p - \delta p$

$$4^{\delta p} \leq \dim(M^b) \leq 4^{\delta p+1}.$$

Wir werden diese „effektive“ Tiefe im Folgenden mit $p_{eff} = p - \delta p$ bezeichnen. Die reguläre Struktur des quadratischen Referenzgitters, das für die Blockstruktur der $\mathcal{H}$-Matrizen verantwortlich ist, erlaubt eine direkte rekursive Definition der $\mathcal{H}$-Matrizen durch 9 hierarchische Formate, einem Diagonalformat $\square$ und 8 Nachbarformaten $\leftarrow, \rightarrow, \uparrow, \downarrow, \nearrow, \searrow, \swarrow$ und $\nwarrow$ (siehe [20]). Diese werden bei der nachfolgenden Behandlung einiger hierarchischer Matrixoperationen eine gewichtige Rolle spielen.
Im Diagonalformat

$$\square = \begin{array}{|c|c|c|c|} \hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline \end{array}$$

spiegelt sich die Abarbeitung eines 2×2-Teilquadrates vom linken oberen im Uhrzeigersinn zum linken unteren Viertel durch die Richtung der Pfeile wieder. Es kann wie auch die übrigen 8 Pfeilformate durch einen scharfen Blick auf die $\mathcal{H}$-Matrix in Abbildung 3.7 gut eingesehen werden.

Bemerkung: Im 1D Fall liegen insgesamt nur 3 $\mathcal{H}$-Formate vor: das Diagonalformat $\square = \begin{array}{|c|c|} \hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array}$ und die zwei Nachbarformate $\rightarrow = \begin{array}{|c|c|} \hline R & R \\ \hline \rightarrow & R \\ \hline \end{array}$ und $\leftarrow = \begin{array}{|c|c|} \hline R & \leftarrow \\ \hline R & R \\ \hline \end{array}$ (siehe [19]). In Abbildung 3.8 ist eine zugehörige 1D $\mathcal{H}$-Matrix dargestellt.

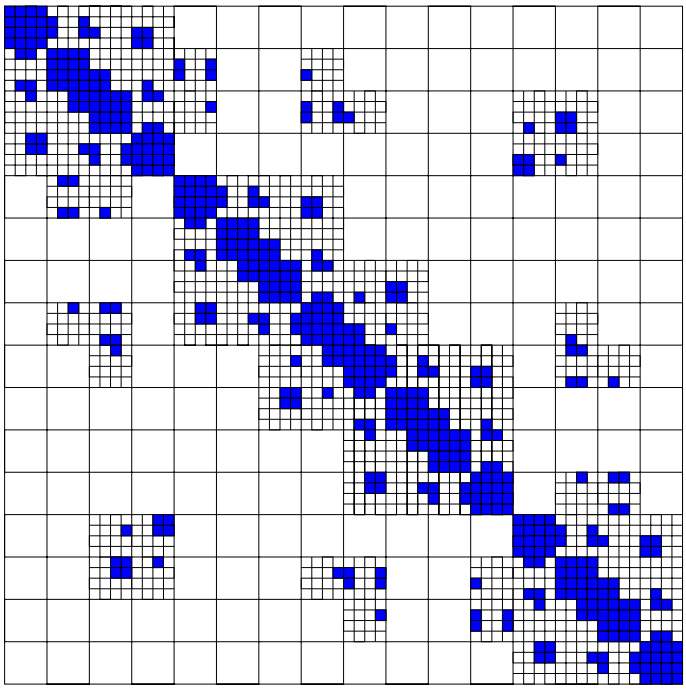

Abbildung 3.7: Beispiel einer hierarchischen FE-Matrix: Die weißen Blöcke dieser $\mathcal{H}$-Matrix sind Rk-Matrizen, die dunklen Blöcke ihrerseits hierarchische 4×4-Teilblöcke für $p_{eff} > 3$ bzw. vollbesetzte Teilmatrizen für $p_{eff} = 3$.

3.3.2 Komplexitätsbetrachtungen

Zur Darstellung bzw. Speicherung einer $N \times N$-$\mathcal{H}$-Matrix $\square$ werden $\mathcal{O}(N \log N)$ Elemente benötigt (siehe [20]). Wir wollen nun zusätzlich beweisen, dass die Anzahl der Blöcke von $\square$ gleich $\mathcal{O}(N)$ ist. Der Beweis hierfür verläuft analog zum Beweis für den Speicheraufwand von $\square$ in [20, Kapitel 4.4.1]:

Bezeichne $\mathcal{N}_{bl}^{\square}(p)$ die Anzahl aller Blöcke einer $\square$-Matrix der Tiefe p, $\mathcal{N}_{bl}^{+}(p)$ die Anzahl aller Blöcke einer $\leftarrow$-, $\rightarrow$-, $\uparrow$- oder $\downarrow$-Matrix sowie $\mathcal{N}_{bl}^{\times}(p)$ die Anzahl aller Blöcke einer $\nearrow$-, $\searrow$-, $\swarrow$- oder $\nwarrow$-Matrix der Tiefe p. Dann gilt

$$\left.\begin{aligned} &\mathcal{N}_{bl}^{\times}(p) = \mathcal{N}_{bl}^{\times}(p-1) + 15, \;\; p \geq 1 \\ &\mathcal{N}_{bl}^{\times}(0) = 1 \end{aligned}\right\} \implies \mathcal{N}_{bl}^{\times}(p) = 15\,p + 1,$$

$$\left.\begin{aligned} &\mathcal{N}_{bl}^{+}(p) = 2\,\mathcal{N}_{bl}^{+}(p-1) + 2\,\mathcal{N}_{bl}^{\times}(p-1) + 12, \;\; p \geq 1 \\ &\mathcal{N}_{bl}^{+}(0) = 1 \end{aligned}\right\}$$
$$\implies \mathcal{N}_{bl}^{+}(p) = 45 \cdot 2^p - 30\,p - 44 \;\text{ und}$$

$$\left.\begin{aligned} &\mathcal{N}_{bl}^{\square}(p) = 4\,\mathcal{N}_{bl}^{\square}(p-1) + 8\,\mathcal{N}_{bl}^{+}(p-1) + 4\,\mathcal{N}_{bl}^{\times}(p-1), \;\; p \geq 1 \\ &\mathcal{N}_{bl}^{\square}(0) = 1 \end{aligned}\right\}$$
$$\implies \mathcal{N}_{bl}^{\square}(p) = 45 \cdot 4^p - 180 \cdot 2^p + 60\,p + 136.$$

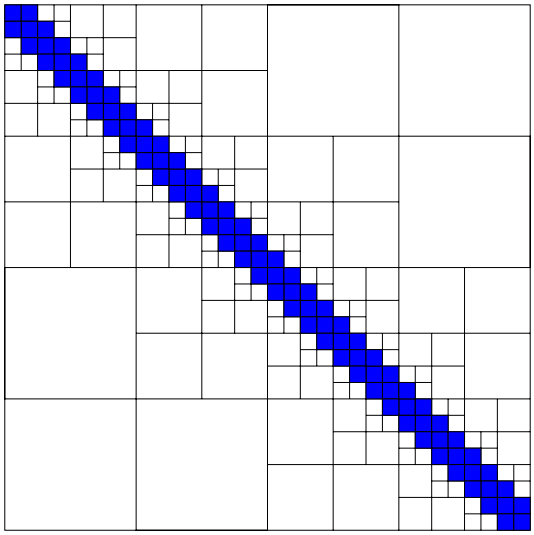

Abbildung 3.8: Beispiel einer 1D $\mathcal{H}$-Matrix: Die weißen Blöcke sind Rk-Matrizen, die dunklen Blöcke ihrerseits hierarchische 2×2-Teilblöcke für $p_{eff} > 5$ bzw. vollbesetzte Teilmatrizen für $p_{eff} = 5$.

Sei nun $\mathcal{N}_M^{\square}(p)$ die Anzahl aller vollbesetzten Matrixblöcke einer $\square$-Matrix der Tiefe p sowie $\mathcal{N}_M^{+}(p)$ und $\mathcal{N}_M^{\times}(p)$ analog dazu definiert. Dann gilt

$$\left.\begin{aligned} &\mathcal{N}_M^{\times}(p) = \mathcal{N}_M^{\times}(p-1), \;\; p \geq 1 \\ &\mathcal{N}_M^{\times}(0) = 1 \end{aligned}\right\} \;\Longrightarrow\; \mathcal{N}_M^{\times}(p) = 1 \;\; \forall\, p \geq 1,$$

$$\left.\begin{aligned} &\mathcal{N}_M^{+}(p) = 2\,\mathcal{N}_M^{+}(p-1) + 2\,\mathcal{N}_M^{\times}(p-1), \;\; p \geq 1 \\ &\mathcal{N}_M^{+}(0) = 1 \end{aligned}\right\}$$
$$\Longrightarrow\; \mathcal{N}_M^{+}(p) = 3 \cdot 2^p - 2 \;\text{ und}$$

$$\left.\begin{aligned} &\mathcal{N}_M^{\square}(p) = 4\,\mathcal{N}_M^{\square}(p-1) + 8\,\mathcal{N}_M^{+}(p-1) + 4\,\mathcal{N}_M^{\times}(p-1), \;\; p \geq 1 \\ &\mathcal{N}_M^{\square}(0) = 1 \end{aligned}\right\}$$
$$\Longrightarrow\; \mathcal{N}_M^{\square}(p) = 9 \cdot 4^p - 12 \cdot 2^p + 4.$$

Für die Anzahl $\mathcal{N}_{Rk}^{\square}(p)$ aller Rk-Blöcke einer $\square$-Matrix der Tiefe p gilt somit

$$\mathcal{N}_{Rk}^{\square}(p) = \mathcal{N}_{bl}^{\square}(p) - \mathcal{N}_M^{\square}(p) = 36 \cdot 4^p - 168 \cdot 2^p + 60\,p + 132.$$

Im Limes $p \to \infty$ gilt also

$$\frac{\mathcal{N}_{Rk}^{\square}(p)}{\mathcal{N}_M^{\square}(p)} \longrightarrow 4,$$

d.h. für große Tiefen p sind rund 20 % aller Matrixblöcke vollbesetzt und der Rest Rk-Blöcke. Man beachte jedoch, dass die vollbesetzten Matrixblöcke nur

auf dem höchsten Level auftreten, d.h. nur kleinste Matrixblöcke ausmachen, wohingegen die Rk-Blöcke auf allen Levels $l \geq 2$ vorkommen.

3.3.3 Adaptivität im Ort

Liegt nicht das regelmäßig triangulierte Einheitsquadrat sondern eine beliebige Verteilung von Knoten in $\Omega = (0,1)^2$ vor, so kann wie vorhin ein regelmäßiges quadratisches Gitter als Referenzgitter genommen und dessen Hierarchie für das unstrukturierte Dreiecksgitter „übernommen“ werden.
Eine andere Strategie wäre, nicht jeweils 4 gleich große Teilquadrate, sondern 4 gleich mächtige Teilquadrate (d.h. mit gleichviel enthaltenen Knoten) zu bilden. Mehr zur hierarchischen Partitionierung und Clusterung von Indexmengen findet sich in [20], [22] oder [17, Kap. 3.1, Seite 28].

3.4 Die approximierte $\mathcal{H}$-Arithmetik

Hackbusch hat in [20] die folgenden Operationen für 2D $\mathcal{H}$-Matrizen eingeführt:

- $\mathcal{H}$-Matrix-Vektor-Multiplikation
- $\mathcal{H}$-Matrix-Addition
- $\mathcal{H}$-Matrix-Multiplikation
- $\mathcal{H}$-Matrix-Invertierung

Diese 4 $\mathcal{H}$-Operationen werden wir nun detailliert ausarbeiten und für die approximierten unter ihnen parallel zum rekursiven Schema mitlaufende a posteriori Fehlerschätzungen konstruieren.

Im Anschluss daran werden wir neue hierarchische Zerlegungen erstellen:

- $\mathcal{H}$-Cholesky-Zerlegung symmetrisch positiv definiter $\mathcal{H}$-Matrizen
- $\mathcal{H}$-LDL^T-Zerlegung symmetrisch indefiniter $\mathcal{H}$-Matrizen
- $\mathcal{H}$-QR-Zerlegung invertierbarer $\mathcal{H}$-Matrizen
- $\mathcal{H}$-Polarzerlegung invertierbarer $\mathcal{H}$-Matrizen

Nach der systematischen Entwicklung linearer Gleichungssysteme in $\mathcal{H}$-Arithmetik von einer Rk-Matrix bis zu einer $\mathcal{H}$-Matrix als rechter Seite wird die Cholesky-Zerlegung hierarchisiert sowie Rückwärtsanalyse für den Approximationsfehler durchgeführt.
Die $\mathcal{H}$-LDL^T-Zerlegung verläuft völlig analog zur $\mathcal{H}$-Cholesky-Zerlegung. Sie wird noch eine wesentliche Rolle in unserem Algorithmus für das $\mathcal{H}$-Eigenwertproblem des diskreten Laplace-Operators spielen.
Die vorgestellte $\mathcal{H}$-QR-Zerlegung einer invertierbaren $\mathcal{H}$-Matrix A hängt stark von deren Kondition ab, da wir den Weg über die $\mathcal{H}$-Cholesky-Zerlegung von A^TA gehen. Im Anschluss an die Entwicklung des Verfahrens zeigen wir eine Möglichkeit der Reorthogonalisierung des orthogonalen Faktors auf. Eine

weitere Verbesserung lässt sich durch vorangestellte $\mathcal{H}$-Polarzerlegung und $\mathcal{H}$-Cholesky-Zerlegung erreichen: Letztere bewirkt nämlich die Reduktion der Kondition der Ausgangsmatrix für die $\mathcal{H}$-QR-Zerlegung!

Die hierarchische Struktur erlaubt eine blockweise rekursive Beschreibung der $\mathcal{H}$-Arithmetik. Bei allen nun folgenden $\mathcal{H}$-Operationen und $\mathcal{H}$-Zerlegungen wird es stets unser vorrangiges Ziel sein zu gewährleisten, dass das Ergebnis wiederum zur Klasse der $\mathcal{H}$-Matrizen gehört. Der approximative Charakter der $\mathcal{H}$-Operationen und $\mathcal{H}$-Zerlegungen basiert dabei ausschließlich auf „Rk-Abschneidungen“, in die eine oder mehrere vollbesetzte oder Rk-Matrizen involviert sind.

3.4.1 Matrix-Vektor-Multiplikation

- Wegen der rekursiven Struktur einer $\mathcal{H}$-Matrix

$$A = \square = \begin{array}{|c|c|c|c|} \hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline \end{array}$$

  wird die Matrix-Vektor-Multiplikation (wie auch die folgenden Matrix-Operationen) rekursiv realisiert: $A\mathbf{v}$ mit $A = (A_{ij})_{1\le i,j\le 4}$, $\mathbf{v} = (\mathbf{v}_j)_{1\le j\le 4}$ wird auf die 16 Matrix-Vektor-Multiplikationen $A_{ij}\mathbf{v}_j$, $1 \le i, j \le 4$ auf dem nächsthöheren Level zurückgeführt.

- Die Multiplikation $\mathbf{v}^T A = (A^T\mathbf{v})^T$ läßt sich auf die obige Matrix-Vektor-Multiplikation zurückführen.

- Die Multiplikation einer $\mathcal{H}$-Matrix A mit einer Rk-Matrix $R = [\mathbf{a}, \mathbf{b}]$ läßt sich auf die Matrix-Vektor-Multiplikation zurückführen: $AR = [A\mathbf{a}, \mathbf{b}]$ (analog für RA).

- Die Komplexität von $A\mathbf{v}$ bzw. AR ist gleich $\mathcal{O}(N \log N)$ mit $N = n^2$ der Anzahl der Unbekannten (siehe Tabelle 3.2).

- Die Operationen $A\mathbf{v}$ bzw. AR sind exakt: Es werden keine Rk-Abschneidungen durchgeführt!

3.4.2 Addition zweier $\mathcal{H}$-Matrizen

- Eine $\mathcal{H}$-Matrix besteht aus vollbesetzten und Rk-Blöcken. Die vollbesetzten Matrixblöcke werden exakt addiert, die Summe zweier Rk-Blöcke auf Rang k abgeschnitten. Damit entspricht die $\mathcal{H}$-Summe insgesamt wieder einer $\mathcal{H}$-Matrix von derselben Rangstruktur der beiden Ausgangsmatrizen.

- Die Komplexität der $\mathcal{H}$-Addition $+_{\mathcal{H}}$ ist gleich $\mathcal{O}(N \log N)$.

k \ N	16^2	32^2	64^2	128^2	256^2
1	0.50	3.28	16.9	78.6	350.7
2	0.50	3.44	18.6	90.9	424.3
3	0.50	3.59	20.3	103.2	497.9
4	0.50	3.74	22.0	115.5	571.5
5	0.50	3.89	23.7	127.8	645.1
6	0.50	4.05	25.4	140.1	718.7
7	0.50	4.20	27.1	152.4	792.3
8	0.50	4.35	28.8	164.7	865.9
9	0.50	4.50	30.5	177.0	939.5
10	0.50	4.65	32.2	189.3	1013.1

Tabelle 3.1: Speicherverbrauch (in MB) einer $N \times N$-$\mathcal{H}$-Matrix mit $\delta p = 3$ und Rang k.

k \ N	16^2	32^2	64^2	128^2	256^2
1	< 0.01	0.03	0.13	0.57	2.55
2	< 0.01	0.03	0.14	0.64	3.04
3	< 0.01	0.03	0.15	0.71	3.55
4	< 0.01	0.03	0.16	0.81	4.03
5	< 0.01	0.03	0.17	0.88	4.50
6	< 0.01	0.03	0.18	0.95	4.96
7	< 0.01	0.03	0.19	1.02	5.41
8	< 0.01	0.03	0.20	1.06	6.14
9	< 0.01	0.03	0.21	1.14	6.71
10	< 0.01	0.03	0.22	1.21	7.35

Tabelle 3.2: Benötigte Zeit (in Sekunden) zur Durchführung der Matrix-Vektor-Multiplikation mit zufälligen $N \times N$-$\mathcal{H}$-Matrizen mit $\delta p = 3$ und Rang k und zufälligen N-dimensionalen Vektoren (Prozessor: Sun Ultra Sparc IIi, 300 MHz).

- **Exakter Fehler der $\mathcal{H}$-Addition in der Frobeniusnorm:**
  Sei $A = (A_{ij})_{1 \le i,j \le 4} \in \mathcal{M}_{\mathcal{H}}$ gegeben.
  Dann gilt $\|A\|_F^2 = \sum_{i,j=1}^4 \|A_{ij}\|_F^2$. Für $A, B \in \mathcal{M}_{\mathcal{H}}$ folgt damit rekursiv

$$\begin{aligned} \|(A+B) - (A +_{\mathcal{H}} B)\|_F^2 &= \sum_{i,j=1}^4 \|(A+B)_{ij} - (A +_{\mathcal{H}} B)_{ij}\|_F^2 \\ &= \ldots = \sum_{\text{Rk-Bl.}b} \sum_{\hat{\sigma}_b} \hat{\sigma}_b^2, \end{aligned}$$

  wobei die zweite Summe über alle abgeschnittenen Singulärwerte $\hat{\sigma}_b$ bei der Rk-Addition $A^b +_{Rk} B^b$ läuft.
  Der exakte Fehler in der Frobeniusnorm ist also gleich der euklidischen Norm des Vektors aller abgeschnittenen Singulärwerte bei sämtlichen durchgeführten Rk-Additionen.

3.4.3 Multiplikation zweier $\mathcal{H}$-Matrizen

Die Matrix-Matrix-Multiplikation zweier $\mathcal{H}$-Matrizen A und B wird rekursiv realisiert durch Rückführung auf 64 Matrix-Matrix-Multiplikationen und geeignete Additionen von je 4 $\mathcal{H}$-Produkten auf dem nächsthöheren Level:

$$A \cdot B = \begin{array}{|c|c|c|c|}\hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline \end{array} \cdot \begin{array}{|c|c|c|c|}\hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline \end{array} =$$

$$\begin{array}{|c|c|c|c|}\hline \square\cdot\square+\rightarrow\cdot\leftarrow+ & \square\cdot\rightarrow+\rightarrow\cdot\square+ & \square\cdot\searrow+\rightarrow\cdot\downarrow+ & \square\cdot\downarrow+\rightarrow\cdot\swarrow+ \\ +\searrow\cdot\nwarrow+\downarrow\cdot\uparrow & +\searrow\cdot\uparrow+\downarrow\cdot\nearrow & +\searrow\cdot\square+\downarrow\cdot\rightarrow & +\searrow\cdot\leftarrow+\downarrow\cdot\square \\ \hline \leftarrow\cdot\square+\square\cdot\leftarrow+ & \leftarrow\cdot\rightarrow+\square\cdot\square+ & \leftarrow\cdot\searrow+\square\cdot\downarrow+ & \leftarrow\cdot\downarrow+\square\cdot\swarrow+ \\ +\downarrow\cdot\nwarrow+\swarrow\cdot\uparrow & +\downarrow\cdot\uparrow+\swarrow\cdot\nearrow & +\downarrow\cdot\square+\swarrow\cdot\rightarrow & +\downarrow\cdot\leftarrow+\swarrow\cdot\square \\ \hline \nwarrow\cdot\square+\uparrow\cdot\leftarrow+ & \nwarrow\cdot\rightarrow+\uparrow\cdot\square+ & \nwarrow\cdot\searrow+\uparrow\cdot\downarrow+ & \nwarrow\cdot\downarrow+\uparrow\cdot\swarrow+ \\ +\square\cdot\nwarrow+\leftarrow\cdot\uparrow & +\square\cdot\uparrow+\leftarrow\cdot\nearrow & +\square\cdot\square+\leftarrow\cdot\rightarrow & +\square\cdot\leftarrow+\leftarrow\cdot\square \\ \hline \uparrow\cdot\square+\nearrow\cdot\leftarrow+ & \uparrow\cdot\rightarrow+\nearrow\cdot\square+ & \uparrow\cdot\searrow+\nearrow\cdot\downarrow+ & \uparrow\cdot\downarrow+\nearrow\cdot\swarrow+ \\ +\rightarrow\cdot\nwarrow+\square\cdot\uparrow & +\rightarrow\cdot\uparrow+\square\cdot\nearrow & +\rightarrow\cdot\square+\square\cdot\rightarrow & +\rightarrow\cdot\leftarrow+\square\cdot\square \\ \hline \end{array}$$

Unser **Ziel** ist $A *_{\mathcal{H}} B \in \mathcal{M}_{\mathcal{H}}$, wobei $*_{\mathcal{H}}$ die im Folgenden entwickelte approximierte $\mathcal{H}$-Multiplikation bezeichnet.

3.4.3.1 Multiplikation eindimensionaler $\mathcal{H}$-Matrizen

Seien A und B eindimensionale $\mathcal{H}$-Matrizen der Form

$$\square = \begin{array}{|c|c|}\hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array} \in \mathcal{M}_{\mathcal{H}}$$

mit den rekursiven Nachbarformaten $\rightarrow = \begin{array}{|c|c|}\hline R & R \\ \hline \rightarrow & R \\ \hline \end{array}$ und $\leftarrow = \begin{array}{|c|c|}\hline R & \leftarrow \\ \hline R & R \\ \hline \end{array}$ (siehe Abschnitt 3.3.1).
Wann gilt nun $A *_{\mathcal{H}} B \in \mathcal{M}_{\mathcal{H}}$ für $A, B \in \mathcal{M}_{\mathcal{H}}$?

$$\begin{aligned} A \cdot B &= \begin{array}{|c|c|}\hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array} \cdot \begin{array}{|c|c|}\hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array} \\ &= \begin{array}{|c|c|}\hline \square\cdot\square+\rightarrow\cdot\leftarrow & \square\cdot\rightarrow+\rightarrow\cdot\square \\ \hline \leftarrow\cdot\square+\square\cdot\leftarrow & \leftarrow\cdot\rightarrow+\square\cdot\square \\ \hline \end{array} \overset{!}{=} \begin{array}{|c|c|}\hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array} \end{aligned}$$

Nun ist

$$\begin{aligned} \rightarrow\cdot\leftarrow &= \begin{array}{|c|c|}\hline R & R \\ \hline \rightarrow & R \\ \hline \end{array} \cdot \begin{array}{|c|c|}\hline R & \leftarrow \\ \hline R & R \\ \hline \end{array} \\ &= \begin{array}{|c|c|}\hline R\cdot R+R\cdot R & R\cdot\leftarrow+R\cdot R \\ \hline \rightarrow\cdot R+R\cdot R & \rightarrow\cdot\leftarrow+R\cdot R \\ \hline \end{array} = \begin{array}{|c|c|}\hline R & R \\ \hline R & \rightarrow\cdot\leftarrow \\ \hline \end{array} \end{aligned}$$

sowie analog $\leftarrow\cdot\rightarrow = \begin{array}{|c|c|}\hline \leftarrow\cdot\rightarrow & R \\ \hline R & R \\ \hline \end{array}$.

Bemerkung: $\rightarrow \cdot \leftarrow$ sowie $\leftarrow \cdot \rightarrow$ sind zwei bei der $\mathcal{H}$-Multiplikation neu entstandene $\mathcal{H}$-Formate!

Weiter ist

$$\begin{aligned}
\square \cdot \rightarrow &= \begin{array}{|c|c|}\hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array} \cdot \begin{array}{|c|c|}\hline R & R \\ \hline \rightarrow & R \\ \hline \end{array} \\
&= \begin{array}{|c|c|}\hline \square \cdot R + \rightarrow \cdot \rightarrow & \square \cdot R + \rightarrow \cdot R \\ \hline \leftarrow \cdot R + \square \cdot \rightarrow & \leftarrow \cdot R + \square \cdot R \\ \hline \end{array} \stackrel{!}{=} \begin{array}{|c|c|}\hline R & R \\ \hline \rightarrow & R \\ \hline \end{array} = \rightarrow \quad \Longleftrightarrow \\
\rightarrow \cdot \rightarrow &= \begin{array}{|c|c|}\hline R & R \\ \hline \rightarrow & R \\ \hline \end{array} \cdot \begin{array}{|c|c|}\hline R & R \\ \hline \rightarrow & R \\ \hline \end{array} = \begin{array}{|c|c|}\hline R & R \\ \hline R & R \\ \hline \end{array} \stackrel{!}{=} R
\end{aligned}$$

Also ist $\square \cdot \rightarrow = \rightarrow$ genau dann, wenn jeder 2×2-Rk-Block $\begin{array}{|c|c|}\hline R & R \\ \hline R & R \\ \hline \end{array}^b$ zu einer Rk'-Matrix mit $k' = k(b)$ abgeschnitten wird.
Ebenso ist $\rightarrow \cdot \square = \rightarrow$, $\square \cdot \leftarrow = \leftarrow$ sowie $\leftarrow \cdot \square = \leftarrow$ und somit insgesamt $A *_{\mathcal{H}} B \in \mathcal{M}_{\mathcal{H}}$ genau dann, wenn $\begin{array}{|c|c|}\hline R & R \\ \hline R & R \\ \hline \end{array}$ auf Rang k' abgeschnitten wird.

• Abschneidung einer 2×2-Blockmatrix $\begin{array}{|c|c|}\hline R & R \\ \hline R & R \\ \hline \end{array}$ zu einer Rk'-Matrix gleicher Dimension:

Fasse

$$\begin{array}{|c|c|}\hline [\mathbf{a}_{11}, \mathbf{b}_{11}] & [\mathbf{a}_{12}, \mathbf{b}_{12}] \\ \hline [\mathbf{a}_{21}, \mathbf{b}_{21}] & [\mathbf{a}_{22}, \mathbf{b}_{22}] \\ \hline \end{array} = \sum_{1 \le i,j \le 2} [\tilde{\mathbf{a}}_{ij}, \tilde{\mathbf{b}}_{ij}]$$

als Summe der 4 Rk-Matrizen $[\tilde{\mathbf{a}}_{ij}, \tilde{\mathbf{b}}_{ij}]$, $1 \le i, j \le 2$, mit $\tilde{\mathbf{a}}_{1j} = \begin{pmatrix} \mathbf{a}_{1j} \\ 0 \end{pmatrix}, \tilde{\mathbf{a}}_{2j} = \begin{pmatrix} 0 \\ \mathbf{a}_{2j} \end{pmatrix}$ $(j = 1, 2)$ und $\tilde{\mathbf{b}}_{i1} = \begin{pmatrix} \mathbf{b}_{i1} \\ 0 \end{pmatrix}, \tilde{\mathbf{b}}_{i2} = \begin{pmatrix} 0 \\ \mathbf{b}_{i2} \end{pmatrix}$ $(i = 1, 2)$ auf und schneide diese Summe analog zum Vorgehen in Abschnitt 3.2 auf Rang k' ab. Diese Abschneidung benötigt die Lösung eines $4k$-dimensionalen Eigenwertproblems bzw. eine $4k$-dimensionale Singulärwertzerlegung.

3.4.3.2 Zurück zu den 2D $\mathcal{H}$-Matrizen

Analog zum 1D Fall lässt sich zeigen:

$$A *_{\mathcal{H}} B \in \mathcal{M}_{\mathcal{H}} \quad \Longleftrightarrow \quad \begin{array}{|c|c|c|c|}\hline R & R & R & R \\ \hline R & R & R & R \\ \hline R & R & R & R \\ \hline R & R & R & R \\ \hline \end{array} \text{ wird auf Rang } k' \text{ abgeschnitten.}$$

Insgesamt entstehen im 2D Fall bei der Matrix-Multiplikation 16 neue hierarchische Formate:

$H11$-Format: $\nwarrow \cdot \searrow =$

R	R	$\nwarrow$	R
R	R	R	R
R	R	R	R
R	R	R	R

$\cdot$

R	R	R	R
R	R	R	R
$\searrow$	R	R	R
R	R	R	R

$=$

$\nwarrow \cdot \searrow$	R	R	R
R	R	R	R
R	R	R	R
R	R	R	R

$H22$-Format: $\nearrow \cdot \swarrow =$

R	R	R	R
R	$\nearrow \cdot \swarrow$	R	R
R	R	R	R
R	R	R	R

$H33$-Format: $\searrow \cdot \nwarrow =$

R	R	R	R
R	R	R	R
R	R	$\searrow \cdot \nwarrow$	R
R	R	R	R

$H44$-Format: $\swarrow \cdot \nearrow =$

R	R	R	R
R	R	R	R
R	R	R	R
R	R	R	$\swarrow \cdot \nearrow$

$H12$-Format: $\uparrow \cdot \swarrow =$

R	$\uparrow \cdot \swarrow$	R	R
R	R	R	R
R	R	R	R
R	R	R	R

und $\nwarrow \cdot \downarrow =$

R	$\nwarrow \cdot \downarrow$	R	R
R	R	R	R
R	R	R	R
R	R	R	R

$H21$-Format: $\uparrow \cdot \searrow =$

R	R	R	R
$\uparrow \cdot \searrow$	R	R	R
R	R	R	R
R	R	R	R

und $\nearrow \cdot \downarrow =$

R	R	R	R
$\nearrow \cdot \downarrow$	R	R	R
R	R	R	R
R	R	R	R

$H14$-Format: $\leftarrow \cdot \nearrow =$

R	R	R	$\leftarrow \cdot \nearrow$
R	R	R	R
R	R	R	R
R	R	R	R

und $\nwarrow \cdot \rightarrow =$

R	R	R	$\nwarrow \cdot \rightarrow$
R	R	R	R
R	R	R	R
R	R	R	R

$H41$-Format: $\leftarrow \cdot \searrow =$

R	R	R	R
R	R	R	R
R	R	R	R
$\leftarrow \cdot \searrow$	R	R	R

und $\swarrow \cdot \rightarrow =$

R	R	R	R
R	R	R	R
R	R	R	R
$\swarrow \cdot \rightarrow$	R	R	R

$H23$-Format: $\rightarrow \cdot \nwarrow =$

R	R	R	R
R	R	$\rightarrow \cdot \nwarrow$	R
R	R	R	R
R	R	R	R

und $\nearrow \cdot \leftarrow =$

R	R	R	R
R	R	$\nearrow \cdot \leftarrow$	R
R	R	R	R
R	R	R	R

$H32$-Format: $\rightarrow \cdot \swarrow =$

R	R	R	R
R	R	R	R
R	$\rightarrow \cdot \swarrow$	R	R
R	R	R	R

und $\searrow \cdot \leftarrow =$

R	R	R	R
R	R	R	R
R	$\searrow \cdot \leftarrow$	R	R
R	R	R	R

$H34$-Format: $\downarrow \cdot \nearrow =$

R	R	R	R
R	R	R	R
R	R	R	$\downarrow \cdot \nearrow$
R	R	R	R

und $\searrow \cdot \uparrow =$

R	R	R	R
R	R	R	R
R	R	R	$\searrow \cdot \uparrow$
R	R	R	R

$H43$-Format: $\swarrow \cdot \uparrow =$

R	R	R	R
R	R	R	R
R	R	R	R
R	R	$\swarrow \cdot \uparrow$	R

und $\downarrow \cdot \nwarrow =$

R	R	R	R
R	R	R	R
R	R	R	R
R	R	$\downarrow \cdot \nwarrow$	R

Hmi-Format: $\rightarrow \cdot \leftarrow =$

R	R	R	R
$\rightarrow$	R	R	$\searrow$
$\nearrow$	R	R	$\rightarrow$
R	R	R	R

$\cdot$

R	$\leftarrow$	$\swarrow$	R
R	R	R	R
R	R	R	R
R	$\nwarrow$	$\leftarrow$	R

$=$

R	R	R	R
R	Hmi	$H32$	R
R	$H23$	Hmi	R
R	R	R	R

($mi \mathrel{\widehat{=}}$ Mitte)

$Heck$-Format: $\leftarrow \cdot \rightarrow =$

R	$\leftarrow$	$\swarrow$	R
R	R	R	R
R	R	R	R
R	$\nwarrow$	$\leftarrow$	R

$\cdot$

R	R	R	R
$\rightarrow$	R	R	$\searrow$
$\nearrow$	R	R	$\rightarrow$
R	R	R	R

$=$

$Heck$	R	R	$H41$
R	R	R	R
R	R	R	R
$H14$	R	R	$Heck$

($eck \mathrel{\widehat{=}}$ Ecken)

Hlo-Format: $\uparrow \cdot \downarrow =$

R	R	$\nearrow$	$\uparrow$
R	R	$\uparrow$	$\nwarrow$
R	R	R	R
R	R	R	R

$\cdot$

R	R	R	R
R	R	R	R
$\swarrow$	$\downarrow$	R	R
$\downarrow$	$\searrow$	R	R

$=$

Hlo	$H21$	R	R
$H12$	Hlo	R	R
R	R	R	R
R	R	R	R

($lo \mathrel{\widehat{=}}$ links oben)

Hru-Format: $\downarrow \cdot \uparrow =$

R	R	R	R
R	R	R	R
$\swarrow$	$\downarrow$	R	R
$\downarrow$	$\searrow$	R	R

$\cdot$

R	R	$\nearrow$	$\uparrow$
R	R	$\uparrow$	$\nwarrow$
R	R	R	R
R	R	R	R

$=$

R	R	R	R
R	R	R	R
R	R	Hru	$H43$
R	R	$H34$	Hru

($ru \mathrel{\widehat{=}}$ rechts unten)

Bemerkung: Die 16 neu entstandenen hierarchischen Formate gehen nur in die Additionen von je 4 Produkten ein. Die Faktoren aller $\mathcal{H}$-Multiplikationen auf jedem Level bestehen nur aus den 9 ursprünglichen Formaten.

• **Abschneidung einer** 4×4**-Blockmatrix** $\begin{array}{|c|c|c|c|}\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline\end{array}$ **zu einer Rk'-Matrix gleicher Dimension:**

Fasse

$$\begin{array}{|c|c|c|c|}\hline [\mathbf{a}_{11},\mathbf{b}_{11}] & \cdots & \cdots & [\mathbf{a}_{14},\mathbf{b}_{14}]\\\hline \vdots & & & \vdots\\\hline \vdots & & & \vdots\\\hline [\mathbf{a}_{41},\mathbf{b}_{41}] & \cdots & \cdots & [\mathbf{a}_{44},\mathbf{b}_{44}]\\\hline\end{array} = \sum_{1\le i,j\le 4} [\tilde{\mathbf{a}}_{ij},\tilde{\mathbf{b}}_{ij}]$$

als Summe der 16 Rk-Matrizen $[\tilde{\mathbf{a}}_{ij},\tilde{\mathbf{b}}_{ij}]$, $1 \le i,j \le 4$, mit $\tilde{\mathbf{a}}_{1j} = \begin{pmatrix}\mathbf{a}_{1j}\\0\\0\\0\end{pmatrix}, \tilde{\mathbf{a}}_{2j} = \begin{pmatrix}0\\\mathbf{a}_{2j}\\0\\0\end{pmatrix}, \tilde{\mathbf{a}}_{3j} = \begin{pmatrix}0\\0\\\mathbf{a}_{3j}\\0\end{pmatrix}, \tilde{\mathbf{a}}_{4j} = \begin{pmatrix}0\\0\\0\\\mathbf{a}_{4j}\end{pmatrix}$ $(j = 1,\ldots,4)$ und $\tilde{\mathbf{b}}_{i1} = \begin{pmatrix}\mathbf{b}_{i1}\\0\\0\\0\end{pmatrix}, \tilde{\mathbf{b}}_{i2} = \begin{pmatrix}0\\\mathbf{b}_{i2}\\0\\0\end{pmatrix}, \tilde{\mathbf{b}}_{i3} = \begin{pmatrix}0\\0\\\mathbf{b}_{i3}\\0\end{pmatrix}, \tilde{\mathbf{b}}_{i4} = \begin{pmatrix}0\\0\\0\\\mathbf{b}_{i4}\end{pmatrix}$ $(i = 1,\ldots,4)$ auf und schneide diese Summe analog zum Vorgehen in Abschnitt 3.2 auf Rang k' ab. Diese Abschneidung benötigt die Lösung eines $16k$-dimensionalen Eigenwertproblems bzw. eine $16k$-dimensionale Singulärwertzerlegung.

Die $\mathcal{H}$**-Additionen von je** 4 $\mathcal{H}$**-Produkten** funktionieren alle nach demselben Schema:
Die $\mathcal{H}$-Formate der 4 Summanden sind jeweils in einem der 4 Formate enthalten. Durch kanonische Erweiterung zum komplexesten dieser 4 $\mathcal{H}$-Formate lassen sich somit all diese $\mathcal{H}$-Additionen auf Additionen ausschließlich vollbesetzter, vollbesetzter und Rk-Matrizen und ausschließlich Rk-Matrizen zurückführen.
Eine Rk-Matrix ist auf jedes andere $\mathcal{H}$-Format erweiterbar: Die $\mathcal{H}$-Summe

$$\begin{array}{|c|c|c|c|}\hline \searrow&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline\end{array} + \begin{array}{|c|c|c|c|}\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline\end{array} + \begin{array}{|c|c|c|c|}\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline\end{array} + \begin{array}{|c|c|c|c|}\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline R&R&R&R\\\hline\end{array}$$

ist beispielsweise rekursiv aufgebaut aus derselben $\mathcal{H}$-Summe und 15 Rk-Additionen auf dem nächsthöheren Level.

Nach diesem einfachstmöglichen Beispiel einer $\mathcal{H}$-Summe mit mindestens einem

echten Blocksummanden stellen wir noch eine der 4 komplexesten $\mathcal{H}$-Additionen dar:

$$\square + Hmi + Hru + H33 \quad = \quad \begin{array}{|c|c|c|c|}\hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline \end{array} + \begin{array}{|c|c|c|c|}\hline R & R & R & R \\ \hline R & Hmi & H32 & R \\ \hline R & H23 & Hmi & R \\ \hline R & R & R & R \\ \hline \end{array} + \begin{array}{|c|c|c|c|}\hline R & R & R & R \\ \hline R & R & R & R \\ \hline R & R & Hru & H43 \\ \hline R & R & H34 & Hru \\ \hline \end{array} + \begin{array}{|c|c|c|c|}\hline R & R & R & R \\ \hline R & R & R & R \\ \hline R & R & H33 & R \\ \hline R & R & R & R \\ \hline \end{array}$$

Diese $\mathcal{H}$-Addition ist rekursiv aus 16 unterschiedlichen $\mathcal{H}$-Additionen auf dem nächsthöheren Level aufgebaut. Wird nun in den 4×4-Blöcken rekursiv vorgegangen und werden die vollbesetzten und Rk-Matrizen in den Blättern des zur hierarchischen Matrixstruktur gehörigen $\mathcal{H}$-Baums erreicht, wird schließlich addiert.

• Die Komplexität der $\mathcal{H}$-Multiplikation $*_{\mathcal{H}}$ ist gleich $\mathcal{O}(N \log^2 N)$ (siehe Tabelle 3.3).

Bemerkung zum Rechenaufwand in Abhängigkeit von δp:
Für größeres δp sind die Matrix-Matrix-Multiplikationen zweier vollbesetzter Blöcke der Dimension d mit $4^{\delta p} \leq d \leq 4^{\delta p+1}$ und die Singulärwertzerlegung ebensolcher Blöcke am zeitintensivsten.
Für kleiner werdendes δp nimmt für konstante Dimension N der $\mathcal{H}$-Matrix die hierarchische Tiefe $p_{eff} = p - \delta p$ zu: Es dominiert dann der Aufwand in den Rk-Additionen.

Steuerung des Rangs k der Rk-Blöcke einer $\mathcal{H}$-Matrix:

$$k = k(b) = \alpha\,(p_{eff} - \text{level}(b)) + \beta,\ \alpha \geq 0,\ \beta \geq 1$$

Die größeren Rk-Matrizen (Blöcke zu einem niedrigeren Level) können größeren Rang als die kleineren Rk-Matrizen (Blöcke zu einem höheren Level) haben.

3.4.3.3 A posteriori Fehlerschätzung von $*_{\mathcal{H}}$ in der Frobeniusnorm

Wir werden nun die Entwicklung des Approximationsfehlers untersuchen und daraus eine a posteriori Fehlerschätzung von $*_{\mathcal{H}}$ in der Frobeniusnorm ableiten.

Gegeben seien zwei $\mathcal{H}$-Matrizen $A = (A_{ij})_{1 \leq i,j \leq 4}$ und $B = (B_{ij})_{1 \leq i,j \leq 4}$.
Gesucht ist der absolute bzw. relative Fehler $\|A *_{\mathcal{H}} B - A \cdot B\|_F$ bzw. $\frac{\|A *_{\mathcal{H}} B - A \cdot B\|_F}{\|A \cdot B\|_F}$ der $\mathcal{H}$-Multiplikation in der Frobeniusnorm.

k \ N	16^2	32^2	64^2	128^2
1	0.13	6.62	53.58	301.7
2	0.13	7.07	62.03	379.5
3	0.13	7.89	75.27	492.4
4	0.13	8.71	96.23	651.3
5	0.13	9.74	130.74	948.5
6	0.13	11.15	171.20	1265.0
7	0.13	13.22	238.70	1823.1
8	0.13	14.71	299.90	2274.7
9	0.13	17.69	397.53	3050.4
10	0.13	20.69	497.82	3833.9

Tabelle 3.3: Benötigte Zeit (in Sekunden) zur Durchführung der Matrix-Matrix-Multiplikation mit zufälligen $N \times N$-$\mathcal{H}$-Matrizen mit $\delta p = 3$ und Rang k inklusive mitlaufender a posteriori Fehlerschätzung (Prozessor: Sun Ultra Sparc IIi, 300 MHz).

Mit

$$A *_{\mathcal{H}} B = \begin{array}{|c|c|c|c|} \hline \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{1k} *_{\mathcal{H}} B_{k1} & \cdots & \cdots & \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{1k} *_{\mathcal{H}} B_{k4} \\ \hline \vdots & & & \vdots \\ \hline \vdots & & & \vdots \\ \hline \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{4k} *_{\mathcal{H}} B_{k1} & \cdots & \cdots & \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{4k} *_{\mathcal{H}} B_{k4} \\ \hline \end{array}$$

gilt

$$\begin{aligned} \|A *_{\mathcal{H}} B - A \cdot B\|_F^2 &= \sum_{i,j=1}^{4} \| \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{ik} *_{\mathcal{H}} B_{kj} - \sum_{k=1}^{4} A_{ik} \cdot B_{kj} \|_F^2 \\ &= \sum_{i,j=1}^{4} \| \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{ik} *_{\mathcal{H}} B_{kj} - \sum_{k=1}^{4} A_{ik} *_{\mathcal{H}} B_{kj} + \\ &\qquad \sum_{k=1}^{4} A_{ik} *_{\mathcal{H}} B_{kj} - \sum_{k=1}^{4} A_{ik} \cdot B_{kj} \|_F^2 \\ &\leq \sum_{i,j=1}^{4} \Bigg(\| \sum_{k=1}^{4}{}_{\mathcal{H}}\, A_{ik} *_{\mathcal{H}} B_{kj} - \sum_{k=1}^{4} A_{ik} *_{\mathcal{H}} B_{kj} \|_F + \\ &\qquad \sum_{k=1}^{4} \| A_{ik} *_{\mathcal{H}} B_{kj} - A_{ik} \cdot B_{kj} \|_F \Bigg)^2 , \end{aligned}$$

wobei $\sum_{\mathcal{H}}$ jeweils die approximierten $\mathcal{H}$-Summen bezeichnet.

Der $\mathcal{H}$-Multiplikationsfehler auf dem Level 0 wurde somit nach oben abgeschätzt durch $\mathcal{H}$-Additionsfehler auf dem nächsthöheren Level 1 und weitere $\mathcal{H}$-Multiplikationsfehler auf dem Level 1. Durch rekursive Fortführung dieser Abschätzung kann schließlich der $\mathcal{H}$-Multiplikationsfehler aus den $\mathcal{H}$-Additionsfehlern auf sämtlichen Levels 1 bis p_{eff} zusammengesetzt werden.

Dies muss auch so sein: Richtig gerechnet wird ja nur mit vollbesetzten und Rk-Matrizen, mit anderen Worten in den Blättern der $\mathcal{H}$-Bäume, und die Multiplikationen solcher Matrizen sind ja exakt. Fehler werden nur in all den Rk-Additionen bzw. nach der Singulärwertzerlegung vollbesetzter Matrizen beim Abschneiden der kleinsten Singulärwerte gemacht.
Die $\mathcal{H}$-Additionsfehler einer $\mathcal{H}$-Summe von 4 $\mathcal{H}$-Produkten ergeben sich wie vorhin für die $\mathcal{H}$-Addition zweier $\mathcal{H}$-Matrizen in Abschnitt 3.4.2 als die euklidische Norm des Vektors, der aus allen Singulärwerten besteht, die bei sämtlichen Rk-Additionen im Laufe der $\mathcal{H}$-Addition abgeschnitten wurden.

Eigenschaften obiger Fehlerschätzung für die $\mathcal{H}$-Multiplikation:

- Die Fehlerschätzung benötigt im Wesentlichen keine zusätzlichen Operationen: Die abgeschnittenen Singulärwerte bei den einzelnen Rk-Additionen wurden alle schon berechnet.
- Sie erfolgt vollkommen parallel zum rekursiven Schema der einzelnen $\mathcal{H}$-Operationen.
- Der Speicheraufwand für die Fehlerschätzung beläuft sich auf lediglich $2\,p_{eff}$ Elementen.
- Die Faktoren A_{ik} und B_{kj} einer jeden Multiplikation sind als Teilmatrizen von A und B exakt!

3.4.4 Invertierung einer $\mathcal{H}$-Matrix

Wir beschreiben nun im Folgenden die approximierte $\mathcal{H}$-Invertierung einer $\mathcal{H}$-Matrix A als 4×4-Blockmatrix aufgefasst mittels des Block-Gauß-Algorithmus, wobei stets die approximierten $\mathcal{H}$-Additionen und $\mathcal{H}$-Multiplikationen verwendet werden.

3.4.4.1 Das rekursive Schema der $\mathcal{H}$-Invertierung

Wie die übrigen $\mathcal{H}$-Operationen besitzt auch die $\mathcal{H}$-Invertierung eine rekursive Struktur: Die $\mathcal{H}$-Invertierung auf dem Level 0 wird zurückgeführt auf 4 $\mathcal{H}$-Invertierungen auf dem nächsthöheren Level 1 sowie $\mathcal{H}$-Multiplikationen und $\mathcal{H}$-Additionen auf dem Level 1.

Im Unterschied zur $\mathcal{H}$-Multiplikation treten bei der $\mathcal{H}$-Invertierung auch hierarchische Summen sowie Rk-Additionen mit 2 oder 3 Summanden auf! Diese Summen werden völlig analog zu den Summen mit 4 Summanden bei der $\mathcal{H}$-Multiplikation gebildet.

• Die Komplexität der $\mathcal{H}$-Matrix-Invertierung $(\ )^{-1_{\mathcal{H}}}$ ist gleich $\mathcal{O}(N\log^2 N)$ (siehe Tabelle 3.4).

k \ N	16^2	32^2	64^2	128^2
1	0.13	4.77	39.2	230.0
2	0.13	4.92	44.2	278.3
3	0.13	5.31	51.8	347.0
4	0.13	5.70	63.9	454.7
5	0.13	6.50	85.1	640.0
6	0.13	6.99	104.0	799.3
7	0.13	8.04	137.3	1088.1
8	0.13	8.22	158.1	1237.5
9	0.13	9.58	193.5	1554.8
10	0.13	10.50	232.0	1882.9

Tabelle 3.4: Benötigte Zeit (in Sekunden) zur Durchführung der $\mathcal{H}$-Invertierung der $N \times N$-$\mathcal{H}$-Steifigkeitsmatrix A_h mit $\delta p = 3$ und Rang k inklusive mitlaufender a posteriori Fehlerschätzung (Prozessor: Sun Ultra Sparc IIi, 300 MHz).

