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. $\diamond$

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.

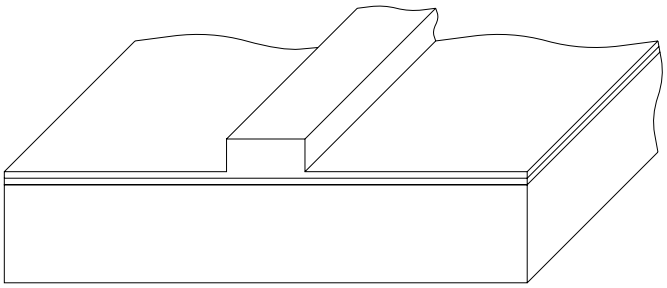

Abbildung 1: Aufbau eines Rippen-Wellenleiters

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

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

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 $\partial\Omega$.

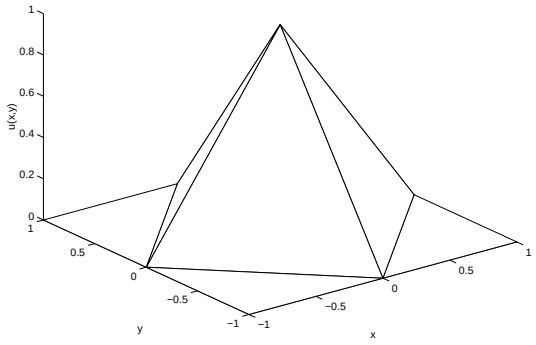

Abbildung 2: Eine Funktion aus $H_0^1(\Omega)$

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

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

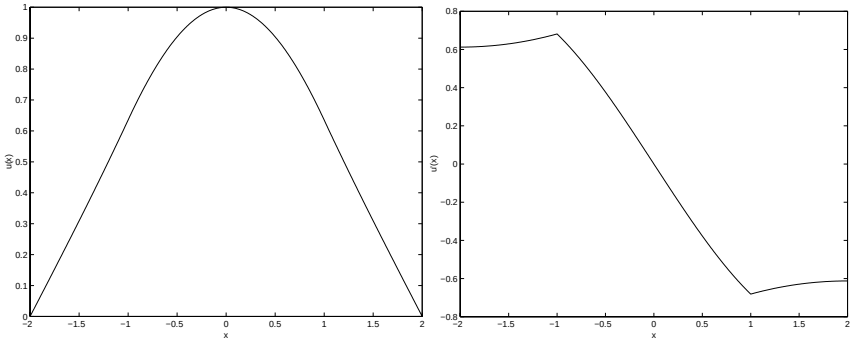

Abbildung 3: Schwache Eigenlösung zum Sprungproblem und erste Ableitung der Lösung

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

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

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*

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

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 $E(0) = \operatorname{span}(w_1)$ und $N(0) = \operatorname{span}(w_j)_{j=1}^{\infty}$, 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

$$V_k = BW_k = \{Bw \,|\, w \in W_k\}$$

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

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

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

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

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

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

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

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

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

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

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

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

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', $l = 1, \dots, m$. Dabei sind die Koeffizienten α_{jl} für $j \le m$ 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

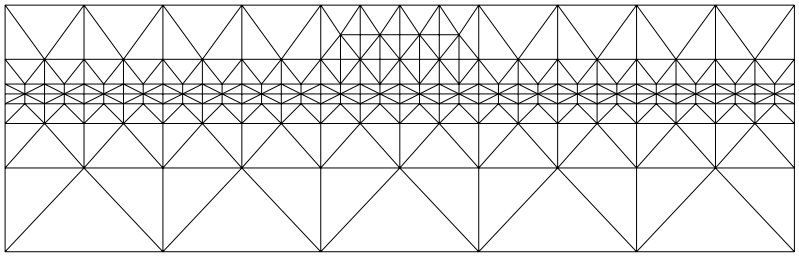

Abbildung 4: Triangulierung eines Gebietes

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 $W_N = \operatorname{span}(w_l)_{l=1}^N =: W_h$ 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 $V_q = \operatorname{span}(u_j)_{j=1}^q$ durch einen von Funktionen $u_{j,h} \in W_h \subset H_0^1(\Omega)$ aufgespannten diskreten invarianten Unterraum $V_{q,h} = \operatorname{span}(u_{j,h})_{j=1}^q$ 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 $u_{1,h}, \ldots, u_{q,h} \in W_h$ 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 $u_{2,h}, \ldots, u_{q,h}$ 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

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

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

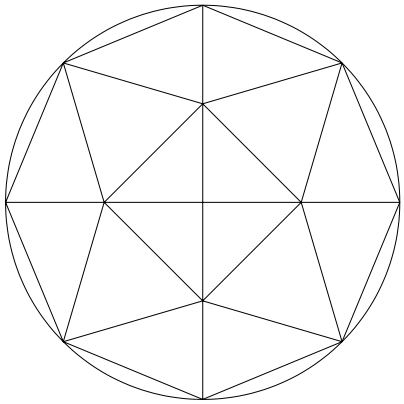

Abbildung 5: Triangulierung eines krummlinig berandeten, konvexen Gebietes

$\Omega_h \subset \Omega$. 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, fw_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 $q \leq n \leq N$ 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

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

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 $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_q$ 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 $\lambda_1, \ldots, \lambda_q$ 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 $(w_1, \ldots, w_q) = W_q$ mit $W_q^*BW_q = I$

- 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

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

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. $\diamond$

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 $w_{m-q}, \ldots, w_{m+q-1}$ 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 $\widehat{W}^* B W = I$ 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 $W^T B W = I$ zu

$$\begin{aligned} \big(W^T A W\big) Q &= Q D \\ Q^T Q &= I \quad , \end{aligned} \tag{38}$$

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

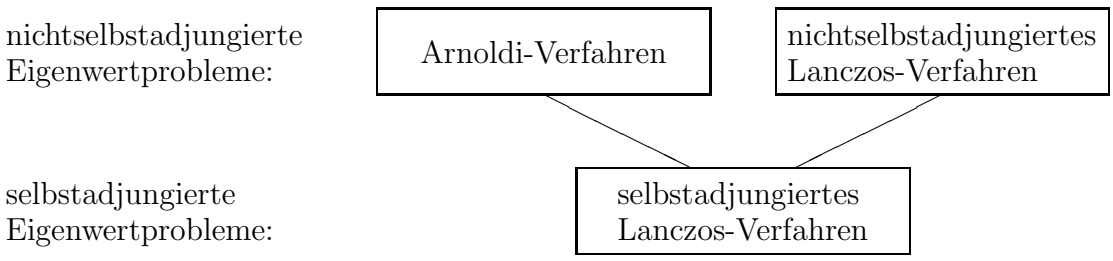

Abbildung 6: Krylovraum-Methoden für Eigenwertprobleme

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.