Eigenwertprobleme

Lothar Nannen

Sommersemester 2018

Version: 16. September 2020

Inhaltsverzeichnis

1. Einleitung

Diese Einleitung orientiert sich in Teilen an [Kre02] und soll die Bedeutung der Eigenwerte des Laplace-Operators bei der Lösung von akustischen Problemen verdeutlichen. Dazu betrachten wir Schallwellen als kleine Störungen im Geschwindigkeitsfeld $v = v(x,t)$, dem Druck $p = p(x,t)$ und der Dichte $\rho = \rho(x,t)$ von Flüssigkeiten oder Gasen, wobei $x \in \mathbb{R}^3$ den Ort und t die Zeit bezeichnet. Seien $v_0 = 0$, $p_0 = \text{const}$ und $\rho_0 = \rho_0(x)$ der stationäre Zustand. Dann wird der Zusammenhang zwischen v, p und ρ beschrieben durch die linearisierte Eulergleichung

$$\rho_0 \frac{\partial v}{\partial t} + \operatorname{grad} p = 0, \tag{1.1}$$

die linearisierte Kontinuitätsgleichung

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho_0 v) = 0 \tag{1.2}$$

und die linearisierte Zustandsgleichung

$$p = c^2 \rho. \tag{1.3}$$

$c = c(x)$ bezeichnet die lokale Schallgeschwindigkeit und beschreibt das konkrete Material, z.B. ist für Luft bei 20°C $c \approx 343 m/s$ oder für Wasser mit der Temperatur T, dem Salzgehalt S und der Tiefe h $c(T,S,h) \approx 1404.85 + 4.618T - 0.0523T^2 + 1.25S + 0.017h$ $(c(4°C, 3.5\%, 100m) \approx 1424 m/s)$.

Einsetzen von (1.3) in (1.2) liefert

$$\frac{1}{c^2}\frac{\partial p}{\partial t} + \operatorname{div}(\rho_0 v) = 0. \tag{1.4}$$

Die Zeitableitung von (1.4) subtrahiert von der Divergenz von (1.1) führt mit $\operatorname{div}\operatorname{grad} = \Delta$ zur Wellengleichung für den Druck p:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = \Delta p. \tag{1.5}$$

Für zeit-harmonische Wellen der Form

$$p(x,t) = \Re\left(u(x)e^{-i\omega t}\right) \tag{1.6}$$

mit der Kreisfrequenz $\omega > 0$ und der komplexen, ortsabhängigen Amplitude $u = u(x)$ folgt die reduzierte Wellengleichung

$$\Delta u + \frac{\omega^2}{c^2} u = 0. \tag{1.7}$$

Im Falle eines homogenen Mediums, d.h. $c = \text{const}$, heisst $\kappa = \omega/c > 0$ die Wellenzahl und (1.7) wird zur Helmholtz-Gleichung

$$\Delta u + \kappa^2 u = 0. \tag{1.8}$$

Zu beachten ist, dass die eigentliche physikalische Größe der Druck p und nicht die komplexe Amplitude u ist.

Werden Lösungen von (1.8) auf einem beschränkten Lipschitz-Gebiet Ω gesucht, so benötigen wir noch Randbedingungen auf dem Rand $\partial\Omega$. In der Akustik spricht man von schallweichen (Dirichlet) und schallharten (Neumann) Randbedingungen. Im folgenden Beispiel nehmen wir homogene Neumann-Randbedingungen und sehen eine Quelle f im Gebiet als Anregung vor:

$$-\Delta u - \kappa^2 u = f \qquad \text{auf } \Omega := (0,\pi)^2, \tag{1.9a}$$

$$\frac{\partial u}{\partial n} = 0 \qquad \text{auf } \partial\Omega. \tag{1.9b}$$

Aus der Theorie selbstadjungierter Operatoren ist bekannt, dass der negative Laplace-Operator $-\Delta : \mathcal{D}(-\Delta) \subset L^2(\Omega) \to L^2(\Omega)$ ein reines Punktspektrum aus reellen Eigenwerten besitzt und die zugehörigen Eigenfunktionen eine Orthonormalbasis des $L^2(\Omega)$ bilden. In diesem Fall sind die Wurzeln der Eigenwerte durch

$$\kappa_{\nu,\mu} = \sqrt{\nu^2 + \mu^2}, \qquad \nu,\mu = 0,1,... \tag{1.10}$$

und die zugehörigen Eigenfunktionen durch

$$u_{\nu,\mu} = \frac{2}{\pi}\cos(\nu x_1)\cos(\mu x_2), \qquad (x_1,x_2) \in \Omega,\ \nu,\mu = 0,1,... \tag{1.11}$$

gegeben und (1.9) ist genau dann eindeutig lösbar, wenn κ^2 kein Eigenwert von $-\Delta$ ist. Die Lösung ist gegeben durch

$$u_\kappa = \sum_{\nu,\mu=0}^{\infty} \frac{(f, u_{\nu,\mu})_{L^2(\Omega)}}{\kappa_{\nu,\mu}^2 - \kappa^2} u_{\nu,\mu}. \tag{1.12}$$

Das Verhalten der Lösung in der Nähe der Wurzel κ^* eines isolierten Eigenwertes wird in der Regel von der zughörigen Eigenfunktion u_{κ^*} dominiert:

$$u_\kappa \approx \frac{(f, u_{\kappa^*})_{L^2(\Omega)}}{\kappa^{*2} - \kappa^2} u_{\kappa^*}, \qquad |\kappa^* - \kappa| \text{ hinreichend klein.} \tag{1.13}$$

2. Helmholtz-Probleme auf beschränkten Gebieten

Sei Ω eine offene, beschränkte Teilmenge des $\mathbb{R}^d$ mit $d = 1, 2, 3$ mit Lipschitz-Rand $\partial\Omega$ und äußerem Normalenvektor n. Wir betrachten zunächst das Quell-Problem: Gesucht ist eine Lösung u von

$$-\Delta u - \kappa^2 p u = l \qquad \text{auf } \Omega, \tag{2.1a}$$

$$\frac{\partial u}{\partial n} = g \qquad \text{auf } \partial\Omega, \tag{2.1b}$$

wobei eine Wellenzahl $\kappa > 0$, eine strikt positive Koeffizientenfunktion p , eine Quelle l und Neumann-Randdaten g gegeben sind. Für positive $p \in L^\infty(\Omega)$, $l \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$ und $u, v \in H^1(\Omega)$ sind folgende Ausdrücke wohldefiniert

$$a(u,v) := \int_\Omega \nabla u \cdot \overline{\nabla v}\, dx, \tag{2.2a}$$

$$b(u,v) := \int_\Omega p\, u\, \overline{v}\, dx, \tag{2.2b}$$

$$f(v) := \int_\Omega l\, \overline{v}\, dx + \int_{\partial\Omega} g\, \overline{v} ds. \tag{2.2c}$$

und es ergibt sich die schache Formulierung des Helmholtz-Problems:

Definition 2.1 (Helmholtz-Quellproblem)**.** *Sei $u \in H^1(\Omega)$ Lösung von*

$$a(u,v) - \kappa^2 b(u,v) = f(v), \qquad v \in H^1(\Omega). \tag{2.3}$$

$a(\bullet,\bullet)$ und $b(\bullet,\bullet)$ sind auf $H^1(\Omega) \times H^1(\Omega)$ symmetrische Sesquilinearformen (auch hermitesche Formen), d.h. sie sind

- linear im ersten Argument: $s(\alpha u_1 + u_2, v) = \alpha s(u_1, v) + s(u_2, v)$,
- antilinear im zweiten Argument: $s(u, \alpha v_1 + v_2) = \overline{\alpha} s(u, v_1) + s(u, v_2)$ und
- symmetrisch/ hermitesch: $s(u,v) = \overline{s(v,u)}$.

$b(\bullet,\bullet)$ ist auf $L^2(\Omega)$ stetig und koerzitiv (sogar positiv definit), d.h.

- $|b(u,v)| \leq \|p\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}$ und

- $b(v,v) \geq C\|v\|^2_{L^2(\Omega)}$ mit $C := \inf_{x\in\Omega} p(x) > 0$.

$a(\bullet,\bullet) + \epsilon\,(\bullet,\bullet)_{L^2(\Omega)}$ mit beliebigem $\epsilon > 0$ und dem $L^2(\Omega)$-Skalarprodukt $(\bullet,\bullet)_{L^2(\Omega)}$ ist ebenfalls stetig und koerzitiv auf $H^1(\Omega)$, d.h. es existiert ein $C = \min\{1,\epsilon\} > 0$ mit

$$|a(v,v) + \epsilon\,(\bullet,\bullet)_{L^2(\Omega)}| \geq C\|v\|^2_{H^1(\Omega)}, \qquad v \in H^1(\Omega). \tag{2.4}$$

f ist ein antilineares, stetiges Funktional auf $H^1(\Omega)$.

Lemma 2.2 (Lax-Milgram)**.** *Sei $(V,(\cdot,\cdot)_V)$ ein Hilbert-Raum und s eine V-koerzitive Sesquilinearform. Dann existiert für alle antilinearen, stetigen Funktionale f eine eindeutige Lösung $u \in V$ des Variationsproblems*

$$s(u,v) = f(v), \qquad v \in V,$$

und die Lösung hängt stetig von f ab.

Beweis. Siehe z.B. [Alt99]. □

Definition 2.3 (Eigenwertproblem)**.** *$(\lambda,u) \in \mathbb{C}\times H^2(\Omega)\setminus\{0\}$ heisst ein Eigenpaar, wenn es eine Lösung das Eigenwertproblems*

$$-\Delta u = \lambda p\, u \qquad \text{auf } \Omega, \tag{2.5a}$$

$$\frac{\partial u}{\partial n} = 0 \qquad \text{auf } \partial\Omega \tag{2.5b}$$

ist.

In variationeller Form suchen wir Lösungen $(\lambda,u) \in \mathbb{C}\times H^1(\Omega)\setminus\{0\}$ von

$$a(u,v) = \lambda b(u,v), \qquad v \in H^1(\Omega), \tag{2.6}$$

mit den Definitionen aus (2.2).

(2.3) ist genau dann eindeutig lösbar, wenn κ^2 kein Eigenwert von (2.6) ist. Dazu hier einige Grundlagen, welche [DL90, Chapt. VIII] entnommen sind.

Definition 2.4 (Spektrum)**.** *Seien $(V,(\bullet,\bullet))$ ein komplexer Hilbert-Raum, $A : V \supset \mathcal{D}(A) \to V$ ein linearer Operator und* id *die Identität in V. Dann heisst $\rho(A) \subset \mathbb{C}$ die Resolventenmenge von A und für jedes $\lambda \in \rho(A)$*

$$R_\lambda(A) := (A - \lambda\,\mathrm{id})^{-1} : (A-\lambda\,\mathrm{id})(\mathcal{D}(A)) \to \mathcal{D}(A)$$

die Resolvente von A, wenn

1. *das Bild $(A-\lambda\,\mathrm{id})(\mathcal{D}(A))$ dicht ist in V,*
2. *die Inverse $R_\lambda(A)$ existiert und*

3. *$R_\lambda(A)$ ist stetig.*

Das Spektrum $\sigma(A) := \mathbb{C} \setminus \rho(A)$ von A wird unterteilt in ein Punktspektrum

$$\sigma_\mathrm{p}(A) := \{\lambda \in \sigma(A) \mid A - \lambda \,\mathrm{id} \text{ ist nicht invertierbar}\},$$

ein kontinuierliches Spektrum

$$\sigma_\mathrm{c}(A) := \{\lambda \in \sigma(A) \mid R_\lambda(A) \text{ ist unstetig und } \overline{(A - \lambda \,\mathrm{id})(\mathcal{D}(A))} = V\}$$

und in das Residualspektrum

$$\sigma_\mathrm{r}(A) := \{\lambda \in \sigma(A) \mid A - \lambda \,\mathrm{id} \text{ ist invertiertbar, aber } \overline{(A - \lambda \,\mathrm{id})(\mathcal{D}(A))} \neq V\}.$$

Wenn keine Unklarheiten zu befürchten sind, schreiben wir R_λ an Stelle von $R_\lambda(A)$.

Satz 2.5. *Zusätzlich zu den Voraussetzugnen aus Def. 2.4 sei A abgeschlossen. Dann ist*

$$\rho(A) = \{\lambda \in \mathbb{C} \mid R_\lambda \in L(V)\},$$

wobei $L(V)$ den Raum der linearen, stetigen Abbildungen auf V bezeichent.

Beweis. $\{\lambda \in \mathbb{C} \mid R_\lambda \in L(V)\} \subset \rho(A)$ ist trivial. Sei also $\lambda \in \rho(A)$. Zu zeigen ist $(A - \lambda \,\mathrm{id})(\mathcal{D}(A)) = V$. Da $(A - \lambda \,\mathrm{id})(\mathcal{D}(A))$ dicht in V ist, existiert zu jedem $y \in V$ eine Urbildfolge $x_n \in \mathcal{D}(A)$, sodass

$$\lim_{n\to\infty} (A - \lambda \,\mathrm{id})(x_n) = y \quad \text{in } V.$$

Aus der Stetigkeit von R_λ folgt

$$\|x_n - x_m\|_X \leq C\|(A - \lambda \,\mathrm{id})(x_n - x_m)\|$$

und daraus die Konvergenz $x_n \to x \in V$. Da $A - \lambda \,\mathrm{id}$ ebenso wie A abgeschlossen ist, folgt $x \in \mathcal{D}(A - \lambda \,\mathrm{id})$ und $y = (A - \lambda \,\mathrm{id})(x) \in (A - \lambda \,\mathrm{id})(\mathcal{D}(A))$. □

Satz 2.6. *Sei $(V, (\bullet, \bullet))$ ein unendlich-dimensionaler, komplexer Hilbert-Raum und $A \in L(V)$ ein kompakter, hermitescher Operator. Dann besteht $\sigma(A)\setminus\{0\}$ aus abzählbar vielen reellen Eigenwerten mit endlicher Vielfachheit, die eine Nullfolge bilden. Die Eigenräume $\ker(A - \lambda \,\mathrm{id})$ zu den Eigenwerten $\lambda \in \sigma_\mathrm{p}$ sind paarweise orthogonal und bilden eine Basis des Raumes V.*

Beweis. Siehe z.B. [DL90, Chapt. VIII, Theorem 3] oder [Kre99, Theorem 3.11]. □

Satz 2.7 (Fredholmsche Alternative)**.** *Sei $(V, (\bullet, \bullet))$ ein komplexer Hilbert-Raum, $A \in L(V)$ ein kompakter, hermitescher Operator, $f \in V$ und $\mu \in \mathbb{C}$ mit $\mu \neq 0$. Weiters seien (λ_j, v_j), $j = 0, 1, \ldots$ die orthonormalisierte Eigenpaare von A, wobei jeder Eigenwert seiner Vielfachheit nach aufgeführt wird. Dann gibt es zur Lösbarkeit von*

$$(A - \mu \,\mathrm{id})(u) = f. \tag{2.7}$$

zwei Alternativen:

1. *$\mu \notin \sigma(A)$: Dann existiert eine eindeutige Lösung $u \in V$ von (2.7), welche gegeben ist durch*
$$u = \sum_{j=0}^{\infty} \frac{(f, v_j)}{\lambda_j - \mu} v_j. \tag{2.8}$$

2. *$\mu \in \sigma(A)$: (2.7) ist genau dann lösbar, wenn $f \in \ker(A - \mu \,\mathrm{id})^{\perp}$. In diesem Fall gibt es unendlich viele Lösungen der Form*
$$u = \sum_{\substack{j=0 \\ \lambda_j \neq \mu}}^{\infty} \frac{(f, v_j)}{\lambda_j - \mu} v_j + u_\mu, \qquad u_\mu \in \ker(A - \mu \,\mathrm{id}). \tag{2.9}$$

Beweis. Folgt direkt aus Satz 2.6. □

Satz 2.8. *Sei $s(\bullet, \bullet)$ eine koerzitive, stetige, hermitesche Form auf dem Hilbertraum $(V, (\bullet, \bullet)_V)$, und V dicht und kompakt eingebettet in den Hilbert-Raum $(H, (\bullet, \bullet)_H)$. Dann ist der Operator $S : H \supset \mathcal{D}(S) \to H$ mit*
$$\begin{aligned} &s(u, v) = (Su, v)_H, \qquad u \in \mathcal{D}(S), \quad v \in V, \\ &\mathcal{D}(S) := \{u \in V \mid Su \in H\} \end{aligned} \tag{2.10}$$
selbstadjungiert und $\mathcal{D}(S)$ dicht in H. Weiterhin ist $\mathcal{D}(S)$ versehen mit der Graphennorm $\|u\|_{\mathcal{D}(S)} := \sqrt{(u, u)_H + (Su, Su)_H}$ ein Hilbert-Raum und die Einbettung von $\mathcal{D}(S)$ nach V ist stetig und nach H kompakt.

Beweis.

- Dichtheit: Sei $f \in H$ mit $(u, f)_H = 0$ für alle $u \in \mathcal{D}(S)$. Dann existiert nach dem Lax-Milgram Lemma ein $u_f \in V$ sodass $s(u, u_f) = (u, f)_H$ für alle $u \in \mathcal{D}(S)$. Es gilt
$$0 = (u, f)_H = s(u, u_f) = (Su, u_f)_H = (y, u_f)_H$$
mit $y := Su \in S(\mathcal{D}(S))$. Wiederum mit dem Lax-Milgram Lemma folgt $S(\mathcal{D}(S)) = H$ und daraus $u_f = f = 0$.

- Selbstadjungiert: $\mathcal{D}(S)$ ist dicht in H. Daher existiert ein eindeutiger, maximaler adjungierter Operator $S^* : \mathcal{D}(S^*) \to H$. Aus
$$(Su, v)_H = s(u, v) = \overline{s(v, u)} = \overline{(Sv, u)_H} = (u, Sv)_H, \qquad u, v \in \mathcal{D}(S)$$
folgt, dass S symmetrisch ist. Da $\mathcal{D}(S)$ maximal ist, gilt auch $\mathcal{D}(S) = \mathcal{D}(S^*)$.

- Abgeschlossenheit in der Graphennorm: Adjungierte Operatoren sind abgeschlossen und somit folgt die Behauptung aus $S^* = S$.

- Die stetige Einbettung in V folgt mit der Koerzitivität und der stetigen Einbettung von V in H:
$$\|v\|_V^2 \leq C s(v, v) = C(Sv, v)_H \leq C \|Sv\|_H \|v\|_H \leq \tilde{C} \|Sv\|_H \|v\|_V.$$

□

Satz 2.9. *Sei $(V,(\bullet,\bullet))$ ein unendlich-dimensionaler komplexer Hilbert-Raum, der kompakt in den Hilbert-Raum $(H,(\bullet,\bullet)_H)$ eingebettet ist und $s(\bullet,\bullet)$ ein stetige, koerzitive (mit Konstante $\alpha>0$), hermitesche Form auf $V\times V$. $S:H\supset\mathcal{D}(S)\to H$ sei der in (2.10) definierte Operator. Dann hat S ein reines Punktspektrum aus Eigenwerten λ_k, $k=0,...$ mit endlicher Vielfachheit, die sich nur im Unendlichen häufen, d.h.*

$$\sigma(S)=\sigma_{\mathrm{p}}(S)=\{\lambda_k\}_{k\in\mathbb{N}_0}\quad \text{mit } 0<\alpha\leq\lambda_k\overset{k\to\infty}{\longrightarrow}\infty.$$

Die gemäß der Vielfachheit der Eigenwerte aufgeführten Eigenvektoren $v_k\in V$, $k=0,...$, können als Orthonormalbasis von V oder auch von H gewählt werden.

Beweis. Aus der Koerzitivität folgt mit dem Lemma von Lax-Milgram die Existenz einer stetigen Inversen $S^{-1}:H\to\mathcal{D}(S)$, welche wegen des vorigen Satzes kompakt von H nach H ist. Somit sind die Voraussetzungen von Satz 2.6 und der Fredholmschen Alternative Satz 2.7 erfüllt, da

$$(S^{-1}u,v)=(S^{-1}u,S(S^{-1}v))=(S(S^{-1}u),S^{-1}v)=(u,S^{-1}v).$$

Sei nun $\lambda^{-1}\notin\sigma_{\mathrm{p}}(S^{-1})$, $f\in H$ und $g=S^{-1}f\in\mathcal{D}(S)$. Dann ist wegen der Fredholmschen Alternative Satz 2.7

$$\left(\frac{1}{\lambda}\,\mathrm{id}-S^{-1}\right)u=\frac{1}{\lambda}g$$

eindeutig lösbar. Aus $S^{-1}u$, $\lambda^{-1}g\in\mathcal{D}(S)$ schließen wir $u\in\mathcal{D}(S)$ und damit ist u auch die eindeutige Lösung der Gleichung

$$(S-\lambda\,\mathrm{id})u=f.$$

Somit folgt $\rho(S)\setminus\{0\}\subset\{\lambda\in\mathbb{C}\setminus\{0\}:\lambda^{-1}\notin\sigma_{\mathrm{p}}(S^{-1})\}$. Für $\lambda^{-1}\in\sigma_{\mathrm{p}}(S^{-1})$ mit Eigenvektor $v\in\ker(S^{-1}-\lambda^{-1}\,\mathrm{id})\subset H$ gilt $S^{-1}v=\lambda^{-1}v$ und damit $v\in\mathcal{D}(S)$. Nach Anwendung von S gilt also $Sv=\lambda v$, d.h. λ ist Eigenwert von S und v ein zugehöriger Eigenvektor. Die Aussage des Satzes folgt damit im Wesentlichen aus Satz 2.6.

Einzig zu zeigen ist noch, dass die Eigenvektoren dicht in V liegen. Dazu sei $\tilde{S}:V\to V$ der mit Hilfe des Rieszschen Darstellungssatzes eindeutig definierte Isomorphismus mit $s(u,v)=(\tilde{S}u,v)_V$, für $u,v\in V$. Sei nun $f\in V$ orthogonal zu allen Eigenfunktionen, d.h. $(v_k,f)_V=0$, $k=0,...$, und $g=\tilde{S}^{-1}f\in V$. Dann folgt

$$0=(f,v_k)_V=(\tilde{S}g,v_k)_V=s(g,v_k)=\overline{s(v_k,g)}=\overline{(Sv_k,k)_H}=\overline{\lambda_k(v_k,g)_H}$$

und daraus wegen $\overline{\lambda_k}\neq 0$, dass g in H orthogonal auf allen Eigenfunktionen steht. Wiederum mit Satz 2.6 ist $g=0$ und daraus $f=\tilde{S}g=0$. Die Orthogonalität zweier Eigenfunktionen zu paarweise verschiedenen Eigenwerten erhält man analog aus der Orthogonalität in H, indem anstelle von f eine Eigenfunktion verwendet wird. □

Korollar 2.10. *Das Helmholtz-Problem (2.3) ist genau dann eindeutig lösbar, wenn κ^2 kein Eigenwert des Eigenwertproblems (2.6) ist. Alle Eigenwerte von (2.6) sind nicht negativ und haben eine endliche Vielfachheit.*

Beweis. Folgt direkt aus (2.4) und dem vorigen Satz mit $V := H^1(\Omega)$, $H := L^2(\Omega)$, $s(\bullet,\bullet) := a(\bullet,\bullet) + \epsilon b(\bullet,\bullet)$ aus (2.2) mit $\epsilon > 0$. □

Allgemeiner gilt: Sei $V \hookrightarrow H \hookrightarrow V'$ ein Gelfand-Triple und $s(\bullet,\bullet) : V \times V \to \mathbb{C}$ eine stetige, hermitesche Sesquilinarform, welche die Gårding-Ungleichung

$$|s(u,u) + \gamma(u,u)_H| \geq \alpha(u,u)_V, \qquad u \in V, \tag{2.11}$$

mit $\alpha, \gamma > 0$ erfüllt. Dann gelten die Aussagen aus Satz 2.9 mit dem Spektralverschiebung

$$\sigma(S) = \sigma_{\mathrm{p}}(S) = \{\lambda_k\}_{k\in\mathbb{N}_0} \quad \text{mit } \alpha - \gamma \leq \lambda_k \overset{k\to\infty}{\longrightarrow} \infty.$$

Zur Charakterisierung der Eigenwerte kann im Falle von hermiteschen Sequilinearformen der Rayleigh-Quotient verwendet werden.

Lemma 2.11. *Sei $(V,(\bullet,\bullet))$ ein unendlich-dimensionaler komplexer Hilbert-Raum, der kompakt in den Hilbert-Raum $(H,(\bullet,\bullet)_H)$ eingebettet ist und $s(\bullet,\bullet)$ ein stetige, koerzitive, hermitesche Form auf $V \times V$ mit Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \dots$ (gemäß ihrer Vielfachheit aufgeführt) mit den zugehörigen Eigenfunktionen v_j, $j \in \mathbb{N}_0$, welche als Orthonormalbasis von H gewählt werden. Dann gilt*

$$\lambda_1 = \min_{v\in V\setminus\{0\}} \frac{s(v,v)}{(v,v)_H}, \qquad \lambda_k = \min_{\substack{v\in V\setminus\{0\}\\ (v,v_j)_H=0, j=1,\dots,k-1}} \frac{s(v,v)}{(v,v)_H}, \quad k = 2,\dots. \tag{2.12}$$

Beweis. Sei $v \in V$ beliebig. Dann gilt $v = \sum_{j=1}^{\infty}(v,v_j)_H v_j$ und es gilt

$$s(v,v) = \sum_{j=1}^{\infty}(v,v_j)_H \underbrace{s(v_j,v)}_{=\lambda_j(v_j,v)_H} = \sum_{j=1}^{\infty}\lambda_j|(v,v_j)_H|^2 \geq \lambda_1\sum_{j=1}^{\infty}|(v,v_j)_H|^2 = \lambda_1(v,v)_H.$$

Daraus folgen die Behauptungen, wobei die Minima jeweils von den Eigenfunktionen $v = v_k$ angenommen werden. □

Diese Darstellung lässt sich zumindest für die Eigenwerte λ_k mit $k \geq 2$ nicht so ohne weiteres zu einer Konvergenzanalysis verwenden. Folgende Charakterisierung ist da deutlich geeigneter.