3.4.4.2 A posteriori Fehlerschätzung von $(\ )^{-1_{\mathcal{H}}}$ in der Frobeniusnorm

Gegeben sei die invertierbare $\mathcal{H}$-Matrix $A = (A_{ij})_{1 \le i,j \le 4}$.
Gesucht ist der absolute bzw. relative Fehler $\|A^{-1_{\mathcal{H}}} - A^{-1}\|_F$ bzw. $\frac{\|A^{-1_{\mathcal{H}}} - A^{-1}\|_F}{\|A^{-1}\|_F}$.
Wir verfolgen nun dieselbe Strategie wie bei der $\mathcal{H}$-Multiplikation: Wir entwickeln den $\mathcal{H}$-Fehler, der durch das wiederholte Abschneiden von Singulärwerten entsteht, parallel zur rekursiven Berechnung von $A^{-1_{\mathcal{H}}}$.

Das Fehlerquadrat lässt sich als Summe der Fehlerquadrate über alle Teilblöcke schreiben:

$$\|A^{-1_{\mathcal{H}}} - A^{-1}\|_F^2 = \sum_{i,j=1}^{4} \| \left(A^{-1_{\mathcal{H}}}\right)_{ij} - \left(A^{-1}\right)_{ij} \|_F^2$$

Dabei ist $\| \left(A^{-1_{\mathcal{H}}}\right)_{44} - \left(A^{-1}\right)_{44} \|_F$ selbst von der Gestalt $\|B^{-1_{\mathcal{H}}} - B^{-1}\|_F$, wobei B eine fehlerhafte $\mathcal{H}$-Matrix auf dem nächsthöheren Level ist. Die restlichen 15 Blöcke sind entweder $\mathcal{H}$-Summen oder $\mathcal{H}$-Produkte fehlerhafter $\mathcal{H}$-Matrizen. Aufgrund der Fehlerhaftigkeit der Eingabedaten der soeben erwähnten $\mathcal{H}$-Summen, $\mathcal{H}$-Produkte und $\mathcal{H}$-Inversen in den einzelnen Teilblöcken ist eine Konditionsanalyse durchzuführen. Insgesamt setzt sich der $\mathcal{H}$-Fehler der $\mathcal{H}$-Inversen auf dem Level 0 also aus Additionsfehlern und transportierten Fehlern höherer Levels zusammen.

Sensitivitätsanalyse für die $\mathcal{H}$-Operationen:

1) Die $\mathcal{H}$-Addition:
Gegeben seien fehlerhafte Summanden $\tilde{A}_i \in \mathcal{M}_{\mathcal{H}}$ samt den Fehlerschranken $\|\tilde{A}_i - A_i\|_F$.
Dann gilt

$$\| \sum\nolimits_{\mathcal{H}} \tilde{A}_i - \sum A_i \|_F \le \| \sum\nolimits_{\mathcal{H}} \tilde{A}_i - \sum \tilde{A}_i \|_F + \sum \|\tilde{A}_i - A_i\|_F.$$

Die Eingabefehler werden also unverändert an die Ausgabe weitergereicht und addieren sich zum Abschneidefehler aller durchgeführten Rk-Additionen.

2) Die $\mathcal{H}$-Multiplikation:
Gegeben seien zwei fehlerhafte Faktoren $\tilde{A},\ \tilde{B} \in \mathcal{M}_{\mathcal{H}}$ samt Fehlerschranken $\|\tilde{A} - A\|_F$ und $\|\tilde{B} - B\|_F$.
Dann gilt

$$\|\tilde{A} *_{\mathcal{H}} \tilde{B} - A \cdot B\|_F \dot{\leq} \|\tilde{A} *_{\mathcal{H}} \tilde{B} - \tilde{A} \cdot \tilde{B}\|_F + \|\tilde{A}\|_F \|\tilde{B} - B\|_F + \|\tilde{B}\|_F \|\tilde{A} - A\|_F.$$

Der Gesamtfehler lässt sich also nach oben abschätzen durch den $\mathcal{H}$-Fehler der $\mathcal{H}$-Matrix-Multiplikation bei exaktem Input und den transportierten Eingabefehlern.

3) Die $\mathcal{H}$-Invertierung:
Gegeben sei eine fehlerhafte $\mathcal{H}$-Matrix $\tilde{A}$ samt Fehlerschranke $\|\tilde{A} - A\|_F$.
Dann gilt

$$\|\tilde{A}^{-1_{\mathcal{H}}} - A^{-1}\|_F \dot{\leq} \|\tilde{A}^{-1_{\mathcal{H}}} - \tilde{A}^{-1}\|_F + \|\tilde{A}^{-1_{\mathcal{H}}}\|_F^2 \|\tilde{A} - A\|_F.$$

Dabei wurde die exakte absolute Konditionszahl für die Matrixinvertierung $\|\tilde{A}^{-1}\|_F^2$ durch $\|\tilde{A}^{-1_{\mathcal{H}}}\|_F^2$ zur Beschreibung des Fehlertransportes genähert.

Beobachtung in der Praxis:
Mit den Verstärkungsfaktoren $\|\tilde{A}\|_F$, $\|\tilde{B}\|_F$ und $\|\tilde{A}^{-1_{\mathcal{H}}}\|_F^2$ in der Frobeniusnorm explodiert der Fehlerschätzer für die $\mathcal{H}$-Invertierung. Das liegt einfach daran, dass die Frobeniusnorm zur Beschreibung der Fehlerverstärkungsfaktoren völlig ungeeignet ist. Man beachte nur $\|I\|_F = \sqrt{N}$ als Faktor für die Multiplikation mit der N-dimensionalen Einheitsmatrix I.
Verwendet man stattdessen die Fehlerverstärkungsfaktoren $\|\tilde{A}\|_2$, $\|\tilde{B}\|_2$ und $\|\tilde{A}^{-1_{\mathcal{H}}}\|_2^2$ in der Spektralnorm, so lassen sich zum Teil sehr gute Fehlerschätzungen in der Frobeniusnorm erzielen. Nachdem sich die Spektralnorm einer Matrix $A \in \mathbb{R}^{m \times n}$ jedoch nicht rekursiv aus ihren Teilblöcken gewinnen lässt, gehen wir einfach zu deren unteren Schranke $\frac{\|A\|_F}{\sqrt{l}} \leq \|A\|_2$ über, wobei $l = \min\{m, n\}$ ist.
Zur Berechnung der Frobeniusnorm eines hierarchischen Matrixblocks ist nur zu erwähnen, dass sich deren Quadrat rekursiv durch Summation der Quadrate der Frobeniusnormen der einzelnen Teilblöcke ergibt.

Die in den Abbildungen 3.9 und 3.10 dargestellten numerischen Ergebnisse der a posteriori Fehlerschätzer für die $\mathcal{H}$-Invertierung der Steifigkeitsmatrix A_h sowie der Massenmatrix M_h wurden mit diesen „gedämpften“ Fehlerverstärkungsfaktoren erzielt.

3.4.4.3 Rückwärtsfehleranalyse für die $\mathcal{H}$-Invertierung

Eine andere Strategie zur Schätzung des $\mathcal{H}$-Fehlers der $\mathcal{H}$-Inversen besteht in der Analyse des Rückwärtsfehlers:
Nach Wilkinson (1954) gilt für den normweisen a posteriori Rückwärtsfehler für $AX = I$ mit der Näherung $\hat{X} = A^{-1_{\mathcal{H}}}$:

Das kleinste η, für das $(A + \Delta A)\hat{X} = I$ mit $\|\Delta A\| \leq \eta \|A\|$ gilt, ist gegeben durch

$$\eta = \frac{\|I - AA^{-1\mathcal{H}}\|}{\|A\| \, \|A^{-1\mathcal{H}}\|} \tag{3.3}$$

(siehe z.B. [5, Seite 54]).
Zur Berechnung von η kann nun das exakte Produkt $AA^{-1\mathcal{H}}$ durch das $\mathcal{H}$-Produkt $A *_{\mathcal{H}} A^{-1\mathcal{H}}$ ersetzt werden.
Andererseits folgt aus

$$A^{-1} - A^{-1\mathcal{H}} = A^{-1}\left(I - AA^{-1\mathcal{H}}\right)$$

durch Anwendung der Dreiecksungleichung in einer beliebigen Norm $\|\cdot\|$ folgende obere Schranke für den relativen normweisen Vorwärtsfehler:

$$\frac{\|A^{-1} - A^{-1\mathcal{H}}\|}{\|A^{-1}\|} \leq \|I - AA^{-1\mathcal{H}}\| \tag{3.4}$$

(siehe [17]).
Gleichung (3.3) für den Rückwärtsfehler und Abschätzung (3.4) für den Vorwärtsfehler ergeben zusammen

$$\frac{\|A^{-1} - A^{-1\mathcal{H}}\|}{\|A^{-1}\|} \leq \|A\| \, \|A^{-1\mathcal{H}}\| \, \eta = \kappa(A) \, \eta,$$

also völlig im Rahmen der Faustformel

„ Vorwärtsfehler $\leq$ Kondition $\cdot$ Rückwärtsfehler “.

In den Abbildungen 3.9 und 3.10 wurden neben den mitlaufenden a posteriori Vorwärtsfehlerschätzern auch die Schätzungen

$$\frac{\|I - A *_{\mathcal{H}} A^{-1\mathcal{H}}\|_F}{\|I\|_F} = \frac{\|I - A *_{\mathcal{H}} A^{-1\mathcal{H}}\|_F}{n}$$

für die relativen normweisen Vorwärtsfehler der $\mathcal{H}$-Inversen $A_h^{-1\mathcal{H}}$ und $M_h^{-1\mathcal{H}}$ in der Frobeniusnorm dargestellt.

Bemerkung:
Zur Approximation der Spektralnorm $\|A\|_2$ einer $\mathcal{H}$-Matrix A mit fast linearem Aufwand bietet sich folgende Vorgehensweise an:
Wegen $\|A\|_2 = \sqrt{\rho(A^T A)}$ bilde man zunächst das symmetrisch positiv semidefinite $\mathcal{H}$-Produkt $A^T *_{\mathcal{H}} A$ und approximiere dann mittels Vektoriteration dessen größten Eigenwert $\rho(A^T *_{\mathcal{H}} A)$.
Die Ausarbeitung dieser Idee zur Approximation der Spektralnorm einer $\mathcal{H}$-Matrix und Bemerkungen zum Konvergenzverhalten finden sich in [17, Kapitel 4.6, Seite 62].

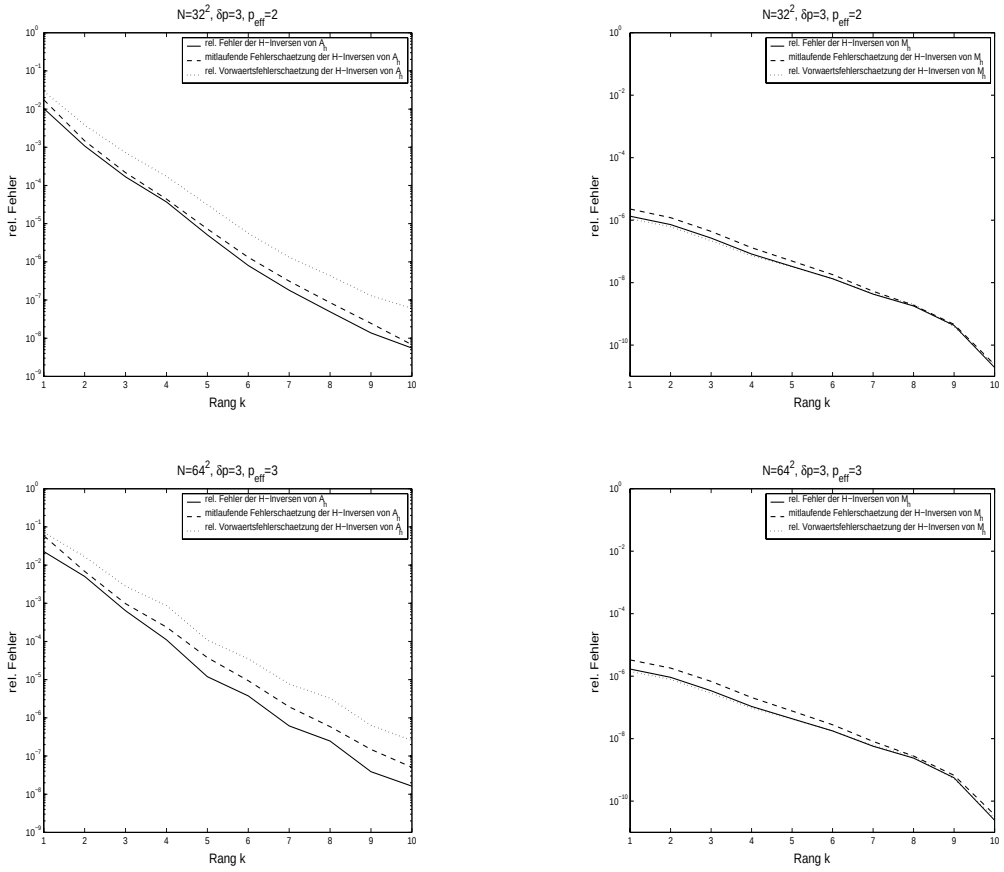

Abbildung 3.9: Relativer Fehler sowie mitlaufende a posteriori Fehlerschätzung und relative Vorwärtsfehlerschätzung der $\mathcal{H}$-Inversen von A_h (links) und von M_h (rechts) in $\|\cdot\|_F$ für $N = 32^2$ (oben) und $N = 64^2$ (unten) Unbekannte und $\delta p = 3$.

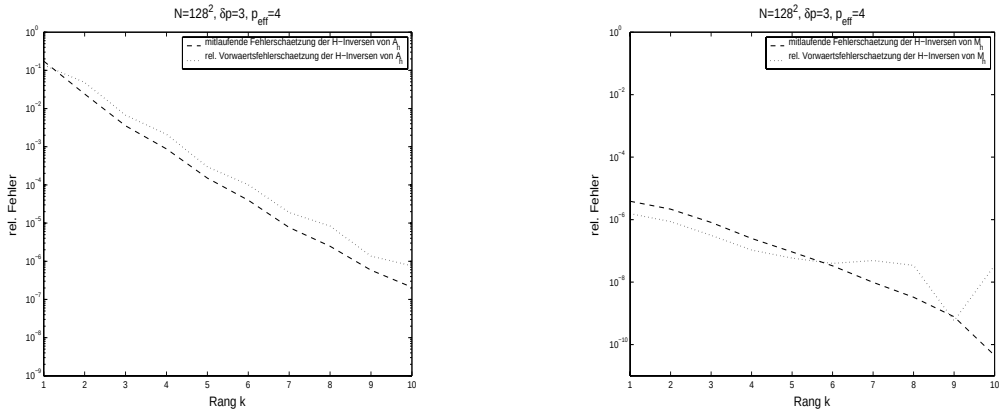

Abbildung 3.10: Mitlaufende a posteriori Fehlerschätzung und relative Vorwärtsfehlerschätzung der $\mathcal{H}$-Inversen von A_h (links) und von M_h (rechts) in $\|\cdot\|_F$ für $N = 128^2$ Unbekannte und $\delta p = 3$.

Für äußerst schlecht konditionierte Matrizen ist der exakte $\mathcal{H}$-Fehler der $\mathcal{H}$-Invertierung mittels approximativer Block-Gauß-Elimination i. Allg. zu groß: Die $\mathcal{H}$-Fehler der einzelnen Teiloperationen werden durch die großen Konditionszahlen enorm verstärkt und jeweils von einem Level zum nächsthöheren weitergereicht. In solchen Fällen wird der tatsächliche $\mathcal{H}$-Fehler durch die oben konstruierte a posteriori Schranke maßlos überschätzt, da die großen Konditionszahlen den tatsächlichen Fehlertransport viel zu pessimistisch wiedergeben. Der Vergleich von A_h mit den äußerst schlecht konditionierten Matrizen A_h^2 und A_h^4 gibt diese Situation deutlich wieder (siehe die Abbildungen 3.11 und 3.12).

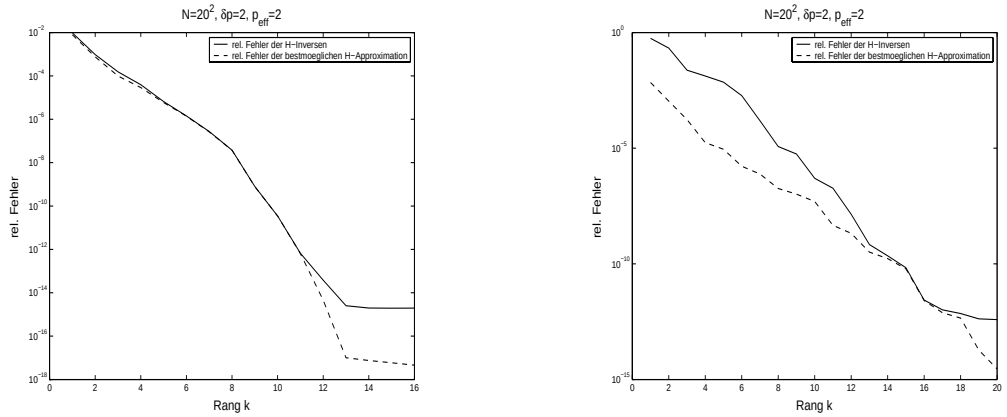

Abbildung 3.11: Relativer Fehler der $\mathcal{H}$-Inversen von A_h (links) und von A_h^2 (rechts) sowie der bestmöglichen $\mathcal{H}$-Approximation an die exakte Inverse von A_h (links) und von A_h^2 (rechts) in $\|\cdot\|_F$.

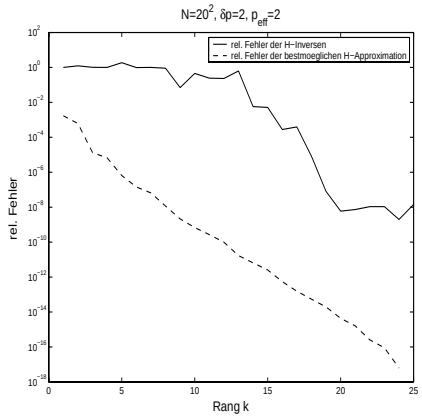

Abbildung 3.12: Relativer Fehler der $\mathcal{H}$-Inversen von A_h^4 sowie der bestmöglichen $\mathcal{H}$-Approximation an die exakte Inverse von A_h^4 in $\|\cdot\|_F$.

Beobachtung: Je größer die Kondition der zu invertierenden Matrix, umso größer muß der Rang der Rk-Blöcke gewählt werden, um mit dem approximierten rekursiven Block-Gauß-Algorithmus eine bestimmte Approximationsgüte zu erreichen.

3.4.4.4 Zusammenfassung

- In den Kapiteln 3.4.1 bis 3.4.4 wurden nach detaillierter Ausarbeitung der hierarchischen Grundoperationen für 2D $\mathcal{H}$-Matrizen aus [20] geeignete Fehlerschätzer für die approximierten $\mathcal{H}$-Operationen neu entwickelt und in der numerischen Praxis erprobt.

- Die Approximationsgüte der $\mathcal{H}$-Inversen $A^{-1_{\mathcal{H}}}$ einer $\mathcal{H}$-Matrix A hängt (natürlich neben der Darstellbarkeit der exakten Inversen A^{-1} als $\mathcal{H}$-Matrix) stark von der Kondition $\kappa(A)$ ab.

3.4.5 $\mathcal{H}$-Cholesky-Zerlegung

Der bereits behandelten $\mathcal{H}$-Invertierung mittels Gaußscher Elimination liegen die zwei folgenden Prinzipien zugrunde: die Formulierung des Gaußschen Eliminationsalgorithmus als 4×4-Blockalgorithmus sowie die Ersetzung der exakten Blockoperationen durch ihre approximierten hierarchischen Pendants.
Diese bilden auch die Grundlage für die folgenden hierarchischen Zerlegungen.

Zunächst wenden wir unsere Aufmerksamkeit der Lösung linearer Gleichungssysteme (LGS) mit hierarchischen unteren Block-Dreiecksmatrizen L zu. (Man beachte, dass die 4 Diagonalblöcke $L_{11}, \ldots, L_{44}$ von L wiederum untere Block-Dreiecksmatrizen sind.)
Hierarchische LGS werden nämlich sowohl gleich danach für die hierarchische Cholesky-Zerlegung als auch in den späteren Kapiteln wie beispielsweise für das hierarchische Eigenwertproblem des diskreten Laplace-Operators benötigt. Die $\mathcal{H}$-Cholesky-Zerlegung hingegen wird noch die zentrale Rolle in unserem Algorithmus zur $\mathcal{H}$-QR-Zerlegung spielen.

3.4.5.1 Lineare Gleichungssysteme in $\mathcal{H}$-Arithmetik

Wie der $\mathcal{H}$-Multiplikation die $\mathcal{H}$-Matrix-Rk-Matrix-Multiplikation liegt dem $\mathcal{H}$-LGS mit hierarchischen 4×4-Blöcken als rechten Seiten das **Rk-LGS**

$$LR_1 = R_2 \text{ bzw. } L^T R_1 = R_2$$

mit einer Rk-Matrix R_1 als Lösung sowie einer Rk-Matrix R_2 als rechter Seite zugrunde.
Mit $R_1 = [\mathbf{x}, \mathbf{y}] = \mathbf{x}\mathbf{y}^T$ und $R_2 = [\mathbf{a}, \mathbf{b}] = \mathbf{a}\mathbf{b}^T$ schreibt sich $LR_1 = R_2$ in der Form $L\mathbf{x}\mathbf{y}^T = \mathbf{a}\mathbf{b}^T$, was folgenden Lösungsweg nahelegt:
Löse $L\mathbf{x} = \mathbf{a}$ nach $\mathbf{x}$ und setze $\mathbf{y} = \mathbf{b}$.
Nun ist

$$L\mathbf{x} = \mathbf{a} \quad \Longleftrightarrow \quad \begin{array}{|c|c|c|c|} \hline L_{11} & & & \\ \hline L_{21} & L_{22} & & \\ \hline L_{31} & L_{32} & L_{33} & \\ \hline L_{41} & L_{42} & L_{43} & L_{44} \\ \hline \end{array} \cdot \begin{array}{|c|} \hline \mathbf{x}_1 \\ \hline \mathbf{x}_2 \\ \hline \mathbf{x}_3 \\ \hline \mathbf{x}_4 \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{a}_1 \\ \hline \mathbf{a}_2 \\ \hline \mathbf{a}_3 \\ \hline \mathbf{a}_4 \\ \hline \end{array}$$

$$\iff \quad \begin{aligned} L_{11}\mathbf{x}_1 &= \mathbf{a}_1 \\ L_{22}\mathbf{x}_2 &= \mathbf{a}_2 - L_{21}\mathbf{x}_1 \\ L_{33}\mathbf{x}_3 &= \mathbf{a}_3 - L_{31}\mathbf{x}_1 - L_{32}\mathbf{x}_2 \\ L_{44}\mathbf{x}_4 &= \mathbf{a}_4 - L_{41}\mathbf{x}_1 - L_{42}\mathbf{x}_2 - L_{43}\mathbf{x}_3 \end{aligned}$$

Die rekursive Lösung der Vorwärtssubstitution $L\mathbf{x} = \mathbf{a}$ besteht also in deren Rückführung auf 4 Vorwärtssubstitutionen und 6 Matrix-Vektor-Multiplikationen mit Außerdiagonalblöcken auf dem nächsthöheren Level.

Die Rückwärtssubstitution $L^T\mathbf{x} = \mathbf{a}$ und somit die Lösung von $L^T R_1 = R_2$ werden völlig analog gehandhabt.

Man beachte, dass die Vorwärtssubstitution $L\mathbf{x} = \mathbf{a}$ wie auch die Rückwärts-substitution $L^T\mathbf{x} = \mathbf{a}$ mit hierarchischen unteren Block-Dreiecksmatrizen L wie bereits die hierarchische Matrix-Vektor-Multiplikation aus Abschnitt 3.4.1 exakt ist.

• Komplexität der Vorwärts- und Rückwärtssubstitution $L\mathbf{x} = \mathbf{a}$ und $L^T\mathbf{x} = \mathbf{a}$ sowie der Rk-LGS $LR_1 = R_2$ und $L^T R_1 = R_2$:
Mit $N = 4^p$ und der Tatsache, dass die hierarchische Matrix-Vektor-Multiplikation mit Außerdiagonalblöcken nur $\mathcal{O}(N)$ Operationen benötigt, folgt aus

$$\mathcal{N}_{Rk-LGS}(p) = 4\,\mathcal{N}_{Rk-LGS}(p-1) + \mathcal{O}(N)$$

für die Anzahl der flops eines Rk-LGS vom Level p

$$\mathcal{N}_{Rk-LGS}(p) = \mathcal{O}(pN) = \mathcal{O}(N \log N).$$

In einem zweiten Schritt behandeln wir nun die **×-LGS** der Form

$$LX_\times = B_\times \text{ bzw. } L^T X_\times = B_\times,$$

wobei $X_\times$ und $B_\times$ einen $\nearrow$-, $\searrow$-, $\swarrow$- oder $\nwarrow$-Block bezeichnen.
Wir veranschaulichen deren rekursive Struktur anhand einer $\mathcal{H}$-Matrix-Vorwärtssubstitution $LX_\nwarrow = B_\nwarrow$ mit einer $\nwarrow$-Matrix als rechter Seite:

Das Matrix-Gleichungssystem

L_{11}			
L_{21}	L_{22}		
L_{31}	L_{32}	L_{33}	
L_{41}	L_{42}	L_{43}	L_{44}

$\cdot$

X_{11}	X_{12}	$\nwarrow$	X_{14}
X_{21}	X_{22}	X_{23}	X_{24}
X_{31}	X_{32}	X_{33}	X_{34}
X_{41}	X_{42}	X_{43}	X_{44}

$=$

R_{11}	R_{12}	$\nwarrow$	R_{14}
R_{21}	R_{22}	R_{23}	R_{24}
R_{31}	R_{32}	R_{33}	R_{34}
R_{41}	R_{42}	R_{43}	R_{44}

(zeilenweise für X_{ij} von links oben nach rechts unten gelöst) ist äquivalent zu einem $\nwarrow$-LGS und 15 Rk-LGS auf dem nächsthöheren Level:

$$
\begin{aligned}
L_{11}\,X_{11} &= R_{11}\\
L_{11}\,X_{12} &= R_{12}\\
L_{11}\,\nwarrow &= \nwarrow\\
L_{11}\,X_{14} &= R_{14}
\end{aligned}
$$

$$
\begin{aligned}
L_{22}\,X_{21} &= R_{21} - L_{21}\,X_{11}\\
L_{22}\,X_{22} &= R_{22} - L_{21}\,X_{12}\\
L_{22}\,X_{23} &= R_{23} - L_{21}\,\nwarrow\\
L_{22}\,X_{24} &= R_{24} - L_{21}\,X_{14}
\end{aligned}
$$

$$
\begin{aligned}
L_{33}\,X_{31} &= R_{31} - L_{31}\,X_{11} - L_{32}\,X_{21}\\
L_{33}\,X_{32} &= R_{32} - L_{31}\,X_{12} - L_{32}\,X_{22}\\
L_{33}\,X_{33} &= R_{33} - L_{31}\,\nwarrow\, - L_{32}\,X_{23}\\
L_{33}\,X_{34} &= R_{34} - L_{31}\,X_{14} - L_{32}\,X_{24}
\end{aligned}
$$

$$
\begin{aligned}
L_{44}\,X_{41} &= R_{41} - L_{41}\,X_{11} - L_{42}\,X_{21} - L_{43}\,X_{31}\\
L_{44}\,X_{42} &= R_{42} - L_{41}\,X_{12} - L_{42}\,X_{22} - L_{43}\,X_{32}\\
L_{44}\,X_{43} &= R_{43} - L_{41}\,\nwarrow\, - L_{42}\,X_{23} - L_{43}\,X_{33}\\
L_{44}\,X_{44} &= R_{44} - L_{41}\,X_{14} - L_{42}\,X_{24} - L_{43}\,X_{34}
\end{aligned}
$$

Man beachte, dass die $\mathcal{H}$-Matrix-Produkte $L_{21} *_{\mathcal{H}} \nwarrow = \leftarrow *_{\mathcal{H}} \nwarrow$, $L_{31} *_{\mathcal{H}} \nwarrow = \nwarrow *_{\mathcal{H}} \nwarrow$ sowie $L_{41} *_{\mathcal{H}} \nwarrow = \uparrow *_{\mathcal{H}} \nwarrow$ Rk-Matrizen ergeben und alle Subtraktionen von Rk-Matrizen jeweils auf geeigneten Rang abgeschnitten werden.

• Komplexität der $\times$-LGS $LX_\times = B_\times$ und $L^T X_\times = B_\times$:
Es ist

$$
\begin{aligned}
\mathcal{N}_{\times-LGS}(p) &= \mathcal{N}_{\times-LGS}(p-1) + 15\,\mathcal{N}_{Rk-LGS}(p-1) + \mathcal{O}(N)\\
&= \mathcal{N}_{\times-LGS}(p-1) + CpN + \mathcal{O}(N)
\end{aligned}
$$

und deshalb

$$
\mathcal{N}_{\times-LGS}(p) = \frac{4}{3}\,CpN + \mathcal{O}(N) = \mathcal{O}(pN) = \mathcal{O}(N \log N).
$$

Dabei ist zu beachten, dass sämtliche L_{ij}-Rk-Matrix-Multiplikationen mit Außerdiagonalblöcken L_{ij}, die drei oben genannten $\mathcal{H}$-Block-Multiplikationen sowie sämtliche Rk-Additionen im $\mathcal{O}(N)$-Term enthalten sind.

In einem dritten Schritt kommen wir nun zu den **+-LGS**

$$
LX_+ = B_+ \text{ bzw. } L^T X_+ = B_+,
$$

wobei + für eines der Formate $\uparrow$, $\downarrow$, $\leftarrow$ oder $\rightarrow$ steht.

Zu lösen sei beispielsweise $LX_\leftarrow = B_\leftarrow$:

L_{11}			
L_{21}	L_{22}		
L_{31}	L_{32}	L_{33}	
L_{41}	L_{42}	L_{43}	L_{44}

$\cdot$

X_{11}	$\leftarrow$	$\swarrow$	X_{14}
X_{21}	X_{22}	X_{23}	X_{24}
X_{31}	X_{32}	X_{33}	X_{34}
X_{41}	$\nwarrow$	$\leftarrow$	X_{44}

$=$

R_{11}	$\leftarrow$	$\swarrow$	R_{14}
R_{21}	R_{22}	R_{23}	R_{24}
R_{31}	R_{32}	R_{33}	R_{34}
R_{41}	$\nwarrow$	$\leftarrow$	R_{44}

Rückführung auf zwei $\leftarrow$-LGS, je ein $\nwarrow$- und $\swarrow$-LGS sowie 12 Rk-LGS auf dem nächsthöheren Level liefert wiederum

$$\begin{aligned}\mathcal{N}_{+-LGS}(p) &= 2\,\mathcal{N}_{+-LGS}(p-1) + 2\,\mathcal{N}_{\times-LGS}(p-1) + \\ &\quad 12\,\mathcal{N}_{Rk-LGS}(p-1) + \mathcal{O}(N) \\ &= 2\,\mathcal{N}_{+-LGS}(p-1) + CpN + \mathcal{O}(N)\end{aligned}$$

mit der Komplexität $\mathcal{O}(N)$ für die restlichen durchzuführenden $\mathcal{H}$-Matrix-Multiplikationen und $\mathcal{H}$-Additionen und damit

$$\mathcal{N}_{+-LGS}(p) = 2\,CpN + \mathcal{O}(N) = \mathcal{O}(pN) = \mathcal{O}(N \log N).$$

Im vierten und letzten Schritt können wir nun endlich ein **$\square$-LGS**

$$LX_\square = B_\square \text{ bzw. } L^T X_\square = B_\square$$

durch Rückführung auf 4 $\square$-LGS, je 2 $\rightarrow$-, $\leftarrow$-, $\uparrow$- und $\downarrow$-LGS und je ein $\nearrow$-, $\searrow$-, $\swarrow$- und $\nwarrow$-LGS auf dem nächsthöheren Level lösen. Es ist nämlich $LX_\square = B_\square$ äquivalent zu

$$\begin{array}{|c|c|c|c|}\hline L_{11} & & & \\ \hline L_{21} & L_{22} & & \\ \hline L_{31} & L_{32} & L_{33} & \\ \hline L_{41} & L_{42} & L_{43} & L_{44} \\ \hline\end{array} \cdot \begin{array}{|c|c|c|c|}\hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline\end{array} = \begin{array}{|c|c|c|c|}\hline \square & \rightarrow & \searrow & \downarrow \\ \hline \leftarrow & \square & \downarrow & \swarrow \\ \hline \nwarrow & \uparrow & \square & \leftarrow \\ \hline \uparrow & \nearrow & \rightarrow & \square \\ \hline\end{array}.$$

Es gilt also

$$\begin{aligned}\mathcal{N}_{\square-LGS}(p) &= 4\,\mathcal{N}_{\square-LGS}(p-1) + 8\,\mathcal{N}_{+-LGS}(p-1) + \\ &\quad 4\,\mathcal{N}_{\times-LGS}(p-1) + \mathcal{O}(pN) \\ &= 4\,\mathcal{N}_{\square-LGS}(p-1) + \mathcal{O}(pN)\end{aligned}$$

und damit

$$\mathcal{N}_{\square-LGS}(p) = \mathcal{O}(p^2 N) = \mathcal{O}(N \log^2 N)$$

für die Komplexität eines $\square$-LGS.

3.4.5.2 Das rekursive Schema der $\mathcal{H}$-Cholesky-Zerlegung

Gegeben sei die symmetrisch positiv definite $\mathcal{H}$-Matrix $A = (A_{ij})_{1\le i,j\le 4} \in \mathbb{R}^{N\times N}$. Gesucht ist eine $\mathcal{H}$-Approximation $L_\mathcal{H}$ an L, wobei $A = LL^T$ die exakte Cholesky-Zerlegung von A ist.
Mit

$$A = \begin{array}{|c|c|c|c|}\hline A_{11} & A_{21}^T & A_{31}^T & A_{41}^T \\ \hline A_{21} & A_{22} & A_{32}^T & A_{42}^T \\ \hline A_{31} & A_{32} & A_{33} & A_{43}^T \\ \hline A_{41} & A_{42} & A_{43} & A_{44} \\ \hline\end{array} \quad \text{und} \quad L = \begin{array}{|c|c|c|c|}\hline L_{11} & & & \\ \hline L_{21} & L_{22} & & \\ \hline L_{31} & L_{32} & L_{33} & \\ \hline L_{41} & L_{42} & L_{43} & L_{44} \\ \hline\end{array}$$

ist $LL^T = A$ äquivalent zu den folgenden Gleichungen für die 10 zu bestimmenden Blöcke von L:

$$
\begin{aligned}
L_{11}\, L_{11}^T &= A_{11} \\
L_{21}\, L_{11}^T &= A_{21} \\
L_{31}\, L_{11}^T &= A_{31} \\
L_{41}\, L_{11}^T &= A_{41}
\end{aligned}
$$

$$
\begin{aligned}
L_{22}\, L_{22}^T &= A_{22} - L_{21}\, L_{21}^T \\
L_{32}\, L_{22}^T &= A_{32} - L_{31}\, L_{21}^T \\
L_{42}\, L_{22}^T &= A_{42} - L_{41}\, L_{21}^T
\end{aligned}
$$

$$
\begin{aligned}
L_{33}\, L_{33}^T &= A_{33} - L_{31}\, L_{31}^T - L_{32}\, L_{32}^T \\
L_{43}\, L_{33}^T &= A_{43} - L_{41}\, L_{31}^T - L_{42}\, L_{32}^T
\end{aligned}
$$

$$
L_{44}\, L_{44}^T = A_{44} - L_{41}\, L_{41}^T - L_{42}\, L_{42}^T - L_{43}\, L_{43}^T
$$

Daraus lässt sich sofort die rekursive Struktur der hierarchischen Cholesky-Zerlegung ablesen: Sie besteht in der Rückführung von $LL^T = A$ auf 4 $\mathcal{H}$-Cholesky-Zerlegungen und 6 $\mathcal{H}$-Matrix-Vorwärtssubstitutionen mit Außerdiagonalblöcken als rechten Seiten auf dem nächsthöheren Level.

Eine $\mathcal{H}$-Matrix-Vorwärtssubstitution $LX = B$ bzw. eine $\mathcal{H}$-Matrix-Rückwärtssubstitution $L^T X = B$ benötigt $\mathcal{O}(N \log N)$ Operationen für einen Außerdiagonalblock B sowie $\mathcal{O}(N \log^2 N)$ Operationen für einen Diagonalblock B (siehe Abschnitt 3.4.5.1).

k \ N	16^2	32^2	64^2	128^2
1	0.04	0.55	4.90	31.3
2	0.04	0.63	5.88	38.5
3	0.04	0.68	6.74	48.1
4	0.04	0.71	7.82	57.0
5	0.04	0.84	10.03	78.1
6	0.04	0.90	11.45	92.0
7	0.04	1.09	14.60	120.5
8	0.04	1.13	14.87	122.6
9	0.04	1.25	18.50	155.6
10	0.04	1.44	21.44	182.1

Tabelle 3.5: Benötigte Zeit (in Sekunden) zur Durchführung der $\mathcal{H}$-Cholesky-Zerlegung der $N \times N$-$\mathcal{H}$-Steifigkeitsmatrix A_h mit $\delta p = 3$ und Rang k (Prozessor: Sun Ultra Sparc IIi, 300 MHz).

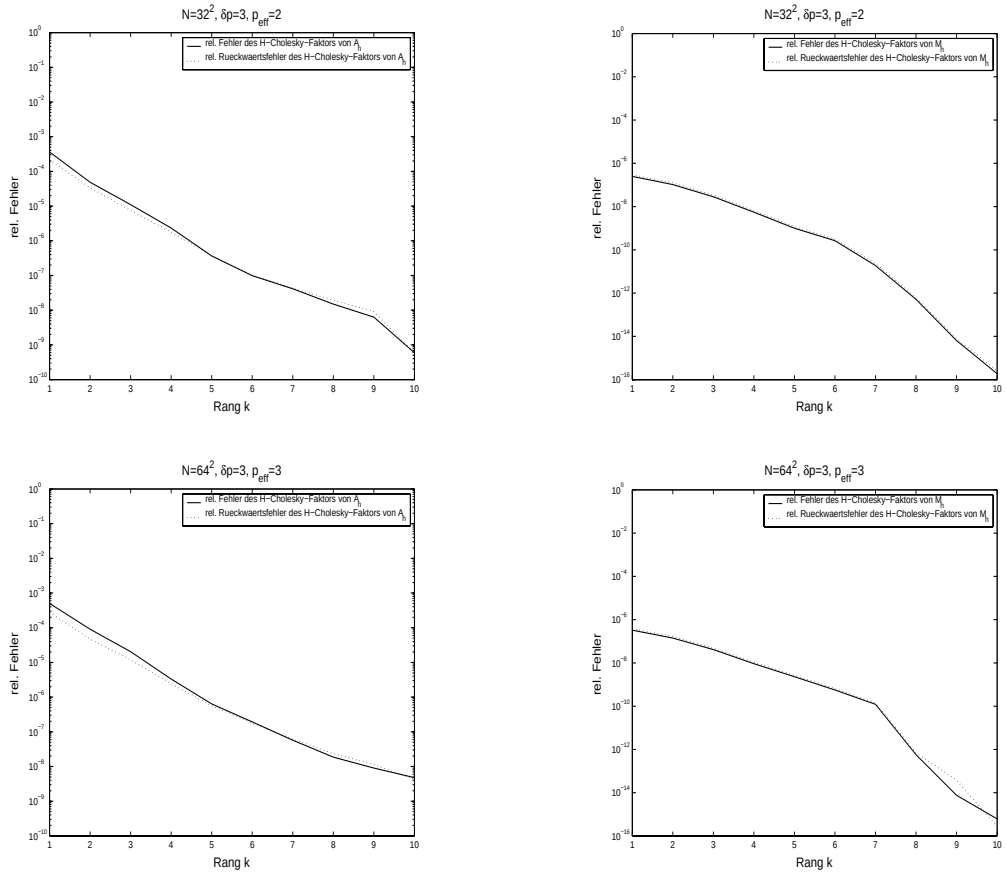

Abbildung 3.13: Relativer Fehler sowie relativer Rückwärtsfehler des $\mathcal{H}$-Cholesky-Faktors L_h von A_h (links) und von M_h (rechts) in $\|\cdot\|_F$ für $N = 32^2$ (oben) und $N = 64^2$ (unten) Unbekannte und $\delta p = 3$.

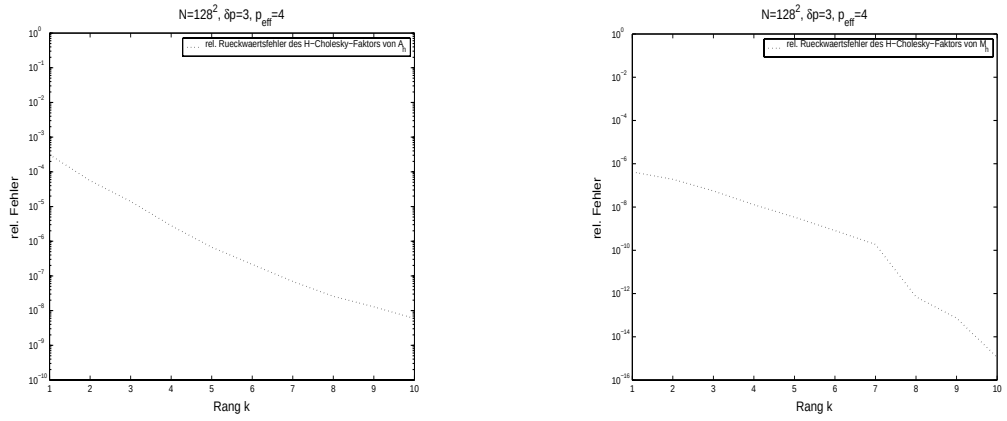

Abbildung 3.14: Relativer Rückwärtsfehler des $\mathcal{H}$-Cholesky-Faktors L_h von A_h (links) und von M_h (rechts) in $\|\cdot\|_F$ für $N = 128^2$ Unbekannte und $\delta p = 3$.

Damit gilt mit der in diesem Abschnitt aufgezeigten rekursiven Struktur der $\mathcal{H}$-Cholesky-Zerlegung

$$\mathcal{N}_{Chol}(p) = 4\,\mathcal{N}_{Chol}(p-1) + \mathcal{O}(pN)$$

und somit

$$\mathcal{N}_{Chol}(p) = \mathcal{O}(p^2 N) = \mathcal{O}(N \log^2 N)$$

(siehe Tabelle 3.5).

3.4.5.3 Normweise Rückwärtsanalyse für den $\mathcal{H}$-Fehler der $\mathcal{H}$-Cholesky-Zerlegung

In exakter Arithmetik gilt $A = LL^T$. Nun gibt es ein eindeutig bestimmtes symmetrisches $\Delta A \in \mathbb{R}^{N\times N}$ mit $A + \Delta A = \hat{L}\hat{L}^T$ in exakter Arithmetik, wobei $\hat{L} = L_{\mathcal{H}}$ die berechnete Approximation an L und ΔA den Rückwärtsfehler bezeichnen.
Damit gilt

$$\frac{\|\Delta A\|}{\|A\|} = \frac{\|A - \hat{L}\hat{L}^T\|}{\|A\|} \tag{3.5}$$

für den relativen normweisen Rückwärtsfehler.

Zur Schätzung von $\|A - \hat{L}\hat{L}^T\|$ ersetzen wir die exakte Matrixmultiplikation $\cdot$ einfach durch die approximierte $\mathcal{H}$-Matrix-Multiplikation $*_{\mathcal{H}}$.

In den Abbildungen 3.13 und 3.14 sind die Rückwärtsfehler aus (3.5) für die $\mathcal{H}$-Cholesky-Zerlegung der Steifigkeitsmatrix A_h und der Massenmatrix M_h dargestellt.

Die Approximationsgüte des $\mathcal{H}$-Cholesky-Faktors $L_{\mathcal{H}}$:

Der oben beschriebene rekursive Algorithmus liefert eine untere $\mathcal{H}$-Dreiecksmatrix $L_{\mathcal{H}}$ als Approximation an die untere Dreiecksmatrix L, wobei L die exakte Matrixgleichung $LL^T = A$, $A \in \mathbb{R}^{N\times N}$ symmetrisch positiv definit, löst. Dabei ist – wie bereits bei der $\mathcal{H}$-Invertierung – $\kappa(A)$ verantwortlich für die Güte des Erreichens der Bestapproximation an L in der Menge aller $\mathcal{H}$-Matrizen bzgl. der Frobeniusnorm.

Je größer $\kappa(A)$, umso größer muss der Rang der Rk-Blöcke gewählt werden, um mit dem oben beschriebenen rekursiven Algorithmus eine bestimmte Approximationsgüte zu erreichen. Je größer dieser Rang, umso kleiner ist der Fehler, der von Level zu Level und innerhalb eines Levels weitertransportiert wird.

Die numerischen Beispiele in den Abbildungen 3.15 und 3.16 verdeutlichen die Abhängigkeit der Approximationsgüte von $L_{\mathcal{H}}$ von der Kondition der zu zerlegenden $\mathcal{H}$-Matrix.

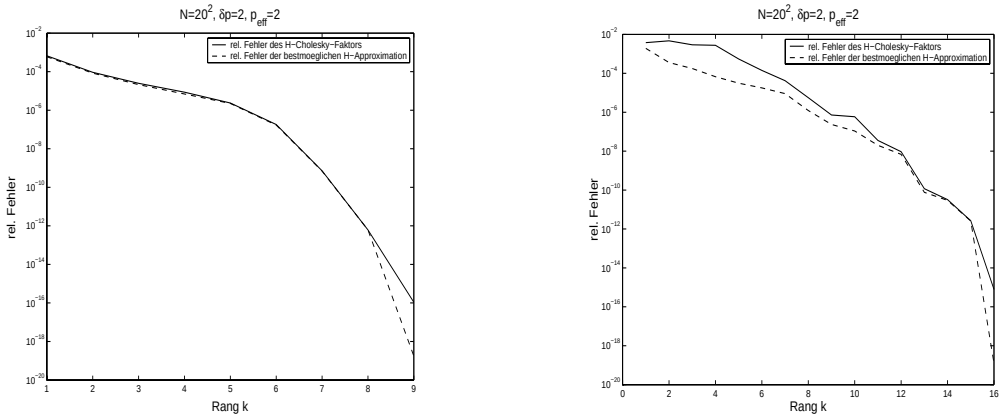

Abbildung 3.15: Relativer Fehler des $\mathcal{H}$-Cholesky-Faktors von A_h (links) und von A_h^2 (rechts) sowie der bestmöglichen $\mathcal{H}$-Approximation an den exakten Cholesky-Faktor von A_h (links) und von A_h^2 (rechts) in $\|\cdot\|_F$.

3.4.5.4 Lösung der Poisson-Gleichung mittels hierarchischer Matrizen

Die 2D Poisson-Gleichung

$$-\triangle u(x) = f(x)\,, \quad x \in \Omega = (0,1)^2$$
$$u|_{\partial\Omega} \equiv 0$$

mit $f \in L^2(\Omega)$ besitzt die schwache Formulierung

$$a(u,v) = (f,v)_{L^2} \quad \forall v \in H_0^1(\Omega). \tag{3.7}$$

Diskretisierung mittels stückweise linearer C^0-Dreieckselemente auf dem regelmäßig triangulierten Einheitsquadrat liefert das N-dimensionale LGS

$$A_h \mathbf{u}_h = M_h \mathbf{f}_h, \tag{3.8}$$

wobei A_h die Steifigkeitsmatrix und M_h die Massenmatrix in der Knotenbasis $\{\psi_j^h\}_{j=1}^N$ bezeichnen und $\mathbf{f}_h = (f_1, \ldots, f_N)^T$ mit $Q_h f = \sum_{j=1}^N f_j \psi_j^h$ ist.

Lösungsvorschläge für $A_h \mathbf{u}_h = M_h \mathbf{f}_h$:

1. Zunächst $\mathcal{H}$-Invertierung von A_h, danach Matrix-Vektor-Multiplikationen $A_h^{-1_{\mathcal{H}}}(M_h \mathbf{f}_h)$.

2. Zunächst $\mathcal{H}$-Cholesky-Zerlegung $A_h = L_h L_h^T$, danach Vorwärtssubstitution $L_h \mathbf{z}_h = M_h \mathbf{f}_h$ und Rückwärtssubstitution $L_h^T \mathbf{u}_h = \mathbf{z}_h$.

