Eine Mehrgitter-Methode zur Lösung des Eigenwertproblems der komplexen Helmholtzgleichung

Vom Fachbereich Mathematik und Informatik
der Freien Universität Berlin
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
genehmigte Dissertation

vorgelegt von
Diplom-Mathematiker Tilmann Friese

Berlin, 1998

Gutachter:	Prof. Dr. Peter Deuflhard, Berlin Prof. Dr. Harry Yserentant, Tübingen
Disputation:	27. November 1998

Inhaltsverzeichnis

1 Einleitung

Diese Arbeit hat die numerische Lösung des Eigenwertproblems der skalaren Helmholtzgleichung

$$-\Delta u(x,y) - f(x,y)u(x,y) = \lambda u(x,y)$$

auf einem beschränkten, zweidimensionalen Gebiet mit Dirichlet-Randbedingung zum Thema, wobei die beschränkte Funktion f sowohl reell- als auch komplexwertig sein kann. Die Aufgabe der Berechnung von Eigenlösungen dieser partiellen Differentialgleichung unter den genannten Voraussetzungen tritt in vielen Bereichen der Natur- und Ingenieurwissenschaften auf. In der integrierten Optik läßt sich die Bestimmung ausbreitungsfähiger Lichtmoden von in einer Raumrichtung invarianten optischen Strukturen auf diese Problemstellung zurückführen. In der Quantenmechanik beschreiben bestimmte Eigenfunktionen stationäre Zustände eines quantenmechanischen Systems, in der Akustik schließlich werden die Eigenlösungen der Helmholtzgleichung zur Untersuchung des Resonanzverhaltens mechanischer Komponenten benötigt. Dabei ist es in vielen Anwendungen, nämlich bei der Beschreibung von Systemen mit Erhaltungseigenschaften, ausreichend, nur den Fall einer reellen Funktion f, das heißt eines selbstadjungierten Helmholtzoperators, zu betrachten. Modelliert man jedoch auch Systeme mit Gewinn bzw. Verlust, so wird man auf ein Eigenwertproblem mit komplexwertiger Funktion f, also mit einem nichtselbstadjungierten Operator, geführt. Ein typisches Beispiel dafür ist eine wellenleitende optische Struktur mit Dämpfung oder Verstärkung in Ausbreitungsrichtung. Während nun das Studium der analytischen Eigenschaften des Problems mit reeller Funktion f (selbstadjungiertes Eigenwertproblem) wie auch mit komplexer Funktion f (nichtselbstadjungiertes Eigenwertproblem) als abgeschlossenes Forschungsgebiet angesehen werden kann, stellt die Entwicklung von Algorithmen zur numerischen Lösung des Eigenwertproblems, insbesondere im nichtselbstadjungierten Fall, ein aktuelles Thema mathematischer Forschungsarbeit dar.

Den Ausgangspunkt der Konstruktion numerischer Verfahren zur Behandlung des Eigenwertproblems bildet eine geeignete Diskretisierung des kontinuierlichen Problems. Dabei wird man auf ein algebraisches Eigenwertproblem der Form

$$Au = \lambda Bu$$

mit den im allgemeinen hochdimensionalen und dünnbesetzten Matrizen A und B geführt. Bei der in dieser Arbeit benutzten Finite-Elemente-Diskretisierung über Dreiecksgittern ist die Matrix A für reell- bzw. komplexwertige Funktionen f selbstadjungiert bzw. nichtselbstadjungiert, die Matrix B stets selbstadjungiert und positiv definit. Um eine möglichst gute Approximation der gesuchten kontinuierlichen Eigenwerte und Eigenfunktionen zu erreichen, wird die näherungsweise Berechnung dieser Lösungen nicht auf der Grundlage eines einzigen festen Gitters, sondern unter Zugrundelegung einer Folge von adaptiv verfeinerten Triangulierungen durchgeführt. Aufgrund

der so entstehenden hierarchischen Mehrgitter-Struktur bieten sich in natürlicher Weise Mehrgitter-Methoden zur Lösung der auftretenden diskreten Probleme an. Diese Verfahren zählen zu den schnellsten Lösern der im Zusammenhang mit der Diskretisierung partieller Differentialgleichungen vorkommenden algebraischen Systeme, da der Rechenaufwand nur proportional zur Anzahl der zu bestimmenden Unbekannten ansteigt. Das konkrete Anliegen dieser Arbeit besteht nun in der Verallgemeinerung eines für den selbstadjungierten Fall und die Berechnung einer einfachen Eigenlösung existierenden Mehrgitter-Verfahrens von Mandel und McCormick [42] auf den nichtselbstadjungierten Fall und die simultane Bestimmung mehrerer Eigenlösungen. Die entwickelte Methode basiert auf der Schur-Zerlegung beliebiger Matrizen und gestattet die direkte Bestimmung der gewünschten Lösungen des diskreten Eigenwertproblems. Der gewählte Zugang erlaubt eine einheitliche Behandlung des selbstadjungierten und nichtselbstadjungierten Falls, wobei (insbesondere für Probleme mit komplexer Funktion f) häufig benutzte Techniken wie Störungsrechnung oder Pfadverfolgung nicht benötigt werden.
Die vorliegende Arbeit gliedert sich im wesentlichen in drei Kapitel. Im einführenden Kapitel 2 wird zunächst gezeigt, daß die Bestimmung geführter Lichtmoden optischer Komponenten unter geeigneten Modellannahmen durch Lösung des Eigenwertproblems der skalaren Helmholtzgleichung erfolgen kann. Anschließend wird die Variationsformulierung dieses Eigenwertproblems motiviert und hergeleitet. In den beiden folgenden Abschnitten werden zum einen die analytischen Eigenschaften des Problems untersucht, wobei Punkte wie die Struktur des Spektrums des Helmholtzoperators und die Vollständigkeit der Eigenlösungen behandelt werden, zum anderen wird darauf aufbauend eine Sensitivitätsanalyse in Bezug auf die Funktion f durchgeführt, das heißt die Abhängigkeit der Eigenlösungen von Störungen in f diskutiert. Der Gegenstand des nächsten Abschnittes ist die Diskretisierung des kontinuierlichen Problems aus der Variationsformulierung mittels finiter Elemente. Es werden Abschätzungen für die Approximationsfehler sowohl in den Eigenwerten als auch in den korrespondierenden Eigenfunktionen angegeben. Im letzten Abschnitt des Kapitels 2 werden ein abstraktes Verfahren, die Projektionsmethode, zur Lösung von algebraischen Eigenwertproblemen und einige darauf beruhende Lösungsverfahren, die Krylovraum-Methoden, dargestellt.
Im sich anschließenden Kapitel 3, welches die Mehrgitter-Methoden zur Lösung von Eigenwertproblemen beinhaltet, wird einführend ein Überblick über die allgemeine Mehrgitteridee und existierende Mehrgitter-Verfahren gegeben. Der zweite Abschnitt behandelt die Lösung des selbstadjungierten Problems und hat die Verallgemeinerung der Rayleigh-Quotienten-Mehrgitter-Minimierung von Mandel und McCormick [42] auf die simultane Berechnung mehrerer Eigenlösungen zum Inhalt. Dazu wird zunächst vom Rayleigh-Quotienten zu einem allgemeineren Funktional zur Charakterisierung des von den gesuchten Eigenvektoren aufgespannten invarianten Unterraumes übergegangen. Es werden verschiedene Minimierungsverfahren zur simultanen Berechnung der kleinsten Eigenwerte und der korrespondierenden Eigenvektoren auf einem Gitter

vorgestellt, wobei speziell die Konvergenzeigenschaften der Methode der konjugierten Gradienten, insbesondere deren Glättungseigenschaft, näher untersucht werden. Nach diesen vorbereitenden Betrachtungen wird ein Mehrgitter-Verfahren zur Bestimmung des zu den kleinsten Eigenwerten gehörenden invarianten Unterraumes angegeben. Dieses läßt sich als Rekursion in Form von Eigenwertproblemen realisieren. Diese Formulierung erlaubt dann die Verallgemeinerung der Mehrgitter-Methode auf den nichtselbstadjungierten Fall, welche im dritten Abschnitt von Kapitel 3 vorgenommen wird. Abschließend wird gezeigt, daß das entwickelte Mehrgitter-Verfahren der Klasse der Projektionsmethoden angehört.
Im letzten Kapitel 4 dieser Arbeit werden einige numerische Beispielrechnungen (darunter anspruchsvolle Anwendungen aus der integrierten Optik), welche mit Hilfe der konstruierten Mehrgitter-Methode durchgeführt worden sind, dokumentiert. Insbesondere zeigt sich in diesen Tests die optimale Komplexität des Mehrgitter-Algorithmus, wobei ausschließlich adaptiv erzeugte Triangulierungen zugrunde gelegt wurden.

Danksagung. An dieser Stelle möchte ich all jenen herzlich danken, die mich bei der Anfertigung dieser Arbeit unterstützt und gefördert haben. Dabei geht mein erster Dank an Prof. Peter Deuflhard für die Betreuung dieser Arbeit und die hervorragenden Arbeitsbedingungen am Konrad-Zuse-Zentrum. Mein nächster Dank gilt meinem Kollegen Dr. Frank Schmidt für die aufgebrachte Geduld und Mühe bei der Beantwortung jeder meiner Fragen. Ich danke weiterhin Dr. Folkmar Bornemann, Prof. David E. Edmunds (University of Sussex at Brighton) und Dr. Dirk Werner (Freie Universität Berlin) für die gewährte Unterstützung bei der Erstellung des Abschnittes 2.3. Weiterer Dank gebührt Dr. Reinhard März (Siemens AG München) für die stets gute und anregende Kooperation und die Bereitstellung der Daten zur Simulation des optischen Verstärkers.
Die vorliegende Arbeit entstand im Rahmen des Programmes „Förderung von anwendungsorientierten Verbundprojekten auf dem Gebiet der Mathematik“ des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie in dem Projekt „Entwicklung von adaptiven Mehrgitter-Methoden zur Berechnung von Eigenlösungen der Helmholtzgleichung für Halbleiterstrukturen der integrierten Optik“ (Projektnummer 03-DE7ZIB-5). Dem BMBF möchte ich abschließend für die finanzielle Förderung meiner Tätigkeit danken.

2 Vorüberlegungen

2.1 Problemstellung

In diesem Abschnitt soll zunächst gezeigt werden, daß sich die Bestimmung von ausbreitungsfähigen Moden in optischen Strukturen, welche invariant bezüglich einer Raumrichtung sind, unter geeigneten Modellvereinfachungen auf die Lösung des Eigenwertproblems der skalaren Helmholtzgleichung zurückführen läßt.
Die Ausbreitung von Lichtwellen in optischen Komponenten wird allgemein durch die Maxwellschen Gleichungen für nichtmagnetische und ladungsfreie Medien beschrieben. Setzt man für die elektrische und die magnetische Feldstärke eine harmonische Zeitabhängigkeit voraus, wird man auf die zeitstationären Maxwellschen Gleichungen

$$\begin{aligned} \nabla \times \vec{E} &= -i\omega\mu\vec{H} \\ \nabla \times \vec{H} &= (i\omega\epsilon + \sigma)\vec{E} \\ \nabla \cdot \epsilon\vec{E} &= 0 \\ \nabla \cdot \vec{H} &= 0 \end{aligned} \tag{1}$$

geführt [18, S. 19 ff.]. Dabei bezeichnet $\vec{E}$ den elektrischen und $\vec{H}$ den magnetischen Feldstärkevektor im Ort, die Größe ω ist die Kreisfrequenz. Die (reelle) Dielektrizitätszahl ϵ und die Leitfähigkeit σ des Mediums hängen vom Ort und der Kreisfrequenz ab, die Permeabilität μ ist im gesamten Raum konstant.
Für die Analyse der meisten integriert optischen Komponenten wird jedoch ein vereinfachtes Modell verwendet [41, S. 12 ff.]. Durch Elimination der elektrischen Feldstärke aus den ersten beiden Maxwellschen Gleichungen erhält man unter Einführung der komplexen Dielektrizitätszahl $\tilde{\epsilon} = \epsilon - i\omega^{-1}\sigma$ die stationäre vektorielle Wellengleichung

$$\nabla \times \frac{1}{\tilde{\epsilon}} \left(\nabla \times \vec{H} \right) = \omega^2 \mu \vec{H}$$

für die magnetische Feldstärke, die unter Benutzung der Rechenregeln der Vektoranalysis in die äquivalente Gleichung

$$-\nabla \log \tilde{\epsilon} \times \nabla \times \vec{H} + \nabla \left(\nabla \cdot \vec{H} \right) - \Delta \vec{H} = \omega^2 \tilde{\epsilon} \mu \vec{H}$$

überführt werden kann. Unter Beachtung der letzten Maxwellschen Gleichung ergibt sich daraus zur Bestimmung der magnetischen Feldstärke die Gleichung

$$-\Delta \vec{H} - \omega^2 \tilde{\epsilon} \mu \vec{H} = \nabla \log \tilde{\epsilon} \times \nabla \times \vec{H} \quad . \tag{2}$$

Die elektrische Feldstärke $\vec{E}$ läßt sich nun über die zweite Maxwellsche Gleichung aus der magnetischen Feldstärke $\vec{H}$ zurückgewinnen. Neben den physikalisch interessierenden Lösungen erhält man so unter Umständen noch weitere, sogenannte „Geisterlösungen“, die die Divergenzbedingungen, das heißt die dritte und die vierte Maxwellsche Gleichung, verletzen.

Beispiel 1. In homogenen Medien hängt die komplexe Dielektrizitätszahl $\tilde{\epsilon}$ nur von der Kreisfrequenz ω ab. Legt man kartesische Koordinaten zugrunde, sind daher einfache Lösungen der Gleichung (2) durch ebene Wellen

$$\vec{H}(x,y,z) = \vec{A}e^{-i\vec{k}\cdot\vec{r}}$$

gegeben. Dabei bezeichnet $\vec{r} = (x,y,z)$ den Ortsvektor, $\vec{k} = (k_x, k_y, k_z)$ den Wellenvektor und $\vec{A} = (A_x, A_y, A_z)$ den Amplitudenvektor. Durch Einsetzen in (2) erhält man die Dispersionsrelation

$$\vec{k} \cdot \vec{k} = \omega^2 \tilde{\epsilon} \mu$$

für den Wellenvektor. Der elektrische Feldstärkevektor $\vec{E}$ ergibt sich aus der zweiten Maxwellschen Gleichung zu

$$\vec{E}(x,y,z) = -\frac{1}{\omega\tilde{\epsilon}} \left(\vec{k} \times \vec{A}\right) e^{-i\vec{k}\cdot\vec{r}} \quad .$$

Dabei ist in homogenen wie auch in rein dielektrischen Medien ($\sigma = 0$) aufgrund der Relation $\nabla \cdot (\nabla \times \vec{H}) = 0$ die dritte Maxwellsche Gleichung automatisch erfüllt. Die Bedingung der Divergenzfreiheit für $\vec{H}$ geht über in die Relation

$$\vec{k} \cdot \vec{A} = 0 \quad ,$$

liefert also zusätzlich eine Beschränkung in der Wahl des Amplitudenvektors. ◇

Da die Dielektrizitätszahl $\tilde{\epsilon}$ bei den betrachteten optischen Strukturen im allgemeinen nur sehr schwach vom Ort abhängt, kann der Kopplungsterm auf der rechten Seite von Gleichung (2) vernachlässigt werden. Man erhält so die vektorielle Helmholtzgleichung

$$-\Delta\vec{H} - \omega^2\tilde{\epsilon}\mu\vec{H} = 0 \quad , \tag{3}$$

die die Grundlage für die Simulation der wichtigsten optischen Bauteile bildet. In kartesischen Koordinaten ist die Gleichung (3) für jede einzelne Komponente des magnetischen Feldstärkevektors gültig. Somit braucht zwischen diesen nicht mehr unterschieden zu werden und es genügt, die skalare Helmholtzgleichung

$$-\Delta H - \omega^2\tilde{\epsilon}\mu H = 0 \tag{4}$$

zu betrachten.

Ein weiteres Charakteristikum der zu untersuchenden optischen Komponenten ist ihre Invarianz in einer Raumrichtung. Das bedeutet, die Geometrie der Struktur ändert sich in einer Richtung bezogen auf die Wellenlänge des Lichtes wenig oder gar nicht.

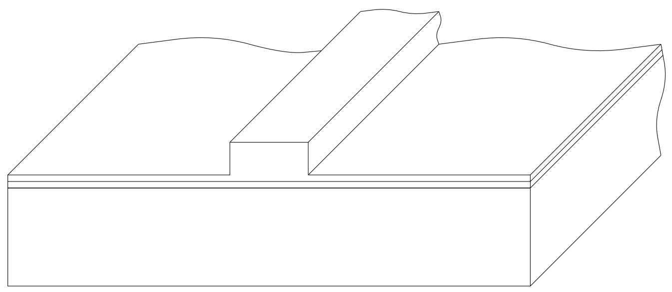

Abbildung 1: Aufbau eines Rippen-Wellenleiters

Beispiel 2. Eine typische Struktur der integrierten Optik ist der Rippen-Wellenleiter, dessen Aufbau in Abbildung 1 dargestellt ist. ◇

Im folgenden wird das Koordinatensystem so gelegt, daß diese ausgezeichnete Raumrichtung durch die z-Komponente des Ortsvektors beschrieben ist. Die Dielektrizitätszahl $\tilde{\epsilon}$ hängt daher nur noch von den Ortskoordinaten x und y und der Kreisfrequenz ω ab. Zur Bestimmung ausbreitungsfähiger Moden wird nun für die magnetische Feldstärkekomponente H der Separationsansatz

$$H(x, y, z) = Au(x, y)e^{-ikz} \tag{5}$$

gemacht. Dabei bestimmt der Realteil der Wellenzahl k die Phasengeschwindigkeit, der Imaginärteil von k gibt Aufschluß über die Dämpfung bzw. Verstärkung in Ausbreitungsrichtung. Setzt man diesen Separationsansatz in Gleichung (4) ein, so ergibt sich das Eigenwertproblem der skalaren Helmholtzgleichung

$$-\Delta u - \omega^2 \tilde{\epsilon} \mu u = -k^2 u \quad . \tag{6}$$

Da die Feldstärkekomponenten in den interessierenden Fällen eine starke räumliche Konzentration aufweisen, wird die Gleichung (6) nur auf einer beschränkten zweidimensionalen Teilmenge Ω betrachtet. Um ein mathematisch wohlgestelltes Problem zu erhalten, muß (6) noch um eine passende Randbedingung ergänzt werden. Prinzipiell sollte diese Bedingung so gewählt werden, daß das Verhalten der Lösung nicht wesentlich von der Vorgabe des Gebietes Ω abhängt. Das kann beispielsweise durch Verwendung transparenter Randbedingungen [52] erreicht werden. Im folgenden wollen wir jedoch von einem bereits dem Problem angepaßten Gebiet Ω ausgehen. Als Randbedingung wird die Dirichlet-Bedingung $u = 0$ auf dem Rand $\partial\Omega$ angesetzt.

Zusammengefaßt erhalten wir die folgende Problemstellung, die den Mittelpunkt dieser Arbeit bildet: gesucht sind Funktionen $u \neq 0$ und Zahlen λ, die Lösungen des Eigenwertproblems

$$\begin{aligned} -\Delta u(x, y) - f(x, y)u(x, y) &= \lambda u(x, y), \qquad (x, y) \in \Omega \\ u(x, y) &= 0, \qquad\qquad\quad (x, y) \in \partial\Omega \end{aligned} \tag{7}$$

sind, wobei zur Abkürzung $f(x, y) = \omega^2\tilde{\epsilon}(x, y, \omega)\mu$ und $\lambda = -k^2$ gesetzt wurde. Da in Medien mit elektrischer Leitfähigkeit die Dielektrizitätszahl $\tilde{\epsilon}$ und damit die Funktion f komplexe Werte annimmt, sind im allgemeinen die Eigenfunktionen u und die Eigenwerte λ ebenfalls komplexwertig.

2.2 Variationsformulierung des Eigenwertproblems

Das Eigenwertproblem in der Formulierung (7) ist im klassischen Sinn, also in der punktweisen Übereinstimmung der linken und rechten Seiten auf dem Gebiet Ω beziehungsweise dessen Rand zu verstehen. Dabei soll Ω eine offene, beschränkte und zusammenhängende Teilmenge des $\mathbb{R}^2$ mit polygonalem oder glattem Rand $\partial\Omega$ sein. Da der Laplace-Operator $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ auf die Funktion u angewandt wird, muß zum einen die Existenz zweiter Ableitungen von u in Ω gefordert werden, zum anderen verlangt die Vorgabe von Randwerten die Stetigkeit von u auf $\overline{\Omega} = \Omega \cup \partial\Omega$. Diese Bedingungen an die Eigenfunktion erweisen sich in vielen Fällen, wie zum Beispiel bei Funktionen f mit Sprungstellen in Ω, als zu restriktiv. Eine Erweiterung des Lösungsbegriffes wird durch den Übergang zur Variationsformulierung in einem geeigneten Hilbertraum, dem Sobolevraum $H_0^1(\Omega)$, erreicht.

Bemerkung 1. Im Zusammenhang mit der im vorigen Abschnitt besprochenen Problemklasse erweist sich dies als natürliche Erweiterung, da die so festgesetzten Differenzierbarkeitseigenschaften der Funktion u und damit der Feldkomponente H genau denen der Maxwellschen Gleichungen (1) entsprechen.

Zur Einführung des Sobolevraumes orientieren wir uns an der Darstellung in Braess [7, S. 28 ff.] und Hackbusch [28, S. 109 ff.]. Die Grundlage bildet der Hilbertraum $L^2(\Omega)$ der über Ω quadratisch Lebesgue-integrierbaren Funktionen mit dem zugehörigen Skalarprodukt

$$(v, u) = (v, u)_{L^2(\Omega)} = \int_\Omega \overline{v(x, y)} u(x, y) \, \mathrm{d}(x, y)$$

und der induzierten Norm

$$\|u\| = \|u\|_{L^2(\Omega)} = \sqrt{(u, u)} \quad .$$

Eine Funktion $u \in L^2(\Omega)$ heißt einmal schwach differenzierbar, falls Elemente

$$\partial_x u \in L^2(\Omega) \quad \text{und} \quad \partial_y u \in L^2(\Omega)$$

existieren, so daß die Gleichungen

$$(v, \partial_x u) = -(\partial_x v, u) \quad \text{und} \quad (v, \partial_y u) = -(\partial_y v, u) \tag{8}$$

für alle Funktionen $v \in C_0^\infty(\Omega)$ erfüllt sind. Dabei bezeichnet $C_0^\infty(\Omega)$ den Raum der beliebig oft differenzierbaren Funktionen, die nur in einer kompakten Teilmenge von Ω von Null verschiedene Werte annehmen.

Beispiel 3. Die Funktion $u(x,y) = |x| + |y|$ besitzt auf $\Omega = (-1,1)^2$ die schwachen Ableitungen

$$\partial_x u(x,y) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \quad \text{und} \quad \partial_y u(x,y) = \begin{cases} 1, & y > 0 \\ 0, & y = 0 \\ -1, & y < 0 \end{cases} .$$

Da die schwachen Ableitungen über die Integralrelation (8) definiert sind, können wir $\partial_x u(x,y)$ bzw. $\partial_y u(x,y)$ für $x = 0$ bzw. $y = 0$ jeden beliebigen Wert zuordnen. ◇

Bemerkung 2. Ist eine Funktion im klassischen Sinne einmal auf ganz $\overline{\Omega}$ differenzierbar, so existieren auch die schwachen Ableitungen und stimmen mit den klassischen überein. Die Gleichungen (8) beinhalten dann die Greenschen Formeln für die partielle Integration.

Die Menge aller einmal schwach differenzierbarer Funktionen $u \in L^2(\Omega)$ bildet den Sobolevraum $H^1(\Omega)$. Zusammen mit dem Skalarprodukt

$$\begin{aligned} (v,u)_{H^1(\Omega)} &= (v,u) + (\partial_x v, \partial_x u) + (\partial_y v, \partial_y u) \\ &= (v,u) + (\nabla v, \nabla u) \end{aligned}$$

und der daraus induzierten Norm

$$\|u\|_{H^1(\Omega)} = \sqrt{(u,u)_{H^1(\Omega)}}$$

wird $H^1(\Omega)$ zu einem Hilbertraum. Weiterhin liegt der Raum der beliebig oft differenzierbaren Funktionen $C^\infty(\Omega) \subset H^1(\Omega)$ dicht in $H^1(\Omega)$, das heißt $H^1(\Omega)$ ist die Vervollständigung von $C^\infty(\Omega)$ bezüglich der Sobolevnorm $\|u\|_{H^1(\Omega)}$. Diese Tatsache legt die entsprechende Verallgemeinerung für Funktionen mit Nullrandbedingungen fest: der Sobolevraum $H_0^1(\Omega)$ wird als die Vervollständigung von $C_0^\infty(\Omega)$ bezüglich $\|u\|_{H^1(\Omega)}$ definiert.

Beispiel 4. Die in Abbildung 2 dargestellte Funktion

$$u(x,y) = \begin{cases} 1 - |x| - |y|, & |x| + |y| < 1 \\ 0, & \text{sonst} \end{cases}$$

auf $\Omega = (-1,1)^2$ ist Element des Hilbertraumes $H_0^1(\Omega)$. ◇

Bemerkung 3. Eine einmal im klassischen Sinne auf ganz $\overline{\Omega}$ differenzierbare Funktion gehört genau dann zu $H_0^1(\Omega)$, wenn $u = 0$ auf $\partial\Omega$.

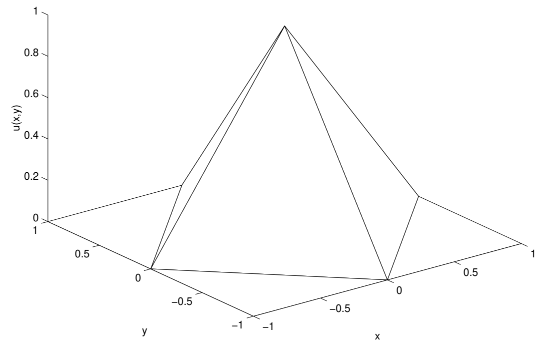

Abbildung 2: Eine Funktion aus $H_0^1(\Omega)$

In Analogie zum eben beschriebenen Vorgehen können auch schwache Ableitungen höherer Ordnung $m > 1$ eingeführt werden. Die zugehörigen Sobolevräume werden dann mit $H^m(\Omega)$ beziehungsweise $H_0^m(\Omega)$ bezeichnet.
Nach diesen vorbereitenden Betrachtungen soll jetzt die Variationsformulierung für das Eigenwertproblem hergeleitet werden, wobei wir den Ausführungen in HACKBUSCH [28, S. 132 ff., S. 227] folgen. Es wird angenommen, daß mit dem Eigenpaar (u, λ) eine klassische Lösung von (7) gegeben ist. Für beliebiges $w \in C_0^\infty(\Omega)$ wird nun das Skalarprodukt

$$(w, -\Delta u - fu) = -(w, \Delta u) - (w, fu)$$

gebildet. Durch partielle Integration kann der erste Term auf der rechten Seite gemäß

$$\begin{aligned} -(w, \Delta u) &= -(w, \partial_x \partial_x u) - (w, \partial_y \partial_y u) \\ &= (\partial_x w, \partial_x u) + (\partial_y w, \partial_y u) \\ &= (\nabla w, \nabla u) \end{aligned}$$

umgeformt werden, wobei wegen $w \in C_0^\infty(\Omega)$ keine Randterme auftreten. Da (u, λ) Lösung des Eigenwertproblems ist, erhalten wir so die Variationsformulierung

$$(\nabla w, \nabla u) - (w, fu) = \lambda(w, u) \quad \forall w \in C_0^\infty(\Omega) \quad .$$

Ist umgekehrt durch (u, λ) mit zweimal im klassischen Sinne auf $\overline{\Omega}$ differenzierbarem u und Randbedingung $u = 0$ auf $\partial\Omega$ eine Lösung des Variationsproblems gegeben, kann die partielle Integration rückgängig gemacht werden. Aus

$$-(w, \Delta u) - (w, fu) = \lambda(w, u) \quad \forall w \in C_0^\infty(\Omega)$$

folgt dann, daß (u, λ) auch Lösung von (7) ist, das heißt bezüglich klassischer Lösungen sind die ursprüngliche Formulierung und die Variationsformulierung äquivalent. Wie oben bemerkt, ist die Randbedingung für klassisch differenzierbare Funktionen gleichbedeutend zu $u \in H_0^1(\Omega)$. Führen wir noch die Sesquilinearform

$$a(w, u) = (\nabla w, \nabla u) - (w, fu) \tag{9}$$

ein, lautet die Variationsformulierung des Eigenwertproblems also wie folgt: gesucht sind $u \in H_0^1(\Omega) \setminus \{0\}$ und $\lambda \in \mathbb{C}$, so daß

$$a(w, u) = \lambda(w, u) \quad \forall w \in H_0^1(\Omega) \tag{10}$$

gilt. Dabei wurde noch benutzt, daß die Form $a(\cdot, \cdot)$ auf $H_0^1(\Omega) \times H_0^1(\Omega)$ fortgesetzt werden kann und die Variationsformulierungen wegen der Dichtheit von $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$ in beiden Räumen äquivalent sind. Die Lösungen des Variationsproblems nennt man schwache Lösungen, u heißt (rechte) Eigenfunktion und λ Eigenwert.

Beispiel 5. Wir betrachten das Eigenwertproblem in einer Raumdimension mit

$$f(x) = \begin{cases} 1, & x \in [-1, 1] \\ 0, & \text{sonst} \end{cases}$$

auf dem Intervall $\Omega = (-2, 2)$, welches strukturell den in der integrierten Optik auftretenden Problemen entspricht. Durch die Funktion

$$u(x) = \begin{cases} A \sinh(\alpha(x + 2)), & x \in (-2, -1) \\ B \cos(\beta x), & x \in [-1, 1] \\ -A \sinh(\alpha(x - 2)), & x \in (1, 2) \end{cases}$$

mit $\alpha \doteq 0.4705$, $\beta \doteq 0.8824$, $A \doteq 1.3018$ und $B = 1$ ist eine Lösung des Variationsproblems zum Eigenwert $\lambda \doteq -0.2213$ gegeben, die jedoch nicht zweimal im klassischen Sinne auf ganz Ω differenzierbar ist (siehe auch Abbildung 3). ◇

Dem Eigenwertproblem in Variationsformulierung (10) wird, wie auch bei Eigenwertaufgaben für Matrizen [24, S. 333] üblich, ein adjungiertes Problem zur Seite gestellt: gesucht ist eine Funktion $v \in H_0^1(\Omega) \setminus \{0\}$ zu $\lambda \in \mathbb{C}$, so daß

$$a(v, w) = \lambda(v, w) \quad \forall w \in H_0^1(\Omega) \tag{11}$$

gilt. Die Funktion v wird als linke Eigenfunktion bezeichnet. Unter Beachtung der Eigenschaften des Skalarproduktes in $L^2(\Omega)$ kann das adjungierte Problem äquivalent wie folgt formuliert werden: gesucht ist $v \in H_0^1(\Omega) \setminus \{0\}$ zu $\lambda \in \mathbb{C}$ mit

$$a^*(w, v) = \overline{\lambda}(w, v) \quad \forall w \in H_0^1(\Omega) \quad ,$$

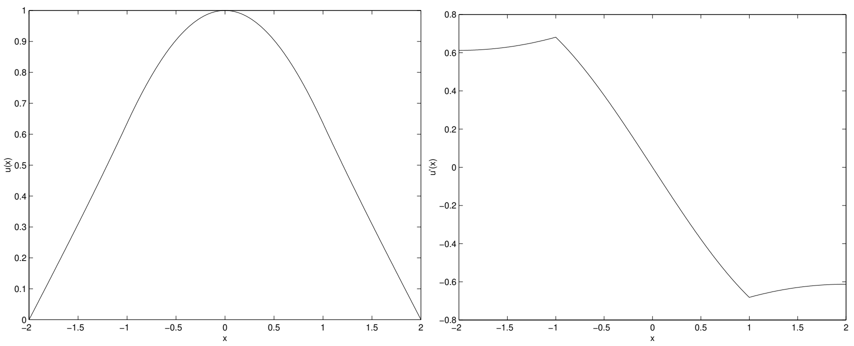

Abbildung 3: Schwache Eigenlösung zum Sprungproblem und erste Ableitung der Lösung

wobei durch

$$a^*(w,v) = \overline{a(v,w)} = (\nabla w, \nabla v) - (w, \overline{f} v)$$

die adjungierte Sesquilinearform definiert ist. Im folgenden Lemma sind einige grundlegende Eigenschaften der Eigenfunktionen und Eigenwerte zusammengestellt.