Lemma 2.12. *Sei $(V,(\bullet,\bullet))$ ein unendlich-dimensionaler komplexer Hilbert-Raum, der kompakt in den Hilbert-Raum $(H,(\bullet,\bullet)_H)$ eingebettet ist und $s(\bullet,\bullet)$ ein stetige,*

koerzitive, hermitesche Form auf $V \times V$ mit Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ (gemäß ihrer Vielfachheit aufgeführt). Dann gilt

$$\lambda_k = \min_{\substack{E \subset V \\ \dim(E)=k}} \max_{v \in E \setminus \{0\}} \frac{s(v,v)}{(v,v)_H}, \qquad k = 1, \ldots. \tag{2.13}$$

Beweis. Sei $\{v_k, k \in \mathbb{N}\}$ die Orthonormalbasis (in H) aus den zugehörigen Eigenfunktionen. Für $v \in \hat{E} := \mathbf{span}\{v_1, \ldots, v_k\}$ folgt wie oben

$$s(v,v) = \sum_{j=1}^{k} \lambda_j |(v, v_j)_H|^2 \leq \lambda_k (v,v)_H$$

und daraus $\min \max s(v,v)/(v,v)_H \leq \lambda_k$, wobei für $v = v_k \in \hat{E} \setminus \{0\}$ Gleichheit gegeben ist.

Sei nun $E \neq \hat{E}$ mit $\dim E = k$. Dann existiert ein $v \in E \setminus \{0\}$ mit $(v, v_j)_H = 0$ für $j = 1, \ldots, k$. Für dieses v folgt $s(v,v) \geq \lambda_k (v,v)_H$. Das heisst für beliebige solche E ist $\max s(v,v)/(v,v)_H \geq \lambda_k$. □

3. Finite Elemente Methode

Die Grundlage konformer Finite Elemente Methoden (FEM) bildet die Lösung des Variationsproblems

$$\text{Suche } u \in V: \qquad s(u,v) = f(v), \qquad v \in V, \tag{3.1}$$

auf endlich-dimensionalen Unterräumen $V_h \subset V$ des Hilbertraumes $(V, (\bullet,\bullet)_V)$:

$$\text{Suche } u_h \in V_h: \qquad s(u_h,v_h) = f(v_h), \qquad v_h \in V_h. \tag{3.2}$$

Zur Konstruktion geeigneter Unterräume sei hier auf [BS08] verwiesen. Typischerweise wird nach Ciarlet ein einzelnes Finites Element als Triple aus einem hinreichend glatt berandeten Gebiet T mit nicht-leerem Inneren, einem Funktionenraum P und einer Basis des Dualraums P' bezeichnet (siehe [BS08, Def. 3.1.1]). Dies ermöglicht die Definition eines lokalen Interpolationsoperators für jedes Finite Element und unter gewissen Voraussetzungen auch die Abschätzung des lokalen Interpolationsfehlers (siehe [BS08, Def. 4.4.4]). Weiters wird das Rechengebiet Ω so in finite Elemente zerlegt, dass jedes Element affin interpolations-äquivalent zu einem gegegebenen Referenzelement ist. Dann kann man unter weiteren Voraussetzungen folgende Abschätzung für den Interpolationsfehler herleiten (siehe [BS08, (4.4.28)])

$$\|v - \mathcal{I}_h\, v\|_{H^s(\Omega)} \le C h^{m-s} |v|_{H^m(\Omega)}, \qquad v \in H^m(\Omega). \tag{3.3}$$

$|\bullet|_{H^m(\Omega)}$ bezeichnet dabei die m-te Sobolev Halbnorm. Die Konstante $C > 0$ hängt nicht vom maximalen Durchmesser h der Finiten Elemente ab. Eine wesentliche Voraussetzung ist jedoch, dass der verwendete Funktionenraum die Polynome bis zum Grade $m-1$ enthält, d.h. $\Pi_{m-1} \subset P$. Da $\mathcal{I}_h : V \to V_h$, kann der Bestapproximationsfehler in $V := H^1(\Omega)$ bei Verwendung von lokalen Polynomräumen des Grades p abgeschätzt werden durch

$$\inf_{v_h \in V_h} \|v - v_h\|_V \le C h^p |v|_{H^{p+1}(\Omega)}, \qquad v \in H^{p+1}(\Omega). \tag{3.4}$$

Da $H^2(\Omega)$ für beschränkte Gebiete Ω dicht in $H^1(\Omega)$ ist, folgt insbesondere für eine Familie von V_h mit $h \to 0$

$$\lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0, \qquad v \in H^1(\Omega). \tag{3.5}$$

Für V-koerzive Sesquilinearformen folgt daraus bereits die Konvergenz von u_h gegen u aus dem Céa-Lemma:

Lemma 3.1 (Céa). *Sei s eine koerzitive ($|s(u,u)| \geq \alpha(u,u)_V$ mit $\alpha > 0$) und stetige $|s(u,v)| \leq C\|u\|_V\|v\|_V$ mit $C > 0$ Sesquilinearform. Dann gilt für die Lösungen $u \in V$ und $u_h \in V_h$ von (3.1) bzw. (3.2)*

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V. \tag{3.6}$$

Beweis. Es gilt die Garlerkin-Orthogonalität für konforme ($V_h \subset V$) Galerkin-Verfahren, d.h.

$$s(u - u_h, v_h) = f(v_h) - f(v_h) = 0, \qquad v_h \in V_h, \tag{3.7}$$

und daraus mit $v_h - u_h \in V_h$ für beliebiges $v_h \in V_h$

$$\alpha\|u - u_h\|_V^2 \leq |s(u - u_h, u - u_h)| \leq |s(u - u_h, u - v_h)| \leq C\|u - u_h\|_V\|u - v_h\|_V.$$

□

Wir bereits gesehen, ist bei Helmholtz-Problemen die Sequilinearform $s(\bullet,\bullet) := a(\bullet,\bullet) - \kappa^2 b(\bullet,\bullet)$ aus (2.3) jedoch nicht koerzitiv. Aber es existiert eine Konstante $K > 0$, sodass die Gårding Ungleichung

$$\left|s(u,u) + K(u,u)_{L^2(\Omega)}\right| \geq \alpha\|u\|_{H^1(\Omega)}^2, \qquad u \in H^1(\Omega) \tag{3.8}$$

für ein $\alpha > 0$ erfüllt ist.

Satz 3.2. *Sei $s : H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$ eine hermitesche Sesquilinearform mit folgenden Eigenschaften:*

- $|s(u,v)| \leq C\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}$,
- $\left|s(u,u) + K(u,u)_{L^2(\Omega)}\right| \geq \alpha\|u\|_{H^1(\Omega)}^2$,
- *die Lösungen u des adjungierten Variationsproblems $s(v,u) = (v,\tilde{f})_{L^2(\Omega)}, v \in H^1(\Omega)$, mit $\tilde{f} \in L^2(\Omega)$ sind eindeutig und es gilt $|u|_{H^2(\Omega)} \leq C_R\|\tilde{f}\|_{L^2(\Omega)}$.*

Weiters seien $V_h \subset V := H^1(\Omega)$ für $h > 0$ endlich-dimensionale Ansatzräume, sodass (3.4) erfüllt ist mit $p = 1$.

Wenn das Variationsproblem (3.1) eindeutig lösbar ist, dann existiert ein $h_0 > 0$ sodass für alle $h \leq h_0$ das diskrete Problem (3.2) eindeutig lösbar ist und es gilt

$$\|u - u_h\|_V \leq \tilde{C} \min_{v_h \in V_h} \|u - v_h\|_V \tag{3.9}$$

mit einer Konstanten $\tilde{C} > 0$ unabhängig von h.

Beweis. Wir nehmen zunächst an, u_h seine eine Lösung zu (3.2) (die Existenz ist noch nicht sichergestellt). Mit Hilfe der Galerkin-Orthogonalität (3.7) und den Voraussetzungen an s folgt daraus wie beim Céa-Lemma

$$\begin{aligned} \alpha\|u - u_h\|_{H^1(\Omega)}^2 &\leq \left|s(u - u_h, u - u_h) + K(u - u_h, u - u_h)_{L^2(\Omega)}\right| \\ &\leq C\|u - u_h\|_{H^1(\Omega)}\|u - v_h\|_{H^1(\Omega)} + K\|u - u_h\|_{L^2(\Omega)}^2 \end{aligned} \tag{3.10}$$

für beliebige $v_h \in V_h$.

Wir verwenden nun Dualitätstechniken zum Abschätzen des zweiten Terms. Sei dazu $w \in V$ die eindeutige Lösung des adjungierten Problems $s(v, w) = (v, u - u_h)_{L^2(\Omega)}$, $v \in V$, mit $|w|_{H^2(\Omega)} \leq C_R \|u - u_h\|_{L^2(\Omega)}$. Dann folgt für beliebiges $w_h \in V_h$

$$\|u - u_h\|^2_{L^2(\Omega)} = s(u - u_h, w) = s(u - u_h, w - w_h) \leq C\|u - u_h\|_{H^1(\Omega)}\|w - w_h\|_{H^1(\Omega)}.$$

Wegen (3.4) für $p = 1$ folgt daraus für geeignetes w_h

$$\|u - u_h\|^2_{L^2(\Omega)} \leq CC_R C_a h\|u - u_h\|_{H^1(\Omega)}\|u - u_h\|_{L^2(\Omega)},$$

d.h. $\|u - u_h\|_{L^2(\Omega)} \leq CC_R C_a h\|u - u_h\|_{H^1(\Omega)}$. Einsetzen in (3.10) liefert

$$\left(\alpha - KC^2 C_R^2 C_a^2 h^2\right) \|u - u_h\|_V \leq C \min_{v_h \in V_h} \|u - v_h\|_V.$$

Da K, C, C_R und C_a unabhängig von h sind, folgt daraus die Existenz eines $h_0 > 0$, sodass für alle $h \leq h_0$ (3.9) mit einer Konstanten $\tilde{C}$ unabhängig von h erfüllt ist.

Bleibt zu zeigen, dass eine Lösung u_h von (3.2) in diesem Fall existiert. Da (3.2) ein lineares Problem auf einem endlich-dimensionalen Raum V_h ist, reicht es, die Eindeutigkeit zu zeigen. Sei also u_h eine Lösung von (3.2) mit $f = 0$. Da (3.1) eindeutig lösbar ist, folgt daraus $u = 0$ und damit wegen $0 \in V_h$ und (3.9) $u_h = 0$. Daraus folgt die Behauptung. □

Die Voraussetzungen an das adjungierte Problem lassen sich mit einer ausgefeilteren Beweistechnik vermeiden.

Satz 3.3. *Seien $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$ und $b : L^2(\Omega) \times L^2(\Omega) \to \mathbb{C}$ hermitesche, stetige und positiv definite Sesquilinearformen (im einfachsten Fall ist $a(\cdot, \cdot)$ das H^1-Skalarprodukt und $b(\cdot, \cdot)$ das L^2-Skalarprodukt). Weiter seien $(\lambda_j, u_j) \in \mathbb{R}_{>0} \times H^1(\Omega) \setminus \{0\}$ die Eigenpaare von (2.6) mit $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ und der Orthogonalbasis (bzgl. $a(\cdot, \cdot)$) aus den zugehörigen Eigenfunktionen $\{u_j : j \in \mathbb{N}\}$. Sei weiter $V_h \subset H^1(\Omega)$ mit $\dim V_h = N(h)$ eine Familie von endlich-dimensionalen Ansatzräumen sodass*

$$\inf_{v_h \in V_h} \|u_j - v_h\|_{H^1(\Omega)} \leq C_j h^p, \qquad h \to 0, \quad p > 0, \quad j = 1, \ldots, J. \tag{3.11}$$

Dann gilt für die diskreten Eigenwerte $\lambda_1^{(h)} \leq \cdots \leq \lambda_N^{(h)}$ von

$$a(u_j^{(h)}, v_h) = \lambda_j^{(h)} b(u_j^{(h)}, v_h), \qquad v_h \in V_h, \tag{3.12}$$

für hinreichend kleine $h \leq h_0$ die Abschätzung

$$0 \leq \frac{\lambda_k^{(h)} - \lambda_k}{\lambda_k} \leq Ch^{2p}, \qquad k = 1, \ldots, J, \tag{3.13}$$

mit einer Konstanten $C > 0$, welche nicht von h aber u.a. von den Eigenfunktionen $u_1, \ldots, u_k$ abhängt.

Beweis. Sei $k \in \{1, \dots, J\}$ fix und $E = \mathbf{span}\{u_1, \dots, v_k\}$. Dann gilt nach (2.13)

$$\lambda_k = \max_{v \in E \setminus \{0\}} \frac{a(v,v)}{b(v,v)}.$$

Sei $\Pi_h : H^1(\Omega) \to V_h$ der orthogonale Projektor bzgl. des Skalarproduktes $a(\cdot,\cdot)$, d.h. für $u \in H^1(\Omega)$ gilt

$$a(u - \Pi_h u, v_h) = 0, \quad v_h \in V_h.$$

Mit Hilfe der umgekehrten Dreiecksungleichung und der Stetigkeit von $a(\cdot,\cdot)$ folgt für alle $v = \sum_{j=1}^k a(v, v_j) v_j \in E$ aus (3.11)

$$\|\Pi_h v\|_{H^1(\Omega)} \geq \|v\|_{H^1(\Omega)} - \|v - \Pi_h v\|_{H^1(\Omega)} \geq (1 - C \left(\max_{j=1}^k C_j \right) h^p) \|v\|_{H^1(\Omega)}. \quad (3.14)$$

Somit ist Π_h als Abbildung von E nach $E_h := \Pi_h E$ für hinreichend kleine h bijektiv und es gilt $\dim E = \dim E_h = k$.

Sei nun $P : H^1(\Omega) \to E$ die orthogonale Projektion bzgl. des Skalarproduktes $a(\cdot,\cdot)$, d.h. für $u \in H^1(\Omega)$ gilt

$$a(u - Pu, v) = 0, \quad v \in E.$$

P ist auch orthogonaler Projektor bezüglich $b(\cdot,\cdot)$, da für $u - Pu = \sum_{j=k+1}^\infty \alpha_j u_j$ und beliebiges $v \in E$ gilt

$$b(u - Pu, v) = \sum_{j=k+1}^\infty \alpha_j b(u_j, v) = \sum_{j=k+1}^\infty \frac{\alpha_j}{\lambda_j} \underbrace{a(u_j, v)}_{=0} = 0.$$

Damit gilt sowohl für P bezüglich $a(\cdot,\cdot)$ und $b(\cdot,\cdot)$ der Satz des Pythargoras für $v = (v - Pv) + Pv \in V$:

$$a(v,v) = a(v - Pv, v - Pv) + a(Pv, Pv), \quad b(v,v) = b(v - Pv, v - Pv) + b(Pv, Pv)$$

und daraus

$$a(v,v) \leq a(Pv, Pv) + C\|v - Pv\|^2_{H^1(\Omega)} \quad \text{und} \quad b(v,v) \geq b(Pv, Pv).$$

Analog zu Π_h ist auch $P : E_h \to E$ für hinreichend kleine h bijektiv, da für alle $v_h = \Pi_h v \in E_h$ mit $v \in E$ ein $C > 0$ existiert mit

$$\|v_h - Pv_h\|_{H^1(\Omega)} \leq \tilde{C} \left(\|\Pi_h v - v\|_a + \underbrace{\|P\|_{a \to a}}_{=1} \|v - \Pi_h v\|_a \right) = C\|v\|_{H^1(\Omega)} h^p \quad (3.15)$$

und daraus folgt

$$\|Pv_h\|_{H^1(\Omega)} \geq \|\Pi_h v\|_{H^1(\Omega)} - \|v_h - Pv_h\|_{H^1(\Omega)} \geq \tilde{C}\|v\|_{H^1(\Omega)}. \quad (3.16)$$

Mit diesen Vorarbeiten sind wir in der Lage $\lambda_k^{(h)}$ abzuschätzen. Wegen (2.13) gilt

$$\begin{aligned}\lambda_k \leq \lambda_k^{(h)} \leq & \max_{v_h \in E_h \setminus \{0\}} \frac{a(v_h, v_h)}{b(v_h, v_h)} \leq \max_{v_h \in E_h \setminus \{0\}} \left\{ \frac{a(Pv_h, Pv_h)}{b(Pv_h, Pv_h)} \left(1 + C \frac{\|v_h - Pv_h\|^2_{H^1(\Omega)}}{a(Pv_h, Pv_h)} \right) \right\} \\ \leq & \max_{\tilde{v} \in E \setminus \{0\}} \left\{ \frac{a(\tilde{v}, \tilde{v})}{b(\tilde{v}, \tilde{v})} \right\} \max_{v_h \in E_h \setminus \{0\}} \left\{ 1 + \tilde{C} \left(\frac{\|v_h - Pv_h\|_{H^1(\Omega)}}{\|Pv_h\|_{H^1(\Omega)}} \right)^2 \right\} \\ \leq & \lambda_k \left(1 + \hat{C} h^{2p} \right).\end{aligned}$$

□

Die Abschätzung (3.13) ist für $k > 2$ nicht optimal, da der Projektionsfehler von allen Eigenfunktionen $v_1, \ldots, v_k$ eingeht. Sobald eine Eigenfunktion dabei ist, welche z.B. wegen fehlender Regularität eine schlechtere Konvergenzrate hat, wirkt sich das auch auf die größeren Eigenwerte zu anderen Eigenfunktionen aus. Dies lässt sich mit verfeinerten Beweistechniken vermeiden. Offen sind an dieser Stelle auch noch Abschätzungen für den Fehler in den Eigenfunktionen.

4. Arnoldi-Verfahren

Sei $\{b_1, \dots, b_N\}$ eine Basis des Raumes $V_h \subset H^1(\Omega)$ und $(\lambda_j, u_j) \in \mathbb{R}_{>0} \times V_h \setminus \{0\}$ ein Eigenpaar des diskreten Eigenwertproblems (3.12), wobei die dortigen Voraussetzungen an die Sesquilinearformen $a(\cdot,\cdot)$ und $b(\cdot,\cdot)$ gelten sollen. Weiter sei $u_j = \sum_{k=1}^{N} \alpha_j^{(k)} b_k$ mit $\alpha_j := (\alpha_j^{(1)}, \dots, \alpha_j^{(N)})^\top \in \mathbb{C}^N$. Dann ist (3.12) äquivalent zum Eigenwertproblem : Finde $(\lambda_j, \alpha_j) \in \mathbb{R}_{>0} \times \mathbb{C}^N \setminus \{0\}$ mit

$$A\alpha_j = \lambda_j B\alpha_j, \qquad j = 1, \dots, N, \tag{4.1}$$

wobei die Matrizen $A, B \in \mathbb{C}^{N\times N}$ mit $A_{jk} := a(b_j, b_k)$, $B_{jk} := b(b_j, b_k)$ für $j, k = 1, \dots, N$, dünn besetzt, symmetrisch und positiv definit sind. Im Folgenden werden wir numerische Löser für solche Eigenwertprobleme betrachten. Da die Raumdimension N und die Matrizen A und B konstant bleiben, unterdrücken wir in diesem Kapitel die Abhängigkeit von V_h.

In den meisten Fällen sind wir nicht an allen Eigenwerten sondern nur an wenigen Eigenwerten in der Nähe eines zu wählenden Parameters $\rho \in \mathbb{R}$ interessiert, der selber kein Eigenwert sein sollte. Dann ist (4.1) äquivalent zum Eigenwertproblem

$$\text{Finde } (\mu, \alpha) \in \mathbb{C} \times \mathbb{C}^N \setminus \{0\} : \quad C\alpha = \mu\alpha, \quad C := (A - \rho B)^{-1} B, \mu := \frac{1}{\lambda - \rho}. \tag{4.2}$$

Die betragsgrößten Eigenwerte μ von (4.2) führen dann zu den Eigenwerten $\lambda = \rho + 1/\mu$ von (4.1), die am nächsten an ρ liegen.

Im Folgenden werden wir das Lanczos-Verfahren für hermitesche Matrizen kennenlernen. $(A - \rho B)^{-1} B$ ist jedoch bezüglich des euklidischen Skalarproduktes nicht hermitesch. Eine Möglichkeit wäre, die Cholesky-Zerlegung $B = LL^*$ der hermitesch, positiv definiten Matrix B zu verwenden und (4.2) umzuschreiben in ein Problem mit einer hermiteschen Matrix:

$$\text{Finde } (\mu, y) \in \mathbb{R} \times \mathbb{C}^N \setminus \{0\} : \quad L^* (A - \rho B)^{-1} L y = \mu y, \qquad y := L^* \alpha.$$

In diesem Fall würden wir aber eine Cholesky-Zerlegung von B und eine Faktorisierung von $A - \rho B$ benötigen. Ersteres können wir vermeiden, denn $(A - \rho B)^{-1} B$ ist hermitesch bezüglich des durch B induzierten Skalarproduktes $(\cdot,\cdot)_B := (B\cdot,\cdot)_2$.Wir werden im Folgenden solche Eigenwertprobleme betrachten und orientieren uns dabei teilweise an [Saa11].

Lemma 4.1. *Sei $C \in \mathbb{C}^{N\times N}$ hermitesch bezüglich des Skalarproduktes $(\cdot,\cdot)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ die Eigenwerte von (4.2) (gemäß ihrer Vielfachheit gezählt) und*

$u_1, \dots, u_N$ die zugehörigen Eigenvektoren. Dann gilt

$$\lambda_1 = \max_{v \in \mathbb{C}^N \setminus \{0\}} \frac{(Cv, v)}{(v, v)}, \qquad \lambda_k = \max_{\substack{v \in \mathbb{C}^N \setminus \{0\} \\ (u_j, v) = 0, j = 1, \dots, k-1}} \frac{(Cv, v)}{(v, v)}, \quad k = 2, \dots, N \tag{4.3a}$$

und

$$\lambda_k = \max_{\substack{S \in \mathbb{C}^N \setminus \{0\} \\ \dim S = k}} \min_{v \in S \setminus \{0\}} \frac{(Cv, v)}{(v, v)}, \quad k = 1, \dots, N. \tag{4.3b}$$

Beweis. Siehe Lemma 2.11. □

4.1. Projektion auf Krylov-Raum

Definition 4.2 (Krylov-Raum)**.** *Sei $v_0 \in \mathbb{K}^N$ und $C \in \mathbb{K}^{N \times N}$. Dann bezeichnet*

$$\mathcal{K}_m(C, v_0) := \mathbf{span}\{v_0, Cv_0, \dots, C^{m-1}v_0\}, \qquad m \in \mathbb{N},$$

den Krylov-Raum von C und v. Wenn keine Unklarheiten zu befürchten sind, schreiben wir kurz $\mathcal{K}_m$. Mit $\mathcal{P}_m : \mathbb{K}^N \to \mathcal{K}_m$ bezeichnen wir den orthogonalen Projektor bezüglich des Skalarproduktes $(\cdot, \cdot)$.

Für Fehlerabschätzung wird wesentlich sein, welchen Abstand ein Eigenvektor u_i vom Krylov-Raum $\mathcal{K}_m$ hat. Diesen können wir als Winkel θ zwischen u_i und $\mathcal{K}_m$ wie folgt angeben:

$$\tan \theta(u_i, \mathcal{K}_m) = \frac{\|(\mathrm{id} - \mathcal{P}_m)u_i\|}{\|\mathcal{P}_m u_i\|}. \tag{4.4}$$

Wir werden diesen Abstand auf den Abstand von u_i zum Startvektor v_0 zurückführen. Dabei wird uns folgender Zusammenhang zu Polynomen sehr nützlich sein.

Lemma 4.3. *Sei Π_m der Raum der Polynome in einer Veränderlichen mit maximalem Grad m. Dann ist $v \in \mathcal{K}_m \subset \mathbb{K}^N$ genau dann wenn ein Polynom $p \in \Pi_{m-1}$ existiert mit $v = p(C)v_0$. Ist C diagonalisierbar mit Eigenwerten $\lambda_1, \dots, \lambda_N$ und zugehörigen Eigenvektoren $u_1, \dots, u_N$, dann existiert eine eindeutige Darstellung $v_0 = \sum_{j=1}^N \alpha_j u_j$ und es gilt*

$$v \in \mathcal{K}_m \quad \Leftrightarrow \quad \exists p \in \Pi_{m-1} : v = \sum_{j=1}^N p(\lambda_j)\alpha_j u_j.$$

Beweis. Nachrechnen. □

Definition 4.4 (Chebyshev-Polynome)**.** *Für $m \in \mathbb{N}_0$ sind die* Chebyshev-Polynome *$T_m \in \Pi_m$ definiert durch*

$$T_m(x) := \frac{1}{2}\left(\left(x + \sqrt{x^2 - 1}\right)^m + \left(x - \sqrt{x^2 - 1}\right)^m\right), \qquad x \in \mathbb{R}, \tag{4.5}$$

Es ist leicht nachzurechnen, dass T_m tatsächlich ein Polynom ist und dass es auf die Definition der (komplexen) Wurzel nicht ankommt. Mit Hilfe der binomischen Formel gilt nämlich

$$T_m(x) = \frac{1}{2}\sum_{k=0}^{m}\binom{m}{k}x^{m-k}\sqrt{x^2-1}^{k}\left(1+(-1)^k\right) = \sum_{k=0}^{\lfloor\frac{m}{2}\rfloor}\binom{m}{2k}x^{m-2k}(x^2-1)^k.$$

Es gibt weitere Darstellungen der Chebyshev-Polynome. Die Bekannteste ist

$$T_m(x) := \cos(m\arccos x), \qquad x \in [-1,1], \tag{4.6}$$

welche aus der Identität $T_m(\cos\varphi) = \cos(m\varphi)$ für $\varphi \in \mathbb{R}$ folgt.