Wie die numerische Praxis zeigt, ist die zweite Methode zur Lösung von (3.8) sowohl schneller als auch ein wenig genauer als die erste. Dies liegt im Wesentlichen daran, dass die Cholesky-Zerlegung schneller und ein wenig genauer als die $\mathcal{H}$-Invertierung mittels Gaußscher Elimination ist.

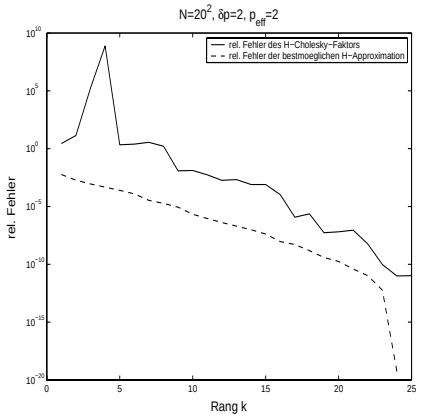

Abbildung 3.16: Relativer Fehler des $\mathcal{H}$-Cholesky-Faktors von A_h^4 sowie der bestmöglichen $\mathcal{H}$-Approximation an den exakten Cholesky-Faktor von A_h^4 in $\|\cdot\|_F$.

3.4.6 $\mathcal{H}$-LDL^T-Zerlegung

Analog zur $\mathcal{H}$-Cholesky-Zerlegung einer symmetrisch positiv definiten $\mathcal{H}$-Matrix kann auch die $\mathcal{H}$-LDL^T-Zerlegung einer symmetrischen $\mathcal{H}$-Matrix durchgeführt werden.

3.4.6.1 Das rekursive Schema der $\mathcal{H}$-LDL^T-Zerlegung

Sei $A = (A_{ij})_{1\le i,j\le 4} \in \mathbb{R}^{N\times N}$ eine symmetrische $\mathcal{H}$-Matrix.
Dann ist $A = LDL^T = L \cdot DL^T$ äquivalent zu

$$\begin{array}{|c|c|c|c|}\hline A_{11} & A_{21}^T & A_{31}^T & A_{41}^T \\ \hline A_{21} & A_{22} & A_{32}^T & A_{42}^T \\ \hline A_{31} & A_{32} & A_{33} & A_{43}^T \\ \hline A_{41} & A_{42} & A_{43} & A_{44} \\ \hline \end{array} =$$

$$\begin{array}{|c|c|c|c|}\hline L_{11} & & & \\ \hline L_{21} & L_{22} & & \\ \hline L_{31} & L_{32} & L_{33} & \\ \hline L_{41} & L_{42} & L_{43} & L_{44} \\ \hline \end{array} \cdot \begin{array}{|c|c|c|c|}\hline D_{11} & & & \\ \hline & D_{22} & & \\ \hline & & D_{33} & \\ \hline & & & D_{44} \\ \hline \end{array} \cdot \begin{array}{|c|c|c|c|}\hline L_{11}^T & L_{21}^T & L_{31}^T & L_{41}^T \\ \hline & L_{22}^T & L_{32}^T & L_{42}^T \\ \hline & & L_{33}^T & L_{43}^T \\ \hline & & & L_{44}^T \\ \hline \end{array} =$$

$$\begin{array}{|c|c|c|c|}\hline L_{11} & & & \\ \hline L_{21} & L_{22} & & \\ \hline L_{31} & L_{32} & L_{33} & \\ \hline L_{41} & L_{42} & L_{43} & L_{44} \\ \hline \end{array} \cdot \begin{array}{|c|c|c|c|}\hline D_{11}L_{11}^T & D_{11}L_{21}^T & D_{11}L_{31}^T & D_{11}L_{41}^T \\ \hline & D_{22}L_{22}^T & D_{22}L_{32}^T & D_{22}L_{42}^T \\ \hline & & D_{33}L_{33}^T & D_{33}L_{43}^T \\ \hline & & & D_{44}L_{44}^T \\ \hline \end{array}$$

Die $\mathcal{H}$-LDL^T-Zerlegung von A lässt sich zurückführen auf 4 $\mathcal{H}$-LDL^T-Zerlegungen und 6 $\mathcal{H}$-Matrix-Vorwärtssubstitutionen mit Außerdiagonalblöcken als rechten Seiten auf dem nächsthöheren Level:

$$
\begin{aligned}
L_{11}D_{11}L_{11}^T &= A_{11} && \longrightarrow && D_{11}, L_{11}\\
L_{21}D_{11}L_{11}^T &= A_{21} && \longrightarrow && L_{21}\\
L_{31}D_{11}L_{11}^T &= A_{31} && \longrightarrow && L_{31}\\
L_{41}D_{11}L_{11}^T &= A_{41} && \longrightarrow && L_{41}\\
L_{22}D_{22}L_{22}^T &= A_{22} - L_{21}D_{11}L_{21}^T && \longrightarrow && D_{22}, L_{22}\\
L_{32}D_{22}L_{22}^T &= A_{32} - L_{31}D_{11}L_{21}^T && \longrightarrow && L_{32}\\
L_{42}D_{22}L_{22}^T &= A_{42} - L_{41}D_{11}L_{21}^T && \longrightarrow && L_{42}\\
L_{33}D_{33}L_{33}^T &= A_{33} - L_{31}D_{11}L_{31}^T - L_{32}D_{22}L_{32}^T && \longrightarrow && D_{33}, L_{33}\\
L_{43}D_{33}L_{33}^T &= A_{43} - L_{41}D_{11}L_{31}^T - L_{42}D_{22}L_{32}^T && \longrightarrow && L_{43}\\
L_{44}D_{44}L_{44}^T &= A_{44} - L_{41}D_{11}L_{41}^T - L_{42}D_{22}L_{42}^T - L_{43}D_{33}L_{43}^T && \longrightarrow && D_{44}, L_{44}
\end{aligned}
$$

Dabei ist die Diagonalmatrix D nach vollendeter $\mathcal{H}$-LDL^T-Zerlegung von A nicht nur blockdiagonal, sondern eine echte Diagonalmatrix.

• Für die Komplexität der $\mathcal{H}$-LDL^T-Zerlegung gilt völlig analog zur $\mathcal{H}$-Cholesky-Zerlegung

$$\mathcal{N}_{LDL^T}(p) = 4\,\mathcal{N}_{LDL^T}(p-1) + \mathcal{O}(pN)$$

und somit

$$\mathcal{N}_{LDL^T}(p) = \mathcal{O}(p^2 N) = \mathcal{O}(N \log^2 N)$$

(siehe Tabelle 3.6).

k \ N	16^2	32^2	64^2	128^2
1	0.04	0.64	5.07	32.5
2	0.04	0.65	6.08	40.4
3	0.04	0.71	7.15	50.3
4	0.04	0.79	8.20	59.8
5	0.04	0.86	10.42	81.4
6	0.04	1.06	12.01	95.8
7	0.04	1.11	15.20	124.2
8	0.04	1.15	15.63	127.2
9	0.04	1.36	19.50	162.2
10	0.04	1.49	23.20	189.5

Tabelle 3.6: Benötigte Zeit (in Sekunden) zur Durchführung der $\mathcal{H}$-LDL^T-Zerlegung der symmetrisch indefiniten $N \times N$-$\mathcal{H}$-Matrix $\mu M_h - A_h$ mit $\mu = 100$, $\delta p = 3$ und Rang k (Prozessor: Sun Ultra Sparc IIi, 300 MHz).

Bemerkung: Die $\mathcal{H}$-LDL^T-Zerlegung wird später zur Lösung linearer Gleichungssysteme mit den symmetrisch indefiniten $\mathcal{H}$-Matrizen $\mu M_h - A_h$, $\mu \in \mathbb{R}$, im Zuge der simultanen Iteration zur Lösung des $\mathcal{H}$-Eigenwertproblems des diskreten Laplace-Operators benötigt werden.

3.4.6.2 Normweise Rückwärtsanalyse für den $\mathcal{H}$-Fehler der $\mathcal{H}$-LDL^T-Zerlegung

Analog zur $\mathcal{H}$-Cholesky-Zerlegung ergibt sich der relative normweise Rückwärtsfehler für die $\mathcal{H}$-LDL^T-Zerlegung zu

$$\frac{\|\Delta A\|}{\|A\|} = \frac{\|A - \hat{L}\hat{D}\hat{L}^T\|}{\|A\|} \tag{3.9}$$

mit $A + \Delta A = \hat{L}\hat{D}\hat{L}^T$ in exakter Arithmetik.
Dabei wird das exakte Produkt $\hat{L}\left(\hat{D}\hat{L}^T\right)$ wiederum durch das $\mathcal{H}$-Produkt $\hat{L} *_{\mathcal{H}} \left(\hat{D}\hat{L}^T\right)$ ersetzt.

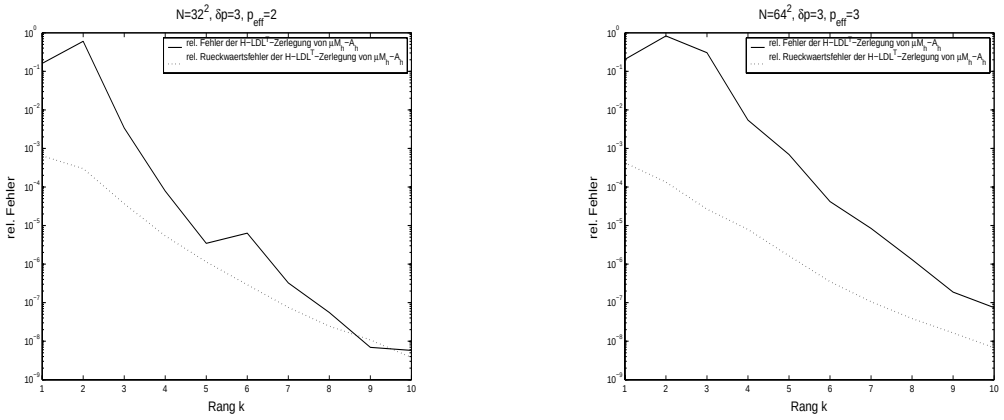

Abbildung 3.17: Relativer Fehler $\frac{\|L_{\mathcal{H}}-L\|_F+\|D_{\mathcal{H}}-D\|_F}{\|(L-I)+D\|_F}$ sowie relativer Rückwärtsfehler der $\mathcal{H}$-LDL^T-Zerlegung von $\mu M_h - A_h$ in $\|\cdot\|_F$ mit $\mu = 100$ für $N = 32^2$ (links) und $N = 64^2$ (rechts) Unbekannte und $\delta p = 3$, wobei $L_{\mathcal{H}}$ und $D_{\mathcal{H}}$ die berechneten $\mathcal{H}$-Approximationen an die exakten Faktoren L und D bezeichnen.

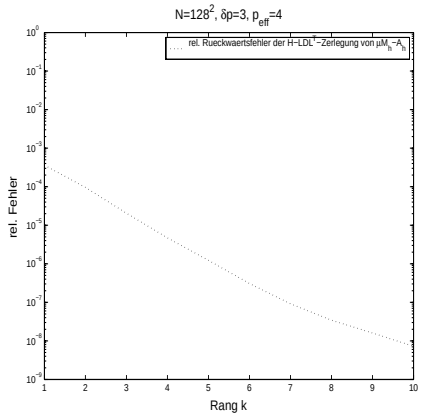

Abbildung 3.18: Relativer Rückwärtsfehler der $\mathcal{H}$-LDL^T-Zerlegung von $\mu M_h - A_h$ in $\|\cdot\|_F$ mit $\mu = 100$ für $N = 128^2$ Unbekannte und $\delta p = 3$.

Numerische Experimente haben gezeigt, dass die Lösung symmetrisch indefiniter Gleichungssysteme mittels $\mathcal{H}$-LDL^T-Zerlegung und je einer Vorwärts- und Rückwärtssubstitution zum einen schneller und zum anderen etwas genauer als die Lösung über die $\mathcal{H}$-Inverse ist (vgl. mit der Lösung symmetrisch positiv definiter LGS mittels $\mathcal{H}$-Cholesky-Zerlegung bzw. $\mathcal{H}$-Invertierung in Abschnitt 3.4.5.4).

3.4.7 $\mathcal{H}$-QR-Zerlegung einiger $\mathcal{H}$-Matrizen

Nachdem all die unternommenen Versuche einer „direkten“ hierarchischen QR-Zerlegung mittels orthogonaler Block-Transformationen am Einbringen der approximierten rekursiven Blockoperationen gescheitert waren, lag es nahe, die bereits entwickelte $\mathcal{H}$-Cholesky-Zerlegung zur Berechnung der $\mathcal{H}$-QR-Zerlegung zu verwenden.
Für den Faktor R der QR-Zerlegung $A = QR$ gilt nämlich $R = L^T$, wobei $A^TA = LL^T$ die Cholesky-Zerlegung von A^TA ist (zum Beweis siehe z.B. [16]).

3.4.7.1 QR-Zerlegung in $\mathcal{H}$-Arithmetik

Sei eine nichtsinguläre $\mathcal{H}$-Matrix $A \in \mathbb{R}^{N\times N}$ gegeben. Gesucht seien $\mathcal{H}$-Approximationen $Q_\mathcal{H}$, $R_\mathcal{H}$ an Q, R, wobei $A = QR$ die exakte QR-Zerlegung von A ist.
Aus dem oben erwähnten Zusammenhang zwischen der QR- und der Cholesky-Zerlegung ergibt sich folgender $\mathcal{H}$-Algorithmus zur Berechnung von $Q_\mathcal{H}$ und $R_\mathcal{H}$:

Algorithmus 3.5 ($\mathcal{H}$-QR-Zerlegung)

1. *Berechne das $\mathcal{H}$-Produkt $B = A^T *_\mathcal{H} A$ sowie die $\mathcal{H}$-Cholesky-Zerlegung $B = LL^T$.*
2. *Setze $R = L^T$.*
3. *Löse $QR = A$ durch $\mathcal{H}$-Matrix-Vorwärtssubstitution nach Q.*

Dieses Verfahren zur Gewinnung der $\mathcal{H}$-QR-Zerlegung benötigt eine $\mathcal{H}$-Matrix-Multiplikation, eine $\mathcal{H}$-Cholesky-Zerlegung sowie eine $\mathcal{H}$-Matrix-Vorwärtssubstitution und besitzt somit die Komplexität $\mathcal{O}(N \log^2 N)$.

Da der Faktor $Q_\mathcal{H}$ indirekt berechnet wurde, ist mit einem **Verlust der Orthogonalität von** $Q_\mathcal{H}$ zu rechnen. Je größer $\kappa(A)$, umso größer wird der $\mathcal{H}$-Fehler der $\mathcal{H}$-Cholesky-Zerlegung von $A^T *_\mathcal{H} A$, und umso mehr kann $Q_\mathcal{H}$ von der Orthogonalität abweichen, d.h. umso größer kann $\|Q_\mathcal{H}^TQ_\mathcal{H} - I\|$ sein. (Man beachte, dass die Kondition der Cholesky-Zerlegung von A^TA gleich $\kappa(A)^2$ ist!)

Reorthogonalisierung von $Q_\mathcal{H}$:
Es ist $A \approx Q_\mathcal{H}R_\mathcal{H}$ und i. Allg. $\kappa(Q_\mathcal{H}) \ll \kappa(A)$.
Mit der mittels Algorithmus 3.5 berechneten $\mathcal{H}$-QR-Zerlegung

$$Q_\mathcal{H} \approx Q'_\mathcal{H}R'_\mathcal{H}$$

von $Q_{\mathcal{H}}$ folgt

$$A \approx Q_{\mathcal{H}} R_{\mathcal{H}} \approx Q'_{\mathcal{H}} \left(R'_{\mathcal{H}} R_{\mathcal{H}} \right)$$

mit den $\mathcal{H}$-Approximationen $Q'_{\mathcal{H}}$ und $R'_{\mathcal{H}} R_{\mathcal{H}}$ an Q und R. Nun gilt i. Allg.

$$\| Q'^T_{\mathcal{H}} Q'_{\mathcal{H}} - I \| \ll \| Q^T_{\mathcal{H}} Q_{\mathcal{H}} - I \|.$$

Man beachte, dass diese Reorthogonalisierung des Faktors $Q_{\mathcal{H}}$ iterativ bis zum Erreichen einer „bestmöglichen“ $\mathcal{H}$-Genauigkeit durchführbar ist.

3.4.7.2 Fehlerschätzung für die approximierten Faktoren $Q_{\mathcal{H}}$ und $R_{\mathcal{H}}$

Nach Bestimmung von $Q_{\mathcal{H}}$ und $R_{\mathcal{H}}$ mittels Algorithmus 3.5 ist $A \approx Q_{\mathcal{H}} R_{\mathcal{H}}$ noch lange kein Garant für ausreichende Orthogonalität von $Q_{\mathcal{H}}$. Beobachtungen in der Praxis zeigen, dass der relative Rückwärtsfehler

$$\frac{\| A - Q_{\mathcal{H}} *_{\mathcal{H}} R_{\mathcal{H}} \|_F}{\| A \|_F}$$

annähernd angibt, wie gut die orthogonale Matrix Q in der vorgegebenen Klasse $\mathcal{M}_{\mathcal{H},k}$ von hierarchischen Matrizen überhaupt approximiert werden kann.
Der Orthogonalitätsverlust von $Q_{\mathcal{H}}$ lässt sich wegen $I + \Delta I = Q^T_{\mathcal{H}} Q_{\mathcal{H}}$ für ein symmetrisches $\Delta I \in \mathbb{R}^{N \times N}$ über den normweisen Rückwärtsfehler abschätzen durch

$$\frac{\| \Delta I \|}{\| I \|} = \frac{\| I - Q^T_{\mathcal{H}} Q_{\mathcal{H}} \|}{\| I \|}. \tag{3.10}$$

Die Approximationsgüte von $R_{\mathcal{H}}$ lässt sich wegen $A^T A = R^T R$ in normweiser Rückwärtsanalyse durch

$$\frac{\| A^T *_{\mathcal{H}} A - R^T_{\mathcal{H}} *_{\mathcal{H}} R_{\mathcal{H}} \|}{\| A^T *_{\mathcal{H}} A \|} \tag{3.11}$$

abschätzen (vgl. Abschätzung (3.5) in Kapitel 3.4.5). Dabei wird - wie immer - die exakte Multiplikation $\cdot$ durch die approximierte $\mathcal{H}$-Multiplikation $*_{\mathcal{H}}$ zum Erhalt der fast linearen Komplexität ersetzt.

Beobachtung in der Praxis:
Die Berechnung von $R = L^T$ mit $A^T A = L L^T$ besitzt die Kondition $\kappa(A^T A) = \kappa(A)^2$ der Cholesky-Zerlegung von $A^T A$, jene von Q aus $QR = A$ die Kondition der Vorwärtssubstitution $\kappa(R) = \kappa(A)$.

Für moderate Konditionszahlen konnten gute $\mathcal{H}$-Approximationen mittels iterativer Reorthogonalisierung erzielt werden (siehe Abbildungen 3.19 und 3.20 (oben)). Für größere Konditionszahlen explodiert die iterative Reorthogonalisierung für kleinen Rang k der Rk-Blöcke, ergibt jedoch gute Approximationen für größere k (siehe Abbildung 3.20 (Mitte und unten)). Für Konditionszahlen nahe 10^8 liegt schließlich für einige $\mathcal{H}$-Matrizen Explosion für jeden Rang k kleiner als die Dimension der Rk-Blöcke vor. Man beachte, dass für solche Konditionszahlen die Matrix $A^T *_{\mathcal{H}} A$, deren Cholesky-Zerlegung berechnet werden soll, bei doppelter Rechengenauigkeit numerisch singulär wird!

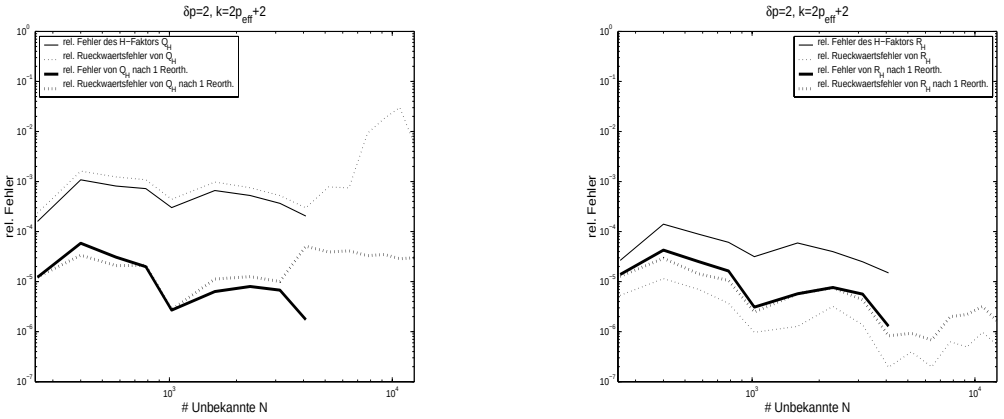

Abbildung 3.19: Relativer Fehler und relativer Rückwärtsfehler der $\mathcal{H}$-Faktoren $Q_{\mathcal{H}}$ (links) und $R_{\mathcal{H}}$ (rechts) der $\mathcal{H}$-QR-Zerlegung von A_h vor und nach einer Reorthogonalisierung in $\|\cdot\|_F$ mit $\delta p = 2$ und Rang $k = 2p_{eff} + 2$.

3.4.7.3 $\mathcal{H}$-Polarzerlegung

Nachdem der Erfolg von Algorithmus 3.5 für die approximative QR-Zerlegung empfindlich von $\kappa(A)^2$ (der Kondition der Cholesky-Zerlegung) abhängt, liegt es nahe, vor der eigentlichen QR-Zerlegung irgendeine HR-Zerlegung $A = HR$ mit $\kappa(H) \ll \kappa(A)$ durchzuführen und dann erst die QR-Zerlegung von H vorzunehmen.

Betrachten wir zunächst den **Spezialfall** symmetrisch positiv definiter $\mathcal{H}$-Matrizen: Für symmetrisch positiv definites A lässt sich sofort die Cholesky-Zerlegung $A = LL^T$ durchführen und anschließend die QR-Zerlegung von L mittels unseres Algorithmus berechnen. Wegen $\kappa(L) = \sqrt{\kappa(A)}$ erhalten wir dadurch eine Reduktion der Kondition von $\kappa(A)^2$ auf die Quadratwurzel $\kappa(A) = \kappa(L)^2$.

Verallgemeinerung:
Man berechne zunächst die $\mathcal{H}$-Polarzerlegung der nichtsingulären Matrix A

$$A = QM$$

mit der eindeutigen symmetrisch positiv definiten Quadratwurzel M von A^TA und einer orthogonalen Matrix Q mittels Highams Methode in $\mathcal{H}$-Arithmetik in $\mathcal{O}(N \log^2 N)$ Operationen:
Die durch

$$\begin{aligned} B_0 &= A, \\ B_{i+1} &= \tfrac{1}{2}\left(\gamma_i B_i + \tfrac{1}{\gamma_i} B_i^{-T}\right), \quad i = 0, 1, \ldots \end{aligned}$$

definierte Folge von Matrizen B_i konvergiert quadratisch gegen Q, wobei γ_i geeignete Skalare zur Konvergenzbeschleunigung sind (siehe [27]). Nach Higham reichen i. Allg. 5 bis 6 Iterationen aus.
Nach erfolgter Berechnung von Q setze man $\tilde{M} = Q^TA$ sowie $M = \frac{1}{2}(\tilde{M} + \tilde{M}^T)$, um die Symmetrie von M zu gewährleisten.

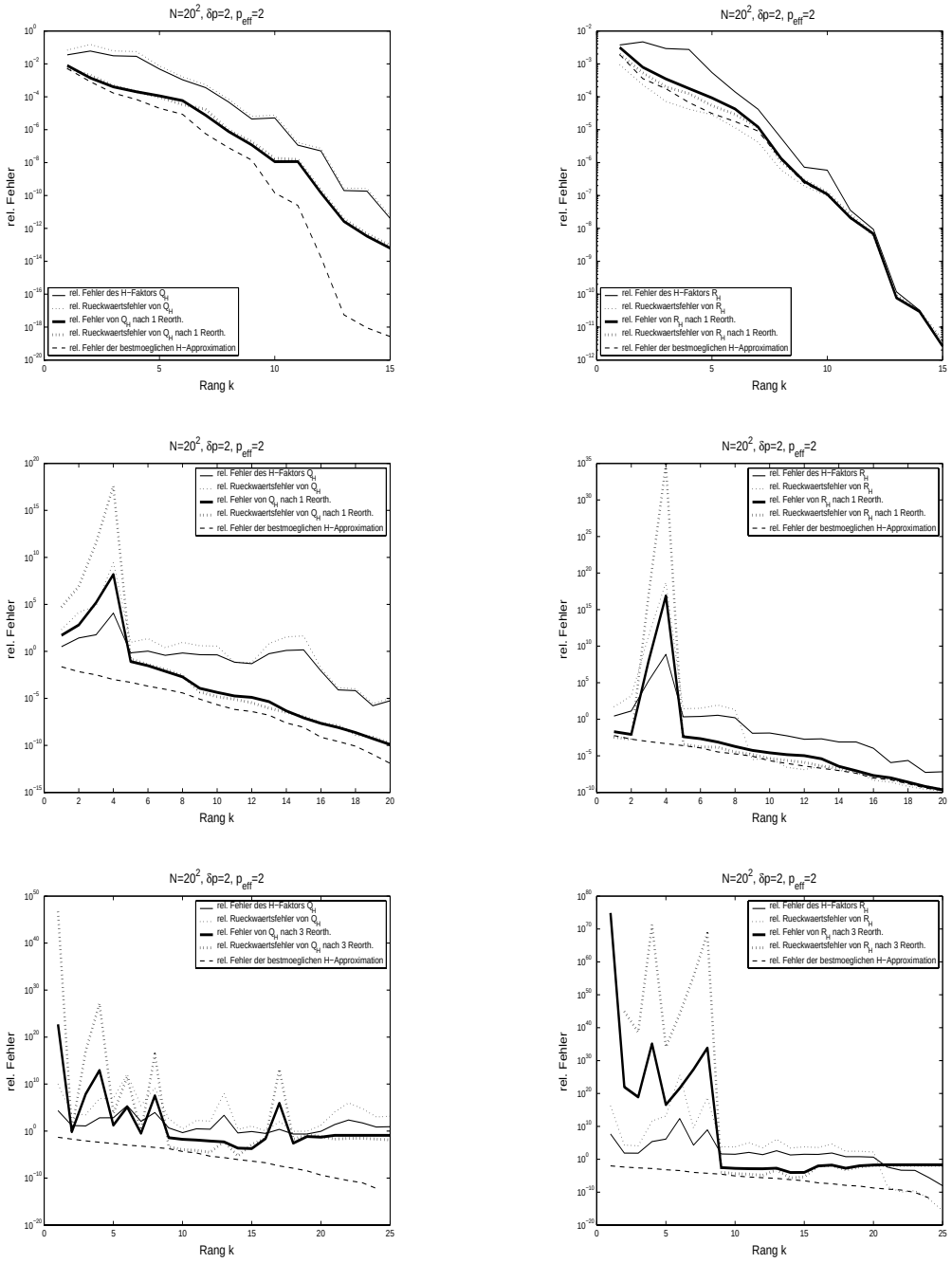

Abbildung 3.20: Relativer Fehler der $\mathcal{H}$-Faktoren $Q_{\mathcal{H}}$ (links) und $R_{\mathcal{H}}$ (rechts) der $\mathcal{H}$-QR-Zerlegung von A_h (oben), A_h^2 (Mitte) und A_h^3 (unten) und relativer Rückwärtsfehler von $Q_{\mathcal{H}}$ (links) und von $R_{\mathcal{H}}$ (rechts) vor und nach einer Reorthogonalisierung sowie relativer Fehler der bestmöglichen $\mathcal{H}$-Approximation an die Faktoren Q (links) und R (rechts) der exakten QR-Zerlegung von A_h (oben), A_h^2 (Mitte) und A_h^3 (unten) in $\|\cdot\|_F$.

Damit ergibt sich folgender Algorithmus:

Algorithmus 3.6 ($\mathcal{H}$-QR-Zerlegung nach $\mathcal{H}$-Polarzerlegung)

1. *Berechne die $\mathcal{H}$-Polarzerlegung und $\mathcal{H}$-Cholesky-Zerlegung $A = QM = QLL^T$.*
2. *Berechne die $\mathcal{H}$-QR-Zerlegung $L = Q'R'$ nach Algorithmus 3.5 (evtl. mit Reorthogonalisierung von Q').*
3. *Bilde die $\mathcal{H}$-Produkte $A = (Q *_{\mathcal{H}} Q')(R' *_{\mathcal{H}} L^T)$.*

Wegen $\kappa(A) = \kappa(M) = \kappa(L)^2$ erhalten wir eine Reduktion der Kondition auf deren Quadratwurzel wie im obigen Spezialfall symmetrisch positiv definiter $\mathcal{H}$-Matrizen.

Dieser Algorithmus besitzt immer noch die Komplexität $\mathcal{O}(N \log^2 N)$, wobei die Konstante in führender Ordnung allerdings von der Anzahl der Iterationen in Highams Algorithmus und von der Anzahl der Reorthogonalisierungen abhängt.

In Abbildung 3.21 sind die Ergebnisse für die $\mathcal{H}$-QR-Zerlegung von A_h^2 mit vorangestellter $\mathcal{H}$-Polarzerlegung dargestellt: Der Vergleich mit Abbildung 3.20 (Mitte) ($\mathcal{H}$-QR-Zerlegung von A_h^2 ohne $\mathcal{H}$-Polarzerlegung) lässt erkennen, dass mit vorangehender $\mathcal{H}$-Polarzerlegung zum einen mittels Algorithmus 3.5 bereits für kleineren Rang k gute Approximationen erzielt werden können und zum anderen eine nachfolgende Reorthogonalisierung fast überflüssig wird.

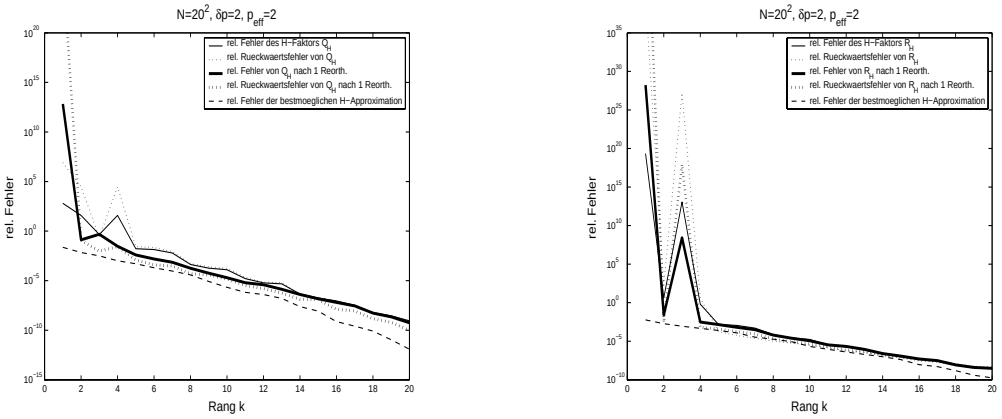

Abbildung 3.21: Relativer Fehler der $\mathcal{H}$-Faktoren $Q_{\mathcal{H}}$ (links) und $R_{\mathcal{H}}$ (rechts) der $\mathcal{H}$-QR-Zerlegung von A_h^2 und relativer Rückwärtsfehler von $Q_{\mathcal{H}}$ (links) und von $R_{\mathcal{H}}$ (rechts) vor und nach einer Reorthogonalisierung, wobei vor der $\mathcal{H}$-QR-Zerlegung eine $\mathcal{H}$-Polarzerlegung durchgeführt wurde, sowie relativer Fehler der bestmöglichen $\mathcal{H}$-Approximation an die Faktoren Q (links) und R (rechts) der exakten QR-Zerlegung von A_h^2 in $\|\cdot\|_F$.

Beobachtung: Eine neuerliche Explosion der Reorthogonalisierung trotz vorangestellter $\mathcal{H}$-Polarzerlegung liegt – wie schon vorhin – an der für die $\mathcal{H}$-Cholesky-Zerlegung in Algorithmus 3.5 zu großen Kondition der Ausgangsmatrix.

3.4.7.4 Zusammenfassung

- In den Kapiteln 3.4.5, 3.4.6 und 3.4.7 wurden die hierarchische Cholesky- und LDL^T-Zerlegung sowie ein Verfahren zur Berechnung einer hierarchischen QR-Zerlegung in $\mathcal{O}(N \log^2 N)$ Operationen entwickelt.
- Die Approximationsgüte der Faktoren der jeweiligen Zerlegungen einer $\mathcal{H}$-Matrix A ist abhängig von deren Kondition $\kappa(A)$.
- Für nicht allzu große Konditionszahlen $\kappa(A)$ ist eine Reorthogonalisierung von $Q_{\mathcal{H}}$ der QR-Zerlegung möglich.
- Die Kondition der $\mathcal{H}$-QR-Zerlegung aus Algortithmus 3.5 kann durch vorangestellte $\mathcal{H}$-Polarzerlegung (entfällt für symmetrisch positiv definite Matrizen) auf deren Quadratwurzel reduziert werden.

Kapitel 4

Transzendente Matrixfunktionen in $\mathcal{H}$-Arithmetik

Gavrilyuk, Hackbusch und Khoromskij haben in [14] gezeigt, wie die Matrixexponentialfunktion $\exp(-tM_h^{-1}A_h)$ mittels hierarchischer Matrizen approximiert werden kann. Diese Matrixfunktion resultiert aus der Projektion der parabolischen Differentialgleichung

$$\dot{u} + Au = f, \;\; u(0) = u_0 \tag{4.1}$$

in einem Gebiet $(0,T) \times \Omega$ mit der Lösung

$$u(t) = \exp(-tA)\, u_0 + \int_0^t \exp(-(t-s)A) f(s) ds \tag{4.2}$$

(siehe [14]) auf Finite Elemente:

$$\dot{u}_h + A_h^{op} u_h = f_h, \;\; u_h(0) = u_{h0}. \tag{4.3}$$

Die Darstellung von (4.3) in der Knotenbasis lautet

$$M_h \dot{\mathbf{u}}_h + A_h \mathbf{u}_h = M_h \mathbf{f}_h, \;\; \mathbf{u}_h(0) = \mathbf{u}_{h0}, \tag{4.4}$$

und dessen zugehöriger Lösungsoperator $\exp(-tM_h^{-1}A_h)$ wird durch das Dunford-Taylor-Integral dargestellt und dann in $\mathcal{H}$-Arithmetik approximiert.

In [15] wurde von denselben Autoren die elliptische Differentialgleichung

$$u_{xx} - Au = -f(x), \;\; u(0) = u_0, u(1) = u_1 \tag{4.5}$$

in einem Gebiet $\Omega \times (0,1)$ mittels hierarchischer Matrizen gelöst. Die Lösung von (4.5) lautet

$$u(x) = E(x;A)\, u_1 + E(1-x;A)\, u_0 + \int_0^1 G(x,s;A) f(s) ds \tag{4.6}$$

mit dem Lösungsoperator

$$E(x) = E(x; A) := \sinh^{-1}(\sqrt{A}) \sinh(x\sqrt{A}) \tag{4.7}$$

der elliptischen Differentialgleichung

$$\frac{d^2E}{dx^2} - AE = 0, \;\; E(0) = 0, E(1) = I \tag{4.8}$$

und der Greenschen Funktion $G(x, s; A)$ des Problems (4.5) (siehe [15]), die durch $E(x; A)$ darstellbar ist. Der operatorwertige normalisierte Sinus hyperbolicus $E(x; A)$ des elliptischen Operators A wird nach Darstellung durch das Dunford-Taylor-Integral wiederum durch $\mathcal{H}$-Matrizen approximiert.

Gavrilyuk und Makarov haben in [10] und [13] explizite Lösungsformeln für die homogene Wellengleichung

$$\ddot{u} + Au = 0, \;\; u(0) = u_0, \dot{u}(0) = \dot{u}_0 \tag{4.9}$$

– womit wir schließlich bei den hyperbolischen Differentialgleichungen angelangt wären – in geschlossener Form angegeben durch Darstellung des Cosinusoperators $\cos(t\sqrt{A})$ und des Sinusoperators $\sin(t\sqrt{A})$ durch die Cayley-Transformation:

$$\begin{aligned} u(t) &= \cos(t\sqrt{A})\, u_0 + A^{-\frac{1}{2}} \sin(t\sqrt{A})\, \dot{u}_0 \\ &= e^{-\delta t} \sum_{n=0}^{\infty} (L_n(t) - L_{n-1}(t)) \, (u_n - \dot{u}_n) \end{aligned} \tag{4.10}$$

mit $\delta < 1$, den Laguerre-Polynomen $L_n(t)$ und Vektoren u_n und $\dot{u}_n$, die mittels geeigneter Drei-Term-Rekursionen ohne explizite Kenntnis von $\sqrt{A}$ berechenbar sind.
Dann ist die exakte Lösung (4.10) von (4.9) durch die abgeschnittene Summe

$$u^N(t) = e^{-\delta t} \sum_{n=0}^{N} (L_n(t) - L_{n-1}(t)) \, (u_n - \dot{u}_n) \tag{4.11}$$

approximierbar, wobei $u^N(t)$ nur mit polynomialer Konvergenzgeschwindigkeit gegen $u(t)$ strebt und die Approximationsgüte zudem von der Regularität der Anfangsdaten abhängt.
Nachdem nun die abgeschnittenen Summen aus (4.11) gerade die (N, N)-Padé-Approximationen der entsprechenden oszillatorischen Matrixfunktionen darstellen (siehe [2, Theorem 2.1]), ist diese Methode für große Wellengeschwindigkeiten c und nur stetige lineare FE völlig ungeeignet.

Unser Ziel ist nun die Lösung der hyperbolischen Differentialgleichung

$$\ddot{u} + c^2 Au = f, \;\; u(0) = u_0, \dot{u}(0) = \dot{u}_0 \tag{4.12}$$

im Gebiet $\Omega \times (0, T)$ mit $\Omega = (0, 1)^2$ und $c \gg 1$.
Wir haben in Abschnitt 2.1 die kontinuierliche Wellengleichung (4.12) im Ort

auf Finite Elemente projiziert und in Abschnitt 2.5 die resultierende ODE in der Zeit mittels des exponentiellen Gautschi-Verfahrens diskretisiert.
Um die dadurch erhaltene Drei-Term-Rekursion in der Knotendarstellung zu lösen, werden die transzendenten Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})$, $\psi(\tau^2 c^2 M_h^{-1}A_h)$ und $\sigma(\tau^2 c^2 M_h^{-1}A_h)$ benötigt.

Die Berechnung von Matrixfunktion-Vektor-Produkten mit den obigen Matrixfunktionen durch Krylovraummethoden unterliegt der Beschränkung von τ durch h und c^{-1} (siehe 3.1). Daher lautet unser Ziel im Folgenden, die Matrixfunktionen selbst mittels $\mathcal{H}$-Matrizen zu approximieren.

Wir werden zunächst in Abschnitt 4.1 eine Duplikationsstrategie verfolgen, nämlich die Berechnung der Matrixfunktion eines gegebenen Arguments jeweils rekursiv aus derselben Matrixfunktion des halben Arguments unter Anwendung der Additionstheoreme für trigonometrische Funktionen. Dabei werden wir aufzeigen, dass die Duplikationsstrategie schlecht konditioniert ist und darüberhinaus feststellen, dass bestimmte oszillatorische Matrixfunktionen gar nicht durch $\mathcal{H}$-Matrizen approximierbar sind.

In Abschnitt 4.2 werden wir dafür einen heuristischen Beweis für die 1D Matrix-Sinusfunktion und -Cosinusfunktion geben, der sich auf den Fall der entsprechenden 2D Matrixfunktionen verallgemeinern lässt, indem wir die Kondition sämtlicher Matrixblöcke nach oben beschränken. Daraus lässt sich nämlich sofort auf die Nichtapproximierbarkeit dieser Matrixblöcke durch Niedrigrang-Matrizen schließen.
Die Nichtapproximierbarkeit durch $\mathcal{H}$-Matrizen wird schließlich noch durch einen Blick auf die zu den oszillatorischen Matrixfunktionen gehörigen Integralkerne verdeutlicht. Diese zeichnen nämlich für die $\mathcal{H}$-Approximierbarkeit der Inversen der Steifigkeitsmatrix verantwortlich (siehe [20, Seite 6])!

In Abschnitt 4.3 werden wir eine neue Methode entwickeln, um Eigenpaare der Matrix $M_h^{-1}A_h$ mit fast linearem Aufwand zu berechnen. Diese spielt sowohl in der Konstruktion von Niedrigrang-Approximationen an bestimmte Matrixfunktionen im darauffolgenden Abschnitt als auch bei der Lösung der 2D Wellengleichung durch Spektralzerlegung von $M_h^{-1}A_h$ in Kapitel 5 eine zentrale Rolle. Analog zu den Eigenwerten und Eigenvektoren lassen sich auch die Singulärwerte und Singulärvektoren von $M_h^{-1}A_h$ mit fast linearem Aufwand berechnen.

In Abschnitt 4.4 werden wir dann unser Ziel, die Approximation transzendenter Matrixfunktionen unter Verwendung von $\mathcal{H}$-Matrizen, für eine bestimmte Klasse von Matrixfunktionen erreichen: Alle Matrizen mit genügend raschem Abklingverhalten der Singulärwerte lassen sich durch Niedrigrang-Matrizen approximieren. Des Weiteren können hervorragende Fehlerschätzer für diese Niedrigrang-Approximationen angegeben werden.

Zum Abschluss des Kapitels wird in Abschnitt 4.5 die Exponentialfunktion des Laplace-Operators durch Niedrigrang-Matrizen genähert. Dabei stellt sich heraus, dass für genügend große Zeiten bereits eine Rang-1-Matrix (!) für hohe

Approximationsgenauigkeit ausreicht.

4.1 Berechnung von Matrixfunktionen mittels Duplikation der Argumente

4.1.1 Duplikation in $\mathcal{H}$-Arithmetik

Wir verwenden im Folgenden den Potenzreihenkalkül

$$f(A) = \sum_{k=0}^{\infty} \alpha_k A^k$$

zur Berechnung einer Matrixfunktion $f(A)$, wobei f eine ganze Funktion sei. Wir betrachten nun exemplarisch die Matrixfunktion $\sigma(\tau^2 c^2 M_h^{-1} A_h)$.
Die ganze Funktion

$$\sigma(x^2) = \left(\frac{\sin \frac{1}{2}x}{\frac{1}{2}x} \right)^2 = 2\,\frac{1-\cos x}{x^2}$$

besitzt die aus der Cosinusreihe abgeleitete Taylorentwicklung

$$\sigma(x^2) = \sum_{k=0}^{\infty} (-1)^k \frac{2}{(2k+2)!} x^{2k} = 1 - \frac{2}{4!}x^2 + \frac{2}{6!}x^4 - \frac{2}{8!}x^6 + \ldots$$

mit dem Abbrechfehler

$$\left| \sigma(x^2) - \sum_{k=0}^{n} (-1)^k \frac{2}{(2k+2)!} x^{2k} \right| \;\leq\; \frac{2}{(2n+4)!}\, x^{2n+2}$$

für $\left(\frac{x^{2k}}{(2k+2)!} \right)_{k \in \mathbb{N}_0}$ monoton fallend nach dem Leibnizschen Konvergenzkriterium für alternierende Reihen.

Das Argument $\tau^2 c^2 M_h^{-1} A_h$ besitzt die Spektralnorm

$$\|\tau^2 c^2 M_h^{-1} A_h\|_2 = \tau^2 c^2 \mathcal{O}(h^{-2}) = \tau^2 c^2 \mathcal{O}(N)$$

(siehe Abschnitt 3.1), weshalb zur Berechnung von $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ mittels obiger Taylorreihe für große N und c viel zu viele Glieder notwendig wären.

Daher wählen wir die folgende Vorgehensweise zur Berechnung von $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ in drei Schritten:

Algorithmus 4.1 (Duplikation in $\mathcal{H}$-Arithmetik)

1. *Skalierung des Arguments* $\tau^2 c^2 M_h^{-1} A_h$ *mit* 4^{-l}, $l \in \mathbb{N}_0$, *so dass* $\|4^{-l} \tau^2 c^2 M_h^{-1} A_h\|_2 \leq \frac{1}{2}$. *Daraus folgt*

$$\begin{aligned} 4^l &\sim \tau^2 c^2 N \qquad &\Longleftrightarrow \\ l &\sim \log(\tau^2 c^2 N) \qquad &\Longleftrightarrow \\ l &\sim \log(\tau^2 c^2) + \log N \end{aligned}$$

2. Kurze Taylorentwicklung von $\sigma(4^{-l}\tau^2c^2M_h^{-1}A_h)$ in $\mathcal{H}$-Arithmetik.

3. l Duplikationsschritte

$$\sigma(4^{-l}\tau^2c^2M_h^{-1}A_h) \longrightarrow \sigma(4^{-l+1}\tau^2c^2M_h^{-1}A_h) \longrightarrow \\ \longrightarrow \cdots \longrightarrow \sigma(\tau^2c^2M_h^{-1}A_h)$$

in $\mathcal{H}$-Arithmetik.

Der Duplikationsalgorithmus für $\sigma(x^2)$ lautet

$$\sigma(4x^2) = \sigma(x^2)\left(1 - \frac{1}{4}\,x^2\sigma(x^2)\right). \tag{4.13}$$

Da er für große Argumente x^2 instabil wird, verwenden wir stattdessen das äquivalente stabile System

$$\begin{aligned} \sigma(4x^2) &= \sigma(x^2)\left(1 - \sin^2(\tfrac{1}{2}x)\right) \\ \sin^2(2\tfrac{1}{2}x) &= 4\,\sin^2(\tfrac{1}{2}x)\left(1 - \sin^2(\tfrac{1}{2}x)\right) \end{aligned} \tag{4.14}$$

mit $\sin^2(\frac{1}{2}x) = \frac{1}{4}\,x^2\sigma(x^2)$ als Startwert neben $\sigma(x^2)$.

Man beachte, dass beide Varianten (4.13) und (4.14) je eine $\mathcal{H}$-Matrix-Skalierung, eine $\mathcal{H}$-Matrix-Subtraktion von der Einheitsmatrix und zwei $\mathcal{H}$-Matrix-Multiplikationen benötigen und somit gleich aufwendig sind. Die Komplexität eines Duplikationsschrittes beläuft sich also auf $\mathcal{O}(N\log^2 N)$, jene zur Berechnung von $\sigma(\tau^2c^2M_h^{-1}A_h)$ wegen $l \sim \log N + \log(\tau^2c^2)$ insgesamt auf $\mathcal{O}(N\log^2 N(\log N + \log(\tau^2c^2)))$.

Schätzung der Spektralnorm von $M_h^{-1}A_h$:
Um l mit $\|4^{-l}\tau^2c^2M_h^{-1}A_h\|_2 \leq \frac{1}{2}$ zu bestimmen, muss eine Schätzung von $\|M_h^{-1}A_h\|_2$ vorliegen. Diese kann man sich wie folgt verschaffen:
Es ist $\|M_h^{-1}A_h\|_2 = \mathcal{O}(N) = \mathcal{O}(n^2)$ und $\|M_h^{-1}A_h\|_F = \mathcal{O}(n^3)$. Durch Beobachtung des Konvergenzverhaltens der Differenz $\frac{\|\cdot\|_2}{n^2} - 2\frac{\|\cdot\|_F}{n^3}$ lässt sich $\|M_h^{-1}A_h\|_2$ durch $\|M_h^{-1}A_h\|_F$ mit fast linearem Aufwand ($\mathcal{O}(N\log N)$ flops für die Berechnung von $\|M_h^{-1}A_h\|_F$) bis auf einen relativen Fehler kleiner als 1% abschätzen, und das reicht für die Bestimmung von l allemal.

4.1.2 Fehleranalyse des Duplikationsalgorithmus

Sei $C_{\mathcal{H}} := M_h^{-1_{\mathcal{H}}} *_{\mathcal{H}} A_h$ die $\mathcal{H}$-Approximation an die exakte Matrix $C := M_h^{-1}A_h$ und $\|\Delta C\|$ eine Schätzung für den absoluten Fehler (zur approximierten Arithmetik samt Fehlerschätzung in der Frobeniusnorm siehe Abschnitt 3.4). Dieser Fehler setzt sich zusammen aus dem transportierten Fehler der $\mathcal{H}$-Invertierung $M_h^{-1_{\mathcal{H}}}$ und dem $\mathcal{H}$-Multiplikationsfehler von $M_h^{-1_{\mathcal{H}}} *_{\mathcal{H}} A_h$.