Lemma 1.

(i) Die linken Eigenfunktionen sind die konjugiert komplexen rechten Eigenfunktionen, also $v = \overline{u}$.

(ii) Linke und rechte Eigenfunktionen zu verschiedenen Eigenwerten stehen senkrecht aufeinander, das heißt $(v_\mu, u_\lambda) = 0$, falls $\lambda \neq \mu$.

(iii) Die Eigenwerte liegen in dem Halbstreifen

$$\Sigma = \{z \in \mathbb{C} \,|\, \Re z > -\sup_\Omega \Re f, -\inf_\Omega \Im f \geq \Im z \geq -\sup_\Omega \Im f\}$$

der komplexen Ebene.

Beweis.

(i) Es sei u rechte Eigenfunktion zum Eigenwert λ. Dann gilt

$$\begin{aligned} a(w,u) &= \lambda(w,u) \quad \forall w \in H_0^1(\Omega) \qquad &\Longleftrightarrow \\ a(\overline{u},\overline{w}) &= \lambda(\overline{u},\overline{w}) \quad \forall w \in H_0^1(\Omega) \qquad &\Longleftrightarrow \\ a(\overline{u},w) &= \lambda(\overline{u},w) \quad \forall w \in H_0^1(\Omega) \quad , & \end{aligned}$$

das heißt $\overline{u}$ ist linke Eigenfunktion.

(ii) Aus (10) bzw. (11) ergibt sich für die rechte bzw. linke Eigenfunktion speziell

$$a(v_\mu, u_\lambda) = \lambda(v_\mu, u_\lambda) \quad \text{bzw.} \quad a(v_\mu, u_\lambda) = \mu(v_\mu, u_\lambda) \quad .$$

Subtraktion der beiden Gleichungen liefert

$$0 = (\lambda - \mu)(v_\mu, u_\lambda) \quad ,$$

also für $\lambda \neq \mu$ die behauptete Orthogonalität.

(iii) Es sei zunächst w eine beliebige Funktion aus $H_0^1(\Omega) \setminus \{0\}$. Wegen $w \neq 0$ können wir den Rayleigh-Quotienten

$$R(w) = \frac{a(w,w)}{(w,w)}$$

bilden und erhalten für den Real- bzw. Imaginärteil von $R(w)$ die in [51] angegebenen Gleichungen

$$\Re R(w) = \frac{(\nabla w, \nabla w)}{(w,w)} - \frac{(w, \Re f\, w)}{(w,w)} \quad \text{bzw.} \quad \Im R(w) = -\frac{(w, \Im f\, w)}{(w,w)} \quad .$$

Eine einfache Rechnung unter Benutzung der Poincaré-Friedrichs-Ungleichung (15) zeigt nun

$$\Re R(w) \geq \frac{1}{s^2} - \sup_\Omega \Re f > -\sup_\Omega \Re f \quad \text{und} \quad -\inf_\Omega \Im f \geq \Im R(w) \geq -\sup_\Omega \Im f \quad ,$$

also $R(w) \in \Sigma$. Aus (10) ergibt sich schließlich für die rechte Eigenfunktion

$$a(u,u) = \lambda(u,u) \quad ,$$

das heißt $\lambda = R(u)$ und damit $\lambda \in \Sigma$. □

Zum Schluß dieses Abschnitts wollen wir noch auf einen wichtigen Spezialfall des Eigenwertproblems eingehen. Das Eigenwertproblem heißt selbstadjungiert, wenn

$$a^*(w,v) = a(w,v) \quad \forall v, w \in H_0^1(\Omega)$$

gilt. Dies ist offenbar genau dann der Fall, wenn $f = \overline{f}$ gilt, das heißt, wenn f auf Ω nur reelle Werte annimmt. Die zu Lemma 1 korrespondierenden Eigenschaften für das selbstadjungierte Problem beschreibt

Lemma 2.

(i) Alle Eigenwerte sind reell.

(ii) Die Eigenfunktionen können reell gewählt werden.

(iii) Linke und rechte Eigenfunktionen stimmen überein.

(iv) Eigenfunktionen zu verschiedenen Eigenwerten sind orthogonal zueinander, das heißt $(u_\mu, u_\lambda) = 0$*, falls* $\lambda \neq \mu$*.*

Beweis.

(i) Da $\Im f = 0$ gilt, folgt dies aus Lemma 1 (iii).

(ii) Da f und λ reell sind, erhalten wir aus (10)

$$a(w, \Re u) = \lambda(w, \Re u) \quad \text{und} \quad a(w, \Im u) = \lambda(w, \Im u) \quad \forall w \in H_0^1(\Omega) \quad .$$

Wegen $u \neq 0$ ist nun $\Re u \neq 0$ oder $\Im u \neq 0$, das heißt zu λ existiert mindestens eine reelle Eigenfunktion.

(iii) Da wir u als reelle Funktion voraussetzen können, folgt dies aus Lemma 1 (i).

(iv) Das ergibt sich aus Lemma 1 (ii) unter Verwendung von $v_\mu = u_\mu$. □

In den vorangegangenen Untersuchungen haben wir stets die Existenz von Eigenlösungen vorausgesetzt. Im Falle selbstadjungierter Eigenwertprobleme lassen sich diesbezügliche Aussagen direkt aus der Variationsformulierung ableiten. Das folgende Theorem gibt neben der Existenzaussage zusätzliche wichtige qualitative Eigenschaften der Eigenwerte und -funktionen.

Theorem 1. *Das Eigenwertproblem sei selbstadjungiert. Dann gibt es eine unendliche Folge von Eigenwerten und Eigenfunktionen* λ_j *und* u_j*. Die Eigenwerte sind dabei durch das Courantsche Minimaxprinzip*

$$\lambda_j = \min_{\substack{W_j \subset H_0^1(\Omega) \\ \dim W_j = j}} \max_{u \in W_j \setminus \{0\}} R(u) = R(u_j), \quad j = 1, 2, \dots , \tag{12}$$

mit dem Rayleigh-Quotienten $R(u) = a(u,u)/(u,u)$ *charakterisiert und ergeben eine gegen unendlich strebende Folge* $\lambda_1 \leq \lambda_2 \leq \ldots \to \infty$*. Die zugehörigen Eigenfunktionen* u_j *bilden eine Orthonormalbasis in* $L^2(\Omega)$*, das heißt mit der Normierung* $(u_j, u_j) = 1$ *und den Fourierkoeffizienten* $\alpha_j = (u_j, u)$ *gilt für jedes* $u \in L^2(\Omega)$ *die Relation*

$$u = \sum_{j=1}^{\infty} \alpha_j u_j \quad .$$

Beweis. Die Resultate sind dem Buch von Courant und Hilbert [11, S. 486 ff.] und dem Artikel von Babuška und Osborn [3, S. 672] entnommen. □

Im folgenden Abschnitt wird es darum gehen, eine entsprechende Aussage für das allgemeine Eigenwertproblem herzuleiten.

2.3 Operatordarstellung und Spektralzerlegung

Wir beginnen mit der Konstruktion des natürlichen Darstellungsoperators der Sesquilinearform $a(\cdot,\cdot)$ in $L^2(\Omega)$. Dabei handelt es sich um einen linearen und in $L^2(\Omega)$ dicht definierten Operator $A : D_A \to L^2(\Omega)$ mit Definitionsgebiet $D_A \subset H_0^1(\Omega)$, für den

$$a(w,v) = (w, Av) \quad \forall v \in D_A, w \in H_0^1(\Omega)$$

gilt. Mit diesem Operator A können wir dann das folgende Eigenwertproblem in $L^2(\Omega)$ formulieren: gesucht sind $v \in D_A \setminus \{0\}$ und $\lambda \in \mathbb{C}$, so daß

$$Av = \lambda v \quad . \tag{13}$$

Der Zusammenhang zwischen diesem und dem Problem in Variationsformulierung ist der folgende: wegen $D_A \subset H_0^1(\Omega)$ ist zunächst jede Lösung des Eigenwertproblems (13) auch Lösung des Variationsproblems (10). Ist umgekehrt durch $u \in D_A$ eine Lösung des Eigenwertproblems in Variationsformulierung gegeben, folgt aus der Dichtheit von $H_0^1(\Omega)$ in $L^2(\Omega)$, daß dieses u auch Eigenfunktion zum Operatorproblem ist. Die beiden Eigenwertprobleme sind also äquivalent, wenn wir für jede Eigenlösung des Variationsproblems zeigen können, daß diese in D_A liegt. Dies wird sich unter anderem im folgenden ergeben.
Betrachten wir zunächst den ersten Summanden von $a(\cdot,\cdot)$ aus Gleichung (9). Wir wollen einen linearen und dicht definierten Operator $L : D_L \to L^2(\Omega)$ erzeugen, so daß

$$(\nabla w, \nabla v) = (w, Lv) \quad \forall v \in D_L, w \in H_0^1(\Omega) \tag{14}$$

erfüllt ist. Zusätzlich fordern wir noch, daß L selbstadjungiert ist. Dabei heißt ein dicht definierter Operator L selbstadjungiert, wenn der Definitionsbereich

$$D_{L^*} = \{w \in L^2(\Omega) \,|\, \exists L^* w \in L^2(\Omega) : (w, Lv) = (L^* w, v) \quad \forall v \in D_L\}$$

des adjungierten Operators $L^* : D_{L^*} \to L^2(\Omega)$ gleich D_L ist und $Lv = L^* v$ für alle $v \in D_L$ gilt [30, S. 565]. Zunächst ist die Form $(\nabla w, \nabla v)$ auf ganz $H_0^1(\Omega) \times H_0^1(\Omega)$ definiert, wegen der Poincaré-Friedrichs-Ungleichung

$$\frac{1}{s^2}\|v\|^2 \leq (\nabla v, \nabla v) \quad \forall v \in H_0^1(\Omega) \tag{15}$$

positiv und aufgrund der Vollständigkeit von $H_0^1(\Omega)$ bezüglich $\|\cdot\|_{H^1(\Omega)}$ abgeschlossen [45, S. 276 f.]. Die Größe s bezeichnet dabei die Seitenlänge eines Quadrats, in dem Ω vollständig enthalten ist. Aus $(\nabla w, \nabla v) = \overline{(\nabla v, \nabla w)}$ folgt weiterhin, daß die Sesquilinearform $(\nabla w, \nabla v)$ hermitesch ist. Nach dem Darstellungssatz von Friedrichs [34, S. 322 f.] existiert nun genau ein linearer und selbstadjungierter Operator $L : D_L \to L^2(\Omega)$ mit Definitionsgebiet $D_L \subset H_0^1(\Omega)$, so daß die geforderte Gleichheit (14) gilt. Dabei liegt D_L dicht in $H_0^1(\Omega)$ bezüglich $\|\cdot\|_{H^1(\Omega)}$ und somit dicht in $L^2(\Omega)$ bezüglich $\|\cdot\|$. Aus der Poincaré-Friedrichs-Ungleichung erhalten wir für L die Relation

$$\frac{1}{s^2}\|v\|^2 \leq (v, Lv) \quad \forall v \in D_L \quad ,$$

das heißt L ist auch ein positiver Operator. Unter Benutzung der Cauchy-Schwarz-Ungleichung folgt daraus

$$\frac{1}{s^2}\|v\| \leq \|Lv\| \quad \forall v \in D_L \tag{16}$$

und damit die Existenz und die Stetigkeit der Inversen L^{-1} auf dem Bild $R_L \subset L^2(\Omega)$ von L. Wir wollen jetzt noch zeigen, daß $R_L = L^2(\Omega)$ gilt und damit die Inverse L^{-1} auf ganz $L^2(\Omega)$ definiert und stetig ist. Dazu sei zunächst bemerkt, daß selbstadjungierte Operatoren stets abgeschlossen sind [30, S. 565]. Daraus und aus der Stetigkeit der Inversen L^{-1} auf R_L ergibt sich die Abgeschlossenheit von R_L und somit die Existenz der Orthogonalzerlegung $L^2(\Omega) = R_L \oplus R_L^{\perp}$ [30, S. 170]. Nehmen wir nun an, es wäre $L^2(\Omega) \neq R_L$, so gäbe es ein $w \perp R_L$ mit $w \neq 0$. Für dieses Element gilt dann

$$(w, Lv) = 0 = (0, v) \quad \forall v \in D_L$$

und damit $w \in D_{L^*} = D_L$ mit $L^*w = Lw = 0$. Aus (16) folgt dann schließlich $w = 0$, obwohl $w \neq 0$ vorausgesetzt wurde. Dieser Widerspruch zeigt die Unzulässigkeit der oben gemachten Annahme und wir erhalten $L^2(\Omega) = R_L$. Der Operator L ist die Fortsetzung des Laplace-Operators $-\Delta$ auf D_L.
Wenden wir uns nun dem zweiten Summanden in $a(\cdot,\cdot)$ zu. Dabei wollen wir voraussetzen, daß die Funktion f in $L^\infty(\Omega)$ liegt, also über Ω beschränkt ist. Die Sesquilinearform (w, fv) ist so auf ganz $L^2(\Omega) \times L^2(\Omega)$ definiert und stetig bezüglich der $L^2(\Omega)$-Norm. Für festes $v \in L^2(\Omega)$ existiert daher nach dem Darstellungssatz von Fréchet-Riesz [30, S. 183] ein eindeutiges $Mv \in L^2(\Omega)$, so daß

$$(w, fv) = (w, Mv) \quad \forall w \in L^2(\Omega)$$

erfüllt ist. Die so gewonnene Abbildung $M : L^2(\Omega) \to L^2(\Omega)$ ist linear und wegen der Stetigkeit von (w, fv) stetig auf $L^2(\Omega)$. Der Operator M wird Multiplikationsoperator genannt.
Der gesuchte Darstellungsoperator $A : D_A \to L^2(\Omega)$ zu $a(\cdot,\cdot)$ ergibt sich nun einfach aus $A = L - M$ mit Definitionsgebiet $D_A = D_L$. Der Operator A ist dabei selbstadjungiert, falls der Multiplikationsoperator M selbstadjungiert ist [34, S. 287], das heißt, wenn die Funktion f nur reelle Werte annimmt. Wir wollen jetzt noch den Nachweis erbringen, daß jede Eigenlösung u des Variationsproblems (10) in D_A enthalten ist. Nehmen wir also an, wir haben eine Eigenfunktion $u \in H_0^1(\Omega)$ zum Eigenwert λ des Problems in Variationsformulierung gefunden. Dann gilt auch

$$(\nabla w, \nabla u) = (w, fu) + (w, \lambda u) = (w, g) \quad \forall w \in H_0^1(\Omega) \quad , \tag{17}$$

wobei wir $g = (f + \lambda)u$ gesetzt haben. Die Funktion g liegt wegen $f + \lambda \in L^\infty(\Omega)$ und $u \in H_0^1(\Omega) \subset L^2(\Omega)$ in $L^2(\Omega)$. Die Theorie linearer Probleme in Variationsformulierung [28, S. 132 ff.] lehrt nun, daß für jedes $g \in L^2(\Omega)$ genau eine Lösung $u \in H_0^1(\Omega)$ von (17) existiert. Andererseits haben wir oben gezeigt, daß die Inverse des Operators L auf ganz $L^2(\Omega)$ definiert ist und erhalten so das Element $L^{-1}g = v \in D_L \subset H_0^1(\Omega)$. Für diese Funktion v gilt nun insbesondere die Relation (14) und damit

$$(\nabla w, \nabla v) = (w, Lv) = (w, g) \quad \forall w \in H_0^1(\Omega) \quad .$$

Aus der Eindeutigkeit der Lösung ergibt sich daraus schließlich $u = v \in D_L = D_A$. Somit sind die beiden Eigenwertprobleme (10) und (13) tatsächlich äquivalent. Dieses Resultat können wir nun noch speziell auf den Operator L anwenden und erhalten unter Berücksichtigung von Theorem 1, daß die Eigenwerte von L eine gegen unendlich strebende Folge bilden und somit der Operator L unbeschränkt sein muß.

Bemerkung 4. Mit derselben Argumentation erhalten wir auch eine Aussage über die Regularität der Eigenfunktionen. Für konvexes Gebiet Ω beispielsweise folgt aus der Regularitätstheorie für lineare Probleme [28, S. 200 f.], daß die Lösung u der Gleichung (17) für jedes $g \in L^2(\Omega)$ und damit jede Eigenfunktion zu beliebigem $f \in L^\infty(\Omega)$ in $H^2(\Omega) \cap H_0^1(\Omega)$ enthalten ist.

Bevor nun das den weiteren Betrachtungen zugrundeliegende Theorem angegeben wird, sollen noch einige Begriffe erklärt werden (siehe beispielsweise Kato [34, S. 172 ff.]). Als Resolventenmenge $\rho(A)$ bezeichnet man die Menge aller komplexen Zahlen $z \in \mathbb{C}$, für die der Operator $P_z = (A - zI)^{-1}$ auf ganz $L^2(\Omega)$ definiert und stetig ist. Das Spektrum von A ist die Menge $\sigma(A) = \mathbb{C} \setminus \rho(A)$. Ein Element $\lambda \in \sigma(A)$ heißt Eigenwert von A, wenn eine Funktion $v \in D_A \setminus \{0\}$ existiert, so daß

$$(A - \lambda I)v = 0 \tag{18}$$

gilt. Das Element v wird dann Eigenvektor genannt. Der von der Menge aller Eigenvektoren aufgespannte Unterraum ist der zu λ korrespondierende Eigenraum $E(\lambda)$, seine Dimension die geometrische Vielfachheit des Eigenwertes. In Verallgemeinerung von (18) untersucht man für einen Eigenwert λ weiterhin Gleichungsketten der Form

$$v_0 = 0, \quad (A - \lambda I)v_l = v_{l-1}, \quad l = 1, 2, \ldots$$

auf ihre Lösbarkeit in $D_A \setminus \{0\}$. Dabei erhält man zunächst wiederum einen Eigenvektor v_1 zu λ. Die eventuell existierenden Lösungen $v_2, v_3, \ldots$ bezeichnet man als verallgemeinerte Eigenvektoren oder Hauptvektoren und die maximale Länge einer solchen Kette als Index des Eigenwertes λ. Der von der Menge aller zu λ gehörenden Eigen- und Hauptvektoren aufgespannte Unterraum heißt verallgemeinerter Eigenraum oder Hauptraum $N(\lambda)$, seine Dimension die algebraische Vielfachheit von λ. Die beiden zu einem Eigenwert assoziierten Unterräume $E(\lambda)$ und $N(\lambda)$ sind offenbar invariant unter A, das heißt für jedes $v \in E(\lambda)$ bzw. $v \in N(\lambda)$ gilt $Av \in E(\lambda)$ bzw. $Av \in N(\lambda)$. Weiterhin gilt die Inklusion $E(\lambda) \subseteq N(\lambda)$. Damit folgt, daß die geometrische Vielfachheit immer kleiner oder gleich der algebraischen Vielfachheit ist. Außerdem sind beide Größen für einen Eigenwert $\lambda \in \sigma(A)$ stets größer oder gleich 1.

Beispiel 6. Wir betrachten die lineare und stetige Abbildung $S : L^2(\Omega) \to L^2(\Omega)$ mit $w = \sum_{j=1}^{\infty} \alpha_j w_j \mapsto Sw = \sum_{j=1}^{\infty} \alpha_{j+1} w_j$, wobei $(w_j)_{j=1}^{\infty}$ eine Orthonormalbasis von $L^2(\Omega)$ ist. Das Spektrum von S bestimmt sich zu $\sigma(S) = \{\lambda \in \mathbb{C} : |\lambda| \leq 1\}$. Jedes Element $\lambda \in \sigma(S)$ mit $|\lambda| < 1$ ist Eigenwert mit der zugehörigen normierten Eigenfunktion $v(\lambda) = \sqrt{1-|\lambda|^2} \sum_{j=1}^{\infty} \lambda^{j-1} w_j$. Betrachten wir speziell den Eigenwert $\lambda = 0$, so erhalten wir $E(0) = \mathrm{span}(w_1)$ und $N(0) = \mathrm{span}(w_j)_{j=1}^{\infty}$, das heißt die geometrische Vielfachheit dieses Eigenwertes ist 1 und seine algebraische Vielfachheit gleich ∞. $\diamond$

Das Spektrum eines Operators A wird diskret genannt, wenn jedes Element $\lambda \in \sigma(A)$ ein Eigenwert mit endlicher algebraischer Vielfachheit ist und sich die Eigenwerte nur im Unendlichen häufen können. Kommen wir nun zum angekündigten

Theorem 2. *Es sei T ein abgeschlossener Operator in einem Hilbertraum H, L ein positiver, selbstadjungierter Operator in H mit diskretem Spektrum und Definitionsgebiet $D_L \subset\subset D_T$ sowie $A = L + T$. Für ein p mit $0 < p \leq 1$ seien ferner die folgenden Bedingungen erfüllt:*

1. *Der Operator $L^{\frac{p-1}{2}} T L^{\frac{p-1}{2}}$ ist auf D_L definiert und beschränkt.*

2. *Für die Eigenwerte $\mu_1 \leq \mu_2 \leq \ldots \to \infty$ des Operators L gilt die Relation*

$$\overline{\lim}_{j\to\infty} j \mu_j^{-p} < \infty \quad .$$

Dann ist das Spektrum des Operators A diskret, und die Menge $\Lambda = \{\lambda_j\}_{j=1}^{\infty}$ der Eigenwerte des Operators A kann in endliche Teilmengen Λ_k unterteilt werden, so daß das System $(V_k)_{k=1}^{\infty}$ der korrespondierenden spektralen Unterräume des Operators A eine Basis in H bildet. Diese ist äquivalent zu einer Orthogonalbasis von H.

Der *Beweis* dieses Resultats ist in der Arbeit [35] von Katsnelson zu finden. $\square$

Bemerkung 5. Die Teilmengen Λ_k sind durch $\Lambda_k = \{\lambda \in \sigma(A) : t_{k-1} < \Re\lambda \leq t_k\}$ mit $-\infty = t_0 < t_1 < t_2 < \ldots \to \infty$ gegeben.

Wir wollen zunächst zeigen, daß die Voraussetzungen von Theorem 2 in unserem Fall erfüllt sind. Dazu setzen wir $p = 1$, $T = -M$ und $D_T = H = L^2(\Omega)$. Der Operator T ist als stetiger Operator mit vollständigem Definitionsgebiet $L^2(\Omega)$ auch abgeschlossen [30, S. 244] und natürlich auf $D_L \subset L^2(\Omega)$ definiert und beschränkt. Vom Operator

L wissen wir bereits, daß er positiv und selbstadjungiert ist und eine auf ganz $L^2(\Omega)$ definierte und stetige Inverse besitzt. Wir wollen jetzt nachweisen, daß das Spektrum von L diskret ist. Aus der Poincaré-Friedrichs-Ungleichung folgt die Abschätzung

$$\frac{1}{s^2+1}\|v\|^2_{H^1(\Omega)} \leq (\nabla v, \nabla v) \quad \forall v \in H^1_0(\Omega)$$

und damit speziell für $v \in D_L$ die Ungleichung

$$\frac{1}{s^2+1}\|v\|^2_{H^1(\Omega)} \leq (v, Lv) \quad .$$

Unter Verwendung der Cauchy-Schwarz-Ungleichung und unter Beachtung der Relation $\|v\| \leq \|v\|_{H^1(\Omega)}$ für alle $v \in H^1(\Omega)$ erhalten wir daraus

$$\frac{1}{s^2+1}\|v\|_{H^1(\Omega)} \leq \|Lv\| \quad \forall v \in D_L \quad .$$

Setzen wir noch $v = L^{-1}g$ mit $g \in L^2(\Omega)$, so ergibt sich

$$\|L^{-1}g\|_{H^1(\Omega)} \leq (s^2+1)\,\|g\| \quad \forall g \in L^2(\Omega) \quad ,$$

das heißt die Inverse ist auch stetig bezüglich der $H^1(\Omega)$-Norm. Verwenden wir nun den Auswahlsatz von Rellich [7, S. 32], so sehen wir wegen $D_L \subset H^1_0(\Omega)$ die kompakte Einbettung $D_L \subset\subset L^2(\Omega)$ und damit, daß die Inverse ein kompakter Operator in $L^2(\Omega)$ ist. Des weiteren ist L^{-1} als die auf ganz $L^2(\Omega)$ definierte Inverse eines selbstadjungierten Operators ebenfalls selbstadjungiert. Damit können wir nun auf den Operator L^{-1} den Spektralsatz für kompakte, selbstadjungierte Operatoren [17, S. 905] anwenden und erhalten unter Zuhilfenahme des Spektralabbildungssatzes für unbeschränkte, abgeschlossene Operatoren [16, S. 602], daß jedes Element aus dem Spektrum $\sigma(L)$ ein Eigenwert ist. Die geometrische und die algebraische Vielfachheit zu jedem Eigenwert stimmen überein und sind endlich. Wie schon erwähnt, bilden die Eigenwerte von L eine gegen unendlich strebende Folge. Sind sie ihrer Größe nach geordnet, wobei jeder Eigenwert so oft vorkommt, wie es seine Vielfachheit angibt, so gilt nach [11, S. 384] für die asymptotische Verteilung im Falle eines zweidimensionalen Gebietes Ω die Relation

$$\lim_{\mu\to\infty}\frac{n(\mu)}{\mu} = \frac{|\Omega|}{4\pi} \quad . \tag{19}$$

Dabei ist $n(\mu)$ die Anzahl der Eigenwerte kleiner oder gleich μ und $|\Omega|$ der Flächeninhalt des Gebietes Ω. Offenbar gilt $n(\mu_j) = j$ und wir können (19) auch in der Form

$$\lim_{j\to\infty}\frac{j}{\mu_j} = \frac{|\Omega|}{4\pi}$$

schreiben. Da Ω beschränkt ist, erhalten wir einen endlichen Grenzwert, und es sind alle Voraussetzungen von Theorem 2 erfüllt.
Auf dieser Grundlage können wir nun eine zum Theorem 1 korrespondierende Aussage für das allgemeine Eigenwertproblem ableiten. Zu diesem Zweck wollen wir die spektralen Unterräume V_k, $k = 1, 2, \ldots$, des Operators A näher untersuchen. Wir wissen aus Theorem 2, daß diese Räume eine Basis in $L^2(\Omega)$ bilden, das heißt zu jedem $v \in L^2(\Omega)$ existieren eindeutig bestimmte Elemente $v_k \in V_k$, so daß

$$v = \sum_{k=1}^{\infty} v_k$$

gilt. Die Äquivalenz zu einer Orthogonalbasis bedeutet dabei [23, S. 334], daß es eine stetige und bijektive lineare Abbildung $B : L^2(\Omega) \to L^2(\Omega)$ und eine Basis von $L^2(\Omega)$ aus paarweise orthogonalen Unterräumen W_k mit der Eigenschaft

$$V_k = BW_k = \{Bw \mid w \in W_k\}$$

gibt. Die Räume V_k sind durch

$$V_k = N(\lambda_{j_k}) \oplus \ldots \oplus N(\lambda_{j_{k+1}-1})$$

mit $1 = j_1 < j_2 < \ldots \to \infty$ gegeben, das heißt als direkte Summe der Haupträume zu den Eigenwerten in

$$\Lambda_k = \{\lambda_{j_k}, \ldots, \lambda_{j_{k+1}-1}\} \quad .$$

Da nun die Teilmengen Λ_k endlich sind und alle Eigenwerte endliche algebraische Vielfachheit haben, sind die Unterräume V_k von endlicher Dimension. Wie wir oben gesehen haben, bilden die Haupträume und damit auch die Räume V_k invariante Unterräume von A. Betrachten wir nun die Einschränkung A_k des Operators A auf V_k, so erhalten wir eine lineare Abbildung eines endlichdimensionalen Raumes in sich selbst, die durch eine $(d_k \times d_k)$-Matrix mit $d_k = \dim(V_k)$ beschrieben werden kann. Unter Verwendung der Schur-Zerlegung [24, S. 335] für beliebige Matrizen können wir jetzt eine Orthonormalbasis $(u_{k,l})_{l=1}^{d_k}$ des Unterraumes V_k finden, bezüglich der der Operator A_k obere Dreiecksgestalt annimmt, das heißt die Relation

$$Au_{k,l} = \lambda_{k,l}\, u_{k,l} + \sum_{m=1}^{l-1} \tau_{k,ml}\, u_{k,m}, \quad l = 1, \ldots, d_k, \tag{20}$$

erfüllt ist. Die Häufigkeit des Auftretens eines Eigenwertes in der Folge $(\lambda_{k,l})_{l=1}^{d_k}$ nennen wir ab jetzt kurz Vielfachheit dieses Eigenwertes. Wichtig für das Weitere ist nun das folgende

Lemma 3. *Das System von Unterräumen* $(V_k)_{k=1}^{\infty}$ *sei eine Basis des Hilbertraumes* H *und äquivalent zu einer Orthogonalbasis. Dann ist jede Folge* $(u_j)_{j=1}^{\infty}$*, die aus der Vereinigung von Orthonormalbasen der Unterräume* V_k*,* $k = 1, 2, \ldots,$ *entsteht, eine Basis des Hilbertraumes* H*, die äquivalent zu einer Orthonormalbasis ist.*

Beweis. Das Resultat ist Gohberg und Kreĭn [23, S. 344] entnommen. □

Wir indizieren nun die Eigenwerte und mit ihnen die korrespondierenden Vektoren $u_{k,l}$ ein zweites Mal um und erhalten

$$
\begin{aligned}
(\lambda_1, \lambda_2, \ldots) &:= (\lambda_{1,1}, \ldots, \lambda_{1,d_1}, \lambda_{2,1}, \ldots, \lambda_{2,d_2}, \ldots) \\
(u_1, u_2, \ldots) &:= (u_{1,1}, \ldots, u_{1,d_1}, u_{2,1}, \ldots, u_{2,d_2}, \ldots) \quad .
\end{aligned}
$$

Wählen wir die Schur-Zerlegung der zu A_k assoziierten Matrix so, daß die Eigenwerte nach der Größe ihrer Realteile sortiert sind, so gilt wegen Bemerkung 5 die Relation $\Re\lambda_1 \leq \Re\lambda_2 \leq \ldots \to \infty$. Für die zugehörigen Vektoren folgt wegen (20) die Gleichung

$$
Au_j = \lambda_j \, u_j + \sum_{k=k_j}^{j-1} \tau_{kj} \, u_k, \quad j = 1, 2, \ldots,
$$

mit $1 \leq k_j \leq j$ und $k_j \leq k_{j+1}$.

Bemerkung 6. Ordnen wir die Eigenwerte λ_j und die Werte τ_{kj} in einem quadratischen Schema an, so ergibt sich analog zur Block-Diagonal-Zerlegung [24, S. 386] einer Matrix beispielsweise

$$
\begin{array}{cccccccccc}
\lambda_1 & \tau_{12} & \tau_{13} & & & & & & & \\
 & \lambda_2 & \tau_{23} & & & & & & & \\
 & & \lambda_3 & & & & & & & \\
 & & & \lambda_4 & \tau_{45} & & & & & \\
 & & & & \lambda_5 & & & & & \\
 & & & & & \lambda_6 & \tau_{67} & \tau_{68} & \tau_{69} & \\
 & & & & & & \lambda_7 & \tau_{78} & \tau_{79} & \\
 & & & & & & & \lambda_8 & \tau_{89} & \\
 & & & & & & & & \lambda_9 & \\
 & & & & & & & & & \ddots
\end{array} \quad .
$$

Die Folge $(k_j)_{j=1}^{\infty}$ gibt an, in welcher Zeile im obigen Schema das erste Element der j-ten Spalte steht. Weiter unten werden wir noch die dazu korrespondierende Folge $(l_j)_{j=1}^{\infty}$ benutzen, die angibt, in welcher Spalte sich das letzte Element der j-ten Zeile befindet.

Da die Basis $(u_j)_{j=1}^{\infty}$ äquivalent zu einer Orthonormalbasis $(w_j)_{j=1}^{\infty}$ ist, existiert analog zu oben eine stetige und bijektive lineare Abbildung $B : L^2(\Omega) \to L^2(\Omega)$, so daß $u_j = Bw_j$ für alle $j \in \mathbb{N}$ gilt. Wir wollen abschließend noch auf die Berechnung der Koeffizienten α_j in der Basisdarstellung

$$u = \sum_{j=1}^{\infty} \alpha_j u_j$$

für ein vorgegebenes $u \in L^2(\Omega)$ eingehen. Wir haben

$$\begin{aligned} B^{-1}u &= \sum_{j=1}^{\infty} (w_j, B^{-1}u)w_j &= \sum_{j=1}^{\infty} (B^{-*}w_j, u)w_j \qquad \Longleftrightarrow \\ u &= \sum_{j=1}^{\infty} (B^{-*}w_j, u)Bw_j &= \sum_{j=1}^{\infty} (u_j^*, u)u_j \end{aligned}$$

und damit $\alpha_j = (u_j^*, u)$ mit $u_j^* = B^{-*}w_j$. Die Folge $(u_j^*)_{j=1}^{\infty}$ ist wegen

$$(u_k^*, u_j) = (B^{-*}w_k, Bw_j) = (w_k, w_j) = \delta_{kj} = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}$$

biorthogonal zu $(u_j)_{j=1}^{\infty}$ und ebenfalls eine Basis von $L^2(\Omega)$ (siehe dazu [23, S. 309 f.]).