Lemma 4.5. *Sei $[a,b]$ ein nicht-leeres Intervall in $\mathbb{R}$ und sei $c \geq b$. Dann gilt mit $\gamma := 1 + 2(c-b)/(b-a) > 0$*

$$\min_{\substack{p\in\Pi_m\\ p(c)=1}}\max_{x\in[a,b]}|p(x)| \leq \frac{1}{|T_m(\gamma)|} \leq \frac{1}{2}\left(\gamma+\sqrt{\gamma^2-1}\right)^{-m}. \tag{4.7}$$

Beweis. Wir verwenden die affin lineare Abbildung $\Psi : [-a,b] \to [-1,1]$ mit

$$\Psi(x) = 1 + 2\frac{x-b}{b-a}, \qquad x \in [a,b].$$

Dann folgt die erste Ungleichung aus (4.6) mit dem Polynom

$$p = \frac{T_m \circ \Psi}{|T_m(\Psi(c))|}$$

und die zweite Ungleichung aus der Definition der Chebyshev-Polynome (4.5). □

Für hermitesche Matrizen sind wir nun in der Lage, die Konvergenz des Projektionsverfahren nachzuweisen.

Satz 4.6 (Konvergenz Eigenwerte hermitescher Matrizen)**.** *Sei $C \in \mathbb{K}^{N\times N}$ eine hermitesche Matrix mit paarweise verschiedenen Eigenwerten $\lambda_1 > \lambda_2 > \cdots > \lambda_N$ und der zugehörigen Orthonormalbasis aus Eigenvektoren $u_1,\ldots,u_N$. Für $1 \leq m < N$ seien $\lambda_1^{(m)} \geq \cdots \geq \lambda_m^{(m)}$ die Eigenwerte der linearen, selbstadjungierten Abbildung $\mathcal{C}_m : \mathcal{K}_m(C,v_0) \to \mathcal{K}_m(C,v_0)$, welche durch $v \mapsto \mathcal{P}_m C v$ definiert ist. Der Startvektor v_0 des Krylov-Raumes sei nicht orthonogal zu einem der ersten $m-1$ Eigenvektoren $u_1,\ldots,u_{m-1}$.*

Dann gilt

$$0 \leq \lambda_i - \lambda_i^{(m)} \leq (\lambda_i - \lambda_N)\,(\tan\theta_i)^2\,\kappa_i^{(m)}\left(\frac{1}{T_{m-i}(\gamma_i)}\right)^2, \qquad i = 1,\ldots,m-1 < N, \tag{4.8}$$

wobei θ_i der Winkel zwischen u_i und v_0 ist, $\gamma_i := 1 + 2(\lambda_i - \lambda_{i+1})/(\lambda_{i+1} - \lambda_N)$ und

$$\kappa_1^{(m)} = 1, \quad \kappa_i^{(m)} = \left(\prod_{j=1}^{i-1} \frac{\lambda_j^{(m)} - \lambda_n}{\lambda_j^{(m)} - \lambda_i} \right)^2, \qquad i = 2, \dots, m.$$

Beweis. Die erste Ungleichung folgt direkt aus (4.3b). Für die zweite Ungleichung betrachten wir zunächst nur den Fall $i = 1$ und verwenden (4.3a) für $\lambda_1^{(m)}$.

Sei $v_0 = \sum_{j=1}^N \alpha_j u_j$, wobei $\alpha_1, \dots, \alpha_m$ aufgrund der Voraussetzung nicht verschwinden. Dann gilt mit Lemma 4.3

$$\begin{aligned} \lambda_1 - \lambda_1^{(m)} =& \lambda_1 - \max_{v \in \mathcal{K}_m \setminus \{0\}} \frac{(Cv, v)}{(v, v)} = \min_{p \in \Pi_{m-1} \setminus \{0\}} \frac{((\lambda_1 - C)p(C)v_0, p(C)v_0)}{(p(C)v_0, p(C)v_0)} \\ =& \min_{p \in \Pi_{m-1} \setminus \{0\}} \frac{\sum_{j=2}^N (\lambda_1 - \lambda_j)|\alpha_j p(\lambda_j)|^2}{\sum_{j=1}^N |\alpha_j p(\lambda_j)|^2} \\ \leq& (\lambda_1 - \lambda_N) \frac{\sum_{j=2}^N |\alpha_j|^2}{|\alpha_1|^2} \min_{p \in \Pi_{m-1} \setminus \{0\}} \max_{j=2,\dots,N} \frac{|p(\lambda_j)|^2}{|p(\lambda_1)|^2}. \end{aligned} \tag{4.9}$$

Nun ist $\tan \theta_1 = \sqrt{\sum_{j=2}^N |\alpha_j|^2 / |\alpha_1|^2}$ und die Behauptung folgt aus (4.7) mit $[a, b] = [\lambda_N, \lambda_2]$ und $c = \lambda_1$.

Sei nun $i > 1$. Wegen (4.3a) muss das Maximum bzw. Minimum über diejenigen $p \in \Pi_{m-1} \setminus \{0\}$ gebildet werden, sodass $(p(C)v_0, u_j^{(m)}) = 0$ für $j = 1, \dots, i-1$, wobei $u_j^{(m)}$ die projezierten, orthonormierten Eigenvektoren zu den Eigenwerten $\lambda_j^{(m)}$ sind. Dazu entwickeln wir $v_0 \in \mathcal{K}_m$ in diese Eigenvektoren, d.h $v_0 = \sum_{l=1}^m \beta_l u_l^{(m)}$. Dann gilt

$$0 = (p(C)v_0, u_j^{(m)}) = \sum_{l=1}^m \beta_l (p(C)u_l^{(m)}, u_j^{(m)}) = \sum_{l=1}^m \beta_l p(\lambda_l^{(m)})(u_l^{(m)}, u_j^{(m)}) = \beta_j p(\lambda_j^{(m)}).$$

Seien zunächst $\beta_1, \dots, \beta_{i-1}$ nicht null. Dann folgt mit $\lambda_i - \lambda_j < 0$ für $j = 1, \dots, i-1$

$$\begin{aligned} \lambda_i - \lambda_i^{(m)} =& \min_{\substack{p \in \Pi_{m-1} \setminus \{0\} \\ p(\lambda_j^{(m)}) = 0, j = 1, \dots, i-1}} \frac{\sum_{j=1}^N (\lambda_i - \lambda_j)|\alpha_j p(\lambda_j)|^2}{\sum_{j=1}^N |\alpha_j p(\lambda_j)|} \\ \leq& \min_{\substack{p \in \Pi_{m-1} \setminus \{0\} \\ p(\lambda_j^{(m)}) = 0, j = 1, \dots, i-1}} \frac{\sum_{j=i+1}^N (\lambda_i - \lambda_j)|\alpha_j p(\lambda_j)|^2}{\sum_{j=1}^N |\alpha_j p(\lambda_j)|} \\ \leq& (\lambda_i - \lambda_N) \frac{\sum_{j=i+1}^N |\alpha_j|^2}{|\alpha_i|^2} \min_{q \in \Pi_{m-i} \setminus \{0\}} \max_{j=i+1,\dots,N} \left\{ \left| \prod_{l=1}^{i-1} \frac{\lambda_l^{(m)} - \lambda_j}{\lambda_l^{(m)} - \lambda_i} \right| \frac{|q(\lambda_j)|}{|q(\lambda_i)|} \right\}^2 \end{aligned}$$

wobei

$$p(\lambda) = \prod_{l=1}^{i-1} \frac{\lambda_l^{(m)} - \lambda}{\lambda_l^{(m)} - \lambda_i} q(\lambda), \qquad q \in \Pi_{m-i}.$$

Mit $\lambda_j > \lambda_N$ folgt die Behauptung aus (4.7) mit $[a, b] = [\lambda_N, \lambda_{i+1}]$ und $c = \lambda_i$.

Abschließend betrachten wir den Fall, dass ein oder mehrere $\beta_1, \dots, \beta_{i-1}$ verschwinden. Dann entfällt im Folgenden die zugehörige Nebenbedingung $p(\lambda_j^{(m)}) = 0$ und die Behauptung gilt sogar mit einem Chebyshev-Polynom höheren Grades. Beachte, dass für $\beta_j = 0$ gilt $\lambda_j^{(m)} = \lambda_{j+1}^{(m)}$. □

Dieser Konvergenzsatz alleine gibt noch keinen Aufschluss darauf, wie man das projezierte Problem effektiv lösen kann. Damit werden wir uns im nächsten Unterabschnitt beschäftigen. Aber er zeigt, dass zumindest für paarweise verschiedene Eigenwerte hermitescher Matrizen das Projektionsverfahren auf den Krylov-Raum exponentiell in der Dimension des Krylov-Raumes konvergiert, wobei die Konvergenzgeschwindigkeit vom größten zu den kleineren Eigenwerten abnehmend ist. Wenn man sich die Konstanten $\kappa_i^{(m)}$ anschaut, erkennt man aber Probleme, falls Eigenwerte nahe beieinander liegen. Denn da $\lambda_j^{(m)}$ gegen λ_j konvergiert, wird der Nenner für $\lambda_i \approx \lambda_j$ sehr groß.

4.2. Arnoldi-Verfahren

Numerisch lassen sich die Eigenpaare $(\lambda_j^{(m)}, u_j^{(m)})$ der projezierten Abbildung $\mathcal{C}_m : \mathcal{K}_m(C, v_0) \to \mathcal{K}_m(C, v_0)$ mit $v \mapsto \mathcal{P}_m C v$ relativ einfach berechnen. Sei dazu $\{v_0, \dots, v_{m-1}\}$ eine Orthonormalbasis von $\mathcal{K}_m$ bezüglich des Skalarproduktes $(\cdot, \cdot) = (B\cdot, \cdot)_2$ mit einer hermitesch, positiv definiten Matrix $B \in \mathbb{K}^{N\times N}$ und dem euklidischen Skalarprodukt $(\cdot, \cdot)_2$. Zur Vereinfachung haben wir hier zunächst angenommen, dass $\dim \mathcal{K}_m = m$ gilt. Weiter sei $V_m := (v_0, \dots, v_{m-1}) \in \mathbb{K}^{N\times m}$ die Matrix aus den Basisvektoren. Dann ist $u_j^{(m)} = V_m \tilde{u}_j^{(m)}$ mit $\tilde{u}_j^{(m)} \in \mathbb{C}^m$ und $(\lambda_j^{(m)}, \tilde{u}_j^{(m)})$ ein Eigenpaar von

$$H_m \tilde{u}_j^{(m)} = \lambda_j^{(m)} \tilde{u}_j^{(m)}, \qquad H_m := V_m^* B C V_m \in \mathbb{K}^{m\times m}. \tag{4.10}$$

Wie üblich ist dabei $V_m^* = \overline{V_m^\top}$ die adjungierte Matrix bezüglich des euklidischen Skalarproduktes.

Eine Orthogonalbasis von $\mathcal{K}_m$ kann mit Hilfe des Orthogonalisierungsverfahrens von Gram-Schmidt gewonnen werden (siehe Alg.4.7 für eine effiziente Umsetzung). Es ist leicht nachzurechnen, dass die dort erzeugten Vektoren $v_0, \dots, v_{m-1}$ eine Orthonormalbasis des $\mathcal{K}_m$ bilden und dass gilt

$$C V_j = V_j H_j + h_{j+1,j} v_j e_j^\top, \qquad j = 0, \dots, m, \tag{4.11}$$

wobei $e_j \in \mathbb{K}^N$ der j-te Einheitsvektor ist und H_j die obere Teilmatrix der im Algorithmus erzeugten Matrix H_m ist. Wegen $V_m^* B V_m = \mathrm{id}_{m\times m}$ und $V_m^* B v_m = 0$ folgt, dass der Algorithmus tatsächlich die projezierte Matrix H_m erzeugt und dass diese eine Hessenberg-Matrix ist. Letzteres ist von Vorteil bei der Berechnung der Eigenwerte von H_m mittels QR-Verfahren (siehe Alg. A.1 im Anhang).

Algorithmus 4.7 Arnoldi-Iteration

Input: $C \in \mathbb{K}^{n\times n}$, $v_0 \in \mathbb{K}^n$ mit $(v_0, v_0) = 1$.

for $j = 1, \dots, m$ **do**
  $w := Cv_{j-1}$
  for $l = 1, \dots, j$ **do**
    $h_{lj} := (w, v_{l-1})$
    $w := w - h_{lj}v_{l-1}$
  end for
  $h_{j+1,j} := \sqrt{(w, w)}$
  if $h_{j+1,j} = 0$ **then**
    STOP: Invarianter Unterraum gefunden
  else
    $v_j := w/h_{j+1,j}$
  end if
end for

Output: Orthonormalbasis $v_0, \dots, v_{m-1}$ von $\mathcal{K}_m$, wobei die projezierte Matrix $H_m \in \mathbb{K}^{m\times m}$ eine Hessenberg-Matrix mit Einträgen h_{lj} ist.

Lemma 4.8. *Sei $(\lambda_j^{(m)}, \tilde{u}_j^{(m)})$ ein Eigenpaar von (4.10) und $u_j^{(m)} := V_m\tilde{u}_j^{(m)}$. Dann gilt*

$$\left\| \left(C - \lambda_j^{(m)} \,\mathrm{id}_{N\times N} \right) u_j^{(m)} \right\| = h_{m+1,m} |e_m^\top \tilde{u}_j^{(m)}|. \tag{4.12}$$

Bricht der Algorithmus 4.7 wegen $h_{j+1,j} = 0$ ab, so sind die Eigenwerte von H_j auch Eigenwerte von C.

Beweis. Die Behauptung ergibt sich aus einer Multiplikation von (4.11) von rechts mit $\tilde{u}_j^{(m)}$:

$$\begin{aligned} CV_m\tilde{u}_j^{(m)} &= V_mH_m\tilde{u}_j^{(m)} + h_{m+1,m}v_me_m^\top\tilde{u}_j^{(m)} \\ &= \lambda_j^{(m)}V_m\tilde{u}_j^{(m)} + h_{m+1,m}v_me_m^\top\tilde{u}_j^{(m)}. \end{aligned}$$

□

(4.12) kann als a-posteriori Fehlerabschätzung verwendet werden. Dazu muss der Algorithmus 4.7 insoweit erweitert werden, dass in jeder Iteration das kleine Eigenwertproblem gelöst wird. Ist für eine gewünschte Anzahl von Eigenpaaren das Residuum unterhalb einer vorgegebenen Toleranz, dann wird der Algorithmus abgebrochen. Man sollte bei diesem Vorgehen jedoch beachten, dass damit nur ein kleines Residuum garantiert wird. Möchte man den Fehler in den Eigenwerten kontrollieren, so ist folgendes Lemma hilfreich.

Lemma 4.9. *Sei $C \in \mathbb{K}^{N\times N}$ hermitesch und $(\lambda_j^{(m)}, \tilde{u}_j^{(m)})$ ein Eigenpaar von (4.10).Dann existiert ein Eigenwert λ_l von C mit*

$$|\lambda_j^{(m)} - \lambda_l| \le h_{m+1,m} \frac{|e_m^\top \tilde{u}_j^{(m)}|}{\|\tilde{u}_j^{(m)}\|_2}. \tag{4.13}$$

Beweis. Seien $\lambda_1, \dots, \lambda_N$ die Eigenwerte von C. Die Aussage ist trivialerweise erfüllt, wenn $\lambda_j^{(m)}$ ein Eigenwert von C ist. Andernfalls ist $C - \lambda_j^{(m)}\,\mathrm{id}$ invertierbar, $\left(C - \lambda_j^{(m)}\,\mathrm{id}\right)^{-1}$ ist hermitesch mit Eigenwerten $\left(\lambda_l - \lambda_j^{(m)}\right)^{-1}$ und es gilt für alle $z \in \mathbb{K}^N$

$$\| \left(C - \lambda_j^{(m)}\,\mathrm{id}\right)^{-1} z\| \leq \|z\| \max_{l=1}^{N} |\lambda_l - \lambda_j^{(m)}|^{-1}.$$

Folglich gilt für $u_j^{(m)} := V_m \tilde{u}_j^{(m)}$ und $z := Cu_j^{(m)} - \lambda_j^{(m)} u_j^{(m)}$ die Ungleichung

$$\max_{l=1}^{N} |\lambda_l - \lambda_j^{(m)}|^{-1} \geq \frac{\left\| u_j^{(m)} \right\|}{\left\| \left(C - \lambda_j^{(m)}\,\mathrm{id}_{N\times N}\right) u_j^{(m)} \right\|} = \frac{\|\tilde{u}_j^{(m)}\|_2}{h_{m+1,m} |e_m^\top \tilde{u}_j^{(m)}|}$$

□

Im letzten Lemma wurde bereits angenommen, dass C hermitesch bezüglich des Skalarproduktes $(\cdot,\cdot)$ ist. Für $(\cdot,\cdot) = (B\cdot,\cdot)_2$ mit der hermitesch, positiv definiten Matrix B bdeutet dies, dass $BC = C^*B$ und damit $H_m = H_m^*$ gilt. H_m ist also hermitesch bezüglich des euklidischen Skalarproduktes und dadurch eine Tridiagonalmatrix mit positiven Einträgen in den Nebendiagonalen. Dadurch kann Alg. 4.7 deutlich vereinfacht werden, da die Schleife über $l = 1, \dots, j$ nur für $l = j$ nicht-triviale, neue Einträge liefert. Diese Variante wird Lanczos-Verfahren genannt. Eine Orthonormalisierung (z.B. durch Gram-Schmidt) erfolgt dabei implizit, sofern alle Rechenoperationen exakt ausgeführt werden. Da dies in aller Regel nicht möglich ist, sind die durch das Lanczos-Verfahren erzeugten Basisvektoren von $\mathcal{K}_m$ leider häufig nicht mehr orthonogal, sodass eine zusätzliche Reorthogonalisierung wie im Arnoldi-Verfahren sinnvoll ist.

4.3. Konvergenz von Eigenvektoren hermitescher Matrizen

Zur Konvergenz von Eigenvektoren betrachten wir zunächst folgendes, algemeines Resultat.

Satz 4.10. *Sei $C \subset \mathbb{K}^{N\times N}$ eine hermitesche Matrix und (λ, u) ein beliebiges Eigenpaar von C mit $\|u\| = 1$. Weiter sei $\mathcal{P} : \mathbb{K}^N \to P \subset \mathbb{K}^N$ eine orthogonale Projektion auf den Unterraum P und $(\tilde{\lambda}_j, \tilde{u}_j)$ die Eigenpaare der projezierten Abbildung $P \to P$ definiert durch $x \mapsto \mathcal{P}Cx$ für $x \in P$. Zur Vereinfachung nehmen wir an, dass die Eigenwerte $\tilde{\lambda}_j$ paarweise verschieden sein sollen.*

Dann existiert zu (λ, u) ein approximierendes Eigenpaar $(\tilde{\lambda}, \tilde{u})$ mit minimalem Abstand $|\tilde{\lambda} - \lambda|$, sodass für den Winkel zwischen u und $\tilde{u}$ gilt

$$\sin\theta(u, \tilde{u}) \leq \sqrt{1 + \frac{\gamma^2}{\delta^2}} \sin\theta(u, P), \tag{4.14}$$

wobei $\theta(u, P)$ der Winkel zwischen u und dem Unterraum P, $\gamma := \|\mathcal{P}C(\mathrm{id} - \mathcal{P})\|$ und $\delta > 0$ der Abstand zwischen λ und den anderen Eigenwerten $\tilde{\lambda}_j$ ist.

Beweis. Sei ϕ der Winkel zwischen u und $\mathcal{P}u$, d.h. $\phi = \theta(u, P)$. Falls $u \in P$, so ist die Aussage trivial. Andernfalls gilt wegen $\|u\| = 1$

$$u = \underbrace{\|\mathcal{P}u\|}_{=\cos\phi} \underbrace{\frac{\mathcal{P}u}{\|\mathcal{P}u\|}}_{=:v} + \underbrace{\|u - \mathcal{P}u\|}_{=\sin\phi} \underbrace{\frac{u - \mathcal{P}u}{\|u - \mathcal{P}u\|}}_{=:w}$$

und daraus mit $Cu = \lambda u$, $\|w\| = 1$, $(\mathrm{id} - \mathcal{P})w = w$ und $Pw = 0$

$$\|\mathcal{P}(C - \lambda\,\mathrm{id})v\| \cos\phi = \|\mathcal{P}(C - \lambda\,\mathrm{id})w\| \sin\phi \leq \gamma \sin\phi. \tag{4.15}$$

Für den Vektor $v \in P$ auf der linken Seite verwenden wir eine weitere orthonogale Zerlegung. Sei dazu $\tilde{u} \in P$ mit $\|\tilde{u}\| = 1$ ein Eigenvektor zu $\tilde{\lambda}$ und ω der Winkel zwischen v und $\tilde{u}$ sodass wie oben gilt

$$v = \tilde{u} \cos\omega + z \sin\omega \tag{4.16}$$

mit einem normierten Vektor $z \in P$, welcher in P orthogonal zu $\tilde{u}$ ist. Betrachten wir nun die Abbildung vom Orthogonalraum $\{y \in P : (y, \tilde{u}) = 0\}$ in den Raum P, welche durch $y \mapsto \mathcal{P}(C - \lambda\,\mathrm{id})y$ definiert ist. Diese ist hermitesch mit den Eigenwerten $\{\tilde{\lambda}_j - \lambda, \tilde{\lambda}_j \neq \tilde{\lambda}\}$. Daher gilt

$$\|\mathcal{P}(C - \lambda\,\mathrm{id})z\| \geq \delta > 0.$$

Da auch $\mathcal{P}(C - \lambda\,\mathrm{id})z$ orthogonal zu $\tilde{u}$ ist, folgt aus (4.16) mit Hilfe des Satzes des Pythargoras

$$\|\mathcal{P}(C - \lambda\,\mathrm{id})v\|^2 = |\tilde{\lambda} - \lambda|^2 \cos^2\omega + \sin^2\omega \|\mathcal{P}(C - \lambda\,\mathrm{id})z\|^2 \geq \delta^2 \sin^2\omega.$$

Zusammen mit (4.15) erhalten wir

$$\cos\phi \sin\omega \leq \frac{\gamma}{\delta} \sin\phi.$$

Wegen $\phi = \theta(u, P)$ bleibt noch, den Winkel $\theta(u, \tilde{u})$ zu berechnen. Betrachten wir uns dazu nochmal die Konstruktion der Vektoren. Die erste Projektion erfolgte von u zu $\mathcal{P}u = v \cos\phi$. Weiter wurde v projeziert auf $\tilde{u} \cos\omega$, insgesamt also die orthogonale Projektion $u \mapsto \cos\phi \cos\omega \tilde{u}$. Wegen $\|u\| = \|\tilde{u}\| = 1$ folgt somit $\cos\theta(u, \tilde{u}) = \cos\phi\cos\omega$ und damit wegen

$$\begin{aligned}\sin^2\theta &= 1 - \cos^2\theta = 1 - \cos^2\phi\left(1 - \sin^2\omega\right) = \sin^2\phi + \sin^2\omega\cos^2\phi \\ &\leq \sin^2\phi\left(1 + \frac{\gamma^2}{\delta^2}\right)\end{aligned}$$

die Behauptung. □

Im Falle des Lanczos-Verfahrens folgt nun die Konvergenz der Eigenvektoren.

Satz 4.11. *Unter den Voraussetzungen der Sätze 4.6 und 4.10 gilt*

$$\sin\theta(u_i,\tilde{u}_i) \le \frac{\kappa_i\sqrt{1+\beta_m^2/\delta_m^2}}{T_{m-i}(\gamma_i)}\tan\theta(v_0,u_i), \tag{4.17}$$

wobei δ_m wie in Satz 4.10, γ_i wie in Satz 4.6, $\beta_m = h_{m+1,m}$ aus Alg. 4.7 und κ_i durch

$$\kappa_1 = 1, \quad \kappa_i = \prod_{j=1}^{i-1}\frac{\lambda_j-\lambda_n}{\lambda_j-\lambda_i}, \qquad i=2,\dots,m,$$

definiert sind.

Beweis. Betrachten wir zunächst die Konstante $\gamma = \|\mathcal{P}_m C(\mathrm{id}-\mathcal{P}_m)\|$ aus Satz 4.10. Für das durch B induzierte Skalarprodukt gilt $\mathcal{P}_m = V_m V_m^* B$ und somit wegen (4.11)

$$(\mathrm{id}-\mathcal{P}_m)C\mathcal{P}_m = (\mathrm{id}-\mathcal{P}_m)\left(V_mH_m+\beta_m v_m e_m^\top\right)V_m^*B = \beta_m v_m v_{m-1}^* B.$$

Da $((\mathrm{id}-\mathcal{P}_m)C\mathcal{P}_m)^* = \mathcal{P}_m C(\mathrm{id}-\mathcal{P}_m)$ und da die Spektralnorm der adjungierten Matrix gleich der Spektralnorm der ursprünglichen Matrix ist, folgt $\gamma=\beta_m$ und mit Hilfe von Satz 4.10

$$\sin\theta(u_i,\tilde{u}_i) \le \sqrt{1+\beta_m^2/\delta_m^2}\sin\theta(u_i,\mathcal{K}_m).$$

Ähnlich zum Beweis von Satz 4.6 lässt sich zeigen

$$\tan\theta(u_i,\mathcal{K}_m) \le \frac{\kappa_i}{T_{m-i}(\gamma_i)}\tan\theta(v_0,v_i)$$

und die Behauptung folgt aus $\sin\theta(u_i,\mathcal{K}_m)\le\tan\theta(u_i,\mathcal{K}_m)$. □

5. Approximationen von Eigenwertproblemen kompakter Operatoren

Im folgenden wird die von Babuška und Osborn in [BO91] entwickelte Theorie zur Konvergenz von Galerkin-Verfahren für Eigenwertproblemen dargestellt. Diese gilt zunächst für kompakte Operatoren auf Hilbert-Räumen (in der Originalarbeit reichen Banach-Räume). Wie wir bereits in Abschnitt 2 gesehen haben, kann man Eigenwertprobleme basierend auf dem Laplace-Operator auf den inversen Operator zurückführen. Dieser inverse Operator ist kompakt und für diesen ist die folgende Theorie anwendbar. Durch Kehrwertbildung erreicht man dann auch Resultate für die Laplace-Eigenwerte selber.