Berechnung der Startmatrizen für die Duplikation:

Nach erfolgter Skalierung von $\tau^2c^2C_{\mathcal{H}}$ auf $B_{\mathcal{H}} := 4^{-l}\tau^2c^2C_{\mathcal{H}}$ mit $\|B_{\mathcal{H}}\|_2 \leq \frac{1}{2}$ wird das n-te Taylorpolynom von $\sigma(B_{\mathcal{H}})$ mittels Hornerschema in $\mathcal{H}$-Arithmetik

berechnet. Dabei reichen i. Allg. $n = 3$ oder 4 Glieder.
Für diese $\mathcal{H}$-Approximation

$$\sigma_{\mathcal{H}}(B_{\mathcal{H}}) = \sigma_{\mathcal{H}}^{(n)}(B_{\mathcal{H}}) = c_0 I + c_1 B_{\mathcal{H}} *_{\mathcal{H}} (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots *_{\mathcal{H}} (I + c_n B_{\mathcal{H}}) \cdots))$$

an $\sigma(B)$ gilt dann

$$\begin{aligned} & \|\Delta\sigma(B)\| \\ = \; & \|\sigma_{\mathcal{H}}(B_{\mathcal{H}}) - \sigma(B)\| \\ = \; & \|\sigma_{\mathcal{H}}^{(n)}(B_{\mathcal{H}}) - \sigma^{(n)}(B_{\mathcal{H}}) + \sigma^{(n)}(B_{\mathcal{H}}) - \sigma^{(n)}(B) + \sigma^{(n)}(B) - \sigma(B)\| \\ \leq \; & \|\sigma_{\mathcal{H}}^{(n)}(B_{\mathcal{H}}) - \sigma^{(n)}(B_{\mathcal{H}})\| + \|\sigma^{(n)}(B_{\mathcal{H}}) - \sigma^{(n)}(B)\| + \|\sigma^{(n)}(B) - \sigma(B)\|. \end{aligned}$$

Dabei bezeichnet der **erste Term** in obiger Ungleichung den akkumulierten $\mathcal{H}$-Fehler des Taylorpolynoms bei exaktem Input $B_{\mathcal{H}}$, der **zweite Term** den verstärkten Eingabefehler $\|\Delta B\|$ und der **dritte Term** den Abbrechfehler der Taylorreihe von σ.
Für den ersten Term gilt

$$\begin{aligned} & \|\sigma_{\mathcal{H}}^{(n)}(B_{\mathcal{H}}) - \sigma^{(n)}(B_{\mathcal{H}})\| \\ = \; & |c_1| \cdot \|B_{\mathcal{H}} *_{\mathcal{H}} (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots)) - B_{\mathcal{H}} \cdot (I + c_2 B_{\mathcal{H}} \cdot (\cdots))\| \\ = \; & |c_1| \cdot \|B_{\mathcal{H}} *_{\mathcal{H}} (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots)) - B_{\mathcal{H}} \cdot (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots)) + \\ & \qquad B_{\mathcal{H}} \cdot (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots)) - B_{\mathcal{H}} \cdot (I + c_2 B_{\mathcal{H}} \cdot (\cdots))\| \\ \leq \; & |c_1| \cdot \|B_{\mathcal{H}} *_{\mathcal{H}} (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots)) - B_{\mathcal{H}} \cdot (I + c_2 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots))\| + \\ & |c_1| \cdot \|B_{\mathcal{H}}\| \cdot |c_2| \cdot \|B_{\mathcal{H}} *_{\mathcal{H}} (I + c_3 B_{\mathcal{H}} *_{\mathcal{H}} (\cdots)) - B_{\mathcal{H}} \cdot (I + c_3 B_{\mathcal{H}} \cdot (\cdots))\|. \end{aligned}$$

Der $\mathcal{H}$-Fehler des n-ten $\mathcal{H}$-Taylorpolynoms lässt sich also rekursiv auf die reinen $\mathcal{H}$-Fehler der $n-1$ $\mathcal{H}$-Matrix-Multiplikationen im Hornerschema zurückspielen.
Für den transportierten Eingabefehler gilt unterdessen

$$\begin{aligned} \|\sigma^{(n)}(B_{\mathcal{H}}) - \sigma^{(n)}(B)\| & \dot{=} \kappa_{\sigma^{(n)}}(B) \cdot \|B_{\mathcal{H}} - B\| \\ & \approx |c_1| \cdot \|\Delta B\| \end{aligned}$$

wegen

$$\begin{aligned} \kappa_{\sigma^{(n)}}(B) & = \|\sigma^{(n)\prime}(B)\| = \|\textstyle\sum_{k=1}^{n} k c_k B^{k-1}\| \\ & \leq \textstyle\sum_{k=1}^{n} k|c_k|\|B\|^{k-1} \approx |c_1| \end{aligned}$$

mit $|c_1| = \frac{1}{12}$.

Die $\mathcal{H}$-Matrix

$$\sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}) = \frac{1}{4} B_{\mathcal{H}} *_{\mathcal{H}} \sigma_{\mathcal{H}}(B_{\mathcal{H}})$$

erhält man nun durch eine $\mathcal{H}$-Matrix-Multiplikation mit anschließender Skalierung. Der Fehler

$$\|\Delta \sin^2(\frac{1}{2}\sqrt{B})\| = \|\sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}) - \sin^2(\frac{1}{2}\sqrt{B})\|$$

lässt sich analog zu vorhin aus dem $\mathcal{H}$-Multiplikationsfehler und den transportierten Eingabefehlern ableiten.

Somit liegen uns die $\mathcal{H}$-Matrizen $\sigma_{\mathcal{H}}(B_{\mathcal{H}})$ und $\sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}})$ samt der Schätzungen für die Fehler $\Delta\sigma(B)$ und $\Delta\sin^2(\frac{1}{2}\sqrt{B})$ in der Frobeniusnorm vor, und wir können uns nun der Betrachtung eines Duplikationsschrittes zuwenden.

Untersuchung eines Duplikationsschrittes:

Wir suchen den Gesamtfehler $\|\sigma_{\mathcal{H}}(4B_{\mathcal{H}}) - \sigma(4B)\|$ nach einer Duplikation. Mit

$$\begin{array}{rcl} \begin{pmatrix} \sigma(4x^2) \\ \sin^2(2\frac{1}{2}x) \end{pmatrix} & = & \begin{pmatrix} \sigma(x^2)\left(1-\sin^2(\frac{1}{2}x)\right) \\ 4\,\sin^2(\frac{1}{2}x)\left(1-\sin^2(\frac{1}{2}x)\right) \end{pmatrix} \\ & =: & f\left(\sigma(x^2), \sin^2(\frac{1}{2}x)\right) = f\left(\sigma, \sin^2\right) \end{array} \tag{4.15}$$

ist

$$\begin{array}{l} \left\| \begin{pmatrix} \sigma_{\mathcal{H}}(4\,B_{\mathcal{H}}) - \sigma(4\,B) \\ \sin^2_{\mathcal{H}}(2\frac{1}{2}\sqrt{B_{\mathcal{H}}}) - \sin^2(2\frac{1}{2}\sqrt{B}) \end{pmatrix} \right\| = \\ \left\| f_{\mathcal{H}}\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f\left(\sigma, \sin^2\right) \right\| \leq \\ \left\| f_{\mathcal{H}}\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) \right\| + \left\| f\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f\left(\sigma, \sin^2\right) \right\|. \end{array} \tag{4.16}$$

Der **erste Fehleranteil** entspricht dem $\mathcal{H}$-Fehler bei exaktem Input, das ist der durch die beiden approximierten $\mathcal{H}$-Multiplikationen einer Duplikation verursachte Fehler. Der **zweite Fehleranteil** ist der durch die Kondition der Funktion f verstärkte Eingabefehler.

Mit

$$Df\left(\sigma(x^2), \sin^2(\frac{1}{2}x)\right) = \begin{pmatrix} \cos^2(\frac{1}{2}x) & -\sigma(x^2) \\ 0 & 4\left(1-2\,\sin^2(\frac{1}{2}x)\right) \end{pmatrix}$$

ergeben sich somit die transportierten Eingabefehler in linearer Näherung zu

$$\begin{array}{l} \|f_1\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f_1\left(\sigma, \sin^2\right)\| \leq \\ \|\cos^2(\frac{1}{2}\sqrt{B})\|\,\|\sigma_{\mathcal{H}}(B_{\mathcal{H}}) - \sigma(B)\| + \\ \|\sigma(B)\|\,\|\sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}) - \sin^2(\frac{1}{2}\sqrt{B})\| \end{array} \tag{4.17}$$

und

$$\begin{array}{l} \|f_2\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f_2\left(\sigma, \sin^2\right)\| \leq \\ 4\,\|\cos(\sqrt{B})\|\,\|\sin^2_{\mathcal{H}}(2\frac{1}{2}\sqrt{B_{\mathcal{H}}}) - \sin^2(2\frac{1}{2}\sqrt{B})\|. \end{array} \tag{4.18}$$

Für die $\mathcal{H}$-Fehler bei exaktem Input $\sigma_{\mathcal{H}}(B_{\mathcal{H}})$ und $\sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}})$ erhalten wir schließlich

$$\begin{array}{l} \|f_{1\mathcal{H}}\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f_1\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right)\| = \\ \|\sigma_{\mathcal{H}}(B_{\mathcal{H}}) *_{\mathcal{H}} (I - \sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}})) - \\ \quad \sigma_{\mathcal{H}}(B_{\mathcal{H}}) \cdot (I - \sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}))\| \end{array} \tag{4.19}$$

und

$$\begin{array}{l} \|f_{2\mathcal{H}}\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right) - f_2\left(\sigma_{\mathcal{H}}, \sin^2_{\mathcal{H}}\right)\| = \\ 4 \cdot \|\sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}) *_{\mathcal{H}} (I - \sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}})) - \\ \quad \sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}) \cdot (I - \sin^2_{\mathcal{H}}(\frac{1}{2}\sqrt{B_{\mathcal{H}}}))\|, \end{array} \tag{4.20}$$

wobei die rechts stehenden Differenzen den jeweiligen $\mathcal{H}$-Multiplikationsfehlern bei exaktem Input entsprechen.

Fazit: Von einem Duplikationsschritt zum nächsten ist wegen (4.18) und (4.20) mit einer Vervierfachung des jeweils resultierenden absoluten Fehlers $\|\sigma_{\mathcal{H}}(4^j B_{\mathcal{H}}) - \sigma(4^j B)\|$, $1 \leq j \leq l$, zu rechnen.
Nachdem die Norm $\|\sigma_{\mathcal{H}}(4^j B_{\mathcal{H}})\|$ nicht mit einem Faktor 4 mitwächst sondern beschränkt bleibt, ist dieselbe Entwicklung auch für den relativen Fehler $\frac{\|\sigma_{\mathcal{H}}(4^j B_{\mathcal{H}}) - \sigma(4^j B)\|}{\|\sigma(4^j B)\|}$, $1 \leq j \leq l$, zu erwarten.

Beobachtung in der Praxis: Die durchgeführten numerischen Berechnungen bestätigen die angestellte Vermutung vollauf; es ist teilweise sogar ein deutlich größerer Fehlerzuwachs zu beobachten (siehe Abbildung 4.1). Die Duplikationsmethoden für die Matrixfunktionen $\cos_{\mathcal{H}}(\sqrt{B_{\mathcal{H}}})$ und $\psi_{\mathcal{H}}(B_{\mathcal{H}})$ besitzen ein völlig analoges Fehlerverstärkungsverhalten.

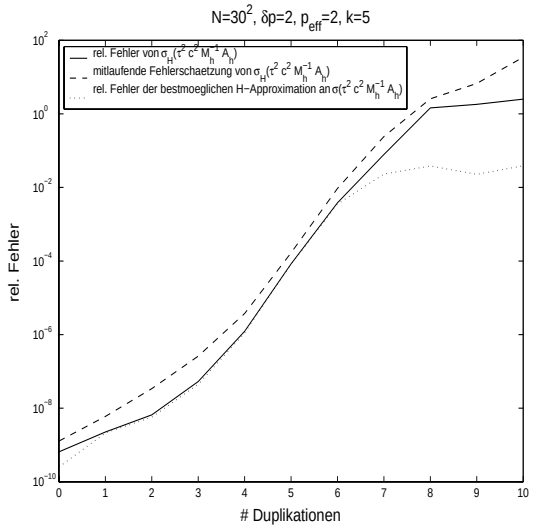

Abbildung 4.1: Relativer Fehler und mitlaufende Fehlerschätzung der duplizierten $\mathcal{H}$-Matrix $\sigma_{\mathcal{H}}(\tau^2 c^2 M_h^{-1} A_h)$ sowie relativer Fehler der bestmöglichen $\mathcal{H}$-Approximation an die exakte Matrix $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ in $\|\cdot\|_F$.

Dabei dominiert bei den ersten Duplikationen der Fehlertransport gegenüber den in den $\mathcal{H}$-Multiplikationen neu entstehenden $\mathcal{H}$-Fehlern. Ab $\|\tau^2 c^2 M_h^{-1} A_h\|_2 \gtrapprox \pi^2$ wird dann der $\mathcal{H}$-Fehler der $\mathcal{H}$-Multiplikationen größer als der mit einem Faktor 4 verstärkte transportierte Fehler.

Dies lässt die Frage aufkommen, ob die transzendenten Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1} A_h})$ bzw. $\sin(\tau c\sqrt{M_h^{-1} A_h})$, $\psi(\tau^2 c^2 M_h^{-1} A_h)$ und $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ überhaupt durch $\mathcal{H}$-Matrizen approximierbar sind!

Um dies nachzuprüfen, berechnen wir zunächst die exakten Matrizen $\sigma(\tau^2 c^2 M_h^{-1} A_h)$, $\psi(\tau^2 c^2 M_h^{-1} A_h)$ und $\cos(\tau c\sqrt{M_h^{-1} A_h})$ mit A_h und M_h in $\mathcal{H}$-Gestalt und dann die jeweiligen Bestapproximationen $\sigma_{\mathcal{H}}^{best}(\tau^2 c^2 M_h^{-1} A_h)$, $\psi_{\mathcal{H}}^{best}(\tau^2 c^2 M_h^{-1} A_h)$ und $\cos_{\mathcal{H}}^{best}(\tau c\sqrt{M_h^{-1} A_h})$ in der Menge aller $\mathcal{H}$-Matrizen, indem wir all jene Teilblöcke der exakten Matrixfunktionen, die in einer $\mathcal{H}$-Matrix Rk-Blöcken entsprechen, auf den geforderten Rang kürzen.

In den Abbildungen 4.2 und 4.3 ist der relative Fehler dieser Bestapproximationen für $N = 32^2$ und $\tau c = 10^{-1}, 10^0$ und 10^1 über dem Rang der Rk-Blöcke aufgetragen.

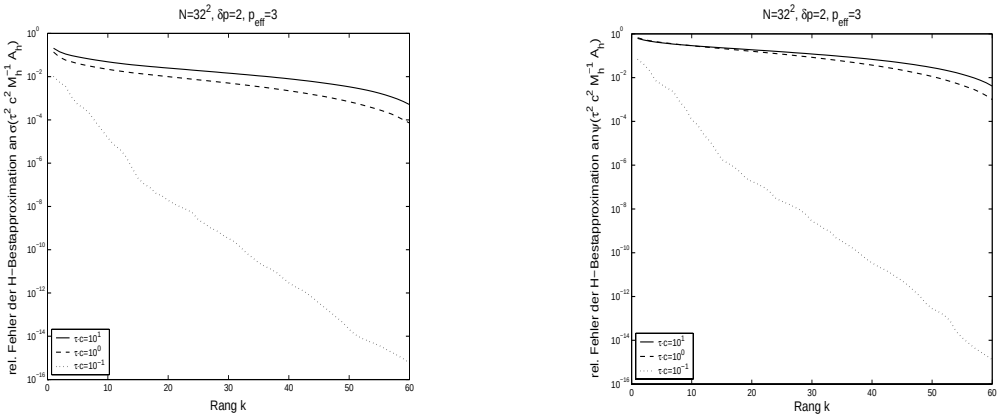

Abbildung 4.2: Relativer Fehler der bestmöglichen $\mathcal{H}$-Approximation an die exakte Matrixfunktion $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ (links) und $\psi(\tau^2 c^2 M_h^{-1} A_h)$ (rechts) in $\|\cdot\|_F$ für $\tau c = 10^{-1}, 10^0$ und 10^1.

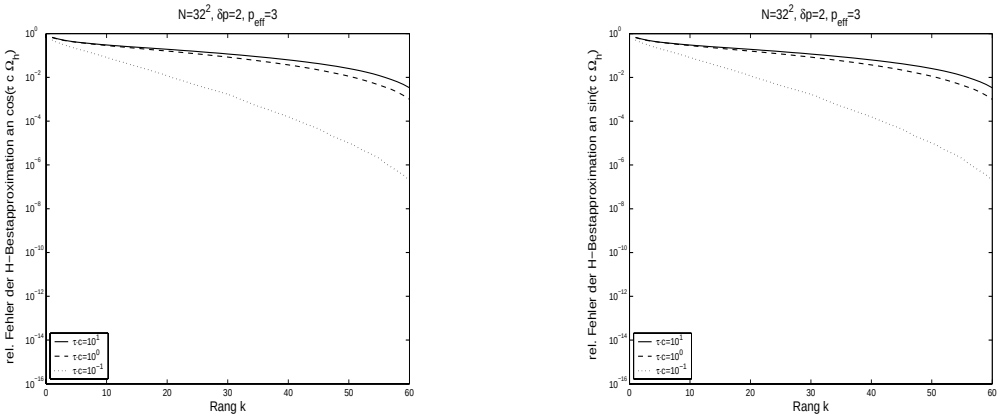

Abbildung 4.3: Relativer Fehler der bestmöglichen $\mathcal{H}$-Approximation an die exakte Matrixfunktion $\cos(\tau c\,\Omega_h)$ (links) und $\sin(\tau c\,\Omega_h)$ (rechts) in $\|\cdot\|_F$ für $\tau c = 10^{-1}, 10^0$ und 10^1, wobei $\Omega_h = \sqrt{M_h^{-1} A_h}$ ist.

Die Ergebnisse in den Abbildungen 4.2 und 4.3 deuten darauf hin, dass die

Matrixfunktionen $\sigma(\tau^2 c^2 M_h^{-1} A_h)$, $\psi(\tau^2 c^2 M_h^{-1} A_h)$ und $\cos(\tau c\sqrt{M_h^{-1} A_h})$ bzw. $\sin(\tau c\sqrt{M_h^{-1} A_h})$ für $\|\tau^2 c^2 M_h^{-1} A_h\|_2 \gtrapprox \pi^2$ nicht mehr durch $\mathcal{H}$-Matrizen darstellbar sind.

4.2 $\mathcal{H}$-Darstellbarkeit von Matrixfunktionen

Unser Ziel ist es nun zu zeigen, dass die Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1} A_h})$ bzw. $\sin(\tau c\sqrt{M_h^{-1} A_h})$ nicht mittels $\mathcal{H}$-Matrizen approximiert werden können. Auf die Matrixfunktionen $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ und $\psi(\tau^2 c^2 M_h^{-1} A_h)$ werden wir dann später separat eingehen.

4.2.1 Der 1D Fall

Wir betrachten zunächst der Einfachheit halber den 1D Fall. Dazu unterteilen wir das Einheitsintervall $\Omega = (0,1)$ in $n+1$ äquidistante Teilintervalle (x_i, x_{i+1}), $x_i = ih$, $0 \leq i \leq n$, $h = \frac{1}{n+1}$, und verwenden stückweise lineare C^0-Funktionen auf Ω, welche wir in der üblichen Knotenbasis $\{\psi_i^h\}_{i=1}^n$, $\psi_i^h(x_j) = \delta_{ij}$, $1 \leq i, j \leq n$, darstellen.

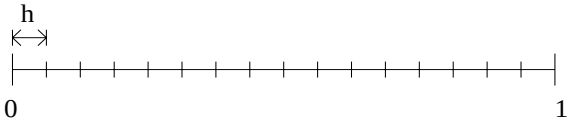

Abbildung 4.4: Äquidistante Unterteilung von $\Omega = (0,1)$

Die 1D $\mathcal{H}$-Matrizen bzgl. obiger Diskretisierung haben die rekursive 2×2-Blockgestalt $\square = \begin{array}{|c|c|}\hline \square & \rightarrow \\ \hline \leftarrow & \square \\ \hline \end{array}$ mit $\rightarrow$ und $\leftarrow$ wie in Abschnitt 3.3.1 (siehe [19]).

4.2.1.1 Eigenwerte und Eigenvektoren des 1D Laplace-Operators

Die zum 1D Laplace-Operator gehörige Steifigkeitsmatrix A_h bzgl. der nodalen Basisfunktionen ψ_i^h, $1 \leq i \leq n$, hat die Form

$$A_h = \frac{1}{h} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix},$$

die zugehörige Massenmatrix M_h ergibt sich zu

$$M_h = \frac{h}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{pmatrix}.$$

Man beachte, dass die Darstellungen von A_h und M_h als $\mathcal{H}$-Matrizen exakt sind.

Die Eigenwerte (EW) von A_h lauten bekanntermaßen

$$\lambda_k = \frac{2}{h}\left(1 - \cos\left(\frac{k\pi}{n+1}\right)\right), \;\; 1 \le k \le n, \tag{4.21}$$

die zugehörigen normierten Eigenvektoren (EV)

$$\mathbf{v}_k = \left(\sqrt{\frac{2}{n+1}}\sin\left(\frac{ik\pi}{n+1}\right)\right)_{1\le i\le n} = \left(\sqrt{\frac{2}{n+1}}\sin\left(i\alpha_k\right)\right)_{1\le i\le n} \tag{4.22}$$

mit $\alpha_k = \frac{k\pi}{n+1}$.

Damit lässt sich unter Ausnutzung der Beziehung $M_h - hI = -\frac{1}{6}h^2 A_h$ sofort das verallgemeinerte Eigenwertproblem

$$A_h\tilde{\mathbf{v}}_k = \tilde{\lambda}_k M_h\tilde{\mathbf{v}}_k, \;\; 1 \le k \le n,$$

mit

$$\tilde{\lambda}_k = \frac{1}{h^2}\frac{2(1-\cos\alpha_k)}{1-\frac{1}{3}(1-\cos\alpha_k)} \tag{4.23}$$

und

$$\tilde{\mathbf{v}}_k = \mathbf{v}_k \tag{4.24}$$

lösen. $(\tilde{\lambda}_k, \tilde{\mathbf{v}}_k)$, $1 \le k \le n$, bilden also die Eigenpaare der Matrix $M_h^{-1}A_h$.

Mit $\tilde{\mathbf{v}}_k$ Rechts-EV (REV) von $M_h^{-1}A_h$ zum EW $\tilde{\lambda}_k$ ergibt sich sofort $\tilde{\mathbf{u}}_k = M_h\tilde{\mathbf{v}}_k$ für den zugehörigen Links-EV (LEV) $\tilde{\mathbf{u}}_k$, denn

$$\begin{aligned} &\tilde{\mathbf{u}}_k^T M_h^{-1}A_h = \tilde{\lambda}_k\tilde{\mathbf{u}}_k^T \quad \Longleftrightarrow \\ &\left(M_h\tilde{\mathbf{v}}_k\right)^T M_h^{-1}A_h = \tilde{\lambda}_k\left(M_h\tilde{\mathbf{v}}_k\right)^T \quad \Longleftrightarrow \\ &A_h\tilde{\mathbf{v}}_k = \tilde{\lambda}_k M_h\tilde{\mathbf{v}}_k. \end{aligned} \tag{4.25}$$

Man beachte, dass $\tilde{\lambda}_k \doteq \left(1 + \frac{1}{6}k^2\pi^2h^2\right)k^2\pi^2$ für $h \to 0$ ist (Die $\tilde{\lambda}_k$ konvergieren natürlich für $h \to 0$ gegen die EW $k^2\pi^2$ des 1D Laplace-Operators).

Für den größten EW $\tilde{\lambda}_n$ von $M_h^{-1}A_h$ folgt aus (4.23) im Limes $h \to 0$

$$\tilde{\lambda}_n \doteq \frac{2\cdot 2}{1-\frac{1}{3}\cdot 2}\,n^2 = 12n^2. \tag{4.26}$$

4.2.1.2 $\mathcal{H}$-Darstellbarkeit von 1D Matrixfunktionen

Sei also $M_h^{-1}A_h = VDV^T$ mit der symmetrischen orthogonalen Matrix $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ und der Diagonalmatrix $D = \mathrm{diag}(d_1, \ldots, d_n)$ mit $d_i = \tilde{\lambda}_i$, $1 \le i \le n$. Damit lässt sich für jedes stetige f

$$\begin{aligned} f(M_h^{-1}A_h) &= Vf(D)V^T \text{ mit } f(D) = \mathrm{diag}(f(d_1), \ldots, f(d_n)) \\ &= [\mathbf{v}_1, \ldots, \mathbf{v}_n] \cdot \mathrm{diag}(f(d_1), \ldots, f(d_n)) \cdot \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\ &= \sum_{k=1}^{n} f(d_k)\, \mathbf{v}_k \mathbf{v}_k^T \end{aligned} \tag{4.27}$$

als Summe von n Rang-1-Matrizen schreiben.

Für jeden Block $b_{\mathbf{ij}}$ von $f(M_h^{-1}A_h)$ gilt nun

$$b_{\mathbf{ij}} = \left(f(M_h^{-1}A_h)\right)_{\substack{i_1 \le i \le i_2 \\ j_1 \le j \le j_2}} = \sum_{k=1}^{n} f(d_k)\, \mathbf{v}_k^{\mathbf{i}} \mathbf{v}_k^{\mathbf{j}\,T} \tag{4.28}$$

mit $\mathbf{v}_k^{\mathbf{i}} = \begin{bmatrix} v_{i_1 k} \\ \vdots \\ v_{i_2 k} \end{bmatrix}$ und $\mathbf{v}_k^{\mathbf{j}} = \begin{bmatrix} v_{j_1 k} \\ \vdots \\ v_{j_2 k} \end{bmatrix}$, und wegen der Symmetrie von V

$$b_{\mathbf{ij}} = \begin{bmatrix} \mathbf{v}_{i_1}^T \\ \vdots \\ \mathbf{v}_{i_2}^T \end{bmatrix} \cdot f(D) \cdot [\mathbf{v}_{j_1}, \ldots, \mathbf{v}_{j_2}]\,.$$

Theorem 4.2 (Nicht-$\mathcal{H}$-Darstellbarkeit von $\cos(\tau c\,\Omega_h)$ und $\sin(\tau c\,\Omega_h)$)
Die oszillatorischen Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})$ und $\sin(\tau c\sqrt{M_h^{-1}A_h})$ sind für geeignet gewähltes τ nicht durch $\mathcal{H}$-Matrizen approximierbar.

Beweis: Wir wollen zeigen, dass $b_{\mathbf{ij}} = \left(\cos(\tau c\sqrt{M_h^{-1}A_h})\right)_{\substack{i_1 \le i \le i_2 \\ j_1 \le j \le j_2}}$ durch keine Rang-k-Matrix kleinen Ranges approximierbar ist.
(Für sin anstelle von cos wird exakt analog vorgegangen.)

Aus $\kappa(b_{\mathbf{ij}}) = \frac{\sigma_{\max}(b_{\mathbf{ij}})}{\sigma_{\min}(b_{\mathbf{ij}})}$ klein mit $\sigma_{\max}(b_{\mathbf{ij}})$ und $\sigma_{\min}(b_{\mathbf{ij}})$ den größten und kleinsten Singulärwerten von $b_{\mathbf{ij}}$ folgt sofort die Nichtapproximierbarkeit durch eine Rk-Matrix kleinen Ranges.

Sei $b_{\mathbf{ij}} \in \mathbb{R}^{l \times m}$. Dann gilt die grobe Abschätzung

$$
\begin{aligned}
\sigma_{\max}(b_{\mathbf{ij}}) &= \sqrt{\lambda_{\max}(b_{\mathbf{ij}}^T b_{\mathbf{ij}})} = \max_{\mathbf{x}\neq 0} \sqrt{\frac{\mathbf{x}^T b_{\mathbf{ij}}^T b_{\mathbf{ij}} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}} = \max_{\substack{\mathbf{x}\in\mathbb{R}^m,\\ \|\mathbf{x}\|_2=1}} \|b_{\mathbf{ij}}\mathbf{x}\|_2 \\
&= \max_{\substack{\mathbf{x}\in\mathbb{R}^m,\\ \|\mathbf{x}\|_2=1}} \| \begin{bmatrix} \mathbf{v}_{i_1}^T \\ \vdots \\ \mathbf{v}_{i_2}^T \end{bmatrix} \cdot \cos(\tau c\sqrt{D}) \cdot \underbrace{[\mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}]\,\mathbf{x}}_{\text{Norm 1}} \|_2 \\
&= \max_{\substack{\mathbf{y}\in\langle \mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}\rangle,\\ \|y\|_2=1}} \| \begin{bmatrix} \mathbf{v}_{i_1}^T \\ \vdots \\ \mathbf{v}_{i_2}^T \end{bmatrix} \cdot \underbrace{\cos(\tau c\sqrt{D})\mathbf{y}}_{\text{Norm} \leq 1} \|_2 \\
&\leq \left\| \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_2 = \sqrt{l},
\end{aligned}
$$

wobei im letzten Schritt

$$
|\mathbf{v}_i^T \mathbf{w}| = |\cos \sphericalangle(\mathbf{v}_i, \mathbf{w})| \underbrace{\|\mathbf{v}_i\|_2}_{=1} \underbrace{\|\mathbf{w}\|_2}_{\leq 1} \leq 1
$$

für $i_1 \leq i \leq i_2$ und $\mathbf{w} = \cos(\tau c\sqrt{D})\mathbf{y}$ mit $\|\mathbf{y}\|_2 = 1$ benutzt wurde.

Weiter ist

$$
\begin{aligned}
\sigma_{\min}(b_{\mathbf{ij}}) &= \sqrt{\lambda_{\min}(b_{\mathbf{ij}}^T b_{\mathbf{ij}})} = \min_{\mathbf{x}\neq 0} \sqrt{\frac{\mathbf{x}^T b_{\mathbf{ij}}^T b_{\mathbf{ij}} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}} = \min_{\substack{\mathbf{x}\in\mathbb{R}^m,\\ \|\mathbf{x}\|_2=1}} \|b_{\mathbf{ij}}\mathbf{x}\|_2 \\
&= \min_{\substack{\mathbf{x}\in\mathbb{R}^m,\\ \|\mathbf{x}\|_2=1}} \| \begin{bmatrix} \mathbf{v}_{i_1}^T \\ \vdots \\ \mathbf{v}_{i_2}^T \end{bmatrix} \cdot \cos(\tau c\sqrt{D}) \cdot \underbrace{[\mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}]\,\mathbf{x}}_{\text{Norm 1}} \|_2 \\
&= \min_{\substack{\mathbf{y}\in\langle \mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}\rangle,\\ \|y\|_2=1}} \| \begin{bmatrix} \mathbf{v}_{i_1}^T \\ \vdots \\ \mathbf{v}_{i_2}^T \end{bmatrix} \cdot \cos(\tau c\sqrt{D})\mathbf{y} \|_2 \geq C > 0
\end{aligned}
$$

für geeignet gewähltes τ:
τ lässt sich nämlich so wählen, dass die n Zahlen $\cos(\tau c\sqrt{d_i})$, $1 \leq i \leq n$, einigermaßen zufällig und gleichmäßig im Intervall $[-1, 1]$ verteilt werden. Damit besitzen alle Vektoren $\mathbf{w} = \cos(\tau c\sqrt{D})\mathbf{y}$ mit $\mathbf{y} = [\mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}]\,\mathbf{x} \in \langle \mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}\rangle \subset \mathbb{R}^n$ und $\|\mathbf{y}\|_2 = 1$ auch nicht-vernachlässigbare Komponenten $\alpha_i = \mathbf{v}_i^T \mathbf{w}$ in Richtung $\mathbf{v}_i$, $i_1 \leq i \leq i_2$, so dass schließlich $\left\| \begin{pmatrix} \alpha_{i_1} \\ \vdots \\ \alpha_{i_2} \end{pmatrix} \right\|_2 \geq C$ unabhängig von $\mathbf{x} \in \mathbb{R}^m$, $\|\mathbf{x}\|_2 = 1$, folgt. Man beachte, dass es hier neben der gleichmäßigen Verteilung der „Diagonalelemente“ $\cos(\tau c\sqrt{d_i})$, $1 \leq i \leq n$, auch wesentlich auf die Gestalt der orthonormalen Vektoren $\mathbf{v}_{j_1},\ldots,\mathbf{v}_{j_2}$ ankommt. $V = I$, d.h. eine transzendente Matrixfunktion einer Diagonalmatrix selbst, ist natürlich wiederum eine Diagonalmatrix und somit trivialerweise stets eine exakte $\mathcal{H}$-Matrix.

Wegen $\kappa(b_{\mathbf{ij}}) = \frac{\sigma_{\max}(b_{\mathbf{ij}})}{\sigma_{\min}(b_{\mathbf{ij}})} \leq C\sqrt{l}$ darf also kein einziger Singulärwert von $b_{\mathbf{ij}}$ abgeschnitten werden, ohne größeren Informationsverlust zu erleiden. □

Wir wenden uns nun den Matrixfunktionen $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ und $\psi(\tau^2 c^2 M_h^{-1} A_h)$ zu. Wegen

$$\sigma(\tau^2 c^2 M_h^{-1} A_h) = V \sigma(\tau^2 c^2 D) V^T \tag{4.29}$$

und der Orthogonalität von V entspricht (4.29) bereits der Singulärwertzerlegung von $\sigma(\tau^2 c^2 M_h^{-1} A_h)$, womit

$$\|\sigma(\tau^2 c^2 M_h^{-1} A_h)\|_F = \sqrt{\sum_{k=1}^{n} \sigma(\tau^2 c^2 d_k)^2} = \frac{4}{\tau^2 c^2} \sqrt{\sum_{k=1}^{n} \frac{\sin(\frac{1}{2}\tau c \sqrt{d_k})^4}{d_k^2}}$$

gilt. Analog folgt für $\psi(\tau^2 c^2 M_h^{-1} A_h)$

$$\|\psi(\tau^2 c^2 M_h^{-1} A_h)\|_F = \sqrt{\sum_{k=1}^{n} \psi(\tau^2 c^2 d_k)^2} = \frac{1}{\tau c} \sqrt{\sum_{k=1}^{n} \frac{\sin(\tau c \sqrt{d_k})^2}{d_k}}\,.$$

Die Matrizen $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ und $\psi(\tau^2 c^2 M_h^{-1} A_h)$ sind also für $n \gg 1$ je nach gewünschter Genauigkeit durch Rang-k-Matrizen großen Ranges „approximierbar“ und daher auch als $\mathcal{H}$-Matrizen (mit höchstens diesem Rang in den einzelnen Rk-Blöcken) „darstellbar“. Numerische Beispiele analog zu jenen in Abbildung 4.2 haben jedoch gezeigt, dass dieser Rang in der Praxis viel zu groß ist. Er beläuft sich je nach gewählter Approximationsgüte für große Werte von τc auf mehrere Zehnerpotenzen, weshalb wohl kaum von $\mathcal{H}$-Darstellbarkeit gesprochen werden kann. Zu $\mathcal{H}$-Matrizen gehören nämlich per definitionem Niedrigrang-Matrizen in den Rk-Blöcken.

Dass die den Laplace-Operator darstellende Matrix $M_h^{-1} A_h$ – sie besitzt ein analoges abklingendes Verhalten der Singulärwerte – hingegen schon als $\mathcal{H}$-Matrix darstellbar ist, liegt wohl daran, dass die den Singulärwerten von $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ und $\psi(\tau^2 c^2 M_h^{-1} A_h)$ innewohnende Information über die hochfrequenten Oszillationen in den Außerdiagonalblöcken nicht auf einige wenige Singulärwerte beschränkbar ist.

4.2.1.3 Gewichtete 1D Matrixfunktionen

Wir betrachten jetzt statt der Matrixfunktion $\sigma(\tau^2 c^2 M_h^{-1} A_h)$ die gewichtete Matrixfunktion $\sigma(\tau^2 c^2 M_h^{-1} A_h)\,(M_h^{-1} A_h)^{-q}$ für $q \in \mathbb{N}$. Dann ist

$$\|\sigma(\tau^2 c^2 M_h^{-1} A_h)\,(M_h^{-1} A_h)^{-q}\|_F = \frac{4}{\tau^2 c^2} \sqrt{\sum_{k=1}^{n} \frac{\sin(\frac{1}{2}\tau c \sqrt{d_k})^4}{d_k^{2(q+1)}}}\,.$$

Für $q = 3$ oder 4 lässt sich nun $\sigma(\tau^2 c^2 M_h^{-1} A_h)\,(M_h^{-1} A_h)^{-q}$ bereits durch Niedrigrang-Matrizen gut approximieren und ist somit als $\mathcal{H}$-Matrix darstellbar. Dasselbe gilt auch für die gewichteten Matrixfunktionen $\psi(\tau^2 c^2 M_h^{-1} A_h)\,(M_h^{-1} A_h)^{-q}$, $\cos(\tau c \sqrt{M_h^{-1} A_h})\,(M_h^{-1} A_h)^{-q}$ und $\sin(\tau c \sqrt{M_h^{-1} A_h})\,(M_h^{-1} A_h)^{-q}$.

Fazit: Die Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})$, $\sin(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma(\tau^2c^2M_h^{-1}A_h)$ und $\psi(\tau^2c^2M_h^{-1}A_h)$ lassen sich nicht als $\mathcal{H}$-Matrizen darstellen. Nach Gewichtung mit $(M_h^{-1}A_h)^{-q}$ existieren jedoch Niedrigrang-Approximationen und somit auch $\mathcal{H}$-Approximationen.

4.2.2 Der 2D Fall

Was die 2D Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})$, $\sin(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma(\tau^2c^2M_h^{-1}A_h)$ und $\psi(\tau^2c^2M_h^{-1}A_h)$ betrifft, wählen wir dieselbe Vorgehensweise wie im 1D Fall.

4.2.2.1 Eigenwerte und Eigenvektoren des 2D Laplace-Operators

Die 2D Steifigkeitsmatrix A_h und Massenmatrix M_h der Dimension $N = n^2$ mit $h = \frac{1}{n+1}$ haben die Gestalt

$$A_h = \begin{pmatrix} B & -I & & \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & B \end{pmatrix} \text{ mit } B = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}$$

und

$$M_h = \tfrac{h^2}{12}\begin{pmatrix} C & D_- & & \\ D_+ & \ddots & \ddots & \\ & \ddots & \ddots & D_- \\ & & D_+ & C \end{pmatrix} \text{ mit } C = \begin{pmatrix} 6 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 6 \end{pmatrix},$$

$$D_+ = \begin{pmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 1 \end{pmatrix} \text{ und } D_- = \begin{pmatrix} 1 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix}.$$

Die EW von A_h lauten nun

$$\lambda_{kl} = 4\left(1 - \frac{1}{2}\cos\alpha_k - \frac{1}{2}\cos\alpha_l\right) = 4\left(\sin^2\frac{\alpha_k}{2} + \sin^2\frac{\alpha_l}{2}\right) \tag{4.30}$$

mit $\alpha_k = \frac{k\pi}{n+1}$, $\alpha_l = \frac{l\pi}{n+1}$, $1 \le k, l \le n$, die zugehörigen normierten EV

$$\mathbf{v}_{kl} = \left(\frac{2}{n+1}\sin(i\alpha_k)\sin(j\alpha_l)\right)_{1\le i,j\le n}. \tag{4.31}$$

Mit der Beziehung

$$12M_h = 10h^2I + h^2\begin{pmatrix} 0 & N_- & & \\ N_+ & \ddots & \ddots & \\ & \ddots & \ddots & N_- \\ & & N_+ & 0 \end{pmatrix} - h^2A$$

mit

$$N_+ = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} \text{ und } N_- = \begin{pmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}$$

lässt sich analog zum 1D Fall das verallgemeinerte Eigenwertproblem

$$A_h \tilde{\mathbf{v}}_{kl} = \tilde{\lambda}_{kl} M_h \tilde{\mathbf{v}}_{kl}, \;\; 1 \le k, l \le n,$$

näherungsweise lösen. Die approximierten EW von $M_h^{-1} A_h$ ergeben sich unter Vernachlässigung der Terme der Ordnung $\mathcal{O}(f(k,l)h^2)$ zu

$$\begin{aligned} \tilde{\lambda}_{kl} &= \frac{12}{h^2} \, \frac{2 - \cos\alpha_k - \cos\alpha_l}{3 + \cos\alpha_k + \cos\alpha_l + \cos\alpha_k \cos\alpha_l} && (4.32) \\ &= \frac{6}{h^2} \, \frac{2\,(2 - \cos\alpha_k - \cos\alpha_l)}{6 + \mathcal{O}(f(k,l)h^2)} \\ &= \frac{1}{h^2} \, \lambda_{kl} \, (1 + \mathcal{O}(f(k,l)h^2)) && (4.33) \end{aligned}$$

mit den zugehörigen EV

$$\tilde{\mathbf{v}}_{kl} = \mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2). \qquad (4.34)$$

Dabei bezeichnet f ($\mathbf{f}$) jeweils eine geeignete reellwertige (vektorwertige) Funktion von k und l.

Wie in 1D gilt auch jetzt für den LEV $\tilde{\mathbf{u}}_{kl}$ zum EW $\tilde{\lambda}_{kl}$ von $M_h^{-1} A_h$

$$\tilde{\mathbf{u}}_{kl} = M_h \tilde{\mathbf{v}}_{kl}$$

mit dem REV $\tilde{\mathbf{v}}_{kl}$ zu $\tilde{\lambda}_{kl}$.

Man beachte, dass

$$\tilde{\lambda}_{kl} \doteq \left(1 + \frac{1}{6}(k^2 + l^2)\pi^2 h^2\right)(k^2 + l^2)\pi^2$$

für $h \to 0$ ist (Die $\tilde{\lambda}_{kl}$ konvergieren natürlich für $h \to 0$ gegen die EW $(k^2+l^2)\pi^2$ des 2D Laplace-Operators).

Für den größten EW $\tilde{\lambda}_N$ von $M_h^{-1} A_h$ folgt aus (4.32) im Limes $h \to 0$

$$\tilde{\lambda}_N \doteq \frac{12 \cdot (2+1+1)}{3-1-1+1} \, N = 24N. \qquad (4.35)$$

4.2.2.2 $\mathcal{H}$-Darstellbarkeit von 2D Matrixfunktionen

Analog zum 1D Fall gilt $M_h^{-1} A_h = V D V^{-1}$, wobei $V = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$ und $D = \mathrm{diag}(d_1, \ldots, d_N)$ ist mit $d_i = \tilde{\lambda}_{kl}$, $\mathbf{v}_i = \tilde{\mathbf{v}}_{kl}$ und $i = (l-1)n + k$, $1 \le k, l \le n$. Dann ist

$$\begin{aligned} f(M_h^{-1} A_h) &= V f(D) V^{-1} \text{ für stetiges } f \\ &= \sum_{k=1}^{N} f(d_k) \, \mathbf{v}_k \mathbf{w}_k^T \end{aligned}$$

mit $V^{-1} = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_N^T \end{bmatrix}$ erneut als Summe von N Rang-1-Matrizen darstellbar.

Analog zum 1D Fall kann nun die Strategie verfolgt werden, die Kondition eines Matrixblocks von $\cos(\tau c\sqrt{M_h^{-1}A_h})$ bzw. $\sin(\tau c\sqrt{M_h^{-1}A_h})$ nach oben zu beschränken und daraus die Nichtapproximierbarkeit von $\cos(\tau c\sqrt{M_h^{-1}A_h})$ und $\sin(\tau c\sqrt{M_h^{-1}A_h})$ durch $\mathcal{H}$-Matrizen abzuleiten.

Für die Matrixfunktionen $\sigma(\tau^2c^2M_h^{-1}A_h)$ und $\psi(\tau^2c^2M_h^{-1}A_h)$ ergibt sich wiederum ein abklingendes Verhalten deren approximativen Singulärwerte $\sigma(\tau^2c^2d_i)$ und $\psi(\tau^2c^2d_i)$, $1 \leq i \leq N$, welches rein theoretisch Approximierbarkeit durch Rk-Matrizen ermöglicht. Diese ist jedoch in der Praxis wegen des viel zu großen Ranges nicht realisierbar.

4.2.2.3 Gewichtete 2D Matrixfunktionen

Gewichtung der untersuchten 2D Matrixfunktionen führt wie in 1D auf Matrizen der Form $f(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ mit $q \in \mathbb{N}$, die sich ab $q = 3$ oder 4 gut als Rk-Matrizen und somit auch als $\mathcal{H}$-Matrizen darstellen lassen, wie die numerischen Resultate in Abbildung 4.5 belegen.

4.2.3 Glattheitseigenschaften von Integralkernen

Die Eignung von $\mathcal{H}$-Matrizen zur Approximation von Integraloperatoren

$$(Au)(x) = \int_\Omega \kappa(x,y)u(y)dy, \quad x \in \Omega \subset \mathbb{R}^d,\ d = 2,3$$

basiert hauptsächlich auf bestimmten Glattheitsanforderungen an den Kern $\kappa(x,y)$ (siehe z.B. [20, Kapitel 3.5], [21] oder [23]).
Wir setzen voraus, dass die Singularitätenfunktion $\kappa(x,y)$ die asymptotische Glattheitsbedingung

$$\left|\partial_x^\alpha \partial_y^\beta \kappa(x,y)\right| \leq C(|\alpha|,|\beta|)\,\|x-y\|_2^{-|\alpha|-|\beta|}\,|\kappa(x,y)| \tag{4.36}$$

für alle $\alpha,\beta \in \mathbb{N}_0^d$, $x,y \in \Omega$, $x \neq y$, erfüllt, wobei α,β Multi-Indizes mit $|\alpha| = \alpha_1 + \ldots + \alpha_d$ sind.

Die Taylorentwicklung von $\kappa(x,y)$ um y_* lautet nun

$$\kappa(x,y) = \sum_{|\nu|=0}^{m-1} \frac{1}{\nu!}(y_*-y)^\nu \partial_y^\nu \kappa(x,y_*) + R_m \tag{4.37}$$

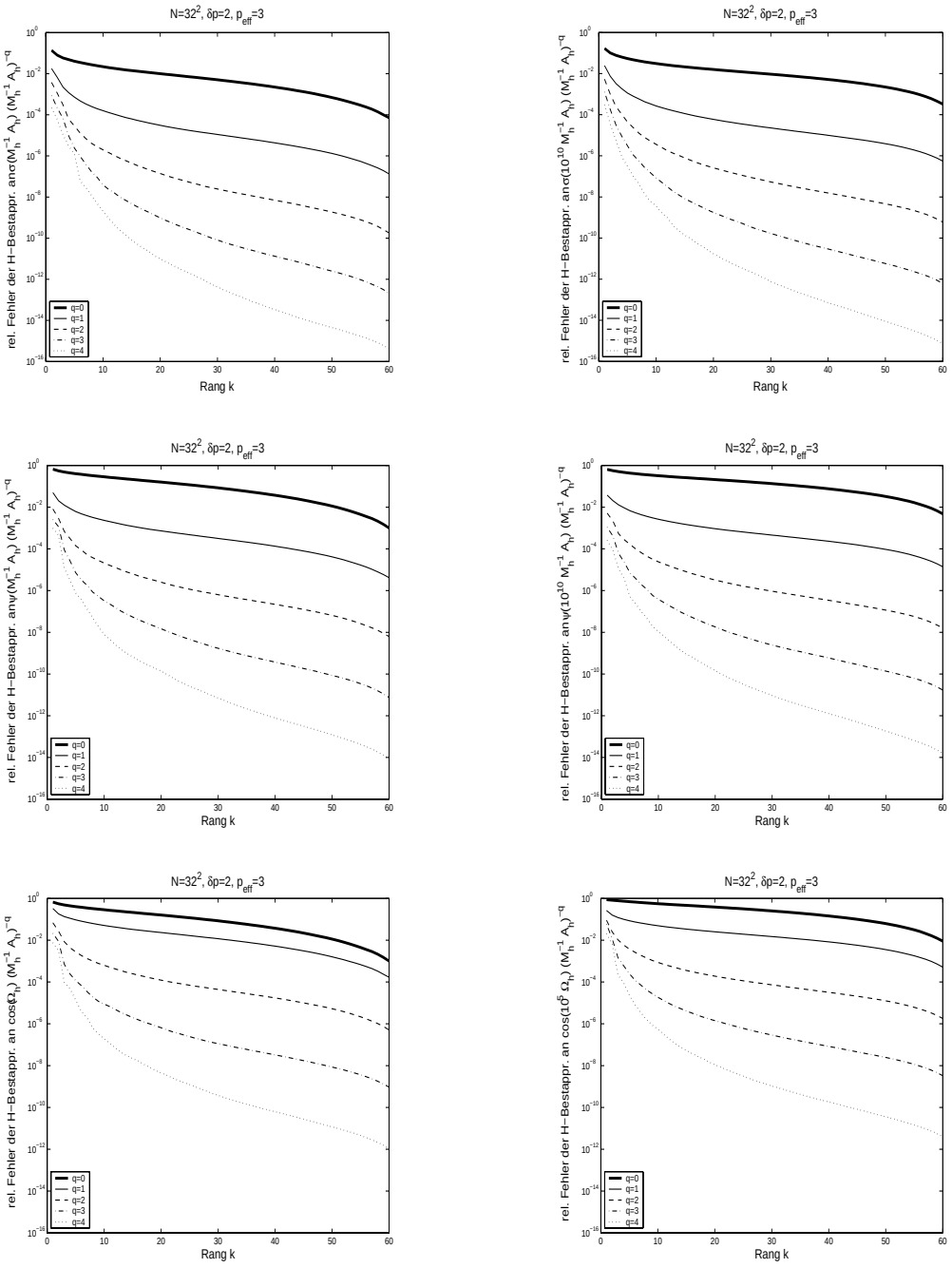

Abbildung 4.5: Relativer Fehler der bestmöglichen $\mathcal{H}$-Approximation an die exakten gewichteten Matrixfunktionen $\sigma(\tau^2 c^2 M_h^{-1} A_h)\,(M_h^{-1} A_h)^{-q}$ (oben), $\psi(\tau^2 c^2 M_h^{-1} A_h)\,(M_h^{-1} A_h)^{-q}$ (Mitte) und $\cos(\tau c\,\Omega_h)\,(M_h^{-1} A_h)^{-q}$ mit $\Omega_h = \sqrt{M_h^{-1} A_h}$ (unten) in $\|\cdot\|_F$ für $\tau c = 10^0$ (links) und $\tau c = 10^5$ (rechts) sowie $q = 0, 1, 2, 3$ und 4.

mit dem Rest R_m, der abgeschätzt werden kann durch

$$|R_m| \leq \frac{1}{m!} \|y_* - y\|_2^m \max_{\eta \in [y,y_*], |\nu|=m} |\partial_y^\nu \kappa(x,\eta)| .$$

Die separierte Entwicklung von $\kappa(x,y)$ aus (4.37) liefert nun eine Matrix vom Rang $k = \#\{\nu \in \mathbb{N}_0^d : 0 \leq |\nu| \leq m-1\}$, und damit werden die Rang-k-Blöcke der $\mathcal{H}$-Matrix zur Approximation des kontinuierlichen Operators gebildet (siehe [20, Seite 6]).