Bemerkung 7. Ein Maß für die Abweichung der Basis $(u_j)_{j=1}^{\infty}$ von einer Orthonormalbasis und damit ein Maß für die Abhängigkeit der Koeffizienten α_j von der Funktion u ist durch die Größe

$$\kappa((u_j)_{j=1}^{\infty}) = \|B\|\|B^{-1}\|$$

gegeben. In Verallgemeinerung des Konditionsbegriffes [14, S. 35] nennen wir $\kappa((u_j)_{j=1}^{\infty})$ die Konditionszahl der Basis $(u_j)_{j=1}^{\infty}$.

Aufgrund der Äquivalenz des Variationsproblems (10) und des Operatorproblems (13) können wir die Ergebnisse der vorangegangenen Betrachtungen in dem folgenden Theorem zusammenfassen.

Theorem 3. *Das Gebiet $\Omega \subset \mathbb{R}^2$ sei beschränkt und die Funktion $f : \Omega \to \mathbb{C}$ ein Element des Funktionenraumes $L^{\infty}(\Omega)$. Dann existiert eine Basis $(u_j)_{j=1}^{\infty}$ von $L^2(\Omega)$, so daß für die Basisfunktionen $u_j \in H_0^1(\Omega)$ die Relation*

$$a(w, u_j) = \lambda_j(w, u_j) + \sum_{k=k_j}^{j-1} \tau_{kj}(w, u_k) \quad \forall w \in H_0^1(\Omega) \tag{21}$$

mit $1 \leq k_j \leq j$ und $k_j \leq k_{j+1}$ für alle $j \in \mathbb{N}$ erfüllt ist. Für die Eigenwerte λ_j gilt dabei $\Re\lambda_1 \leq \Re\lambda_2 \leq \ldots \to \infty$.

Bemerkung 8. Für selbstadjungierte Probleme, also für reellwertige Funktionen f, ist nach Theorem 1 die Basis $(u_j)_{j=1}^{\infty}$ sogar eine Orthonormalbasis. Diese besteht ausschließlich aus Eigenfunktionen, das heißt alle τ_{kj} sind gleich Null.

2.4 Sensitivität der Eigenlösungen

In diesem Abschnitt wollen wir den Einfluß von Störungen auf Lösungen der Eigenwertaufgabe (21) untersuchen. Die dabei interessierenden „Eingabegrößen" sind zum einen das Gebiet Ω und zum anderen die Funktion $f \in L^\infty(\Omega)$. Im folgenden wollen wir nur die Abhängigkeit der Lösungen von der Funktion f studieren und uns zunächst auf den ersten Eigenwert konzentrieren. Dabei orientieren wir uns an der Sensitivitätsanalyse für Matrixeigenwertprobleme [24, S. 341 ff.].
Wir nehmen an, der Eigenwert λ_1 sei einfach, das heißt ein Eigenwert mit Vielfachheit 1 und zugehöriger (rechter) Eigenfunktion u_1 und linker Eigenfunktion v_1, so daß

$$a(w, u_1) = \lambda_1(w, u_1) \quad \text{und} \quad a(v_1, w) = \lambda_1(v_1, w) \quad \forall w \in H_0^1(\Omega)$$

gilt. Wir ersetzen nun f durch die Funktion $f + \varepsilon g$ mit reellem ε und $g \in L^\infty(\Omega)$ und betrachten das Eigenwertproblem

$$a_\varepsilon(w, u_1(\varepsilon)) = \lambda_1(\varepsilon)(w, u_1(\varepsilon)) \quad \forall w \in H_0^1(\Omega) \tag{22}$$

mit

$$a_\varepsilon(w, u) = (\nabla w, \nabla u) - (w, (f + \varepsilon g)u) = a(w, u) - \varepsilon(w, gu) \quad .$$

Die Lösungen des gestörten Problems (22) sollen in asymptotische Reihen

$$u_1(\varepsilon) = u_1 + \varepsilon u_1' + \frac{\varepsilon^2}{2} u_1'' + \dots \quad \text{und} \quad \lambda_1(\varepsilon) = \lambda_1 + \varepsilon \lambda_1' + \frac{\varepsilon^2}{2} \lambda_1'' + \dots$$

mit $u_1', u_1'', \ldots \in D_A$ entwickelbar sein. Als Maß für die Abhängigkeit (oder Sensitivität) der Lösung von der Störung in der Funktion f werden nun die Terme u_1' und λ_1' dienen. Setzen wir die asymptotischen Entwicklungen in (22) ein, so erhalten wir für diese Größen durch Koeffizientenvergleich vor dem linearen Term die Bestimmungsgleichung

$$a(w, u_1') - (w, gu_1) = \lambda_1(w, u_1') + \lambda_1'(w, u_1) \quad \forall w \in H_0^1(\Omega) \quad . \tag{23}$$

Verwendet man hier speziell $w = v_1$, so ergibt sich für die Störung im Eigenwert die Relation

$$\lambda_1' = -\frac{(v_1, gu_1)}{(v_1, u_1)} \quad .$$

Durch Verwendung der Cauchy-Schwarz-Ungleichung folgt für den Betrag von λ_1' die Abschätzung

$$|\lambda_1'| \leq \|g\|_{L^\infty(\Omega)} \frac{\|v_1\| \|u_1\|}{|(v_1, u_1)|} \quad .$$

Die Größe

$$\kappa(\lambda_1) = \frac{\|v_1\| \|u_1\|}{|(v_1, u_1)|} = \frac{1}{|\cos \angle[v_1, u_1]|}$$

heißt Konditionszahl des Eigenwertes λ_1. Wir erhalten also, daß der Winkel zwischen linker und rechter Eigenfunktion maßgebend für die Sensitivität des Eigenwertes gegenüber Störungen in der Funktion f ist. Für selbstadjungierte Probleme haben wir oben gesehen, daß linke und rechte Eigenfunktion übereinstimmen, das heißt es gilt $\kappa(\lambda_1) = 1$. Die Berechnung eines einfachen Eigenwertes von selbstadjungierten Problemen ist stets gut konditioniert.
Wenden wir uns nun der Eigenfunktion zu. Wir stellen die Funktionen u_1' und gu_1 bezüglich der in Theorem 3 genannten Basis $(u_j)_{j=1}^{\infty}$ von $L^2(\Omega)$ gemäß

$$u_1' = \sum_{j=1}^{\infty} \alpha_j u_j \quad \text{und} \quad gu_1 = \sum_{j=1}^{\infty} \beta_j u_j$$

dar und gehen von Gleichung (23) zur Operatordarstellung

$$Au_1' - gu_1 = \lambda_1 u_1' + \lambda_1' u_1 \tag{24}$$

über. Unter Berücksichtigung von

$$\begin{aligned} Au_1' = \sum_{j=1}^{\infty} \alpha_j A u_j &= \sum_{j=1}^{\infty} \alpha_j \lambda_j u_j + \sum_{j=1}^{\infty} \sum_{k=k_j}^{j-1} \alpha_j \tau_{kj} u_k \\ &= \sum_{j=1}^{\infty} \alpha_j \lambda_j u_j + \sum_{j=1}^{\infty} \sum_{k=j+1}^{l_j} \alpha_k \tau_{jk} u_j = \sum_{j=1}^{\infty} \left(\alpha_j \lambda_j + \sum_{k=j+1}^{l_j} \alpha_k \tau_{jk} \right) u_j \end{aligned}$$

und unter Benutzung der zu $(u_j)_{j=1}^{\infty}$ biorthogonalen Basis $(u_j^*)_{j=1}^{\infty}$ erhalten wir aus (24) für die Koeffizienten $\alpha_j = (u_j^*, u_1')$ die Beziehung

$$\alpha_j \lambda_j + \sum_{k=j+1}^{l_j} \alpha_k \tau_{jk} - \beta_j = \alpha_j \lambda_1 + \delta_{j1} \lambda_1' \tag{25}$$

für alle $j \in \mathbb{N}$. Der Wert α_1 ist somit beliebig, und für die weiteren Koeffizienten ergibt sich

$$\alpha_j = \frac{1}{\lambda_j - \lambda_1}\left(\beta_j - \sum_{k=j+1}^{l_j} \alpha_k \tau_{jk}\right), \quad j > 1 \quad . \tag{26}$$

Die Sensitivität der Eigenfunktion u_1 gegenüber Störungen wird also im wesentlichen durch den Abstand des korrespondierenden Eigenwertes λ_1 von den übrigen Eigenwerten und durch die Kondition der Basis $(u_j)_{j=1}^{\infty}$ bestimmt.
Im Falle eines mehrfachen Eigenwertes und auch für dicht benachbarte Eigenwerte führt man die beschriebene Analyse nicht für jede Lösung einzeln, sondern wie bei Matrixproblemen [24, S. 347] bezogen auf den zugehörigen invarianten Unterraum $U_m = \operatorname{span}(u_l)_{l=1}^{m}$ durch. Den Ausgangspunkt der Betrachtungen bildet hier das gestörte Problem

$$a_\varepsilon(w, u_l(\varepsilon)) = \sum_{n=1}^{m} \tau_{nl}(\varepsilon)(w, u_n(\varepsilon)) \quad \forall w \in H_0^1(\Omega), \quad l = 1, \ldots, m \quad .$$

Wir nehmen analog zu oben an, daß sich die Lösungen $u_l(\varepsilon)$ und die zugehörigen Werte $\tau_{nl}(\varepsilon)$ in asymptotische Reihen entwickeln lassen. Durch Koeffizientenvergleich erhalten wir das System

$$Au_l' - gu_l = \lambda_l u_l' + \sum_{n=1}^{l-1} \tau_{nl} u_n' + \sum_{n=1}^{m} \tau_{nl}' u_n, \quad l = 1, \ldots, m$$

in Operatorform. Wir stellen nun die Funktionen u_l' und gu_l wieder in der Basis $(u_j)_{j=1}^{\infty}$ dar und erhalten so zu (25) analoge Bestimmungsgleichungen für die Koeffizientenfolge $(\alpha_{jl})_{j=1}^{\infty}$ von u_l', $l = 1, \ldots, m$. Dabei sind die Koeffizienten α_{jl} für $j \leq m$ frei wählbar, und die Werte τ_{jl}' bestimmen sich in Abhängigkeit von dieser Wahl.

Beispiel 7. Wir betrachten den Fall $m = 2$ für die Anordnung wie in Bemerkung 6. Die Werte τ_{jl}' sind hier durch

$$\begin{pmatrix} \tau_{11}' & \tau_{12}' \\ \tau_{21}' & \tau_{22}' \end{pmatrix} = \begin{pmatrix} \lambda_1 & \tau_{12} \\ 0 & \lambda_2 \end{pmatrix} \overbrace{\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}}^{\substack{\text{beliebig}\\ \text{wählbar}}} - \overbrace{\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}}^{\substack{\text{beliebig}\\ \text{wählbar}}} \begin{pmatrix} \lambda_1 & \tau_{12} \\ 0 & \lambda_2 \end{pmatrix} + \begin{pmatrix} \tau_{13} \\ \tau_{23} \end{pmatrix} \underbrace{\begin{pmatrix} \alpha_{31} & \alpha_{32} \end{pmatrix}}_{\substack{\text{berechenbar}\\ \text{wie in (26)}}} - \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}$$

gegeben. ◇

Für die Koeffizienten α_{jl} mit $j > m$ ergeben sich strukturell dieselben Gleichungen wie in (26), wobei im Nenner nun die Eigenwertdifferenzen $\lambda_j - \lambda_l$, $j > m \geq l$, auftreten. Damit erkennen wir, daß die Sensitivität des invarianten Unterraumes U_m von der Separation der Eigenwerte $\lambda_1, \ldots, \lambda_m$ zum Rest des Spektrums (und natürlich wiederum von der Kondition der Basis $(u_j)_{j=1}^{\infty}$) bestimmt wird. *Insbesondere erhalten wir als wichtige Folgerung, daß der von sensitiven Eigenlösungen aufgespannte invariante Unterraum bezüglich Störungen unempfindlich sein kann, falls die korrespondierenden Eigenwerte entsprechend isoliert vom übrigen Spektrum liegen.* Diese Einsicht wird eine wesentliche Rolle bei der approximativen Lösung des Eigenwertproblems spielen.

2.5 Finite-Elemente-Approximation

Wir wollen uns nun der Bestimmung von Lösungen der Gleichung (21) zuwenden. Bei den betrachteten Problemen aus der integrierten Optik sind dabei nur einige wenige Eigenlösungen, und zwar die zu den Eigenwerten mit kleinstem Realteil, von Interesse. Da sich diese Funktionen in den meisten Fällen nur schwer oder gar nicht mit analytischen Mitteln wie in Beispiel 5 finden lassen, ist man auf Approximationsmethoden angewiesen. Dabei ist es im Kontext von in Variationsform gestellten Aufgaben üblich, die Methode der finiten Elemente anzuwenden.

Dazu zerlegt man das Gebiet Ω, welches zunächst als polygonal begrenzt angenommen wird, in eine endliche Anzahl von Dreiecken t_j (die finiten Elemente), wie es in Abbildung 4 für den Rippen-Wellenleiter aus Beispiel 2 (Seite 6) gezeigt ist. Die Menge

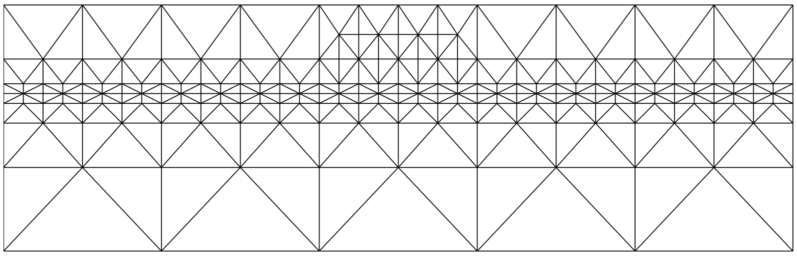

Abbildung 4: Triangulierung eines Gebietes

der Dreiecke $t = \{t_1, \ldots, t_{N_t}\}$ heißt Triangulierung von Ω. Formal stellen wir an eine Triangulierung die folgenden Anforderungen [28, S. 158 f.]:

(i) Die Dreiecke t_j sind offen.

(ii) Es gilt $t_j \cap t_k = \emptyset$ für $j \neq k$, das heißt die Dreiecke sind disjunkt.

(iii) Der Schnitt $\overline{t}_j \cap \overline{t}_k$ mit $j \neq k$ ist entweder

(a) leer,

(b) eine gemeinsame Kante der Dreiecke t_j und t_k oder

(c) ein gemeinsamer Punkt der Dreiecke t_j und t_k.

(iv) Es ist $\overline{\Omega} = \bigcup_{j=1}^{N_t} \overline{t}_j$.

Die Größe h gibt im folgenden die maximale Seitenlänge der Dreiecke aus t an. Ein Punkt aus $\overline{\Omega}$ heißt Knoten der Triangulierung, falls er Eckpunkt eines Dreiecks aus t ist. Die Menge aller Knoten $p = \{p_1, \ldots, p_{N_p}\}$ mit $p_l = (x_l, y_l)$ bildet die Punkt- oder Knotenmenge der Triangulierung, wobei zwischen inneren Knoten aus Ω und Randknoten aus $\partial\Omega$ unterschieden wird. Die Anzahl der inneren Knoten bezeichnen wir mit N. Zu einem inneren Knoten p_l assoziiert man nun eine auf jedem Dreieck lineare und auf ganz $\overline{\Omega}$ stetige Funktion $w_l(x, y)$, für die die Relation

$$w_l(x_m, y_m) = \delta_{lm}, \quad m = 1, \ldots, N_p,$$

gilt. Ein Beispiel für eine solche Funktion ist in Abbildung 2 auf Seite 9 gezeigt. Der Träger dieser Funktion besteht aus allen Dreiecken, die den Knoten p_l als Eckpunkt haben, und ist damit stark lokalisiert. Der von den Funktionen w_l aufgespannte Raum $W_N = \mathrm{span}(w_l)_{l=1}^N =: W_h$ ist ein N-dimensionaler linearer Unterraum von $H_0^1(\Omega)$ und wird Finite-Elemente-Raum genannt. Er enthält stückweise lineare und auf $\overline{\Omega}$ stetige Funktionen, die auf dem Rand von Ω verschwinden. Jedes $u_h \in W_h$ kann eindeutig in der Form

$$u_h = \sum_{l=1}^{N} u_h^{(l)} w_l$$

dargestellt werden, wobei $u_h^{(l)} = u_h(x_l, y_l)$ der Funktionswert von u_h im Knoten p_l ist. Die näherungsweise Berechnung von Eigenlösungen u_j der Gleichung (21) zu den Eigenwerten λ_j mit $j = 1, \ldots, q$ wird jetzt mit dem Ritz-Galerkin-Verfahren [28, S. 228] durchgeführt. Dabei wollen wir uns auf konvexe Gebiete beschränken. Wie in Bemerkung 4 gilt dann $u_j \in H^2(\Omega) \cap H_0^1(\Omega)$ für jedes $f \in L^\infty(\Omega)$, und aus dem Sobolev-Lemma [28, S. 117] folgt die Stetigkeit von u_j auf $\overline{\Omega}$. Das Ritz-Galerkin-Verfahren besteht nun im einzelnen darin, den von den Funktionen $u_j \in H_0^1(\Omega)$ aufgespannten invarianten Unterraum $V_q = \mathrm{span}(u_j)_{j=1}^q$ durch einen von Funktionen $u_{j,h} \in W_h \subset H_0^1(\Omega)$ aufgespannten diskreten invarianten Unterraum $V_{q,h} = \mathrm{span}(u_{j,h})_{j=1}^q$ und die Eigenwerte λ_j durch Werte $\lambda_{j,h} \in \mathbb{C}$ zu approximieren. Dabei werden die Funktionen $u_{j,h} \neq 0$ und die komplexen Zahlen $\lambda_{j,h}$ als Lösungen des diskreten Problems

$$a(w_h, u_{j,h}) = \lambda_{j,h}(w_h, u_{j,h}) + \sum_{k=k_j}^{j-1} \tau_{kj,h}(w_h, u_{k,h}) \quad \forall w_h \in W_h \tag{27}$$

bestimmt. Wir setzen

$$u_{j,h} = \sum_{l=1}^{N} u_{j,h}^{(l)} w_l = \sum_{l=1}^{N} u_{lj,h} w_l, \quad j = 1, \ldots, q,$$

und erhalten so die zu (27) äquivalente Gleichung

$$\sum_{l=1}^{N} a(w_m, w_l) u_{lj,h} = \sum_{l=1}^{N} (w_m, w_l) \sum_{k=k_j}^{j} u_{lk,h} \tau_{kj,h}, \quad m = 1, \ldots, N, \tag{28}$$

mit $\tau_{jj,h} = \lambda_{j,h}$. Da uns die Größe k_j nicht zugänglich ist, wählen wir ab jetzt stets $k_j = 1$. Führen wir nun die Matrizen

$$A_h = \big(a(w_m, w_l)\big)_{\substack{m=1,\ldots,N \\ l=1,\ldots,N}} \quad , \quad U_h = \big(u_{lk,h}\big)_{\substack{l=1,\ldots,N \\ k=1,\ldots,q}} \quad ,$$
$$B_h = \big((w_m, w_l)\big)_{\substack{m=1,\ldots,N \\ l=1,\ldots,N}} \quad , \quad T_h = \big(\tau_{kj,h}\big)_{\substack{k=1,\ldots,q \\ j=1,\ldots,q}}$$

mit $\tau_{kj,h} = 0$ für $k > j$ ein, so können wir (28) auch kompakt in der Form

$$A_h U_h = B_h U_h T_h, \tag{29}$$

also als Matrixeigenwertproblem, schreiben. Dabei sind die Matrizen A_h und B_h aufgrund der Lokalität der Funktionen w_l sehr dünn besetzt („sparse“), die Matrix T_h ist eine obere Dreiecksmatrix. Die Matrix A_h wird Systemmatrix und die Matrix B_h Massenmatrix genannt. Die Massenmatrix ist als Gramsche Matrix der Basisfunktionen w_l stets selbstadjungiert und positiv definit.

Zur weiteren Untersuchung des Matrixeigenwertproblems (29) nutzen wir die Cholesky-Zerlegung $B_h = R_h^* R_h$ der Massenmatrix. Wir führen die $(N \times q)$-Matrix der Unbekannten $X_h = R_h U_h$ ein und erhalten so das äquivalente Problem

$$C_h X_h = X_h T_h \quad \text{mit} \quad C_h = R_h^{-*} A_h R_h^{-1} \quad . \tag{30}$$

Die Grundlage für die nun folgenden Betrachtungen bildet die schon in Abschnitt 2.3 verwendete Schur-Zerlegung [24, S. 335] von Matrizen: zu jeder $(N \times N)$-Matrix C_h existiert eine unitäre $(N \times N)$-Matrix Q_h und eine obere $(N \times N)$-Dreiecksmatrix S_h, so daß

$$C_h Q_h = Q_h S_h \tag{31}$$

gilt. In der Diagonalen von S_h stehen dabei die N Eigenwerte von C_h, und die Matrix Q_h kann so gewählt werden, daß die Eigenwerte nach der Größe ihrer Realteile, also gemäß $\Re\lambda_{1,h} \leq \Re\lambda_{2,h} \leq \ldots \leq \Re\lambda_{N,h}$ sortiert sind. Um die Eigenwerte $\lambda_1, \ldots, \lambda_q$

möglichst gut zu approximieren, wählen wir als Lösung X_h des Problems (30) die ersten q Spalten der Matrix Q_h. Die obere Dreiecksmatrix T_h besteht folglich aus den ersten q Zeilen und Spalten der Matrix S_h. Die gesuchte Matrix U_h der Gleichung (29) bestimmt sich nun aus $U_h = R_h^{-1} X_h$. Da Q_h eine unitäre Matrix ist, gilt $U_h^* B_h U_h = I$. Die aus der Koeffizientenmatrix U_h gebildeten Funktionen $u_{1,h}, \dots, u_{q,h} \in W_h$ bilden also ein Orthonormalsystem in W_h.

Bemerkung 9. Aufgrund der Orthonormalität der Funktionen $u_{j,h}$ erhalten wir aus (27) für die diskreten Eigenwerte die Relation

$$a(u_{j,h}, u_{j,h}) = \lambda_{j,h} \quad ,$$

das heißt $\lambda_{j,h}$ ist gleich dem Rayleigh-Quotienten $R(u_{j,h})$. Da $u_{j,h}$ eine Funktion aus $H_0^1(\Omega)$ ist, ergibt sich nun aus dem Beweis von Lemma 1 (iii), daß auch die diskreten Eigenwerte in dem in Lemma 1 angegebenen Halbstreifen Σ der komplexen Ebene liegen.

Die Funktion $u_{1,h}$ ist wegen

$$a(w_h, u_{1,h}) = \lambda_{1,h}(w_h, u_{1,h}) \quad \forall w_h \in W_h$$

immer eine diskrete Eigenfunktion und damit eine direkte Approximation von u_1. Ist das Eigenwertproblem selbstadjungiert, die Funktion $f \in L^\infty(\Omega)$ also reellwertig, so haben wir

$$\overline{a(w_l, w_m)} = a^*(w_m, w_l) = a(w_m, w_l) \quad ,$$

das heißt die Matrix A_h und somit auch die Matrix C_h sind selbstadjungiert. Aus der Schur-Zerlegung (31) ergibt sich stets $S_h = Q_h^* C_h Q_h$, und wegen $S_h^* = Q_h^* C_h^* Q_h$ erhalten wir im selbstadjungierten Fall $S_h^* = S_h$. Die Matrix S_h und damit die Matrix T_h bilden hier also Diagonalmatrizen mit den reellen diskreten Eigenwerten in der Diagonalen. Bei selbstadjungierten Eigenwertproblemen sind die Funktionen $u_{2,h}, \dots, u_{q,h}$ folglich ebenfalls diskrete Eigenfunktionen. Im allgemeinen Fall nennen wir die Funktionen $u_{j,h}$ diskrete Schurfunktionen.

Wir wollen nun für das selbstadjungierte Eigenwertproblem eine äquivalente Formulierung des diskreten Problems (27) beschreiben. Dabei gehen wir vom Courantschen Minimaxprinzip (12) in Theorem 1 aus. Zur approximativen Berechnung der Eigenwerte λ_j und der Eigenfunktionen u_j wird jetzt das Minimum nur über alle j-dimensionalen Unterräume von $W_h \subset H_0^1(\Omega)$ gemäß

$$\lambda_{j,h} = \min_{\substack{W_{j,h} \subset W_h \\ \dim W_{j,h} = j}} \max_{u_h \in W_{j,h} \setminus \{0\}} R(u_h) = R(u_{j,h}), \quad j = 1, \dots, q, \tag{32}$$

mit dem Rayleigh-Quotienten $R(u_h) = a(u_h, u_h)/(u_h, u_h)$ bestimmt [57, S. 223]. Daraus erhalten wir unmittelbar die Relation

$$\lambda_j \leq \lambda_{j,h} \quad ,$$

das heißt die approximativen Eigenwerte sind obere Schranken für die exakten Eigenwerte.
Unter Benutzung genereller Resultate zur Finite-Elemente-Approximation läßt sich mit Hilfe des Minimaxprinzips (32) auch eine Aussage über den Fehler in den Eigenwerten treffen [57, S. 228 ff.]. Setzen wir eine quasiuniforme Triangulierung des konvexen Gebietes Ω voraus, das heißt gilt für die Innenwinkel α aller Dreiecke die Relation $\alpha \geq \alpha_0 > 0$, so ergibt sich für hinreichend kleines h die Ungleichung

$$\lambda_{j,h} - \lambda_j \leq Ch^2 \|u_j\|^2_{H^2(\Omega)} \quad .$$

Für den Fehler in den Eigenfunktionen erhält man die Abschätzungen

$$\|u_{j,h} - u_j\|_{H^1(\Omega)} \leq C'h\|u_j\|_{H^2(\Omega)} \quad \text{bzw.} \quad \|u_{j,h} - u_j\| \leq C''h^2\|u_j\|_{H^2(\Omega)} \quad .$$

Im Falle eines mehrfachen Eigenwertes beziehen sich diese Fehlerschranken jeweils auf spezielle Linearkombinationen der zu diesem mehrfachen Eigenwert korrespondierenden Eigenfunktionen. Dabei ist stets $\|u_{j,h}\| = \|u_j\| = 1$ zu beachten.
Für nichtselbstadjungierte Eigenwertprobleme lassen sich mit Hilfe einer anderen Beweistechnik analoge Resultate für die Approximationsfehler herleiten [3, S. 692 ff.]. Wir betrachten den allgemeinen Fall eines m-fachen Eigenwertes $\lambda_{k+1} = \ldots = \lambda_{k+m}$ mit dem korrespondierenden verallgemeinerten Eigenraum N_k. Die assoziierten diskreten Eigenwerte $\lambda_{k+1,h}, \ldots, \lambda_{k+m,h}$ sind nicht notwendigerweise gleich, der zugehörige diskrete invariante Unterraum sei mit $N_{k,h}$ bezeichnet. Für die Fehler in den Eigenwerten gelten nun unter den obigen Voraussetzungen an die Triangulierung die Abschätzungen

$$\left|\lambda_{k+l} - \lambda_{k+l,h}\right| \leq C\delta_k^{2/\gamma_k} \quad \text{und} \quad \left|\lambda_{k+l} - \frac{1}{m}\sum_{n=1}^{m} \lambda_{k+n,h}\right| \leq C'\delta_k^2 \quad ,$$

für den Abstand zwischen N_k und $N_{k,h}$ in $H_0^1(\Omega)$ ergibt sich

$$\operatorname{dist}(N_k, N_{k,h}) \leq C''\delta_k \quad .$$

Dabei ist γ_k der Index des Eigenwertes λ_{k+l}. Die Größe δ_k ist durch

$$\delta_k = \sup_{\substack{u \in N_k \\ \|u\|=1}} \inf_{u_h \in W_h} \|u - u_h\|_{H^1(\Omega)}$$

gegeben und erfüllt in unserem Fall die in [7, S. 74] angegebene Ungleichung

$$\delta_k \leq c\, h \sup_{\substack{u \in N_k \\ \|u\|=1}} \|u\|_{H^2(\Omega)} \quad .$$

Bemerkung 10. Im nichtselbstadjungierten Fall erhalten wir also durch Übergang zum arithmetischen Mittel der approximativen Eigenwerte und zum korrespondierenden invarianten Unterraum die gleichen Fehlerordnungen in h wie bei selbstadjungierten Problemen.

Bemerkung 11. Anstelle von stückweise linearen Funktionen kann der Raum W_h auch durch Funktionen, die auf jedem Dreieck mit einem Polynom vom Grad $p > 1$ übereinstimmen, aufgespannt werden. Unter der zusätzlichen Voraussetzung der $H^{p+1}(\Omega)$-Regularität der Eigenlösungen gilt dann für δ_k die Abschätzung (siehe [7, S. 74])

$$\delta_k \leq c\, h^p \sup_{\substack{u \in N_k \\ \|u\|=1}} \|u\|_{H^{p+1}(\Omega)} \quad .$$

Die Erweiterung der eben beschriebenen Finite-Elemente-Approximationsmethode auf krummlinig berandete, konvexe Gebiete erfolgt wie in Abbildung 5 für ein Kreisgebiet dargestellt: man ersetzt das Gebiet Ω durch ein approximierendes polygonales Gebiet

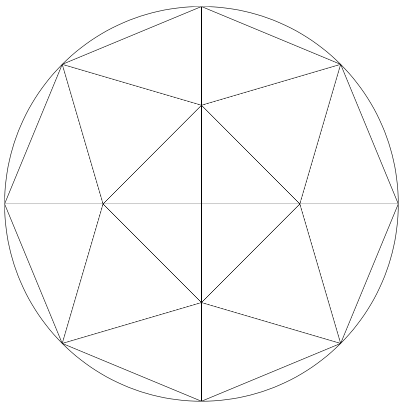

Abbildung 5: Triangulierung eines krummlinig berandeten, konvexen Gebietes

$\Omega_h \subset \Omega$. Die zu Ω_h bestimmten Näherungen der Eigenlösungen werden dann durch die Fortsetzung $u_{j,h} = 0$ in $\Omega \setminus \Omega_h$ auf ganz Ω erklärt. Der dabei auftretende Approximationsfehler liegt für konvexe Gebiete und stückweise lineare Basisfunktionen in der Größenordnung des ohnehin auftretenden Diskretisierungsfehlers [28, S. 177 f.].

Abschließend wollen wir noch auf die Berechnung der Matrixelemente $a(w_m, w_l)$ und (w_m, w_l) der System- und Massenmatrix eingehen. Die dabei auftretenden Integralausdrücke werden hierbei nicht knoten-, sondern elementorientiert ausgewertet und dann entsprechend aufsummiert [54, S. 158 ff.]. Da die Funktionen w_l auf jedem Dreieck linear sind, können die Skalarprodukte $(\nabla w_m, \nabla w_l)$ und (w_m, w_l) für jedes Dreieck leicht berechnet werden. Die Funktion $f \in L^\infty(\Omega)$ setzen wir als stückweise stetig voraus, wobei im Falle von Sprungstellen die Triangulierung so gewählt werden soll, daß jede Sprungstelle auf einer Dreieckskante liegt und somit die Funktion f auf jedem Element stetig ist. Die Integrale $(w_m, f w_l)$ werden nun durch eine Mittelpunktregel gemäß $(w_m, f_s w_l)$ approximiert, wobei f_s den Funktionswert von f im Schwerpunkt des jeweiligen Dreiecks bezeichnet. Diese Quadraturformel ist für lineare Basisfunktionen optimal, da der auftretende Quadraturfehler die gleiche Ordnung in h wie der Approximationsfehler der Finite-Elemente-Methode hat [57, S. 181 ff.].