Für die meisten nachfolgenden Resultate ist nicht relevant, ob der Operator selbstadjungiert ist.

5.1. Projektion auf Eigenraum

In diesem Abschnitt benötigen wir ein wenig Theorie zu holomorphen, Operatorwertigen Funktionen. Sei dazu $A(\lambda) : V \to V$ für $\lambda \in \mathbb{C}$ eine Familie stetiger, linearer Operatoren im Banach-Raum V. Dann kann die Ableitung nach λ definiert werden durch den üblichen Grenzwert

$$A'(\lambda_0) = \lim_{\lambda \to \lambda_0} \frac{1}{\lambda - \lambda_0}(A(\lambda) - A(\lambda_0)),$$

wobei der Grenzwert im Sinne der Operatornorm auf V zu verstehen ist. Existiert dieser Grenzwert, so bezweichnet man $\lambda_0 \mapsto A(\lambda_0)$ als holomorph. Viele der Eigenschaften holomorpher Funktionen gelten auch für Operatorwertige, holomorphe Funktionen, inbesondere der Integralsatz von Cauchy. Details dazu finden sich z.B. in [Kal15].

Wir benötigen im Folgenden ein kleines Hilfslemma.

Lemma 5.1. *Sei $A(\lambda) : V \to V$ mit $\lambda \in \mathbb{C}$ eine Familie beschränkter linearer Operatoren und sei $\lambda \mapsto A(\lambda)$ stetig (holomorph). Ist $A(\lambda_0)$ für ein λ_0 stetig invertierbar,*

so existiert die Inverse $A(\lambda)^{-1}$ in einer hinreichend kleinen Umgebung von λ_0 und $\lambda \mapsto A(\lambda)^{-1}$ ist dort stetig (holomorph).

Beweis. Wir zeigen die Aussage für $\lambda \mapsto A(\lambda)$ stetig, d.h. wenn $A(\lambda + h) = A(\lambda) - B(h)$ mit $\|B(h)\| \to 0$ für $|h| \to 0$. Für hinreichend kleine $|h|$ ist dann $\|A(\lambda_0)^{-1}B(h)\| < 1$ und mit Hilfe der Neumannschen Reihe folgt

$$\begin{aligned} A(\lambda_0 + h)^{-1} &= \big(A(\lambda_0)(\mathrm{id} - A(\lambda_0)^{-1}B(h)\big)^{-1} = \left(\sum_{k=0}^{\infty} \big(A(\lambda_0)^{-1}B(h)\big)^k\right) A(\lambda_0)^{-1} \\ &= A(\lambda_0)^{-1} + A(\lambda_0)^{-1}B(h)A(\lambda_0)^{-1} + \sum_{k=2}^{\infty} \big(A(\lambda_0)^{-1}B(h)\big)^k A(\lambda_0)^{-1} \end{aligned}.$$

Wegen $\| \left(A(\lambda_0)^{-1}B(h)\right)^k \| \leq \|A(\lambda_0)^{-1}B(h)\|^k$ und $\|A(\lambda_0)^{-1}B(h)\| < 1$ folgt daraus die Stetigkeit von $\lambda_0 \mapsto A(\lambda_h)^{-1}$. □

Sei nun $A : V \to V$ ein stetiger, linearer Operator. $\Gamma \subset \rho(A)$ sei eine geschlossene, positiv orientierte Jordan-Kurve in der Resolventenmenge von A. Weiters sei auch das Innere mit Ausnahme eines einzelnen, isolierten Eigenwertes λ_0 in der Resolventenmenge. Dann heisst

$$P_{\lambda_0} := \frac{1}{2\pi \mathrm{i}} \int_\Gamma (A - \lambda \,\mathrm{id})^{-1} d\lambda \tag{5.1}$$

Riesz-Projekter (oder spektraler Projektor). Das Integral lässt sich dabei z.B. als Grenzwert bezüglich der Operatornorm von Riemann-Summen defnieren. Der Operator $P_{\lambda_0} : V \to V$ ist unabhängig von der Wahl der Kurve Γ, solange die oben angebenen Voraussetzungen bestehen bleiben. Dies folgt aus dem Ingralsatz von Cauchy für die holomorphe, Operator-wertige Funktion $\lambda \mapsto R_\lambda := (A - \lambda \,\mathrm{id})^{-1}$ und wird z.B. in [HS96, Lemma 6.1] nachgewiesen. Weitere Details zu diesen Dunford Integralen findet sich z.B. in [Yos74, Chapter VIII.].

Lemma 5.2. *P_{λ_0} ist ein stetiger Projektor mit abgeschlossenen Bild $P_{\lambda_0}(V)$, in dem der Eigenraum* $\ker(A - \lambda_0 \,\mathrm{id})$ *von λ_0 enthalten ist.*

Beweis. Aus der Definition von P_{λ_0} folgt mit Lemma 5.1

$$\|P_{\lambda_0}\| \leq \frac{1}{2\pi} \,\mathrm{length}(\Gamma) \sup_{\lambda \in \Gamma} \|R(\lambda)\| < \infty,$$

da Γ kompakt ist. Somit ist P_{λ_0} stetig.

Sei $\tilde{\Gamma}$ eine weitere geeignete Kurve um λ_0, die echt im Inneren von Γ enthalten ist. Dann gilt

$$P_{\lambda_0}^2 = (2\pi \mathrm{i})^{-2} \int_{\tilde{\Gamma}} \int_\Gamma (A - \mu \,\mathrm{id})^{-1} (A - \lambda \,\mathrm{id})^{-1} d\lambda d\mu.$$

Wegen

$$R(\mu) - R(\lambda) = R(\mu)\underbrace{(A - \lambda \operatorname{id})R(\lambda)}_{\operatorname{id}} - \underbrace{R(\mu)(A - \mu \operatorname{id})}_{\operatorname{id}} R(\lambda) = (\lambda - \mu)R(\mu)R(\lambda)$$

folgt

$$\begin{aligned} P_{\lambda_0}^2 &= (2\pi\mathfrak{i})^{-2} \int_{\tilde{\Gamma}} \int_{\Gamma} (\mu - \lambda)^{-1}(R(\mu) - R(\lambda))d\lambda d\mu \\ &= (2\pi\mathfrak{i})^{-2} \left(\int_{\tilde{\Gamma}} R(\mu) \left(\int_{\Gamma} (\mu - \lambda)^{-1} d\lambda \right) d\mu - \int_{\Gamma} R(\lambda) \left(\int_{\tilde{\Gamma}} (\mu - \lambda)^{-1} d\mu \right) d\lambda \right). \end{aligned}$$

Mit Hilfe des Residuensatzes

$$\int_{\Gamma} (\mu - \lambda)^{-1} d\lambda = 2\pi\mathfrak{i}, \qquad \int_{\tilde{\Gamma}} (\mu - \lambda)^{-1} d\mu = 0.$$

ergibt sich $P_{\lambda_0}^2 = P_{\lambda_0}$, d.h. P_{λ_0} ist eine Projektion.

Das Bild des stetigen Projektors P_{λ_0} ist abgeschlossen, da es das Urbild der abgeschlossenen Menge $\{0\}$ der stetigen Abbildung $\operatorname{id} - P_{\lambda_0}$ ist. Bleibt noch zu zeigen, dass alle $f \in \ker(A - \lambda_0 \operatorname{id})$ in $P_{\lambda_0}(V)$ liegen. Aus $(A - \lambda \operatorname{id})f = (\lambda_0 - \lambda)f$ folgert man

$$P_{\lambda_0} f = \frac{1}{2\pi\mathfrak{i}} \int_{\Gamma} (A - \lambda \operatorname{id})^{-1} f d\lambda = \frac{1}{2\pi\mathfrak{i}} \int_{\Gamma} (\lambda_0 - \lambda)^{-1} f d\lambda = f,$$

d.h. $f \in P_{\lambda_0}(V)$. □

Lemma 5.3. *Sei A selbstadjungiert. Dann ist P_{λ_0} eine orthogonale Projektion auf den Eigenraum* $\ker(A - \lambda_0 \operatorname{id})$.

Beweis. Der Beweis findet sich in [HS96, Prop. 6.3]. Wir zeigen hier nur, dass $P_{\lambda_0}(V) = \ker(A - \lambda_0 \operatorname{id})$. Nach dem vorigen Lemma ist jedoch nur noch $P_{\lambda_0}(V) \subset \ker(A - \lambda_0 \operatorname{id})$ zu zeigen. Wegen

$$(A - \lambda_0 \operatorname{id})(A - \lambda \operatorname{id})^{-1} = \operatorname{id} + (\lambda - \lambda_0)(A - \lambda \operatorname{id})^{-1}$$

gilt

$$(A - \lambda_0 \operatorname{id})P_{\lambda_0} = \frac{1}{2\pi\mathfrak{i}} \int_{\Gamma} (A - \lambda_0 \operatorname{id})(A - \lambda)^{-1} d\lambda = \frac{1}{2\pi\mathfrak{i}} \int_{\Gamma} (\lambda - \lambda_0)(A - \lambda)^{-1} d\lambda.$$

Nun ist die Funktion $\lambda \mapsto (\lambda - \lambda_0)(A - \lambda \operatorname{id})^{-1}$ im Inneren von Γ mit Ausnahme von λ_0 holomorph. Weiters ist sie dort gleichmäßig beschränkt, da für selbstadjungierte Operatoren A und alle $\lambda \in \rho(A)$ gilt

$$\|R(\lambda)\| \leq (\operatorname{dist}(\lambda, \sigma(A)))^{-1}.$$

Wenn nun also Γ hinreichend nah an λ_0 gewählt wird, sodass der Abstand zwischen Γ und $\sigma(A)$ immer durch den Abstand zwischen λ und λ_0 gegeben ist, so ist $\lambda \mapsto \|(\lambda - \lambda_0)(A - \lambda \operatorname{id})^{-1}\|$ durch 1 nach oben beschränkt. Damit muss $\lambda \mapsto (\lambda - \lambda_0)(A - \lambda \operatorname{id})^{-1}$ jedoch holomorph nach λ_0 fortsetzbar sein und es folgt $(A - \lambda_0 \operatorname{id})P_{\lambda_0} = 0$. □

Obiges Lemma ist zumindest teilweise auf nicht selbstadjungierte Operatoren fortsetzbar. Sei A dazu ein kompater Operator. Dann wissen wir aus der Riesz-Theorie, dass für alle $\lambda \in \mathbb{C} \setminus \{0\}$ eine Riesz-Zahl $r \in \mathbb{N}_0$ existiert, sodass die Folge der Nullräume stationär wird, d.h.

$$\{0\} = \ker(A-\lambda \operatorname{id})^0 \subset \ker(A-\lambda \operatorname{id})^1 \subset \cdots \subset \ker(A-\lambda \operatorname{id})^r = \ker(A-\lambda \operatorname{id})^{r+1} = \ldots.$$

Z.B. in [Kat95, Chapt.3, §3.6] wird gezeigt, dass dann $P_{\lambda_0}(V) = \ker(A-\lambda_0 \operatorname{id})^r$ gilt und endlich-dimensional ist. $\ker(A-\lambda_0 \operatorname{id})^r$ heisst dann verallgemeinerter Eigenraum. Für halbeinfache (engl. semisimple) Eigenwerte ist $r = 1$ und P_{λ_0} projeziert auf den Eigenraum von λ_0. Wie bei Matrizen üblich wird die Dimension des Eigenraums $\ker(A-\lambda_0 \operatorname{id})$ als geometrische Vielfachheit und die Dimension des verallgemeinerten Eigenraums $\ker(A-\lambda_0 \operatorname{id})^r$ als algebraische Vielfachheit bezeichnet.

Bemerkung 5.4. *Auf endlich-dimensionalen Räumen ist A eine Matrix aus $\mathbb{C}^{N\times N}$. Wenn diese Matrix hermitesch ist, so ist sie diagonalisierbar und der spektrale Projektor projeziert auf den Eigenraum. Das ist das Analogon zu Lemma 5.3. Falls A nicht diagonalisierbar ist, so definiert man auch dort verallgmeinerte Eigenräume und erhält Jordan-Blöcke. Dies ist genau das Verhalten von den Eigenwerten eines kompakten Operators. Mit Ausnahme der 0 verhält sich ein solcher Operator also wie eine Matrix, welche ja eh als kompakter (weil endlich-dimensionaler), linearer Operator aufgefasst werden kann.*

Bemerkung 5.5. *Die Theorie lässt sich unproblematisch verallgemeinern, wenn im Inneren der Kurve Γ endlich viele Eigenwerte liegen. Der spektrale Projektor projeziert dann auf die Summe der verallgemeinerten Eigenräume all dieser Eigenwerte.*

Lemma 5.6. *Der verallgemeinerte Eigenraum $\ker(A-\lambda_0 \operatorname{id})^r$ ist invariant gegenüber A und gegenüber einer beliebigen Resolvente $R(\lambda)$ mit $\lambda \in \rho(A)$.*

Beweis. Sei $f \in \ker(A-\lambda_0 \operatorname{id})^r$. Es gilt

$$(A-\lambda_0 \operatorname{id})^r Af = \sum_{j=0}^{r} \binom{r}{j} A^j (-\lambda_0)^{r-j} Af = A(A-\lambda_0 \operatorname{id})^r f = 0,$$

d.h. $Af \in \ker(A-\lambda_0 \operatorname{id})^r$ und $\ker(A-\lambda_0 \operatorname{id})^r$ ist invariant gegenüber A. Für die Resolvente folgt die Behauptung aus der Tatsache, dass $A-\lambda \operatorname{id}$ ein bijektive Abbildung von $\ker(A-\lambda_0 \operatorname{id})^r$ nach $\ker(A-\lambda_0 \operatorname{id})^r$ ist. □

Lemma 5.7. *Die Resolvente $R(\lambda)$ und damit auch der Projektor P_{λ_0} kommutieren mit A, d.h. $R(\lambda)A = AR(\lambda)$ und $P_{\lambda_0}A = AP_{\lambda_0}$.*

Beweis. Dies folgt aus

$$AR(\lambda) = R(\lambda)(A-\lambda \operatorname{id})AR(\lambda) = R(\lambda)A(A-\lambda \operatorname{id})R(\lambda) = R(\lambda)A.$$

□

5.2. Approximationstheorie für Eigenräume kompakter Operatoren

Sei V ein Banach-Raum, $A : V \to V$ ein kompakter Operator und $A_h : V \to V$ eine Folge kompakter Operatoren mit $\lim_{h\to 0} \|A_h - A\| = 0$.

Bemerkung 5.8. *Wenn wir mit $\Pi_h : V \to V_h$ die Projektion auf einen endlichdimensionalen Teilraum $V_h \subset V$ bezeichnen, dann wäre $A_h := \Pi_h A$ die Galerkin-Projektion von A. Falls V_h so gewählt wird, dass die Projektion für jedes $v \in V$ punktweise konvergiert, dann konvergiert A_h in der Operatornorm gegen A (pktw. Konvergenz + Kompakt = glm. Konvergenz).*

Lemma 5.9. *Für hinreichend kleine h ist $\Gamma \in \rho(A_h)$. Die Resolvente $R_h(\lambda) := (A_h - \lambda \,\mathrm{id})^{-1}$ ist gleichmäßig für $\lambda \in \Gamma$ beschränkt und konvergiert in der Operatornorm gleichmäßig für $\lambda \in \Gamma$ gegen $R(\lambda)$ für $h \to 0$.*

Beweis. Wegen

$$A_h - \lambda \,\mathrm{id} = A_h - A + A - \lambda \,\mathrm{id} = (A - \lambda \,\mathrm{id})\,(\mathrm{id} - R(\lambda)(A - A_h))$$

ist für h hinreichend klein (genauer $\|R(\lambda)(A - A_h)\| < 1$) der Operator $A_h - \lambda \,\mathrm{id}$ stetig invertierbar, d.h. $\Gamma \in \rho(A_h)$. Wie im Beweis von Lemma 5.1 folgt $R_h(\lambda) = R(\lambda) + B_h(\lambda)$ mit $\|B_h(\lambda)\| \to 0$ für $h \to 0$, wobei die Konvergenz gleichmäßig in $\lambda \in \Gamma$ ist. □

Wir können nun die spektralen Projektoren

$$P := \frac{1}{2\pi \mathrm{i}} \int_\Gamma R(\lambda) d\lambda, \qquad P_h := \frac{1}{2\pi \mathrm{i}} \int_\Gamma R_h(\lambda) d\lambda \tag{5.2}$$

definieren. $E := P(V)$ ist der verallgemeinerte Eigenraum zum Eigenwert λ_0 von A. $E_h := P_h(V)$ ist die Summe der verallgemeinerten Eigenräume zu den Eigenwerten von A_h, die innerhalb von Γ liegen.

Bemerkung 5.10. *Der Raum E_h enhält die Eigenfunktionen des projezierten Operators A_h. Im Gegensatz dazu haben wir im Beweis von Satz 3.3 mit E_h die orthogonale Projektion des kontinuierlichen Eigenraums in den diskreten Raum V_h bezeichnet. Diese beiden Definitionen stimmen nicht notwendig überein.*

Lemma 5.11. *Es gilt $\|P - P_h\| \to 0$ für $h \to 0$ und für hinreichend kleine h ist $\dim E = \dim E_h < \infty$.*

Beweis. Mit dem vorigen Lemma und

$$P - P_h = \frac{1}{2\pi \mathrm{i}} \int_\Gamma (R(\lambda) - R_h(\lambda))\, d\lambda$$

folgt die Konvergenz $P_h \to P$. Da A ein kompakter Operator ist, ist $\dim E$ endlich. Sei nun $v \in E_h$. Dann folgt aus der umgekehrten Dreiecksungleichung

$$\|Pv\| \geq \|v\| - \|Pv - P_h v\| \geq (1 - \|P - P_h\|)\,\|v\|,$$

dass P als Abbildung von E_h nach E für hinreichend kleine h injektiv ist. Aus dem Rangsatz linearer Abbildungen folgt daher $\dim E_h = \dim \ker P + \dim P(E_h) \leq \dim E$. Umgekehrt ist P_h als Abbildung von E nach E_h ebenfalls für hinreichend kleine h injektiv und es folgt $\dim E = \dim E_h$. □

Lemma 5.12. *Seien $M, N \subset V$ abgeschlossene Unterräume von V. Dann definieren wir den Abstand zwischen M und N durch*

$$\delta(M,N) := \sup_{x \in M, \|x\|=1} \operatorname{dist}(x,N), \qquad \hat{\delta}(M,N) := \max\{\delta(M,N), \delta(N,M)\}. \tag{5.3}$$

Falls $\dim M = \dim N < \infty$, so ist $\delta(N,M) \leq \delta(M,N)\,(1 - \delta(M,N))^{-1}$. Weiter ist $\delta(M,N) = \delta(N,M)$, wenn V ein Hilbert-Raum ist und $\hat{\delta}(M,N) < 1$ gilt.

Beweis. Siehe [Kat58, Lemma 213 und 221]. □

Satz 5.13. *Es existiert ein $h_0 > 0$ und eine Konstante $C > 0$ unabhängig von h, sodass für alle $h \leq h_0$ gilt*

$$\hat{\delta}(E, E_h) \leq C\,\|(A - A_h)\,|_E\|\,, \tag{5.4}$$

wobei $(A - A_h)\,|_E$ die Einschränkung von $A - A_h$ auf den verallgemeinerten Eigenraum E bezeichnet.

Beweis. Sei $v \in E$ mit $\|v\| = 1$. Dann gilt wegen

$$R(\lambda) - R_h(\lambda) = R_h(\lambda)(A_h - \lambda\,\mathrm{id})R(\lambda) - R_h(\lambda)(A - \lambda\,\mathrm{id})R(\lambda) = R_h(\lambda)(A_h - A)R(\lambda)$$

und mit Lemma 5.6 und Lemma 5.11

$$\begin{aligned}\|v - P_h v\| =& \|Pv - P_h v\| = \left\| \frac{1}{2\pi \mathrm{i}} \int_\Gamma (R(\lambda) - R_h(\lambda))\, v d\lambda \right\| \\ =& \frac{1}{2\pi} \left\| \int_\Gamma R(\lambda)(A_h - A)R(\lambda) v d\lambda \right\| \\ \leq& \frac{1}{2\pi} \operatorname{length}(\Gamma) \sup_{\lambda \in \Gamma} \|R_h(\lambda)\| \|\,(A - A_h)\,|_E\| \sup_{\lambda \in \Gamma} \|R(\lambda)\|.\end{aligned} \tag{5.5}$$

Da $R_h(\lambda)$ und $R(\lambda)$ für $\lambda \in \Gamma$ gleichmäßig in λ und in h beschränkt sind (sofern h hinreichend klein ist), existiert die geforderte Konstante $\hat{C} > 0$ mit

$$\delta(E, E_h) \leq \hat{C} \|\,(A - A_h)\,|_E\|.$$

Insbesondere folgt $\delta(E, E_h) \to 0$ für $h \to 0$. Da für hinreichend kleine h gilt $\dim E = \dim E_h$, folgt aus Lemma 5.12 $\delta(E_h, E) \leq \delta(E, E_h)(1 - \delta(E, E_h))^{-1} \leq 2\delta(E, E_h)$ und daraus die Behauptung mit $C = 2\hat{C}$. □

Falls $\dim E = 1$ gilt und wenn $A_h = \Pi_h A$, so ist der Abstand zwischen E und E_h beschränkt durch den Projektionsfehler $v - \Pi_h v$ für eine normierte Eigenfunktion $Av = \lambda_0 v$. In diesem Fall erhalten wir somit bereits die Konvergenz (im Sinne von $\hat{\delta}(E, E_h)$) der diskreten Eigenvektoren v_h von A_h gegen einen Eigenvektor v von A.

Korollar 5.14. *Unter den Voraussetzungen des vorigen Satzes gilt*

$$\|(P - P_h)|_E\| \leq C\|(A - A_h)|_E\|.$$

Beweis. Das ist (5.5). □

5.3. Das adjungierte Eigenwertproblem

Für die Konvergenz der Eigenwerte benötigen wir einige Aussagen zu den adjungierten Eigenwertproblemen. Sei dazu $(V, (\cdot, \cdot))$ ein Hilbert-Raum über $\mathbb{C}$ und $A : V \to V$ ein kompakter Operator. Sei weiter $A^* : V \to V$ der adjungierte Operator, d.h. für alle $u, v \in V$ gilt $(Au, v) = (u, A^* v)$.

Lemma 5.15. *Es gilt $\rho(A^*) = \overline{\rho(A)}$ und $\sigma(A^*) = \overline{\sigma(A)}$, d.h. Spektrum und Resolventenmenge des adjungierten Operators sind jeweils die kunjugierten Mengen des ursprünglichen Operators.*

Beweis. Wegen

$$(u, v) = ((A - \lambda \,\mathrm{id})R_A(\lambda)u, v) = (u, R_A^*(\lambda)(A^* - \overline{\lambda}\,\mathrm{id})v)$$

ist der adjungierte Operator $R_A^*(\lambda)$ zur Resolvente $R_A(\lambda)$ von A für $\lambda \in \rho(A)$ genau die Resolvente $R_{A^*}(\overline{\lambda})$ von A^*, d.h. $R_A^*(\lambda) = R_{A^*}(\overline{\lambda})$. Daraus folgt die Behauptung für die Resolventenmengen und damit auch für das Spektrum. □

Lemma 5.16. *Sei $\lambda_0 \in \sigma(A) \setminus \{0\}$ ein isolierter Eigenwert von A mit spektralem Projektor*

$$P = \frac{1}{2\pi \mathrm{i}} \int_\Gamma R_A(\lambda) d\lambda$$

und verallgemeinertem Eigenraum $\ker(A - \lambda_0 \,\mathrm{id})^r$ der Dimension m. Dann ist $\overline{\lambda_0} \in \sigma(A^*) \setminus \{0\}$ ein isolierter Eigenwert von A^* mit spektralem Projektor P^*. λ_0 und $\overline{\lambda_0}$ haben die gleichen algebraischen und geometrischen Vielfachheiten und der verallgemeinerte Eigenraum von $\overline{\lambda_0}$ ist gegeben durch $\ker(A^* - \overline{\lambda_0}\,\mathrm{id})^r = P^* \ker(A - \lambda_0 \,\mathrm{id})^r$.