Die Singularitätenfunktion des Laplace-Operators $\triangle$ lautet beispielsweise

$$\kappa(x,y) = \begin{cases} \frac{1}{2\pi} \ln \|x-y\|_2 & \text{falls } d = 2 \\ -\frac{1}{(n-2)\omega_d} \frac{1}{\|x-y\|_2^{d-2}} & \text{falls } d > 2 \end{cases}, \tag{4.38}$$

wobei ω_d die Oberfläche der d-dimensionalen Einheitskugel ist. κ erfüllt die Gleichung $\triangle \kappa = \delta(x-y)$. Es gilt $\kappa \in C^\infty$ für $x \in \mathbb{R}^d \setminus \{y\}$, und κ ist singulär für $x = y$ (siehe [31, Seite 80]). κ erfüllt die asymptotische Glattheitsbedingung (4.36). Man beachte dabei, dass das Vorhandensein logarithmischer Terme (z.B. hier für $d = 2$) eine leichte Modifizierung von (4.36) erfordert, z.B. $\|x-y\|_2^{2-|\alpha|-|\beta|-d}$ anstelle von $\|x-y\|_2^{-|\alpha|-|\beta|} |\kappa(x,y)|$ (siehe [23, Kapitel 2.3]).

Als Nächstes betrachten wir die Wärmeleitungsgleichung

$$u_t - \triangle u = 0 \text{ im } \mathbb{R}^{d+1}. \tag{4.39}$$

Die Fundamentallösung des Differentialoperators $\partial_t - \triangle_x$ lautet

$$\Phi(x,t) = \begin{cases} \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|x\|_2^2}{4t}} & \text{falls } t > 0 \\ 0 & \text{falls } t \leq 0 \end{cases} \tag{4.40}$$

mit $(x,t) \in \mathbb{R}^d \times \mathbb{R}$ (siehe [31, Seite 81]), d.h. es gilt $\Phi \in C^\infty(\mathbb{R}^{d+1} \setminus \{0\})$ und $(\partial_t - \triangle_x)\Phi = \delta_0$. Φ ist singulär im Punkt $(0,0)$.
Es lässt sich leicht nachprüfen, dass auch $\Phi(x,t)$ die asymptotische Glattheitsbedingung (4.36) vollauf erfüllt. Zugehörige $\mathcal{H}$-Approximationen an den Lösungsoperator $\exp(-tA)$ finden sich in [14] oder in Abschnitt 4.5 dieser Arbeit.

Die Helmholtz-Gleichung

$$\triangle u + \kappa^2 u = 0 \tag{4.41}$$

mit einer positiven Wellenzahl κ besitzt den Helmholtz-Kern

$$\kappa(x,y) = \frac{i}{4\pi} H_0^{(1)}(-\kappa \|x-y\|_2) \tag{4.42}$$

für $d = 2$ mit der Hankelschen Funktion $H_0^{(1)}$ erster Gattung der Ordnung 0 (siehe [37, Seite 210] oder [36]) bzw.

$$\kappa(x,y) = \frac{1}{4\pi} \frac{e^{i\kappa\|x-y\|_2}}{\|x-y\|_2} \tag{4.43}$$

für $d = 3$ (siehe [25] oder [36]).
Für $d = 3$ gilt beispielsweise

$$\begin{aligned}\partial_{x_i} 4\pi\kappa(x,y) &= \partial_{x_i}\left(\tfrac{e^{i\kappa\|x-y\|_2}}{\|x-y\|_2}\right) = \tfrac{e^{i\kappa\|x-y\|_2}\cdot i\kappa\frac{x_i-y_i}{\|x-y\|_2}\cdot\|x-y\|_2 - e^{i\kappa\|x-y\|_2}\cdot\frac{x_i-y_i}{\|x-y\|_2}}{\|x-y\|_2^2} \\ &= 4\pi\kappa(x,y)\left(\tfrac{i\kappa(x_i-y_i)}{\|x-y\|_2} - \tfrac{x_i-y_i}{\|x-y\|_2^2}\right)\end{aligned}$$

und damit

$$|\partial_{x_i}\kappa(x,y)| \le C\,|\kappa(x,y)| \ \text{ mit } \ C \sim \kappa,$$

weiter

$$\begin{aligned}\partial_{y_j}\partial_{x_i} 4\pi\kappa(x,y) &= \partial_{y_j}\left(\partial_{x_i}\left(\tfrac{e^{i\kappa\|x-y\|_2}}{\|x-y\|_2}\right)\right) \\ &= \partial_{y_j}\left(4\pi\kappa(x,y)\right)\cdot\left(\tfrac{i\kappa(x_i-y_i)}{\|x-y\|_2} - \tfrac{x_i-y_i}{\|x-y\|_2^2}\right) + \\ &\quad 4\pi\kappa(x,y)\cdot\underbrace{i\kappa\frac{-\delta_{ij}\cdot\|x-y\|_2 - (x_i-y_i)\frac{x_j-y_j}{\|x-y\|_2}}{\|x-y\|_2^2}}_{\sim\frac{1}{\|x-y\|_2}} - \\ &\quad 4\pi\kappa(x,y)\cdot\underbrace{\frac{-\delta_{ij}\cdot\|x-y\|_2^2 - (x_i-y_i)\,2\|x-y\|_2\,\frac{x_j-y_j}{\|x-y\|_2}}{\|x-y\|_2^4}}_{\sim\frac{1}{\|x-y\|_2^2}} \\ &= -\tfrac{i^2\kappa^2(x_i-y_i)(x_j-y_j)}{\|x-y\|_2^2}\cdot 4\pi\kappa(x,y) + \mathcal{O}\left(\tfrac{1}{\|x-y\|_2}\right)\cdot\kappa(x,y)\end{aligned}$$

und damit

$$|\partial_{y_j}\partial_{x_i}\kappa(x,y)| \le C\,|\kappa(x,y)| \ \text{ mit } \ C \sim \kappa^2,$$

usw., was auf die in [25, Kapitel 4] gezeigte Abhängigkeit des Ranges k von κ hinweist.

Die Schrödinger-Gleichung

$$iu_t + \triangle u = 0 \ \text{ in } \ \mathbb{R}^d \times (0,\infty) \tag{4.44}$$

besitzt die Fundamentallösung

$$\Phi(x,t) = \begin{cases} \frac{1}{(4\pi i t)^{\frac{d}{2}}}\, e^{\frac{i\|x\|_2^2}{4t}} & \text{falls } t > 0 \\ 0 & \text{falls } t \le 0 \end{cases} \tag{4.45}$$

(siehe [31, Seite 81 ff.] bzw. [9, Seite 189]), die wiederum singulär in $(0,0)$ ist. Es ist leicht einzusehen, dass diese Kernfunktion die asymptotische Glattheitsbedingung nicht erfüllt.
Wir haben in Theorem 4.2 gezeigt, dass die Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})$ und $\sin(\tau c\sqrt{M_h^{-1}A_h})$ nicht durch $\mathcal{H}$-Matrizen approximierbar sind. Der Lösungsoperator der Schrödinger-Gleichung (4.44) ist formal durch $\exp(-itA)$

gegeben. Somit verhindert auch hier das oszillierende und größenordnungsmäßig nicht abklingende Verhalten der Singulärwerte von $\exp(-itM_h^{-1}A_h)$ in Verbindung mit der Form der Singulärvektoren die Approximierbarkeit durch $\mathcal{H}$-Matrizen.

Wir kommen nun schließlich zur 2D Wellengleichung, die ja für uns von vorrangigem Interesse ist.
Der Wellenoperator in $\mathbb{R}^{d+1}$

$$\Box = c^{-2}\partial_{tt} - \triangle_x \tag{4.46}$$

besitzt nach [31] eine Fundamentallösung

$$E = \pi^{\frac{1-d}{2}} \frac{c}{4} A^* \chi_+^{\frac{1-d}{2}} \text{ mit } A = c^2\, t^2 - \|x\|_2^2, \tag{4.47}$$

wobei A^* durch $A^*u := u \circ A$ für alle $u \in C^0(\mathbb{R})$ definiert ist und

$$\chi_+^a = \frac{x_+^a}{\Gamma(a+1)} \text{ für } \Re a > -1$$

mit

$$x_+^a = \begin{cases} x^a & \text{falls } x > 0 \\ 0 & \text{falls } x \leq 0 \end{cases} \quad \text{für } \Re a > -1$$

ist. Der Träger von E besteht aus dem Doppelkegel $A \geq 0$.

$$E_+ = \begin{cases} 2E & \text{falls } t > 0 \\ 0 & \text{falls } c\,t < \|x\|_2 \end{cases}$$

besitzt den Vorwärtskegel mit $t \geq 0$ als Träger, und es gilt $\Box E_+ = \delta_0$.
Für $d = 2$ lautet die Fundamentallösung E aus Gleichung (4.47)

$$\begin{aligned} E(x,t) &= \tfrac{c}{4\pi} x_+^{-\frac{1}{2}} (c^2\, t^2 - \|x\|_2^2) \\ &= \tfrac{c}{4\pi} \begin{cases} (c^2\, t^2 - \|x\|_2^2)^{-\frac{1}{2}} & \text{falls } A = c^2\, t^2 - \|x\|_2^2 > 0 \\ 0 & \text{sonst} \end{cases} \end{aligned} \tag{4.48}$$

Es lässt sich leicht nachprüfen, dass $E(x,t)$ aus (4.48) die asymptotische Glattheitsbedingung (4.36) nicht erfüllen kann. Jede partielle Ableitung

$$\partial_x^\alpha \partial_y^\beta E(x-y, t-s)$$

divergiert bei Annäherung an den Lichtkegel $A = c^2\,(t-s)^2 - \|x-y\|_2^2 = 0$. Diese Tatsache verdeutlicht die Unmöglichkeit der $\mathcal{H}$-Approximierbarkeit der in diesem Kapitel untersuchten Matrixfunktionen.

4.2.4 Konsequenzen für die 2D Wellengleichung

Wegen

$$\|M_h^{-1}A_h\|_2 = \mathcal{O}(N) = \mathcal{O}(h^{-2})$$

und der Tatsache, dass $\cos(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma(\tau^2 c^2 M_h^{-1}A_h)$ und $\psi(\tau^2 c^2 M_h^{-1}A_h)$ nur dann durch $\mathcal{H}$-Matrizen approximierbar sind, falls $\|\tau^2 c^2 M_h^{-1}A_h\|_2 \leq C\pi^2$ ist, folgt:

1. Die Zeitschrittweite τ ist zu beschränken durch

$$\tau = \mathcal{O}(c^{-1}h) = \mathcal{O}(c^{-1}N^{-\frac{1}{2}}).$$

2. Die Komplexität des hierarchischen FE-Gautschi-Algorithmus unter Verwendung von $\mathcal{H}$-Approximationen an obengenannte transzendente Matrixfunktionen zur Lösung der inhomogenen 2D Wellengleichung beträgt $\mathcal{O}(cN^{\frac{3}{2}}\log N)$:
Es sind nämlich $\mathcal{O}(cN^{\frac{1}{2}})$ Zeitschritte notwendig zum Erreichen einer Endzeit t, und in jedem Zeitschritt sind Matrix-Vektor-Multiplikationen mit einem Aufwand von $\mathcal{O}(N\log N)$ flops durchzuführen. Die dafür benötigten und eingangs ein für alle Mal berechneten Matrixfunktionen $\cos_{\mathcal{H}}(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma_{\mathcal{H}}(\tau^2c^2M_h^{-1}A_h)$ und $\psi_{\mathcal{H}}(\tau^2c^2M_h^{-1}A_h)$ kosten $\mathcal{O}(N\log^2 N)$ Operationen.

4.2.5 Übersicht zur Lösung der hochfrequenten 2D Wellengleichung

- **Explizite** Zeitschrittverfahren liefern die CFL-Bedingung $\tau = \mathcal{O}(c^{-1}h)$ aus Stabilitätsgründen.
- **Implizite** Zeitschrittverfahren liefern $\tau = \mathcal{O}(c^{-1}h)$ wegen ansonsten großer Phasenfehler.
- **Transzendentale** Verfahren:
  - Die Darstellung von $\sigma(\tau^2c^2M_h^{-1}A_h)\,\mathbf{v}_h$ mittels **Krylovraummethoden** liefert eine Beschränkung von τ der Form $\tau = \mathcal{O}(c^{-1}h)$ aus Genauigkeitsgründen (bei konstanter Dimension der Krylovunterräume).
  - Die Darstellung von $\sigma(\tau^2c^2M_h^{-1}A_h)$ durch **$\mathcal{H}$-Matrizen** liefert $\tau = \mathcal{O}(c^{-1}h)$ aus Gründen der $\mathcal{H}$-Darstellbarkeit.

Während bei expliziten und impliziten Verfahren die Beschränkung der Zeitschrittweite von der Zeitdiskretisierung herrührt, ist bei exponentiellen Zeitintegratoren die Kopplung von τ an h und c^{-1} nicht durch die Diskretisierung, sondern durch die explizite Berechnung der mit sin und cos gebildeten Matrixfunktionen verursacht.

Um die 2D Wellengleichung mit fast linearem Aufwand in $\mathcal{H}$-Arithmetik zu lösen, müssen wir also nach einem neuen Ansatz suchen.

4.3 Das $\mathcal{H}$-Eigenwertproblem für den diskreten 2D Laplace-Operator

Im vorhergehenden Abschnitt haben wir gesehen, dass die Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma(\tau^2c^2M_h^{-1}A_h)$ und $\psi(\tau^2c^2M_h^{-1}A_h)$ nicht durch $\mathcal{H}$-Matrizen

approximierbar sind. Ebenso haben wir beobachtet, dass sich die gewichteten Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})\,(M_h^{-1}A_h)^{-q}$, $\sigma(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ und $\psi(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ mit geeignetem $q \in \mathbb{N}$ wegen des hinreichend schnellen Abklingverhaltens deren Singulärwerte bereits durch Niedrigrang-Matrizen gut approximieren lassen.

Im nächsten Abschnitt werden wir ein Verfahren entwickeln, diese Niedrigrang-Approximationen mit fast linearem Aufwand unter Verwendung von $\mathcal{H}$-Matrizen zu berechnen. Dazu werden jedoch die kleinsten Eigenwerte von $M_h^{-1}A_h$ samt Eigenvektoren benötigt.

Daher werden wir in diesem Abschnitt eine Methode erstellen, das Eigenwertproblem für $M_h^{-1}A_h$ in $\mathcal{H}$-Arithmetik mit fast linearem Aufwand zu lösen.

4.3.1 Eigenwerte und Eigenvektoren des diskreten 2D Laplace-Operators in $\mathcal{H}$-Arithmetik

Für die EW $\tilde{\lambda}_{kl}$ von $M_h^{-1}A_h$ gilt

$$\tilde{\lambda}_{kl} \underset{h\to 0}{\longrightarrow} (k^2+l^2)\,\pi^2, \quad k,l \geq 1,$$

d.h. die überwiegende Mehrheit der EW von $M_h^{-1}A_h$ konvergiert für $h \to 0$ gegen mehrfache EW des Operators $A = -\triangle$. Daher bedienen wir uns eines sogenannten simultanen Iterationsverfahrens: Nach jedem Schritt orthogonaler Iteration bauen wir einen sogenannten Schur-Rayleigh-Ritz-Schritt (SRR-Schritt) ein. Dieses Verfahren geht auf Stewart zurück und stellt eine natürliche Verallgemeinerung des Rayleigh-Ritz-Verfahrens zur Approximation von EV hermitescher Matrizen dar (siehe [43]).

Sei m die Vielfachheit des zu berechnenden EW im Limes $h \to 0$. Sei weiter $Q^{(0)} = [\mathbf{q}_1^{(0)}, \ldots, \mathbf{q}_m^{(0)}]$ die Startmatrix mit den orthogonalen Spalten $\mathbf{q}_j^{(0)}$, wobei $\mathbf{q}_j^{(0)}$ dem hierarchisierten Einheitsvektor $\mathbf{e}_j$ entspricht, d.h. $\mathbf{q}_j^{(0)} = P\mathbf{e}_j$ mit der Permutationsmatrix P aus Abschnitt 3.3.1.

Damit ergibt sich mit dem Shift μ der folgende Algorithmus zur Berechnung von EW und EV der Matrix $M_h^{-1}A_h$, wobei alle Operationen mit $\mathcal{H}$-Matrizen in $\mathcal{H}$-Arithmetik erfolgen:

Algorithmus 4.3 (Simultane Iteration in $\mathcal{H}$-Arithmetik)
Bestimmung von μ nach (4.49).
$Q^{(0)} = P\,[\mathbf{e}_1, \ldots, \mathbf{e}_m]$ *mit P aus Abschnitt 3.3.1.*
Festlegung einer Mindestanzahl $k_{\min}$ an Iterationsschritten
und des Parameters $\epsilon > 0$ fürs Abbruchkriterium.
$k = 0$
`while` $\left(\left(\frac{err_j^{(k)}}{err_j^{(k+1)}} > 1+\epsilon \right)$ *für ein* $j \in \{1,\ldots,m\}$ *ODER* $k < k_{\min} \right)$
1. $k = k+1$
2. *LGS* $(\mu I - M_h^{-1}A_h)Z^{(k)} = Q^{(k-1)}$

3. *QR-Zerlegung von* $Z^{(k)}$: $Z^{(k)} = Q^{(k)}R^{(k)}$

4. *SRR-Schritt:*

   LGS $(\mu I - M_h^{-1}A_h)C^{(k)} = Q^{(k)}$

   $B^{(k)} = Q^{(k)T}C^{(k)}$

   Schur-Zerlegung von $B^{(k)}$: $Y^{(k)T}B^{(k)}Y^{(k)} = T^{(k)}$

   $Q^{(k)} := Q^{(k)}Y^{(k)}$

5. *Rayleigh-Quotienten und relative Rückwärtsfehler:*

   `for` $j = 1, \ldots, m$

   $\lambda_j^{(k)} = \mathbf{q}_j^{(k)T} M_h^{-1} A_h \mathbf{q}_j^{(k)}$

   $err_j^{(k)} = \frac{\|M_h^{-1}A_h\mathbf{q}_j^{(k)} - \lambda_j^{(k)}\mathbf{q}_j^{(k)}\|_2}{\|M_h^{-1}A_h\|_2}$

   `end`

`end`

Lineare Ausgleichsprobleme und Rayleigh-Quotienten
nach l *Iterationsschritten:*

`for` $j = 1, \ldots, m$

$$\|M_h^{-1}A_h\left(\mathbf{q}_j^{(l)} + \sum_{\substack{i=1\\ i\neq j}}^{m} \alpha_i \mathbf{q}_i^{(l)}\right) - \lambda_j^{(l)}\left(\mathbf{q}_j^{(l)} + \sum_{\substack{i=1\\ i\neq j}}^{m} \alpha_i \mathbf{q}_i^{(l)}\right)\|_2 = \min_{\alpha_i}!$$

$$\mathbf{q}_j^{(l)} = \mathbf{q}_j^{(l)} + \sum_{\substack{i=1\\ i\neq j}}^{m} \alpha_i \mathbf{q}_i^{(l)}$$

$$\mathbf{q}_j^{(l)} = \frac{\mathbf{q}_j^{(l)}}{\|\mathbf{q}_j^{(l)}\|_2}$$

$$\lambda_j^{(l)} = \mathbf{q}_j^{(l)T} M_h^{-1} A_h \mathbf{q}_j^{(l)}$$

`end`

Da die Spalten von $Q^{(k)}Y^{(k)}$ i. Allg. bessere Approximationen an die zu berechnenden EV als die Spalten von $Q^{(k)}$ sind (siehe [43]), ersetzen wir am Ende des SRR-Schritts in Algorithmus 4.3 $Q^{(k)}$ durch $Q^{(k)}Y^{(k)}$.

Die LGS

$$(\mu I - M_h^{-1}A_h)Z^{(k)} = Q^{(k-1)} \quad \Longleftrightarrow \quad (\mu M_h - A_h)Z^{(k)} = M_h Q^{(k-1)}$$

und

$$(\mu I - M_h^{-1}A_h)C^{(k)} = Q^{(k)} \quad \Longleftrightarrow \quad (\mu M_h - A_h)C^{(k)} = M_h Q^{(k)}$$

werden durch $\mathcal{H}$-LDL^T-Zerlegung der symmetrischen Matrix $\mu M_h - A_h$ in $\mathcal{O}(N \log^2 N)$ flops mit anschließender Vorwärts- und Rückwärtssubstitution in $\mathcal{O}(N \log N)$ Operationen gelöst. Da die LDL^T-Zerlegung nur einmal für den gesamten Algorithmus 4.3 gebildet werden muss, beläuft sich der Gesamtaufwand für die simultane Iteration aus Algorithmus 4.3 auf $\mathcal{O}(N \log^2 N)$ flops.

Für die Matrix-Vektor-Produkte $M_h^{-1}A_h\mathbf{q}_j^{(k)}$, $1 \leq j \leq m$, $1 \leq k \leq l$, benötigen

wir nach anfangs durchgeführter Cholesky-Zerlegung von M_h nur jeweils eine Matrix-Vektor-Multiplikation sowie eine Vorwärts- und Rückwärtssubstitution.

Als Shift μ verwenden wir die in Abschnitt 4.2.2.1 berechneten Näherungen für die Eigenwerte von $M_h^{-1}A_h$

$$\mu_{kl} = \frac{12}{h^2}\,\frac{2-\cos\alpha_k-\cos\alpha_l}{3+\cos\alpha_k+\cos\alpha_l+\cos\alpha_k\cos\alpha_l}. \tag{4.49}$$

Ist der Shift μ zu genau und damit die Kondition der Matrix $\mu M_h - A_h$ zu groß (und daher die $\mathcal{H}$-Operationen zu ungenau), kann er zur Vermeidung allzu großer Konditionszahlen gesteuert werden (z.B. $\mu_{kl} - 1$ statt μ_{kl} aus (4.49) verwendet werden). Da die μ_{kl} die exakten EW von $M_h^{-1}A_h$ bis auf $\mathcal{O}(f(k,l)h^2)$ approximieren, taugen für kleiner werdendes h auch immer mehr Shifts zur Approximation der exakten EW. Der numerischen Praxis zufolge liegt die Anzahl der mit den Shifts μ aus (4.49) approximierbaren EW in der Größenordnung $\mathcal{O}(\sqrt{N}) = \mathcal{O}(h^{-1})$. Die mittels Algorithmus 4.3 berechneten $\mathcal{H}$-EW und $\mathcal{H}$-EV sind dabei so genau, wie es die $\mathcal{H}$-Matrix $L_{\mathcal{H}}$ der $\mathcal{H}$-LDL^T-Zerlegung von $\mu M_h - A_h$ ist: Nach erfolgter approximativer LDL^T-Zerlegung sind nämlich sämtliche Operationen von Algorithmus 4.3 exakt!

Die Spalten $\mathbf{q}_j^{(l)}$ der nach l Schritten simultaner Iteration gewonnenen $N \times m$-Matrix $Q^{(l)} = [\mathbf{q}_1^{(l)}, \ldots, \mathbf{q}_m^{(l)}]$ spannen den zu den $\mathcal{H}$-EW $\lambda_1^{(l)}, \ldots, \lambda_m^{(l)}$ gehörigen m-dimensionalen invarianten Unterraum des $\mathbb{R}^N$ auf. Dabei entspricht i. Allg. nur die erste Spalte $\mathbf{q}_1^{(l)}$ einem $\mathcal{H}$-EV, dem zum $\mathcal{H}$-EW $\lambda_1^{(l)}$.

Zur Nachbesserung der übrigen $\mathcal{H}$-EV lösen wir nun für den Fall $m > 1$ die folgenden m linearen Ausgleichsprobleme:
Suche $\alpha_{(j)} = (\alpha_i)_{\substack{1\le i\le m\\ i\neq j}} \in \mathbb{R}^{m-1}$, so dass

$$\begin{aligned}
&\|M_h^{-1}A_h\left(\mathbf{q}_j^{(l)} + \sum_{\substack{i=1\\ i\neq j}}^{m}\alpha_i\mathbf{q}_i^{(l)}\right) - \lambda_j^{(l)}\left(\mathbf{q}_j^{(l)} + \sum_{\substack{i=1\\ i\neq j}}^{m}\alpha_i\mathbf{q}_i^{(l)}\right)\|_2 = \min_{\alpha_i}! \quad \Longleftrightarrow\\
&\|\left(M_h^{-1}A_h - \lambda_j^{(l)}I\right)Q_{(j)}^{(l)}\alpha_{(j)} - \left(\lambda_j^{(l)}I - M_h^{-1}A_h\right)\mathbf{q}_j^{(l)}\|_2 = \min_{\alpha_i}!
\end{aligned}$$

mit $Q_{(j)}^{(l)} = [\mathbf{q}_1^{(l)}, \ldots, \mathbf{q}_{j-1}^{(l)}, \mathbf{q}_{j+1}^{(l)}, \ldots, \mathbf{q}_m^{(l)}]$ unter Verwendung der QR-Zerlegung der $N \times (m-1)$-Matrix $\left(M_h^{-1}A_h - \lambda_j^{(l)}I\right)Q_{(j)}^{(l)}$. Die Lösung dieser linearen Ausgleichsprobleme erfordert nach bereits erfolgter Berechnung des $\mathcal{H}$-Cholesky-Faktors von M_h – schon benötigt für die Bestimmung der Rayleigh-Quotienten und der Rückwärtsfehler in Algorithmus 4.3 – nur mehr $\mathcal{O}(N \log N)$ flops.

Im Anschluss daran werden die optimierten $\mathcal{H}$-EV nachnormiert und abschließend die zugehörigen Rayleigh-Quotienten – zur Optimierung der $\mathcal{H}$-EW – berechnet.

4.3.2 Rückwärtsfehleranalyse des Eigenwertproblems in $\mathcal{H}$-Arithmetik

Sei $(\lambda, \mathbf{v})$ das fehlerbehaftete Eigenpaar von $M_h^{-1}A_h$ aus Algorithmus 4.3 mit normiertem EV $\mathbf{v}$. Dieses hinterlässt in der Eigenwertgleichung das Residuum $\mathbf{r} = \lambda\mathbf{v} - M_h^{-1}A_h\mathbf{v}$.
Die Rang-1-Matrix $\Delta(M_h^{-1}A_h) = \mathbf{r}\mathbf{v}^T$ erfüllt

$$(M_h^{-1}A_h + \Delta(M_h^{-1}A_h))\mathbf{v} = \lambda\mathbf{v},$$

und für den Rückwärtsfehler $\Delta(M_h^{-1}A_h)$ gilt

$$\|\Delta(M_h^{-1}A_h)\|_2 = \max_{\|\mathbf{w}\|_2=1} \|\Delta(M_h^{-1}A_h)\mathbf{w}\|_2 = \max_{\|\mathbf{w}\|_2=1} \|\mathbf{r}\mathbf{v}^T\mathbf{w}\|_2 = \|\mathbf{r}\|_2.$$

Damit ergibt sich der relative Rückwärtsfehler in der Spektralnorm zu

$$\frac{\|\Delta(M_h^{-1}A_h)\|_2}{\|M_h^{-1}A_h\|_2} = \frac{\|\mathbf{r}\|_2}{\|M_h^{-1}A_h\|_2}. \tag{4.50}$$

Als Abbruchkriterium für die simultane Iteration in Algorithmus 4.3 verwenden wir

$$\frac{err_j^{(k)}}{err_j^{(k+1)}} \leq 1 + \epsilon \text{ für alle } j \in \{1, \ldots, m\}$$

mit einem $\epsilon > 0$ und

$$err_j^{(k)} = \frac{\|M_h^{-1}A_h\mathbf{q}_j^{(k)} - \lambda_j^{(k)}\mathbf{q}_j^{(k)}\|_2}{\|M_h^{-1}A_h\|_2} \tag{4.51}$$

für die zu berechnenden REV $\mathbf{q}_j^{(k)}$, wobei wir $\|M_h^{-1}A_h\|_2$ durch $24N$ (siehe Abschnitt 4.2.2.1) approximieren.

In den Abbildungen 4.6 bis 4.8 sind die numerischen Ergebnisse für die mittels Algorithmus 4.3 in $\mathcal{H}$-Arithmetik berechneten und im Limes $h \to 0$ einfachen EW λ_1, 2-fachen EW λ_7, λ_8 sowie 4-fachen EW λ_{42}, λ_{43}, λ_{44}, λ_{45} dargestellt.
Mit Hilfe des Abstands der beiden Linien, die den exakten relativen Vorwärtsfehler und den geschätzten relativen Rückwärtsfehler der $\mathcal{H}$-EV kennzeichnen, lassen sich Rückschlüsse auf die Kondition der jeweiligen EV ziehen.
Bei den $\mathcal{H}$-EV $\mathbf{v}_{42}$ und $\mathbf{v}_{43}$ fällt auf, dass erst für größeres N die in den anderen Beispielen gegebene Approximationsgüte erreicht wird. Das liegt einfach daran, dass der Shift zur simultanen Berechnung der 4 EW λ_{42} bis λ_{45} erst ab einem bestimmten N gut genug für das Erreichen der gewünschten Genauigkeit ist.

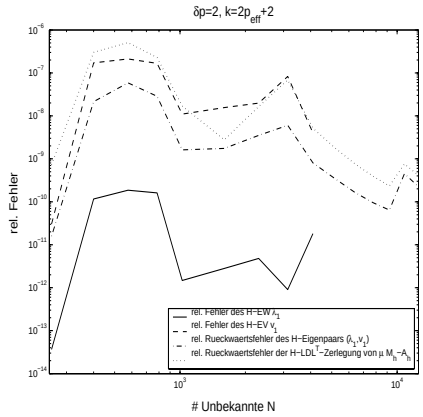

Abbildung 4.6: Relativer Fehler des $\mathcal{H}$-EW λ_1 und des zugehoerigen $\mathcal{H}$-EV $\mathbf{v}_1$ sowie relativer Rückwärtsfehler des $\mathcal{H}$-Eigenpaars $(\lambda_1, \mathbf{v}_1)$ in $\|\cdot\|_2$ und relativer Rückwärtsfehler der $\mathcal{H}$-LDL^T-Zerlegung von $\mu M_h - A_h$ in $\|\cdot\|_F$ für $\delta p = 2$ und Rang $k = 2p_{eff} + 2$.

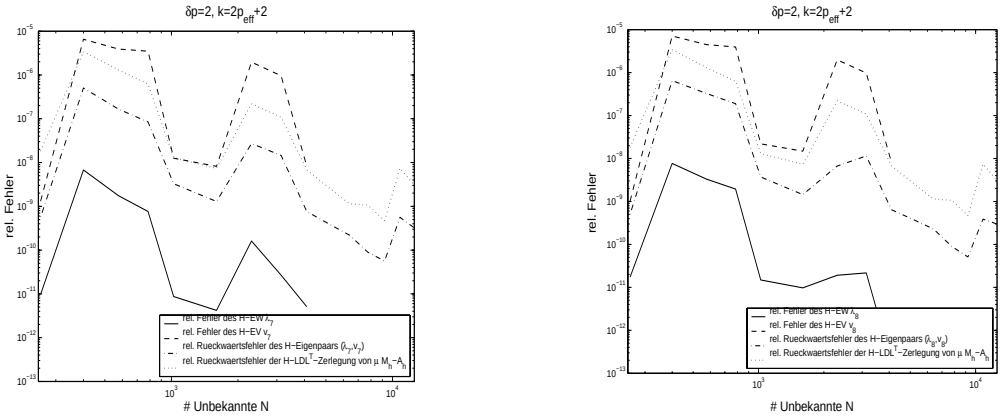

Abbildung 4.7: Relativer Fehler der $\mathcal{H}$-EW λ_7 (links) und λ_8 (rechts) und der zugehoerigen $\mathcal{H}$-EV $\mathbf{v}_7$ (links) und $\mathbf{v}_8$ (rechts) sowie relativer Rückwärtsfehler der $\mathcal{H}$-Eigenpaare $(\lambda_7, \mathbf{v}_7)$ (links) und $(\lambda_8, \mathbf{v}_8)$ (rechts) in $\|\cdot\|_2$ und relativer Rückwärtsfehler der $\mathcal{H}$-LDL^T-Zerlegung von $\mu M_h - A_h$ in $\|\cdot\|_F$ für $\delta p = 2$ und Rang $k = 2p_{eff} + 2$.

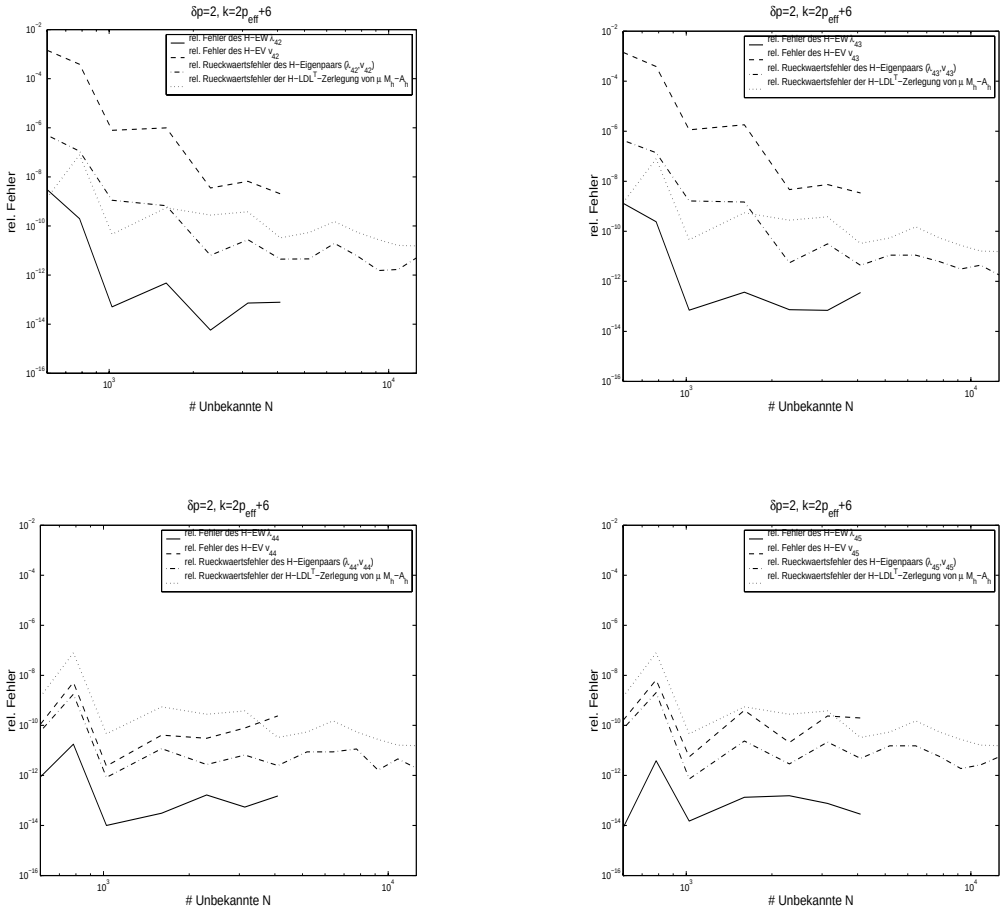

Abbildung 4.8: Relativer Fehler der $\mathcal{H}$-EW λ_{42} (oben links), λ_{43} (oben rechts), λ_{44} (unten links) und λ_{45} (unten rechts) und der zugehoerigen $\mathcal{H}$-EV $\mathbf{v}_{42}$ (oben links), $\mathbf{v}_{43}$ (oben rechts), $\mathbf{v}_{44}$ (unten links) und $\mathbf{v}_{45}$ (unten rechts) sowie relativer Rückwärtsfehler der $\mathcal{H}$-Eigenpaare $(\lambda_{42}, \mathbf{v}_{42})$ (oben links), $(\lambda_{43}, \mathbf{v}_{43})$ (oben rechts), $(\lambda_{44}, \mathbf{v}_{44})$ (unten links) und $(\lambda_{45}, \mathbf{v}_{45})$ (unten rechts) in $\|\cdot\|_2$ und relativer Rückwärtsfehler der $\mathcal{H}$-LDL^T-Zerlegung von $\mu M_h - A_h$ in $\|\cdot\|_F$ für $\delta p = 2$ und Rang $k = 2p_{eff} + 6$.

4.3.3 Singulärwerte und Singulärvektoren des diskreten 2D Laplace-Operators in $\mathcal{H}$-Arithmetik

Analog zu den kleinsten EW und EV von $M_h^{-1}A_h$ lassen sich auch die kleinsten Singulärwerte (SW) σ_j und Singulärvektoren (SV) $\mathbf{u}_j$ und $\mathbf{v}_j$ von $M_h^{-1}A_h = \sum_{j=1}^{N} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$ im Rahmen der $\mathcal{H}$-Genauigkeit berechnen: Die Quadrate der SW σ_j^2 sind nämlich gerade die EW der Matrix $(M_h^{-1}A_h)^T(M_h^{-1}A_h)$, deren EV den SV $\mathbf{v}_j$ von $M_h^{-1}A_h$ entsprechen. Die zugehörigen SV $\mathbf{u}_j$ lassen sich dann wegen $M_h^{-1}A_h\,V = U\,\Sigma$ mit $V = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ und $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$ leicht durch QR-Zerlegung von $M_h^{-1}A_h\,V$ bestimmen.

Daher schlagen wir folgenden Algorithmus zur Berechnung der vorderen SW samt SV der Matrix $M_h^{-1}A_h$ vor:

Algorithmus 4.4 (SW und SV in $\mathcal{H}$-Arithmetik)

1. *Berechne die $\mathcal{H}$-Matrizen $M_h^{-1}A_h$ und $(M_h^{-1}A_h)^T(M_h^{-1}A_h)$ durch eine $\mathcal{H}$-Invertierung und zwei $\mathcal{H}$-Multiplikationen.*

2. *Berechne die Eigenwerte λ_j von $(M_h^{-1}A_h)^T(M_h^{-1}A_h)$ samt zugehörigen Eigenvektoren $\mathbf{v}_j$ mittels simultaner Iteration in $\mathcal{H}$-Arithmetik analog zu Algorithmus 4.3 mit der $\mathcal{H}$-Matrix $(M_h^{-1}A_h)^T(M_h^{-1}A_h)$ anstelle von $M_h^{-1}A_h$.*
   Dann sind $\sigma_j = \sqrt{\lambda_j}$ die gesuchten SW und $\mathbf{v}_j$ SV zu σ_j.

3. *Bilde das $\mathcal{H}$-Matrix-Vektor-Produkt $M_h^{-1}A_h\,V$, wobei die Spalten von V aus den in 2. berechneten SV $\mathbf{v}_j$ bestehen. Dann liefert QR-Zerlegung von $M_h^{-1}A_h\,V$ die gesuchten SV $\mathbf{u}_j$.*

Zur Berechnung der EW und EV von $(M_h^{-1}A_h)^T(M_h^{-1}A_h)$ verwenden wir als Shift μ die Quadrate der Näherungen für die EW von $M_h^{-1}A_h$ aus (4.49).

Die bei der EW-Berechnung von $(M_h^{-1}A_h)^T(M_h^{-1}A_h)$ auftretenden LGS

$$\left((M_h^{-1}A_h)^T(M_h^{-1}A_h) - \mu I\right) Z^{(k)} = Q^{(k-1)}$$

und

$$\left((M_h^{-1}A_h)^T(M_h^{-1}A_h) - \mu I\right) C^{(k)} = Q^{(k)}$$

lösen wir wiederum durch einmalige $\mathcal{H}$-LDL^T-Zerlegung der symmetrischen Matrix $\left((M_h^{-1}A_h)^T(M_h^{-1}A_h) - \mu I\right)$ in $\mathcal{O}(N \log^2 N)$ flops mit anschließender Vorwärts- und Rückwärtssubstitution in $\mathcal{O}(N \log N)$ flops.

Der Shift μ lässt sich nun wie vorhin zur Verringerung der großen Konditionszahlen von $\left((M_h^{-1}A_h)^T(M_h^{-1}A_h) - \mu I\right)$ gezielt steuern. Um jedoch dieselbe relative Genauigkeit wie für die $\mathcal{H}$-EW und $\mathcal{H}$-EV von $M_h^{-1}A_h$ zu erreichen, muss wegen

$$\kappa((M_h^{-1}A_h)^T(M_h^{-1}A_h)) \lessapprox \kappa(M_h^{-1}A_h)^2 \sim N^2$$

der Rang der Rk-Blöcke in den $\mathcal{H}$-Matrizen deutlich vergrößert werden.

4.3.4 Rückwärtsfehleranalyse für die $\mathcal{H}$-Singulärwerte und $\mathcal{H}$-Singulärvektoren

Seien σ ein fehlerbehafteter $\mathcal{H}$-SW von $M_h^{-1}A_h$ und $\mathbf{u}$, $\mathbf{v}$ mit $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$ die zugehörigen $\mathcal{H}$-SV aus Algorithmus 4.4. Dann gilt für die mit dem Residuum $\mathbf{r} = \sigma\mathbf{u} - M_h^{-1}A_h\mathbf{v}$ gebildete Rang-1-Matrix $\Delta(M_h^{-1}A_h) = \mathbf{r}\mathbf{v}^T$

$$(M_h^{-1}A_h + \Delta(M_h^{-1}A_h))\mathbf{v} = \sigma\mathbf{u},$$

und der Rückwärtsfehler $\Delta(M_h^{-1}A_h)$ erfüllt

$$\|\Delta(M_h^{-1}A_h)\|_2 = \max_{\|\mathbf{w}\|_2=1} \|\Delta(M_h^{-1}A_h)\mathbf{w}\|_2 = \max_{\|\mathbf{w}\|_2=1} \|\mathbf{r}\mathbf{v}^T\mathbf{w}\|_2 = \|\mathbf{r}\|_2.$$

Damit ergibt sich der relative Rückwärtsfehler in der Spektralnorm zu

$$\frac{\|\Delta(M_h^{-1}A_h)\|_2}{\|M_h^{-1}A_h\|_2} = \frac{\|\mathbf{r}\|_2}{\|M_h^{-1}A_h\|_2}. \tag{4.52}$$

In der numerischen Praxis approximieren wir $\|M_h^{-1}A_h\|_2$ wiederum durch $24N$ (siehe Abschnitt 4.2.2.1).

4.4 Niedrigrang-Approximationen bestimmter Matrixfunktionen in $\mathcal{H}$-Arithmetik

Mit der im letzten Abschnitt entwickelten $\mathcal{H}$-Methode zur Berechnung von EW und EV von $M_h^{-1}A_h$ gelangen wir nun endlich zum erklärten Ziel dieses Kapitels, der Konstruktion von Niedrigrang-Approximationen an geeignet gewichtete Matrixfunktionen unter Verwendung von $\mathcal{H}$-Matrizen. Aufwand und Approximationsgüte der im Folgenden konstruierten Näherungen an die gewichteten Matrixfunktionen $\cos(\tau c\sqrt{M_h^{-1}A_h})\,(M_h^{-1}A_h)^{-q}$, $\sigma(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ und $\psi(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ sind dabei völlig unabhängig vom Produkt τc!

4.4.1 Konstruktion der Niedrigrang-Approximationen

Seien $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ und $V = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$ die EW, normierten LEV und normierten REV von $M_h^{-1}A_h$, d.h. es gelte

$$M_h^{-1}A_h\,V = V\,D \tag{4.53}$$

und

$$U^T\,M_h^{-1}A_h = D\,U^T. \tag{4.54}$$

Dann folgt aus (4.53)

$$f(\tau^2c^2M_h^{-1}A_h)\,V = V\,f(\tau^2c^2D)$$

für alle in diesem Kapitel behandelten transzendenten Matrixfunktionen f. Aus

$$V^{-1}\,f(\tau^2c^2M_h^{-1}A_h) = f(\tau^2c^2D)\,V^{-1}$$

folgt weiter, dass die Zeilen von V^{-1} bis auf Skalierung durch die LEV $\mathbf{u}_j^T$ gegeben sind:

$$V^{-1} = \begin{bmatrix} c_1\mathbf{u}_1^T \\ \vdots \\ c_N\mathbf{u}_N^T \end{bmatrix}$$

Aus

$$I = V^{-1}V = \begin{bmatrix} c_1\mathbf{u}_1^T\mathbf{v}_1 & & \\ & \ddots & \\ & & c_N\mathbf{u}_N^T\mathbf{v}_N \end{bmatrix}$$

folgt $c_j = \frac{1}{\mathbf{u}_j^T\mathbf{v}_j}, 1 \leq j \leq N$. Damit gilt

$$f(\tau^2c^2M_h^{-1}A_h) = \sum_{j=1}^{N} f(\tau^2c^2\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}, \tag{4.55}$$

und

$$\sum_{j=1}^{J} f(\tau^2c^2\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j} \tag{4.56}$$

stellt eine Rang-J-Approximation an $f(\tau^2c^2M_h^{-1}A_h)$ dar.

Mit den mittels Algorithmus 4.3 rückwärtsstabil berechneten EW λ_j von $M_h^{-1}A_h$ samt den normierten REV $\mathbf{v}_j$ und LEV $\mathbf{u}_j = \frac{M_h\mathbf{v}_j}{\|M_h\mathbf{v}_j\|_2}$ ergibt sich nun der folgende Algorithmus zur Berechnung der Rang-J-Approximation (4.56) an die Matrixfunktion $f(\tau^2c^2M_h^{-1}A_h)$:

Algorithmus 4.5 (Rang-J-Approximation an Matrixfunktionen)

1. *Berechne die J kleinsten Eigenwerte λ_j von $M_h^{-1}A_h$ sowie die zugehörigen normierten Rechts-Eigenvektoren $\mathbf{v}_j$ in $\mathcal{H}$-Arithmetik mit Algorithmus 4.3.*

2. *Berechne die normierten Links-Eigenvektoren $\mathbf{u}_j$ durch Multiplikation der $\mathbf{v}_j$ mit M_h und anschließende Normierung.*

3. *Stelle die Rang-J-Matrix $\sum_{j=1}^{J} f(\tau^2c^2\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}$ auf.*

4.4.2 Fehlerschätzung der Niedrigrang-Approximationen

$$\begin{aligned} & \|f(\tau^2c^2M_h^{-1}A_h) - \textstyle\sum_{j=1}^{J} f(\tau^2c^2\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}\|_F \\ = \; & \|\textstyle\sum_{j=J+1}^{N} f(\tau^2c^2\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}\|_F \\ \leq \; & \textstyle\sum_{j=J+1}^{N} |f(\tau^2c^2\lambda_j)| \cdot \|\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}\|_F \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=J+1}^{N} |f(\tau^2 c^2 \lambda_j)| \cdot \tfrac{1}{|\mathbf{v}_j^T \mathbf{u}_j|} \cdot \sqrt{\sum_{i=1}^{N} \Big(\mathbf{u}_j \underbrace{\mathbf{v}_j^T \mathbf{v}_j}_{=1} \mathbf{u}_j^T \Big)_{ii}} \\
&= \sum_{j=J+1}^{N} |f(\tau^2 c^2 \lambda_j)| \cdot \tfrac{1}{|\mathbf{v}_j^T \mathbf{u}_j|} \cdot \underbrace{\|\mathbf{u}_j\|_2}_{=1} \\
&= \sum_{j=J+1}^{N} |f(\tau^2 c^2 \lambda_j)| \cdot \underbrace{\frac{1}{|\mathbf{v}_j^T \mathbf{u}_j|}}_{=\kappa(\lambda_j)}
\end{aligned}
$$

Dabei bezeichnet $\kappa(\lambda_j)$ die Kondition des EW λ_j von $M_h^{-1} A_h$. Nachdem für die Konditionszahl eines jeden EW λ_j von $M_h^{-1} A_h$ $\kappa(\lambda_j) \to 1$ für $h \to 0$ gilt, stellt $\sum_{j=J+1}^{N} |f(\tau^2 c^2 \lambda_j)|$ für betragsmäßig hinreichend schnell fallendes f eine gute Approximation an die Summe der abgeschnittenen SW von $f(\tau^2 c^2 M_h^{-1} A_h)$ dar: Die berechneten EW $f(\tau^2 c^2 \lambda_j)$ von $f(\tau^2 c^2 M_h^{-1} A_h)$ konvergieren nämlich für $h \to 0$ betragsmäßig gegen die SW von $f(\tau^2 c^2 M_h^{-1} A_h)$.