2.6 Krylovraum-Methoden zur Lösung des Eigenwertproblems

In diesem Abschnitt sollen das wesentliche Prinzip zur numerischen Lösung des diskreten Eigenwertproblems in Matrixform (29) und einige darauf basierende, gängige Lösungsverfahren, die Krylovraum-Methoden, dargestellt werden. Den Diskretisierungsparameter h werden wir dabei im folgenden als Index unterdrücken. Das numerisch zu lösende Problem lautet damit wie folgt: gesucht sind eine $(N \times q)$-Matrix U und eine obere $(q \times q)$-Dreiecksmatrix T, so daß

$$\begin{aligned} AU &= BUT \\ U^* BU &= I \end{aligned} \tag{33}$$

gilt, wobei in der Diagonalen von T die q Eigenwerte mit kleinstem Realteil des Matrixpaares (A, B) stehen. Die Spalten $u_1, \ldots, u_q$ der Matrix U bilden im selbstadjungierten Fall Eigenvektoren, im allgemeinen Fall Schurvektoren.

Bemerkung 12. Zur Illustration des Unterschiedes zwischen Schurvektoren und Eigenvektoren bei nichtselbstadjungierten Eigenwertproblemen betrachten wir das Matrixpaar

$$A = \begin{pmatrix} i\,z & 1 \\ 1 & 0 \end{pmatrix} \quad , \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

mit dem reellen Parameter $z \in (-2, 2)$. Die Matrix U der Schurvektoren und die obere Dreiecksmatrix T aus (33) bestimmen sich zu

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -\lambda_1 \\ -\lambda_2 & 1 \end{pmatrix} \quad \text{und} \quad T = \begin{pmatrix} \lambda_1 & \tau_{12} \\ 0 & \lambda_2 \end{pmatrix} \quad ,$$

wobei die Eigenwerte λ_1, λ_2 und die Größe τ_{12} durch

$$\lambda_1 = -\sqrt{1 - \frac{z^2}{4}} + i\,\frac{z}{2} \quad , \quad \lambda_2 = \sqrt{1 - \frac{z^2}{4}} + i\,\frac{z}{2} \quad , \quad \tau_{12} = z\left(\frac{z}{2} + i\sqrt{1 - \frac{z^2}{4}}\right)$$

gegeben sind. Die Matrix V der Eigenvektoren ergibt sich gemäß

$$V = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -\lambda_2 & -\lambda_1 \end{pmatrix} \quad .$$

Für $z \to \pm 2$ erhalten wir $\lambda_{1,2} \to \pm i$, das heißt die Eigenvektoren werden im Grenzfall linear abhängig, während die Schurvektoren stets ein Orthonormalsystem bilden. *Im Sinne des Konditionsbegriffes einer Basis wie in Bemerkung 7 erhalten wir also, daß die aus Schurvektoren bestehende Basis immer gut konditioniert ist, während die aus Eigenvektoren gebildete Basis eine schlechte Kondition haben kann.* Im speziellen Beispiel von oben haben wir

$$\kappa(u_1, u_2) = 1 \quad \text{und} \quad \kappa(v_1, v_2) = \sqrt{\frac{1 + |z|/2}{1 - |z|/2}} \xrightarrow[z \to \pm 2]{} \infty \quad .$$

Daher bilden im allgemeinen Fall die Schurvektoren die Grundlage zur Berechnung invarianter Unterräume.

Die Matrizen A und B sind in den von uns betrachteten Fällen stets schwach besetzte $(N \times N)$-Matrizen, die Größe N liegt typischerweise zwischen 10^3 und 10^5. Die Anzahl q der gewünschten Eigenlösungen bewegt sich in der Größenordnung $q = 1, \ldots, 10$, das heißt es gilt $q \ll N$. Damit scheiden die „klassischen“ Iterationsmethoden zur Lösung des Eigenwertproblems, wie das Verfahren von JACOBI [31] für selbstadjungierte Matrizen oder der QR-Algorithmus von FRANCIS [20] und KUBLANOVSKAYA [36] für den allgemeinen Fall aus, da diese sämtliche N Eigenlösungen liefern.

Eine weitere Restriktion in der Wahl der Lösungsmethode wird durch die Art der q gesuchten Eigenwerte vorgegeben. Das einfachste Verfahren zur Berechnung des zu bestimmten Eigenwerten gehörenden invarianten Unterraumes ist die simultane Vektoriteration von RUTISHAUSER [49] und STEWART [56]. In der Variante der inversen Iteration mit Spektralverschiebung wird dabei der invariante Unterraum zu den Eigenwerten, die betragsmäßig am dichtesten zum gewählten Verschiebungsparameter liegen, bestimmt. Da wir an den Eigenwerten mit kleinstem Realteil interessiert sind, können wir diesen einfachen Algorithmus zur numerischen Lösung des Problems (33) im allgemeinen nicht nutzen.

Das den „modernen“ Iterationsverfahren zur Berechnung invarianter Unterräume zugrundeliegende Prinzip ist das diskrete Analogon zu der im letzten Abschnitt beschriebenen Finite-Elemente-Approximation. Wir wählen einen n-dimensionalen Unterraum des zugrundeliegenden unitären Vektorraumes $\mathbb{C}^N$ mit $q \le n \le N$ und bestimmen durch Lösung einer Eigenwertaufgabe der Dimension n Näherungen aus diesem Unterraum zu den in (33) gesuchten Matrizen U und T. Sei dazu durch $(w_1, \ldots, w_n) = W$ eine Basis des entsprechenden Unterraumes gegeben. Die Berechnung der Approximationen $\widetilde{U}$ und $\widetilde{T}$ wird nun wiederum mit dem Ritz-Galerkin-Verfahren durchgeführt. Wir setzen $\widetilde{U} = WQ$ mit einer $(n \times q)$-Matrix Q und lösen anstelle von (33) das Problem

$$\begin{aligned} \big(W^*AW\big)Q &= \big(W^*BW\big)Q\widetilde{T} \\ Q^*\big(W^*BW\big)Q &= I \quad . \end{aligned} \tag{34}$$

Setzen wir zusätzlich voraus, daß die Spalten von W ein Orthonormalsystem im Sinne des durch B induzierten Skalarproduktes bilden, so vereinfacht sich (34) zu dem Standardproblem

$$\begin{aligned}(W^*AW)Q &= Q\widetilde{T} \\ Q^*Q &= I \quad .\end{aligned} \tag{35}$$

Um die gewünschten Matrizen U und T möglichst gut zu approximieren, wählen wir Q und $\widetilde{T}$ so, daß in der Diagonalen von $\widetilde{T}$ die Eigenwerte $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_q$ von W^*AW mit den kleinsten Realteilen stehen. Dieses allgemeine Prinzip zur näherungsweisen Lösung von Eigenwertproblemen wird orthogonale Projektionsmethode [50, S. 127 ff.] genannt. Da der Koeffizientenmatrix W ein n-dimensionaler Unterraum des im vorherigen Abschnitt eingeführten Finite-Elemente-Raumes entspricht, ist das Verfahren auch als orthogonale Projektionsmethode im Funktionenraum $H_0^1(\Omega)$ zu verstehen.
Die Idee der iterativen Löser ist jetzt, durch wiederholte Lösung von Problemen der Form (35), sei es durch Erhöhung der Dimension n oder Verwendung immer „besser“ werdender Ansatzmatrizen W zu festem n, eine Folge U_k und T_k zu konstruieren, die gegen die gesuchten Lösungen U und T aus (33) streben. Zu dieser Klasse von Verfahren gehören die Krylovraum-Methoden, bei denen sukzessive die Dimension des von den Spaltenvektoren der Matrix W aufgespannten Unterraumes (des Krylovraumes) erhöht wird. Diese Verfahren sollen im folgenden kurz beschrieben werden.
Wir beginnen mit dem Block-Arnoldi-Verfahren zur Berechnung des zu den Eigenwerten $\lambda_1, \dots, \lambda_q$ korrespondierenden invarianten Unterraumes des Matrixpaares (A, B). Da wir an Eigenwerten aus dem „unteren“ Teil des Spektrums interessiert sind, legen wir der Herleitung des Verfahrens ein inverses und spektralverschobenes Eigenwertproblem zugrunde. Die Spektralverschiebung wird dabei gemäß $\mu = \sup_\Omega \Re f$ gewählt. Aus Bemerkung 9 folgt dann für die Eigenwerte die Relation $\Re\lambda_j + \mu > 0$ und damit die Invertierbarkeit der Matrix $A + \mu B$. Unter Verwendung der Matrizen $C = R(A + \mu B)^{-1}R^*$, $X = RU$ und $S = (T + \mu I)^{-1}$, wobei R aus der Cholesky-Zerlegung der Matrix $B = R^*R$ bestimmt wird, ergibt sich nun das zu (33) äquivalente Standard-Eigenwertproblem

$$\begin{aligned}CX &= XS \\ X^*X &= I \quad .\end{aligned}$$

Aus dem Block-Arnoldi-Verfahren von Saad [50, S. 195 ff.] zur Lösung dieses speziellen Problems erhalten wir zur Berechnung der Matrix W den folgenden

Algorithmus 1.

- Initialisierung:
  - Vorgabe von $(w_1, \dots, w_q) = W_q$ mit $W_q^* B W_q = I$

- Iteration für $m = 1, 2, \ldots$:
  - $*$ bestimme v aus $(A + \mu B)v = Bw_m$
  - $*$ berechne $h_{l\,m} = w_l^* Bv$ für $l = 1, \ldots, m + q - 1$
  - $*$ setze $u = v - \sum_{l=1}^{m+q-1} h_{l\,m} w_l$
  - $*$ berechne $h_{m+q\,m} = \sqrt{u^* Bu}$
  - $*$ setze $w_{m+q} = \dfrac{u}{h_{m+q\,m}}$, falls $h_{m+q\,m} \neq 0$

Ergibt sich im Laufe der Iteration einmal $h_{m+q\,m} = 0$, so wird q durch $q - 1$ ersetzt und der Algorithmus mit diesem neuen Wert fortgesetzt. Ist $q = 1$ und $h_{m+q\,m} = 0$, so wurde ein invarianter Unterraum des Matrixpaares (A, B), aufgespannt durch die Vektoren $w_1, \ldots, w_m$, gefunden, und die Iteration wird beendet. In jedem Schritt des Algorithmus ist zur Bestimmung des Vektors v ein lineares Gleichungssystem mit der Koeffizientenmatrix $A + \mu B$ zu lösen. Daher ist zu Beginn der Iteration eine Faktorisierung dieser Matrix vorzunehmen. Ganz allgemein ist in jeder Variante des Arnoldi-Verfahrens zur Lösung des Eigenwertproblems (33) eine Zerlegung der Matrix A oder B oder einer Linearkombination beider Matrizen erforderlich. Die aus den iterierten Vektoren gebildete Matrix $W_m = (w_1, \ldots, w_m)$ erfüllt in exakter Arithmetik die Orthogonalitätsrelation $W_m^* B W_m = I$. Die numerische Orthogonalisierung der Vektoren wird mit den von Ruhe in [48] angegebenen iterativen Gram-Schmidt-Verfahren durchgeführt. Die Berechnung von Approximationen T_k und U_k der Lösungen des Eigenwertproblems (33) erfolgt nun in unserem speziellen Fall nicht durch Lösung eines Problems der Form (35) mit $W = W_m$, sondern durch Lösung von

$$\begin{aligned} H_m Q &= Q\widetilde{S} \\ Q^* Q &= I \quad , \end{aligned} \tag{36}$$

wobei die Matrix $H_m = (h_{j\,k})_{j,k=1,\ldots,m}$ aus den in Algorithmus 1 bestimmten Koeffizienten $h_{j\,k}$ gebildet wird. Aus $H_m = W_m^* B(A + \mu B)^{-1} B W_m = Y_m^* C Y_m$ mit $Y_m = RW_m$ sehen wir, daß (36) ein projiziertes Problem zur näherungsweisen Bestimmung von Eigenlösungen der Matrix C ist. Um die q Eigenwerte des Matrixpaares (A, B) mit kleinstem Realteil möglichst gut zu approximieren, wählen wir die obere Dreiecksmatrix $\widetilde{S}$ so, daß in der Diagonalen von $\widetilde{S}$ die q Eigenwerte von H_m stehen, deren Kehrwerte die kleinsten Realteile aufweisen. Die Näherungen T_k und U_k ergeben sich dann aus $T_k = \widetilde{S}^{-1} - \mu I$ und $U_k = W_m Q$. Dazu müssen wir noch zeigen, daß in unserem Fall die Matrix $\widetilde{S}$ stets invertierbar ist. Zunächst haben wir aus (36) die Darstellung $\widetilde{S} = Q^* H_m Q = (Y_m Q)^* C (Y_m Q)$. Da $\widetilde{S}$ eine obere Dreiecksmatrix ist, folgt also die Invertierbarkeit von $\widetilde{S}$ aus der Relation $|x^* C x| > 0$ für alle $x \in \mathbb{C}^N \setminus \{0\}$. Zum Nachweis dieser Ungleichung beachten wir $x^* C x = x^* R (A + \mu B)^{-1} R^* x = g^* (A + \mu B)^{-1} g$

mit $g = R^*x$. Die Matrix $A + \mu B$ ist nun in den von uns betrachteten Fällen komplex symmetrisch, das heißt wir haben $A + \mu B = L + iM$, wobei die Matrizen L und M reell und symmetrisch sind. Aus dem Beweis von Lemma 1 (iii) erhalten wir $u^*Lu = \Re u^*(A + \mu B)u > 0$ für $u \neq 0$, das heißt die Matrix L ist positiv definit und daher invertierbar. Die Links- und Rechtsinverse von $A + \mu B$ ergeben sich damit formal zu $(L + ML^{-1}M)^{-1}(I - iML^{-1})$ bzw. $(I - iL^{-1}M)(L + ML^{-1}M)^{-1}$. Die in beiden Formeln vorkommende Matrix $L + ML^{-1}M$ ist reell und symmetrisch, wegen $u^*(L + ML^{-1}M)u = u^*Lu + (Mu)^*L^{-1}(Mu) > 0$ für $u \neq 0$ positiv definit und daher tatsächlich invertierbar. Aus der bereits oben bemerkten Invertierbarkeit von $A + \mu B$ ergibt sich nun die Übereinstimmung von Links- und Rechtsinverser und damit speziell $\Re(A + \mu B)^{-1} = (L + ML^{-1}M)^{-1}$. Da sowohl der Realteil als auch der Imaginärteil von $(A + \mu B)^{-1}$ reell symmetrische Matrizen sind, erhalten wir $\Re x^*Cx = \Re g^*(A+\mu B)^{-1}g = g^*\Re(A+\mu B)^{-1}g$. Daraus ergibt sich schließlich wegen der positiven Definitheit von $(L+ML^{-1}M)^{-1}$ die Relation $\Re x^*Cx > 0$ für $x \neq 0$, das heißt in unserem Spezialfall die zu zeigende Invertierbarkeit der Matrix $\widetilde{S}$. Diese Eigenschaft kann im allgemeinen Fall nicht gewährleistet werden, wie folgendes Beispiel zeigt.

Beispiel 8. Unter Benutzung von Algorithmus 1 erhalten wir für das Matrixpaar

$$A = \begin{pmatrix} \frac{1}{8} & \frac{19}{16} & -\frac{19}{8} & \frac{27}{8} \\ 0 & \frac{1}{4} & -\frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad , \quad B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

mit $q = 1$ und $\mu = 0$ zum Startvektor $w_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 & -1 & -1 \end{pmatrix}^*$ die Matrizen

$$W_2^*AW_2 = \begin{pmatrix} \frac{41}{64} & \frac{1}{7}\sqrt{14} \\ \frac{17}{448}\sqrt{14} & 0 \end{pmatrix} \quad \text{und} \quad H_2 = \begin{pmatrix} 0 & 0 \\ \frac{1}{2}\sqrt{14} & 0 \end{pmatrix} \quad .$$

Aus

$$\sigma(W_2^*AW_2) = \left\{\frac{287 \mp \sqrt{143297}}{896}\right\} \quad \text{und} \quad \sigma(H_2) = \{0\}$$

ergibt sich als Lösung von (35) bzw. (36)

$$\widetilde{T} = \widetilde{\lambda}_1 \doteq -0.1022 \quad \text{bzw.} \quad \widetilde{S} = \widetilde{\mu}_1 = 0 \quad ,$$

das heißt die Matrix $\widetilde{S}$ kann im allgemeinen Fall nicht invertierbar sein. ◇

Da zur Berechnung der Koeffizienten $h_{l\,m}$ im Gram-Schmidt-Verfahren und zur Bestimmung der Approximationen U_k alle bereits berechneten Vektoren w_l benötigt werden, müssen diese zunächst während der gesamten Iteration im Rechenspeicher gehalten werden. Eine Beschränkung des benötigten Speicherplatzes ist durch einen Neustart des Verfahrens nach einer bestimmten Anzahl $m_{\max}$ von Iterationen möglich. Ein solcher Neustart wirkt auch dem durch Rundungsfehler unvermeidlich auftretenden langsamen Verlust der Orthogonalität der Basisvektoren entgegen. Man setzt nach einer zusätzlichen Orthonormalisierung der Spaltenvektoren der zuletzt bestimmten Approximation U_k von U zur Initialisierung $W_q = U_k$ und beginnt die Iteration mit $m = 1$ von neuem. Andere Varianten zur speichersparenden und numerisch stabilen Realisierung des Arnoldi-Verfahrens sind in den Arbeiten von Sorensen [55] und Lehoucq [37] dargestellt und in dem Programmpaket Arpack [38] implementiert.
Im Fall eines selbstadjungierten Eigenwertproblems ergeben sich in exakter Arithmetik Vereinfachungen in der Arnoldi-Iteration. Aus der Selbstadjungiertheit der Matrix A folgt die Relation $C^* = C$. Damit ergibt sich, daß die Matrix H_m selbstadjungiert und demzufolge eine Bandmatrix der Bandbreite q ist. Zur Bestimmung des Vektors w_{m+q} im m-ten Iterationsschritt werden somit nur die zuvor berechneten Vektoren $w_{m-q}, \dots, w_{m+q-1}$ benötigt, das heißt die Iteration läßt sich in einer $(2q+1)$-Term-Rekursion realisieren. Ist man nur an der Bestimmung von Eigenwerten interessiert, müssen also zur Durchführung des Verfahrens in exakter Arithmetik im wesentlichen nur die $2q+1$ zuletzt bestimmten Basisvektoren w_l und die Bandmatrix H_m gespeichert werden. In endlicher Arithmetik stellt man jedoch fest, daß die theoretisch vorhandene Orthogonalität der Vektoren w_l mit wachsender Iterationszahl sehr rasch verlorengeht, insbesondere sobald ein approximatives Eigenpaar $(\widetilde{\lambda}_j, \widetilde{u}_j)$ gegen ein Eigenpaar (λ_j, u_j) von (A, B) konvergiert [44, S. 262 ff.]. Dabei können zu einem einfachen Eigenwert des Matrixpaares (A, B) mehrere korrespondierende approximative Eigenwerte auftreten [12, S. 121 ff.]. Eine Möglichkeit, diesem Phänomen zu begegnen, besteht in der Methode der partiellen oder selektiven Reorthogonalisierung [44, S. 272 ff.]. Das Arnoldi-Verfahren zur Lösung des selbstadjungierten Eigenwertproblems wird auch als Lanczos-Algorithmus bezeichnet.
Aus dem Lanczos-Verfahren für selbstadjungierte Probleme läßt sich eine weitere Krylovraum-Methode zur Lösung allgemeiner Eigenwertprobleme ableiten, das sogenannte nichtselbstadjungierte Lanczos-Verfahren [50, S. 186 ff.]. Dieses Verfahren beruht nicht auf einer orthogonalen, sondern auf einer „schiefen“ Projektionsmethode und hat die Bestimmung von Eigenwerten mit korrespondierenden rechten und linken Eigenvektoren zum Ziel. Die projizierten Eigenwertprobleme nehmen dabei die Gestalt

$$
\begin{aligned}
\big(\widehat{W}^* A W\big) Q &= Q D \\
\big(W^* A^* \widehat{W}\big) \widehat{Q} &= \widehat{Q}\,\overline{D} \\
\widehat{Q}^* Q &= I
\end{aligned}
\tag{37}
$$

an, wobei die Biorthogonalitätsrelation $\widehat{W}^*BW = I$ vorausgesetzt wurde. Da in unserem speziellen Fall die Matrix A komplex symmetrisch und die Matrix B reell symmetrisch ist, setzen wir $\widehat{W} = \overline{W}$ und erhalten $\widehat{Q} = \overline{Q}$. Das zu lösende projizierte Problem (37) vereinfacht sich daher unter der Annahme $W^TBW = I$ zu

$$\begin{aligned} \left(W^TAW\right)Q &= QD \\ Q^TQ &= I \quad , \end{aligned} \tag{38}$$

wobei die Diagonalmatrix D Approximationen der Eigenwerte mit kleinstem Realteil des Matrixpaares (A, B) enthält. Die Bestimmung der Basisvektoren w_l erfolgt nun in exakter Arithmetik wiederum durch eine $(2q+1)$-Term-Rekursion, das resultierende Verfahren ist der komplex symmetrische Lanczos-Algorithmus [12, S. 199 ff.]. Auf die Schwierigkeiten, die bei der Berechnung von Eigenvektoren komplex symmetrischer Matrizen auftreten können, sind wir bereits in Bemerkung 12 eingegangen. Ein weiterer möglicher Artefakt dieser Methode besteht in einem „serious breakdown“ [58, S. 389 ff.], das heißt es ergibt sich im Laufe der Iteration bei $q = 1$ der Wert $h_{m+q\,m} = 0$, obwohl kein invarianter Unterraum gefunden worden ist. Eine Abhilfe für diesen Fall ist durch die „look-ahead“-Strategien gegeben [21].
Eine Übersicht über die Krylovraum-Methoden zur Lösung des Eigenwertproblems (33) ist in Abbildung 6 dargestellt. Für selbstadjungierte Eigenwertprobleme gehen sowohl

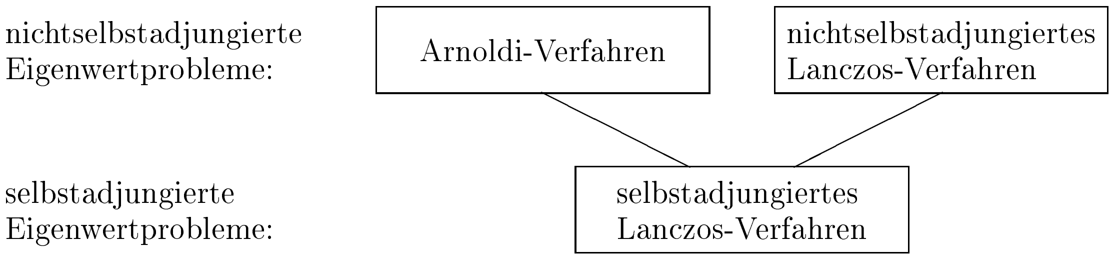

Abbildung 6: Krylovraum-Methoden für Eigenwertprobleme

das Arnoldi-Verfahren als auch das nichtselbstadjungierte Lanczos-Verfahren in das selbstadjungierte Lanczos-Verfahren über.
Der wesentliche Rechenaufwand der Krylovraum-Methoden zur Lösung des Eigenwertproblems (33) besteht in der Auflösung der linearen Gleichungssysteme in jedem Iterationsschritt. Im optimalen Fall läßt sich diese bei den in unserem Zusammenhang auftretenden Matrizen mit einem Aufwand von $O(N^{\frac{3}{2}})$ Operationen direkt durchführen. Anstelle eines direkten Verfahrens kann zur Lösung der linearen Gleichungssysteme natürlich auch eine iterative Methode zum Einsatz kommen. Bei Verwendung einer linearen Mehrgitter-Methode erhält man so beispielsweise optimale Komplexität, das heißt die Anzahl der benötigten Operationen zur näherungsweisen Lösung des Eigenwertproblems ist dann proportional zur Größe N. Das gesamte Verfahren besteht jetzt

aus einer äußeren Iteration zur Bestimmung der Basisvektoren des Krylovraumes und einer inneren Iteration zur Lösung der linearen Systeme.
Die im nächsten Kapitel dargestellten Mehrgitter-Methoden zur Lösung des Eigenwertproblems bilden in unserem Kontext eine Alternative zu den eben beschriebenen Krylovraum-Methoden. Dabei wird das Mehrgitter-Konzept direkt auf das zu lösende Problem (33) angewandt. Der resultierende Algorithmus läßt sich in einer einfachen Iteration mit optimaler Komplexität realisieren. Es wird sich zeigen, daß die entwickelten Verfahren ebenfalls der Klasse der Projektionsmethoden angehören.

3 Mehrgitter-Methoden zur Lösung des Eigenwertproblems

3.1 Überblick

Die Mehrgitter-Methoden zählen zu den schnellsten Verfahren zur Lösung des diskreten Eigenwertproblems (33), insbesondere bei Problemen mit sehr vielen Unbekannten. Der Aufwand zur Lösung des Systems (33) steigt bei diesen Algorithmen nur proportional zur Anzahl N der Freiheitsgrade an. Das bedeutet, die Verfahren haben optimale Komplexität. Zur Anwendung dieser Methoden wird eine Gitterhierarchie benötigt, das heißt eine Folge von Triangulierungen $t_{h_0}, t_{h_1}, \ldots, t_{h_{l_{\max}}}$ mit korrespondierenden Finite-Elemente-Räumen $W_{h_0}, W_{h_1}, \ldots, W_{h_{l_{\max}}}$, wie sie beispielsweise durch fortgesetzte adaptive Gitterverfeinerung entsteht. Wir wollen im Rahmen dieser Arbeit nur konforme Methoden betrachten, also ineinander geschachtelte Triangulierungen, wobei die zugehörigen Räume die Inklusionsrelation

$$W_{h_0} \subset W_{h_1} \subset \ldots \subset W_{h_{l_{\max}}} \subset H_0^1(\Omega)$$

erfüllen.
Das prinzipielle Vorgehen soll zuerst für den Zweigitter-Algorithmus dargestellt werden. Jeder Iterationsschritt der Methode gliedert sich in zwei Etappen, die Glättung und die Bestimmung einer Grobgitter-Korrektur. Im Glättungsschritt auf dem Feingitter werden durch ein einfaches Iterationsverfahren, den Glätter, die hochfrequenten Fehleranteile gedämpft. Zur Behandlung der niederfrequenten Fehleranteile wird danach ein Grobgitter-Korrekturproblem (von wesentlich kleinerer Dimension als das eigentliche Problem) aufgestellt und gelöst. Durch abwechselnde Wiederholung der Glättungs- und Grobgitter-Korrekturschritte entsteht ein iteratives Verfahren über zwei Gittern. Da das zur Bestimmung der Grobgitter-Korrektur zu lösende Problem von der gleichen Bauart wie das ursprüngliche Problem ist, kann dieses nun ebenfalls nach dem eben beschriebenen Prinzip von Glättung und Grobgitter-Korrektur bearbeitet werden. Dadurch entsteht in jedem Iterationsschritt eine Rekursion über mehrere Gitter, das resultierende Verfahren ist eine Mehrgitter-Methode. Die jeweils auftretenden Grobgitterprobleme werden dabei nur noch auf dem gröbsten Gitter exakt, ansonsten typischerweise durch einen oder zwei Iterationsschritte (V-Zyklus oder W-Zyklus) näherungsweise gelöst. Die Glättungsschritte verteilt man üblicherweise gleichmäßig auf die Phase vor und nach der (approximativen) Bestimmung der Grobgitter-Korrektur.
Aus diesem allgemeinen Konzept sind verschiedene Varianten von Mehrgitter-Methoden zur Lösung des Eigenwertproblems (33) entwickelt worden. Diese Verfahren lassen sich grob in zwei Klassen einteilen. Bei den Algorithmen der ersten Klasse wird eine Mehrgitter-Methode zur Lösung auftretender linearer Teilprobleme benutzt, während bei den Methoden der zweiten Klasse das Mehrgitterkonzept direkt auf das zu lösende

Problem (33) angewandt wird. Wir beginnen mit einer kurzen Beschreibung einiger Verfahren der ersten Klasse.

Dem Algorithmus von BANK [4] liegt die inverse Vektoriteration zugrunde. Die dabei in jedem Teilschritt entstehenden Gleichungssysteme werden mit einem Mehrgitter-Verfahren für lineare Probleme gelöst. Bei der Methode von BANK und CHAN [5] wird die Bestimmung einer einfachen Eigenlösung als parameterabhängiges Problem mit dem skalaren Parameter λ angesehen, zu dessen Lösung eine Mehrgitter-Fortsetzungsmethode angewandt wird. HACKBUSCH [27] betrachtet das Eigenwertproblem (33) als nichtlineares Problem in den Unbekannten U und T und benutzt zur Lösung das Newton-Verfahren, wobei die auftretenden linearen Teilprobleme mit einer der Struktur der Jacobi-Matrix angepaßten Mehrgitter-Methode behandelt werden.

Häufig wird das Mehrgitterkonzept auch zur Vorkonditionierung benutzt. Die im nächsten Abschnitt für den selbstadjungierten Fall dargestellte Rayleigh-Quotienten-Minimierung auf dem feinsten Gitter wird bei JUNG et al. [32] mit einem linearen Mehrgitter-Verfahren zur Vorkonditionierung verknüpft. Die von BRAMBLE et al. [8] und LEINEN et al. [39] zur Lösung des selbstadjungierten Eigenwertproblems vorgeschlagene Methode läßt sich in diesem Zusammenhang als spezielle Variante einer vorkonditionierten simultanen Rayleigh-Quotienten-Minimierung auffassen.

Kommen wir nun zu den Verfahren der zweiten Klasse. Dabei wollen wir zunächst die Methode von HACKBUSCH [26, S. 257 ff.] für den Fall der Berechnung eines einfachen Eigenwertes darstellen. Die Gleichung

$$(A - \lambda_1 B)u_1 = 0 \tag{39}$$

wird als parameterabhängiges lineares Problem betrachtet, auf das die Mehrgitter-Idee direkt angewandt wird. Vor jedem Mehrgitter-Zyklus wird dabei eine Näherung des Eigenwertes λ_1 durch Bildung des Rayleigh-Quotienten bzw. verallgemeinerten Rayleigh-Quotienten

$$\widetilde{\lambda}_1 = \frac{\widetilde{u}_1^* A \widetilde{u}_1}{\widetilde{u}_1^* B \widetilde{u}_1} \quad \text{bzw.} \quad \widetilde{\lambda}_1 = \frac{\widetilde{v}_1^* A \widetilde{u}_1}{\widetilde{v}_1^* B \widetilde{u}_1}$$

bestimmt, wobei $\widetilde{u}_1$ und $\widetilde{v}_1$ Approximationen des rechten und linken Eigenvektors u_1 und v_1 sind. Dieser Parameterwert wird dann während des gesamten Zyklus konstant gehalten. Als Glätter wird ein „klassisches" Iterationsverfahren zur Lösung linearer Probleme, beispielsweise das Jacobi- oder das Gauß-Seidel-Verfahren, eingesetzt. Die zur Berechnung der Grobgitter-Korrekturen zu lösenden Probleme sind von der Form

$$(A_l - \widetilde{\lambda}_{1,l} B_l) w_l = d_l, \quad l = 0, \ldots, l_{\max} - 1 \quad .$$

Dabei sind A_l und B_l die zur Triangulierung t_{h_l} korrespondierenden Matrizen, der Wert $\widetilde{\lambda}_{1,l}$ eine Näherung des zu diesem Gitter gehörenden Eigenwertes $\lambda_{1,l}$ mit kleinstem

Realteil und die Größe d_l der Defekt, der sich aus dem aktuellen Residuum bezüglich der nächstfeineren Triangulierung ergibt. Um die Existenz und Eindeutigkeit der Lösung w_l auch im Falle $\widetilde{\lambda}_{1,l} = \lambda_{1,l}$ sichern zu können, sind Projektionen von d_l und w_l bezüglich des zu $\widetilde{\lambda}_{1,l}$ gehörenden (approximativen) Eigenvektors $\widetilde{u}_{1,l}$ gemäß

$$
\begin{aligned}
d_l \mapsto d_l^{\perp} &= d_l - \big(\widetilde{v}_{1,l}^{*} d_l\big) B_l \widetilde{u}_{1,l} \\
w_l \mapsto w_l^{\perp} &= w_l - \big(\widetilde{v}_{1,l}^{*} B_l w_l\big) \widetilde{u}_{1,l}
\end{aligned}
$$

erforderlich, wobei $\widetilde{v}_{1,l}$ eine Näherung des korrespondierenden linken Eigenvektors mit $\widetilde{v}_{1,l}^{*} B_l \widetilde{u}_{1,l} = 1$ ist. In unserem speziellen Fall sind die Matrizen A_l komplex symmetrisch und die Matrizen B_l reell symmetrisch, das heißt die linken Eigenvektoren sind die konjugiert komplexen rechten Eigenvektoren. Damit können wir hier $\widetilde{v}_{1,l}^{*}$ durch $\widetilde{u}_{1,l}^{T}$ ersetzen. Die in jedem Zyklus des Mehrgitter-Verfahrens benötigten Approximationen der Eigenpaare $(\lambda_{1,l}, u_{1,l})$ werden mit der „nested-iteration“-Technik (einer geschachtelten Iteration) bestimmt. Dabei löst man das Problem zunächst für die Grundtriangulierung t_{h_0}, danach für die folgende Triangulierung t_{h_1} usw., wobei als Startwerte stets die berechneten Lösungen vom letzten Gitter benutzt werden. Im Falle der Berechnung mehrerer Eigenwerte werden q Gleichungen der Form (39) gelöst. Die Bestimmung der Eigenwertapproximationen zu Beginn jedes Zyklus erfolgt dabei durch Lösung eines projizierten Eigenwertproblems (38), wobei $W = \widetilde{U} = (\widetilde{u}_1, \ldots, \widetilde{u}_q)$ mit den aktuellen Eigenvektornäherungen $\widetilde{u}_j$ gesetzt wird. Die entsprechend auftretenden q Grobgitterprobleme behandelt man simultan, das heißt die oben beschriebenen Projektionen werden für jedes $d_{j,l}$ und $w_{j,l}$ bezüglich des von den (approximativen) Eigenvektoren $\widetilde{u}_{1,l}, \ldots, \widetilde{u}_{q,l}$ aufgespannten Unterraumes durchgeführt. Eine Konvergenztheorie für diese Mehrgitter-Methode ist in der Arbeit [25] dargestellt.