Beweis. Γ ist eine positiv orientierte Kurve um λ_0, welche voll in $\rho(A)$ liegt und welche im Inneren nur λ_0 als Eigenwert hat. Dann ist $\overline{\Gamma}$ eine negativ orientierte Kurve um

$\overline{\lambda_0}$, welche voll in $\rho(A^*)$ liegt und welche im Inneren nur $\overline{\lambda_0}$ als Eigenwert hat. Der zugehörige spektrale Projektor ist dann gegeben durch

$$P(\overline{\lambda_0}, A^*) = -\frac{1}{2\pi \mathrm{i}} \int_{\overline{\Gamma}} R_{A^*}(\lambda) d\lambda = -\frac{1}{2\pi \mathrm{i}} \int_{\Gamma} R_{A^*}(\overline{\lambda}) d\lambda = -\frac{1}{2\pi \mathrm{i}} \int_{\Gamma} R_A^*(\lambda) d\lambda = P^*.$$

Sei nun ϕ_j für $j = 1, \dots, m$ eine Orthonormalbasis von $\ker(A - \lambda_0 \operatorname{id})^r$ und $\phi_j^* := P^* \phi_j$. Dann gilt für beliebiges $v \in V$

$$((A^* - \overline{\lambda_0} \operatorname{id})^r \phi_j^*, v) = (P^* \phi_j, (A - \lambda_0 \operatorname{id})^r v) = (\phi_j, (A - \lambda_0 \operatorname{id})^r P v) = 0,$$

da wegen Lemma 5.7 P und $(A - \lambda_0 \operatorname{id})^r$ kommutieren. Damit ist $\phi_j^* \in \ker(A^* - \overline{\lambda_0} \operatorname{id})^r$. Aus

$$(\phi_j, \phi_k^*) = (\phi_j, P^* \phi_k) = (P \phi_j, \phi_k) = (\phi_j, \phi_k) = \delta_{jk} \tag{5.6}$$

folgt die lineare Unabhängigkeit von $\phi_1^*, \dots, \phi_m^*$ und damit $\dim \ker(A^* - \overline{\lambda_0} \operatorname{id})^r \geq \dim \ker(A - \lambda_0 \operatorname{id})^r$. Durch Vertauschung der Rollen von A und A^* folgt, dass beide Räume die gleiche Dimension haben und damit $\ker(A^* - \overline{\lambda_0} \operatorname{id})^r = P^* \ker(A - \lambda_0 \operatorname{id})^r$.

Die Aussagen zu den Vielfachheiten folgen analog. □

5.4. Approximationstheorie für Eigenwerte kompakter Operatoren

Satz 5.17. *Sei $(V, (\cdot, \cdot))$ ein komplexer Hilbert-Raum und $A : V \to V$ ein kompakter Operator. Weiter sei $\Gamma \in \rho(A)$ eine geschlossene Jordan-Kurve, in deren inneren nur ein Eigenwert λ von A liegt. P sei der zugehörige spektrale Projektor auf den verallgemeinerten Eigenraum $E := \ker(A - \lambda \operatorname{id})^r$, $\phi_1, \dots, \phi_m$ eine Orthonormalbasis von E und $\phi_j^* = P^* \phi_j$, $j = 1, \dots, m$, eine Basis des Eigenraums E^* zum Eigenwert $\overline{\lambda_0}$ des adjungierten Operators A^*. Zuletzt sei $A_h : V \to V$ eine Folge kompakter Operatoren mit $\lim_{h \to 0} \|A_h - A\| = 0$ und $\lambda_j^{(h)}$ die Eignwerte von A_h im Inneren von Γ.*

Dann existierten für hinreichend kleine h genau m Eigenwerte $\lambda_1^{(h)}, \dots, \lambda_j^{(h)}$ (gemäß ihrer algebraischen Vielfachheit gezählt) und für das arithmetische Mittel

$$\hat{\lambda}_h := \frac{1}{m} \sum_{j=1}^{m} \lambda_j^{(h)}$$

gilt die Fehlerabschätzung

$$|\lambda - \hat{\lambda}_h| \leq \frac{1}{m} \sum_{j=1}^{m} |((A - A_h)\phi_j, \phi_j^*)| + C \|(A - A_h)|_E\| \|(A^* - A_h^*)|_{E^*}\|. \tag{5.7}$$

Beweis. Sei P_h die spektrale Projektion bezüglich der Kurve Γ von A_h und $E_h := P_h V$. Wegen Lemma 5.11 und der dortigen Beweistechnik giltdim $E_h = \dim E$ für

hinreichend kleine h und die Operatoren $\tilde{P}_h := P_h|_E : E \to E_h$ sind bijektiv. Somit ist $\tilde{P}_h^{-1} : E_h \to E$ wohldefiniert und stetig, da für $v \in E \setminus \{0\}$ und hinreichend kleine h gilt

$$\|v\| - \|\tilde{P}_h v\| = \|Pv\| - \|\tilde{P}_h v\| \leq \|(P - P_h)|_E\| \leq \frac{1}{2}\|v\| \quad \Rightarrow \|v\| \leq 2\|\tilde{P}_h v\|. \tag{5.8}$$

$\tilde{P}_h \tilde{P}_h^{-1}$ ist auf E_h die Identität und $\tilde{P}_h^{-1}\tilde{P}_h$ ist auf E die Identität. Wir definieren nun

$$T_h := \tilde{P}_h^{-1} A_h \tilde{P}_h : E \to E.$$

Da P_h und A_h wegen Lemma 5.7 permutieren, ist $T_h = \tilde{P}_h^{-1} P_h A_h|_E$. Zu beachten ist hierbei jedoch, dass E nicht invariant unter A_h ist, d.h. das Bild $A_h(E)$ liegt nicht notwendig in E. Daher ist $\tilde{P}_h^{-1} P_h$ auf $A_h(E)$ auch nicht notwendig die Identität.

Die algebraischen Vielfachheiten der Eigenwerte von T_h als lineare Abbildung von einem Raum der Dimension m in einen Raum der Dimension m müssen sich auf m summieren. Da E_h invariant unter A_h ist (Lemma 5.6), gilt für alle $r \in \mathbb{N}$

$$(T_h - \lambda \,\mathrm{id}_{E\to E})^r = \tilde{P}_h^{-1} (A_h - \lambda \,\mathrm{id})^r \tilde{P}_h.$$

Somit sind die Eigenwerte von T_h genau die Eigenwerte innerhalb von Γ der Operatoren A_h und die jeweiligen Vielfachheiten stimmen überein. (Sei z.B. $\lambda^{(h)}$ ein solcher Eigenwert und $\phi_1^{(h)}, \ldots, \phi_l^{(h)}$ mit $l \leq m$ eine Orthonormalbasis des zugehörigen verallgmeinerten Eigenraums. Dann sind $\tilde{P}_h^{-1}\phi_1^{(h)}, \ldots, \tilde{P}_h^{-1}\phi_l^{(h)}$ linear unabhängig und in $\ker(T_h - \lambda^{(h)} \,\mathrm{id})^l$.) Für $T := A|_E : E \to E$ gilt ebenso, dass λ der einzige Eigenwert von T ist und dass die Vielfachheiten von λ mit denen von A übereinstimmen.

Betrachten wir nun die Matrix-Darstellung von $T - T_h : E \to E$ in der Orthonormalbasis $\phi_1, \ldots, \phi_m$. Die Spur dieser Matrix ist gegeben durch die Summe ihrer Eigenwerte (gemäß der algebraischen Vielfachheit gezählt), da zueinander ähnliche Matrizen die gleiche Spur besitzen und jede Matrix ähnlich zu ihrer Jordan-Normalenform ist. Somit folgt

$$\lambda - \hat{\lambda}_h = \frac{1}{m} \mathbf{Spur}(T - T_h). \tag{5.9}$$

Zur Berechnung der Spur betrachten wir die Diagonaleinträge der Matrix, welche für $j = 1, \ldots, m$ gegeben sind durch

$$\begin{aligned}((T - T_h)\phi_j, \phi_j) =& (P(T - T_h)\phi_j, \phi_j) = ((T - T_h)\phi_j, P^*\phi_j) = ((T - T_h)\phi_j, \phi_j^*) \\ =& (A\phi_j - \tilde{P}_h^{-1} A_h \tilde{P}_h \phi_j, \phi_j^*) = (\tilde{P}_h^{-1} P_h (A - A_h)\phi_j, \phi_j^*) \\ =& ((A - A_h)\phi_j, \phi_j^*) + ((\tilde{P}_h^{-1} P_h - \mathrm{id})(A - A_h)\phi_j, \phi_j^*).\end{aligned} \tag{5.10}$$

Den letzten Term können wir noch weiter umschreiben. Sei dazu $L_h := \tilde{P}_h^{-1} P_h : V \to E$ die Projektion auf E mit Nullraum $\ker(P_h)$. Daher ist

$$P_h(L_h - \mathrm{id})v = 0$$

für beliebiges $v \in V$ und mit $v = (A - A_h)\phi_j$ erhalten wir

$$((\tilde{P}_h^{-1} P_h - \mathrm{id})(A - A_h)\phi_j, \phi_j^*) = ((L_h - \mathrm{id})(A - A_h)\phi_j, (P^* - P_h^*)\phi_j^*).$$

Wegen Lemma 5.9 ist die Resolvente $R_h(\mu)$ und damit auch die Projektion P_h gleichmäßig in h beschränkt. Gleiches gilt für $\tilde{P}_h^{-1}$ wegen (5.8). Somit existiert eine positive Konstante $\tilde{C} := \sup_{0<h\leq h_0} \|L_h - \mathrm{id}\|$, sodass mit Cor. 5.14 (angewendet auf $P^* - P_h^*$) folgt

$$\left|((\tilde{P}_h^{-1}P_h - \mathrm{id})(A - A_h)\phi_j, \phi_j^*)\right| \leq C\|(A - A_h)|_E\|\|(A^* - A_h^*)|_{E^*}\| \tag{5.11}$$

Der Betrag von (5.9) zusammen mit der Dreiecksungleichung über die Summe von (5.10) mit $j = 1, \ldots, m$ und der letzten Ungleichung ergibt nun die Behauptung. □

Bemerkung 5.18. *Mit (5.6) und der gleichmäßigen Beschränktheit von $\|(A^* - A_h^*)|_{E^*}\|$ folgt aus (5.7) unmittelbar*

$$|\lambda - \hat{\lambda}_h| \leq C\|(A - A_h)|_E\|. \tag{5.12}$$

Dies ist mit Ausnahme der Konstante die gleiche Abschätzung wie für die Eigenräume in Satz 5.13. In vielen Fällen werden wir jedoch eine bessere Abschätzung erwarten können. Zumindest der zweite Term erlaubt direkt eine Abschätzung höherer Ordnung, wenn man die Konvergenz $A_h^* \to A^*$ auf dem adjungierten Eigenraum E^* berücksichtigt.

Betrachten wir nun noch eine Abschätzung für $|\lambda - \lambda_j^{(h)}|$. Dazu erstmal ein technisches Hilfslemma.

Lemma 5.19. *Seien a, b, c Konstanten und $\alpha \in \mathbb{N}$. Dann gilt*

$$(a-b)^\alpha - (a-c)^\alpha = \sum_{k=0}^{\alpha-1}(a-b)^k(a-c)^{\alpha-1-k}(c-b). \tag{5.13}$$

Beweis. Für $\alpha = 1$ ist die Aussage offensichtlich korrekt. Danach zeigt man per Induktion über α

$$\begin{aligned}(a-b)^{\alpha+1} - (a-c)^{\alpha+1} &= (a-b)^{\alpha+1} - (a-c)^\alpha(a-b+b-c)\\ &= \left((a-b)^\alpha - (a-c)^\alpha\right)(a-b) + (c-b)(a-c)^\alpha\\ &= (a-b)\sum_{k=0}^{\alpha-1}(a-b)^k(a-c)^{\alpha-1-k}(c-b) + (c-b)(a-c)^\alpha\\ &= \sum_{k=0}^{\alpha}(a-b)^k(a-c)^{\alpha-k}(c-b).\end{aligned}$$

□

Satz 5.20. *Unter den Voraussetzungen von Satz 5.17 gilt für alle $j = 1, \ldots, m$*

$$|\lambda - \lambda_j^{(h)}| \leq C\left(\sum_{l,k=1}^{m}|((A - A_h)\phi_l, \phi_k^*)| + \|(A - A_h)|_E\|\|(A^* - A_h^*)|_{E^*}\|\right)^{1/r}, \tag{5.14}$$

wobei r die Riesz-Zahl des verallgemeinerten Eigenraums $E := \ker(A - \lambda\,\mathrm{id})^r$ ist.

Beweis. Wir verwenden die gleichen Bezeichnungen wie im Beweis von Satz 5.17. Für die Abbildung $T_h := \tilde{P}_h^{-1} A_h \tilde{P}_h : E \to E$ haben wir dort bewiesen, dass $\lambda_j^{(h)}$ ein Eigenwert von T_h ist. Sei $w_h \in E$ mit $\|w_h\| = 1$ eine zugehörige Eigenfunktion und $w_h^* := P^* w_h \in \ker(A^* - \overline{\lambda}\,\mathrm{id})^r$. Dann ist $(w_h, w_h^*) = (Pw_h, w_h) = (w_h, w_h) = 1$ und $\|w_h^*\| \le \|P\|$. Damit gilt mit dem vorigen Lemma

$$\begin{aligned}
|\lambda - \lambda_j^{(h)}|^r &= \left|\left((\lambda - \lambda_j^{(h)})^r w_h, w_h^*\right)\right| = \left|\left(\left\{(\lambda - \lambda_j^{(h)})^r\,\mathrm{id} - (\lambda\,\mathrm{id} - A)^r\right\} w_h, w_h^*\right)\right| \\
&= \left|\left(\sum_{k=0}^{r-1} \left\{(\lambda - \lambda_j^{(h)})^k (\lambda\,\mathrm{id} - A)^{r-1-k}\right\} (\lambda_j^{(h)}\,\mathrm{id} - A) w_h, w_h^*\right)\right| \\
&\le \sum_{k=0}^{r-1} |\lambda - \lambda_j^{(h)}|^k \left|\left((\lambda_j^{(h)}\,\mathrm{id} - A) w_h, (\overline{\lambda}\,\mathrm{id} - A^*)^{r-1-k} w_h^*\right)\right| \\
&\le \sum_{k=0}^{r-1} |\lambda - \lambda_j^{(h)}|^k \left\|(\overline{\lambda}\,\mathrm{id} - A^*)^{r-1-k} w_h^*\right\| \max_{v^* \in E^*, \|v^*\|=1} \left|\left((\lambda_j^{(h)}\,\mathrm{id} - A) w_h, v^*\right)\right|.
\end{aligned}$$

Der letzte Term lässt sich wie im vorigen Beweis (beachte $v^* = P^* v^*$) weiter abschätzen durch

$$\begin{aligned}
\left|\left((\lambda_j^{(h)}\,\mathrm{id} - A) w_h, v^*\right)\right| &\le |((T_h - A) w_h, v^*)| = \left|\left(\tilde{P}_h^{-1} P_h (A_h - A) w_h, v^*\right)\right| \\
&\le \left|((A_h - A) w_h, v^*) + \left((\tilde{P}_h^{-1} P_h - \mathrm{id})(A_h - A) w_h, v^*\right)\right| \\
&\le |((A_h - A) w_h, v^*)| + C \|(A - A_h)|_E\| \|(A^* - A_h^*)|_{E^*}\|
\end{aligned}$$

Da $w_h \in E = \mathbf{span}\{\phi_1, \ldots, \phi_m\}$, $v^* \in E^* = \mathbf{span}\{\phi_1^*, \ldots, \phi_m^*\}$ und $\|w_h\| = \|v^*\| = 1$, existiert eine Konstante $C > 0$ mit

$$|((A_h - A) w_h, v^*)| \le C \sum_{l,k=1}^{m} |((A_h - A)\phi_l, \phi_k^*)|.$$

Zusammen erhalten wir

$$|\lambda - \lambda_j^{(h)}|^r \le \delta(h) \sum_{k=0}^{r-1} |\lambda - \lambda_j^{(h)}|^k$$

mit

$$\delta(h) := \tilde{C} \left(\sum_{l,k=1}^{m} |((A - A_h)\phi_l, \phi_k^*)| + \|(A - A_h)|_E\| \|(A^* - A_h^*)|_{E^*}\| \right) \to 0 \qquad \text{für } h \to 0.$$

Solange $|\lambda - \lambda_j^{(h)}| \ge 1$, können wir die Terme auf der rechten Seite durch $|\lambda - \lambda_j^{(h)}|^r$ abschätzen und erhalten

$$|\lambda - \lambda_j^{(h)}|^r \le \frac{\delta(h)}{1 - r\delta(h)} \to 0 \qquad \text{für } h \to 0.$$

Sobald $|\lambda - \lambda_j^{(h)}| \le 1$ schätzen wir sie durch 1 ab und erhalten die Behauptung mit $C = (r\tilde{C})^{1/r}$. □

Für halbeinfache Eigenwerte ($r = 1$) ist abgesehen von der Konstante die Konvergenzgeschwindigkeit der Eigenwerte die gleiche wie die des arithmetischen Mittels. Für $r > 1$ sehen wir jedoch eine schlechtere Konvergenzgeschwindigkeit. In diesem Fall sollte man also besser das arithmetische Mittel der diskreten Eigenwerte verwenden.

6. Approximationen von Eigenwertproblemen in variationeller Form

Die im vorigen Kapitel vorgestellte Theorie wollen wir nun auf Eigenwertprobleme der Form

$$\text{Suche } (\lambda, u) \in \mathbb{C} \times V \setminus \{0\} : a(u,v) = \lambda b(u,v), \qquad v \in V, \tag{6.1}$$

anwenden. Wir halten uns dabei wieder an [BO91]. Zur Vereinfachung nehmen wir jedoch an, dass $(V, (\cdot,\cdot)_V)$ ein komplexer Hilbert-Raum ist und wir Sesquilinearformen (oder Bilinearformen) auf $V \times V$ vorliegen haben. Wir verwenden im Wesentlichen die Voraussetzungen, welche auch in Lemma 2.11 verwendet wurden. Einziger Unterschied ist, dass a, b nicht hermitesch sein müssen.

Sei also a eine Bilinear- oder Sesquilinearform auf $V \times V$ mit

$$|a(u,v)| \leq C_a \|u\|_V \|v\|_V, \qquad u,v \in V, \quad C_a > 0, \tag{6.2a}$$
$$|a(u,u)| \geq \alpha \|u\|_V^2, \qquad u \in V, \quad \alpha > 0. \tag{6.2b}$$

Weiter sei $A : V \to V$ der durch

$$a(u,v) = (Au, v)_V, \qquad u,v \in V \tag{6.3}$$

erzeugte lineare, stetige und stetig-invertierbare Operator.

Sei b eine Bilinear- oder Sesquilinearform auf $V \times V$, $(H, (\cdot,\cdot)_H)$ ein weiterer Hilbert-Raum mit kompakter Einbettung $V \hookrightarrow H$ und

$$|b(u,v)| \leq C_b \|u\|_H \|v\|_H, \qquad u,v \in V, \quad C_b > 0. \tag{6.4}$$

Dann existieren Operatoren $T, T_* : V \to V$, welche durch

$$a(Tu, v) = b(u,v), \qquad u,v \in V, \tag{6.5a}$$
$$a(u, T_* v) = b(u,v), \qquad u,v \in V, \tag{6.5b}$$

definiert sind. Aus

$$\alpha \|Tu\|_V^2 \leq |a(Tu, Tu)| = |b(u, Tu)| \leq C_b \|u\|_H \|Tu\|_H$$

und der kompakten (und damit stetigen) Einbettung $V \hookrightarrow H$ folgt

$$\|Tu\|_V \leq C \|u\|_H \tag{6.6}$$

und damit die Stetigkeit und Kompaktheit von $T : V \to V$.

Lemma 6.1. *Sei T^* der zu T adjungierte Operator. Dann ist $T^* = A^* T_* A^{-*}$. Die Eigenwerte von T_* sind genau die Eigenwerte von T^* und der verallgemeinerte Eigenraum $\ker(T_* - \lambda \,\mathrm{id})^r$ von T_* ist gegeben durch $A^{-*}\left(\ker(T^* - \lambda \,\mathrm{id})^r\right)$.*

Beweis. Dies folgt aus

$$\begin{aligned}(Tu, v)_V =& (A^{-1}ATu, v)_V = (ATu, A^{-*}v)_V = a(Tu, A^{-*}v) = b(u, A^{-*}v)\\ =& a(u, T_* A^{-*}v) = (Au, T_* A^{-*}v)_V = (u, A^* T_* A^{-*}v)_V.\end{aligned}$$

□

Lemma 6.2. *Für einen Eigenwert $\lambda \in \mathbb{C} \setminus \{0\}$ von (6.1) definieren wir*

1. *$E(\lambda)$: Verallgemeinerter Eigenraum $\ker(T - \lambda^{-1}\,\mathrm{id})^r$ von T,*
2. *$P(\lambda)$: Spektraler Projektor auf den Eigenraum $E(\lambda)$ von T zum Eigenwert λ,*
3. *$E^*(\overline{\lambda})$: Verallgemeinerter Eigenraum $\ker(T^* - \overline{\lambda}^{-1}\,\mathrm{id})^r$ des adjungierten Operatores T^*,*
4. *$E_*(\overline{\lambda})$: Verallgemeinerter Eigenraum $\ker(T_* - \overline{\lambda}^{-1}\,\mathrm{id})^r = A^{-*}\left(\ker(T^* - \lambda\,\mathrm{id})^r\right)$ von T_*.*

Folgende Aussagen sind äquivalent

1. *$(\lambda, u) \in \mathbb{C} \setminus \{0\} \times E(\lambda) \setminus \{0\}$ ist ein verallgemeinertes Eigenpaar von (6.1),*
2. *$(\lambda^{-1}, u) \in \mathbb{C} \setminus \{0\} \times E(\lambda) \setminus \{0\}$ ist ein verallgemeinertes Eigenpaar von T,*
3. *$(\overline{\lambda}^{-1}, P^* u) \in \mathbb{C} \setminus \{0\} \times E^*(\overline{\lambda}) \setminus \{0\}$ ist ein verallgemeinertes Eigenpaar von T^*,*
4. *$(\overline{\lambda}^{-1}, A^{-*}P^* u) \in \mathbb{C} \setminus \{0\} \times E_*(\overline{\lambda}) \setminus \{0\}$ ist ein verallgemeinertes Eigenpaar von T_*,*
5. *$(\overline{\lambda}, A^{-*}P^* u) \in \mathbb{C} \setminus \{0\} \times E_*(\overline{\lambda}) \setminus \{0\}$ ist ein verallgemeinertes Eigenpaar des adjungierten, variationellen Eigenwertproblems*

$$\text{Suche } (\lambda, v) \in \mathbb{C} \times V \setminus \{0\} : a(u, v) = \lambda b(u, v), \qquad u \in V. \tag{6.7}$$

Die verallgemeinerten Eigenräume von (6.1) und (6.7) werden dabei über die verallgemeinerten Eigenräume von T und T_* definiert.

Beweis. Die erste Äquivalenz folgt aus

$$a(u, v) = \lambda b(u, v) = \lambda a(Tu, v) \quad \Leftrightarrow \quad a((T - \lambda^{-1}\,\mathrm{id})u, v) = 0, \qquad v \in V.$$

Die weiteren Aussagen folgen mit dem vorigen Lemma, Lemma 5.15 und Lemma 5.16. □

Sei nun $\{V_h\}_{h>0}$ eine Familie von endlich-dimensionalen Teilräumen von V mit monoton fallendem Index h sodass

$$\lim_{h\to 0} \inf_{v\in V_h} \|u - v\|_V = 0, \qquad u \in V. \tag{6.8}$$

Mit anderen Worten die a-orthogonale Projektion $\Pi_h : V \to V_h$ konvergiert punktweise gegen die Identität für alle $u \in V$.

Lemma 6.3. *$(\lambda_h, u_h) \in \mathbb{C} \setminus \{0\} \times V_h \setminus \{0\}$ ist genau dann ein Eigenpaar von*

$$a(u_h, v) = \lambda_h b(u_h, v), \qquad v \in V_h, \tag{6.9}$$

wenn (λ_h^{-1}, u_h) ein Eigenpaar von $T_h := \Pi_h T$ ist.