Die numerische Praxis hat nun gezeigt, dass die bereits gute Fehlerschätzung

$$
\sum_{j=J+1}^{N} |f(\tau^2 c^2 \lambda_j)| = \| \begin{pmatrix} f(\tau^2 c^2 \lambda_{J+1}) \\ \vdots \\ f(\tau^2 c^2 \lambda_N) \end{pmatrix} \|_1 \tag{4.57}
$$

durch Ersetzung von $\|\cdot\|_1$ durch $\|\cdot\|_2$ noch wesentlich verbessert wird: Dann liegt nämlich wirklich die Frobeniusnorm des Fehlers als euklidische Norm des Vektors der approximierten abgeschnittenen SW vor. Man beachte, dass der Verlust von $\|\cdot\|_2$ in der Herleitung obiger Fehlerschätzung durch die Anwendung der Dreiecksungleichung verursacht wird. Da die EV $\mathbf{v}_j$ von $M_h^{-1} A_h$ nicht notwendigerweise exakt orthogonal zueinander sind, ist ihre Anwendung jedoch unumgänglich.

Beweis für $\kappa(\lambda_j) \underset{h \to 0}{\longrightarrow} 1$:
Mit dem normierten REV $\mathbf{v}_j$ zum EW λ_j ist $\mathbf{u}_j = \frac{M_h \mathbf{v}_j}{\|M_h \mathbf{v}_j\|_2}$ normierter LEV zu λ_j. Somit gilt

$$
\kappa(\lambda_j) = \frac{1}{|\mathbf{v}_j^T \mathbf{u}_j|} = \frac{\|M_h \mathbf{v}_j\|_2}{|\mathbf{v}_j^T M_h \mathbf{v}_j|}. \tag{4.58}
$$

In Abschnitt 4.2.2.1 wurden die EW $\tilde{\lambda}_{kl}$ von $M_h^{-1} A_h$ approximativ bis auf $\mathcal{O}(f(k,l)h^2)$ zu

$$
\tilde{\lambda}_{kl} = \frac{1}{h^2} \, \lambda_{kl} \, (1 + \mathcal{O}(f(k,l)h^2))
$$

bestimmt mit den EW λ_{kl} von A_h. Für die REV $\tilde{\mathbf{v}}_{kl}$ zu den EW $\tilde{\lambda}_{kl}$ von $M_h^{-1} A_h$ ergab sich

$$
\tilde{\mathbf{v}}_{kl} = \mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)
$$

mit den normierten REV $\mathbf{v}_{kl}$ zu den EW λ_{kl} von A_h.

Nun entsprechen hier λ_j und $\mathbf{v}_j$ gerade einem $\tilde{\lambda}_{kl}$ und einem $\frac{\tilde{\mathbf{v}}_{kl}}{\|\tilde{\mathbf{v}}_{kl}\|_2}$, und wegen

$M_h\mathbf{v}_j = \frac{1}{\lambda_j} A_h\mathbf{v}_j$ und $\|\tilde{\mathbf{v}}_{kl}\|_2 = 1 + \mathcal{O}(f(k,l)h^2)$ gilt somit

$$\begin{aligned} M_h\mathbf{v}_j &= \tfrac{1}{\lambda_j} A_h\mathbf{v}_j = \tfrac{1}{\tilde{\lambda}_{kl}} A_h \left(\mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)\right) \cdot (1 + \mathcal{O}(f(k,l)h^2)) \\ &= \tfrac{1}{\lambda_{kl}}\, h^2 \left(1 + \mathcal{O}(f(k,l)h^2)\right) \cdot \left(\lambda_{kl}\mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)\right) \\ &= h^2 \left(1 + \mathcal{O}(f(k,l)h^2)\right) \cdot \left(\mathbf{v}_{kl} + \tfrac{1}{\lambda_{kl}}\, \mathcal{O}(\mathbf{f}(k,l)h^2)\right) \\ &= h^2 \left(\mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)\right), \end{aligned}$$

wobei in der zweiten Zeile die Eigenwertgleichung $A_h\mathbf{v}_{kl} = \lambda_{kl}\mathbf{v}_{kl}$ und die Beschränktheit von $\|A_h\|_2 \leq 8$ unabhängig von h benutzt wurden.

Mit $\|\mathbf{v}_{kl}\|_2 = 1$ ergibt sich schließlich der Zähler von (4.58) zu

$$\|M_h\mathbf{v}_j\|_2 = h^2 \|\mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)\|_2 = h^2 \left(1 + \mathcal{O}(f(k,l)h^2)\right).$$

Für den Nenner von (4.58) gilt

$$\begin{aligned} & |\mathbf{v}_j^T M_h \mathbf{v}_j| \\ = \;& |\left(1 + \mathcal{O}(f(k,l)h^2)\right) \left(\mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)\right)^T \cdot h^2 \left(\mathbf{v}_{kl} + \mathcal{O}(\mathbf{f}(k,l)h^2)\right)| \\ = \;& h^2 \left(1 + \mathcal{O}(f(k,l)h^2)\right). \end{aligned}$$

Insgesamt folgt also wie behauptet

$$\kappa(\lambda_j) = \frac{\|M_h\mathbf{v}_j\|_2}{|\mathbf{v}_j^T M_h \mathbf{v}_j|} = \frac{h^2 \left(1 + \mathcal{O}(f(k,l)h^2)\right)}{h^2 \left(1 + \mathcal{O}(f(k,l)h^2)\right)} \underset{h\to 0}{\longrightarrow} 1.$$

□

N	$\kappa(\lambda_1)$
16^2	1.0000123333
24^2	1.0000029010
32^2	1.0000010006
40^2	1.0000004315
48^2	1.0000002153
56^2	1.0000001190
64^2	1.0000000710

Tabelle 4.1: $\kappa(\lambda_1) \to 1$ für $h \to 0$.

Unter Verwendung der Näherungen (4.49) für die EW von $M_h^{-1}A_h$ zur Berechnung der Werte $f_{kl} = f(\tau^2 c^2 \mu_{kl})$ erhalten wir also die hervorragende Näherung

$$\epsilon_{rel} = \sqrt{\frac{\sum'_{k,l} f_{kl}^2}{\sum_{k,l} f_{kl}^2}} \tag{4.59}$$

für den relativen Fehler der Rk-Approximation (4.56) in der Frobeniusnorm. Dabei werden in der Summe $\sum'$ im Zähler nur die Quadrate der abgeschnittenen $N - J$ approximierten Singulärwerte aufsummiert, während die Summe $\sum$

im Nenner über alle N Werte f_{kl}^2 läuft.

In Abbildung 4.9 sind nun die numerischen Ergebnisse für die Rk-Approximationen an die gewichteten Matrixfunktionen $\cos_q(\tau c\sqrt{M_h^{-1}A_h}) := \cos(\tau c\sqrt{M_h^{-1}A_h})\,(M_h^{-1}A_h)^{-q}$ (jeweils links) und $\sigma_q(\tau^2c^2M_h^{-1}A_h) := \sigma(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ (jeweils rechts) für $q = 1, 3$ und 5 dargestellt.

Die mit den Schätzungen (4.49) für die EW von $M_h^{-1}A_h$ konstruierten Fehlerschätzer (4.59) stimmen mit den exakten Fehlern der mit den $\mathcal{H}$-EW, $\mathcal{H}$-LEV und $\mathcal{H}$-REV konstruierten Rk-Approximationen (4.56) perfekt überein und liegen nur geringfügig über dem Fehler der mit den exakten SW und SV gebildeten Best-Rk-Approximationen.

Wären die $\mathcal{H}$-EW und $\mathcal{H}$-EV nur weniger genau bestimmt worden (durch kleineren Rang der Rk-Blöcke in den $\mathcal{H}$-Matrizen), so wäre der exakte Fehler der Rk-Approximationen (4.56) bei der $\mathcal{H}$-Ungenauigkeit stehengeblieben, während die anderen beiden Linien unverändert geblieben wären.

Fazit:

1. Die Rk-Approximation (4.56) stellt eine approximative Singulärwertzerlegung der Matrixfunktion $f(\tau^2c^2M_h^{-1}A_h)$ dar.

2. Dabei ist nur eine $\mathcal{H}$-LDL^T-Zerlegung pro EW der Vielfachheit m zu berechnen (siehe Abschnitt 4.3.1) und somit die gesamte Rang-J-Approximation mit einem Aufwand von $\mathcal{O}(JN\log^2 N)$ flops realisierbar.

3. Aufwand und Güte der Rk-Approximationen an die gewichteten Matrixfunktionen $\cos_q(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma_q(\tau^2c^2M_h^{-1}A_h)$ und $\psi_q(\tau^2c^2M_h^{-1}A_h)$ sind völlig unabhängig vom Produkt τc (siehe (4.56) und (4.57))! Die relative Genauigkeit hängt einzig und allein von der Anzahl J der beibehaltenen größten EW von $\cos_q(\tau c\sqrt{M_h^{-1}A_h})$, $\sigma_q(\tau^2c^2M_h^{-1}A_h)$ und $\psi_q(\tau^2c^2M_h^{-1}A_h)$ ab (siehe (4.57)).

4. Will man die Genauigkeit in (4.56) erhöhen, können an die bereits bestehende Rk-Approximation die mit ein paar weiteren Eigenpaaren gebildeten Rang-1-Matrizen einfach hinzugefügt werden.

5. Die Auswertung der Rk-Approximation (4.56) an $f(\tau^2c^2M_h^{-1}A_h)$ für verschiedene Werte von τ und c erfordert überhaupt keinen zusätzlichen $\mathcal{H}$-Aufwand.

6. Die gewichteten Matrixfunktionen $\sigma_q(\tau^2c^2M_h^{-1}A_h)$, $\psi_q(\tau^2c^2M_h^{-1}A_h)$, $\cos_q(\tau c\sqrt{M_h^{-1}A_h})$, $\sin_q(\tau c\sqrt{M_h^{-1}A_h})$, ... sind alle simultan als Rk-Matrizen gemäß (4.56) berechenbar. Dabei sind die Eigenpaar-Berechnungen in $\mathcal{H}$-Arithmetik nur einmal durchzuführen!

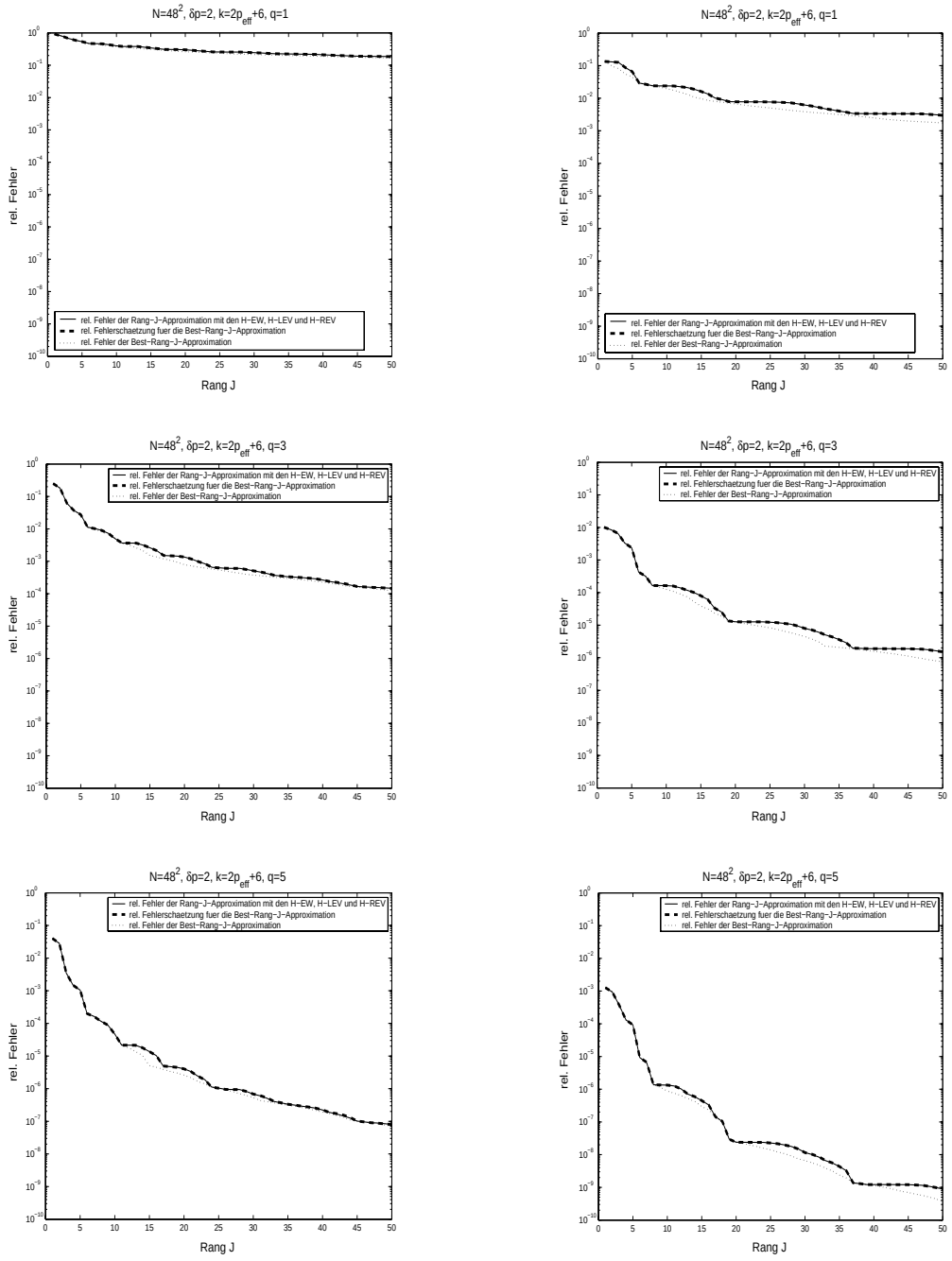

Abbildung 4.9: Relativer Fehler der Rang-J-Approximation an die Matrixfunktionen $\cos_q(\tau c\sqrt{M_h^{-1}A_h})$ (links) und $\sigma_q(\tau^2 c^2 M_h^{-1}A_h)$ (rechts) mit Hilfe der $\mathcal{H}$-EW, $\mathcal{H}$-Links-EV und $\mathcal{H}$-Rechts-EV sowie Schätzung des relativen Fehlers und exakter relativer Fehler der Best-Rang-J-Approximation an dieselben Matrixfunktionen in $\|\cdot\|_F$ für $N = 48^2$, $\delta p = 2$, $k = 2p_{eff} + 6$ und $q = 1$ (oben), $q = 3$ (Mitte) und $q = 5$ (unten).

4.5 Niedrigrang-Approximation an die Exponentialfunktion des 2D Laplace-Operators

Analog zu den gewichteten transzendenten Matrixfunktionen aus dem vorhergehenden Abschnitt lässt sich auch die den Lösungsoperator der Wärmeleitungsgleichung darstellende Matrixfunktion $e^{-tM_h^{-1}A_h}$ für hinreichend großes $t > 0$ durch Algorithmus 4.5 als Rk-Matrix darstellen. Für $t = 1$ reicht beispielsweise bereits die Rang-1-Approximation

$$e^{-\lambda_1} \frac{\mathbf{v}_1 \mathbf{u}_1^T}{\mathbf{v}_1^T \mathbf{u}_1} \tag{4.60}$$

an $\exp(-M_h^{-1}A_h)$, denn es gilt

$$\frac{e^{-\lambda_2}}{e^{-\lambda_1}} = \frac{e^{-5\pi^2}}{e^{-2\pi^2}} \approx 1.4e-13$$

(siehe Abbildung 4.10). Hier ist der relative Fehler der mit dem $\mathcal{H}$-EW λ_1, dem $\mathcal{H}$-REV $\mathbf{v}_1$ und dem $\mathcal{H}$-LEV $\mathbf{u}_1$ von $M_h^{-1}A_h$ gebildeten Rang-1-Approximation tatsächlich durch die $\mathcal{H}$-Fehler von λ_1, $\mathbf{v}_1$ und $\mathbf{u}_1$ gegeben! (Die Fehlerschätzung ist wiederum nach (4.59) gebildet.)

Während nach [14] 40 $\mathcal{H}$-Inversionen zum Erreichen einer relativen Approximationsgüte von 10^{-7} für $\exp(-M_h^{-1}A_h)$ durchgeführt werden müssen (siehe [14, Table 5]), reicht mit unserem Zugang bereits eine $\mathcal{H}$-LDL^T-Zerlegung für eine relative Genauigkeit von 10^{-10} (siehe Abbildung 4.10).
Da der Aufwand für eine $\mathcal{H}$-Invertierung um einen Faktor 7 bis 10 höher liegt als der für eine $\mathcal{H}$-LDL^T-Zerlegung (vgl. die Tabellen 3.4 und 3.6), ist unsere Methode zur Approximation von $\exp(-M_h^{-1}A_h)$ um einen Faktor 300 bis 400 schneller als die in [14] angewendete.

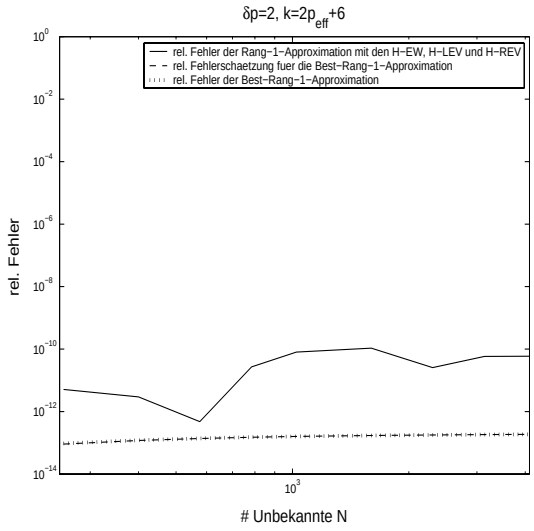

Abbildung 4.10: Relativer Fehler der Rang-1-Approximation an die Matrixfunktion $\exp(-tM_h^{-1}A_h)$ mit Hilfe der $\mathcal{H}$-EW, $\mathcal{H}$-LEV sowie $\mathcal{H}$-REV sowie Schätzung des relativen Fehlers und exakter relativer Fehler der Best-Rang-1-Approximation an dieselbe Matrixfunktion in $\|\cdot\|_F$ für $\delta p = 2$ und $k = 2p_{eff}+6$.

Kapitel 5

Lösung der 2D Wellengleichung in $\mathcal{H}$-Arithmetik

Die hochfrequente 2D Wellengleichung wird in Abschnitt 5.1 durch Spektralzerlegung des diskreten Laplace-Operators in ein entkoppeltes System eindimensionaler Wellengleichungen transformiert. Aus der Berechnung einer bestimmten Anzahl von $\mathcal{H}$-Eigenpaaren des diskreten Laplace-Operators mit einer in Abschnitt 4.3 entwickelten $\mathcal{H}$-Methode und der Lösung der zugehörigen eindimensionalen Wellengleichungen resultieren für genügend glatte Anfangsdaten im Ort gute Approximationen an die Lösung der 2D hochfrequenten Wellengleichung.
Der fast lineare Aufwand und die Approximationsgüte sind dabei unabhängig vom Produkt aus der Zeitschrittweite mit der Wellengeschwindigkeit. Allein die Glattheit der Anfangsdaten und der Inhomogenität ist für die Approximationsgüte der $\mathcal{H}$-Lösungen verantwortlich.

Nach der in Abschnitt 5.2 durchgeführten Konditionsanalyse für die eindimensionalen Drei-Term-Rekursionen des Gautschi-Algorithmus wird in Abschnitt 5.3 die 2D Wärmeleitungsgleichung analog zur Wellengleichung gelöst. Für hinreichend große Zeiten ist die berechnete Lösung völlig unabhängig von der Glattheit der Anfangsdaten im Ort. Das in der Lösungsformel für die inhomogene Wärmeleitungsgleichung auftretende Integral wird mit einer geeigneten Quadraturformel numerisch berechnet. Zum Abschluß dieses Kapitels folgt schließlich eine Diskussion über die Lösung der Schrödingergleichung.

5.1 Die diskrete Lösung der 2D Wellengleichung

Gegeben seien die Anfangswerte $\mathbf{u}_0$ und $\dot{\mathbf{u}}_0$ sowie die Inhomogenität $\mathbf{f}(t)$.
Zur Berechnung der diskreten Lösung $\mathbf{u}_n$ und deren Ableitung $\dot{\mathbf{u}}_n$ sind die inhomogenen Drei-Term-Rekursionen

$$\mathbf{u}_{n+1} = 2\mathbf{u}_n - \mathbf{u}_{n-1} + \tau^2\sigma(\tau^2 c^2 M_h^{-1} A_h)(-c^2 M_h^{-1} A_h \mathbf{u}_n + \mathbf{f}_n), \;\; n \geq 1 \qquad (5.1)$$

mit den Startwerten $\mathbf{u}_0$ und

$$\mathbf{u}_1 = \cos(\tau c\sqrt{M_h^{-1}A_h})\,\mathbf{u}_0 + \tau\psi(\tau^2c^2M_h^{-1}A_h)\,\dot{\mathbf{u}}_0 + \frac{1}{2}\tau^2\sigma(\tau^2c^2M_h^{-1}A_h)\,\mathbf{f}_0$$

sowie

$$\dot{\mathbf{u}}_{n+1} = \dot{\mathbf{u}}_{n-1} + 2\tau\psi(\tau^2c^2M_h^{-1}A_h)(-c^2M_h^{-1}A_h\mathbf{u}_n + \mathbf{f}_n),\ \ n \geq 1 \tag{5.2}$$

mit den Startwerten $\dot{\mathbf{u}}_0$ und

$$\dot{\mathbf{u}}_1 = \cos(\tau c\sqrt{M_h^{-1}A_h})\,\dot{\mathbf{u}}_0 + \tau\psi(\tau^2c^2M_h^{-1}A_h)(-c^2M_h^{-1}A_h\mathbf{u}_0 + \mathbf{f}_0)$$

zu lösen (siehe Abschnitt 2.5).

Die Matrix-Vektor-Produkte $f(\tau^2c^2M_h^{-1}A_h)\,\mathbf{v}$ mit obigen transzendenten Matrixfunktionen f und N-dimensionalen Vektoren $\mathbf{v}$ schreiben wir jetzt in der Form $f_q(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^q\,\mathbf{v}$. Dabei approximieren wir die gewichteten Matrixfunktionen $f_q(\tau^2c^2M_h^{-1}A_h) := f(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^{-q}$ mittels Algorithmus 4.5 durch Rk-Matrizen. $M_h^{-1}A_h$ wird als $\mathcal{H}$-Matrix dargestellt und die je $q+1$ Matrix-Vektor-Produkte sukzessive von rechts nach links abgearbeitet.

Numerische Experimente haben nun gezeigt, dass der relative Fehler der Lösung $\mathbf{u}_n$ unabhängig vom gewählten $q \in \mathbb{N}_0$ ist. Der Grund dafür ist die Abhängigkeit der Approximationsgüte von $\mathbf{u}_n$ allein von der Anzahl der beibehaltenen Eigenwerte bei den gewichteten transzendenten Matrixfunktionen. Die Wahl von q spielt dabei überhaupt keine Rolle.

Dies wollen wir nun anhand des Matrixfunktion-Vektor-Produkts $f(\tau^2c^2M_h^{-1}A_h)\,\mathbf{v}$ auch beweisen:

Satz 5.1 (Unabhängigkeit der Approximationsgüte von q)
Sei

$$f_{q,J}(\tau^2c^2M_h^{-1}A_h) = \sum_{j=1}^{J}\frac{f(\tau^2c^2\lambda_j)}{\lambda_j^q}\,\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}$$

die nach Algorithmus 4.5 berechnete Rang-J-Approximation an die gewichtete Matrixfunktion $f_q(\tau^2c^2M_h^{-1}A_h)$. Dann ist die Güte der Näherung

$$f_{q,J}(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^q\,\mathbf{v}$$

an $f(\tau^2c^2M_h^{-1}A_h)\,\mathbf{v}$ unabhängig von q.

Beweis: Sei $M_h^{-1}A_h = VDV^{-1}$ mit den EW $D = \mathrm{diag}(\lambda_1,\ldots,\lambda_N)$ und den EV $V = [\mathbf{v}_1,\ldots,\mathbf{v}_N]$ von $M_h^{-1}A_h$. Dann gilt mit $\mathbf{v} = \sum_{i=1}^N \rho_i\mathbf{v}_i$

$$\begin{aligned}
&f_{q,J}(\tau^2c^2M_h^{-1}A_h)\,(M_h^{-1}A_h)^q\,\mathbf{v} = \\
&\textstyle\sum_{j=1}^J \frac{f(\tau^2c^2\lambda_j)}{\lambda_j^q}\,\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}\,VD^qV^{-1}\,\sum_{i=1}^N\rho_i\mathbf{v}_i = \\
&\textstyle\sum_{j=1}^J \frac{f(\tau^2c^2\lambda_j)}{\lambda_j^q}\,\frac{\mathbf{v}_j(\mathbf{u}_j^T\mathbf{v}_j)\mathbf{e}_j^T}{\mathbf{v}_j^T\mathbf{u}_j}\,\sum_{i=1}^N\lambda_i^q\rho_i\mathbf{e}_i = \\
&\textstyle\sum_{j=1}^J \frac{f(\tau^2c^2\lambda_j)}{\lambda_j^q}\,\mathbf{v}_j\,\lambda_j^q\rho_j = \\
&\textstyle\sum_{j=1}^J \rho_j f(\tau^2c^2\lambda_j)\mathbf{v}_j
\end{aligned} \tag{5.3}$$

Die letzte Summe in (5.3) ist unabhängig von q, woraus unmittelbar die Behauptung folgt. In (5.3) wurde die leicht nachzuweisende Tatsache benutzt, dass ein REV und ein LEV zu paarweise verschiedenen EW stets orthogonal zueinander sind. □

Fazit:

1. Die Approximationsgüte des Matrix-Vektor-Produkts $f(\tau^2 c^2 M_h^{-1} A_h)\,\mathbf{v}$ ist allein durch die Anzahl J und die Größe der Koeffizienten ρ_j und der Eigenwerte $f(\tau^2 c^2 \lambda_j)$ bestimmt (siehe die letzte Summe in (5.3)).

2. Der unseren Bedürfnissen am besten angepasste Kalkül besteht in Spektralzerlegung der Operatoren:

$$\begin{aligned} f(\tau^2 c^2 M_h^{-1} A_h) &= \sum_{\lambda \,\mathrm{EW\ von}\, M_h^{-1} A_h} f(\tau^2 c^2 \lambda)\, E(\lambda) \\ &= \sum_{j=1}^{N} f(\tau^2 c^2 \lambda_j)\, E(\lambda_j) \end{aligned}$$

   Dabei bezeichnet $E(\lambda)$ die Projektion auf den Eigenraum zum Eigenwert λ, d.h. $E(\lambda_j)\mathbf{v} = \rho_j \mathbf{v}_j$, $1 \le j \le N$, für $\mathbf{v} = \sum_{j=1}^{N} \rho_j \mathbf{v}_j$.

3. Die Unterräume $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sind viel geeignetere Kandidaten zur Approximation der Lösung der hochfrequenten Wellengleichung als die Krylovunterräume aus Abschnitt 3.1.

4. Es ist dasselbe, ob ich die Projektion auf den Unterraum $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ auf Seiten der Anfangsdaten oder beim Operator durchführe:
   Neben
   $$\sum_{j=1}^{J} f(\tau^2 c^2 \lambda_j) \tfrac{\mathbf{v}_j \mathbf{u}_j^T}{\mathbf{v}_j^T \mathbf{u}_j} \sum_{i=1}^{N} \rho_i \mathbf{v}_i = \sum_{j=1}^{J} \rho_j f(\tau^2 c^2 \lambda_j) \mathbf{v}_j$$
   ist nämlich auch
   $$\sum_{j=1}^{N} f(\tau^2 c^2 \lambda_j) \tfrac{\mathbf{v}_j \mathbf{u}_j^T}{\mathbf{v}_j^T \mathbf{u}_j} \sum_{i=1}^{J} \rho_i \mathbf{v}_i = \sum_{j=1}^{J} \rho_j f(\tau^2 c^2 \lambda_j) \mathbf{v}_j \,.$$
   Somit kann nach erfolgter Projektion der Anfangsdaten auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ auf die explizite Berechnung der Rang-J-Matrizen gemäß (4.56) vollständig verzichtet werden.

Zur Bestimmung der Koeffizienten ρ_j $(1 \le j \le J)$ eines Vektors $\mathbf{v} \in \mathbb{R}^N$ in der Eigenbasis von $M_h^{-1} A_h$ gehen wir nun folgendermaßen vor: Mit den REV $\mathbf{v}_j$ und den LEV $\mathbf{u}_j = M_h \mathbf{v}_j$ zum EW λ_j von $M_h^{-1} A_h$ sowie $\mathbf{v} = \sum_{i=1}^{N} \rho_i \mathbf{v}_i$ ist

$$\langle \mathbf{v}, \mathbf{v}_j \rangle_{M_h} = \langle \mathbf{v}, \mathbf{u}_j \rangle_2 = \sum_{i=1}^{N} \rho_i \langle \mathbf{v}_i, \mathbf{u}_j \rangle_2 = \rho_j \langle \mathbf{v}_j, \mathbf{u}_j \rangle_2 = \rho_j \langle \mathbf{v}_j, \mathbf{v}_j \rangle_{M_h},\ 1 \le j \le J$$

und damit

$$\rho_j = \frac{\langle \mathbf{v}, \mathbf{v}_j \rangle_{M_h}}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle_{M_h}} \ \forall\, 1 \le j \le J. \tag{5.4}$$

Der Aufwand zur Berechnung der Koeffizienten ρ_j, $1 \le j \le J$, beläuft sich – nach ein für alle Mal erfolgter Berechnung der J Matrix-Vektor-Produkte $M_h \mathbf{v}_j$ und Skalarprodukte $\langle \mathbf{v}_j, M_h \mathbf{v}_j \rangle_2$ – auf die J Skalarprodukte $\langle \mathbf{v}, M_h \mathbf{v}_j \rangle_2$ und beträgt somit $J(2N-1)$ flops.

5.1.1 Der Gautschi-Algorithmus im Frequenzraum

Sind die zu den Anfangsdaten und der Inhomogenität gehörigen Koeffizienten bzgl. der EV $\mathbf{v}_1, \ldots, \mathbf{v}_J$ von $M_h^{-1} A_h$ einmal berechnet, kann der Gautschi-Algorithmus vollständig im Frequenzraum abgewickelt werden.
Im Folgenden meinen wir aus Bequemlichkeitsgründen mit dem Koeffizientenvektor $\rho^{(\cdot)}$ nur dessen erste J Koeffizienten $\rho_j^{(\cdot)}$, $1 \le j \le J$.

Algorithmus 5.2 (Gautschi-Algorithmus im Frequenzraum)

1. *Berechnung der zu den Anfangsdaten und zur Inhomogenität gehörigen Koeffizientenvektoren $\rho^{(\mathbf{u}_0)}$, $\rho^{(\dot{\mathbf{u}}_0)}$ und $\rho^{(\mathbf{f}_0)}$*

2. *Berechnung von $\rho^{(\mathbf{u}_1)}$ und $\rho^{(\dot{\mathbf{u}}_1)}$:*

$$\begin{aligned} \rho_j^{(\mathbf{u}_1)} &= \cos(\tau c \sqrt{\lambda_j})\, \rho_j^{(\mathbf{u}_0)} + \tau \psi(\tau^2 c^2 \lambda_j)\, \rho_j^{(\dot{\mathbf{u}}_0)} + \tfrac{1}{2} \tau^2 \sigma(\tau^2 c^2 \lambda_j)\, \rho_j^{(\mathbf{f}_0)}, \\ \rho_j^{(\dot{\mathbf{u}}_1)} &= \cos(\tau c \sqrt{\lambda_j})\, \rho_j^{(\dot{\mathbf{u}}_0)} + \tau \psi(\tau^2 c^2 \lambda_j)(-c^2 \lambda_j \rho_j^{(\mathbf{u}_0)} + \rho_j^{(\mathbf{f}_0)}) \end{aligned} \tag{5.5}$$

   für $1 \le j \le J$

3. *Berechnung von $\rho^{(\mathbf{f}_n)}, n \ge 1$, und damit Lösung der folgenden Drei-Term-Rekursionen für die Koeffizienten der diskreten Lösung der 2D Wellengleichung und ihrer Ableitung:*

$$\begin{aligned} \rho_j^{(\mathbf{u}_{n+1})} &= 2\rho_j^{(\mathbf{u}_n)} - \rho_j^{(\mathbf{u}_{n-1})} + \tau^2 \sigma(\tau^2 c^2 \lambda_j)(-c^2 \lambda_j \rho_j^{(\mathbf{u}_n)} + \rho_j^{(\mathbf{f}_n)}), \\ \rho_j^{(\dot{\mathbf{u}}_{n+1})} &= \rho_j^{(\dot{\mathbf{u}}_{n-1})} + 2\tau \psi(\tau^2 c^2 \lambda_j)(-c^2 \lambda_j \rho_j^{(\mathbf{u}_n)} + \rho_j^{(\mathbf{f}_n)}) \end{aligned} \tag{5.6}$$

   für $1 \le j \le J$ und $n \ge 1$.

Damit sind die Lösungsvektoren $\mathbf{u}_{n+1}$ und $\dot{\mathbf{u}}_{n+1}$ zum Zeitpunkt $t_{n+1} = (n+1)\tau$ approximativ durch

$$\mathbf{u}_{n+1}^{(J)} = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_{n+1})} \mathbf{v}_j \tag{5.7}$$

und

$$\dot{\mathbf{u}}_{n+1}^{(J)} = \sum_{j=1}^{J} \rho_j^{(\dot{\mathbf{u}}_{n+1})} \mathbf{v}_j \tag{5.8}$$

gegeben.

Man beachte, dass die mittels Algorithmus 5.2 konstruierte Lösung $\mathbf{u}_n^{(J)} = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_n)} \mathbf{v}_j$ mit der Projektion der Lösung des N-dimensionalen Gautschi-Algorithmus auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle \subset \mathbb{R}^N$ übereinstimmt. Dies liegt einfach daran, dass die Entwicklung der zu den einzelnen Frequenzen gehörigen Teilwellen unabhängig voneinander durch die Wellengleichung beschrieben wird:
Mit $\mathbf{u}_h = \sum_{j=1}^{N} \rho_j^{(\mathbf{u}_h)} \mathbf{v}_j$ und $\mathbf{f}_h = \sum_{j=1}^{N} \rho_j^{(\mathbf{f}_h)} \mathbf{v}_j$ gilt nämlich in der Knotenbasis

$$\begin{aligned} &\ddot{\mathbf{u}}_h + c^2 M_h^{-1} A_h \mathbf{u}_h = \mathbf{f}_h \quad \Longleftrightarrow \\ &\textstyle\sum_{j=1}^{N} \left(\ddot{\rho}_j^{(\mathbf{u}_h)} + c^2 M_h^{-1} A_h \rho_j^{(\mathbf{u}_h)} \right) \mathbf{v}_j = \sum_{j=1}^{N} \rho_j^{(\mathbf{f}_h)} \mathbf{v}_j \quad \Longleftrightarrow \\ &\ddot{\rho}_j^{(\mathbf{u}_h)} + c^2 \lambda_j \rho_j^{(\mathbf{u}_h)} = \rho_j^{(\mathbf{f}_h)},\ 1 \le j \le N \end{aligned} \tag{5.9}$$

Fazit:

1. Der ursprüngliche Gautschi-Algorithmus im $\mathbb{R}^N$ zerfällt im Frequenzraum in N eindimensionale Drei-Term-Rekursionen (siehe (5.9)). Dies erleichtert die im nächsten Abschnitt vorzunehmende Konditionsanalyse erheblich.

2. Wir lösen nun J dieser N eindimensionalen Drei-Term-Rekursionen und konstruieren daraus die approximativen Lösungsvektoren (5.7) und (5.8).

3. Aufwand und Approximationsgüte sind unabhängig vom Produkt τc (siehe (5.5) und (5.6)). Es besteht keine Einschränkung von τ durch c^{-1} und h. Die Unabhängigkeit der Approximationsgüte von h ist dadurch gegeben, dass die Eigenvektoren $\mathbf{v}_j = \tilde{\mathbf{v}}_{kl}$ von $M_h^{-1} A_h$ für $h \to 0$ gegen die Funktionen $v_{kl}(x, y) = \sin(k\pi x) \sin(l\pi y)$ auf $\Omega = (0, 1)^2$ und damit auch all die Koeffizienten ρ_j gegen die entsprechenden Grenzwerte konvergieren.

4. Allein die Glattheit der Anfangsdaten u_0 und $\dot{u}_0$ sowie der Inhomogenität f ist – abgesehen vom Diskretisierungsfehler durch den Gautschi-Algorithmus für nicht-konstantes f – entscheidend für die Approximationsgüte der Lösungen $\mathbf{u}_n^{(J)}$ und $\dot{\mathbf{u}}_n^{(J)}$: Je glatter die Anfangsdaten, d.h. je schneller die Koeffizienten $\rho_j^{(\mathbf{u}_0)}$ und $\rho_j^{(\dot{\mathbf{u}}_0)}$ mit wachsendem Index j abklingen, umso weniger EW und EV von $M_h^{-1} A_h$ werden benötigt, um eine bestimmte relative Genauigkeit zu erreichen.
   Mit unserem Verfahren können also niederfrequente Anfangsdaten im Ort bei hohen Frequenzen in der Zeit gut behandelt werden. Dabei ist zu beachten, dass zur Auflösung hochfrequenter Funktionen im Ort ein äußerst feines Gitter notwendig ist, und je kleiner h ist, umso mehr EW und EV von $M_h^{-1} A_h$ sind mit den in Abschnitt 4.2.2.1 bestimmten Näherungen als Shift in $\mathcal{H}$-Arithmetik berechenbar.

5. Liegen nicht-glatte Anfangsdaten vor, können diese auf dem vorgegebenen Gitter durch glatte approximiert werden und daraufhin die Wellengleichung für die approximierten glatten Daten gelöst werden.

6. Nach erfolgter Berechnung der J kleinsten EW und EV von $M_h^{-1}A_h$ lässt sich die Lösung der Wellengleichung für verschiedene Anfangswerte, rechte Seiten, Zeitschrittweiten τ und Wellengeschwindigkeiten c mit $\mathcal{O}(JN)$ flops pro Zeitschritt zur Bestimmung von $\rho^{(\mathbf{f}_n)}$ berechnen (siehe (5.6) und die Aufwandsermittlung für die Berechnung des Koeffizientenvektors $\rho^{(\mathbf{f}_n)}$ im Anschluss an (5.4)).

7. Am relativen Fehler $\frac{\|\mathbf{v}-\sum_{j=1}^{J}\rho_j^{(\mathbf{v})}\mathbf{v}_j\|_{M_h}}{\|\mathbf{v}\|_{M_h}}$ in der durch $\langle\cdot,\cdot\rangle_{M_h}$ induzierten Norm $\|\cdot\|_{M_h}=\sqrt{\langle\cdot,\cdot\rangle_{M_h}}$ lässt sich erkennen, ob ein Anfangsdatum den gestellten Glattheitsanforderungen entspricht, oder ob die Anzahl der Frequenzen im Ort zur besseren Approximation von $\mathbf{v}$ erhöht werden muss.

5.1.2 Numerische Beispiele zur 2D Wellengleichung

Die folgenden numerischen Beispiele bestätigen die soeben gezogenen Schlüsse zur Glattheit der Anfangswerte, wobei die Lösung der 2D Wellengleichung zu gegebenen Anfangsdaten und Inhomogenität jeweils für $c = 10^0$ und $c = 10^5$ berechnet wurde:

Beispiel 1: (siehe die Abbildungen 5.1 bis 5.4)
$u_0(x_1,x_2) = x_1(x_1-1)x_2(x_2-1)$,
$\dot{u}_0(x_1,x_2) = 5\,x_1(x_1-1)x_2(x_2-1)$ und
$f \equiv 0$.

Beispiel 2: (siehe die Abbildungen 5.5 bis 5.8)
$u_0(x_1,x_2) = 0.1\exp(-100\,((0.2-x_1)^2+(0.2-x_2)^2))$,
$\dot{u}_0(x_1,x_2) = x_1(x_1-1)x_2(x_2-1)$ und
$f \equiv 0$.

Beispiel 3: (siehe die Abbildungen 5.9 bis 5.12)
$u_0(x_1,x_2) = \sin(2\pi x_1)\sin(3\pi x_2)$,
$\dot{u}_0(x_1,x_2) = \cos(4\pi x_1)\cos(5\pi x_2)$ und
$f \equiv 10$.

Wegen $f \equiv$ const. liefert der Gautschi-Algorithmus für die Beispiele 1 bis 3 die exakte Lösung der semidiskreten 2D Wellengleichung im FE-Raum V_h (siehe Abschnitt 2.4.2, Eigenschaft 3). Der Diskretisierungsfehler im Ort durch Projektion auf V_h ist in den Abbildungen 5.1, 5.2, 5.5, 5.6, 5.9 und 5.10 durch eine gerade, gestrichelt gepunktete Linie gekennzeichnet, während alle übrigen Linien die algebraischen Fehler der $\mathcal{H}$-Approximationen in $\langle\mathbf{v}_1,\ldots,\mathbf{v}_J\rangle$ an die exakten FE-Lösungen darstellen.

Die Tatsache, dass der relative Diskretisierungsfehler im Ort der exakten FE-Lösung für $c = 10^5$ bei 100% liegt, ist eine Folge der zu c direkt proportionalen Kondition der Lösung der Wellengleichung (siehe Abschnitt 2.3).

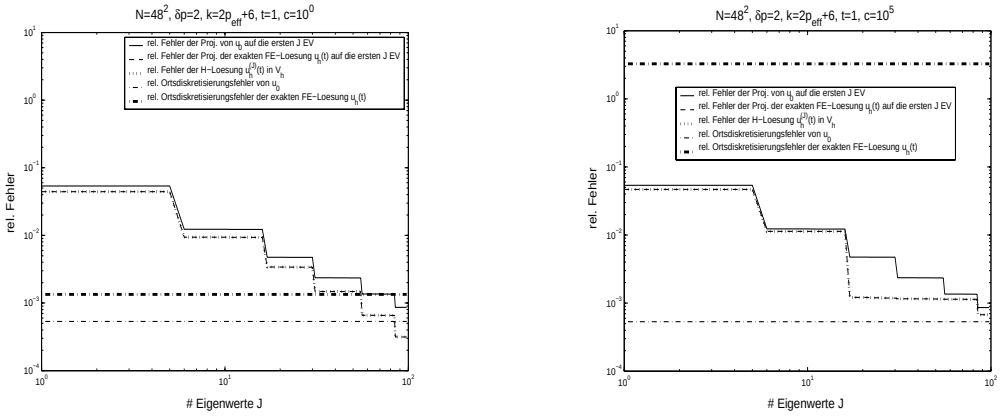

Abbildung 5.1: Relativer algebraischer Fehler der Projektion des Anfangswerts $\mathbf{u}_0$ und der Projektion der exakten FE-Lösung $\mathbf{u}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\mathbf{u}_0$ und der exakten FE-Lösung $\mathbf{u}_h(t)$ zum Zeitpunkt $t = 1$ in $\|\cdot\|_{L^2}$ für $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ und $\dot{u}_0(x_1, x_2) = 5\, x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

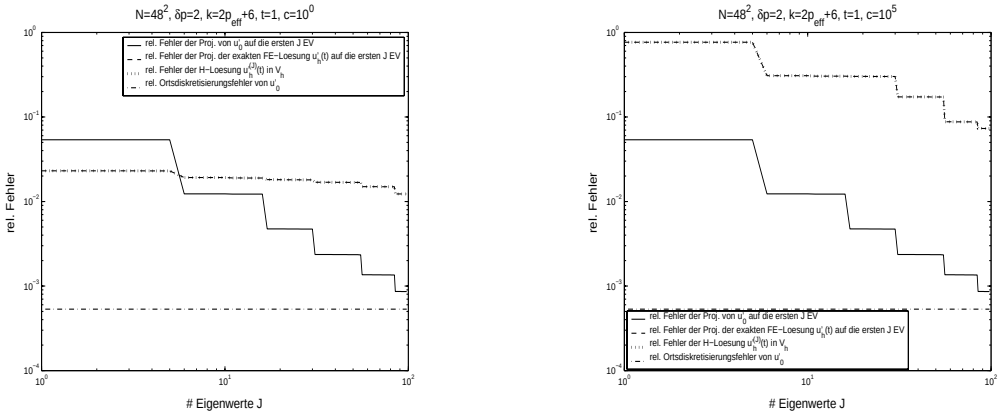

Abbildung 5.2: Relativer algebraischer Fehler der Projektion des Anfangswerts $\dot{\mathbf{u}}_0$ und der Projektion der exakten FE-Lösung $\dot{\mathbf{u}}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\dot{\mathbf{u}}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\dot{\mathbf{u}}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\dot{\mathbf{u}}_0$ zum Zeitpunkt $t = 1$ in $\|\cdot\|_{L^2}$ für $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ und $\dot{u}_0(x_1, x_2) = 5\, x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

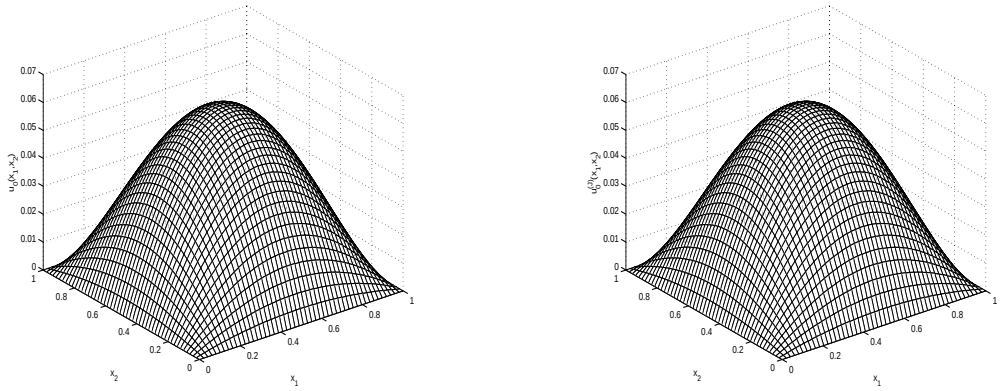

Abbildung 5.3: Koeffizientenvektor $\mathbf{u}_0$ der L^2-Projektion u_{0h} des Anfangswerts $u_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ auf V_h (links) und Projektion von $\mathbf{u}_0$ auf die EV zu den $J = 98$ kleinsten EW von $M_h^{-1}A_h$ (rechts).