Ein zweites Verfahren, bei dem das Mehrgitterkonzept direkt auf das zu lösende Eigenwertproblem angewandt wird, ist die Rayleigh-Quotienten-Mehrgitter-Minimierung von Mandel und McCormick [42] zur Berechnung eines einfachen Eigenwertes und des korrespondierenden Eigenvektors. Diese Methode ist zunächst nur im selbstadjungierten Fall anwendbar. Wesentliches Strukturmerkmal des Algorithmus ist, daß dieser nach Konstruktion eine monoton fallende Folge von Eigenwertnäherungen liefert, welche aufgrund der Beschränktheit des Rayleigh-Quotienten im selbstadjungierten Fall stets konvergent ist. Diese Eigenschaft verleiht dem Verfahren eine hohe numerische Stabilität. Das Anliegen dieser Arbeit ist die Verallgemeinerung der Methode auf den nichtselbstadjungierten Fall. Dazu werden wir zunächst im folgenden Abschnitt die Rayleigh-Quotienten-Mehrgitter-Minimierung in einer äquivalenten Formulierung beschreiben, wobei wir gleich die Erweiterung des Verfahrens zur Berechnung invarianter Unterräume entwickeln. Es stellt sich heraus, daß das Verfahren als Mehrgitter-Rekursion in Form von Eigenwertproblemen dargestellt werden kann. Diese Darstellung bildet die Grundlage der Entwicklung einer direkten Mehrgitter-Methode für den nichtselbstadjungierten Fall. Wir erhalten ein Mehrgitter-Verfahren, welches zur Lösung des allgemeinen Problems (33) eine Rekursion in Form von Schurproblemen realisiert.

Bemerkung 13. Die Darstellung des Verfahrens der Rayleigh-Quotienten-Minimierung als eine Iteration in Form von Eigenwertproblemen geht auf FADDEJEW und FADDEJEWA [19, S. 415] zurück. Die Formulierung der Mehrgitter-Minimierung als Rekursion von Eigenwertproblemen wird unter anderem auch von CHAN und SHARAPOV [10] im Zusammenhang mit Gebietszerlegungsmethoden zur Bestimmung einer einfachen Eigenlösung des selbstadjungierten Problems benutzt.

3.2 Selbstadjungierte Eigenwertprobleme

In diesem Abschnitt soll die Verallgemeinerung der Rayleigh-Quotienten-Mehrgitter-Minimierung von MANDEL und MCCORMICK [42] zur simultanen Berechnung mehrerer Eigenwerte und der korrespondierenden Eigenvektoren dargestellt werden. Im selbstadjungierten diskreten Fall können analog zu Lemma 2 (ii) die Eigenvektoren reell gewählt werden, so daß das zu lösende Problem die Form

$$\begin{aligned} AU &= BU\Lambda \\ U^T BU &= I \end{aligned} \tag{40}$$

annimmt, wobei alle Matrizen reellwertig sind. Die dünn besetzten $(N \times N)$-Matrizen A und B sind symmetrisch, die Matrix B ist positiv definit. Die Matrix Λ ist eine $(q \times q)$-Diagonalmatrix, bestehend aus den gesuchten q kleinsten Eigenwerten.

Wir beginnen mit der Herleitung der Methode der Rayleigh-Quotienten-Minimierung auf einem Gitter. Eine spezielle Variante dieses Algorithmus wird in dem weiter unten beschriebenen Mehrgitter-Verfahren die Rolle des Glätters übernehmen. Das allgemeine Prinzip dieses Verfahrens wollen wir zunächst an der Bestimmung des kleinsten (einfachen) Eigenwertes λ_1 und des korrespondierenden Eigenvektors u_1 des Matrixpaares (A, B) erläutern. Dabei legen wir, wie auch im weiteren dieses Abschnitts, den Euklidischen Vektorraum $\mathbb{R}^N$ zugrunde. Nach dem diskreten Courantschen Minimaxprinzip (32) ist der Eigenwert λ_1 durch

$$\lambda_1 = \min_{u \in \mathbb{R}^N \setminus \{0\}} R(u) = R(u_1)$$

mit dem Rayleigh-Quotienten

$$R(u) = \frac{u^T A u}{u^T B u}$$

charakterisiert. Die wesentliche Idee aller Rayleigh-Quotienten-Minimierungsverfahren zur Berechnung des untersten Eigenpaares (λ_1, u_1) besteht jetzt in folgendem: zu einer

Näherung $(\widetilde{\lambda}_1, \widetilde{u}_1)$ und einer vorgegebenen, von $\widetilde{u}_1$ linear unabhängigen Suchrichtung $\widetilde{p}$ wird über den Minimierungsansatz

$$\widehat{\lambda}_1 = \min_{\delta \in \mathbb{R}} R(\widetilde{u}_1 + \delta \widetilde{p}) = R(\widehat{u}_1) \tag{41}$$

eine neue Approximation $(\widehat{\lambda}_1, \widehat{u}_1)$ bestimmt. Nach Definition von $\widehat{\lambda}_1$ besteht dabei die Relation $\widehat{\lambda}_1 \leq \widetilde{\lambda}_1$. Aufgrund der Homogenität nullter Ordnung $R(\vartheta u) = R(u)$ des Rayleigh-Quotienten läßt sich (41) allgemeiner als zweidimensionale Minimierungsaufgabe gemäß

$$\widehat{\lambda}_1 = \min_{\binom{\varphi}{\psi} \in \mathbb{R}^2 \setminus \{0\}} R(\varphi \widetilde{u}_1 + \psi \widetilde{p}) = R(\widehat{u}_1) \tag{42}$$

schreiben. Aus $R(\varphi \widetilde{u}_1 + \psi \widetilde{p}) = (\xi^T \widetilde{A} \xi)/(\xi^T \widetilde{B} \xi)$ mit

$$\widetilde{A} = \begin{pmatrix} \widetilde{u}_1 & \widetilde{p} \end{pmatrix}^T A \begin{pmatrix} \widetilde{u}_1 & \widetilde{p} \end{pmatrix} \quad , \quad \widetilde{B} = \begin{pmatrix} \widetilde{u}_1 & \widetilde{p} \end{pmatrix}^T B \begin{pmatrix} \widetilde{u}_1 & \widetilde{p} \end{pmatrix} \quad \text{und} \quad \xi = \begin{pmatrix} \varphi \\ \psi \end{pmatrix}$$

sehen wir, daß die Bestimmung von $(\widehat{\lambda}_1, \widehat{u}_1)$ äquivalent ist zur Berechnung des kleineren Eigenwertes und zugehörigen Eigenvektors des projizierten (2×2)-Matrixpaares $(\widetilde{A}, \widetilde{B})$. Die Matrizen $\widetilde{A}$ und $\widetilde{B}$ sind nach Definition reell und symmetrisch, $\widetilde{B}$ ist aufgrund der linearen Unabhängigkeit von $\widetilde{u}_1$ und $\widetilde{p}$ positiv definit. Damit ist zur Bestimmung von $(\widehat{\lambda}_1, \widehat{u}_1)$ strukturell dasselbe wie das ursprüngliche Problem (40) zu lösen, wenn wir noch $\widehat{u}_1$ gemäß $\widehat{u}_1^T B \widehat{u}_1 = 1$ normieren.

Durch wiederholte Anwendung dieses Prinzips erhalten wir ein Iterationsverfahren, das eine monoton fallende Folge von Eigenwertnäherungen mit korrespondierenden Eigenvektorapproximationen liefert. Aus Gründen der Übersichtlichkeit werden wir im folgenden die Tilden und die Hüte über den iterierten Vektoren und den Index 1 unterdrücken. Die verschiedenen Varianten der Rayleigh-Quotienten-Minimierungsverfahren unterscheiden sich nun lediglich in der Wahl der Suchrichtung p in jedem Iterationsschritt. Bei der Koordinatenrelaxation von Faddejew und Faddejewa [19, S. 412 ff.] dienen die Einheitsvektoren e_k als Suchrichtungen, die im Laufe der Iteration zyklisch durchlaufen werden. Die Methode ist das Äquivalent zum Gauß-Seidel-Verfahren für die Lösung linearer Gleichungssysteme. Im Zusammenhang mit der Approximation mittels finiter Elemente läßt sich dieser Algorithmus auch als Minimierungsmethode im Funktionenraum $H_0^1(\Omega)$ deuten, wobei in jedem Schritt in Richtung einer Basisfunktion des zugrunde liegenden Finite-Elemente-Raumes minimiert wird. Eine weitere Variante der Rayleigh-Quotienten-Minimierung, eingeführt von Hestenes und Karush [29], beruht auf der Methode des steilsten Abstiegs. Hier wird der negative Gradient des Rayleigh-Quotienten im Punkt u als Suchrichtung benutzt, das heißt $p = -g(u)$, wobei der Gradient durch

$$g(u) = \nabla R(u) = \frac{2}{u^T B u} \big(Au - R(u) B u \big)$$

gegeben ist. Wegen $u^T g(u) = 0$ sind u und p für $u \neq u_j$ stets linear unabhängig. Auf eine dritte Methode, die Rayleigh-Quotienten-Minimierung mit konjugierten Gradienten, wollen wir im folgenden näher eingehen. Diese Variante geht ursprünglich auf BRADBURY und FLETCHER [6] zurück. Wir wollen hier jedoch der Darstellung in DÖHLER [15] folgen. Den Ausgangspunkt der Herleitung bildet der Ansatz $p = r + \chi s$ für die Suchrichtung. Der Vektor r ist dabei zunächst beliebig, der Vektor s wird jeweils aus der vorangegangenen Eigenvektorapproximation und der dazu korrespondierenden Suchrichtung durch Maximierung des Rayleigh-Quotienten analog zu Gleichung (42) gewonnen. Die aktuelle Eigenvektornäherung u und der Vektor s erfüllen somit die Relationen

$$\begin{pmatrix} u & s \end{pmatrix}^T A \begin{pmatrix} u & s \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} u & s \end{pmatrix}^T B \begin{pmatrix} u & s \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad ,$$

wobei wir λ echt kleiner als μ voraussetzen wollen. Zur Berechnung des reellen Parameters χ benutzen wir die Taylorentwicklung

$$R(u + \Delta u) = \underbrace{R(u) + g(u)^T \Delta u + \frac{1}{2} \Delta u^T H(u) \Delta u}_{= J(u, \Delta u)} + \ldots$$

des Rayleigh-Quotienten im Punkt u, wobei die Hessesche Matrix durch

$$H(u) = \frac{2}{u^T B u} \big(A - R(u) B - g(u)(Bu)^T - (Bu) g(u)^T \big)$$

bestimmt ist. Wir setzen $\Delta u = \delta p = \delta(r + \chi s)$, wobei δ beliebig sein soll, und bestimmen χ durch Minimierung des quadratischen Funktionals

$$J(u, \delta(r + \chi s)) = \lambda + \delta \big(g^T r + \chi g^T s \big) + \frac{\delta^2}{2} \big(r^T H r + 2 \chi s^T H r + \chi^2 s^T H s \big)$$

mit $g = g(u)$ und $H = H(u)$. Da $u^T A s = u^T B s = 0$ gilt, ergibt sich $g^T s = 0$. Notwendige Bedingung für die Existenz eines Minimums ist das Verschwinden der ersten Ableitung

$$\frac{\partial}{\partial \chi} J(u, \delta(r + \chi s)) = \delta^2 \big(s^T H r + \chi s^T H s \big) = \delta^2 s^T H \big(r + \chi s \big) = \delta^2 s^T H p \quad ,$$

also die H-Konjugiertheit der Vektoren s und p. Wegen $s^T H = 2 s^T (A - \lambda B)$ folgt daraus für χ die Bestimmungsgleichung

$$\chi = - \frac{s^T H r}{s^T H s} = - \frac{s^T (A r - \lambda B r)}{\mu - \lambda} \quad . \tag{43}$$

Für die zweite Ableitung erhalten wir

$$\frac{\partial^2}{\partial \chi^2} J(u, \delta(r + \chi s)) = \delta^2 s^T H s = 2\delta^2(\mu - \lambda) \quad ,$$

das heißt das so bestimmte χ ist wegen $\mu > \lambda$ für jede nicht verschwindende Korrektur δ tatsächlich ein Minimum. Wählen wir nun in jedem Iterationsschritt $r = -g$, das heißt $p = -g + \chi s$, wobei χ aus Gleichung (43) bestimmt wird, so erhalten wir das Verfahren der konjugierten Gradienten zur Minimierung des Rayleigh-Quotienten.

Bemerkung 14. Eine andere Wahl des Vektors r, die wir weiter unten im Zusammenhang mit adaptiven Mehrgitter-Verfahren benötigen werden, ergibt sich aus folgender Betrachtung: analog zur Koordinatenrelaxation von Faddejew und Faddejewa werden die einzelnen Komponenten r_k des Vektors r durch Minimierung des quadratischen Funktionals $J(u, \Delta u)$ in Richtung der Einheitsvektoren e_k bestimmt, das heißt wir setzen $\Delta u = r_k e_k$ und minimieren die Funktion

$$J(u, r_k e_k) = \lambda + r_k g_k + \frac{r_k^2}{2} H_{kk}$$

bezüglich r_k. Als notwendige Bedingung erhalten wir

$$r_k = -\frac{g_k}{H_{kk}} = -\frac{1}{2} \frac{g_k}{A_{kk} - \lambda B_{kk} - 2(Bu)_k g_k} \quad ,$$

wobei wir wieder die Normierung $u^T B u = 1$ verwenden. Setzen wir nun voraus, daß die Approximation u schon in einer geeigneten Umgebung des gesuchten Eigenvektors u_1 liegt und für u_1 die Relation $u_1 \neq e_k$ gilt, so erfüllt zum einen die korrespondierende Eigenwertnäherung λ die Ungleichung $A_{kk} - \lambda B_{kk} > 0$, zum anderen sind die Diagonalelemente H_{kk} von H alle positiv. Somit wird bei r_k tatsächlich ein Minimum angenommen, und für hinreichend kleines g_k haben wir die Entwicklung

$$r_k = -\frac{1}{2} \frac{g_k}{A_{kk} - \lambda B_{kk}} \sum_{m=0}^{\infty} \left(\frac{2(Bu)_k g_k}{A_{kk} - \lambda B_{kk}} \right)^m \quad .$$

Berücksichtigen wir hier nur den ersten Summanden, erhalten wir insgesamt für r die Darstellung

$$r = -\frac{1}{2} \big(\mathrm{diag}(A - \lambda B)\big)^{-1} g \quad . \tag{44}$$

Verwendet man allgemein anstelle von $r = -g$ den Vektor $r = -C^{-1} g$ in der Methode der konjugierten Gradienten, so nennt man das Verfahren vorkonditioniert und die Matrix C Vorkonditionierungsmatrix. Die eben beschriebene Methode heißt in Analogie zum Vorgehen bei linearen Gleichungssystemen Jacobi-Vorkonditionierung.

Weitere Varianten der Vorkonditionierung können durch Änderung des zugrundeliegenden Skalarproduktes gewonnen werden. Die Richtung des steilsten Abstiegs des Rayleigh-Quotienten im Punkt u ist beispielsweise bezüglich des diskreten L^2- bzw. Energie-Skalarproduktes durch $r = -B^{-1}g$ bzw. $r = -(A + \mu B)^{-1}g$ (mit einer geeigneten Spektralverschiebung μ) gegeben.

Wenden wir uns jetzt der Verallgemeinerung des Verfahrens zur Berechnung des zu den kleinsten Eigenwerten korrespondierenden invarianten Unterraumes, das heißt der simultanen Bestimmung der Eigenwerte $\lambda_1, \ldots, \lambda_q$ und der korrespondierenden Eigenvektoren $u_1, \ldots, u_q$, zu. Dabei wollen wir voraussetzen, daß die gewünschten Eigenwerte separiert vom restlichen Spektrum liegen, das heißt die Relation $\lambda_j < \lambda_k$ für $j \le q < k$ gilt. Die Grundlage für die Herleitung des Algorithmus bildet die adäquate Erweiterung des Courantschen Minimaxprinzips (32) auf ein allgemeineres Minimalprinzip [15] zur Charakterisierung des gesuchten invarianten Unterraumes, welches im folgenden kurz beschrieben werden soll. Die aus den Eigenvektoren gebildete Matrix $U_1 = \begin{pmatrix} u_1 & \cdots & u_q \end{pmatrix} \in \mathbb{R}^{N \times q}$ erfüllt die Minimaleigenschaft

$$F(U_1) = \min_{\substack{U \in \mathbb{R}^{N \times q} \\ \mathrm{rang}(U) = q}} F(U) = \lambda_1 + \ldots + \lambda_q \tag{45}$$

mit dem Funktional

$$F(U) = \mathrm{spur}\big(\Lambda(U)\big) \quad , \quad \Lambda(U) = (U^T B U)^{-1} (U^T A U) \quad ,$$

wobei $\mathbb{R}^{N \times q}$ den Raum aller reellen $(N \times q)$-Matrizen bezeichnet. Da ähnliche Matrizen dieselbe Spur besitzen, ergibt sich $F(U\Theta) = F(U)$ für jede invertierbare Matrix $\Theta \in \mathbb{R}^{q \times q}$, das heißt der Minimalwert des Funktionals wird von jeder $(N \times q)$-Matrix, deren Spaltenvektoren eine Basis des zu den kleinsten Eigenwerten gehörenden invarianten Unterraumes bilden, angenommen. Umgekehrt sehen wir unter Zuhilfenahme des Gradienten

$$G(U) = 2\big(AU - BU\Lambda(U)\big)(U^T B U)^{-1}$$

von F im Punkt U, daß jede $(N \times q)$-Matrix mit vollem Rang, für die der Minimalwert des Funktionals angenommen wird, zu einer Basis des gesuchten invarianten Unterraumes korrespondiert. Insgesamt erhalten wir also, daß die durch (45) gegebene Charakterisierung den invarianten Unterraum eindeutig bestimmt. Ist nun durch $\widetilde{U}_1 \in \mathbb{R}^{N \times q}$ eine Approximation der Matrix U_1 und durch $\widetilde{P} \in \mathbb{R}^{N \times n}$ eine Matrix, bestehend aus n Suchrichtungen, mit $\mathrm{rang}(\begin{pmatrix} \widetilde{U}_1 & \widetilde{P} \end{pmatrix}) = q + n$ gegeben, so wird analog zu (42) über den Minimierungsansatz

$$F(\widehat{U}_1) = \min_{\substack{\left(\begin{smallmatrix} \Phi \\ \Psi \end{smallmatrix}\right) \in \mathbb{R}^{(q+n) \times q} \\ \mathrm{rang}(\left(\begin{smallmatrix} \Phi \\ \Psi \end{smallmatrix}\right)) = q}} F(\widetilde{U}_1 \Phi + \widetilde{P} \Psi) = \widehat{\lambda}_1 + \ldots + \widehat{\lambda}_q \tag{46}$$

eine neue Näherung $\widehat{U}_1$ berechnet. Es ist $F(\widetilde{U}_1\Phi + \widetilde{P}\Psi) = \mathrm{spur}\big((\Xi^T\widetilde{B}\Xi)^{-1}(\Xi^T\widetilde{A}\Xi)\big)$ mit

$$\widetilde{A} = \begin{pmatrix}\widetilde{U}_1 & \widetilde{P}\end{pmatrix}^T A \begin{pmatrix}\widetilde{U}_1 & \widetilde{P}\end{pmatrix} \quad , \quad \widetilde{B} = \begin{pmatrix}\widetilde{U}_1 & \widetilde{P}\end{pmatrix}^T B \begin{pmatrix}\widetilde{U}_1 & \widetilde{P}\end{pmatrix} \quad \text{und} \quad \Xi = \begin{pmatrix}\Phi \\ \Psi\end{pmatrix} \quad ,$$

wobei $\widetilde{A}, \widetilde{B} \in \mathbb{R}^{(q+n)\times(q+n)}$ und $\Xi \in \mathbb{R}^{(q+n)\times q}$ gilt, so daß die Lösung der Minimierungsaufgabe (46) gleichbedeutend ist zur Bestimmung der q kleinsten Eigenwerte des projizierten Matrixpaares $(\widetilde{A}, \widetilde{B})$, das heißt zur Lösung des Eigenwertproblems

$$\begin{aligned} \widetilde{A}\Xi &= \widetilde{B}\Xi\widehat{\Lambda} \\ \Xi^T\widetilde{B}\Xi &= I \quad , \end{aligned} \tag{47}$$

wenn wir noch die Normierung $\widehat{U}_1^T B \widehat{U}_1 = I$ einführen. Die Matrizen $\widetilde{A}$ und $\widetilde{B}$ sind hierbei reell und symmetrisch, die Matrix $\widetilde{B}$ ist aufgrund des vollen Ranges von $\begin{pmatrix}\widetilde{U}_1 & \widetilde{P}\end{pmatrix}$ positiv definit.

Wie im Falle der einfachen Rayleigh-Quotienten-Minimierung erhalten wir durch sukzessive Wiederholung des dargelegten Schemas ein Iterationsverfahren, welches eine monoton fallende Folge von Funktionalwerten produziert. Diese iterative Methode zur approximativen Berechnung des zu den Eigenwerten $\lambda_1, \ldots, \lambda_q$ korrespondierenden invarianten Unterraumes wird simultanes Rayleigh-Quotienten-Minimierungsverfahren genannt. Es existieren auch hier verschiedene Varianten dieses Algorithmus. Bei der simultanen Koordinatenrelaxation von Schwarz [53] besteht die Suchrichtung P (wir lassen Tilden und Hüte und den Index 1 fortan wieder weg) aus einem n-spaltigen Block der $(N \times N)$-Einheitsmatrix. Wählen wir $P = -G(U)$ und damit $n = q$, so ergibt sich das simultane Gradienten-Verfahren, wie es beispielsweise von Longsine und McCormick in [40] beschrieben ist. Für die Darstellung der Methode der konjugierten Gradienten orientieren wir uns wieder an der Arbeit [15] von Döhler. Für die Suchrichtung $P \in \mathbb{R}^{N\times n}$ mit $n = q$ wird der Ansatz $P = R + SX$ mit einer noch zu spezifizierenden Matrix $R \in \mathbb{R}^{N\times q}$ und der Matrix $S \in \mathbb{R}^{N\times q}$ gemacht, die den Bedingungen

$$\begin{pmatrix}U & S\end{pmatrix}^T A \begin{pmatrix}U & S\end{pmatrix} = \begin{pmatrix}\Lambda & 0 \\ 0 & M\end{pmatrix} \quad \text{und} \quad \begin{pmatrix}U & S\end{pmatrix}^T B \begin{pmatrix}U & S\end{pmatrix} = \begin{pmatrix}I & 0 \\ 0 & I\end{pmatrix}$$

mit $\Lambda = \mathrm{diag}(\lambda_j)_{j=1,\ldots,q}$ und $M = \mathrm{diag}(\mu_k)_{k=1,\ldots,q}$ genügen soll. Diese Matrix S wird im Laufe der Iteration durch vollständige Lösung des projizierten Eigenwertproblems (47) aus den vorherigen Eigenvektorapproximationen und zugehörigen Suchrichtungen bestimmt, wobei wir $\lambda_j < \mu_k$ voraussetzen wollen. Durch Minimierung des quadratischen Taylorfunktionals von $F(U)$ läßt sich nun analog zur einfachen Rayleigh-Quotienten-Minimierung eine Bestimmungsgleichung für die Matrix $X \in \mathbb{R}^{q\times q}$ ableiten. Es stellt sich heraus, daß X als Lösung der Sylvester-Gleichung

$$MX - X\Lambda = -S^T(AR - BR\Lambda) \tag{48}$$

berechnet werden kann. Für die einzelnen Komponenten von X erhalten wir somit

$$\chi_{kj} = -\frac{s_k^T(Ar_j - \lambda_j Br_j)}{\mu_k - \lambda_j} \quad ,$$

wobei s_k bzw. r_j die k-te bzw. j-te Spalte von S bzw. R bezeichnet. Das Verfahren der konjugierten Gradienten ergibt sich jetzt durch die Wahl $R = -G = -G(U)$, das heißt $P = -G + SX$ mit X aus Gleichung (48).

Bemerkung 15. Ersetzen wir hier den Gradienten G durch $C^{-1}G$, so nennt man die Methode ähnlich wie in Bemerkung 14 vorkonditioniert. Die dort beschriebene Jacobi-Vorkonditionierung erfolgt im simultanen Fall durch Lösung des Gleichungssystems

$$2\big(\mathrm{diag}(A)R - \mathrm{diag}(B)R\Lambda\big) = -G \quad , \tag{49}$$

das heißt durch simultane Verwendung von q Gleichungen der Form (44).

Die bisherigen Untersuchungen dieses Abschnitts fassen wir in dem folgenden Algorithmus zusammen, dem vorkonditionierten Verfahren der konjugierten Gradienten zur simultanen Berechnung der kleinsten Eigenwerte und korrespondierenden Eigenvektoren auf einem Gitter. Der Übersichtlichkeit halber lassen wir die Iterationsindizes weg.

Algorithmus 2.

- Initialisierung:
  - ∗ Vorgabe von $U \in \mathbb{R}^{N\times q}$ mit $U^TBU = I$
  - ∗ setze $P = R = -C^{-1}\big(AU - BU\Lambda(U)\big) \in \mathbb{R}^{N\times q}$
- Iteration für $m = 1, \dots, \nu$:
  - ∗ setze $W = \begin{pmatrix} U & P \end{pmatrix}$
  - ∗ bestimme $Q \in \mathbb{R}^{2q\times 2q}$ und $\Lambda \in \mathbb{R}^{2q\times 2q}$ aus

$$\begin{aligned} \big(W^TAW\big)Q &= \big(W^TBW\big)Q\Lambda \\ Q^T\big(W^TBW\big)Q &= I \end{aligned} \tag{50}$$

    mit $\Lambda = \mathrm{diag}(\lambda_j)_{j=1,\dots,2q}$, wobei $\lambda_1 \leq \dots \leq \lambda_q < \lambda_{q+1} \leq \dots \leq \lambda_{2q}$
  - ∗ setze $\begin{pmatrix} U & S \end{pmatrix} = WQ$
  - ∗ berechne $R = -C^{-1}\big(AU - BU\Lambda(U)\big)$
  - ∗ bestimme $X \in \mathbb{R}^{q\times q}$ aus

$$\Lambda(S)X - X\Lambda(U) = -S^T\big(AR - BR\Lambda(U)\big)$$

∗ setze $P = R + SX$

Bemerkung 16. Da in jedem Schritt des Iterationsverfahrens ein projiziertes Eigenwertproblem der Form (50) gelöst wird, zählt der Algorithmus zu der in Abschnitt 2.6 dargestellten Klasse der orthogonalen Projektionsmethoden.

Nach Konstruktion des Verfahrens liefert dieses eine monoton fallende Folge von Funktionalwerten, die aufgrund der Beschränktheit des Funktionals nach unten stets konvergent ist. Genaue Bedingungen für die Konvergenz gegen den untersten Eigenwert λ_1 im Fall $q = 1$ sind in der Arbeit von Yang [59] für eine modifizierte Variante des Verfahrens angegeben. Die numerische Erfahrung zeigt, daß die Konvergenzraten im Fall $C = B$ abhängig sind vom Abstand der gewünschten Eigenwerte $\lambda_1, \ldots, \lambda_q$ zum restlichen Spektrum (jedoch nicht vom Abstand der gesuchten Eigenwerte untereinander) und von der Gitterweite der zugrundeliegenden Triangulierung des Gebietes Ω bei der Finite-Elemente-Approximation. Diese Beobachtung wollen wir im folgenden anhand der einfachen Rayleigh-Quotienten-Minimierung ($q = 1$) mittels konjugierter Gradienten näher untersuchen. Dabei setzen wir voraus, daß sämtliche Eigenwerte des Matrixpaares (A, B) der Größe nach, also gemäß $\lambda_1 < \lambda_2 \leq \ldots \leq \lambda_N$ angeordnet sind. Da sich das Verfahren asymptotisch wie die Methode der konjugierten Gradienten zur Minimierung quadratischer Funktionale verhält, wird die Konvergenz im wesentlichen durch den Quotienten

$$\rho = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

mit der modifizierten spektralen Konditionszahl

$$\kappa = \frac{\lambda_N - \lambda_1}{\lambda_2 - \lambda_1}$$

der Hesseschen Matrix $H(u_1) = 2(A - \lambda_1 B)$ des Rayleigh-Quotienten am Lösungspunkt u_1 bestimmt [47]. Daraus sehen wir zum einen, daß die Konvergenzrate für dicht benachbarte Eigenwerte ($\lambda_1 \approx \lambda_2$) nahe der Eins liegt ($\rho \approx 1$). Aus Theorem 1 wissen wir weiterhin, daß die Eigenwerte des kontinuierlichen Problems eine gegen unendlich strebende Folge bilden. Da die diskreten Eigenwerte obere Schranken für die kontinuierlichen Eigenwerte sind, erhalten wir somit andererseits mit feiner werdender Triangulierung des Gebietes Ω, das heißt für $N \to \infty$, die Relation $\kappa \to \infty$ und damit $\rho \to 1$. Eine von der Gitterweite und damit von der Anzahl N der Unbekannten unabhängige Konvergenzrate kann durch eine andere, geeignetere Vorkonditionierung erreicht werden. Wählen wir als Vorkonditionierungsmatrix $C = A + \mu B$, wobei die Spektralverschiebung μ der Bedingung $0 < \lambda_1 + \mu$ genügen soll, so ergibt sich für die Konditionszahl die in [54, S. 343] angegebene Relation

$$\kappa = \frac{1 - (\lambda_1 + \mu)/(\lambda_N + \mu)}{1 - (\lambda_1 + \mu)/(\lambda_2 + \mu)} \quad ,$$

das heißt im Grenzfall $N \to \infty$ die Beziehung

$$\kappa = \frac{1}{1-\zeta} \quad \text{mit} \quad 0 < \zeta = \frac{\lambda_1 + \mu}{\lambda_2 + \mu} < 1 \quad .$$

Damit erhalten wir als Konvergenzrate für das vorkonditionierte Verfahren

$$\rho_{\mathrm{PCG}} = \frac{(1-\zeta)^{-\frac{1}{2}} - 1}{(1-\zeta)^{-\frac{1}{2}} + 1} = \frac{1}{4}\zeta + \frac{1}{8}\zeta^2 + \ldots \quad ,$$

welche nur vom Quotienten ζ der verschobenen Eigenwerte $\lambda_1 + \mu$ und $\lambda_2 + \mu$ abhängig ist.

Bemerkung 17. In Abbildung 7 ist ρ_{PCG} in Abhängigkeit von ζ dargestellt, wobei wir zum Vergleich zusätzlich die Konvergenzrate der spektralverschobenen, inversen Vektoriteration

$$\rho_{\mathrm{IPM}} = \zeta$$

aufgenommen haben. Aus der Grafik ist insbesondere zu erkennen, daß mit dem vorkonditionierten Verfahren der konjugierten Gradienten bei dicht zusammenliegenden Eigenwerten ($\zeta \approx 1$) durch eine kleine Änderung der Spektralverschiebung eine deutlich verbesserte Konvergenzrate als mit der inversen Vektoriteration erzielt werden kann.