Beweis. Analog zum vorigen Lemma ist $(\lambda_h, u_h) \in \mathbb{C} \setminus \{0\} \times V_h \setminus \{0\}$ genau dann ein Eigenpaar von (6.9) wenn

$$a((T - \lambda_h^{-1}\,\mathrm{id})u_h, v) = 0, \qquad v \in V_h.$$

Aus $a(\tilde{u} - \Pi_h\tilde{u}, v) = 0$ für alle $v \in V_h$ folgt die Behauptung. □

Somit sind wir exakt in dem Kontext, für welchen wir im vorigen Kapitel Konvergenz nachgewiesen haben. In Lemma 6.2 haben wir bereits kontinuierliche Eigenräume definiert. Sei nun Γ eine geschlossene, positiv orientierte Jordan-Kurve in $\rho(T)$ mit einem einzelnen Eigenwert λ^{-1} von T im Inneren. Dann definieren wir für hinreichend kleine $h > 0$

1. $E_h(\Gamma)$: Summe der verallgemeinerter Eigenräme $\ker(T_h - \lambda_h^{-1}\,\mathrm{id})^{r_h}$ von T_h zu den Eigenwerten λ_h im Inneren der Kurve,

2. Projektionsfehler von $E(\lambda)$:

$$\epsilon_h(\lambda) := \sup_{\substack{u\in E(\lambda)\\ \|u\|_V=1}} \inf_{v\in V_h} \|u - v\|_V \tag{6.10a}$$

3. Projektionsfehler von $E_*(\overline{\lambda})$ (des verallgemeinerten Eigenraums des adjungierten, variationellen Problems, d.h. des Eigenraums von T_*)

$$\epsilon_h^*(\overline{\lambda}) := \sup_{\substack{u\in E_*(\overline{\lambda})\\ \|u\|_V=1}} \inf_{v\in V_h} \|u - v\|_V \tag{6.10b}$$

Zunächst noch ein kleines Hilfsresultat:

Lemma 6.4. *Unter den gegebenen Voraussetzungen gilt für alle $u \in E$ mit $\|u\| = 1$*

$$\|(T - T_h)u\| \le \frac{C_a\|T\|}{\alpha}\epsilon_h.$$

Beweis. Wie beim Céa-Lemma 3.1 folgt für beliebiges $v \in V_h$ wegen $T_h u = \Pi_h T u \in V_h$

$$\begin{aligned}\alpha\|(T-T_h)u\|^2 =& |a(Tu-T_hu, Tu-T_hu)| = |a(Tu-T_hu, Tu-v)|\\ \leq& C_a\|(T-T_h)u\|\|Tu-v\|\end{aligned}$$

und damit

$$\|(T-T_h)u\| \leq \frac{C_a\|Tu\|}{\alpha}\left\|\frac{Tu}{\|Tu\|} - \frac{v}{\|Tu\|}\right\|.$$

Da E invariant ist gegenüber T (d.h. $Tu \in E$), folgt mit $\|Tu\| \leq \|T\|$ die Behauptung. □

Satz 6.5. *Unter den gegebenen Voraussetzungen existieren Konstanten $h_0, C > 0$ sodass für alle $0 < h \leq h_0$ die Dimensionen von E und E_h übereinstimmen und sodass gilt*

$$\hat{\delta}(E, E_h) \leq C\epsilon_h. \tag{6.11}$$

Beweis. Aus Satz 5.13 folgen der erste Teil der Aussage und die Abschätzung

$$\hat{\delta}(E, E_h) \leq \tilde{C}\|(T-T_h)|_E\|.$$

Aus dem vorigen Lemma folgt dann die Behauptung. □

Satz 6.6. *Sei m die algebraische Vielfachheit des Eigenwertes λ. Dann existieren für alle hinreichend kleinen h innerhalb der Kurve Γ genau m Eigenwerte $\lambda_j^{(h)}$ (gemäß ihrer algebraischen Vielfachheit gezählt) von (6.9) und es gilt*

$$\left|\lambda - \left(\frac{1}{m}\sum_{j=1}^{m}\frac{1}{\lambda_j^{(h)}}\right)^{-1}\right| \leq C\epsilon_h\epsilon_h^*. \tag{6.12}$$

Beweis. In Satz 5.17 wurde die Fehlerabschätzung

$$|\lambda^{-1} - \hat{\lambda}_h| \leq \frac{1}{m}\sum_{j=1}^{m}|((T-T_h)\phi_j, \phi_j^*)| + C\|(T-T_h)|_E\|\|(T^*-T_h^*)|_{E^*}\|$$

mit $\hat{\lambda}_h := \frac{1}{m}\sum_{j=1}^{m}\lambda_j^{(h)^{-1}}$ bewiesen. Wegen

$$|\lambda - \hat{\lambda}_h^{-1}| = \frac{|\lambda^{-1} - \hat{\lambda}_h|}{|\lambda^{-1}\hat{\lambda}_h|}$$

erhalten wir daraus die Behauptung, sofern wir die rechte Seite abschätzen können.

Betrachten wir zunächst den ersten Teil der rechten Seite. Sei $u \in E$ mit $\|u\| = 1$ und $v^* \in E^*$ (also verallgemeinerte Eigenfunktionen des adjungierten Operators T^*) mit

$\|v^*\| = 1$). Dann ist wegen Lemma 6.1 $A^{-*}v^*$ eine verallgemeinerter Eigenfunktion von T_*. Es gilt für beliebiges $v_h \in V_h$

$$\begin{aligned}((T-T_h)u, v^*) =& (A^{-1}A(T-T_h)u, v^*) = (A(T-T_h)u, A^{-*}v^*)\\ =& a((T-T_h)u, A^{-*}v^*) = a((T-T_h)u, A^{-*}v^* - v_h)\\ \leq& C_a\|A^{-*}\|\,\|(T-T_h)u\|\left\|\frac{A^{-*}v^*}{\|A^{-*}v^*\|} - \frac{v_h}{\|A^{-*}v^*\|}\right\|\\ \leq& C_a\|A^{-*}\|\,\|(T-T_h)u\|\,\epsilon_h^*.\end{aligned}$$

Mit Lemma 6.4 erhalten wir für den ersten Teil der rechten Seite die gewünschte Abschätzung. Für den zweiten Teil betrachten wir erstmal den zweiten Faktor und erhalten analog

$$\begin{aligned}\|(T^*-T_h^*)v^*\| =& \sup_{\substack{w\in V\\ \|w\|=1}} |(w, (T^*-T_h^*)v^*)| = \sup_{\substack{w\in V\\ \|w\|=1}} |((T-T_h)w, v^*)|\\ \leq& C_a\|A^{-*}\|\,\|T-T_h\|\,\epsilon_h^*.\end{aligned}$$

Zusammen mit dem ersten Faktor $\|(T-T_h)|_E\|$ des zweiten Teils der rechten Seite folgt nun die Behauptung. □

Bemerkung 6.7. *Auch Satz 5.20 zur Konvergenz der Eigenwerte selber kann analog übertragen werden zu*

$$\left|\lambda - \lambda_j^{(h)}\right| \leq C\,(\epsilon_h\epsilon_h^*)^{1/r}, \qquad j = 1, \ldots, m. \tag{6.13}$$

Für selbstadjungierte und damit halbeinfache Eigenwerte ($r = 1$) ergibt dies quadratische Konvergenz ϵ_h^2 im Gegensatz zur linearen Konvergenz ϵ_h für die Eigenräume. Dies ist asymptotisch (also abgesehen von den Konstanten) die gleiche Aussage wie in Satz 3.3. Wobei in Satz 3.3 bei höheren Eigenwerten Bestapproximationsfehler auftauchen, welche zu Eigenräumen von kleineren Eigenwerten gehören. Dies ist mit dieser Beweistechnik nicht notwendig.

Zu beachten ist bei allen Abschätzungen, dass diese nur für hinreichende feine Diskretisierungen gelten. Aufgrund der diversen Abschätzungen ist diese Grenze unmöglich praktisch zu berechnen. Auch die Konstanten sind praktisch nicht berechenbar. Daher wird man im konkreten Beispiel den zu erwartenden Fehler praktisch nicht angeben können.

7. Nichtlineare Eigenwertprobleme

Im Folgenden beschäftigen wir uns mit nichtlinearen Eigenwertproblemen. Streng genommen ist diese Bezeichnung unsinnig, da bereits lineare Eigenwertprobleme der Form $Au = \lambda u$ durch das Produkt λu des gesuchten Eigenpaares (λ, u) nichtlinear sind. Dennoch ist diese Bezeichnung gebräuchlich.

In der allgemeinsten Form eines nichtlinearen Eigenwertproblems werden Eigenpaare (λ, u) mit $u \neq 0$ gesucht, sodass $L(\lambda, u) = 0$ gilt. $(\lambda, u) \mapsto L(\lambda, u)$ bezeichnet dabei eine Familie von Operatoren, die nichtlinear von λ und u abhängen können. Letztlich wären in dieser Form also einfach die Nullstellen von L gesucht. Wir beschränken uns in dieser Vorlesung jedoch auf Operatoren L, die linear in der Eigenfunktion sind.

Sei also $(V, (\cdot, \cdot))$ ein Hilbert-Raum und $\lambda \mapsto L(\lambda)$ holomorph von einer Teilmenge des $\mathbb{C}$ in die Menge der beschränkten, linearen Operatoren auf V. Dann suchen wir ein Eigenpaar $(\lambda, u) \in \mathbb{C} \times V \setminus \{0\}$ sodass

$$L(\lambda)u = 0. \tag{7.1}$$

7.1. Polynomielle Eigenwertprobleme

In diesem Abschnitt sei $L(\lambda)$ von der Form

$$L(\lambda) = \sum_{j=0}^{n} \lambda^j L_j, \tag{7.2}$$

mit $n \in \mathbb{N}$ und beschränkten, linearen Operatoren $L_j : V \to V$, $j = 0, \ldots, n$, welche unabhängig von $\lambda \in \mathbb{C}$ sind.

Beispiel 7.1. *Betrachten wir die sogenannte Telegraphen-Gleichung. In dieser wird die Spannung $u(x,t)$ eines Übertragungskabels durch*

$$\frac{\partial^2}{\partial x^2} u = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} u + \gamma \frac{\partial}{\partial t} u + ku$$

mit Konstanten $c, \gamma, k > 0$ beschrieben. Diese Konstanten hängem vom Widerstand, der Induktion, der Kapazität und vor allem den Stromverlusten im Kabel ab. Bei einem zeit-harmonischen Ansatz $u(x,t) = \Re\{\exp(-\mathrm{i}\omega t)v(x)\}$ mit $\omega \in \mathbb{C}$ erhalten wir

$$(-\Delta + k\,\mathrm{id})\, v - \mathrm{i}\gamma\omega v - \frac{\omega^2}{c^2} v = 0. \tag{7.3}$$

Dies entspricht einem quadratischen Eigenwertproblem in ω.

Beispiel 7.2. *Betrachten wir Lösungen der eindimensionalen Helmholtz-Gleichung*

$$-\Delta u(x) - \omega^2 p(x) u(x) = 0, \qquad x > 0 \tag{7.4a}$$

mit $p(x) = 1$ für $x \geq R > 0$ und $u'(0) = 0$. Diese sind nicht eindeutig, da für $x > R$ zwei linear unabhänige Lösungen der Form

$$x \mapsto u_\pm(x) := \exp(\pm \mathrm{i}\omega x)$$

existieren. Physikalisch sinnvoll ist davon typischerweise u_+. Dies führt an der Stelle $x = R$ zu der Randbedingung

$$u'(R) = \mathrm{i}\omega u(R) \tag{7.4b}$$

und somit ebenfalls zu einem quadratischen Eigenwertproblem in ω. In variationeller Form erhalten wir nämlich

$$\int_0^R u'(x)v'(x)dx - \mathrm{i}\omega u(R)v(R) - \omega^2 \int_0^R p(x)u(x)v(x)dx = 0, \qquad v \in H^1((0,R)). \tag{7.5}$$

Polynomielle Eigenwertprobleme können formal immer in verallgemeinerte, lineare Eigenwertprobleme umgeschrieben werden. Sei dazu $\mathbf{v} = (v_1, \ldots, v_n)^\top \in V^n$ mit $v_j := \lambda^{j-1} u$ für $j = 1, \ldots, n$. Dann ist das Eigenwertproblem (7.1) äquivalent zu

$$\mathbf{A}\,\mathbf{v} = \lambda\,\mathbf{B}\,\mathbf{v} \tag{7.6}$$

mit den Operatoren $\mathbf{A}, \mathbf{B} : V^n \to V^n$ definiert durch

$$\mathbf{A} := \operatorname{diag}(L_0, \mathrm{id}, \ldots, \mathrm{id}), \quad \mathbf{B} := \begin{pmatrix} -L_1 & -L_2 & \ldots & -L_n \\ \mathrm{id} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & \mathrm{id} & 0 \end{pmatrix}. \tag{7.7}$$

Dies ist jedoch bei weitem nicht die einzige Möglichkeit, um ein lineares Eigenwertproblem zu erhalten. Betrachten wir dazu Beispiel 7.1. Mit den Bilinearformen

$$a(v,w) = \int_\Omega (\nabla v \cdot \nabla w + k\, v\, w)\, dx \tag{7.8a}$$

$$b_1(v,w) = -\mathrm{i} \int_\Omega \gamma\, v\, w\, dx \tag{7.8b}$$

$$b_2(v,w) = \int_\Omega \frac{1}{c^2}\, v\, w\, dx \tag{7.8c}$$

ist (7.3) zusammen mit homogenen Neumann-Randbedingungen äquivalent zu ($v \in H^1(\Omega)$)

$$a(v,w) + \omega b_1(v,w) - \omega^2 b_2(v,w) = 0, \qquad w \in H^1(\Omega). \tag{7.9}$$

Wir führen nun analog zu (6.5) die kompakten Operatoren $B_1, B_2 : V \to V$ durch

$$a(B_1 v, w) = b_1(v, w), \quad a(B_2 v, w) = b_2(v, w) \tag{7.10}$$

ein. Dann ist (7.9) wie in Lemma 6.2 äquivalent zum Eigenwertproblem

$$\left(\mathrm{id} + \omega B_1 - \omega^2 B_2\right) u = 0. \tag{7.11}$$

Analog ist das diskrete Problem

$$\text{Suche} \quad (\omega_h, v_h) \in \mathbb{C} \times V_h \setminus \{0\} : \quad a(v_h, w) + \omega_h b_1(v, w) - \omega_h^2 b_2(v_h, w) = 0, \qquad w \in V_h \tag{7.12}$$

äquivalent zu

$$\left(\Pi_h + \omega_h \Pi_h B_1 - \omega_h^2 \Pi_h B_2\right) u_h = 0 \tag{7.13}$$

mit der a-orthogonalen Projektion $\Pi_h : V \to V_h$.

Wir wollen die Babuska-Osborn-Theorie für lineare Eigenwertprobleme anwenden. Dazu schreiben wir (7.11) für $\omega \neq 0$ und $\mathbf{y} \in \left(H^1(\Omega)\right)^2 \setminus \{0\}$ in ein lineares Eigenwertproblem um:

$$\begin{pmatrix} \mathrm{id} & B_1 \\ 0 & \mathrm{id} \end{pmatrix} \mathbf{y} = \omega \begin{pmatrix} 0 & B_2 \\ \mathrm{id} & 0 \end{pmatrix} \mathbf{y} \qquad \Leftrightarrow \qquad \begin{pmatrix} -B_1 & B_2 \\ \mathrm{id} & 0 \end{pmatrix} \mathbf{y} = \frac{1}{\omega} \mathbf{y}. \tag{7.14}$$

Analog erhalten wir für $\mathbf{y}_h \in V_h^2 \setminus \{0\}$ zunächst

$$\begin{pmatrix} \Pi_h & \Pi_h B_1 \\ 0 & \Pi_h \end{pmatrix} \mathbf{y}_h = \omega_h \begin{pmatrix} 0 & \Pi_h B_2 \\ \Pi_h & 0 \end{pmatrix} \mathbf{y}_h. \tag{7.15}$$

Nun ist klar, dass Eigenpaare $(\omega_h, \mathbf{y}_h) \in \mathbb{C} \times V_h^2 \setminus \{0\}$ von (7.15) auch Eigenpaare von

$$\begin{pmatrix} \mathrm{id} & \Pi_h B_1 \\ 0 & \mathrm{id} \end{pmatrix} \mathbf{y}_h = \omega_h \begin{pmatrix} 0 & \Pi_h B_2 \\ \mathrm{id} & 0 \end{pmatrix} \mathbf{y}_h \qquad \Leftrightarrow \qquad \begin{pmatrix} -\Pi_h B_1 & \Pi_h B_2 \\ \mathrm{id} & 0 \end{pmatrix} \mathbf{y}_h = \frac{1}{\omega_h} \mathbf{y}_h. \tag{7.16}$$

sind. Andererseits gilt für Eigenpaare von (7.16) sofort aus der ersten Gleichung

$$\mathbf{y}_h^{(1)} = \Pi_h \left(\omega_h B_2 - B_1\right) \mathbf{y}_h^{(2)} \in V_h$$

und damit aus der zweiten Gleichung auch $\mathbf{y}_h^{(2)} = \omega_h \mathbf{y}_h^{(1)} \in V_h$. Somit sind Eigenpaare von (7.16) auch Eigenpaare von (7.15).

Die wichtigste Voraussetzung der Babuska-Osborn-Theorie ist die Normkonvergenz der diskreten gegen die kontinuierlichen Operatoren. Dies folgt hier aus der Kompaktheit von B_1 und B_2, da

$$\lim_{h \to 0} \|(\Pi_h - \mathrm{id}) B_1\| = 0, \qquad \lim_{h \to 0} \|(\Pi_h - \mathrm{id}) B_2\| = 0.$$

Es bleibt lediglich die Frage, welches Spektrum wir für (7.14) erhalten. Dieser Operator ist nämlich nicht kompakt. Da (7.14) äquivalent zu (7.11) ist und B_1 und B_2

kompakt sind, können wir die analytische Fredholm-Theorie anwenden. Aus dieser ergibt sich, dass (7.11) mit Ausnahme des Ursprungs nur isolierte Eigenwerte endlicher Vielfachheit hat. Somit ist die Babuska-Osborn-Theorie anwendbar und wir erhalten auch für das quadratische Eigenwertproblem (7.12) die üblichen Konvergenzraten für die Eigenwerte und die Eigenräume.

Betrachten wir zum Schluß dieses Abschnittes nun noch numerische Löser für (7.12) oder allgemeiner für

$$\text{Suche } (\omega_h, u_h) \in \mathbb{C} \times V_h \setminus \{0\} : \sum_{j=0}^{n} \omega_h^j a_j(u_h, v_h) = 0, \qquad v_h \in V_h.$$

Bei gegebener Basis von V_h ist dies äquivalent zum polynomiellen Matrix-Eigenwertproblem

$$\text{Suche } (\lambda, y) \in \mathbb{C} \times \mathbb{C}^N \setminus \{0\} : P(\lambda)y = 0, \qquad P(\lambda) := \sum_{j=0}^{n} \lambda^j A_j, \tag{7.17}$$

mit Matrizen $A_j \in \mathbb{C}^{N\times N}$, $j = 0, \dots n$. Die Abhängigkeit von h werden wir im Folgenden unterdrücken. Auch dieses Eigenwertproblem können wir in ein lineares Eigenwertproblem der Form

$$\text{Suche } (\lambda, \mathbf{y}) \in \mathbb{C} \times \mathbb{C}^{nN} \setminus \{0\} : \quad \mathbf{A}\,\mathbf{y} = \lambda\,\mathbf{B}\,\mathbf{y} \tag{7.18}$$

mit den Matrizen $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{nN\times nN}$

$$\mathbf{A} := \operatorname{diag}(A_0, \mathrm{id}, \dots, \mathrm{id}), \quad \mathbf{B} := \begin{pmatrix} -A_1 & -A_2 & \dots & -A_n \\ \mathrm{id} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & \mathrm{id} & 0 \end{pmatrix}.$$

umschreiben. Wenn wir das Eigenwertproblems wie in (4.2) (Shift-and-Invert) umschreiben, müssen wir für einen Parameter $\rho \in \mathbb{C}$ die Matrix $\mathbf{A} - \rho\,\mathbf{B}$ invertieren. Für große n wäre dies ohne weitere Vorbereitungen sehr aufwändig. Wir können jedoch die spezielle Struktur

$$\mathbf{A} - \rho\,\mathbf{B} = \begin{pmatrix} A_0 + \rho A_1 & \rho A_2 & \dots & \rho A_n \\ -\rho\,\mathrm{id} & \mathrm{id} & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & -\rho\,\mathrm{id} & \mathrm{id} \end{pmatrix} \tag{7.19}$$

ausnutzen. Dazu verwenden wir das Schur-Komplent

$$\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathrm{id} & 0 \\ -C_{22}^{-1}C_{21} & \mathrm{id} \end{pmatrix} \begin{pmatrix} (C_{11} - C_{12}C_{22}^{-1}C_{21})^{-1} & 0 \\ 0 & C_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathrm{id} & -C_{12}C_{22}^{-1} \\ 0 & \mathrm{id} \end{pmatrix}$$

und starten mit $C_{22} = \mathrm{id}$, also dem Eintrag unten rechts in $\mathbf{A} - \rho\,\mathbf{B}$. Wir erhalten

$$(\mathbf{A} - \rho\,\mathbf{B})^{-1} = \begin{pmatrix} \mathrm{id} & 0 \\ (0, \dots, 0, \rho\,\mathrm{id}) & \mathrm{id} \end{pmatrix} \begin{pmatrix} D_1^{-1} & 0 \\ 0 & \mathrm{id} \end{pmatrix} \begin{pmatrix} \mathrm{id} & \begin{pmatrix} -\rho A_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ 0 & \mathrm{id} \end{pmatrix} \tag{7.20}$$

mit

$$D_1 := \begin{pmatrix} A_0 + \rho A_1 & \rho A_2 & \dots & \rho A_{n-2} & \rho A_{n-1} + \rho^2 A_n \\ -\rho\,\mathrm{id} & \mathrm{id} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -\rho\,\mathrm{id} & \mathrm{id} & 0 \\ 0 & \dots & 0 & -\rho\,\mathrm{id} & \mathrm{id} \end{pmatrix}.$$

Es bleibt somit, die Matrix $D_1 \in \mathbb{C}^{(n-1)N\times(n-1)N}$ zu invertieren. Ein weiteres Mal ein Schur-Komplement führt zur Matrix

$$D_2 := \begin{pmatrix} A_0 + \rho A_1 & \rho A_2 & \dots & \rho A_{n-3} & \rho A_{n-2} + \rho^2 A_{n-1} + \rho^3 A_n \\ -\rho\,\mathrm{id} & \mathrm{id} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -\rho\,\mathrm{id} & \mathrm{id} & 0 \\ 0 & \dots & 0 & -\rho\,\mathrm{id} & \mathrm{id} \end{pmatrix} \in \mathbb{C}^{(n-2)N\times(n-2)N}.$$

Weiter fortführend gelangen wir zur Matrix

$$D_{n-1} = A_0 + \rho A_1 + \dots + \rho^n A_n = P(\rho) \in \mathbb{C}^{N\times N}.$$

Effektiv müssen wir also wie beim linearen Eigenwertproblem nur die Matrix $P(\rho)$ invertieren. Alles zusammen erhalten wir

$$(\mathbf{A} - \rho\,\mathbf{B})^{-1} = \begin{pmatrix} \mathrm{id} & & & & \\ \rho\,\mathrm{id} & \mathrm{id} & & & \\ \rho^2\,\mathrm{id} & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ \rho^{n-1}\,\mathrm{id} & \dots & \rho^2\,\mathrm{id} & \rho\,\mathrm{id} & \mathrm{id} \end{pmatrix} \begin{pmatrix} P(\rho)^{-1} & & & \\ & \mathrm{id} & & \\ & & \ddots & \\ & & & \mathrm{id} \end{pmatrix} \begin{pmatrix} \mathrm{id} & E_2 & \dots & E_n \\ & \mathrm{id} & & \\ & & \ddots & \\ & & & \mathrm{id} \end{pmatrix} \tag{7.21}$$

mit

$$E_j := -\sum_{l=j}^{n} \rho^{l+1-j} A_l.$$

Nun können wir die Arnoldi-Iteration aus 4.7 mit diesem shift-and-invert Ansatz kombinieren. Der Aufwand wird für größere n höher sein als bei linearen Eigenwertproblemen, weil z.B. die Orthonormalisierung im Algorithmus auf $\mathbb{C}^{nN}$ durchgeführt werden muss. Die Matrix-Inversion bzw. das Lösen des linearen Gleichungssystems hat jedoch nahezu den gleichen Aufwand wie im linearen Fall.

7.2. Vereinfachte Konvergenztheorie für nichtlineare Eigenwertprobleme

Wir betrachten nun Operatoren im nichtlinearen Eigenwertproblem (7.1), welche von folgender Form sind:

$$L(\lambda) := A(\lambda) + K.$$

Dabei ist $K : V \to V$ unabhängig von λ (die Vor. ist nicht zwingend, erleichtert aber die Beweise) und $\lambda \mapsto A(\lambda)$ ist für $\lambda \in \Lambda$ holomorph, wobei $\Lambda \subset \mathbb{C}$ offen und zusammenhängend sein soll. Weiters sollen $A(\lambda) : V \to V$ lineare, beschränkte Operatoren sein, welche gleichmäßig für λ in einer kompakten Teilmenge von Λ koerizitiv sind, d.h. es existiert ein $\alpha > 0$ mit

$$|(A(\lambda)u, u)| \geq \alpha \|u\|^2, \qquad u \in V. \tag{7.22}$$

Beispiel 7.3. *Die Beispiele aus dem vorigen Abschnitt sind von dieser Bauart. So ist* (7.11) *äquivalent zu*

$$(A(\lambda) + K)\, v = 0,$$

wobei für beliebiges $C \in \mathbb{C}$ die Operatoren definiert werden durch

$$A(\lambda) := \mathrm{id} + \left(k - \mathrm{i}\gamma\lambda - \frac{\omega^2}{c^2} + C\right) K \tag{7.23}$$

wobei der kompakte Operator K durch die Sesquilinarform $(u, v)_{L^2(\Omega)}$ wie in (6.5) *definiert wird. $A(\lambda)$ ist stetig und für beschränkte Λ und hinreichend große C hat*

$$k - \mathrm{i}\omega - \frac{\omega^2}{c^2} + C$$

positiven Realteil (gleichmäßig für λ in einer kompakten Teilmenge von Λ). Somit gilt (7.22).

Aufgrund der Voraussetzungen ist klar, dass für fixes λ_0 der Operator $L(\lambda_0)$ ein Fredholm-Operator ist. Somit ist der Kern $\ker L(\lambda_0)$ auf jeden Fall endlich-dimensional. Weiter ist mit Lem. 5.1 klar, dass für invertierbares $L(\lambda_0)$ eine Umgebung um λ_0 existiert, sodass für alle λ in dieser Umgebung der Operator invertierbar ist und die Abbildung $\lambda \mapsto L(\lambda)^{-1}$ holomorph ist. Mit Hilfe der analytischen Fredholm-Theorie lässt sich nachweisen, dass $\lambda \mapsto L(\lambda)^{-1}$ für alle λ, die im Inneren einer kompakten Teilmenge von Λ liegen, eine meromorphe Funktion ist, deren diskret viele Pole genau die Eigenwerte von (7.1) sind.

Im Folgenden geht es um die Konvergenztheorie, wenn wir V in endlich-dimensionale Teilräume V_h projezieren. In [Kar96a, Kar96b] wird nachgewiesen, dass die gleichen Resultate wie bei der Babuska-Osborn-Theorie gelten. Der Aufwand dazu ist jedoch sehr hoch. Daher beschränkten wir uns hier auf Teilresultate, die mit angemessenem Aufwand zu beweisen sind.