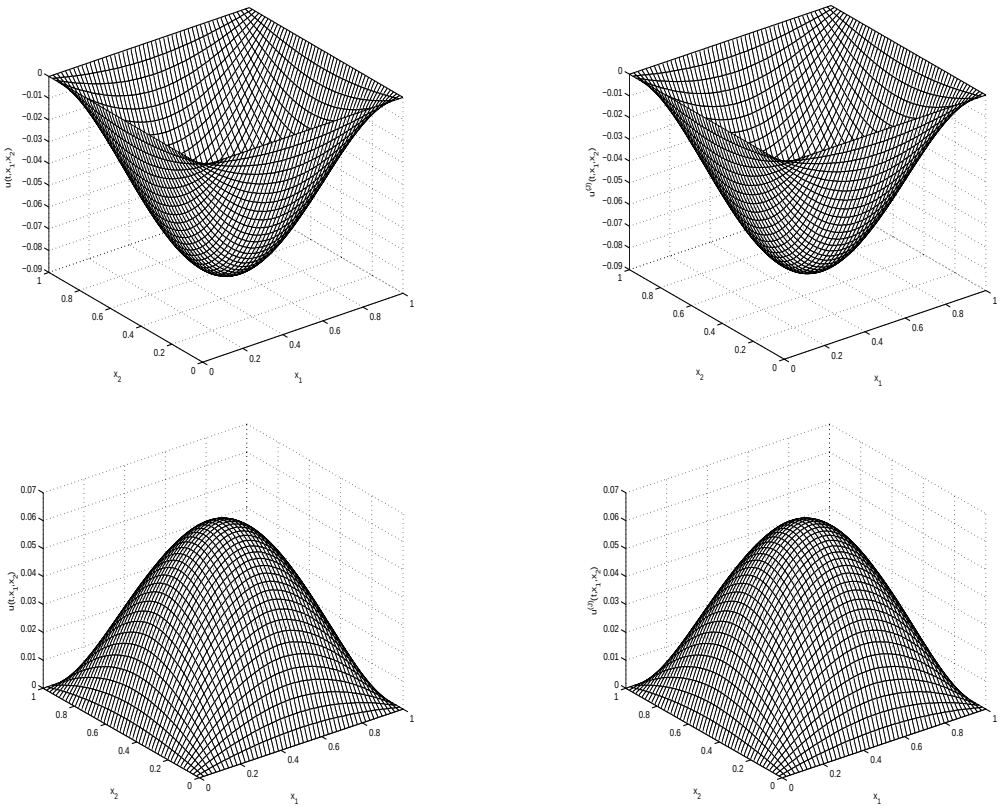

Abbildung 5.4: Exakte FE-Lösung $\mathbf{u}_h(t)$ für $c = 10^0$ (links oben) und $c = 10^5$ (links unten) sowie $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ mit $J = 98$ in V_h für $c = 10^0$ (rechts oben) und $c = 10^5$ (rechts unten) zum Zeitpunkt $t = 1$ zu den Anfangsdaten $u_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ und $\dot{u}_0(x_1, x_2) = 5\, x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

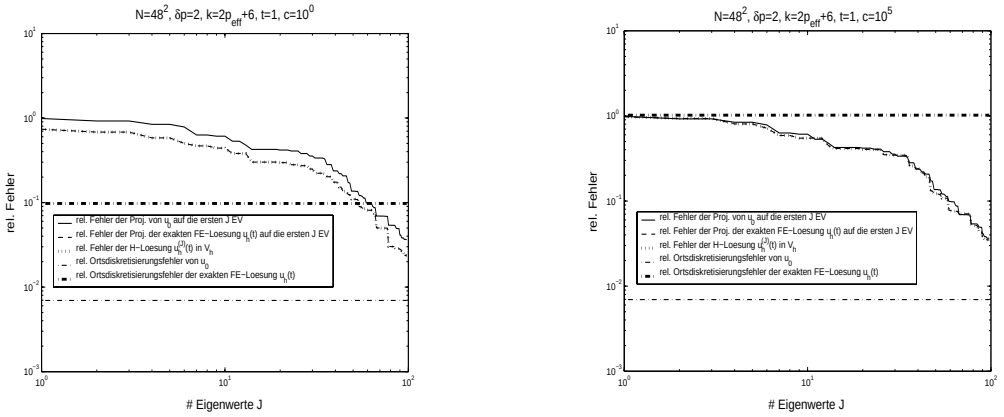

Abbildung 5.5: Relativer algebraischer Fehler der Projektion des Anfangswerts $\mathbf{u}_0$ und der Projektion der exakten FE-Lösung $\mathbf{u}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\mathbf{u}_0$ und der exakten FE-Lösung $\mathbf{u}_h(t)$ zum Zeitpunkt $t = 1$ in $\|\cdot\|_{L^2}$ für $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = 0.1 \exp(-100\,((0.2 - x_1)^2 + (0.2 - x_2)^2))$ und $\dot{u}_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

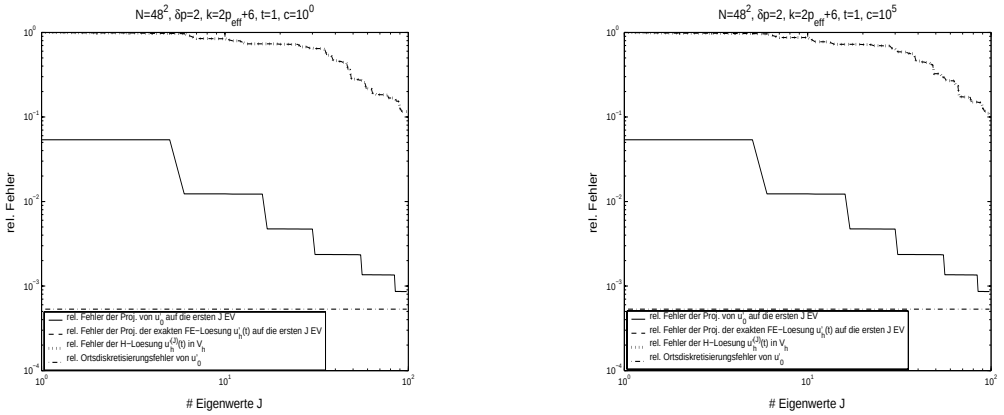

Abbildung 5.6: Relativer algebraischer Fehler der Projektion des Anfangswerts $\dot{\mathbf{u}}_0$ und der Projektion der exakten FE-Lösung $\dot{\mathbf{u}}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\dot{\mathbf{u}}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\dot{\mathbf{u}}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\dot{\mathbf{u}}_0$ zum Zeitpunkt $t = 1$ in $\|\cdot\|_{L^2}$ für $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = 0.1 \exp(-100\,((0.2 - x_1)^2 + (0.2 - x_2)^2))$ und $\dot{u}_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

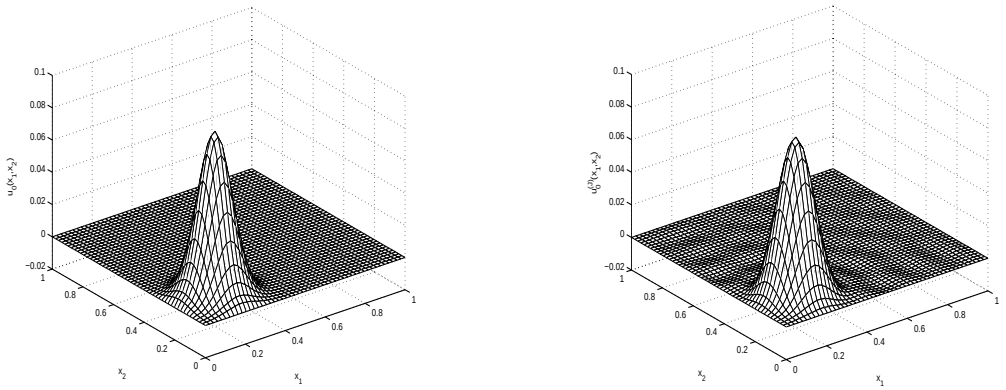

Abbildung 5.7: Koeffizientenvektor $\mathbf{u}_0$ der L^2-Projektion u_{0h} des Anfangswerts $u_0(x_1, x_2) = 0.1 \exp(-100\,((0.2 - x_1)^2 + (0.2 - x_2)^2))$ auf V_h (links) und Projektion von $\mathbf{u}_0$ auf die EV zu den $J = 98$ kleinsten EW von $M_h^{-1} A_h$ (rechts).

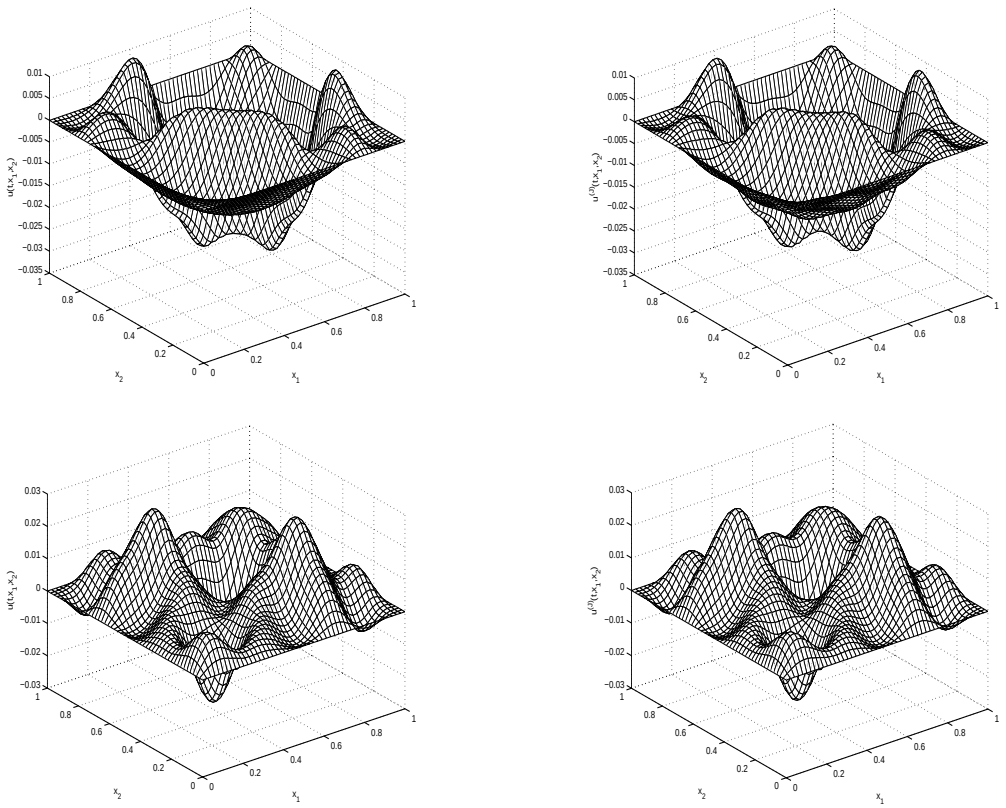

Abbildung 5.8: Exakte FE-Lösung $\mathbf{u}_h(t)$ für $c = 10^0$ (links oben) und $c = 10^5$ (links unten) sowie $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ mit $J = 98$ in V_h für $c = 10^0$ (rechts oben) und $c = 10^5$ (rechts unten) zum Zeitpunkt $t = 1$ zu den Anfangsdaten $u_0(x_1, x_2) = 0.1 \exp(-100\,((0.2 - x_1)^2 + (0.2 - x_2)^2))$ und $\dot{u}_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

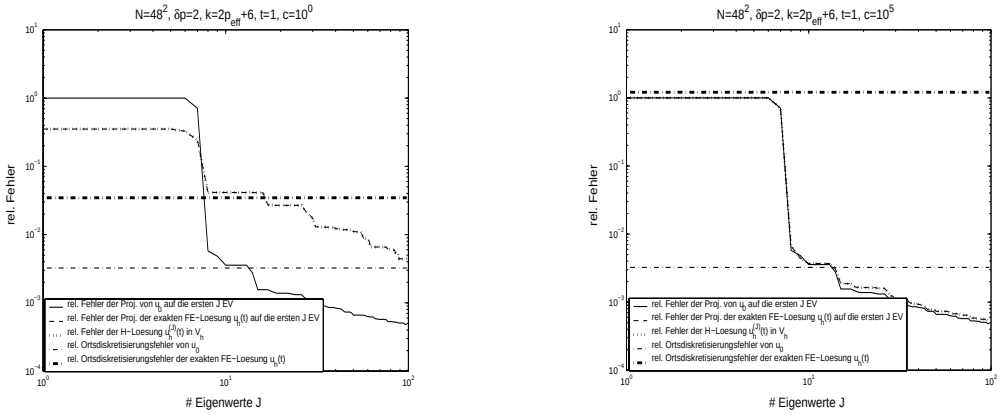

Abbildung 5.9: Relativer algebraischer Fehler der Projektion des Anfangswerts $\mathbf{u}_0$ und der Projektion der exakten FE-Lösung $\mathbf{u}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\mathbf{u}_0$ und der exakten FE-Lösung $\mathbf{u}_h(t)$ zum Zeitpunkt $t = 1$ in $\| \cdot \|_{L^2}$ für $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = \sin(2\pi x_1) \sin(3\pi x_2)$ und $\dot{u}_0(x_1, x_2) = \cos(4\pi x_1) \cos(5\pi x_2)$ und der Inhomogenität $f \equiv 10$.

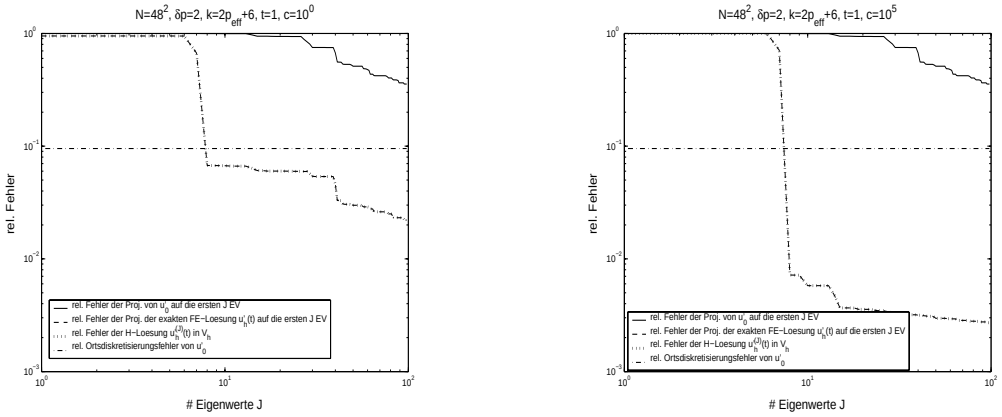

Abbildung 5.10: Relativer algebraischer Fehler der Projektion des Anfangswerts $\dot{\mathbf{u}}_0$ und der Projektion der exakten FE-Lösung $\dot{\mathbf{u}}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\dot{\mathbf{u}}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\dot{\mathbf{u}}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\dot{\mathbf{u}}_0$ zum Zeitpunkt $t = 1$ in $\| \cdot \|_{L^2}$ für $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = \sin(2\pi x_1) \sin(3\pi x_2)$ und $\dot{u}_0(x_1, x_2) = \cos(4\pi x_1) \cos(5\pi x_2)$ und der Inhomogenität $f \equiv 10$.

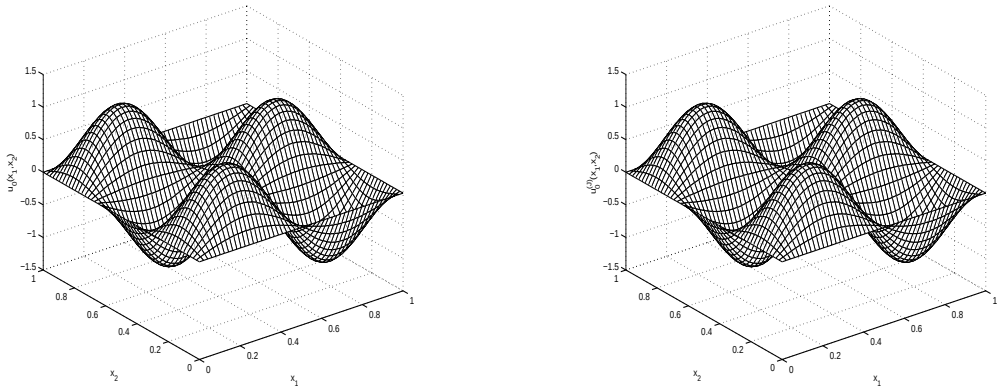

Abbildung 5.11: Koeffizientenvektor $\mathbf{u}_0$ der L^2-Projektion u_{0h} des Anfangswerts $u_0(x_1, x_2) = \sin(2\pi x_1)\sin(3\pi x_2)$ auf V_h (links) und Projektion von $\mathbf{u}_0$ auf die EV zu den $J = 98$ kleinsten EW von $M_h^{-1}A_h$ (rechts).

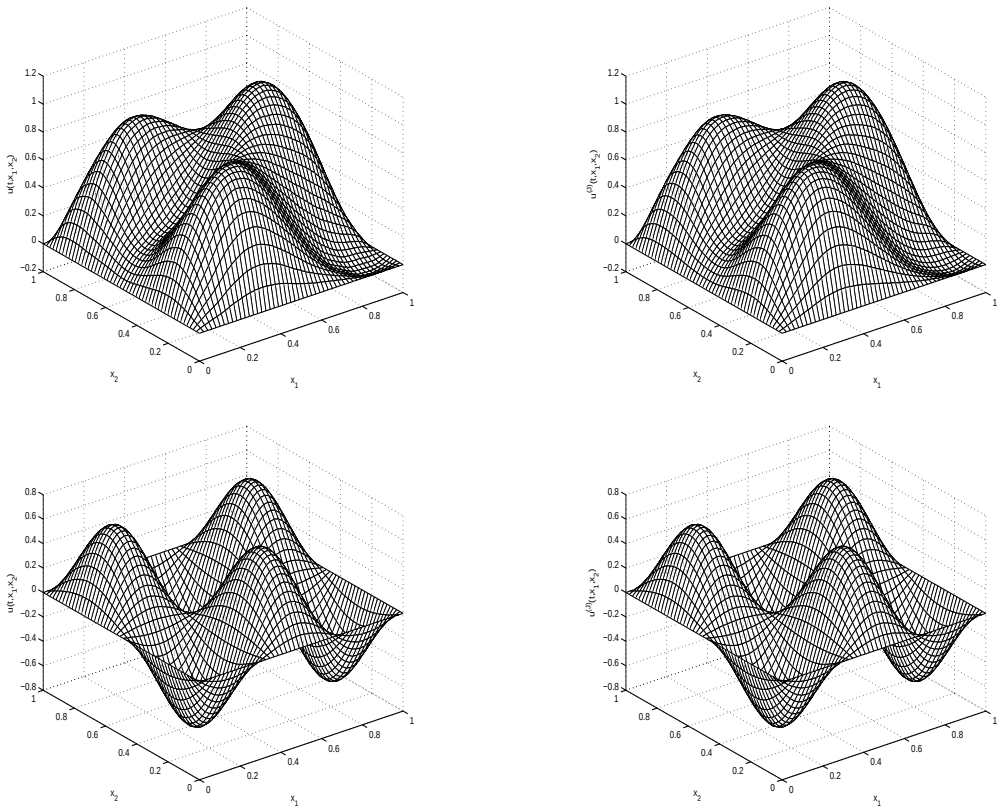

Abbildung 5.12: Exakte FE-Lösung $\mathbf{u}_h(t)$ für $c = 10^0$ (links oben) und $c = 10^5$ (links unten) sowie $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ mit $J = 98$ in V_h für $c = 10^0$ (rechts oben) und $c = 10^5$ (rechts unten) zum Zeitpunkt $t = 1$ zu den Anfangsdaten $u_0(x_1, x_2) = \sin(2\pi x_1)\sin(3\pi x_2)$ und $\dot{u}_0(x_1, x_2) = \cos(4\pi x_1)\cos(5\pi x_2)$ und der Inhomogenität $f \equiv 10$.

5.1.3 Energieerhaltung für die $\mathcal{H}$-Gautschi-Lösung

Wir prüfen nun zum Abschluss dieses Abschnitts die Energieerhaltung der nach Algorithmus 5.2 ermittelten FE-Lösung u_h der homogenen 2D Wellengleichung. Die Wellengleichung erhält die Gesamtenergie

$$\begin{aligned} E_{ges}(t) &= E_{kin}(t) + E_{pot}(t) = \tfrac{1}{2}\|\dot{u}_h(t)\|_{L^2}^2 + \tfrac{1}{2}c^2|u_h(t)|_{H^1}^2 \\ &= \tfrac{1}{2}\langle \dot{\mathbf{u}}_h(t), \dot{\mathbf{u}}_h(t)\rangle_{M_h} + \tfrac{1}{2}c^2\langle \mathbf{u}_h(t), \mathbf{u}_h(t)\rangle_{A_h} \\ &= \tfrac{1}{2}\langle M_h\dot{\mathbf{u}}_h(t), \dot{\mathbf{u}}_h(t)\rangle_2 + \tfrac{1}{2}c^2\langle A_h\mathbf{u}_h(t), \mathbf{u}_h(t)\rangle_2 \end{aligned} \tag{5.10}$$

genau dann, wenn

$$\begin{aligned} \tfrac{dE_{ges}}{dt}(t) &= \langle M_h\ddot{\mathbf{u}}_h(t), \dot{\mathbf{u}}_h(t)\rangle_2 + c^2\langle A_h\mathbf{u}_h(t), \dot{\mathbf{u}}_h(t)\rangle_2 \\ &= \langle \ddot{\mathbf{u}}_h(t) + c^2M_h^{-1}A_h\mathbf{u}_h(t), \dot{\mathbf{u}}_h(t)\rangle_{M_h} = \langle \mathbf{f}_h(t), \dot{\mathbf{u}}_h(t)\rangle_{M_h} = 0 \end{aligned}$$

ist.

Die numerischen Beispiele in Abbildung 5.13 bestätigen die exakte Erhaltung der Gesamtenergie (5.10) für die Gautschi-Lösung (5.7) und (5.8) über je 10000 untersuchte Zeitschritte hinweg für $c = 10^0$ und $c = 10^5$.

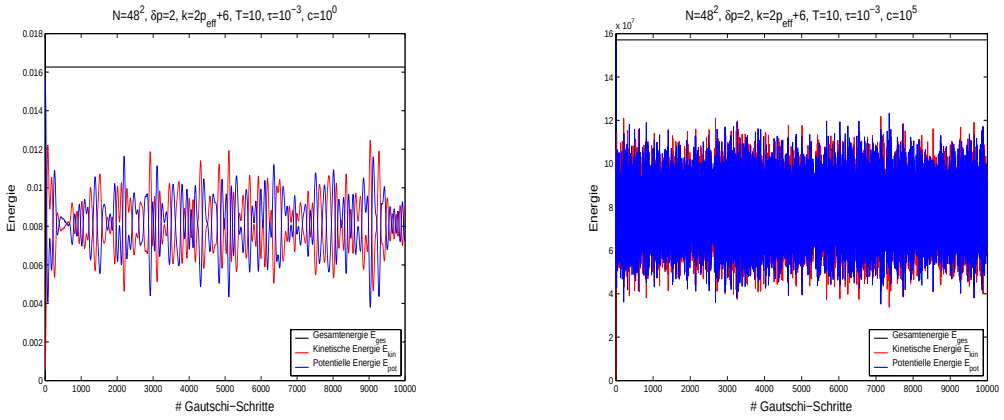

Abbildung 5.13: Kinetische Energie E_{kin}, potentielle Energie E_{pot} und Gesamtenergie $E_{ges} = E_{kin} + E_{pot}$ für $\tau = 10^{-3}$, $t \in [0, 10]$ und $c = 10^0$ (links) sowie $c = 10^5$ (rechts) zu den Anfangsdaten $u_0(x_1, x_2) = 0.1\exp(-100\,((0.2 - x_1)^2 + (0.2 - x_2)^2))$ und $\dot{u}_0(x_1, x_2) = -x_1(x_1 - 1)x_2(x_2 - 1)$ und der Inhomogenität $f \equiv 0$.

5.2 Sensitivitätsanalyse der diskreten Lösung der 2D Wellengleichung

Zur Ermittlung der diskreten Lösung der 2D Wellengleichung sind die eindimensionalen Drei-Term-Rekursionen aus Algorithmus 5.2 zu lösen. Wir werden nun untersuchen, wie sich Störungen in den Anfangsdaten und in den Koeffizienten dieser Drei-Term-Rekursionen auf deren Lösung auswirken.

Seien also die inhomogenen Drei-Term-Rekursionen

$$\rho^{(\mathbf{u}_{n+1})} = 2\rho^{(\mathbf{u}_n)} - \rho^{(\mathbf{u}_{n-1})} + \tau^2\sigma(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)}),\ \ n \geq 1 \tag{5.11}$$

mit den Anfangswerten $\rho^{(\mathbf{u}_0)}$ und $\rho^{(\mathbf{u}_1)}$ sowie

$$\rho^{(\dot{\mathbf{u}}_{n+1})} = \rho^{(\dot{\mathbf{u}}_{n-1})} + 2\tau\psi(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)}),\ \ n \geq 1 \tag{5.12}$$

mit den Anfangswerten $\rho^{(\dot{\mathbf{u}}_0)}$ und $\rho^{(\dot{\mathbf{u}}_1)}$ gegeben. Dabei steht $\rho^{(\cdot)}$ stellvertretend für die Koeffizienten $\rho_j^{(\cdot)}, 1 \leq j \leq J$, und λ für den entsprechenden EW λ_j von $M_h^{-1}A_h$.

Wir betrachten nun analog zur Vorgehensweise in [5, Seite 161 ff.] die Drei-Term-Rekursionen (5.11) und (5.12) als Abbildungen, die den Anfangswerten $\rho^{(\mathbf{u}_0)}$, $\rho^{(\mathbf{u}_1)}$ und $\rho^{(\dot{\mathbf{u}}_0)}$, $\rho^{(\dot{\mathbf{u}}_1)}$ sowie den Größen λ und $\rho^{(\mathbf{f}_n)}$ als Eingabegrößen die Werte $\rho^{(\mathbf{u}_{n+1})}$ und $\rho^{(\dot{\mathbf{u}}_{n+1})}, n \geq 1$, als Resultate zuordnet. Da in jedem Schritt nur zwei bis drei Multiplikationen und Additionen durchgeführt werden, deren Stabilität nachgewiesen ist (siehe [5, Seite 46, 47]), ist die Ausführung der Drei-Term-Rekursionen (5.11) und (5.12) in Gleitkommaarithmetik stabil. Über die numerische Brauchbarkeit entscheidet also nur die Kondition der Drei-Term-Rekursionen (5.11) und (5.12).

Zu deren Analyse seien nun die gestörten Anfangswerte $\tilde{\rho}^{(\mathbf{u}_0)} = \rho^{(\mathbf{u}_0)} + \Delta\rho^{(\mathbf{u}_0)}$, $\tilde{\rho}^{(\mathbf{u}_1)} = \rho^{(\mathbf{u}_1)} + \Delta\rho^{(\mathbf{u}_1)}$ und $\tilde{\rho}^{(\dot{\mathbf{u}}_0)} = \rho^{(\dot{\mathbf{u}}_0)} + \Delta\rho^{(\dot{\mathbf{u}}_0)}$, $\tilde{\rho}^{(\dot{\mathbf{u}}_1)} = \rho^{(\dot{\mathbf{u}}_1)} + \Delta\rho^{(\dot{\mathbf{u}}_1)}$ sowie die gestörten Größen $\tilde{\lambda} = \lambda + \Delta\lambda$ und $\tilde{\rho}^{(\mathbf{f}_n)} = \rho^{(\mathbf{f}_n)} + \Delta\rho^{(\mathbf{f}_n)}$ gegeben. Seien $\tilde{\rho}^{(\mathbf{u}_{n+1})}$ und $\tilde{\rho}^{(\dot{\mathbf{u}}_{n+1})}$ die Lösungen der gestörten Drei-Term-Rekursionen

$$\tilde{\rho}^{(\mathbf{u}_{n+1})} = 2\tilde{\rho}^{(\mathbf{u}_n)} - \tilde{\rho}^{(\mathbf{u}_{n-1})} + \tau^2\sigma(\tau^2c^2\tilde{\lambda})(-c^2\tilde{\lambda}\tilde{\rho}^{(\mathbf{u}_n)} + \tilde{\rho}^{(\mathbf{f}_n)}),\ \ n \geq 1 \tag{5.13}$$

mit den Anfangswerten $\tilde{\rho}^{(\mathbf{u}_0)}$ und $\tilde{\rho}^{(\mathbf{u}_1)}$ sowie

$$\tilde{\rho}^{(\dot{\mathbf{u}}_{n+1})} = \tilde{\rho}^{(\dot{\mathbf{u}}_{n-1})} + 2\tau\psi(\tau^2c^2\tilde{\lambda})(-c^2\tilde{\lambda}\tilde{\rho}^{(\mathbf{u}_n)} + \tilde{\rho}^{(\mathbf{f}_n)}),\ \ n \geq 1 \tag{5.14}$$

mit den Anfangswerten $\tilde{\rho}^{(\dot{\mathbf{u}}_0)}$ und $\tilde{\rho}^{(\dot{\mathbf{u}}_1)}$. Dann genügen die Fehler $\Delta\rho^{(\mathbf{u}_n)} := \tilde{\rho}^{(\mathbf{u}_n)} - \rho^{(\mathbf{u}_n)}$ und $\Delta\rho^{(\dot{\mathbf{u}}_n)} := \tilde{\rho}^{(\dot{\mathbf{u}}_n)} - \rho^{(\dot{\mathbf{u}}_n)}$ den inhomogenen Drei-Term-Rekursionen

$$\begin{aligned}
\Delta\rho^{(\mathbf{u}_{n+1})} &= 2\Delta\rho^{(\mathbf{u}_n)} - \Delta\rho^{(\mathbf{u}_{n-1})} + \tau^2\Delta\left(\sigma(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)})\right) \\
&\doteq 2\Delta\rho^{(\mathbf{u}_n)} - \Delta\rho^{(\mathbf{u}_{n-1})} + \tau^2\Delta\sigma(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)}) + \\
&\quad \tau^2\sigma(\tau^2c^2\lambda)\left(-c^2(\Delta\lambda\rho^{(\mathbf{u}_n)} + \lambda\Delta\rho^{(\mathbf{u}_n)}) + \Delta\rho^{(\mathbf{f}_n)}\right) \\
&= 2\cos(\tau c\sqrt{\lambda})\Delta\rho^{(\mathbf{u}_n)} - \Delta\rho^{(\mathbf{u}_{n-1})} + \tau^2E^{(n)},\ \ n \geq 1
\end{aligned} \tag{5.15}$$

mit

$$\begin{aligned}
E^{(n)} &= \Delta\sigma(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)}) + \\
&\quad \sigma(\tau^2c^2\lambda)\left(-c^2\Delta\lambda\rho^{(\mathbf{u}_n)} + \Delta\rho^{(\mathbf{f}_n)}\right)
\end{aligned} \tag{5.16}$$

und

$$\begin{aligned}
\Delta\rho^{(\dot{\mathbf{u}}_{n+1})} &= \Delta\rho^{(\dot{\mathbf{u}}_{n-1})} + 2\tau\Delta\left(\psi(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)})\right) \\
&\doteq \Delta\rho^{(\dot{\mathbf{u}}_{n-1})} + 2\tau\Delta\psi(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)} + \rho^{(\mathbf{f}_n)}) + \\
&\quad 2\tau\psi(\tau^2c^2\lambda)\left(-c^2(\Delta\lambda\rho^{(\mathbf{u}_n)} + \lambda\Delta\rho^{(\mathbf{u}_n)}) + \Delta\rho^{(\mathbf{f}_n)}\right) \\
&= \Delta\rho^{(\dot{\mathbf{u}}_{n-1})} + 2\tau\dot{E}^{(n)},\ \ n \geq 1
\end{aligned} \tag{5.17}$$

mit

$$\begin{aligned}\dot{E}^{(n)} &= \Delta\psi(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_n)}+\rho^{(\mathbf{f}_n)})+\\ &\quad \psi(\tau^2c^2\lambda)\left(-c^2(\Delta\lambda\rho^{(\mathbf{u}_n)}+\lambda\Delta\rho^{(\mathbf{u}_n)})+\Delta\rho^{(\mathbf{f}_n)}\right).\end{aligned} \tag{5.18}$$

Um die Rekursionen (5.15) und (5.17) zu starten, benötigen wir neben $\Delta\rho^{(\mathbf{u}_0)}$ und $\Delta\rho^{(\dot{\mathbf{u}}_0)}$ auch noch die Fehler $\Delta\rho^{(\mathbf{u}_1)}$ und $\Delta\rho^{(\dot{\mathbf{u}}_1)}$. Diese ergeben sich – analog zu den Rekursionen (5.15) und (5.17) – aus der Konditionsanalyse für die Werte $\mathbf{u}_1$ und $\dot{\mathbf{u}}_1$ zu

$$\begin{aligned}\Delta\rho^{(\mathbf{u}_1)} &\doteq \cos(\tau c\sqrt{\lambda})\,\Delta\rho^{(\mathbf{u}_0)}+\Delta\cos(\tau c\sqrt{\lambda})\,\rho^{(\mathbf{u}_0)}+\\ &\quad \tau\psi(\tau^2c^2\lambda)\,\Delta\rho^{(\dot{\mathbf{u}}_0)}+\tau\Delta\psi(\tau^2c^2\lambda)\,\rho^{(\dot{\mathbf{u}}_0)}+\\ &\quad \tfrac{1}{2}\tau^2\Delta\sigma(\tau^2c^2\lambda)\,\rho^{(\mathbf{f}_0)}+\tfrac{1}{2}\tau^2\sigma(\tau^2c^2\lambda)\,\Delta\rho^{(\mathbf{f}_0)}\end{aligned} \tag{5.19}$$

und

$$\begin{aligned}\Delta\rho^{(\dot{\mathbf{u}}_1)} &\doteq \cos(\tau c\sqrt{\lambda})\,\Delta\rho^{(\dot{\mathbf{u}}_0)}+\Delta\cos(\tau c\sqrt{\lambda})\,\rho^{(\dot{\mathbf{u}}_0)}+\\ &\quad \tau\Delta\psi(\tau^2c^2\lambda)(-c^2\lambda\rho^{(\mathbf{u}_0)}+\rho^{(\mathbf{f}_0)})+\\ &\quad \tau\psi(\tau^2c^2\lambda)\left(-c^2(\Delta\lambda\rho^{(\mathbf{u}_0)}+\lambda\Delta\rho^{(\mathbf{u}_0)})+\Delta\rho^{(\mathbf{f}_0)}\right).\end{aligned} \tag{5.20}$$

Dabei ist

$$\begin{aligned}\Delta\cos(\tau c\sqrt{\lambda}) &= \tfrac{d}{d\lambda}\cos(\tau c\sqrt{\lambda})\cdot\Delta\lambda\\ &= -\sin(\tau c\sqrt{\lambda})\,\tau c\tfrac{1}{2}(\sqrt{\lambda})^{-1}\cdot\Delta\lambda\\ &= -\tfrac{1}{2}\tau^2c^2\psi(\tau^2c^2\lambda)\cdot\Delta\lambda,\end{aligned}$$

$$\begin{aligned}\Delta\sigma(\tau^2c^2\lambda) &= 2\tfrac{d}{d\lambda}\left(\tfrac{1-\cos(\tau c\sqrt{\lambda})}{\tau^2c^2\lambda}\right)\cdot\Delta\lambda\\ &= 2\tfrac{\frac{1}{2}\tau^2c^2\psi(\tau^2c^2\lambda)\,\tau^2c^2\lambda-(1-\cos(\tau c\sqrt{\lambda}))\tau^2c^2}{(\tau^2c^2\lambda)^2}\cdot\Delta\lambda\\ &= \tfrac{1}{\lambda}(\psi(\tau^2c^2\lambda)-\sigma(\tau^2c^2\lambda))\cdot\Delta\lambda\end{aligned}$$

und

$$\begin{aligned}\Delta\psi(\tau^2c^2\lambda) &= \tfrac{d}{d\lambda}\left(\tfrac{\sin(\tau c\sqrt{\lambda})}{\tau c\sqrt{\lambda}}\right)\cdot\Delta\lambda\\ &= \tfrac{\cos(\tau c\sqrt{\lambda})\,\tau c\frac{1}{2}(\sqrt{\lambda})^{-1}\,\tau c\sqrt{\lambda}-\sin(\tau c\sqrt{\lambda})\,\tau c\frac{1}{2}(\sqrt{\lambda})^{-1}}{\tau^2c^2\lambda}\cdot\Delta\lambda\\ &= \tfrac{1}{2\lambda}(\cos(\tau c\sqrt{\lambda})-\psi(\tau^2c^2\lambda))\cdot\Delta\lambda.\end{aligned}$$

Nun gilt für die Anfangswerte der inhomogenen Drei-Term-Rekursionen (5.15) und (5.17)

$$\begin{aligned}\Delta\rho^{(\mathbf{u}_0)} &= \tau^2E^{(-1)},\\ \Delta\rho^{(\mathbf{u}_1)} &= 2\cos(\tau c\sqrt{\lambda})\Delta\rho^{(\mathbf{u}_0)}+\tau^2E^{(0)} \quad\Longrightarrow\\ \tau^2E^{(0)} &= \Delta\rho^{(\mathbf{u}_1)}-2\cos(\tau c\sqrt{\lambda})\Delta\rho^{(\mathbf{u}_0)}\end{aligned}$$

und

$$\begin{aligned}\Delta\rho^{(\dot{\mathbf{u}}_0)} &= 2\tau\dot{E}^{(-1)},\\ \Delta\rho^{(\dot{\mathbf{u}}_1)} &= 2\tau\dot{E}^{(0)}.\end{aligned}$$

Die Lösungen der inhomogenen Rekursionen (5.15) und (5.17) lassen sich nun durch Superposition gemäß

$$\Delta\rho^{(\mathbf{u}_{n+1})} = \tau^2 \sum_{j=-1}^{n} E^{(j)} W_{n-j} \tag{5.21}$$

(vgl. Lemma 2.7) und

$$\Delta\rho^{(\dot{\mathbf{u}}_{n+1})} = 2\tau \sum_{j=-1}^{n} \dot{E}^{(j)} \dot{W}_{n-j} \tag{5.22}$$

(vgl. Lemma 2.10) gewinnen mit den in Abschnitt 2.4.3 bestimmten diskreten Greenschen Funktionen

$$W_{n-j} = (\sin(n-j+1)\tau c\sqrt{\lambda})(\sin\tau c\sqrt{\lambda})^{-1}, \;\; n \geq j$$

der Drei-Term-Rekursion (5.15) und

$$\dot{W}_{n-j} = \begin{cases} 1 & \text{falls } n = j, j+2, \ldots \\ 0 & \text{falls } n = j+1, j+3, \ldots \end{cases}$$

der Drei-Term-Rekursion (5.17).

Die diskreten Greenschen Funktionen W_{n-j} und $\dot{W}_{n-j}$ geben also an, wie die Inhomogenitäten $\tau^2 E^{(j)}$ und $2\tau\dot{E}^{(j)}$, die die absoluten Eingabefehler im j-ten Rekursionsschritt darstellen, in den Endresultaten verstärkt werden.
Nun gilt

$$|W_{n-j}| = |(\sin(n-j+1)\tau c\sqrt{\lambda})(\sin\tau c\sqrt{\lambda})^{-1}| \leq n-j+1,$$

d.h. es liegt schlechtestenfalls lineare Fehlerverstärkung in der Anzahl der Rekursionsschritte vor.
Weiter ist $|\dot{W}_{n-j}| \leq 1$; da in $\dot{E}^{(j)}$ jedoch $\Delta\rho^{(\mathbf{u}_j)}$ linear eingeht, ist die Entwicklung von $\Delta\rho^{(\dot{\mathbf{u}}_j)}$ eng an jene von $\Delta\rho^{(\mathbf{u}_j)}$ gekoppelt.

Die relativen Fehler $\theta^{(\mathbf{u}_{n+1})} := \frac{\Delta\rho^{(\mathbf{u}_{n+1})}}{\rho^{(\mathbf{u}_{n+1})}}$ und $\theta^{(\dot{\mathbf{u}}_{n+1})} := \frac{\Delta\rho^{(\dot{\mathbf{u}}_{n+1})}}{\rho^{(\dot{\mathbf{u}}_{n+1})}}$ lösen die inhomogenen Drei-Term-Rekursionen

$$\theta^{(\mathbf{u}_{n+1})} = \frac{2\cos(\tau c\sqrt{\lambda})\rho^{(\mathbf{u}_n)}}{\rho^{(\mathbf{u}_{n+1})}}\,\theta^{(\mathbf{u}_n)} - \frac{\rho^{(\mathbf{u}_{n-1})}}{\rho^{(\mathbf{u}_{n+1})}}\,\theta^{(\mathbf{u}_{n-1})} + \tau^2\epsilon^{(n)}, \;\; n \geq 1 \tag{5.23}$$

zu den Anfangswerten

$$\begin{aligned} \theta^{(\mathbf{u}_0)} &= \tau^2\epsilon^{(-1)}, \\ \theta^{(\mathbf{u}_1)} &= \frac{2\cos(\tau c\sqrt{\lambda})\rho^{(\mathbf{u}_0)}}{\rho^{(\mathbf{u}_1)}}\,\theta^{(\mathbf{u}_0)} + \tau^2\epsilon^{(0)} \end{aligned}$$

mit $\epsilon^{(n)} = \frac{E^{(n)}}{\rho^{(\mathbf{u}_{n+1})}}$, $n \geq 1$, und

$$\theta^{(\dot{\mathbf{u}}_{n+1})} = \frac{\rho^{(\dot{\mathbf{u}}_{n-1})}}{\rho^{(\dot{\mathbf{u}}_{n+1})}}\,\theta^{(\dot{\mathbf{u}}_{n-1})} + 2\tau\dot{\epsilon}^{(n)}, \;\; n \geq 1 \tag{5.24}$$

zu den Anfangswerten

$$\begin{aligned} \theta^{(\dot{\mathbf{u}}_0)} &= 2\tau\dot{\epsilon}^{(-1)}\,, \\ \theta^{(\dot{\mathbf{u}}_1)} &= 2\tau\dot{\epsilon}^{(0)} \end{aligned}$$

mit $\dot{\epsilon}^{(n)} = \frac{\dot{E}^{(n)}}{\rho^{(\dot{\mathbf{u}}_{n+1})}}$, $n \geq 1$.
Daher gilt

$$\theta^{(\mathbf{u}_{n+1})} = \tau^2 \sum_{j=-1}^{n} \epsilon^{(j)} r_{n-j} \tag{5.25}$$

mit $r_{n-j} = \frac{\rho^{(\mathbf{u}_{j+1})}}{\rho^{(\mathbf{u}_{n+1})}}\, W_{n-j}$ und

$$\theta^{(\dot{\mathbf{u}}_{n+1})} = 2\tau \sum_{j=-1}^{n} \dot{\epsilon}^{(j)} \dot{r}_{n-j} \tag{5.26}$$

mit $\dot{r}_{n-j} = \frac{\rho^{(\dot{\mathbf{u}}_{j+1})}}{\rho^{(\dot{\mathbf{u}}_{n+1})}}\, \dot{W}_{n-j}$.

Die Funktionen r_{n-j} und $\dot{r}_{n-j}$ beschreiben nun offensichtlich die Verstärkung der relativen Fehler und kennzeichnen somit die relative Kondition der Drei-Term-Rekursionen (5.15) und (5.17).

5.3 Lösung der Wärmeleitungsgleichung und der Schrödingergleichung

Wir haben in Abschnitt 5.1 zunächst den Gautschi-Algorithmus zur Lösung der in der Knotenbasis gegebenen diskreten 2D Wellengleichung auf die Eigenräume zu den kleinsten EW von $M_h^{-1}A_h$ projiziert und dann die daraus resultierenden eindimensionalen Drei-Term-Rekursionen separat gelöst.
Wir werden nun versuchen, mit derselben Lösungsstrategie zwei weitere PDEs zu lösen, die Wärmeleitungsgleichung und die Schrödingergleichung.

5.3.1 Die 2D Wärmeleitungsgleichung

Die 2D Wärmeleitungsgleichung

$$\dot{u} + Au = f, \;\; u(0) = u_0 \in L^2(\Omega) \tag{5.27}$$

mit $f \in L^2(\Omega_T)$, $\Omega_T := (0,T) \times \Omega$, besitzt die Lösung

$$u(t) = \exp(-tA)\, u_0 + \int_0^t \exp(-(t-s)A) f(s) ds$$

(siehe Kapitel 4 bzw. [14, Seite 10]).
Wir approximieren als Erstes den Lösungsvektor $\mathbf{u}_{hom}(t) = \exp(-tM_h^{-1}A_h)\, \mathbf{u}_0$ der homogenen Wärmeleitungsgleichung in der Knotenbasis $\{\psi_j^h\}_{j=1}^N$ von V_h

$$\dot{\mathbf{u}}_h + M_h^{-1}A_h\mathbf{u}_h = \mathbf{0}, \;\; \mathbf{u}_h(0) = \mathbf{u}_0 \in \mathbb{R}^N \tag{5.28}$$

durch Projektion auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$:

$$\begin{aligned} \mathbf{u}_{hom}(t) &= \exp(-tM_h^{-1}A_h)\,\mathbf{u}_0 \\ &= \sum_{j=1}^{N} \exp(-t\lambda_j)\rho_j^{(\mathbf{u}_0)}\mathbf{v}_j \end{aligned} \tag{5.29}$$

lässt sich für hinreichend großes $t > 0$ hervorragend durch

$$\mathbf{u}_{hom}^{(J)}(t) = \sum_{j=1}^{J} \exp(-t\lambda_j)\rho_j^{(\mathbf{u}_0)}\mathbf{v}_j \tag{5.30}$$

annähern.

Im Gegensatz zur Lösung der homogenen Wellengleichung liegt hier ein exponentielles Abklingverhalten der Koeffizienten des Lösungsvektors $\mathbf{u}_{hom}(t)$ vor, d.h. die Lösung $u_{hom}(t)$ lässt sich für hinreichend großes $t > 0$ für beliebige Anfangswerte u_0 sehr genau durch den Koeffizientenvektor in (5.30) angeben.

$\mathbf{u}_{hom}^{(J)}(t)$ kann nun analog zu Algorithmus 5.2 folgendermaßen berechnet werden:

Algorithmus 5.3 (Hom. Wärmeleitungsgleichung im Frequenzraum)

1. *Berechnung des Koeffizientenvektors* $\rho^{(\mathbf{u}_0)}$ *zum Anfangswert* $\mathbf{u}_0$
2. *Berechnung der zur Lösung* $\mathbf{u}_{hom}(t)$ *gehörigen Koeffizienten*

   $$\rho_j^{(\mathbf{u}_{hom}(t))} = \exp(-t\lambda_j)\rho_j^{(\mathbf{u}_0)}$$

   für $1 \leq j \leq J$.

Damit ist der Lösungsvektor $\mathbf{u}_{hom}(t)$ zum Zeitpunkt t approximativ durch

$$\mathbf{u}_{hom}^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_{hom}(t))}\mathbf{v}_j \tag{5.31}$$

gegeben.

In Abbildung 5.14 werden die relativen Fehler der mittels Algorithmus 5.3 berechneten Lösungsvektoren verglichen mit den exakten FE-Lösungen zu den Anfangsdaten $u_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ (links) und $u_0(x_1, x_2) = 0.1\exp(-100\,((0.2-x_1)^2+(0.2-x_2)^2))$ (rechts) für $t = 10^0, 10^{-2}$ und 10^{-4} graphisch dargestellt. Der Diskretisierungsfehler im Ort ist wiederum durch eine gerade, gestrichelt gepunktete Linie gekennzeichnet. Während für $t = 10^0$ bereits eine Rang-1-Approximation für hohe relative Genauigkeit reicht, müssen für kleiner werdendes t immer mehr EW zum Erreichen einer bestimmten Approximationsgüte hinzugezogen werden. Für $t \to 0$ gilt nämlich $\exp(-tM_h^{-1}A_h) \to I$, d.h. mit obigem Ansatz nähert sich die relative Genauigkeit von $\mathbf{u}_{hom}^{(J)}(t) = \sum_{j=1}^{J} \exp(-t\lambda_j)\rho_j^{(\mathbf{u}_0)}\mathbf{v}_j$ für $t \to 0$ dem Wert $\frac{\|\mathbf{u}_0 - \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_0)}\mathbf{v}_j\|_{M_h}}{\|\mathbf{u}_0\|_{M_h}}$ (also jener von $\mathbf{u}_0^{(J)} = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_0)}\mathbf{v}_j$) an.

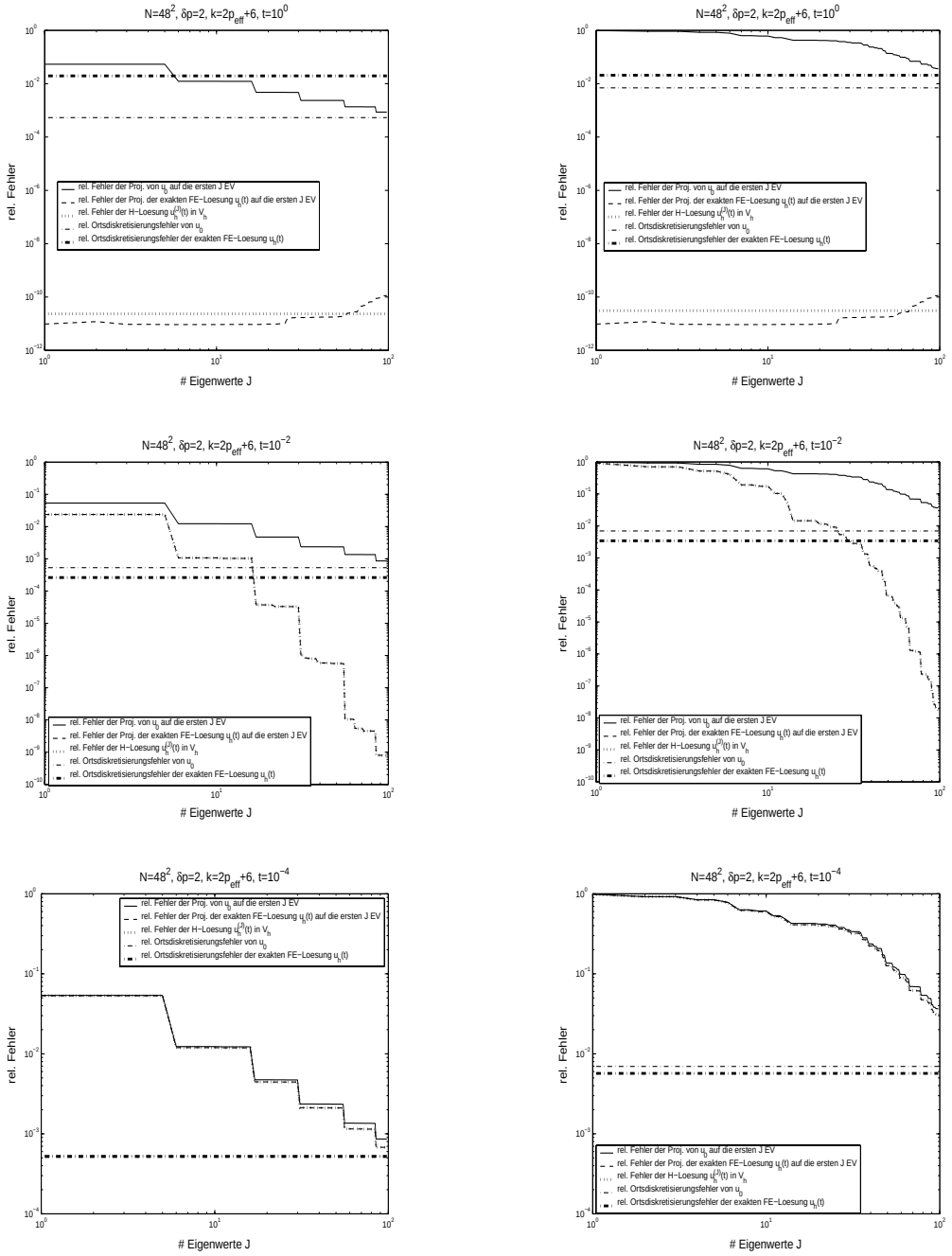

Abbildung 5.14: Relativer algebraischer Fehler der Projektion des Anfangswerts $\mathbf{u}_0$ und der Projektion der exakten FE-Lösung $\mathbf{u}_h(t)$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$ sowie relativer algebraischer Fehler der $\mathcal{H}$-Lösung $\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_h(t))} \mathbf{v}_j$ in V_h und relativer Diskretisierungsfehler im Ort von $\mathbf{u}_0$ und der exakten FE-Lösung $\mathbf{u}_h(t)$ zum Zeitpunkt $t = 10^0$ (oben), $t = 10^{-2}$ (Mitte) und $t = 10^{-4}$ (unten) in $\| \cdot \|_{L^2}$ zu den Anfangsdaten $u_0(x_1, x_2) = x_1(x_1 - 1)x_2(x_2 - 1)$ (links) und $u_0(x_1, x_2) = 0.1 \exp(-100\,((0.2 - x_1)^2 + (0.2 - x_2)^2))$ (rechts).