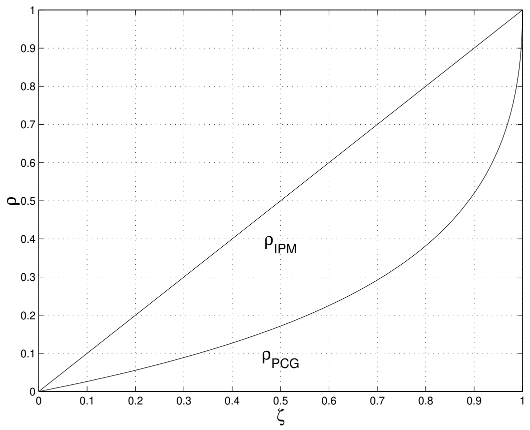

Abbildung 7: Konvergenzraten des vorkonditionierten Verfahrens der konjugierten Gradienten ρ_{PCG} und der spektralverschobenen, inversen Vektoriteration ρ_{IPM} in Abhängigkeit von ζ

Da die effiziente Berechnung möglichst guter Approximationen der kontinuierlichen Eigenlösungen die Verwendung einer Folge von adaptiv verfeinerten Triangulierungen erforderlich macht, wird bei Lösung der entsprechenden diskreten Eigenprobleme mit Algorithmus 2 stets eine vorkonditionierte Variante benutzt. Aufbauend auf die eben durchgeführte Analyse wird dabei meist eine approximative Lösung des linearen Gleichungssystems $(A+\mu B)R = -G$ bestimmt. Eine Untersuchung des vorkonditionierten Verfahrens der konjugierten Gradienten, bei dem zur näherungsweisen Berechnung von R eine lineare Mehrgittermethode angewandt wird, findet sich in der Arbeit von Jung et al. [32].
Alternativ zur eben beschriebenen Vorkonditionierung lassen sich von der Gitterweite unabhängige Konvergenzraten durch direkte Anwendung des Mehrgitterkonzepts auf das zu lösende Eigenwertproblem ohne explizite Verwendung einer Spektralverschiebung gewinnen. Dabei wird das Verfahren der konjugierten Gradienten, zunächst ohne die in Bemerkung 15 beschriebene Jacobi-Vorkonditionierung (das heißt $C = I$), als Glättungsverfahren eingesetzt. Im folgenden Beispiel wollen wir dazu zunächst anhand eines einfachen Modellproblems die Glättungseigenschaft von Algorithmus 2 aufzeigen.

Beispiel 9. Wir betrachten das Eigenwertproblem

$$-u''(x) = \lambda u(x)$$

auf $\Omega = (0,1)$ mit Dirichlet-Randbedingung $u(0) = u(1) = 0$. Die kontinuierlichen Lösungen des Problems sind durch $\lambda_j = \pi^2 j^2$ und $u_j(x) = \sqrt{2}\,\sin(\pi j x)$ mit $j = 1, 2, \ldots$ gegeben. Die Diskretisierung mit linearen finiten Elementen über einem äquidistanten Gitter mit der Gitterweite $h = 1/N$ führt auf die System- bzw. Massenmatrix

$$A = N \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \quad \text{bzw.} \quad B = \frac{1}{6N} \begin{pmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix} \quad ,$$

wobei $A, B \in \mathbb{R}^{(N-1)\times(N-1)}$. Daraus bestimmen sich die diskreten Eigenwerte zu

$$\lambda_j = 12\,N^2\,\frac{\sin^2(\pi j/(2N))}{2+\cos(\pi j/N)}, \quad j = 1, \ldots, N-1$$

und die Komponenten der korrespondierenden normierten Eigenvektoren zu

$$u_{lj} = \sqrt{\frac{6}{2+\cos(\pi j/N)}}\,\sin(\pi j l/N), \quad l = 1, \ldots, N-1$$

(den Diskretisierungsparameter h unterdrücken wir im folgenden wiederum). Zur Illustration der Glättungseigenschaft von Algorithmus 2 im Fall $q = 1$ (einfache Rayleigh-Quotienten-Minimierung) führen wir die Iteration nicht im Funktionenraum für den Vektor u, sondern im Koeffizientenraum für den Vektor $a = (a_j)_{j=1,\dots,N-1}$, der aus den Koeffizienten der Basisdarstellung

$$u = \sum_{j=1}^{N-1} a_j u_j$$

von u bezüglich der diskreten Eigenvektoren u_j gebildet wird, durch. Dabei ist zu beachten, daß das Iterationsverfahren ohne Vorkonditionierung im Funktionenraum in ein Verfahren mit Vorkonditionierung im Koeffizientenraum übergeht, wobei die Komponenten c_{kj} der Vorkonditionierungsmatrix C durch $c_{kj} = u_k^T u_j$ gegeben sind. In unserem speziellen Fall beispielsweise ergibt sich daraus

$$C^{-1} = \operatorname{diag}\Big(\frac{2+\cos(\pi j/N)}{3N}\Big)_{j=1,\dots,N-1} \quad .$$

Für die Iteration im Koeffizientenraum ersetzen wir nun in Algorithmus 2 die Systemmatrix A durch die aus den $N-1$ diskreten Eigenwerten bestehende Diagonalmatrix $D = \operatorname{diag}(\lambda_j)_{j=1,\dots,N-1}$ mit $\lambda_1 \le \lambda_2 \le \dots \le \lambda_{N-1}$, die Massenmatrix B durch die Identität, den iterierten Vektor durch den Koeffizientenvektor a und die Vorkonditionierungsmatrix entsprechend der oben gemachten Bemerkung. Die Komponenten des Startvektors initialisieren wir gemäß $a_j = 1/\sqrt{N-1}$ für $j = 1, \dots, N-1$. In Abbildung 8 sind für die Werte $N = 500$ und $N = 1000$ der Startvektor und die ersten vier Iterierten des Verfahrens im Koeffizientenraum dargestellt. $\diamond$

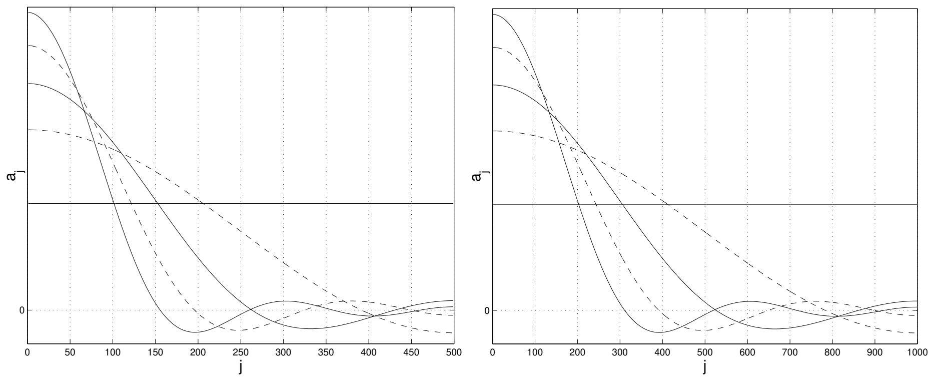

Abbildung 8: Rayleigh-Quotienten-Minimierung mit konjugierten Gradienten im Koeffizientenraum für $N = 500$ und $N = 1000$ (horizontale Linie: Startvektor)

Schon durch wenige Schritte des Iterationsverfahrens werden die Anteile in Richtung der Eigenvektoren zu großen Eigenwerten erheblich gedämpft. Wie aus Abbildung 8 ersichtlich, ist dabei das qualitative Glättungsverhalten von der Anzahl der Freiheitsgrade N unabhängig. Zur Behandlung der Anteile in Richtung der Eigenvektoren zu den kleineren Eigenwerten minimiert man das oben eingeführte Funktional $F(U)$ über den Ansatz (46) simultan in Richtung der Finite-Elemente-Basisfunktionen eines im aktuellen Feingitterraum enthaltenen Grobgitterraumes der Dimension n. Die Matrix $P \in \mathbb{R}^{N\times n}$ der Suchrichtungen besteht dabei spaltenweise aus der Koordinatendarstellung der Grobgitterbasisfunktionen in Bezug auf die Feingitterbasisfunktionen, ist also die im Mehrgitterkontext üblicherweise auftretende Prolongationsmatrix.

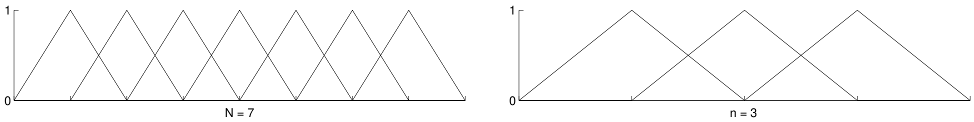

Abbildung 9: Feingitter- und Grobgitterbasisfunktionen über einem Intervall

Beispiel 10. Für die in Abbildung 9 gezeigte Feingitter-Grobgitter-Konstellation erhalten wir beispielsweise die Prolongationsmatrix

$$P = \begin{pmatrix} 1/2 & & \\ 1 & & \\ 1/2 & 1/2 & \\ & 1 & \\ & 1/2 & 1/2 \\ & & 1 \\ & & 1/2 \end{pmatrix} . \qquad \diamond$$

Wie wir weiter oben gesehen haben, ist die Lösung der Minimierungsaufgabe (46) äquivalent zur Berechnung der q kleinsten Eigenwerte des projizierten Matrixpaares (W^TAW, W^TBW) mit $W = \begin{pmatrix} U & P \end{pmatrix} \in \mathbb{R}^{N\times(q+n)}$, wobei U die aktuelle Iterierte bezeichnet. Das bedeutet, daß die Bestimmung der Grobgitter-Korrekturen zur Behandlung der niederfrequenten Anteile strukturell auf dasselbe Problem wie das ursprünglich zu lösende führt. Daher kann dieses bei Vorhandensein einer weiteren Gitterebene nun ebenfalls nach dem Prinzip der Glättung und Bestimmung einer Grobgitter-Korrektur näherungsweise gelöst werden. Aufgrund der speziellen Struktur der projizierten Matrizen wird dabei die Matrix

$$\begin{pmatrix} 0 \\ P' \end{pmatrix} \in \mathbb{R}^{(q+n)\times n'} \tag{51}$$

als Suchrichtungsmatrix gewählt, wobei P' die zum aktuellen und nächsten Grobgitter korrespondierende Prolongationsmatrix ist. Somit erfolgt die Berechnung der nächsten Grobgitter-Korrektur insgesamt auf der Grundlage des projizierten Matrixpaares $\big((WW')^T A(WW'), (WW')^T B(WW')\big)$ mit

$$W' = \begin{pmatrix} \Phi & 0 \\ \Psi & P' \end{pmatrix} \in \mathbb{R}^{(q+n)\times(q+n')} \quad .$$

Unter Beachtung von $WW' = \begin{pmatrix} U\Phi + P\Psi & PP' \end{pmatrix} \in \mathbb{R}^{N\times(q+n')}$ sehen wir zum einen, daß die aktuellen Korrekturen Φ und Ψ im nächsten Grobgitter-Korrekturschritt bereits berücksichtigt werden, und zum anderen, daß die Struktur der projizierten Matrizen erhalten bleibt. Bei Existenz noch weiterer Gitterebenen kann daher das eben beschriebene Vorgehen rekursiv fortgesetzt werden. Die algorithmische Umsetzung dieser Idee ist durch den folgenden Mehrgitter-Zyklus gegeben.

Algorithmus 3.
$[U_l, \Lambda_l] = \mathbf{MGM}(A_l, B_l, U_l, \Lambda_l, l)$

1. Vorglättung mit Algorithmus 2: $U_l \to \widetilde{U}_l$, $\Lambda_l \to \widetilde{\Lambda}_l$
2. Grobgitter-Korrektur: $\widetilde{U}_l \to \widehat{U}_l$, $\widetilde{\Lambda}_l \to \widehat{\Lambda}_l$
   - berechne $A_{l-1} = W_l^T A_l W_l$ und $B_{l-1} = W_l^T B_l W_l$, wobei für
     * $l = l_{\max}$: $W_l = \begin{pmatrix} \widetilde{U}_l & P_l \end{pmatrix}$
     * $l < l_{\max}$: $W_l = \begin{pmatrix} \widetilde{U}_l & \begin{matrix} 0 \\ P_l \end{matrix} \end{pmatrix} = \begin{pmatrix} \widetilde{\Phi}_l & 0 \\ \widetilde{\Psi}_l & P_l \end{pmatrix}$
   - falls
     * $l > 1$:
       - Initialisierung: $U_{l-1} = \begin{pmatrix} I \\ 0 \end{pmatrix}$, $\Lambda_{l-1} = \widetilde{\Lambda}_l$
       - Iteration für $m = 1, \ldots, \mu$:
         $[U_{l-1}, \Lambda_{l-1}] = \mathbf{MGM}(A_{l-1}, B_{l-1}, U_{l-1}, \Lambda_{l-1}, l-1)$
     * $l = 1$:
       - exakte Lösung von $A_0 U_0 = B_0 U_0 \Lambda_0$, $U_0^T B_0 U_0 = I$
   - setze $\widehat{U}_l = W_l U_{l-1}$, $\widehat{\Lambda}_l = \Lambda_{l-1}$
3. Nachglättung mit Algorithmus 2: $\widehat{U}_l \to U_l$, $\widehat{\Lambda}_l \to \Lambda_l$

Bemerkung 18. Die Matrix $\widehat{U}_l$ auf jeder Stufe der Mehrgitterrekursion ergibt sich für

* $l = l_{\max}$ zu $\widehat{U}_l = \begin{pmatrix} \widetilde{U}_l & P_l \end{pmatrix} \begin{pmatrix} \Phi_{l-1} \\ \Psi_{l-1} \end{pmatrix} = \widetilde{U}_l \Phi_{l-1} + P_l \Psi_{l-1}$ und

* $l < l_{\max}$ zu $\widehat{U}_l = \begin{pmatrix} \widetilde{\Phi}_l & 0 \\ \widetilde{\Psi}_l & P_l \end{pmatrix} \begin{pmatrix} \Phi_{l-1} \\ \Psi_{l-1} \end{pmatrix} = \begin{pmatrix} \widetilde{\Phi}_l \Phi_{l-1} \\ \widetilde{\Psi}_l \Phi_{l-1} + P_l \Psi_{l-1} \end{pmatrix} = \begin{pmatrix} \widehat{\Phi}_l \\ \widehat{\Psi}_l \end{pmatrix}$.

In Algorithmus 3 bezeichnet A_l bzw. B_l (mit dem Gitterparameter $0 \leq l \leq l_{\max}$) die (projizierte) System- bzw. Massenmatrix, welche zum Gitter mit der Gitterweite h_l korrespondiert. Die q-spaltige Matrix U_l ist jeweils die gesuchte Koeffizientenmatrix, die Matrix Λ_l eine Diagonalmatrix mit den gesuchten Eigenwerten $\lambda_1, \dots, \lambda_q$ in der Diagonalen, und die Matrix P_l die Prolongationsmatrix zwischen den Gittern der Gitterweite h_l und h_{l-1}. Durch vollständige Induktion erhalten wir für die projizierten Matrizen die Darstellungen

$$A_{l-1} = \begin{pmatrix} V_l & Q_l \end{pmatrix}^T A \begin{pmatrix} V_l & Q_l \end{pmatrix} \quad \text{und} \quad B_{l-1} = \begin{pmatrix} V_l & Q_l \end{pmatrix}^T B \begin{pmatrix} V_l & Q_l \end{pmatrix}$$

mit

* $V_l = \widetilde{U}_l$ und $Q_l = P_l$ für $l = l_{\max}$,

* $V_l = V_{l+1} \widetilde{\Phi}_l + Q_{l+1} \widetilde{\Psi}_l$ und $Q_l = Q_{l+1} P_l$ für $l < l_{\max}$.

Auf jeder Stufe des Mehrgitter-Zyklus wird daher ein projiziertes Eigenwertproblem der Form (47) durch rekursive Benutzung von Algorithmus 2 näherungsweise ($l > 1$) bzw. exakt ($l = 1$) gelöst, wobei sämtliche bereits bestimmten Korrekturen berücksichtigt werden. Da in jedem Schritt des Glättungsverfahrens wiederum die Lösung eines projizierten Eigenwertproblems bestimmt wird, ergibt sich insgesamt beim Durchlaufen des Mehrgitter-Zyklus aufgrund der Äquivalenz des projizierten Eigenproblems (47) und des Minimierungsproblems (46) eine monoton fallende Folge von Funktionalwerten. Im Fall $q = 1$ fällt die Rekursion mit der monotonen Rayleigh-Quotienten-Mehrgitter-Minimierung von Mandel und McCormick [42] zusammen. Abschließend wollen wir noch die vollständige Mehrgitter-Prozedur angeben.

Algorithmus 4.

- Initialisierung:
  * Vorgabe von $U \in \mathbb{R}^{N \times q}$ mit $U^T B U = I$, $U^T A U = \Lambda$
- Iteration:
  * falls $\|AU - BU\Lambda\|_{\mathrm{diag}(B)^{-1}} >$ tol: $[U, \Lambda] = \mathbf{MGM}(A, B, U, \Lambda, l_{\max})$

Bemerkung 19. Das Abbruchkriterium ergibt sich hierbei durch Übergang vom Eigenwertproblem (40) zum speziellen Eigenwertproblem $CX = X\Lambda$, $X^TX = I$ mit $C = R^{-T}AR^{-1}$ und $X = RU$, wobei R aus der Cholesky-Zerlegung $B = R^TR$ bestimmt wird, über die Gleichungskette

$$\begin{aligned}\|CX - X\Lambda\| &= \sqrt{\lambda_{\max}\big((CX - X\Lambda)^T(CX - X\Lambda)\big)} \\ &= \sqrt{\lambda_{\max}\big((AU - BU\Lambda)^T R^{-1}R^{-T}(AU - BU\Lambda)\big)} \\ &= \|AU - BU\Lambda\|_{B^{-1}} \quad .\end{aligned}$$

Die Bestimmung der Lösung der so auftretenden linearen Gleichungssysteme mit der Koeffizientenmatrix B wird nun in Algorithmus 4 nur näherungsweise durch einen Schritt des linearen Jacobi-Verfahrens zum Startwert 0 durchgeführt, das heißt $\|\cdot\|_{B^{-1}}$ durch $\|\cdot\|_{\mathrm{diag}(B)^{-1}}$ ersetzt.

Die numerische Erfahrung zeigt, daß die Konvergenzraten des Verfahrens unabhängig von der Gitterweite sind, das heißt die benötigte Anzahl an Mehrgitter-Zyklen zum Unterschreiten einer gewünschten Genauigkeit tol ist unabhängig von der Verfeinerungstiefe $l_{\max}$. Da jeder Schritt des Glättungsverfahrens (Algorithmus 2) nur einen Aufwand an Operationen proportional zur Anzahl der Freiheitsgrade benötigt und die Berechnung der projizierten Matrizen A_{l-1}, B_{l-1} auf rekursivem Wege erfolgt, ist der benötigte Aufwand pro Zyklus insgesamt proportional zum Parameter N. Zusammen mit der zuvor erwähnten Unabhängigkeit der Iterationszahl von der Gitterweite erhalten wir somit die optimale Komplexität der Mehrgitter-Methode.

Bemerkung 20. Insbesondere im Zusammenhang mit stark nichtuniformen (adaptiven) Gittern und großen Sprüngen in den Funktionswerten der in der Helmholtzgleichung auftretenden Funktion f beobachtet man bei numerischen Experimenten im Vergleich zur Glättung ohne Vorkonditionierung eine wesentliche Verbesserung der Konvergenzraten, falls der Glättungsalgorithmus mit der in Bemerkung 15 angegebenen simultanen Jacobi-Vorkonditionierung versehen wird. In Übereinstimmung mit (51) benutzen wir dabei Gleichung (49) im Fall $l < l_{\max}$ nur zur Berechnung der Zeilen von R mit Index größer als q, während die ersten q Zeilen von R gleich Null gesetzt werden.

Eine theoretische Untersuchung des Konvergenzverhaltens der Mehrgitter-Methode im Fall $q = 1$ (einfache Rayleigh-Quotienten-Mehrgitter-Minimierung) findet sich für den Zweigitter-Algorithmus in dem Artikel von McCormick [43], für den Mehrgitter-Algorithmus in der Arbeit von Cai et al. [9] (wobei allerdings als Glättungsverfahren eine linearisierte Variante von Algorithmus 2 zugrunde gelegt wird).

Bemerkung 21. Im Gegensatz zu der in Abschnitt 3.1 dargestellten Mehrgitter-Methode von Hackbusch sind in Algorithmus 3 in jedem Glättungsschritt und bei jeder Grobgitter-Korrekturbestimmung die Eigenwertnäherungen ebenfalls Unbekannte, die automatisch beim Durchlaufen des Zyklus berechnet werden. Außerdem sind in der durch Algorithmus 3 beschriebenen Mehrgitter-Rekursion keine Projektionen bezüglich der (approximativen) Eigenvektoren gröberer Gitter notwendig. Der wesentliche strukturelle Unterschied beider Verfahren liegt jedoch in der bereits oben bemerkten Tatsache, daß durch Algorithmus 3 eine monoton fallende Folge von Funktionalwerten erzeugt wird, die aufgrund der Beschränktheit des Funktionals nach unten stets konvergent ist. Diese Eigenschaft verleiht dem Mehrgitter-Minimierungsverfahren eine wesentlich höhere Robustheit. Ein numerischer Vergleich zwischen dem Verfahren von Hackbusch und der Rayleigh-Quotienten-Mehrgitter-Minimierung von Mandel und McCormick im Zusammenhang mit adaptiv erzeugten Triangulierungen ist in dem Artikel [13] anhand eines einfachen, eindimensionalen Testbeispiels dokumentiert. Bei der dort angegebenen Testrechnung bestätigte sich zum einen die erwartete höhere Robustheit der Rayleigh-Quotienten-Minimierung gegenüber Hackbuschs Verfahren, zum anderen ergaben sich für beide Methoden asymptotisch gleiche Konvergenzraten. Diese zweite Beobachtung läßt sich dadurch erklären, daß das Verfahren von Hackbusch als Linearisierung der Rayleigh-Quotienten-Mehrgitter-Minimierung gedeutet werden kann [42].

3.3 Nichtselbstadjungierte Eigenwertprobleme

Wir kommen jetzt zur Erweiterung der im vorigen Abschnitt dargestellten Mehrgitter-Methode auf den nichtselbstadjungierten Fall (unter Zugrundelegung des unitären Vektorraumes $\mathbb{C}^N$). Das zu lösende Problem ist nunmehr von der Gestalt

$$\begin{aligned} AU &= BUT \\ U^*BU &= I \quad , \end{aligned} \tag{52}$$

wobei wir die $(N \times N)$-Matrix A als nichtselbstadjungiert, die $(N \times N)$-Matrix B als selbstadjungiert und positiv definit voraussetzen wollen. Die $(q \times q)$-Matrix T ist eine obere Dreiecksmatrix mit den gesuchten q Eigenwerten mit kleinsten Realteilen in der Diagonalen, die $(N \times q)$-Matrix U besteht aus den dazu korrespondierenden Schurvektoren. Strukturell haben wir also dasselbe Problem wie im selbstadjungierten Fall (40), nämlich die Berechnung einer Basis des zu bestimmten Eigenwerten korrespondierenden invarianten Unterraumes, zu lösen (wobei wir wegen Bemerkung 12 auf Seite 31 als gesuchte Basis die aus Schurvektoren bestehende wählen). Im Unterschied zum selbstadjungierten Problem ist diese Basis jedoch nicht durch ein Minimalprinzip charakterisiert. Aus Abschnitt 2.3 wissen wir allerdings andererseits, daß alle weiteren Eigenschaften des selbstadjungierten Problems (wie zum Beispiel diskretes Spektrum, Existenz einer Eigenbasis) im nichtselbstadjungierten Fall erhalten bleiben. Da zwar die Charakterisierung durch die Minimaleigenschaft zur Herleitung der Mehrgitter-Methode für den selbstadjungierten Fall benutzt wurde, jedoch das Minimierungsproblem nicht mehr explizit in den Algorithmen des letzten Abschnittes auftaucht, erwarten wir daher insgesamt, daß die prinzipielle Struktur der Mehrgitter-Methode im nichtselbstadjungierten Fall beibehalten werden kann. Die Strategie zur Verallgemeinerung der Algorithmen 2 und 3 auf den nichtselbstadjungierten Fall besteht nun konkret darin, die jeweils auftretenden projizierten Eigenwertprobleme der Form (47) durch entsprechende projizierte Schurprobleme der Gestalt

$$\begin{aligned} \widetilde{A}\Xi &= \widetilde{B}\Xi\widehat{T} \\ \Xi^*\widetilde{B}\Xi &= I \end{aligned} \tag{53}$$

mit

$$\widetilde{A} = \begin{pmatrix} U & P \end{pmatrix}^* A \begin{pmatrix} U & P \end{pmatrix} \quad \text{und} \quad \widetilde{B} = \begin{pmatrix} U & P \end{pmatrix}^* B \begin{pmatrix} U & P \end{pmatrix}$$

zu ersetzen, wobei $U \in \mathbb{C}^{N\times q}$ die Matrix der aktuellen Schurvektornäherungen und $P \in \mathbb{C}^{N\times n}$ eine Matrix, gebildet aus n Suchrichtungen, mit $\operatorname{rang}(\begin{pmatrix} U & P \end{pmatrix}) = q + n$ ist. Dabei wählen wir $\Xi \in \mathbb{C}^{(q+n)\times q}$ und die obere Dreiecksmatrix $\widehat{T} \in \mathbb{C}^{q\times q}$ so, daß in der Diagonalen von $\widehat{T}$ die q Eigenwerte von $(\widetilde{A}, \widetilde{B})$ mit kleinsten Realteilen stehen.
Wir beginnen mit der Verallgemeinerung der Glättungsroutine (Algorithmus 2). In Analogie zum Verfahren der konjugierten Gradienten wird hierbei für die Suchrichtungsmatrix $P \in \mathbb{C}^{N\times q}$ der Ansatz $P = R + SX$ gemacht. Für die Matrix $U \in \mathbb{C}^{N\times q}$ der approximativen Schurvektoren und die Matrix $S \in \mathbb{C}^{N\times q}$ werden die Relationen

$$\begin{pmatrix} U & S \end{pmatrix}^* A \begin{pmatrix} U & S \end{pmatrix} = \begin{pmatrix} T_1 & T_{12} \\ 0 & T_2 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} U & S \end{pmatrix}^* B \begin{pmatrix} U & S \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

mit den oberen Dreiecksmatrizen T_1 und T_2 und $\Re\tau_{1,jj} < \Re\tau_{2,kk}$ vorausgesetzt. Die Berechnung von S erfolgt dabei durch vollständige Lösung des projizierten Schurproblems (53) aus den vorherigen Schurvektornäherungen und den dazugehörigen Suchrichtungen. Die Matrix $R \in \mathbb{C}^{N\times q}$ ist zunächst beliebig, die Bestimmung der Matrix $X \in \mathbb{C}^{q\times q}$ erfolgt in Verallgemeinerung des im letzten Abschnitt für den selbstadjungierten Fall dargestellten Vorgehens (der Minimierung eines quadratischen Funktionals) mit dem Ritz-Galerkin-Verfahren über das lineare Gleichungssystem

$$S^*A(U + \delta P) = S^*B(U + \delta P)T_1, \quad \delta \in \mathbb{C},$$

welches aufgrund der obigen Voraussetzungen äquivalent in die Sylvester-Gleichung

$$T_2X - XT_1 = -S^*(AR - BRT_1) = Y$$

umgeformt werden kann. Da nun die Matrizen T_1 und T_2 obere Dreiecksgestalt haben, ergibt sich daraus für die Komponenten von X die leicht auswertbare Formel

$$\chi_{kj} = \frac{\upsilon_{kj} - \sum_{l=k+1}^{q} \tau_{2,kl}\, \chi_{lj} + \sum_{l=1}^{j-1} \chi_{kl}\, \tau_{1,lj}}{\tau_{2,kk} - \tau_{1,jj}} \quad ,$$

wobei v_{kj} die entsprechende Komponente von Y bezeichnet (der Index j durchläuft dabei in einer äußeren Schleife die Werte $j = 1, 2, \ldots, q$, der Index k in einer inneren Schleife die Werte $k = q, q-1, \ldots, 1$). Die volle Erweiterung von Algorithmus 2 zur Behandlung nichtselbstadjungierter Eigenwertprobleme erhalten wir schließlich, indem wir $R = -C^{-1}\big(AU - BUT(U)\big)$ mit $T(U) = U^*AU = T_1$ setzen (die Matrix C ist wiederum eine Vorkonditionierungsmatrix, auf deren Wahl wir weiter unten eingehen werden). Zusammengefaßt ergibt sich so der folgende

Algorithmus 5.

- Initialisierung:
  - ∗ Vorgabe von $U \in \mathbb{C}^{N\times q}$ mit $U^*BU = I$
  - ∗ setze $P = R = -C^{-1}\big(AU - BUT(U)\big) \in \mathbb{C}^{N\times q}$
- Iteration für $m = 1, \ldots, \nu$:
  - ∗ setze $W = \begin{pmatrix} U & P \end{pmatrix}$
  - ∗ bestimme $Q \in \mathbb{C}^{2q\times 2q}$ und $T \in \mathbb{C}^{2q\times 2q}$ aus

$$\begin{aligned} \big(W^*AW\big)Q &= \big(W^*BW\big)QT \\ Q^*\big(W^*BW\big)Q &= I \quad , \end{aligned} \tag{54}$$

    wobei T obere Dreiecksmatrix mit Diagonalelementen λ_j, sortiert nach aufsteigenden Realteilen gemäß $\Re\lambda_1 \leq \ldots \leq \Re\lambda_q < \Re\lambda_{q+1} \leq \ldots \leq \Re\lambda_{2q}$
  - ∗ setze $\begin{pmatrix} U & S \end{pmatrix} = WQ$
  - ∗ berechne $R = -C^{-1}\big(AU - BUT(U)\big)$
  - ∗ bestimme $X \in \mathbb{C}^{q\times q}$ aus

$$T(S)X - XT(U) = -S^*\big(AR - BRT(U)\big)$$

  - ∗ setze $P = R + SX$

Bemerkung 22. In jedem Schritt des Iterationsverfahrens sind bei der Lösung des projizierten Schurproblems (54) die Diagonalelemente einer oberen Dreiecksmatrix nach der Größe ihrer Realteile zu sortieren. Da sich die Anzahl q der gewünschten Eigenwerte in den von uns betrachteten Fällen stets in der Größenordnung $q = 1, \ldots, 10$ bewegt, benutzen wir dazu in Algorithmus 5 als Sortierroutine den einfachen „bubblesort"-Algorithmus [2, S. 176]. Nach einer Idee von RUHE [46] erfolgt das dabei eventuell notwendige paarweise Vertauschen benachbarter Diagonalelemente durch komplexe

Givens-Rotationen. Das wesentliche Prinzip wollen wir hier nur am Tausch der Diagonalelemente einer oberen (2×2)-Dreiecksmatrix darstellen. Es ist eine unitäre Matrix Q, das heißt eine Matrix Q mit $Q^*Q = I$ gesucht, so daß

$$Q^* \begin{pmatrix} \lambda_1 & \tau_{12} \\ 0 & \lambda_2 \end{pmatrix} Q = \begin{pmatrix} \lambda_2 & \tau'_{12} \\ 0 & \lambda_1 \end{pmatrix}$$

gilt. Aus dem Ansatz

$$Q = \begin{pmatrix} \alpha & \overline{\beta} \\ -\beta & \overline{\alpha} \end{pmatrix}$$

lassen sich nun leicht die Komponenten von Q bestimmen, wobei zur Vermeidung von Exponentenüberlauf zwischen den Fällen

(i) $|\lambda_1 - \lambda_2| \leq |\tau_{12}|$: $\alpha = \dfrac{1}{\sqrt{1 + |\gamma|^2}}$, $\beta = \gamma\alpha$ mit $\gamma = \dfrac{\lambda_1 - \lambda_2}{\tau_{12}}$ und

(ii) $|\lambda_1 - \lambda_2| > |\tau_{12}|$: $\beta = \dfrac{1}{\sqrt{1 + |\gamma|^2}}$, $\alpha = \gamma\beta$ mit $\gamma = \dfrac{\tau_{12}}{\lambda_1 - \lambda_2}$

zu unterscheiden ist.