Satz 7.4. *Sei $\Lambda \subset \mathbb{C}$ offen, $\hat{\Lambda} \subset \Lambda$ kompakt und sei $A(\lambda) : V \to V$ für $\lambda \in \Lambda$ eine Familie beschränkter linearer Operatoren, sodass die Abbildung $\lambda \mapsto A(\lambda)$ stetig ist. Weiter sei der Operator $K : V \to V$ kompakt und für alle $\lambda \in \hat{\Lambda}$ gelte $|(A(\lambda)v, v)| \geq \alpha\|v\|^2$, $v \in V$, mit einer Konstanten $\alpha > 0$.*

Sei weiter $V_h \subset V$ für $h > 0$ eine Familie abgeschlossene Unterräume von V, sodass die orthogonale Projektion $\Pi_h : V \to V_h$ für $h \to 0$ punktweise gegen die Identität $\mathrm{id} : V \to V$ konvergiert.

Ist $L(\lambda) := A(\lambda) + K$ für alle $\lambda \in \hat{\Lambda}$ invertierbar, dann existiert ein $h_0 > 0$, sodass $L_h := \Pi_h L : V_h \to V_h$ für alle $h \le h_0$ invertierbar ist und

$$\sup_{\lambda\in\hat{\Lambda}} \left\| L_h(\lambda)^{-1} \right\| \le C \tag{7.24}$$

mit einer Konstanten $C > 0$ unabhängig von h.

Beweis. Für $\hat{\Lambda} := \{\lambda_0\}$ folgt der Beweis aus [Kre99, Theorem 13.7]. Wir zeigen hier den etwas allgemeineren Fall.

Sei dazu $\lambda \in \hat{\Lambda}$. $A_h(\lambda) := \Pi_h A(\lambda) : V_h \to V_h$ und $A(\lambda) : V \to V$ sind nach dem Satz von Lax-Milgram stetig invertierbar und wegen Lem. 5.1 sind die Abbildungen $\lambda \mapsto \|A(\lambda)^{-1}\| \le 1/\alpha$ und $\lambda \mapsto \|A_h(\lambda)^{-1}\| \le 1/\alpha$ stetig. Daher existiert ein $\alpha > 0$ sodass $\sup_{\lambda\in\hat{\Lambda}} \|A(\lambda)^{-1}\| \le 1/\alpha$ und $\sup_{\lambda\in\hat{\Lambda}} \|A_h(\lambda)^{-1}\| \le 1/\alpha$ gilt. Analog existiert ein $C_1 > 0$ mit $\sup_{\lambda\in\hat{\Lambda}} \|A(\lambda)\| \le C_1$.

Wie im Beweis des Céa-Lemmas 3.1 folgt für alle $\lambda \in \hat{\Lambda}$ aus der Galerkin-Orthogonalität

$$\|A_h(\lambda)^{-1}\Pi_h f - A(\lambda)^{-1} f\| \le \frac{C_1}{\alpha} \inf_{v_h\in V_h} \|A(\lambda)^{-1} f - v_h\|, \qquad f \in V,$$

und aus der punktweisen Konvergenz von $\Pi_h - \mathrm{id}$

$$\lim_{h\to 0} \sup_{\lambda\in\hat{\Lambda}} \|A_h(\lambda)^{-1}\Pi_h f - A(\lambda)^{-1} f\| = 0, \qquad f \in V. \tag{7.25}$$

Wir zeigen, dass daraus

$$\lim_{h\to 0} \sup_{\lambda\in\hat{\Lambda}} \|A_h(\lambda)^{-1}\Pi_h K - A(\lambda)^{-1} K\|_{L(V)} = 0 \tag{7.26}$$

folgt und verwenden dazu, dass eine punktweise Konvergenz gleichmäßig ist auf kompakten Teilmengen des Raumes.

Genauer sei $\epsilon > 0$. Dann kann die relativ kompakte Menge $U := \{Kf : f \in V, \|f\| \le 1\}$ durch eine endliche Anzahl von Kugeln $B_r(f_m)$, $m = 1, \dots, M(\epsilon)$ mit Radius $r := \alpha\epsilon/3$ überdeckt werden. Wegen (7.25) existiert ein $h_1 > 0$, sodass $\sup_{\lambda\in\hat{\Lambda}} \|A_h(\lambda)^{-1}\Pi_h f_m - A(\lambda)^{-1} f_m\| \le \epsilon/3$ für alle $h \le h_1$ und alle $m = 1, \dots, M$. Da alle $f \in U$ in einem der Bälle $B_r(f_m)$ enthalten sind, folgt

$$\begin{aligned}
&\|A_h(\lambda)^{-1}\Pi_h f - A(\lambda)^{-1} f\| \\
&\le \|A_h(\lambda)^{-1}\Pi_h (f - f_m)\| + \|A_h(\lambda)^{-1}\Pi_h f_m - A(\lambda)^{-1} f_m\| + \|A(\lambda)^{-1}(f - f_m)\| \\
&\le \frac{1}{\alpha} r + \frac{\epsilon}{3} + \frac{1}{\alpha} r = \epsilon
\end{aligned}$$

und daraus (7.26). Nun zerlegen wir L_h in

$$\begin{aligned}
L_h(\lambda) =& A_h(\lambda) + \Pi_h K = A_h(\lambda) \left(\mathrm{id} + A_h(\lambda)^{-1}\Pi_h K\right) \\
=& A_h(\lambda) \left(\mathrm{id} + A(\lambda)^{-1} K + \left(A_h(\lambda)^{-1}\Pi_h - A(\lambda)^{-1}\right) K\right) \\
=& A_h(\lambda) A(\lambda)^{-1} \left(A(\lambda) + K + A(\lambda) \left(A_h(\lambda)^{-1}\Pi_h - A(\lambda)^{-1}\right) K\right) \\
=& A_h(\lambda) A(\lambda)^{-1} L(\lambda) \big(\mathrm{id} + \underbrace{L(\lambda)^{-1} A(\lambda) \left(A_h(\lambda)^{-1}\Pi_h - A(\lambda)^{-1}\right) K}_{=: B_h(\lambda)}\big).
\end{aligned} \tag{7.27}$$

$L(\lambda)$ ist für $\lambda \in \hat{\Lambda}$ invertierbar und nach der Riesz-Theorie ist die Inverse $L(\lambda)^{-1}$ beschränkt. Wegen Lem. 5.1 folgt die Stetigkeit von $\lambda \mapsto L(\lambda)^{-1}$ und daraus $\sup_{\lambda\in\hat{\Lambda}} \|L(\lambda)^{-1}\| \leq C_2$ mit einem $C_2 > 0$. Wegen (7.26) existiert daher ein $h_0 \leq h_1$ sodass $\sup_{\lambda\in\hat{\Lambda}} \|B_h(\lambda)\| \leq C_3 < 1$ für alle $h \leq h_0$. Damit folgt die Behauptung über die Neumannsche Reihe mit

$$\sup_{\lambda\in\hat{\Lambda}} \|L_h(\lambda)^{-1}\| \leq \sup_{\lambda\in\hat{\Lambda}} \frac{\|L(\lambda)^{-1}A(\lambda)A_h(\lambda)^{-1}\|}{1 - \|B_h(\lambda)\|} \leq \frac{C_1C_2\alpha^{-1}}{1 - C_3}. \tag{7.28}$$

□

Satz 7.5. *Sei $(\lambda_h, u_h) \in \Lambda \times V_h \setminus \{0\}$ eine Folge von Eigenpaaren des diskreten Eigenwertproblems*

$$(\Pi_h L)(\lambda_h)u_h = 0 \tag{7.29}$$

und λ_h konvergiere für $h \to 0$ gegen ein $\lambda_0 \in \Lambda$. Dann ist λ_0 ein Eigenwert von (7.1), d.h. der Grenzwert einer Folge von diskreten Eigenwerten ist immer ein Eigenwert des kontinuierlichen Problems (es gibt keine Fehlkonvergenz).

Beweis. Sei $\lambda_0 \in \Lambda$ kein Eigenwert von (7.1), d.h. $L(\lambda_0)$ ist stetig invertierbar. Dann existiert ein hinreichend kleines kompaktes $\hat{\Lambda} \in \Lambda$, sodass $L(\lambda)$ für alle $\lambda \in \hat{\Lambda}$ stetig invertierbar ist. Mit Hilfe von Satz 7.4 folgt daraus, dass für alle hinreichend kleinen h die Operatoren L_h stetig invertierbar sind. Damit ist die eindeutige Lösung u_h von (7.29) für hinreichend kleine h gleich 0 im Widerspruch zur Voraussetzung, dass $u_h \in V_h \setminus \{0\}$. □

Satz 7.6. *Für alle $\lambda_0 \in \Lambda$, die nicht Eigenwerte von (7.1) sind, existieren Konstanten $h_0, \epsilon > 0$, sodass für alle $h \leq h_0$ die Menge $\{\lambda \in \Lambda : |\lambda - \lambda_0| < \epsilon\}$ keine Eigenwerte des diskreten Eigenwertproblems (7.29) enthält.*

Beweis. Folgt direkt aus dem vorigen Satz. □

Lemma 7.7 (Maximumsprinzip)**.** *Sei Γ der Rand eines einfach zusammenhängenden, beschränkten Gebietes $D \subset \Lambda$ und sei $\lambda \mapsto S(\lambda)$ mit einem linearen, beschränkten Operator $S(\lambda) : V \to V$ holomorph für alle $\lambda \in \overline{D}$. Dann gilt*

$$\max_{\lambda\in\overline{D}} \|S(\lambda)\| \leq \max_{\lambda\in\Gamma} \|S(\lambda)\|. \tag{7.30}$$

Beweis. Die Behauptung folgt aus der Integralformel von Cauchy

$$S(z) = \frac{1}{2\pi\mathrm{i}} \int_\Gamma \frac{S(\lambda)}{\lambda - z} d\lambda, \qquad z \in D.$$

□

Satz 7.8. *Für jeden Eigenwert $\lambda_0 \in \Lambda$ von (7.1) existiert eine Folge von diskreten Eigenwerten λ_h von (7.29), welche gegen λ_0 konvergiert.*

Beweis. Wegen der analytischen Fredholm-Theorie existiert ein $\epsilon > 0$ sodass $L(\lambda)$ stetig invertierbar ist für alle $\lambda \in \overline{B_\epsilon(\lambda_0)} \setminus \{\lambda_0\}$. Da die Voraussetzungen des Satzes 7.4 für $\hat{\Lambda} := \partial B_\epsilon(\lambda_0)$ erfüllt sind, existiert ein $h_0 > 0$ und eine Konstante $C > 0$ sodass für alle $h \leq h_0$

$$\max_{\lambda \in \partial B_\epsilon(\lambda_0)} \|L_h(\lambda)^{-1}\| \leq C. \tag{7.31}$$

Sei nun $u_0 \in V \setminus \{0\}$ ein Eigenvektor zu λ_0, d.h. $L(\lambda_0)u_0 = 0$. Insbesondere ist auch $L_h(\lambda_0)u_0 = \Pi_h L(\lambda_0)u_0 = 0$. Es gilt $\Pi_h u_0 \to u_0$ für $h \to 0$, d.h. wir können annehmen, dass $\Pi_h u_0 \neq 0$ für alle $h \leq h_0$. Für alle $h \leq h_0$ mit $L_h(\lambda_0)\Pi_h u_0 = 0$ ist $(\lambda_0, \Pi_h u_0) \in \Lambda \times V_h \setminus \{0\}$ bereits ein Eigenpaar von (7.29) und die Behauptung für diese h gezeigt. Sei also $v := L_h(\lambda_0)(\Pi_h u_0 - u_0) = L_h(\lambda_0)\Pi_h u_0 \in V_h \setminus \{0\}$. Dann folgt

$$\|L_h(\lambda_0)^{-1}\| = \sup_{v \in V_h \setminus \{0\}} \frac{\|L_h(\lambda_0)^{-1} v\|}{\|v\|} \geq \frac{\|\Pi_h u_0\|}{\|L_h(\lambda_0)(\Pi_h u_0 - u_0)\|} \geq \frac{\|\Pi_h u_0\|}{\|L(\lambda_0)\|\|\Pi_h u_0 - u_0\|}. \tag{7.32}$$

Daher gilt für alle hinreichend kleinen $h \leq h_0$

$$\|L_h(\lambda_0)^{-1}\| > C \geq \max_{\lambda \in \partial B_\epsilon(\lambda_0)} \|L_h(\lambda)^{-1}\|. \tag{7.33}$$

Im Widerspruch zur Behauptung existiere nun kein Eigenwert von L_h in $\overline{B_\epsilon(\lambda_0)}$. Da L_h ein Fredholm-Operator und die Abbildung $\lambda \mapsto L_h(\lambda)$ holomorph ist, ist $L_h(\lambda)$ stetig invertierbar und mit Lem. 5.1 ist die Abbildung $\lambda \mapsto L_h(\lambda)^{-1}$ holomorph. Wegen des Maximumsprinzips folgt daraus

$$\|L_h(\lambda_0)^{-1}\| \leq \max_{\lambda \in \partial B_\epsilon(\lambda_0)} \|L_h(\lambda)^{-1}\| \leq C$$

im Widerspruch zu (7.33). □

Korollar 7.9. *Unter den Voraussetzungen des vorigen Satzes existiere für alle $h \leq h_0$ genau ein Eigenwert λ_h mit Riesz-Konstante $r \in \mathbb{N}$ in der Kugel $\overline{B_\epsilon(\lambda_0)}$. Dann existiert eine Konstante $C > 0$ mit*

$$|\lambda_h - \lambda_0| \leq C\|\Pi_h u_0 - u_0\|^{1/r} \tag{7.34}$$

für alle nichttrivialen Eigenvektoren u_0 zum Eigenwert λ_0.

Beweis. Die Abbildungen $\lambda \mapsto (\lambda - \lambda_h)^r L_h(\lambda)^{-1}$ sind in $\overline{B_\epsilon(\lambda_0)}$ holomorph. Mit Hilfe des Maximumsprinzips folgt für $\lambda = \lambda_0$

$$|\lambda_0 - \lambda_h|^r \|L_h(\lambda_0)^{-1}\| \leq \epsilon \sup_{\lambda \in \partial B_\epsilon(\lambda_0)} \|L_h(\lambda)^{-1}\| \tag{7.35}$$

und mit (7.31) und (7.32) die Behauptung. □

7.3. Integralmethode für nichtlineare Eigenwertprobleme

Sei nun für $\lambda \mapsto A(\lambda) \in \mathbb{C}^{N\times N}$ Eigenpaare $(\lambda, u) \in \mathbb{C} \times \mathbb{C}^N \setminus \{0\}$ von

$$A(\lambda)u = 0 \tag{7.36}$$

gesucht. Theoretisch könnten wir die Eigenwerte mit Hilfe eine Nullstellensuche von $\lambda \mapsto \det A(\lambda)$ berechnen. Dies ist aber sehr aufwändig und wird für größere N instabil. Vergleichsweise effizient und mathematisch sehr elegant ist dagegen die u.a. von Beyn in [Bey12] vorgeschlagene Methode über komplexe Wegintegrale. Wir benötigen dazu den Satz von Keldysh, welchen wir in der einfachst möglichen Form hier präsentieren. Für Erweiterungen sei auf [Bey12] verwiesen.

Satz 7.10 (Keldysh)**.** *Sei $\mathbb{C} \supset \Lambda \ni \lambda \mapsto A(\lambda) \in \mathbb{C}^{N\times N}$ holomorph und es existiere mindestens ein $\lambda \in \Lambda$, sodass $A(\lambda)$ invertierbar ist. Weiter sei $\lambda_1 \in \Lambda$ ein halb-einfacher Eigenwert, d.h. es existiere eine L-dimensionale Orthonormalbasis aus (Rechts-)Eigenvektoren $v_{1,1}, \dots, v_{1,L} \in \ker A(\lambda_1)$ und es gelte $A'(\lambda_1)v_{1,l} \notin A(\lambda_1)(C^N)$ für alle $l = 1, \dots, L$.*

Dann ist auch $\dim\ker(A^*(\lambda_1)) = L$, d.h. es existiert eine Basis $w_{1,1}, \dots, w_{1,L}$ von $\ker(A^*(\lambda_1))$, und diese Basis kann so gewählt werden, dass

$$w_{1,l}^* A'(\lambda_1)v_{1,k} = \delta_{lk}, \qquad l, k = 1, \dots L \tag{7.37}$$

gilt. Weiterhin existiert eine Umgebung U um λ_1 sodass

$$A(\lambda)^{-1} = (\lambda - \lambda_1)^{-1} \sum_{l=1}^{L} v_{1,l} w_{1,l}^* + R(\lambda), \qquad \lambda \in U, \tag{7.38}$$

wobei $\lambda \mapsto R(\lambda) \in \mathbb{C}^{N\times N}$ holomorph ist.

Nehmen wir nun an, dass im beschränkten Gebiet Λ nur endlich viele Eigenwerte liegen und dass diese alle halb-einfach sind. Dann ist die Funktion

$$R(\lambda) := A(\lambda)^{-1} - \sum_{n=1}^{k} (\lambda - \lambda_n)^{-1} \sum_{l=1}^{L_n} v_{n,l} w_{n,l}^*$$

in $\Lambda \setminus \{\lambda_1, \dots, \lambda_n\}$ holomorph. Da sie auch holomorph in den Umgebungen der isolierten Eigenwerte ist, folgt

$$A(\lambda)^{-1} = \sum_{n=1}^{k} (\lambda - \lambda_n)^{-1} \sum_{l=1}^{L_n} v_{n,l} w_{n,l}^* + R(\lambda), \qquad \lambda \in \Lambda, \tag{7.39}$$

mit einer auf ganz Λ holomorphen Funktion $\lambda \mapsto R(\lambda)$. Mit Hilfe des Residuen-Satzes folgt daraus für eine geschlosenne, positiv orientierte Jordan-Kurve Γ, welche

ganz in Λ liegt und in dessen Inneren sich die Eignwerte befinden, dass für beliebige holomorphe Funktionen $\lambda \mapsto f(\lambda) \in \mathbb{C}$ gilt

$$\frac{1}{2\pi \mathrm{i}} \int_\Gamma f(\lambda) A(\lambda)^{-1} d\lambda = \sum_{n=1}^{k} f(\lambda_n) \sum_{l=1}^{L_n} v_{n,l} w_{n,l}^*. \tag{7.40}$$

Diese Formel bildet die Grundlage zur Berechnung der Eigenwerte.

Sei dazu $J := \sum_{n=1}^{k} L_n$ die Vielfachheit der halbeinfachen Eigenwerte innerhalb von Γ (und wir nehmen an, es gibt keine weiteren Eigenwerte innerhalb von Γ). Dann definieren wir $V := (v_{1,1}, \dots v_{k,L_k}) \in \mathbb{C}^{N \times J}$, $W := (w_{1,1}, \dots w_{k,L_k}) \in \mathbb{C}^{N \times J}$ und wählen ein (zunächst beliebiges) $\hat{V} \in \mathbb{C}^{N \times j}$ mit $j \geq J$. Zudem nehmen wir an, dass sowohl j als auch J nicht größer als die Dimension N der Matrix sind. Typischerweise gilt $J \leq j \ll N$.

Nun definieren wir

$$A_0 : = \frac{1}{2\pi \mathrm{i}} \int_\Gamma A(\lambda)^{-1} \hat{V} d\lambda \in \mathbb{C}^{N \times j} \tag{7.41a}$$

$$A_1 : = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \lambda A(\lambda)^{-1} \hat{V} d\lambda \in \mathbb{C}^{N \times j} \tag{7.41b}$$

und erhalten mit (7.40)

$$A_0 = V W^* \hat{V}, \qquad A_1 = V D W^* \hat{V} \text{ mit } D := \operatorname{diag}(\lambda_1, \dots, \lambda_n), \tag{7.42}$$

wobei in D die Eigenwerte ihrer Vielfachheit nach aufgeführt werden. Wir erkennen bereits, dass Informationen über die gewünschten Eigenwerte in den Matrizen A_0 und A_1 enthalten sind. Die Frage bleibt, wie diese Matrizen zu berechnen sind und wie man dann an die Eigenwerte kommt.

Betrachten wir zunächst den zweiten Teil. Dazu führen wir eine Singulärwertzerlegung von A_0 durch, d.h. wir erhalten

$$A_0 = \tilde{V} \Sigma \tilde{W}^*$$

mit Matrizen $\tilde{V} \in \mathbb{C}^{N \times N}$, $\tilde{W} \in \mathbb{C}^{j \times j}$ und einer verallgemeinerter Diagonalmatrix $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_j) \in \mathbb{C}^{N \times j}$.

Wir nehmen nun an, dass $V, W \in \mathbb{C}^{N \times J}$ vollen Spaltenrang J haben. Da die Matrizen die Rechts- bzw. Linkseigenvektoren enthalten, erscheint diese Annahme sinnvoll. Leider müssen bei nichtlinearen Eigenwertproblemen die Eigenvektoren zu unterschiedlichen Eigenwerten nicht linear unabhängig sein. Daher ist die hier gemachte Annahme nicht immer erfüllt. Notfalls müsste man eine andere Kurve Γ wählen, sodass weniger Eigenwerte innerhalb der Kurve liegen.

Wählen wir nun $\hat{V}$ als eine hinreichend große Zufallsmatrix, so können wir annehmen, dass A_0 ebenfalls den maximalen Rang J hat. Daher verschwinden die letzten Singulärwerte

$$\sigma_{J+1} = \dots = \sigma_j = 0. \tag{7.43}$$

Somit können wir statt der vollen Singulärwertzerlegung eine reduzierte Singulärwertzerlegung verwenden:

$$A_0 = \tilde{V}\Sigma\tilde{W}^*, \qquad \tilde{V} \in \mathbb{C}^{N\times J}, \tilde{W} \in \mathbb{C}^{j\times J}, \Sigma \in \mathbb{C}^{J\times J}, \tag{7.44}$$

wobei $\tilde{V}^*\tilde{V} = \tilde{W}^*\tilde{W} = \mathrm{id}_{J\times J}$ gilt.

Nun ist $S := \tilde{V}^*V \in \mathbb{C}^{J\times J}$ aufgrund unserer Annahme an den Rang von V regulär wir erhalten

$$\Sigma\tilde{W}^* = \tilde{V}^*A_0 = \tilde{V}^*VW^*\hat{V} = SW^*\hat{V} \quad \Rightarrow W^*\hat{V} = S^{-1}\Sigma\tilde{W}^*. \tag{7.45}$$

Einsetzen in die Darstellung von A_1 aus (7.42) und Multiplikation von links mit $\tilde{V}^*$ und von rechts mit $W\Sigma^{-1}$ ergibt

$$SDS^{-1} = \tilde{V}A_1\tilde{W}\Sigma^{-1}. \tag{7.46}$$

Somit ist die Diagonalmatrix D, welche ja die gesuchten Eigenwerte enthält, ähnlich zur Matrix $\tilde{V}A_1\tilde{W}\Sigma^{-1}$. Sofern wir A_0 und A_1 berechnen (oder zumindest approximieren können), haben wir somit einen Algorithmus gefunden:

1. Berechne $A_0 \in \mathbb{C}^{N\times j}$
2. Berechne reduzierte Singulärwertzerlegung $A_0 = \tilde{V}\Sigma\tilde{W}^*$ mit J Singulärwerten
3. Berechne $A_1 \in \mathbb{C}^{N\times j}$
4. Berechne die Eigenwerte der Matrix $\tilde{V}A_1\tilde{W}\Sigma^{-1} \in \mathbb{C}^{j\times j}$ (z.B. mit QR-Verfahren)

Der Aufwand im letzten Punkt ist vernachlässigbar, wenn j vergleichsweise klein ist. Eine reduzierte Singulärwertzerlegung kann man mit einem Krylovraum-Verfahren erhalten. Hier ist zu beachten, dass ebenfalls nur j und damit wenige Singulärwerte und die dazugehörigen Vektoren benötigt werden. Daher ist auch dieser Aufwand vergleichsweise überschaubar. Anspruchsvoll ist dagegen die (näherungsweise) Berechnung von A_0 und A_1, welche wir nun angehen wollen.

Bemerkung 7.11. *Die Matrix S und die Diagonalmatrix D werden im Algorithmus nicht benötigt. Dies ist wesentlich, da sie ohne Kenntnis der gesuchten Eigenpaare nicht berechenbar sind. Desweiteren haben wir angenommen, dass $V, W, \hat{V}$ und damit auch A_0 vollen Spaltenrang haben. Wäre das nicht der Fall, so wäre die Matrix S nicht regulär und diese Herleitung führt nicht zum gewünschten Ergebnis. Da S in der Berechnung selber nicht vorkommt, könnte man hoffen, dass dies in praktischen Berechnungen keine Rolle spielt. Dem ist aber nicht so. Wählt man z.B. eine zu kleine Matrix $\hat{V}$, so sind die numerischen Resultate in der Tat inkorrekt.*

Zur Berechnung von A_0 und A_1 mit Hilfe einer geeigneten Quadraturformel ist folgendes Lemma sehr hilfreich.