Der Koeffizientenvektor

$$\mathbf{u}_{part}(t) = \int_0^t \exp(-(t-s)M_h^{-1}A_h)\mathbf{f}_h(s)ds \tag{5.32}$$

der partikulären FE-Lösung

$$u_{part}(t) = \int_0^t \exp(-(t-s)A_h^{op})(Q_h f)(s)ds \tag{5.33}$$

der auf V_h projizierten inhomogenen Wärmeleitungsgleichung lässt sich nun mit Hilfe der nach Algorithmus 4.5 konstruierten Rang-J-Approximation

$$\exp_J(-(t-s)M_h^{-1}A_h) = \sum_{j=1}^{J} \exp(-(t-s)\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j} \tag{5.34}$$

an $\exp(-(t-s)M_h^{-1}A_h)$ wie folgt durch $\mathbf{u}_{part}^{(J)}(t)$ nähern:

$$\begin{aligned}
\mathbf{u}_{part}^{(J)}(t) &= \int_0^t \exp_J(-(t-s)M_h^{-1}A_h)\mathbf{f}_h(s)ds \\
&= \int_0^t \sum_{j=1}^{J} \exp(-(t-s)\lambda_j)\frac{\mathbf{v}_j\mathbf{u}_j^T}{\mathbf{v}_j^T\mathbf{u}_j} \sum_{i=1}^{N} \rho_i^{(\mathbf{f}_h(s))}\mathbf{v}_i ds \\
&= \sum_{j=1}^{J} \left(\int_0^t \exp(-(t-s)\lambda_j)\rho_j^{(\mathbf{f}_h(s))} ds \right) \mathbf{v}_j
\end{aligned} \tag{5.35}$$

Zu dessen Berechnung sind also die J Integrale $\int_0^t \exp(-(t-s)\lambda_j)\rho_j^{(\mathbf{f}_h(s))}ds$ mittels numerischer Quadratur auszuwerten. Bedient man sich einer Quadraturformel (QF) mit $K+1$ paarweise verschiedenen Knoten $0 = s_0 < s_1 < \ldots < s_{K-1} < s_K = t$ auf $[0,t]$, müssen pro Auswertung der QF (d.h. für jedes $j \in \{1,\ldots,J\}$) die $K+1$ Koeffizienten $\rho_j^{(\mathbf{f}_h(s_k))}$, $0 \le k \le K$, berechnet werden. Damit beläuft sich der Gesamtaufwand zur Berechnung von (5.35) in führender Ordnung auf $J(K+1)2N$ flops für die Bestimmung der $J(K+1)$ Koeffizienten $\rho_j^{(\mathbf{f}_h(s_k))}$.

Zusammenfassend lässt sich nun der folgende Algorithmus zur Berechnung der Lösung der inhomogenen Wärmeleitungsgleichung angeben:

Algorithmus 5.4 (Inh. Wärmeleitungsgleichung im Frequenzraum)

1. *Berechnung des Koeffizientenvektors $\rho^{(\mathbf{u}_{hom}(t))}$ nach Algorithmus 5.3*

2. *Berechnung der $K+1$ Koeffizienten $\rho_j^{(\mathbf{f}_h(s_k))}$, $0 \leq k \leq K$, und anschließende Anwendung der QF $\hat{I} : C[0,t] \to \mathbb{R}; \hat{I}(f) = \sum_{k=0}^{K} \gamma_k f(s_k)$ zur Approximation von*

$$I_j := \int_0^t \exp(-(t-s)\lambda_j)\rho_j^{(\mathbf{f}_h(s))} ds \tag{5.36}$$

durch

$$\hat{I}_j := \sum_{k=0}^{K} \gamma_k \exp(-(t-s_k)\lambda_j)\rho_j^{(\mathbf{f}_h(s_k))} \tag{5.37}$$

für $1 \leq j \leq J$.

Damit ist der Lösungsvektor $\mathbf{u}_{inhom}(t)$ zum Zeitpunkt t approximativ durch

$$\mathbf{u}_{inhom}^{(J)}(t) = \mathbf{u}_{hom}^{(J)}(t) + \mathbf{u}_{part}^{(J)}(t) = \sum_{j=1}^{J} \left(\rho_j^{(\mathbf{u}_{hom}(t))} + \hat{I}_j\right) \mathbf{v}_j \tag{5.38}$$

gegeben.

5.3.2 Die 2D Schrödingergleichung

Zum Abschluss dieses Kapitels wenden wir uns noch kurz der 2D Schrödingergleichung zu und bestimmen deren Lösung analog zum Vorgehen bei der Wellengleichung und Wärmeleitungsgleichung.

Für die Lösung der Schrödingergleichung

$$i\dot{u} - Au = 0, \;\; u(0) = u_0 \tag{5.39}$$

gilt

$$u(t) = \exp(-itA)\, u_0, \tag{5.40}$$

und für die $\mathcal{H}$-Darstellbarkeit des zugehörigen Lösungsoperators $\exp(-itA)$ dasselbe wie für die Lösungsoperatoren $\cos(tc\sqrt{A})$ und $\sin(tc\sqrt{A})$ der Wellengleichung (vgl. Abschnitt 4.2.3).

Das Matrix-Vektor-Produkt $\mathbf{u}_h(t) = \exp(-itM_h^{-1}A_h)\,\mathbf{u}_0$ kann wie folgt durch Spektralzerlegung der Matrixexponentialfunktion zu

$$\begin{aligned} \mathbf{u}_h(t) &= \exp(-itM_h^{-1}A_h)\,\mathbf{u}_0 \\ &= V\exp(-itD)\,V^{-1}\,\textstyle\sum_{j=1}^{N} \rho_j^{(\mathbf{u}_0)}\,\mathbf{v}_j \\ &= \textstyle\sum_{j=1}^{N} \exp(-it\lambda_j)\rho_j^{(\mathbf{u}_0)}\mathbf{v}_j \end{aligned} \tag{5.41}$$

bestimmt werden. Damit bildet

$$\mathbf{u}_h^{(J)}(t) = \sum_{j=1}^{J} \exp(-it\lambda_j)\rho_j^{(\mathbf{u}_0)}\mathbf{v}_j \tag{5.42}$$

für niederfrequente Anfangswerte $\mathbf{u}_0$ im Ort eine gute Approximation an (5.41).

Der relative Fehler von $\mathbf{u}_h^{(J)}(t)$ liegt dabei im Bereich von $\frac{\|\mathbf{u}_0 - \sum_{j=1}^{J} \rho_j^{(\mathbf{u}_0)}\,\mathbf{v}_j\|_{M_h}}{\|\mathbf{u}_0\|_{M_h}}$, dem relativen Fehler der Projektion von $\mathbf{u}_0$ auf $\langle \mathbf{v}_1, \ldots, \mathbf{v}_J \rangle$.

Kapitel 6

Implementierung in der Programmiersprache C

Zur Überprüfung der Theorie in der Praxis wurde die gesamte $\mathcal{H}$-Arithmetik sowie die Algorithmen zur Lösung der Wellengleichung und Wärmeleitungsgleichung in C implementiert. Sämtliche numerische Berechnungen in dieser Arbeit wurden mittels dieser C-Routinen durchgeführt.

In diesem Kapitel werden wir die Grundzüge der Implementierung in C kurz und prägnant darstellen. Die im Einzelnen behandelten Punkte sind

- das gewählte Speicherschema der $\mathcal{H}$-Matrizen
- die rekursive Beschreibung der $\mathcal{H}$-Operationen und $\mathcal{H}$-Zerlegungen
- die L^2-orthogonale Projektion auf den FE-Raum V_h in $\mathcal{H}$-Arithmetik
- die Algorithmen zur Bestimmung von $\mathcal{H}$-EW und $\mathcal{H}$-EV des diskreten 2D Laplace-Operators
- die Lösung der 2D Wellengleichung und Wärmeleitungsgleichung

Dabei wurden sämtliche von uns geschriebene C-Programme unter Einbindung der BLAS- und LAPACK-Bibliotheken optimiert.

6.1 Die 2D Steifigkeitsmatrix und Massenmatrix als $\mathcal{H}$-Matrizen

Die 2D Steifigkeitsmatrix zum Differentialoperator $-\triangle$ auf dem regelmäßig triangulierten Einheitsquadrat mit homogenen Dirichlet-Randbedingungen bzgl. der nodalen FE-Basis lautet bekanntermaßen

$$A_h = \begin{pmatrix} B & -I & & \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & B \end{pmatrix} \in \mathbb{R}^{N \times N}$$

mit

$$B = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

und $N = n^2$, die zur selben diskreten Basis gehörende Massenmatrix

$$M_h = \frac{h^2}{12} \begin{pmatrix} C & D_- & & \\ D_+ & \ddots & \ddots & \\ & \ddots & \ddots & D_- \\ & & D_+ & C \end{pmatrix} \in \mathbb{R}^{N \times N}$$

mit den $n \times n$-Matrizen

$$C = \begin{pmatrix} 6 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 6 \end{pmatrix}, \quad D_+ = \begin{pmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 1 \end{pmatrix}$$

$$\text{und } D_- = \begin{pmatrix} 1 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix}.$$

Wir werden im Folgenden beschreiben, wie diese beiden FE-Matrizen auf hierarchische Gestalt transformiert und in entsprechend strukturierter Weise abgespeichert werden können. Sei dazu das regelmäßig triangulierte Einheitsquadrat gegeben (siehe Abbildung 3.2), welchem wir die regelmäßigen quadratischen Gitter mit der Gitterbreite $\left(\frac{1}{2}\right)^l$ und dem zugehörigen Level $l \geq 0$ einschreiben.

Die 4^l Teilblöcke

$$t_{ij}^l = \{(x, y) \in \Omega | (i-1)\, 2^{-l} < x \leq i\, 2^{-l},\ \ (j-1)\, 2^{-l} < y \leq j\, 2^{-l}\},$$

$1 \leq i, j \leq 2^l$, des quadratischen Gitters vom Level l enthalten jeweils bestimmte Knoten des Dreiecksgitters, wobei auf dem Rand eines Teilquadrats liegende Knoten dem Quadrat links bzw. unterhalb des jeweiligen Knotens zugewiesen werden: Ein Paar zweier Blöcke t_{ij}^l und t_{km}^l vom Level l ist nun genau dann zulässig, wenn $|i-k| > 1$ oder $|j-m| > 1$ gilt (vgl. Abbildung 6.1).
Was die Anordnung der den einzelnen Teilquadraten von Ω entsprechenden Indizes in einer $\mathcal{H}$-Matrix M betrifft, gilt das in Bemerkung 1 nach Definition 3.4 Gesagte.

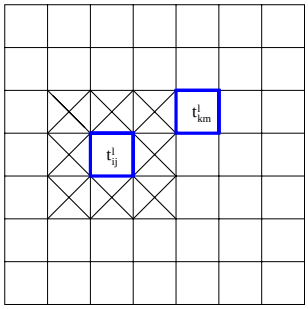

Abbildung 6.1: Zulässiges Paar zweier Blöcke t^l_{ij} und t^l_{km} vom Level l

Einen vollbesetzten Matrixblock $M^b = (a_{ij})_{\substack{1\le i\le m\\ 1\le j\le n}} \in \mathbb{R}^{m\times n}$ speichern wir stets in der Form

$$\begin{bmatrix} m \\ n \\ 0 \\ a_{11} \\ \vdots \\ a_{mn} \end{bmatrix}.$$

Die 0 im dritten Eintrag lässt erkennen, dass es sich bei M^b um eine vollbesetzte Matrix handelt.
Einen Rk-Matrixblock $M^b = \mathbf{a}\mathbf{b}^T \in \mathbb{R}^{m\times n}$ mit $\mathbf{a} = (a_{ij})_{\substack{1\le i\le m\\ 1\le j\le k}} \in \mathbb{R}^{m\times k}$, $\mathbf{b} = (b_{ij})_{\substack{1\le i\le n\\ 1\le j\le k}} \in \mathbb{R}^{n\times k}$ schreiben wir in der Form

$$\begin{bmatrix} m \\ n \\ k \\ a_{11} \\ \vdots \\ a_{mk} \\ b_{11} \\ \vdots \\ b_{nk} \end{bmatrix}.$$

Dabei beschreibt der dritte Eintrag k den Rang der Rk-Matrix M^b.

Die Speicherung einer $\mathcal{H}$-Matrix erfolgt nun als rekursives `struct array`, das durch die Vorschrift

```
struct block {struct block *bp; double *dp;} bl;
```

definiert ist. Durch

```
bl *h;
```

wird der Zeiger `h` auf eine $\mathcal{H}$-Matrix deklariert.

• Für $p_{eff} = 0$ entspricht die $\mathcal{H}$-Matrix $A = (a_{ij})_{1 \le i,j \le N}$ einer vollbesetzten Matrix: `(*h).dp` zeigt auf das erste Element des `double`-Vektors

$$\begin{bmatrix} N \\ N \\ 0 \\ a_{11} \\ \vdots \\ a_{NN} \end{bmatrix}.$$

• Für $p_{eff} = 1$ entspricht die $\mathcal{H}$-Matrix $A = (A_{ij})_{1 \le i,j \le 4}$ einer 4×4-Blockmatrix. Die 16 Teilblöcke eines 4×4-Blocks werden immer spaltenweise von links oben nach rechts unten durchnumeriert, d.h. `(*h).bp[0].dp` zeigt auf den vollbesetzten Matrixblock A_{11}, `(*h).bp[1].dp` auf $A_{21}, \ldots,$ `(*h).bp[15].dp` auf A_{44}.

• Für $p_{eff} = 2$ entspricht die $\mathcal{H}$-Matrix $A = (A_{ij})_{1 \le i,j \le 4}$ einer 4×4-Blockmatrix, wobei jeder Block A_{ij} wiederum aus 16 vollbesetzten bzw. Rk-Matrizen besteht. `(*h).bp[k].bp[m].dp` zeigt auf die $(m+1)$-te Blockmatrix des $(k+1)$-ten Teilblocks von A $(0 \le k, m \le 15)$.

• Für $p_{eff} \ge 3$ werden nun die nicht zulässigen Blöcke `(*h).bp[k].bp[m]` weiter partitioniert in die 16 Teilblöcke `(*h).bp[k].bp[m].bp[n]`, $0 \le n \le 15$, die zulässigen Blöcke durch die entsprechenden Rk-Matrizen `(*h).bp[k].bp[m].dp` besetzt.

• Die Nichtzulässigkeit eines Blocks `(*h).bp[k].bp[m]` wird dabei durch `(*h).bp[k].bp[m].dp[0]=0` gekennzeichnet, während für einen zulässigen Block automatisch `(*h).bp[k].bp[m].dp[0]>0` gilt (siehe Abbildung 6.2).

Die **Speicherung der Matrizen A_h und M_h als $\mathcal{H}$-Matrizen** erfolgt nun blockweise rekursiv, wobei beim Block $I \times I - I$ bezeichne die gesamte Indexmenge – mit der Untersuchung auf Zulässigkeit begonnen wird:

- Ist ein Block $b = \tau \times \sigma \subset I \times I$ zulässig, d.h. entspricht er einer Rk-Matrix, wird er mit einer Rk-Nullmatrix besetzt.
- Ist ein Block $b = \tau \times \sigma \subset I \times I$ nicht zulässig und $\text{level}(b) = p_{eff}$, d.h. entspricht er einer vollbesetzten Matrix, wird er mit dem entsprechenden Teil der FE-Matrix besetzt.
- Ist ein Block $b = \tau \times \sigma \subset I \times I$ nicht zulässig und $\text{level}(b) < p_{eff}$, d.h. entspricht er einem 4×4-Block, werden dessen 16 Teilblöcke weiter auf Zulässigkeit untersucht.

6.2 Die approximierte $\mathcal{H}$-Arithmetik in C

Eine der fundamentalen Eigenschaften aller approximierten $\mathcal{H}$-Operationen ist deren blockweise rekursive Struktur, die sich auch als wesentliches Merkmal

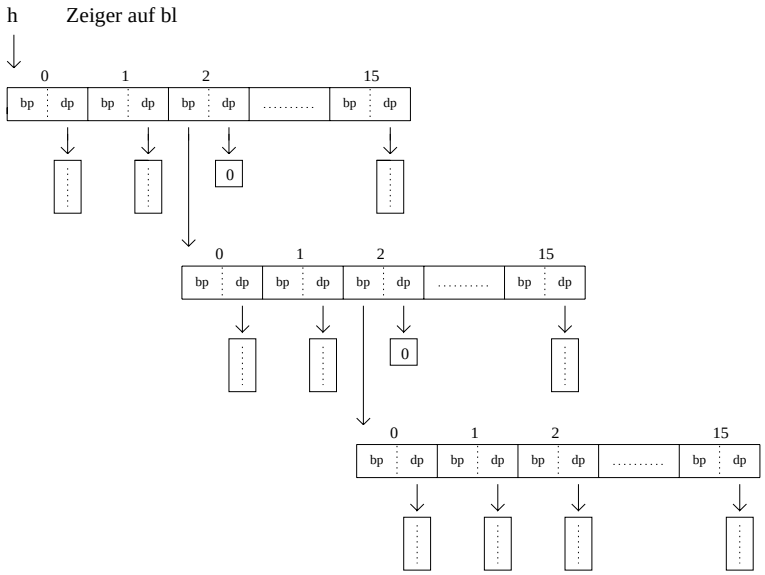

Abbildung 6.2: `h` zeigt auf einen $\searrow$-Block der hierarchischen Tiefe 3

in den entsprechenden C-Routinen wiederfindet. Die jeweiligen Operationen auf den vollbesetzten bzw. Rk-Matrixblöcken werden, soweit möglich, von Funktionen der BLAS- und LAPACK-Bibliotheken ausgeführt.

Die Matrix-Vektor-Multiplikation einer 4×4-$\mathcal{H}$-Blockmatrix mit einer Matrix, deren Spalten den zu multiplizierenden Vektoren entsprechen, wird jeweils auf 16 Matrix-Vektor-Multiplikationen auf dem nächsthöheren Level zurückgeführt. Die „echten“ Multiplikationen mit den zulässigen Rk-Teilblöcken bzw. den vollbesetzten Teilmatrizen werden von der BLAS-Routine `dgemm` durchgeführt.

Die $\mathcal{H}$-Matrix-Addition erfolgt ebenso blockweise rekursiv durch exakte Addition der vollbesetzten Blöcke und Rk-Addition der Rk-Blöcke (siehe Abschnitt 3.4.2). Dabei werden die exakten Matrixsummen mit der BLAS-Routine `daxpy` berechnet. Pro Rk-Summe $R_1 +_{Rk} R_2$ zweier Rk-Matrizen R_1 und R_2 wird ein unsymmetrisches $2k$-dimensionales Eigenwertproblem mit der LAPACK-Routine `dgeev` gelöst. Die damit errechneten Eigenvektoren von $R^T R$, $R = R_1 + R_2$, werden mittels orthogonaler Iteration bzgl. $R^T R$ nachverbessert, wobei die erforderliche QR-Zerlegung durch die LAPACK-Funktionen `dgeqrf` und `dorgqr` erfolgt. Numerische Experimente haben gezeigt, dass ein bis zwei Nachiterationen i. Allg. vollständig ausreichend sind (siehe Abschnitt 3.2).

Die $\mathcal{H}$-Matrix-Multiplikation wird ebenfalls rekursiv realisiert durch Rückführung auf „echte“ Matrix-Matrix-Multiplikationen sowie exakte und Rk-Additionen von je 4 vollbesetzten und Rk-Matrizen (siehe Abschnitt 3.4.3). Dabei werden die Multiplikationen von vollbesetzten und Rk-Matrizen durch die BLAS3-Routine `dgemm` abgehandelt. Die Multiplikation einer Blockmatrix mit einer Rk-Matrix entspricht einer Matrix-Vektor-Multiplikation derselben Blockmatrix mit dem entsprechenden k-spaltigen „Vektor“. Die Rk-Abschneidungen

der Summe von 4 Rk-Matrizen unterschiedlichen Ranges und eines Rk-4 × 4-Blocks zu einer Rk-Matrix sind analog zur Rk-Addition zweier Rk-Matrizen aufgebaut. Die Rk-Abschneidung einer vollbesetzten Teilmatrix wird schließlich mit der LAPACK-Routine `dgesvd` berechnet.

Was die $\mathcal{H}$-Invertierung betrifft ist zu beachten, dass neben der approximierten Addition von 4 $\mathcal{H}$-Formaten bzw. von 4 vollbesetzten und Rk-Matrizen im Zuge des 4×4-Block-Gauß-Algorithmus auch $\mathcal{H}$-Summen von 3 sowie nur 2 Summanden (4×4-Blöcke oder vollbesetzte oder Rk-Matrizen) benötigt werden. Diese sind genau gleich strukturiert wie die entsprechenden $\mathcal{H}$-Additionen von 4 Matrizen. Die exakte Invertierung eines vollbesetzten Matrixblocks erfolgt durch die LAPACK-Routinen `dgetrf` und `dgetri`.

Die Funktionen zur Lösung der $\mathcal{H}$-Vorwärtssubstitution mit einer mehrspaltigen Matrix als rechten Seite sowie der $\mathcal{H}$-Matrix-Vorwärtssubstitution mit einer $\mathcal{H}$-Matrix als rechten Seite verwenden die LAPACK-Routinen `dtrtrs` und `dtrsm`, die Funktionen für die respektiven Rückwärtssubstitutionen kommen allein mit der Routine `dtrtrs` aus.

Die hierarchische Cholesky-Zerlegung einer symmetrisch positiv definiten $\mathcal{H}$-Matrix wird unter Verwendung der LAPACK-Routinen `dpotrf` und `dtrsm` berechnet. Die C-Funktion zur hierarchischen LDL^T-Zerlegung ist jener für die $\mathcal{H}$-Cholesky-Zerlegung nachgebildet, wobei in Abwesenheit einer geeigneten LAPACK-Routine zur LDL^T-Zerlegung eines vollbesetzten symmetrisch indefiniten Matrixblocks eine dementsprechende Subroutine unter Verwendung der LAPACK- und BLAS3-Routinen `dtrsm`, `dgemm` und `dsyr2k` in Anlehnung an die Funktion `dpotrf` selbst geschrieben wurde.

Die in Abschnitt 3.4.7 vorgestellten Algorithmen 3.5 und 3.6 zur Berechnung einer approximierten $\mathcal{H}$-QR-Zerlegung einer nichtsingulären $\mathcal{H}$-Matrix konnten schließlich allein unter Verwendung der C-Funktionen für die einzelnen $\mathcal{H}$-Operationen und $\mathcal{H}$-Zerlegungen implementiert werden.

6.3 Die L^2-orthogonale Projektion von $L^2(\Omega)$ auf den FE-Raum V_h

Um die gegebenen Anfangswerte aus $H_0^1(\Omega)$ bzw. $L^2(\Omega)$ in die in Kapitel 5 erarbeiteten Lösungsalgorithmen auf dem Raum V_h der stückweise linearen C^0-Dreieckselemente einzubauen, benötigen wir noch eine Projektion von $L^2(\Omega)$ auf V_h. In Abschnitt 2.1 haben wir dafür die L^2-orthogonale Projektion $Q_h : L^2(\Omega) \to V_h$ gewählt, d.h.

$$(u, v_h)_{L^2} = (Q_h u, v_h)_{L^2} \quad \forall\, v_h \in V_h$$

für $u \in L^2(\Omega)$.

Wir suchen nun die Koeffizienten $(\mathbf{Q_h u})_j, 1 \le j \le N$, von $Q_h u \in V_h$ in der nodalen FE-Basis $\{\psi_j^h\}_{j=1}^N$ auf dem regelmäßig triangulierten Einheitsquadrat

Ω, d.h.

$$Q_h u = \sum_{j=1}^{N} (\mathbf{Q_h u})_j \psi_j^h.$$

Sei $u \in L^2(\Omega)$ gegeben. Dann gilt

$$(Q_h u, v_h)_{L^2} = \langle M_h \mathbf{Q_h u}, \mathbf{v}_h \rangle_2$$

mit $v_h = \sum_{j=1}^{N} (\mathbf{v}_h)_j \psi_j^h$ sowie

$$\begin{aligned} (Q_h u, v_h)_{L^2} &= (u, v_h)_{L^2} = (u, \textstyle\sum_{j=1}^{N} (\mathbf{v}_h)_j \psi_j^h)_{L^2} \\ &= \textstyle\sum_{j=1}^{N} (\mathbf{v}_h)_j (u, \psi_j^h)_{L^2} = \langle ((u, \psi_j^h)_{L^2}), \mathbf{v}_h \rangle_2. \end{aligned}$$

Also ist

$$M_h \mathbf{Q_h u} = ((u, \psi_j^h)_{L^2}) \tag{6.1}$$

und der Koeffizientenvektor $\mathbf{Q_h u}$ nach Berechnung aller Skalarprodukte $(u, \psi_j^h)_{L^2}, 1 \leq j \leq N$, durch Lösung eines LGS mit der Massenmatrix M_h erhältlich.

Bezeichne $\mathcal{T}_h$ die regelmäßige Triangulierung von $\bar{\Omega}$. Seien mit $\xi_j, 1 \leq j \leq N$, sämtliche inneren Gitterpunkte des regelmäßigen Dreiecksgitters bezeichnet (siehe Abbildung 6.3).

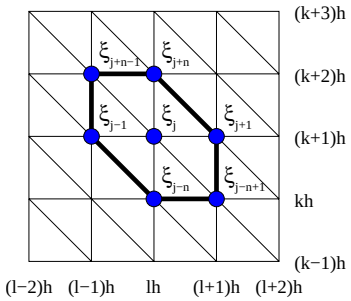

Abbildung 6.3: Träger der nodalen Basisfunktion ψ_j^h

Dann ist $\xi_j = (lh, (k+1)h)^T$, wobei $j = kn + l$ mit geeigneten $0 \leq k \leq n-1$ und $1 \leq l \leq n$ ist. Damit gilt

$$(u, \psi_j^h)_{L^2} = \int_\Omega u \psi_j^h dx = \sum_{T \in \mathcal{T}_h} \int_T u \psi_j^h dx = \sum_{\substack{T \in \mathcal{T}_h, \\ \xi_j \in T}} \int_T u \psi_j^h dx.$$

Wir definieren nun das Referenzdreieck

$$\hat{T} := \{\hat{x} \in \Omega : \hat{x}_1, \hat{x}_2 \geq 0, \hat{x}_1 + \hat{x}_2 \leq 1\}$$

(siehe Abbildung 6.4) sowie die bijektive affine Abbildung

$$\phi_T : \hat{T} \to T; \hat{x} \mapsto B_T \hat{x} + b_T \text{ mit } B_T \in GL(2, \mathbb{R}).$$

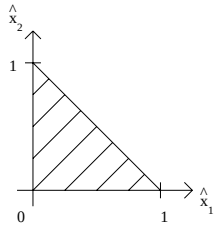

Abbildung 6.4: Standardelement $\hat{T} := \{\hat{x} \in \Omega : \hat{x}_1, \hat{x}_2 \geq 0, \hat{x}_1 + \hat{x}_2 \leq 1\}$

Für die Dreiecke T_l (siehe Abbildung 6.5) gilt

$$B_{T_l} = hI, x = \phi_{T_l}(\hat{x}) = h\hat{x} + b \text{ mit } b = \begin{pmatrix} ih \\ jh \end{pmatrix}, \ 0 \leq i, j \leq n,$$

für die Dreiecke T_r (siehe Abbildung 6.5) ist

$$B_{T_r} = -hI, x = \phi_{T_r}(\hat{x}) = -h\hat{x} + b \text{ mit } b = \begin{pmatrix} ih \\ jh \end{pmatrix}, \ 1 \leq i, j \leq n+1.$$

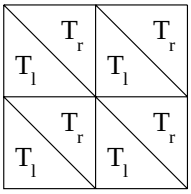

Abbildung 6.5: Dreiecke T_l ($l \mathrel{\widehat{=}}$ links) und T_r ($r \mathrel{\widehat{=}}$ rechts)

Damit ist

$$(u, \psi_j^h)_{L^2} = \sum_{\substack{T \in \mathcal{T}_h, \\ \xi_j \in T}} \int_{\hat{T}} (u \circ \phi_T)(\hat{x})(\psi_j^h \circ \phi_T)(\hat{x}) \underbrace{|\det B_T|}_{=h^2} d\hat{x}. \tag{6.2}$$

Pro Knoten $\xi_j, j = kn + l$ mit geeigneten $0 \leq k \leq n-1$ und $1 \leq l \leq n$, wird in (6.2) nur über die drei Dreiecke T_1, T_3, T_5 der Form T_r mit

$$\begin{aligned}
\phi_{T_1}(\hat{x}) &= -h\hat{x} + b_1 = -h\hat{x} + \begin{pmatrix} l \\ k+2 \end{pmatrix} h, \\
\phi_{T_3}(\hat{x}) &= -h\hat{x} + b_3 = -h\hat{x} + \begin{pmatrix} l+1 \\ k+1 \end{pmatrix} h, \\
\phi_{T_5}(\hat{x}) &= -h\hat{x} + b_5 = -h\hat{x} + \begin{pmatrix} l \\ k+1 \end{pmatrix} h
\end{aligned}$$

und über die drei Dreiecke T_2, T_4, T_6 der Form T_l mit

$$\phi_{T_2}(\hat{x}) = h\hat{x} + b_2 = h\hat{x} + \begin{pmatrix} l \\ k+1 \end{pmatrix} h,$$
$$\phi_{T_4}(\hat{x}) = h\hat{x} + b_4 = h\hat{x} + \begin{pmatrix} l \\ k \end{pmatrix} h,$$
$$\phi_{T_6}(\hat{x}) = h\hat{x} + b_6 = h\hat{x} + \begin{pmatrix} l-1 \\ k+1 \end{pmatrix} h$$

summiert (siehe Abbildung 6.6).

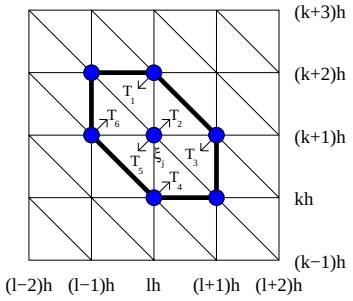

Abbildung 6.6: Sechs Dreiecke des Trägers einer nodalen Basisfunktion

Weiter ist $\psi_j^h \circ \phi_{T_1} = \hat{x}_2$, $\psi_j^h \circ \phi_{T_3} = \hat{x}_1$, $\psi_j^h \circ \phi_{T_5} = 1 - \hat{x}_1 - \hat{x}_2$ sowie $\psi_j^h \circ \phi_{T_2} = 1 - \hat{x}_1 - \hat{x}_2$, $\psi_j^h \circ \phi_{T_4} = \hat{x}_2$, $\psi_j^h \circ \phi_{T_6} = \hat{x}_1$. Damit ist

$$\begin{aligned}
(u, \psi_j^h)_{L^2} &= h^2 \sum_{T \in \{T_1,\ldots,T_6\}} \int_{\hat{T}} (u \circ \phi_T)(\hat{x})(\psi_j^h \circ \phi_T)(\hat{x})\, d\hat{x} \\
&= h^2 \Bigg(\int_{\hat{T}} u\left(h\left(-\hat{x} + \begin{pmatrix} l \\ k+2 \end{pmatrix}\right)\right) \hat{x}_2\, d\hat{x} + \\
&\qquad \int_{\hat{T}} u\left(h\left(-\hat{x} + \begin{pmatrix} l+1 \\ k+1 \end{pmatrix}\right)\right) \hat{x}_1\, d\hat{x} + \\
&\qquad \int_{\hat{T}} u\left(h\left(-\hat{x} + \begin{pmatrix} l \\ k+1 \end{pmatrix}\right)\right) (1 - \hat{x}_1 - \hat{x}_2)\, d\hat{x} + \\
&\qquad \int_{\hat{T}} u\left(h\left(\hat{x} + \begin{pmatrix} l \\ k+1 \end{pmatrix}\right)\right) (1 - \hat{x}_1 - \hat{x}_2)\, d\hat{x} + \\
&\qquad \int_{\hat{T}} u\left(h\left(\hat{x} + \begin{pmatrix} l \\ k \end{pmatrix}\right)\right) \hat{x}_2\, d\hat{x} + \\
&\qquad \int_{\hat{T}} u\left(h\left(\hat{x} + \begin{pmatrix} l-1 \\ k+1 \end{pmatrix}\right)\right) \hat{x}_1\, d\hat{x} \Bigg)
\end{aligned}$$

$$\begin{aligned}
\dot{=} \quad & \tfrac{h^2}{6}\left(u\left(h\begin{pmatrix} l-\frac{1}{3} \\ k+\frac{5}{3}\end{pmatrix}\right)+u\left(h\begin{pmatrix} l+\frac{2}{3} \\ k+\frac{2}{3}\end{pmatrix}\right)+\right. \\
& u\left(h\begin{pmatrix} l-\frac{1}{3} \\ k+\frac{2}{3}\end{pmatrix}\right)+u\left(h\begin{pmatrix} l+\frac{1}{3} \\ k+\frac{4}{3}\end{pmatrix}\right)+ \\
& \left. u\left(h\begin{pmatrix} l+\frac{1}{3} \\ k+\frac{1}{3}\end{pmatrix}\right)+u\left(h\begin{pmatrix} l-\frac{2}{3} \\ k+\frac{4}{3}\end{pmatrix}\right)\right),
\end{aligned}$$

wobei die Integrale über $\hat{T}$ im letzten Schritt durch Mittelwertsquadratur approximiert wurden (Dies reicht für stetige Lagrange-Elemente):

$$\int_{\hat{T}} f\, d\hat{x} \dot{=} |\hat{T}| \cdot f(\xi_{\hat{T}})$$

mit dem Flächeninhalt $|\hat{T}| = \frac{1}{2}$ von $\hat{T}$ und dem Schwerpunkt $\xi_{\hat{T}}$ von $\hat{T}$, der sich zu

$$\begin{aligned}
(\xi_{\hat{T}})_1 &= \tfrac{1}{|\hat{T}|}\int_{\hat{T}} \hat{x}_1 d\hat{x}_1 d\hat{x}_2 = 2\int_{\hat{x}_2=0}^{1}\int_{\hat{x}_1=0}^{1-\hat{x}_2} \hat{x}_1 d\hat{x}_1 d\hat{x}_2 \\
&= 2\int_0^1 \tfrac{(1-\hat{x}_2)^2}{2}\, d\hat{x}_2 = \left[-\tfrac{(1-\hat{x}_2)^3}{3}\right]_0^1 = \tfrac{1}{3}
\end{aligned}$$

und aus Symmetriegründen

$$(\xi_{\hat{T}})_2 = \frac{1}{3}$$

errechnet.

Vor der Lösung des LGS (6.1) in $\mathcal{H}$-Arithmetik muss noch dessen rechte Seite, das ist der zeilenweise von links unten nach rechts oben durchnumerierte Koeffizientenvektor $((u, \psi_j^h)_{L^2})$ hierarchisiert werden, d.h. die einzelnen Komponenten müssen entsprechend der $\mathcal{H}$-Struktur der $\mathcal{H}$-Matrizen angeordnet werden, also mit der Permutationsmatrix P aus Abschnitt 3.3.1 multipliziert werden. Es gilt nämlich

$$Ax = b \iff PAP^T Px = Pb \iff A_{\mathcal{H}} x_{\mathcal{H}} = b_{\mathcal{H}}$$

mit der hierarchisierten Matrix $A_{\mathcal{H}}$ und den hierarchisierten Vektoren $x_{\mathcal{H}}$ und $b_{\mathcal{H}}$.
Die C-Routine, die die Multiplikation eines N-dimensionalen Vektors mit P bewerkstelligt, ist völlig analog zu jener aufgebaut, die die FE-Matrizen hierarchisiert. Für die C-Funktion zur Rücktransformation, das ist die Multiplikation eines N-dimensionalen Vektors mit P^T, sind in der Routine zur Hintransformation nur die Indizes der jeweiligen Input- und Output-Vektoren zu vertauschen.

Was schließlich die L^2-orthogonale Projektion auf V_h betrifft, kann die Lösung von LGS (6.1) entweder durch Matrix-Vektor-Multiplikation mit der bereits berechneten $\mathcal{H}$-Inversen $M_h^{-1_{\mathcal{H}}}$ von M_h oder durch $\mathcal{H}$-Vorwärts- und $\mathcal{H}$-Rückwärtssubstitution mit dem bereits berechneten $\mathcal{H}$-Cholesky-Faktor L_h von M_h erfolgen.

6.4 Auswertung transzendenter Matrixfunktionen und Lösung der 2D Wellengleichung in C

Nach erfolgter Implementierung der gesamten hierarchischen Arithmetik aus Kapitel 3.4 sowie der L^2-Projektion in $\mathcal{H}$-Arithmetik wenden wir uns nun der Approximation transzendenter Matrixfunktionen und der Lösung der 2D Wellengleichung und Wärmeleitungsgleichung zu.

Dazu benötigen wir die kleinsten Eigenwerte der Matrix $M_h^{-1}A_h$ samt zugehörigen Eigenvektoren. In der C-Funktion, die Algorithmus 4.3 implementiert, werden neben den Routinen für die $\mathcal{H}$-LDL^T-Zerlegung, die $\mathcal{H}$-Matrix-Vektor-Multiplikation und die $\mathcal{H}$-Vorwärts- und $\mathcal{H}$-Rückwärtssubstitution noch die LAPACK-Routinen `dgeqrf` und `dorgqr` zur QR-Zerlegung, `dgees` zur Schur-Zerlegung und `dgels` zur Lösung linearer Ausgleichsprobleme sowie die BLAS-Routinen `dcopy`, `dscal`, `daxpy`, `dgemv` und `dgemm` verwendet.

Damit können die durch Algorithmus 4.5 gegebenen Rk-Approximationen an Matrixfunktionen $f(\tau^2 c^2 M_h^{-1}A_h)$ nach Multiplikation der REV mit M_h zum Erhalt der LEV von $M_h^{-1}A_h$ sofort unter Verwendung der BLAS1-Routinen `ddot` und `dscal` bestimmt werden.

Mit den bereits ermittelten $\mathcal{H}$-EW und $\mathcal{H}$-EV sowie durch Anwendung der C-Routine für die L^2-Projektion auf V_h lassen sich schließlich die N-dimensionalen Lösungsvektoren der diskreten 2D Wellengleichung und Wärmeleitungsgleichung in der Knotendarstellung nach Algorithmus 5.2 und 5.4 berechnen.

6.5 Zusammenfassung

- Die Implementierung der hierarchischen Arithmetik in C orientiert sich stark an deren blockweise rekursiven Beschreibung und spiegelt die hierarchische Struktur der einzelnen Operationen sehr genau wieder.
- Die numerischen Resultate bestätigen die fast lineare Komplexität der approximierten $\mathcal{H}$-Operationen.
- Von einer ganzen $\mathcal{H}$-Matrix ausgehend wird jeweils bis zu den einzelnen vollbesetzten und Rk-Matrixblöcken vorgegangen, auf denen die anstehenden Operationen – soweit möglich unter Ausnutzung der optimierten BLAS- und LAPACK-Routinen – durchgeführt werden, um dann die jeweiligen Zwischenresultate zum hierarchischen Gesamtergebnis zusammenzufügen.
- Die L^2-orthogonale Projektion auf den FE-Raum V_h sowie die Bestimmung von Eigenpaaren der Matrix $M_h^{-1}A_h$ und damit die Lösung der 2D Wellengleichung und Wärmeleitungsgleichung erfolgen schließlich im vorhin angelegten hierarchischen Setting mit fast linearem Aufwand.

Literaturverzeichnis

[1] M. Arioli, B. Codenotti, C. Fassino. *The Padé Method for Computing the Matrix Exponential.* Linear Algebra Appl. 240:111-130, 1996.

[2] F. A. Bornemann. *An Adaptive Multilevel Approach to Parabolic Equations in Two Space Dimensions.* Technical Report TR 91-7, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Dissertation, 1991.

[3] D. Braess. *Finite Elemente.* Springer-Verlag, Berlin, Heidelberg, New York, 2. Auflage, 1997.

[4] G. Coté, R. V. M. Zahar. *Computation of the Matrix Exponential.* Congressus Numerantium 75, 21-28, 1990.

[5] P. Deuflhard, A. Hohmann. *Numerische Mathematik - Eine algorithmisch orientierte Einführung.* Walter de Gruyter, Berlin, New York, 1991.

[6] P. Deuflhard, F. Bornemann. *Numerische Mathematik II - Integration gewöhnlicher Differentialgleichungen.* Walter de Gruyter, Berlin, New York, 1994.

[7] V. Druskin, L. Knizhnerman. *Krylov Subspace Approximation of Eigenpairs and Matrix Functions in Exact and Computer Arithmetic.* Numer. Linear Algebra Appl. 2, No.3, 205-217, 1995.

[8] N. Dunford, J. T. Schwartz. *Linear Operators: Part I.* Interscience Publishers, Inc., New York, 1958.

[9] L. C. Evans. *Partial Differential Equations.* AMS Graduate Studies in Mathematics, Vol. 19, 1998.

[10] I. P. Gavrilyuk. *Strongly P-Positive Operators and Explicit Representations of the Solutions of Initial Value Problems for Second-Order Differential Equations in Banach Space.* Journal of Mathematical Analysis and Applications 236, 327-349, 1999.

[11] I. P. Gavrilyuk, V. L. Makarov. *Representation and Approximation of the Solution of an Initial Value Problem for a First Order Differential Equation in Banach Spaces.* Journal for Analysis and its Applications, Vol. 15, No. 2, 495-527, 1996.

[12] I. P. Gavrilyuk, V. L. Makarov. *Exponentially convergent parallel discretization methods for the first order evolution equations.* Preprint NTZ 12/2000, Universität Leipzig.

[13] I. P. Gavrilyuk, V. L. Makarov. *Explicit and Approximate Solutions of Second-Order Evolution Differential Equations in Hilbert Space.* Numer. Methods Partial Differential Eq. 15, 111-131, 1999.

[14] I. P. Gavrilyuk, W. Hackbusch, B. N. Khoromskij. *$\mathcal{H}$-matrix approximation for the operator exponential with applications.* Preprint 42 (2000), Max-Planck-Institute for Mathematics in the Sciences, Leipzig. To appear in Numer. Math., 2001.

[15] I. P. Gavrilyuk, W. Hackbusch, B. N. Khoromskij. *$\mathcal{H}$-matrix approximation for elliptic solution operators in cylindric domains.* East-West J. Numer. Math. 9, No. 1, 25-58, 2001.

[16] G. Golub, C. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore and London, 3rd edition, 1996.

[17] L. Grasedyck. *Theorie und Anwendungen Hierarchischer Matrizen.* Dissertation, Universität Kiel, 2001.

[18] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* B. G. Teubner, Stuttgart, 1986.

[19] W. Hackbusch. *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: Introduction to $\mathcal{H}$-matrices.* Computing 62, No.2, 89-108, 1999.

[20] W. Hackbusch, B. N. Khoromskij. *A sparse $\mathcal{H}$-matrix arithmetic. Part II: Application to multi-dimensional problems.* Computing 64, No.1, 21-47, 2000.

[21] W. Hackbusch, B. N. Khoromskij. *A sparse $\mathcal{H}$-matrix arithmetic: General complexity estimates.* J. Comput. Appl. Math. 125, No.1-2, 479-501, 2000.

[22] W. Hackbusch, B. N. Khoromskij. *$\mathcal{H}$-matrix approximation on graded meshes.* The mathematics of finite elements and applications X, MAFELAP 1999, J.R. Whiteman (ed.), Elsevier, Amsterdam, Chapter 19, 307-316, 2000.

[23] W. Hackbusch, B. N. Khoromskij. *Towards $\mathcal{H}$-matrix approximation of linear complexity.* Operator Theory: Advances and Applications, Vol. 121, Birkhäuser Verlag, 194-220, 2001.

[24] W. Hackbusch, B. N. Khoromskij, S. Sauter. *On $\mathcal{H}^2$-matrices.* In: Lectures on Applied Mathematics (H.-J. Bungartz, R. Hoppe, C. Zenger, eds.), Springer-Verlag, 9-29, 2000.

[25] W. Hackbusch, B. N. Khoromskij. *Blended Kernel Approximation in the $\mathcal{H}$-Matrix Techniques.* Preprint Nr. 66, MPI MIS Leipzig, 2000; Numer. Lin. Alg. with Appl., to appear.

[26] E. Hairer, S. P. Nørsett, G. Wanner. *Solving Ordinary Differential Equations I.* Springer-Verlag, Berlin, Heidelberg, New York, 2nd Revised Edition, 1993.

[27] N. J. Higham. *Computing the polar decomposition – with applications.* SIAM J. Sci. Stat. Comput., Vol. 7, No. 4, 1160-1174, 1986.

[28] M. Hochbruck, Ch. Lubich. *Exponential integrators for large systems of differential equations.* SIAM J. Sci. Comput. 19, No.5, 1552-1574, 1998.

[29] M. Hochbruck, Ch. Lubich. *A Gautschi-type method for oscillatory second-order differential equations.* Numer. Math. 83, 403-426, 1999.

[30] M. Hochbruck, Ch. Lubich. *On Krylov subspace approximations to the matrix exponential operator.* SIAM J. Numer. Anal. 34, No.5, 1911-1925, 1997.

[31] L. Hörmander. *The Analysis of Linear Partial Differential Operators I.* Springer-Verlag, Berlin, Heidelberg, New York, 2nd edition, 1990.

[32] L. Hörmander. *The Analysis of Linear Partial Differential Operators II.* Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[33] R. A. Horn, C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[34] T. Kato. *Perturbation Theory for Linear Operators.* Springer-Verlag, Berlin, Heidelberg, New York, 2nd edition, 1984.

[35] B. W. Kernighan, D. M. Ritchie. *Programmieren in C.* Prentice-Hall Internat., 2. Ausgabe, 1990.

[36] B. N. Khoromskij. *Data-sparse Approximate Inverse in Elliptic Problems: Green's Function Approach.* Preprint Nr. 79, MPI MIS Leipzig, 2001.

[37] R. Kress. *Linear Integral Equations.* Springer-Verlag, Berlin, Heidelberg, New York, 2nd edition, 1999.

[38] C. Moler, C. Van Loan. *Nineteen Dubious Ways to Compute the Exponential of a Matrix.* SIAM Rev. Vol. 14, 801-836, 1978.

[39] J. R. Roche. *On the Sensitivity of the Matrix Exponential Problem.* R.A.I.R.O. Numerical Analysis Vol. 15, No. 3, 249-255, 1981.

[40] Y. Saad. *Iterative methods for sparse linear systems.* PWS Publishing Company, Boston u.a., 1996.

[41] A. A. Samarskii, I. P. Gavrilyuk, V. L. Makarov. *Stability and regularization of three-level difference schemes with unbounded operator coefficients in a Banach space.* SIAM Journal on Numerical Analysis, Vol. 39, No. 2, 708-723, 2001.

[42] M. Schemann, F. A. Bornemann. *An adaptive Rothe method for the wave equation.* Comput. Visual. Sci. 1:137-144, 1998.

[43] G. W. Stewart. *Simultaneous Iteration for Computing Invariant Subspaces of Non-Hermitian Matrices.* Numer. Math. 25, 123-136, 1976.

[44] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems.* Springer-Verlag, Berlin, Heidelberg, New York, 1997.

[45] C. Van Loan. *The Sensitivity of the Matrix Exponential.* SIAM J. Numer. Anal. Vol. 14, No. 6, 971-981, 1977.

[46] R. C. Ward. *Numerical Computation of the Matrix Exponential with Accuracy Estimate.* SIAM J. Numer. Anal. Vol. 14, No. 4, 600-610, 1977.

[47] J. Wloka. *Partielle Differentialgleichungen.* B. G. Teubner, Stuttgart, 1982.