Das durch Algorithmus 5 beschriebene Verfahren zählt zu der Klasse der orthogonalen Projektionsmethoden und kann als spezielle Variante eines (vorkonditionierten) Einschritt-Block-Arnoldi-Verfahrens interpretiert werden (vergleiche dazu auch Algorithmus 1 auf Seite 33). Numerische Experimente zeigen, daß schon durch wenige Schritte des Iterationsverfahrens mit $C = I$ die hochfrequenten Anteile des iterierten Vektors stark gedämpft werden, der Algorithmus also über eine Glättungseigenschaft verfügt. Zur Behandlung der niederfrequenten Anteile greifen wir jetzt wie in Abschnitt 3.2 auf das direkte Mehrgitterkonzept zurück. Wir setzen $W = \begin{pmatrix} U & P \end{pmatrix}$, wobei die Matrix U die aktuellen Schurvektorapproximationen enthält und die Matrix P die Prolongationsmatrix zwischen dem aktuellen Feingitter und dem nächstgröberen Gitter ist, und lösen zur Bestimmung der Grobgitter-Korrektur das Schurproblem zum projizierten Matrixpaar (W^*AW, W^*BW), das heißt berechnen den zu den q Eigenwerten mit kleinsten Realteilen korrespondierenden invarianten Unterraum von (W^*AW, W^*BW). Da nun dieses Grobgitter-Korrekturproblem wiederum von der gleichen Bauart wie das ursprünglich zu lösende Problem ist, kann wie im selbstadjungierten Fall unter Beachtung der speziellen Struktur der projizierten Matrizen eine Mehrgitter-Rekursion, hier in Form von Schurproblemen fallender Dimension, aufgebaut werden. Diese ist im folgenden Algorithmus realisiert.

Algorithmus 6.
$[U_l, T_l] = \mathbf{MGM}(A_l, B_l, U_l, T_l, l)$

1. Vorglättung mit Algorithmus 5: $U_l \to \widetilde{U}_l$, $T_l \to \widetilde{T}_l$
2. Grobgitter-Korrektur: $\widetilde{U}_l \to \widehat{U}_l$, $\widetilde{T}_l \to \widehat{T}_l$
   - berechne $A_{l-1} = W_l^* A_l W_l$ und $B_{l-1} = W_l^* B_l W_l$, wobei für
     * $l = l_{\max}$: $W_l = \begin{pmatrix} \widetilde{U}_l & P_l \end{pmatrix}$
     * $l < l_{\max}$: $W_l = \begin{pmatrix} \widetilde{U}_l & \begin{matrix} 0 \\ P_l \end{matrix} \end{pmatrix}$
   - falls
     * $l > 1$:
       - Initialisierung: $U_{l-1} = \begin{pmatrix} I \\ 0 \end{pmatrix}$, $T_{l-1} = \widetilde{T}_l$
       - Iteration für $m = 1, \dots, \mu$:
         $[U_{l-1}, T_{l-1}] = \mathbf{MGM}(A_{l-1}, B_{l-1}, U_{l-1}, T_{l-1}, l-1)$
     * $l = 1$:
       - exakte Lösung von $A_0 U_0 = B_0 U_0 T_0$, $U_0^* B_0 U_0 = I$
   - setze $\widehat{U}_l = W_l U_{l-1}$, $\widehat{T}_l = T_{l-1}$
3. Nachglättung mit Algorithmus 5: $\widehat{U}_l \to U_l$, $\widehat{T}_l \to T_l$

Für eine Erläuterung der in Algorithmus 6 vorkommenden Matrizen verweisen wir auf die Ausführungen zu Algorithmus 3, wobei die dort auftretenden Diagonalmatrizen Λ_l durch die oberen Dreiecksmatrizen $T_l \in \mathbb{C}^{q \times q}$ (mit den Approximationen der q Eigenwerte mit kleinsten Realteilen in der Diagonalen) zu ersetzen sind. Die projizierten System- bzw. Massenmatrizen sind hier von der Form

$$A_{l-1} = \begin{pmatrix} V_l & Q_l \end{pmatrix}^* A \begin{pmatrix} V_l & Q_l \end{pmatrix} \quad \text{bzw.} \quad B_{l-1} = \begin{pmatrix} V_l & Q_l \end{pmatrix}^* B \begin{pmatrix} V_l & Q_l \end{pmatrix} \quad ,$$

das heißt auf jeder Stufe des durch Algorithmus 6 beschriebenen Mehrgitter-Zyklus wird unter Berücksichtigung aller bereits bestimmten Korrekturen ein projiziertes Schurproblem der Form (53) (näherungsweise) gelöst. Die vollständige Mehrgitter-Prozedur für den nichtselbstadjungierten Fall ist schließlich gegeben durch

Algorithmus 7.

- Initialisierung:
  * Vorgabe von $U \in \mathbb{C}^{N \times q}$ mit $U^* B U = I$, $U^* A U = T$
- Iteration:
  * falls $\|AU - BUT\|_{\mathrm{diag}(B)^{-1}} > \text{tol}$: $[U, T] = \mathbf{MGM}(A, B, U, T, l_{\max})$

Bemerkung 23. Das Abbruchkriterium in Algorithmus 7 ergibt sich dabei analog wie in Bemerkung 19 für den selbstadjungierten Fall dargestellt. Eine Untersuchung zur Güte der Eigenwertnäherungen und des (durch die Spaltenvektoren von U aufgespannten) korrespondierenden approximativen invarianten Unterraumes in Abhängigkeit von der Größe des Residuums $AU - BUT$ ist in dem Artikel von KAHAN et al. [33] enthalten. Dort ist gezeigt, daß die näherungsweisen Eigenlösungen U und T des Matrixpaares (A, B) als exakte Eigenlösungen eines gestörten Matrixpaares angesehen werden können, wobei die Größe der Störung durch die Norm des Residuums gegeben ist.

Wie wir an den in Kapitel 4 durchgerechneten Beispielen sehen werden, ist die Konvergenzrate des Verfahrens und damit die zum Erreichen einer geforderten Genauigkeit tol benötigte Anzahl an Mehrgitter-Zyklen unabhängig von der Gitterweite der zugrundeliegenden Triangulierung. Da weiterhin der Rechenaufwand pro Zyklus proportional zum Parameter N ist, erhalten wir so auch im nichtselbstadjungierten Fall die (allerdings nur experimentell belegte) optimale Komplexität der Mehrgitter-Methode.

Bemerkung 24. Ähnlich wie im selbstadjungierten Fall beobachten wir eine deutliche Verbesserung der Konvergenzraten des Verfahrens, falls der Glättungsalgorithmus 5 mit einer geeigneten Vorkonditionierung ausgestattet wird. In Verallgemeinerung der in Bemerkung 15 angegebenen Jacobi-Vorkonditionierung wird dabei die Berechnung von R auf der Grundlage des Gleichungssystems

$$\operatorname{diag}(A)R - \operatorname{diag}(B)RT = -(AU - BUT) \tag{55}$$

durchgeführt. Wegen der speziellen Struktur der projizierten Matrizen verwenden wir Gleichung (55) im Fall $l < l_{\max}$ nur zur Bestimmung der Komponenten von R mit Zeilenindex größer als q, und setzen die ersten q Zeilen von R gleich Null.

Aus der Struktur der im Mehrgitter-Algorithmus 6 auftretenden Matrizen A_{l-1} und B_{l-1} und aus der bereits oben bemerkten Zugehörigkeit des Glättungsalgorithmus 5 zur Klasse der orthogonalen Projektionsmethoden erkennen wir, daß das Mehrgitter-Verfahren ausschließlich auf der Lösung projizierter Probleme der Form (34) basiert, also insgesamt zu den in Abschnitt 2.6 beschriebenen orthogonalen Projektionsmethoden zählt. Analog zu den dort dargestellten Krylovraum-Methoden läßt sich nun auch ein Mehrgitter-Verfahren konstruieren, welches auf „schiefen" Projektionen, das heißt der sukzessiven Lösung projizierter Probleme der Gestalt (37) beruht. Der Vollständigkeit halber wollen wir für dieses Verfahren noch die Mehrgitter-Rekursion angeben, ohne allerdings weiter auf Details, wie zum Beispiel die Wahl eines geeigneten Glätters, einzugehen.

Algorithmus 8.
$[U_l, V_l, D_l] = \mathbf{MGM}(A_l, B_l, U_l, V_l, D_l, l)$

1. Vorglättung: $U_l \to \widetilde{U}_l$, $V_l \to \widetilde{V}_l$, $D_l \to \widetilde{D}_l$

2. Grobgitter-Korrektur: $\widetilde{U}_l \to \widehat{U}_l$, $\widetilde{V}_l \to \widehat{V}_l$, $\widetilde{D}_l \to \widehat{D}_l$
   - berechne $A_{l-1} = \widehat{W}_l^* A_l W_l$ und $B_{l-1} = \widehat{W}_l^* B_l W_l$, wobei für
     * $l = l_{\max}$: $W_l = \begin{pmatrix} \widetilde{U}_l & P_l \end{pmatrix}$, $\widehat{W}_l = \begin{pmatrix} \widetilde{V}_l & Q_l \end{pmatrix}$
     * $l < l_{\max}$: $W_l = \begin{pmatrix} \widetilde{U}_l & 0 \\ & P_l \end{pmatrix}$, $\widehat{W}_l = \begin{pmatrix} \widetilde{V}_l & 0 \\ & Q_l \end{pmatrix}$
   - falls
     * $l > 1$:
       ◦ Initialisierung: $U_{l-1} = \begin{pmatrix} I \\ 0 \end{pmatrix}$, $V_{l-1} = \begin{pmatrix} I \\ 0 \end{pmatrix}$, $D_{l-1} = \widetilde{D}_l$
       ◦ Iteration für $m = 1, \ldots, \mu$:
       $[U_{l-1}, V_{l-1}, D_{l-1}] = \mathbf{MGM}(A_{l-1}, B_{l-1}, U_{l-1}, V_{l-1}, D_{l-1}, l-1)$
     * $l = 1$:
       ◦ exakte Lösung von $A_0 U_0 = B_0 U_0 D_0$, $A_0^* V_0 = B_0^* V_0 \overline{D}_0$, $V_0^* B_0 U_0 = I$
   - setze $\widehat{U}_l = W_l U_{l-1}$, $\widehat{V}_l = \widehat{W}_l V_{l-1}$, $\widehat{D}_l = D_{l-1}$
3. Nachglättung: $\widehat{U}_l \to U_l$, $\widehat{V}_l \to V_l$, $\widehat{D}_l \to D_l$

Dabei enthalten die q-spaltigen Matrizen U_l jeweils die gesuchten Koeffizienten der rechten Eigenvektoren und die q-spaltigen Matrizen V_l entsprechend die der zugehörigen linken Eigenvektoren. Die korrespondierenden Matrizen D_l sind Diagonalmatrizen mit den gesuchten q Eigenwerten in der Diagonalen. Im Gegensatz zu dem in Algorithmus 6 gegebenen hat dieser Mehrgitter-Zyklus die Berechnung der Eigenvektoren zu den q Eigenwerten mit kleinsten Realteilen zum Ziel. Auf die dabei möglicherweise auftretenden Schwierigkeiten sind wir bereits in Bemerkung 12 in Abschnitt 2.6 eingegangen.
Eine Übersicht über die in den letzten beiden Abschnitten entwickelten Mehrgitter-Methoden zur Lösung von Eigenwertproblemen ist in Abbildung 10 gegeben, wobei die

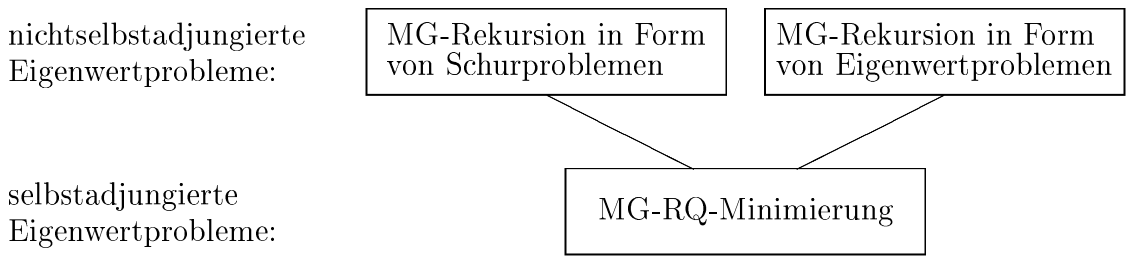

Abbildung 10: Mehrgitter-Projektionsmethoden zur Lösung von Eigenwertproblemen

Kästchen denen in Abbildung 6 auf Seite 37 für Krylovraum-Methoden entsprechen. Im selbstadjungierten Fall gehen sowohl das auf orthogonalen Projektionen beruhende Mehrgitter-Verfahren (Rekursion in Form von Schurproblemen, Algorithmen 5-7) als auch die auf „schiefen" Projektionen basierende Mehrgitter-Methode (Rekursion in Form von Eigenwertproblemen, Algorithmus 8) in die im vorigen Abschnitt dargestellte Mehrgitter-Minimierung (Algorithmen 2-4) über. Insbesondere kann also das durch die Algorithmen 5-7 beschriebene Mehrgitter-Verfahren zur Lösung des allgemeinen Eigenwertproblems (selbstadjungierter und nichtselbstadjungierter Fall) genutzt werden. Im Vergleich zu der in Abschnitt 3.1 vorgestellten Mehrgitter-Methode von Hackbusch sind dabei zum einen die Eigenwertnäherungen in jedem Glättungs- und Grobgitter-Korrekturschritt zusätzliche Unbekannte, die automatisch bestimmt werden, und zum anderen werden keine Projektionen bezüglich (approximativer) Eigenlösungen gröberer Gitter benötigt. Im speziellen selbstadjungierten Fall verfügt das hier entwickelte Mehrgitter-Verfahren im Gegensatz zu Hackbuschs Methode zusätzlich über die wichtige Eigenschaft der Monotonie (siehe dazu den vorherigen Abschnitt).

4 Beispiele

In diesem Kapitel sollen einige numerische Berechnungen, die auf der Grundlage der in dieser Arbeit beschriebenen und in dem MATLAB-Programmpaket [22] implementierten Mehrgitter-Methode durchgeführt worden sind, dokumentiert werden. Dabei wurde die Mehrgitter-Rekursion (Algorithmus 6) stets als V-Zyklus durchlaufen, das heißt mit dem Parameterwert $\mu = 1$, und die Anzahl der Vor- und Nachglättungen, also der Wert ν in Algorithmus 5, jeweils $\nu = 3$ gesetzt. Die Lösung der projizierten Schurprobleme im Glättungsalgorithmus 5 erfolgte mit der MATLAB-Routine `schur` (nach einer B-Orthonormalisierung der Spaltenvektoren der entsprechenden Matrix W), zur Vorkonditionierung wurde die in Bemerkung 24 erwähnte Jacobi-Vorkonditionierung benutzt. Die Korrekturprobleme auf dem gröbsten Gitter in Algorithmus 6 wurden mit dem in Abschnitt 2.6 beschriebenen Block-Arnoldi-Verfahren nicht exakt, sondern nur approximativ bis auf die in Algorithmus 7 vorkommende Genauigkeit tol gelöst.
Zum Aufbau der benötigten Gitterhierarchie kam die „nested-iteration"-Technik zur Anwendung, wobei die Eigenlösungen des aktuellen Gitters einerseits zur adaptiven Gitterverfeinerung herangezogen wurden und andererseits als Startwerte für das verfeinerte Gitter dienten. Der zur adaptiven Verfeinerung benutzte Fehlerindikator beruhte dabei auf der folgenden dreiecksorientierten Strategie: durch Hinzunahme der Kantenmittelpunkte der Dreiecke der aktuellen Triangulierung als zusätzliche Freiheitsgrade und Berechnung der Korrektur in diesen Knoten in Bezug auf den gesamten invarianten Unterraum nach einem Schema analog zu Gleichung (55) wurde für jedes Dreieck t_j ein relativer L^2-Fehler ε_j berechnet. Die Dreiecke t_k, für die die Relation

$$\varepsilon_k > \frac{1}{2} \max_{1 \leq j \leq N_t} \varepsilon_j$$

erfüllt war, erhielten nun eine Markierung. Darauf aufbauend wurde mit der dreiecksbasierten Gitterverfeinerungs-Routine der MATLAB-PDE-Toolbox eine neue Triangulierung erzeugt, indem die markierten Dreiecke uniform in vier Teildreiecke und die an diese angrenzenden mit einer Bisektionsmethode unterteilt wurden, um eventuell „hängende" Knoten zu beseitigen.
Im folgenden werden zunächst zwei „akademische" Beispiele, danach einige kompliziertere Anwendungsbeispiele aus der integrierten Optik dargestellt.

4.1 Komplexes Kastenpotential

Wir betrachten das Eigenwertproblem (7) in der schwachen Formulierung (10) mit

$$f(x,y) = \begin{cases} 100 + 250i, & (x,y) \in \left[\frac{3}{8}, \frac{5}{8}\right]^2 \\ 0, & \text{sonst} \end{cases}$$

auf $\Omega = (0,1)^2$. In einem ersten Programmlauf wurde die Eigenfunktion zum Eigenwert mit kleinstem Realteil, das heißt $q = 1$, bestimmt. Das Startgitter, bestehend

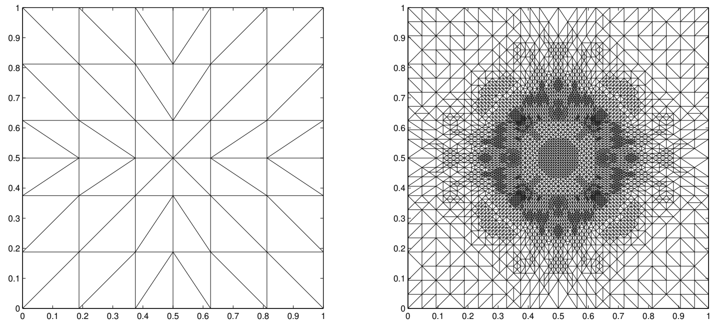

Abbildung 11: Start- und Endgitter zum Kastenpotential

aus 49 Knoten und 72 Dreiecken, ist in Abbildung 11 links, das nach 6 Verfeinerungen erreichte Endgitter mit 4449 Knoten und 8816 Dreiecken in Abbildung 11 rechts aufgetragen. Eine logarithmische Konturdarstellung der berechneten Eigenfunktion u_1

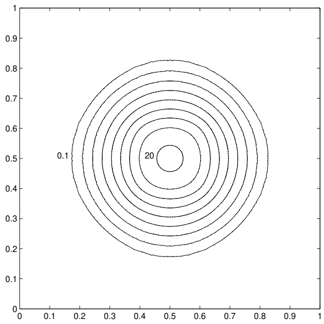

Abbildung 12: Betragsquadrat der Eigenfunktion u_1 zum Kastenpotential (logarithmische Konturlinien)

zum Eigenwert $\lambda_1 \doteq 38.9789 - 183.8185i$ ist in Abbildung 12 gegeben. Die pro Gitter benötigte Anzahl an V-Zyklen zur geforderten Genauigkeit $\mathrm{tol} = 10^{-5}$ ist Tabelle 1 zu entnehmen, der benötigte Rechenaufwand in flops in Abbildung 13 dargestellt. Die CPU-Zeit für die Rechnung auf einer Sun Sparc 4 Workstation betrug 3.6 Minuten.
In einem zweiten Programmlauf waren die ersten 6 Eigenwerte mit kleinsten Realteilen und die dazu korrespondierenden Schurfunktionen gesucht. Als Startgitter diente auch hier die in Abbildung 11 links gezeigte Triangulierung. Die zum (nach 5 Verfeinerungen

Gitter	1	2	3	4	5	6
Anzahl N innerer Punkte	97	237	525	1033	2477	4369
Anzahl V-Zyklen	5	5	6	6	6	6

Tabelle 1: Anzahl N innerer Punkte und V-Zyklen pro Gitter

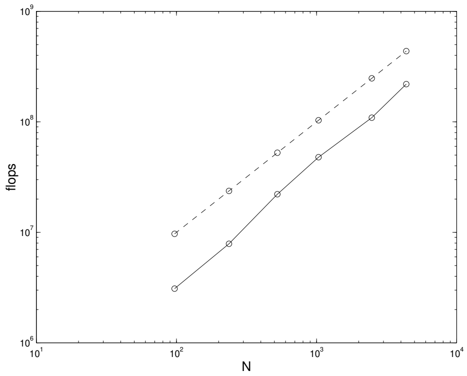

Abbildung 13: Rechenaufwand pro Gitter (zum Vergleich $10^5 N$: gestrichelte Linie)

erreichten) Endgitter berechneten Eigenwerte sind in Abbildung 14 dargestellt, wobei zusätzlich die Eigenwerte des reellen Problems mit

$$f(x,y) = \begin{cases} 100, & (x,y) \in \left[\frac{3}{8}, \frac{5}{8}\right]^2 \\ 0, & \text{sonst} \end{cases}$$

aufgenommen sind. In beiden Fällen ist das Eigenwertproblem entartet, das heißt es existiert jeweils ein Eigenwert mit Vielfachheit 2. Insbesondere ist weiterhin zu beachten, daß im komplexen Fall die Relation $|\lambda_1| \gg |\lambda_j|$ für $j = 2, \dots, 6$ gilt. Die obere Dreiecksmatrix T für das komplexe Problem ergab sich zu

$$T = \begin{pmatrix} \lambda_1 & \tau_{12} & & & & \\ & \lambda_2 & & & & \\ & & \lambda_3 & & & \\ & & & \lambda_4 & & \\ & & & & \lambda_5 & \\ & & & & & \lambda_6 \end{pmatrix}$$

mit $\tau_{12} \doteq 86.6959 - 24.8548i$ (wobei nur die Elemente τ_{kj} mit $|\tau_{kj}| > 10^{-12}$ berücksichtigt wurden), das heißt die bestimmten Schurvektoren $u_1, u_3, \dots, u_6$ sind gleichzeitig auch Eigenvektoren.

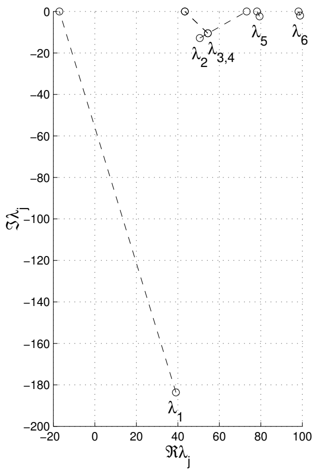

Abbildung 14: Eigenwerte des komplexen und reellen Kastenpotentials

4.2 Harmonischer Oszillator

In diesem Beispiel untersuchen wir das Eigenwertproblem mit der Funktion

$$f(x, y) = -(x^2 + 2ixy + y^2)$$

auf $\Omega = \{(x, y) \in \mathbb{R}^2 : \sqrt{x^2 + y^2} < 5\}$. Ausgehend von dem in Abbildung 15 links

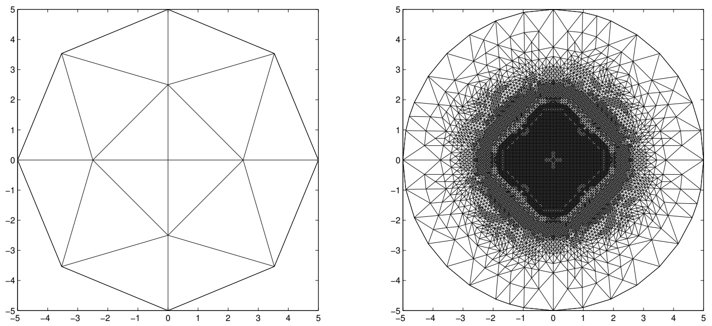

Abbildung 15: Start- und Endgitter zum harmonischen Oszillator

gezeigten Startgitter mit 13 Knoten und 16 Dreiecken wurden die 3 Eigenwerte mit kleinsten Realteilen und die zugehörigen Schurfunktionen bestimmt. Das nach 7 Verfeinerungen erhaltene Endgitter, bestehend aus 11393 Knoten und 22748 Dreiecken, ist in Abbildung 15 rechts dargestellt. Bei Verfeinerung von Randelementen wurden dabei die neu hinzukommenden Randknoten auf den Rand von Ω verschoben.

Die im Mehrgitter-Verfahren zur Berechnung der Grobgitter-Korrekturen benutzten Grobgitter-Basisfunktionen liegen somit nicht in den zu den ursprünglichen groben Gittern gehörenden Finite-Elemente-Räumen, da die aktuell verwendeten Grobgitter keine Triangulierungen mehr sind (es treten Randelemente mit mehr als 3 Ecken auf). Ein logarithmischer Konturplot der zum Endgitter bestimmten Schurfunktionen u_1, u_2 und u_3 ist in Abbildung 16 gegeben, die korrespondierenden Eigenwerte ergaben

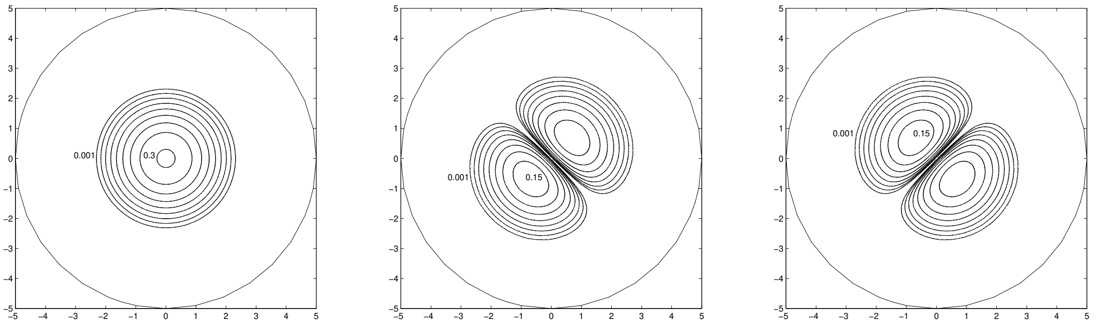

Abbildung 16: Betragsquadrat der Schurfunktionen u_1, u_2 und u_3 zum harmonischen Oszillator (logarithmische Konturlinien)

sich zu $\lambda_1 \doteq 2.1980$ und $\lambda_{2,3} \doteq 4.3962 \pm 0.9106i$. Die Elemente τ_{12}, τ_{13} und τ_{23} der berechneten oberen Dreiecksmatrix T waren betragsmäßig kleiner als 10^{-12}, das heißt die ermittelten Schurfunktionen sind gleichzeitig Eigenfunktionen. Die für eine Genau-

Gitter	1	2	3	4	5	6	7
Anzahl N innerer Punkte	21	81	261	909	2545	4253	11357
Anzahl V-Zyklen	2	3	4	5	4	4	4

Tabelle 2: Anzahl N innerer Punkte und V-Zyklen pro Gitter

igkeit von $\mathrm{tol} = 10^{-5}$ benötigten V-Zyklen pro Gitter sind in Tabelle 2 enthalten, der Rechenaufwand in Abbildung 17 dargestellt. Die CPU-Zeit für die Rechnung betrug 12.3 Minuten auf einer Sun Sparc 4 Workstation. Schließlich sollen noch die durch eine analoge Rechnung bestimmten Eigenwerte $\lambda_1 \doteq 2.0011$ und $\lambda_{2,3} \doteq 4.0030$ des reellen Problems mit

$$f(x, y) = -(x^2 + y^2)$$

angegeben werden. Die vorhandene Entartung im reellen Fall wird im komplexen durch den imaginären „Kopplungsterm“ $2ixy$ aufgehoben.

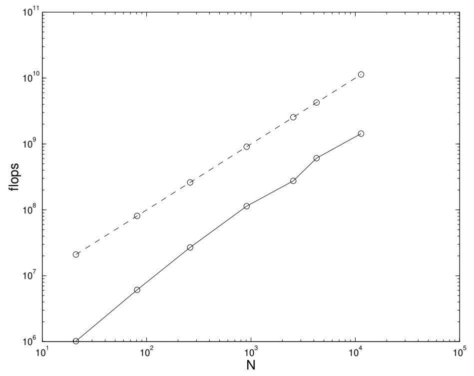

Abbildung 17: Rechenaufwand pro Gitter (zum Vergleich $10^6 N$: gestrichelte Linie)

4.3 Integrierte Optik I: Rippen-Wellenleiter und Koppler

Zwei typische Strukturen der integrierten Optik sind der bereits in Beispiel 2 erwähnte Rippen-Wellenleiter und der Koppler, deren Querschnitte in Abbildung 18 dargestellt sind. Die Komponenten bestehen jeweils aus einem III-V-Mischhalbleitersystem (InP

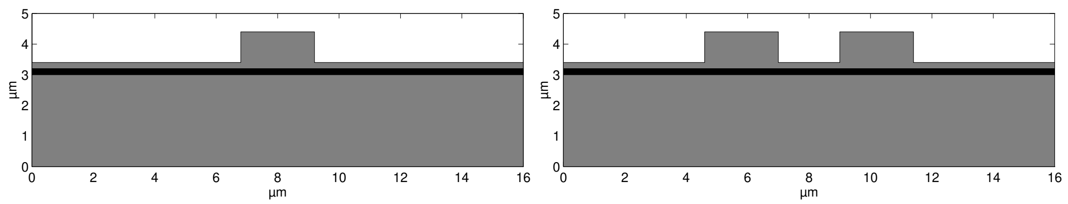

Abbildung 18: Querschnitt des Rippen-Wellenleiters (links) und des Kopplers (rechts)

und GaInAsP). Wie wir in Abschnitt 2.1 gesehen haben, läßt sich die Bestimmung von ausbreitungsfähigen Moden solcher Komponenten auf die Lösung des Eigenwertproblems der skalaren Helmholtzgleichung zurückführen, wobei nur die Eigenlösungen zu den Eigenwerten mit kleinstem Realteil von Interesse sind. Die Funktion f ist hierbei durch $f(x,y) = \omega^2 \tilde{\epsilon}(x,y,\omega)\mu$ gegeben. Üblicherweise werden die komplexe Dielektrizitätszahl $\tilde{\epsilon}$ und die Permeabilität μ in der Form $\tilde{\epsilon}(x,y,\omega) = \epsilon_0 \epsilon_r(x,y,\omega)$ und $\mu = \mu_0 \mu_r$ mit der elektrischen und magnetischen Feldkonstante ϵ_0 und μ_0 geschrieben. Die relative Dielektrizitätszahl ϵ_r ist dabei durch das Quadrat der komplexen Brechungsindizes, das heißt durch $\epsilon_r(x,y,\omega) = n^2(x,y,\omega)$ gegeben, für die relative Permeabilität μ_r gilt in den von uns betrachteten Fällen stets $\mu_r = 1$ (nichtmagnetische Medien). Aus $\omega^2 \epsilon_0 \mu_0 = \omega^2 c^{-2} = k_0^2$ mit der Wellenzahl k_0 im Vakuum erhalten wir so für die

Funktion f die äquivalente Darstellung

$$f(x,y) = k_0^2 n^2(x,y,\omega) \quad .$$

Die Rechnungen dieses Abschnittes basieren nun auf einer Wellenlänge von 1.55 μm im Vakuum, das heißt auf der Kreisfrequenz $\omega \doteq 1.2153 \cdot 10^{15}$ Hz und damit der Wellenzahl $k_0 \doteq 4.0537\ \mu\text{m}^{-1}$. Die entsprechenden Brechungsindizes der vorkommenden Materialien sind in diesem Beispiel reell (keine elektrische Leitfähigkeit) und durch

$$n(x,y,\omega) = \begin{cases} 1.00, & \text{Luft: weiß} \\ 3.17, & \text{InP: grau} \\ 3.38, & \text{GaInAsP: schwarz} \end{cases}$$

gegeben [1]. Auf der Grundlage dieser Werte wurden für den Rippen-Wellenleiter der kleinste Eigenwert mit zugehöriger Eigenfunktion (das heißt $q = 1$) und für den Koppler aufgrund der Symmetrie der Struktur die zwei kleinsten Eigenwerte mit korrespondierenden Eigenfunktionen (also $q = 2$) bestimmt. Die jeweiligen Starttriangulierungen sind in Abbildung 19 oben dargestellt, wobei das Gitter zum Rippen-Wellenleiter aus 206 Knoten und 379 Dreiecken, das zum Koppler aus 242 Knoten und 442 Dreiecken besteht. Die nach 6 Verfeinerungen erreichten Endgitter mit 6489 Knoten und 12924

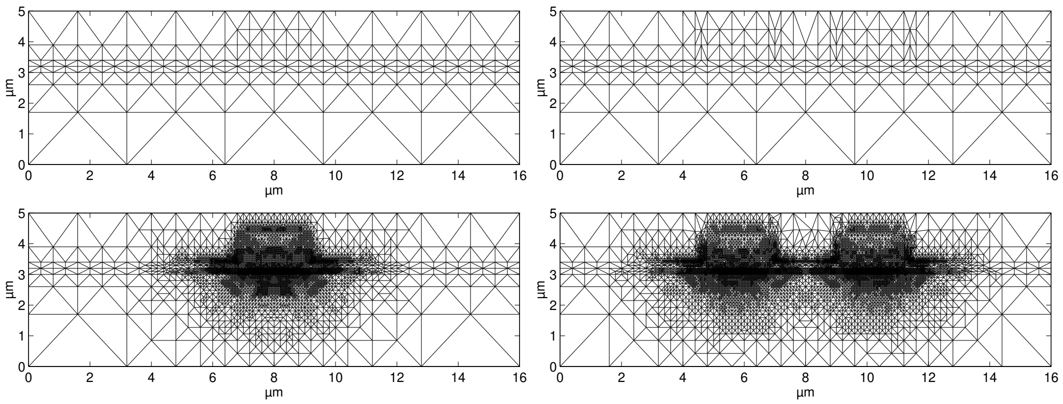

Abbildung 19: Start- und Endgitter zum Rippen-Wellenleiter (links) und zum Koppler (rechts)

Dreiecken (Rippen-Wellenleiter) bzw. 10447 Knoten und 20825 Dreiecken (Koppler) sind in Abbildung 19 unten gezeigt. Logarithmische Konturdarstellungen der zu diesen Endtriangulierungen berechneten Eigenfunktionen finden sich in Abbildung 20. Die Eigenwerte sind für den Rippen-Wellenleiter durch $\lambda_1 \doteq -168.1887\ \mu\text{m}^{-2}$, für den Koppler durch $\lambda_1 \doteq -168.2090\ \mu\text{m}^{-2}$ und $\lambda_2 \doteq -168.1656\ \mu\text{m}^{-2}$ gegeben. Daraus

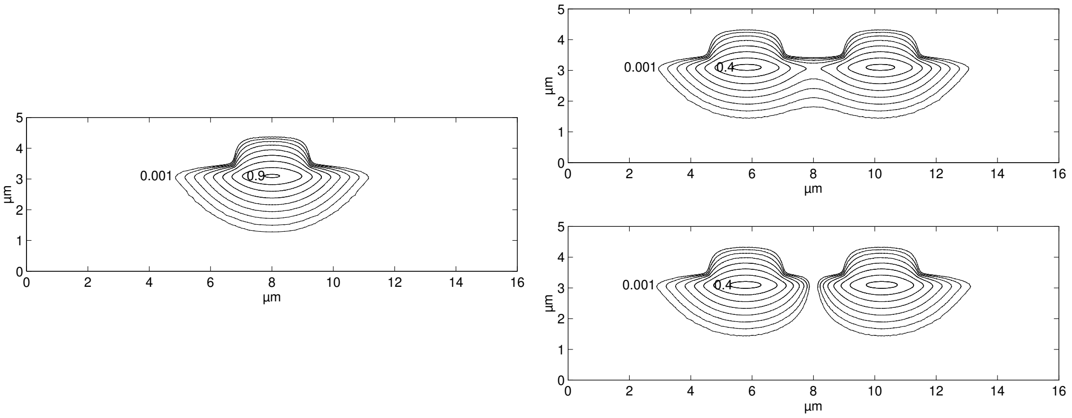

Abbildung 20: Betragsquadrat der Eigenfunktion u_1 zum Rippen-Wellenleiter (links) und der Eigenfunktionen u_1 und u_2 zum Koppler (rechts, logarithmische Konturlinien)

Gitter	1	2	3	4	5	6
Anzahl N innerer Punkte	272	487	885	1743	3388	6437
Anzahl V-Zyklen	6	6	7	7	7	6
Anzahl N innerer Punkte	332	725	1289	2876	4760	10380
Anzahl V-Zyklen	6	8	8	8	8	7

Tabelle 3: Anzahl N innerer Punkte und V-Zyklen pro Gitter (Rippen-Wellenleiter oben und Koppler unten)

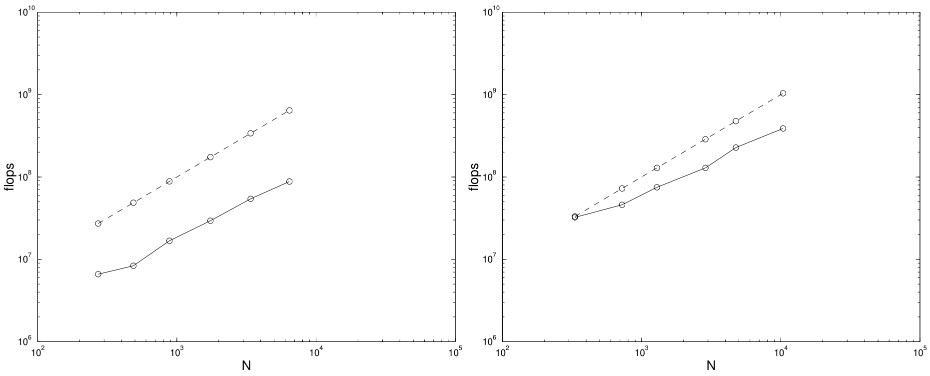

Abbildung 21: Rechenaufwand pro Gitter (Rippen-Wellenleiter links und Koppler rechts, zum Vergleich $10^5 N$: gestrichelte Linie)

ergeben sich für die eigentlich gesuchten Wellenzahlen $k = \sqrt{-\lambda}$ in z-Richtung die Werte $k_1 \doteq 12.9688\ \mu\text{m}^{-1}$ bzw. $k_1 \doteq 12.9695\ \mu\text{m}^{-1}$ und $k_2 \doteq 12.9679\ \mu\text{m}^{-1}$. Die pro Gitter für eine Genauigkeit von tol $= 10^{-5}$ benötigten V-Zyklen sind Tabelle 3 zu entnehmen, der jeweilige Rechenaufwand ist in Abbildung 21 dargestellt. Die CPU-Zeit zur Lösung des Rippen-Wellenleiter-Problems betrug 4.3 Minuten, zur Lösung des Koppler-Problems 11.2 Minuten auf einer Sun Sparc 4 Workstation.

4.4 Integrierte Optik II: Optischer Verstärker

Abschließend soll die in Abbildung 22 dargestellte integriert-optische Struktur untersucht werden. Die genauen Abmessungen der Komponente und die exakten Para-

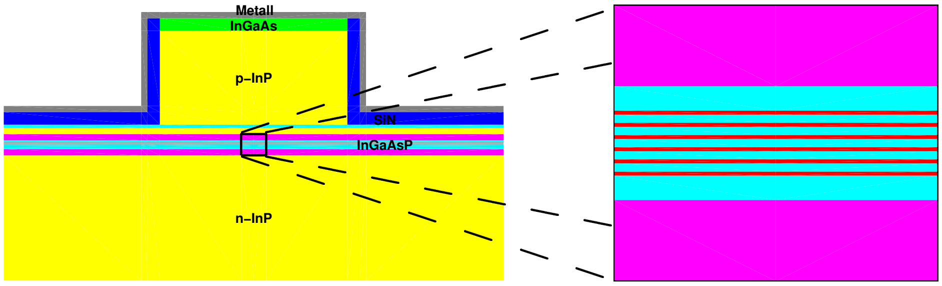

Abbildung 22: Querschnitt des optischen Verstärkers mit Zoom

meter der verwendeten Materialien sind technologisch relevant und daher hier nicht aufgeführt. Die Berechnungen dieses Abschnittes wurden auf der Grundlage einer Wel-

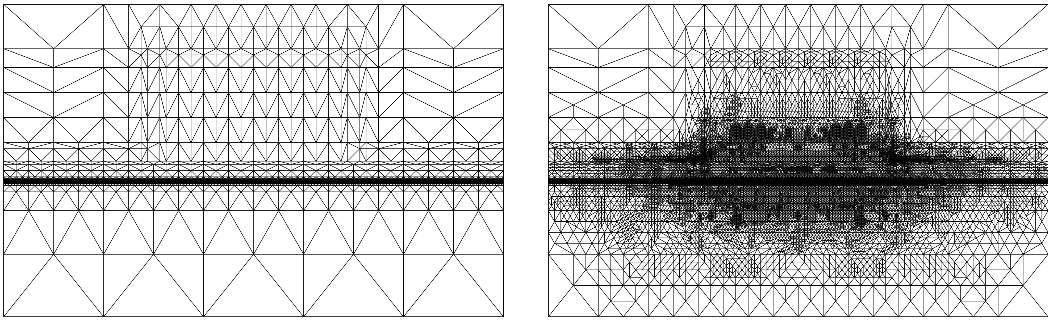

Abbildung 23: Start- und Endgitter zum optischen Verstärker

lenlänge von 1.3 μm im Vakuum, das heißt der Kreisfrequenz $\omega \doteq 1.4490 \cdot 10^{15}$ Hz und somit der Wellenzahl $k_0 \doteq 4.8332\ \mu\mathrm{m}^{-1}$, durchgeführt. Der Brechungsindex ist komplexwertig für die Metallschicht und die Multi-Quantum-Well-Schichten (die dünnen Schichten im Zoom von Abbildung 22), für die anderen Materialien reellwertig. Zur Bestimmung der beiden gesuchten Schurfunktionen zu den Eigenwerten mit kleinsten Realteilen wurde nun das in Abbildung 23 links gezeigte Startgitter benutzt. Da die sehr dünnen, aber funktionsbestimmenden MQW-Schichten (das Verhältnis der Schichtdicke zur Gesamthöhe der Struktur ist 1:1000) mit normal geformten Dreiecken aufgelöst werden mußten, besteht dieses bereits aus 2515 Knoten und 4956 Dreiecken. Das nach 6 Verfeinerungen erreichte Endgitter mit 15246 Knoten und 30396 Drei-

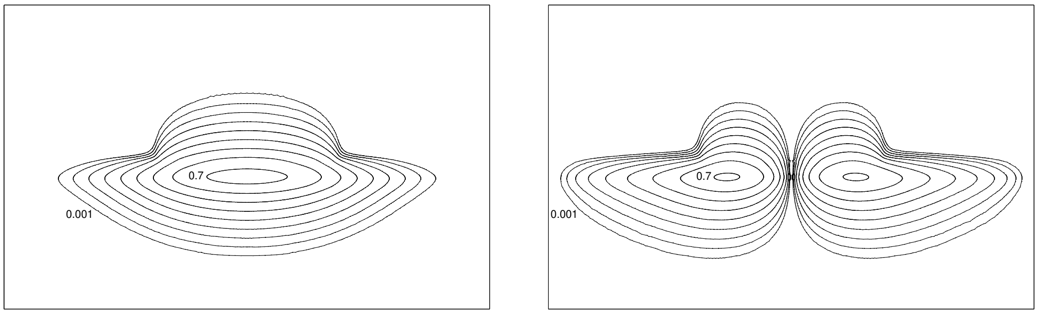

Abbildung 24: Betragsquadrat der Schurfunktionen u_1 und u_2 zum optischen Verstärker (logarithmische Konturlinien)

ecken ist in Abbildung 23 rechts dargestellt, die Betragsquadrate der dazu berechneten Schurfunktionen u_1 und u_2 sind Abbildung 24 zu entnehmen. Die Eigenwerte ergaben sich gemäß $\lambda_1 \doteq (-246.3927 - 0.0725i)\ \mu\mathrm{m}^{-2}$ und $\lambda_2 \doteq (-244.7427 - 0.0452i)\ \mu\mathrm{m}^{-2}$, der Wert τ_{12} war betragsmäßig kleiner als 10^{-5}. Für die Wellenzahlen ergeben sich daraus die Werte $k_1 \doteq (15.6969 + 0.0023i)\ \mu\mathrm{m}^{-1}$ und $k_2 \doteq (15.6443 + 0.0014i)\ \mu\mathrm{m}^{-1}$, das heißt unter Beachtung des Separationsansatzes (5) erhalten wir eine Amplitudenverstärkung um die Faktoren 10 und 4 auf 1 mm in Ausbreitungsrichtung. Die für eine

Gitter	1	2	3	4	5	6
Anzahl N innerer Punkte	2564	2888	3987	5536	9422	15152
Anzahl V-Zyklen	3	3	4	4	4	4

Tabelle 4: Anzahl N innerer Punkte und V-Zyklen pro Gitter

Genauigkeit von $\mathrm{tol} = 10^{-3}$ benötigten V-Zyklen pro Gitter sind Tabelle 4 zu entnehmen, der Rechenaufwand ist in Abbildung 25 aufgetragen. Da die Kosten zur Lösung

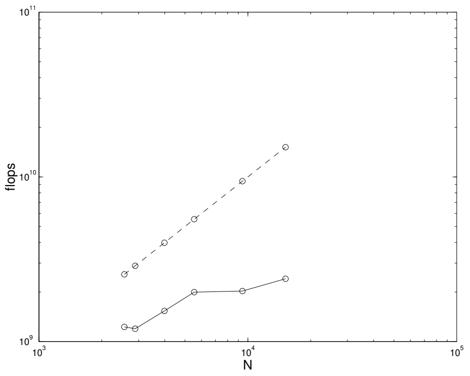

Abbildung 25: Rechenaufwand pro Gitter (zum Vergleich $10^6 N$: gestrichelte Linie)

der jeweils im Mehrgitter-Zyklus auftretenden Korrekturprobleme auf dem gröbsten Gitter aufgrund der hohen Anzahl an Unbekannten vergleichsweise hoch sind, ist die für Mehrgitter-Verfahren typischerweise asymptotisch eintretende direkte Proportionalität von Rechenaufwand und Anzahl an Freiheitsgraden noch nicht zu beobachten. Die CPU-Zeit für die Rechnung betrug 16 Minuten auf einer Sun Ultra 1 Workstation. In weiteren Programmläufen wurde durch eine Parameterstudie die Abhängigkeit der vier Eigenwerte mit kleinsten Realteilen vom Imaginärteil des Brechungsindex in den MQW-Schichten untersucht. Die für fünf ausgewählte Parameter berechneten Eigen-

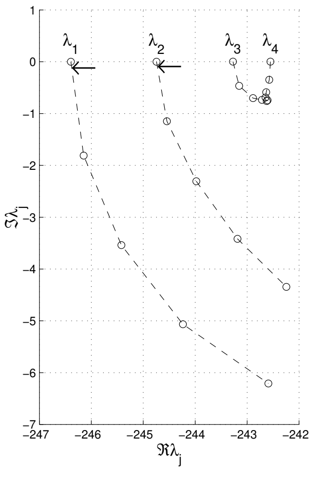

Abbildung 26: Eigenwerte in Abhängigkeit vom Imaginärteil des Brechungsindex in den MQW-Schichten

werte sind in Abbildung 26 dargestellt, wobei die mit der zuvor dokumentierten Rechnung bestimmten Werte durch Pfeile markiert wurden. Diagramme von der Art wie Abbildung 26 sind bei der Entwicklung integriert-optischer Komponenten von technologischem Interesse.

5 Schlußbemerkung und Ausblick

Die Diskretisierung des Eigenwertproblems der skalaren, komplexen Helmholtzgleichung in schwacher Formulierung mittels finiter Elemente führt auf das algebraische Eigenwertproblem

$$Au = \lambda Bu$$

mit der im allgemeinen nichtselbstadjungierten Matrix A und der selbstadjungierten, positiv definiten Matrix B. In dieser Arbeit wurde eine Mehrgitter-Methode zur effizienten Lösung dieses diskreten Problems entwickelt.

Durch eine Sensitivitätsanalyse des kontinuierlichen Eigenproblems konnte zunächst gezeigt werden, daß die stabile Berechnung der gewünschten Eigenwerte und korrespondierenden Eigenfunktionen nur auf der Grundlage der Approximation des zugehörigen invarianten Unterraumes, das heißt der simultanen Bestimmung der zum gesuchten Teil des Spektrums korrespondierenden Eigenlösungen, erfolgen kann. Dabei wurde aus Konditionsgründen zur Darstellung des invarianten Unterraumes nicht die Basis aus Eigenvektoren, sondern die aus Schurvektoren bestehende Orthonormalbasis gewählt. In dem nun zur simultanen Berechnung dieser Schurvektoren hergeleiteten Mehrgitter-Verfahren basiert sowohl jeder Glättungs- als auch Grobgitter-Korrekturschritt analog zur Finite-Elemente-Approximation auf der Lösung eines projizierten Eigenproblems. Damit läßt sich das Verfahren insgesamt in die Klasse der orthogonalen Projektionsmethoden einordnen, der unter anderem auch die Krylovraum-Methoden zur approximativen Lösung von Eigenwertproblemen angehören. Im speziellen selbstadjungierten Fall liegt dem Verfahren außerdem ein Minimalprinzip zugrunde, so daß dieses dann zusätzlich zu den monotonen Mehrgitter-Methoden zählt.

In dem in dieser Arbeit betrachteten Kontext waren stets die Eigenlösungen zu den Eigenwerten mit kleinsten Realteilen von Interesse. Durch eine einfache Modifikation der vorgestellten Algorithmen kann das Verfahren auch zur Bestimmung von anders charakterisierten Eigenwerten, beispielsweise der betragskleinsten oder -größten, und des zugehörigen invarianten Unterraumes eingesetzt werden. Aufgrund der allgemein gehaltenen Herleitung der Methode lassen sich dabei auch Probleme, die nicht notwendigerweise aus der Diskretisierung eines partiellen Differentialoperators entstanden sind, bearbeiten. Voraussetzungen für den Einsatz des Mehrgitter-Algorithmus sind das Vorliegen einer hierarchischen Multilevel-Struktur (ähnlich der hier durch fortgesetzte adaptive Gitterverfeinerung erhaltenen) und gleiche analytische Eigenschaften (diskretes Spektrum, Vollständigkeit der Eigenlösungen) wie im Fall des Eigenwertproblems der Helmholtzgleichung. Zur Behandlung nichtlinearer Eigenwertprobleme kann die Methode zudem, beispielsweise im Rahmen des Newton-Verfahrens, zur Lösung auftretender Teilprobleme benutzt werden.

Die für Mehrgitter-Verfahren typische optimale Komplexität konnte anhand einiger ausgewählter Beispielrechnungen belegt werden. Ein Beweis dieser Beobachtung ist bislang nur für den selbstadjungierten Fall und der Berechnung einer einfachen Eigenlösung bekannt. Die Erweiterung dieser Theorie auf den nichtselbstadjungierten Fall und auf die Bestimmung beliebiger invarianter Unterräume ist eine Aufgabe zukünftiger Forschungsarbeit.

Literatur

[1] R. Accornero, M. Artiglia, G. Coppa, P. Di Vita, G. Lapenta, M. Potenza, P. Ravetto. Finite Difference Methods for the Analysis of Integrated Optical Waveguides. Electronic Letters **26**, S. 1959-1960, 1990.

[2] M. Aigner. Diskrete Mathematik. Vieweg Braunschweig Wiesbaden, 1993.

[3] I. Babuška, J. Osborn. Eigenvalue Problems. In P. G. Ciarlet, J. L. Lions. *Handbook of Numerical Analysis, Vol. II: Finite Element Methods (Part 1)*, S. 641-788. Elsevier Science Publishers B. V. (North-Holland) Amsterdam New York Oxford Tokyo, 1991.

[4] R. E. Bank. Analysis of a Multilevel Inverse Iteration Procedure for Eigenvalue Problems. SIAM J. Numer. Anal. **19**, S. 886-898, 1982.

[5] R. E. Bank, T. F. Chan. PLTMGC: A Multi-Grid Continuation Program for Parameterized Nonlinear Elliptic Systems. SIAM J. Sci. Stat. Comput. **7**, S. 540-559, 1986.

[6] W. W. Bradbury, R. Fletcher. New Iterative Methods for Solution of the Eigenproblem. Numer. Math. **9**, S. 259-267, 1966.

[7] D. Braess. Finite Elemente. Springer-Verlag Berlin Heidelberg New York, 1992.

[8] J. H. Bramble, A. V. Knyazev, J. E. Pasciak. A Subspace Preconditioning Algorithm for Eigenvector/Eigenvalue Computation. Advances Comp. Math. **6**, S. 159-189, 1996.

[9] Z. Cai, J. Mandel, S. McCormick. Multigrid Methods for Nearly Singular Linear Equations and Eigenvalue Problems. SIAM J. Numer. Anal. **34**, S. 178-200, 1997.

[10] T. F. Chan, I. Sharapov. Subspace Correction Multilevel Methods for Elliptic Eigenvalue Problems. CAM Report **96-34**, University of California Los Angeles, 1996.

[11] R. Courant, D. Hilbert. Methoden der mathematischen Physik. Springer-Verlag Berlin Heidelberg New York, 1993.

[12] J. K. Cullum, R. A. Willoughby. Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Vol. I: Theory. Birkhäuser Boston Basel Stuttgart, 1985.

[13] P. Deuflhard, T. Friese, F. Schmidt, R. März, H.-P. Nolting. Effiziente Eigenmodenberechnung für den Entwurf integriert-optischer Chips. In K.-H. Hoffmann, W. Jäger, T. Lohmann, H. Schunck. *Mathematik-Schlüsseltechnologie für die Zukunft*, S. 267-280. Springer-Verlag Berlin Heidelberg New York, 1996.

[14] P. Deuflhard, A. Hohmann. Numerische Mathematik I (2. Auflage). Walter de Gruyter Berlin New York, 1993.

[15] B. Döhler. Ein neues Gradientenverfahren zur simultanen Berechnung der kleinsten oder größten Eigenwerte des allgemeinen Eigenwertproblems. Numer. Math. **40**, S. 79-91, 1982.

[16] N. Dunford, J. T. Schwartz. Linear Operators Part I: General Theory. Wiley Interscience New York, 1958.

[17] N. Dunford, J. T. Schwartz. Linear Operators Part II: Spectral Theory (Self Adjoint Operators in Hilbert Space). Wiley Interscience New York, 1963.

[18] K. J. Ebeling. Integrierte Optoelektronik. Springer-Verlag Berlin Heidelberg New York, 1989.

[19] D. K. Faddejew, W. N. Faddejewa. Numerische Methoden der linearen Algebra (2. Auflage). Oldenbourg Verlag München Wien, 1970.

[20] J. G. F. Francis. The QR Transformation: A Unitary Analogue to the LR Transformation, Parts I and II. Comput. J. **4**, S. 265-271, 332-345, 1961.

[21] R. W. Freund, M. H. Gutknecht, N. M. Nachtigal. An Implementation of the look-ahead Lanczos Algorithm for Non-Hermitian Matrices. SIAM J. Sci. Comput. **14**, S. 137-158, 1993.

[22] T. Friese. Matlab-Programme zur Lösung von Eigenwertproblemen mit einem adaptiven, nichtlinearen Mehrgitter-Verfahren. Technical Report TR **97-01**, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1997.

[23] I. C. Gohberg, M. G. Kreĭn. Introduction to the Theory of Linear Nonselfadjoint Operators. Transl. Math. Monographs Vol. **18**, Amer. Math. Soc. Providence, 1969.

[24] G. H. Golub, C. F. Van Loan. Matrix Computations (Second Edition). The Johns Hopkins University Press Baltimore London, 1989.

[25] W. Hackbusch. On the Computation of Approximate Eigenvalues and Eigenfunctions of Elliptic Operators by Means of a Multi-Grid Method. SIAM J. Numer. Anal. **16**, S. 201-215, 1979.

[26] W. Hackbusch. Multi-Grid Methods and Applications. Springer-Verlag Berlin Heidelberg New York, 1985.

[27] W. Hackbusch. Multi-Grid Eigenvalue Computation. In D. Braess, W. Hackbusch, U. Trottenberg. *Advances in Multi-Grid Methods*, S. 24-32. Notes on Numerical Fluid Mechanics Vol. **11**, Vieweg Braunschweig Wiesbaden, 1985.

[28] W. Hackbusch. Theorie und Numerik elliptischer Differentialgleichungen. B. G. Teubner Stuttgart, 1986.

[29] M. R. Hestenes, W. Karush. A Method of Gradients for the Calculation of the Characteristic Roots and Vectors of a Real Symmetric Matrix. J. Res. Nat. Bur. Stand. **47**, S. 45-61, 1951.

[30] H. Heuser. Funktionalanalysis. B. G. Teubner Stuttgart, 1992.

[31] C. G. J. Jacobi. Über ein leichtes Verfahren, die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen. Crelle's J. **30**, S. 51-94, 1846.

[32] M. Jung, U. Langer, A. Meyer, W. Queck, M. Schneider. Multigrid Preconditioners and their Applications. In G. Telschow. *Third Multigrid Seminar*, S. 11-52. Report R-MATH-**03/89**, Karl-Weierstrass-Institut für Mathematik Berlin, 1989.

[33] W. Kahan, B. N. Parlett, E. Jiang. Residual Bounds on Approximate Eigensystems of Nonnormal Matrices. SIAM J. Numer. Anal. **19**, S. 470-484, 1982.

[34] T. Kato. Perturbation Theory for Linear Operators (Second Edition). Springer-Verlag Berlin Heidelberg New York, 1984.

[35] V. E. Katsnelson. Conditions under which systems of eigenvectors of some classes of operators form a basis. Functional Anal. Appl. **1**, S. 122-132, 1967.

[36] V. N. Kublanovskaya. On Some Algorithms for the Solution of the Complete Eigenvalue Problem. USSR Comp. Math. Phys. **3**, S. 637-657, 1961.

[37] R. B. Lehoucq. Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration. Ph.D. Thesis, Rice University Houston, 1995.

[38] R. B. Lehoucq, D. C. Sorensen, C. Yang. Arpack User's Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods. Rice University Houston, 1997.

[39] P. Leinen, W. Lembach, K. Neymeyr. An Adaptive Subspace Method for Elliptic Eigenproblems with Hierarchical Basis Preconditioning. SFB 382, Report **68**, Universität Tübingen, 1997.

[40] D. E. Longsine, S. F. McCormick. Simultaneous Rayleigh-Quotient Minimization Methods for $Ax = \lambda Bx$. Lin. Alg. Appl. **34**, S. 195-234, 1980.

[41] R. März. Integrated Optics: Design and Modelling. Artech House Boston London, 1995.

[42] J. Mandel, S. F. McCormick. A Multilevel Variational Method for $Au = \lambda Bu$ on Composite Grids. J. Comp. Phys. **80**, S. 442-452, 1989.

[43] S. F. McCormick. Multilevel Adaptive Methods for Elliptic Eigenproblems: A Two-Level Convergence Theory. SIAM J. Numer. Anal. **31**, S. 1731-1745, 1994.

[44] B. N. Parlett. The Symmetric Eigenvalue Problem. Prentice-Hall Englewood Cliffs, 1980.

[45] M. Reed, B. Simon. Methods of Modern Mathematical Physics I: Functional Analysis. Academic Press New York, 1972.

[46] A. Ruhe. An Algorithm for Numerical Determination of the Structure of a General Matrix. BIT **10**, S. 196-216, 1970.

[47] A. Ruhe. Computation of Eigenvalues and Eigenvectors. In V. A. Barker. *Sparse Matrix Techniques*, S. 130-184. Lecture Notes in Mathematics Vol. **572**, Springer-Verlag Berlin Heidelberg New York, 1977.

[48] A. Ruhe. Numerical Aspects of Gram-Schmidt Orthogonalization of Vectors. Lin. Alg. Appl. **52-53**, S. 591-601, 1983.

[49] H. Rutishauser. Computational Aspects of F. L. Bauer's Simultaneous Iteration Method. Numer. Math. **13**, S. 4-13, 1969.

[50] Y. Saad. Numerical Methods for Large Eigenvalue Problems. Manchester University Press, 1992.

[51] F. Schmidt. Simultaneous Computation of the Lowest Eigenvalues and Eigenvectors of the Helmholtz Equation. Technical Report TR **95-13**, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1995.

[52] F. Schmidt, D. Yevick. Discrete Transparent Boundary Conditions for Schrödinger-Type Equations. J. Comp. Phys. **134**, S. 96-107, 1997.

[53] H. R. Schwarz. Simultaneous Rayleigh-Quotient Iteration Methods for Large Sparse Generalized Eigenvalue Problems. In J. Hinze. *Numerical Integration of Differential Equations and Large Linear Systems*, S. 384-398. Lecture Notes in Mathematics Vol. **968**, Springer-Verlag Berlin Heidelberg New York, 1982.

[54] H. R. Schwarz. Methode der finiten Elemente. B. G. Teubner Stuttgart, 1991.

[55] D. C. Sorensen. Implicit Application of Polynomial Filters in a k-step Arnoldi Method. SIAM J. Matr. Anal. Appl. **13**, S. 357-385, 1992.

[56] G. W. Stewart. Simultaneous Iteration for Computing Invariant Subspaces of Non-Hermitian Matrices. Numer. Math. **25**, S. 123-136, 1976.

[57] G. Strang, G. J. Fix. An Analysis of the Finite Element Method. Prentice-Hall Englewood Cliffs, 1973.

[58] J. H. Wilkinson. The Algebraic Eigenvalue Problem. Clarendon Press Oxford, 1965.

[59] H. Yang. Conjugate Gradient Methods for the Rayleigh Quotient Minimization of Generalized Eigenvalue Problems. Computing **51**, S. 79-94, 1993.

A Zusammenfassung

Die Berechnung von Eigenlösungen partieller Differentialoperatoren ist eine in den Natur- und Ingenieurwissenschaften häufig auftretende Problemstellung. In der integrierten Optik beispielsweise läßt sich die Bestimmung geführter Lichtmoden optischer Komponenten unter geeigneten Modellannahmen auf die Lösung des Eigenwertproblems der skalaren Helmholtzgleichung

$$-\Delta u(x,y) - f(x,y)u(x,y) = \lambda u(x,y) \tag{$*$}$$

mit Dirichlet-Randbedingung auf einem beschränkten, zweidimensionalen Gebiet zurückführen. Dabei kann die beschränkte Funktion f, die Materialparameter enthält, sowohl reellwertig (selbstadjungiertes Eigenwertproblem) als auch komplexwertig (nichtselbstadjungiertes Eigenwertproblem) sein. Während die Untersuchung des selbstadjungierten Problems als ein im wesentlichen abgeschlossenes Forschungsgebiet angesehen werden kann, stellt die Behandlung des nichtselbstadjungierten Falls einen wichtigen, aktuellen Forschungsschwerpunkt dar.
In der Dissertation werden zunächst einige bekannte Aussagen über die analytischen Eigenschaften des Problems ($*$) (Struktur des Spektrums des assoziierten Helmholtzoperators, Vollständigkeit der Eigenfunktionen, Abhängigkeit der Eigenlösungen von Störungen in der Funktion f), insbesondere im nichtselbstadjungierten Fall, dargestellt. Den Schwerpunkt der Arbeit bildet daran anschließend die numerische Lösung des Eigenwertproblems ($*$). Ausgehend von der Variationsformulierung des Problems wird die Methode der finiten Elemente über Dreiecksgittern zur approximativen Berechnung der gewünschten Eigenlösungen benutzt. Es werden einige Standardverfahren (zum Beispiel Krylovraum-Methoden) zur Lösung des dabei entstehenden diskreten Eigenwertproblems

$$Au = \lambda Bu \tag{$**$}$$

mit den dünn besetzten und hochdimensionalen Matrizen A und B beschrieben. Da die effiziente Berechnung möglichst guter Approximationen der kontinuierlichen Eigenlösungen die Verwendung einer Folge von adaptiv verfeinerten Triangulierungen erforderlich macht, bieten sich alternativ zur Lösung von ($**$) Mehrgitter-Methoden an. Der Forschungsbeitrag der Dissertation besteht in der Verallgemeinerung eines bekannten Mehrgitter-Verfahrens zur Lösung des selbstadjungierten Eigenwertproblems (Rayleigh-Quotienten-Mehrgitter-Minimierung von Mandel und McCormick) auf den nichtselbstadjungierten Fall. Dazu werden vorbereitend zum einen die Konvergenzeigenschaften des Verfahrens der konjugierten Gradienten zur Minimierung des Rayleigh-Quotienten auf einem Gitter untersucht, zum anderen wird die Methode von Mandel und McCormick in einer äquivalenten Formulierung dargestellt, wobei sich zusätzlich die Erweiterung des Verfahrens zur Berechnung invarianter Unterräume (simultane Berechnung mehrerer Eigenlösungen) ergibt. Diese Formulierung erlaubt dann die Verallgemeinerung der Methode auf den nichtselbstadjungierten Fall. Die für Mehrgitter-Verfahren typische optimale Komplexität wird abschließend anhand einiger repräsentativer Beispiele dokumentiert.

B Lebenslauf

- geboren am 4. Juni 1968 in Schwerin
- 1975-1985: zehnklassige allgemeinbildende polytechnische Oberschule in Schwerin
- 1985-1988: Berufsausbildung mit Abitur zum Facharbeiter für Anlagentechnik im Kabelwerk Nord Schwerin
- 1988-1989: Grundwehrdienst
- 1990-1995: Diplomstudium der Mathematik mit Nebenfach Physik an der Freien Universität Berlin (Diplomarbeit mit dem Titel „Eine adaptive Spektralmethode zur Berechnung periodischer Orbits“, Betreuer: Prof. Dr. P. Deuflhard)
- 1994-1995: Forschungstutor am Konrad-Zuse-Zentrum für Informationstechnik Berlin
- seit 1995: wissenschaftlicher Angestellter am Konrad-Zuse-Zentrum für Informationstechnik Berlin