Lemma 7.12. *Sei $R > 0$, $1 < a_- < a_+$ und $D := \{z \in \mathbb{C} : a_-^{-1} < |z|/R < a_+\}$ ein Ringgebiet. Weiter sei $D \ni \lambda \mapsto f(\lambda) \in \mathbb{C}^N$ eine holormorphe Funktion und*

$$Q(f) := \frac{1}{2\pi\mathrm{i}} \int_{|\lambda|=R} f(\lambda)d\lambda.$$

Dann gilt für die summierte Rechtecksregel (hier gleichbedeutend mit einer summierten Trapezregel)

$$Q_m(f) := \frac{R}{m} \sum_{\nu=0}^{m-1} \omega_m^\nu f(R\omega_m^\nu), \qquad \omega_m := \exp\left(\frac{2\pi\mathfrak{i}}{m}\right),$$

mit $m \in \mathbb{N}$ Punkten für alle $\rho_\pm \in (1, a_\pm)$ die Fehlerabschätzung

$$|Q_m(f) - Q(f)| \leq \max_{|\lambda|=\rho_+ R} \|f(\lambda)\| \frac{\rho_+^{-m}}{1-\rho_+^{-m}} + \max_{|\lambda|=\rho_- R} \|f(\lambda)\| \frac{\rho_-^{-m}}{1-\rho_-^{-m}}. \tag{7.47}$$

Beweis. Wir verwenden die Laurent-Entwicklung

$$f(\lambda) = \sum_{\mu=-\infty}^{\infty} \lambda^\mu f_\mu, \qquad f_\mu := \frac{1}{2\pi\mathfrak{i}} \int_{|\lambda|=R} \lambda^{-\mu-1} f(\lambda) d\lambda, \tag{7.48}$$

welche aufgrund der Holomorphie für alle λ in einer kompakten Teilmenge von D gleichmäßig konvergiert. Durch direktes Nachrechnen erhält man

$$Q_m(\lambda \mapsto \lambda^\mu) - Q(\lambda \mapsto \lambda^\mu) = \begin{cases} R^{\mu+1}, & \mu + 1 = m\mathbb{Z} \setminus \{0\} \\ 0, & \text{sonst} \end{cases}.$$

Aufgrund der Linearität und Stetigkeit des Integrals und der Quadraturformel erhalten wir

$$Q_m(f) - Q(f) = \sum_{l=1}^{\infty} \left(f_{lm} R^{lm} + f_{-lm} R^{-lm} \right). \tag{7.49}$$

Mit Hilfe des Integralsatzes von Cauchy können wir die Kurve in der Berechnung von $f_{\pm lm}$ innerhalb von D beliebig verändern. So erhalten wir

$$\left| f_{lm} R^{lm} \right| = \left| \frac{R^{lm}}{2\pi\mathfrak{i}} \int_{|\lambda|=\rho_+ R} \lambda^{-lm-1} f(\lambda) d\lambda \right| \leq \max_{|\lambda|=\rho_+ R} \|f(\lambda)\| \rho_+^{-lm}$$

und analog

$$\left| f_{lm} R^{lm} \right| \leq \max_{|\lambda|=\rho_- R} \|f(\lambda)\| \rho_-^{-lm}.$$

Zusammen ergibt sich die Behauptung. □

Das Lemma besagt, dass die zusammengesetzte Rechtecksregel für holomorphe Funktionen auf Kreisen exponentiell konvergiert, wenn nur die Quadraturpunkte gleichmäßig auf dem Kreis verteilt werden. Es lässt sich problemlos auf andere geschlossene Kurven übertragen, sofern diese (und deren Inneres) das Bild des abgeschlossenen Einheitskreises unter einer holomorphen Abbildung sind. Die Konvergenzgeschwindigkeit hängt offenbar vom Abstand der Kurve zu den am nächsten liegenden Polen der Funktion im Inneren (ρ_-) bzw. im Äußeren (ρ_+).

Die Details dazu sind in [Bey12] ausgeführt. Wir zitieren hier nur das wesentliche Resultat.

Satz 7.13. *Werden die Kurven in der Definition (7.41) von A_0 und A_1 als Kreise der Form $z_0 + R\exp(\mathrm{i}t)$ mit $t \in [0, 2\pi)$ gewählt und durch eine Rechtecksregel mit m äquidistant verteilten Punkten diskretisiert, so erhalten wir für die angenäherten Matrizen $A_{0,1}^{(m)} \in \mathbb{C}^{N\times j}$ die Fehlerabschätzung*

$$\left\| A_{0,1}^{(m)} - A_{0,1} \right\| \leq C \left(\rho_-^{m-r+1} + \rho_+^{m-r+1} \right),$$

wobei r die maximale Ordnung der Pole von $A(\lambda)^{-1}$ und

$$\rho_- := \max_{\lambda\in\sigma(A),|\lambda-z_0|<R} \frac{|\lambda - z_0|}{R}, \qquad \rho_+ := \max_{\lambda\in\sigma(A),|\lambda-z_0|>R} \frac{R}{|\lambda - z_0|}$$

ist.

Die Matrizen A_0 und A_1 können wir somit exponentiell schnell mit Hilfe der Rechtecksregel annähern, sofern wir nur glatte Kurve und passende Quadraturpunkte wählen. Dennoch ist auch die Berechnung der angenäherten Matrizen $A_{0,1}^{(m)}$ sehr aufwändig. Für jeden Quadraturpunkt und jede Spalte von $\hat{V}$ muss ein lineares Gleichungssystem gelöst werden. Dies entspricht somit mj linearen Gleichungssystemen der Größe $N \times N$. Verwendet man eine LU-Zerlegung von $A(\lambda)$ und nehmen wir für diese den Aufwand N^3 an, so erhalten wir in etwa einen Gesamtaufwand zur Berechnung von $A_0^{(m)}$ und $A_1^{(m)}$ von

$$m(N^3 + 2N^2 j).$$

Die Integralmethode zur Berechnung der Eigenwerte nichtlinearer Eigenwertprobleme ist vergleichsweise neu und Gegenstand aktueller Forschung. Ein wesentliches Problem der Methode (neben dem Aufwand) ist, dass sich durch die Verwendung von $A_{0,1}^{(m)}$ die Anzahl der nicht-verschwindenden Singulärwerte (siehe (7.43)) deutlich erhöht. Solange die "korrekten" Singulärwerte um Größenordnungen über den zusätzlichen liegen, können letztere einfach aussortiert werden. Gerade bei großen Fehlern, wie sie z.B. durch Verwendung von Matrix-Kompressionstechniken (z.B. H-Matrizen) oder durch iterative Gleichungslöser entstehen, sind die zusätzlichen Singulärwerte aber häufig nicht mehr von den "korrekten" zu unterscheiden und führen dann zu zusätzlichen falschen Eigenwerten und zu großen Fehlern bei den eigentlich gesuchten Eigenwerten.

Dennoch hat die Integralmethode die Lösung von nichtlinearen Eigenwertproblemen massiv vereinfacht. Sie ermöglicht bei entsprechender Genauigkeit die zuverlässige Berechnung aller Eigenwerte innerhalb einer gewählten Kurve. Dies ist ein sehr großer Vorteil zu iterativen Verfahren, bei denen typischerweise die Lage der gefundenen Eigenwerte nicht zuverlässig kontrolliert werden kann.

A. QR-Verfahren

Der Basisalgorithmus des QR-Verfahrens ist in Alg. A.1 dargestellt. Sollte die Matrix keine Hessenberg-Matrix sein, ist es zur Reduktion des Aufwandes sinnvoll, sie zunächst mit Ähnlichkeitstransformationen (z.B. mit Householder-Matrizen) in eine solche zu bringen. Zum Nachweis der Konvergenz benötigen wir zunächst zwei Hilfsresultate.

Lemma A.2. *Die Matrizen $H^{(t)}$ des QR-Verfahrens sind paarweise ähnlich und besitzen dieselben Eigenwerte.*

Beweis. Per Induktion über $t \in \mathbb{N}_0$ wegen

$$H^{(t+1)} = \underbrace{{Q^{(t)}}^{-1} Q^{(t)}}_{=\mathrm{id}} R^{(t)} Q^{(t)} = {Q^{(t)}}^{-1} H^{(t)} Q^{(t)}.$$

□

Lemma A.3. *Seien $A = Q_1 R_1$ und $A = Q_2 R_2$ zwei QR-Zerlegungen einer regulären Matrix $A \in \mathbb{C}^{n\times n}$. Dann existiert eine unitäre Diagonalmatrix $S \in \mathbb{C}^{n\times n}$ mit $Q_1 = Q_2 S^*$ und $R_1 = S R_2$.*

Beweis. Da A und somit auch R_2 regulär und $Q_1^{-1} = Q_1^*$ ist, gilt

$$Q_1^* Q_2 = R_1 R_2^{-1},$$

d.h. die obere Dreiecksmatrix $S := R_1 R_2^{-1}$ ist unitär. Da die Spalten von S orthogonal zueinander sind, ist S eine Diagonalmatrix. □

Algorithmus A.1 QR-Verfahren: Basisalgorithmus

Input: Hessenberg-Matrix $H \in \mathbb{K}^{m\times m}$

1: $H^{(0)} = H$
2: **for** $t = 1, \ldots$ **do**
3: Berechne QR-Zerlegung von $H^{(t-1)} = Q^{(t-1)} R^{(t-1)}$ mittels Givens-Rotationen
4: $H^{(t)} = R^{(t-1)} Q^{(t-1)}$
5: **end for**

Output: $H^{(t)}$ hat unter gewissen Voraussetzungen Approximationen an die Eigenwerte von H auf der Hauptdiagonalen

Satz A.4 (QR-Algorithmus). *Sei $H \in \mathbb{K}^{m\times m}$ diagonalisierbar mit Eigenwerten $\lambda_1, \dots, \lambda_m \in \mathbb{K}$ und $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_m| > 0$. Weiter seien $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m) \in \mathbb{K}^{m\times m}$ und $V = (v_1, \dots, v_m) \in \mathbb{K}^{m\times m}$ die Matrix aus Eigenvektoren v_j zu den Eigenwerten λ_j für $j = 1, \dots, m$. Zudem existiere eine untere, normierte Dreiecksmatrix L und eine obere Dreiecksmatrix U sodass $V^{-1} = LU$.*

Dann konvergieren die Hauptdiagonaleneinträge der Matrizen $A^{(t)}$ aus Alg. A.1 linear gegen die Eigenwerte von A.

Beweis. Wegen Lemma A.2 genügt es zu zeigen, dass $H^{(t)}$ im gewissen Sinne gegen eine obere Dreiecksmatrix konvergiert. Dann stehen die Eigenwerte von $H^{(t)}$ und damit auch die von H auf der Hauptdiagonalen von $H^{(t)}$. Der Beweis erfolgt in mehreren Schritten.

1. QR-Zerlegung von H^t: Wir unterscheiden die Matrizen $H^{(t)}$ aus dem QR-Verfahren und die t-te Potenz H^t. Für H^t gilt mit $V^{-1} = LU$

$$H^t = \left(V\Lambda V^{-1}\right)^t = V\Lambda^t V^{-1} = V\Lambda^t LU = \underbrace{V\Lambda^t L\Lambda^{-t}}_{=:V_t}\Lambda^t U = V_t\Lambda^t U.$$

Da die Eigenwerte nicht verschwinden, ist Λ regulär und somit V_t wohldefiniert und regulär. Sei $V_t = \tilde{Q}_t\tilde{R}_t$ eine QR-Zerlegung von V_t. Dann ist auch $\tilde{R}_t$ regulär, $\tilde{R}_t\Lambda^t U$ eine obere Dreiecksmatrix und

$$H^t = \tilde{Q}_t\left(\tilde{R}_t\Lambda^t U\right) \tag{A.1}$$

eine QR-Zerlegung von H^t. Insbesondere ist H^t regulär.

2. QR-Zerlegung von H^t: Sei $Q_t := Q^{(0)}\cdots Q^{(t-1)}$ und $R_t := R^{(t-1)}\cdots R^{(0)}$. Dann folgt aus Lemma A.2 induktiv $H^{(t)} = Q_t^* H Q_t$ und

$$H^t = Q_t R_t \tag{A.2}$$

ist eine weitere QR-Zerlegung von H^t, da

$$Q_t R_t = \underbrace{Q^{(0)}\cdots Q^{(t-2)}}_{=Q_{t-1}}\ \underbrace{Q^{(t-1)}R^{(t-1)}}_{=H^{(t-1)}=Q_{t-1}^*HQ_{t-1}}\ \underbrace{R^{(t-2)}\cdots R^{(0)}}_{=R_{t-1}} = H\ \underbrace{Q_{t-1}R_{t-1}}_{=H^{t-1}\text{ per Induktion}} = H^t.$$

Darstellung von $H^{(t)}$: Mit Hilfe von Lemma A.3 existieren unitäre Diagonalmatrizen S_t mit $Q_t = \tilde{Q}_t S_t^*$ und $R_t = S_t\tilde{R}_t\Lambda^t U$, da H^t regulär ist. Es folgt

$$\begin{aligned} Q^{(t)} &= Q_t^* Q_{t+1} = S_t\tilde{Q}_t^*\tilde{Q}_{t+1}S_{t+1}^*, \\ R^{(t)} &= R_{t+1}R_t^{-1} = \left(S_{t+1}\tilde{R}_{t+1}\Lambda^{t+1}U\right)\left(U^{-1}\Lambda^{-t}\tilde{R}_t^{-1}S_t^*\right) = S_{t+1}\tilde{R}_{t+1}\Lambda\tilde{R}_t^{-1}S_t^* \end{aligned}$$

und daraus

$$\begin{aligned} H^{(t)} =& Q^{(t)}R^{(t)} = S_t\tilde{Q}_t^*\tilde{Q}_{t+1}\underbrace{S_{t+1}^* \quad S_{t+1}}_{=\mathrm{id}}\tilde{R}_{t+1}\Lambda\tilde{R}_t^{-1}S_t^* \\ =& S_t\left(\tilde{R}_t\tilde{R}_t^{-1}\right)\tilde{Q}_t^*\underbrace{\tilde{Q}_{t+1} \quad \tilde{R}_{t+1}}_{=V_{t+1}}\Lambda\tilde{R}_t^{-1}S_t^* \\ =& S_t\tilde{R}_t\underbrace{(\tilde{Q}_t\tilde{R}_t)^{-1}}_{=V_t^{-1}}V_{t+1}\Lambda\tilde{R}_t^{-1}S_t^* = S_t\tilde{R}_tV_t^{-1}V_{t+1}\Lambda\tilde{R}_t^{-1}S_t^*. \end{aligned} \tag{A.3}$$

Asymptotik von V_t: Nach Definition gilt $V_t = V\Lambda^t L\Lambda^{-t}$ mit der normierten unteren Dreiecksmatrix L und der Diagonalmatrix Λ, bei der die Diagonaleinträge der Größe des Betrages nach fallend geordnet sind. Somit gilt

$$\Lambda^t L\Lambda^{-t} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21}\left(\frac{\lambda_2}{\lambda_1}\right)^t & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{m1}\left(\frac{\lambda_m}{\lambda_1}\right)^t & \dots & l_{m(m-1)}\left(\frac{\lambda_m}{\lambda_{m-1}}\right)^t & 1 \end{pmatrix}, \tag{A.4}$$

d.h. $\Lambda^t L\Lambda^{-t} = \mathrm{id} + E_t$ mit $\|E_t\| = \mathcal{O}(q^t)$ für $t \to \infty$ und einem $q < 1$. Entsprechend folgt

$$V_t = V + VE_t \qquad \text{mit} \quad \|VE_t\| = \mathcal{O}(q^t), \qquad t \to \infty.$$

Asymptotik von $H^{(t)}$: Mit Hilfe der Asymptotik von V_t erhält man

$$V_t^{-1}V_{t+1} = (V + VE_t)^{-1}(V + VE_{t+1}) = \mathrm{id} + F_t \quad \text{mit} \quad \|F_t\| = \mathcal{O}(q^t), \quad t \to \infty.$$

Die Darstellung von $H^{(t)}$ aus (A.3) wird damit zu

$$H^{(t)} = S_t\tilde{R}_t\Lambda\tilde{R}_t^{-1}S_t^* + S_t\tilde{R}_tF_t\Lambda\tilde{R}_t^{-1}S_t^*.$$

Der zweite Term konvergiert gegen Null, da wegen $\|V_t\| = \|\tilde{Q}_t\tilde{R}_t\| = \|\tilde{R}_t\|$ und $\|V_t^{-1}\| = \|\tilde{R}_t^{-1}\|$ gilt

$$\left\|S_t\tilde{R}_tF_t\Lambda\tilde{R}_t^{-1}S_t^*\right\| \le |\lambda_1|\,\mathrm{cond}(V_t)\|F_t\| = \mathcal{O}(q^t), \qquad t \to \infty.$$

Da die Hauptdiagonaleneinträge der oberen Dreiecksmatrix $S_t\tilde{R}_t\Lambda\tilde{R}_t^{-1}S_t^*$ von t unabhängig und gleich den Eigenwerten λ_j von A sind, konvergieren die Hauptdiagonaleneinträge von $A^{(t)}$ linear gegen $\lambda_1, \dots, \lambda_m$ und die Behauptung ist gezeigt. □

Die Voraussetzungen des vorigen Beweises sind zum Teil nicht notwendig, wie folgende Überlegungen zeigen:

1. Die Voraussetzung, dass V^{-1} eine LU-Zerlegung besitzt, ist lediglich für diese Version des Konvergenzbeweises des QR-Verfahrens notwendig. V ist invertierbar, sodass immer eine Permutationsmatrix P existiert mit $PV^{-1} = LU$.

2. Wenn $H \in \mathbb{R}^{m\times m}$ so wird gefordert, dass alle Eigenwerte ebenfalls reell sind. Dies ist im Allgemeinen nicht der Fall. Zudem haben konjugiert komplexe Eigenwerte den gleichen Betrag und verstoßen damit gegen die Bedingung $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_m|$. Da das asymptotische Verhalten von $H^{(t)}$ auf der Gleichung (A.4) beruht, werden die Matrizen $H^{(t)}$ z.B. im Falle $m = 6$ und $|\lambda_1| > |\lambda_2| > |\lambda_3| = \cdots = |\lambda_5| > |\lambda_6|$ gegen eine eine Matrix folgender Form konvergieren:

$$\begin{pmatrix} \lambda_1 & \star & \dots & & & \star \\ 0 & \lambda_2 & \star & \dots & & \star \\ 0 & 0 & D_{11} & D_{12} & D_{13} & \star \\ 0 & 0 & D_{21} & D_{22} & D_{23} & \star \\ 0 & 0 & D_{31} & D_{32} & D_{33} & \star \\ 0 & 0 & 0 & 0 & 0 & \lambda_6 \end{pmatrix}.$$

   Dies ist eine Block-Dreiecksmatrix mit Diagonalblöcken in der Größe der Anzahl der betragsgleichen Eigenwerte. Die Eigenwerte von H können dann aus den Eigenwerten dieser Diagonalblöcke berechnet werden.

3. Die Voraussetzung $|\lambda_m| > 0$ ist nicht einschränkend. Falls nötig, kann wie im vorigen Abschnitt an Stelle der Matrix H der QR-Algorithmus auf die Matrix $H - \rho\,\mathrm{id}$ mit einem geeigneten shift-Parameter ρ angewendet werden. Der gleiche Trick kann verwendet werden, wenn die Matrix unterschiedliche Eigenwerte mit gleichem Betrag hat. Durch einen komplexen shift ist typischerweise zu erreichen, dass sich die verschobenen Eigenwerte vom Betrag unterscheiden.

Die letzte Überlegung kann auch zur Beschleunigung der Konvergenz eingesetzt werden. Diese hängt maßgeblich von den Quotienten der Eigenwerte $|\lambda_j/\lambda_{j+1}|$ ab. Durch geschickte Wahl des shiftes, kann dieser erheblich verkleinert werden. Betrachen wir dazu wieder (A.4) und darin die letzte Zeile. Da $\left(H^{(t)}\right)_{mm}$ eine Näherung an den betragskleinsten Eigenwert von H darstellt, könnten wir in jedem Schritt $\rho^{(t)} = \left(H^{(t)}\right)_{mm}$ wählen und erhalten so eine sehr schnelle Konvergenz der letzten Zeile von $H^{(t)}$ gegen $(0,\dots,0,\lambda_m)$, da die Quotienten $|\frac{\lambda_m-\rho^{(t)}}{\lambda_j-\rho^{(t)}}|$ für $j = 1,\dots,m-1$ sehr klein sind. Sobald der mutmaßlich größte Eintrag $\left(H^{(t)}\right)_{m(m-1)}$ in der letzten Zeile von $H^{(t)}$ unter einer vorgegebenen Toleranz ist, können wir eine Zeile nach oben rücken und $\rho^{(t)} = \left(H^{(t)}\right)_{(m-1)(m-1)}$ wählen.

In Alg. A.5 ist dieses Vorgehen mit einer kleinen Änderung zur Wahl des shifts beschrieben. Für diese ist die Konvergenzbeschleunigung noch größer als bei der skizzierten Wahl des shifts.

Aufwand: Der Hauptaufwand des QR-Verfahrens ist die Berechnung der QR-Zerlegung in jedem Schritt. Für eine beliebige Matrix $A \in \mathbb{K}^{m\times m}$ benötigt diese $\mathcal{O}(m^3)$ Rechenoperationen und ist somit sehr aufwendig. Bei Hessenberg-Matrizen lässt sich der Aufwand der QR-Uerlegung auf $\mathcal{O}(m^2)$ Rechenoperationen reduzieren.

Algorithmus A.5 QR-Verfahren mit shifts

Input: $H \in \mathbb{C}^{m\times m}$ mit m Eigenwerten mit paarweise unterschiedlichem Betrag, **TOL** > 0 vorgegebene Toleranz

for $j = m, m-1, \ldots, 2$ **do**
  while $|H(j-1,j)| >$ **TOL** $(|H(j-1,j-1)| + |H(j,j)|)$ **do**
    Berechne Eigenwerte ρ_1, ρ_2 von $\begin{pmatrix} H(j-1,j-1) & H(j-1,j) \\ H(j,j-1) & H(j,j) \end{pmatrix}$
    if $|\rho_1 - H(j,j)| < |\rho_2 - H(j,j)|$ **then**
      $\rho = \rho_1$
    else
      $\rho = \rho_2$
    end if
    Berechne QR-Zerlegung von $H - \rho\,\mathrm{id} = QR$
    $H = RQ + \rho\,\mathrm{id}$
  end while
end for

Output: H wird mit einer unitär äquivalenten oberen Dreiecksmatrix überschrieben

Definition A.6. *Eine $m \times m$-Matrix der Form*

$$G(j,k,c,s) = \left(\begin{array}{c|c|c} \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} & & \\ \hline & \begin{matrix} \overline{c} & & \overline{s} \\ & 1 & \\ & & \ddots \\ & & & 1 \\ -s & & & c \end{matrix} & \\ \hline & & \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} \end{array}\right) \tag{A.5}$$

heißt Givens-Rotation*, falls $|c|^2 + |s|^2 = 1$. Dabei grenzen die durchgezogenen Linien den Bereich von der j-ten bis zur k-ten Zeile, bzw. Spalte ab.*

Givens-Rotationen sind unitär, da

$$\begin{pmatrix} \overline{c} & \overline{s} \\ -s & c \end{pmatrix} \begin{pmatrix} c & -\overline{s} \\ s & \overline{c} \end{pmatrix} = \begin{pmatrix} |c|^2 + |s|^2 & 0 \\ 0 & |c|^2 + |s|^2 \end{pmatrix}.$$

Der wichtigste Fall ist, dass c und s reell sind. In diesem Fall gibt es ein $\theta \in [0, 2\pi)$ mit $c = \cos\theta$ und $s = \sin\theta$, und die Matrix $\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$ beschreibt eine Drehung in der Ebene $\mathbb{R}^2$ um den Winkel θ.

Offenbar bewirkt eine Multiplikation GH einer Givens-Rotation $G = G(j,k,c,s)$ von links an eine Matrix $H \in \mathbb{K}^{n\times n}$ eine Ersetzung der j-ten und k-ten Zeile h_j^*, bzw. h_k^* von H durch $\overline{c}h_j^* + \overline{s}h_k^*$, bzw. $-sh_j^* + ch_k^*$.

Die Bedeutung der Givens-Rotationen liegt darin, dass man zu gebenen Indizes j,k,l mit $j \neq k$ und gegebener Matrix $H \in \mathbb{K}^{m\times m}$ eine Givens-Rotation $G(j,k,c,s)$ der Form (A.5) finden kann, so dass $(GH)_{k,l} = 0$. Dazu ist die Gleichung $-sa_{jl} + ca_{kl} = 0$ unter der Nebenbedingung $|c|^2 + |s|^2 = 1$ zu lösen. Dieses Problem besitzt zwei Lösungen, die sich durch das Vorzeichen unterscheiden. Eine der Lösungen lautet

$$\begin{pmatrix} c \\ s \end{pmatrix} = \mathbf{rot}(h_{jl}, h_{kl}) := \frac{1}{\sqrt{|h_{jl}|^2 + |h_{kl}|^2}} \begin{pmatrix} h_{jl} \\ h_{kl} \end{pmatrix}. \tag{A.6}$$

Um einen over- oder underflow zu vermeiden, benutzt man üblicherweise die äquivalenten Formeln

$$\begin{aligned} c &= \frac{h_{jl}/|h_{jl}|}{\sqrt{1+|t|^2}}, \quad s = \frac{t}{\sqrt{1+|t|^2}}, \quad t = \frac{h_{kl}}{|h_{jl}|}, && \text{für } |h_{jl}| \geq |h_{kl}|, \\ c &= \frac{t}{\sqrt{1+|t|^2}}, \quad s = \frac{h_{kl}/|h_{kl}|}{\sqrt{1+|t|^2}}, \quad t = \frac{h_{jl}}{|h_{kl}|}, && \text{für } |h_{jl}| < |h_{kl}|. \end{aligned}$$

Um eine QR-Zerlegung einer Hessenberg-Matrix $H \in \mathbb{K}^{m\times m}$ zu berechnen, definieren wir nun

$$H^{(k)} := G(k,k+1,c_k,s_k)H^{(k-1)} \qquad \text{mit} \qquad (c_k,s_k)^\top = \mathbf{rot}(H_{k,k}^{(k-1)}, H_{k+1,k}^{(k-1)})$$

für $k = 1,\ldots,m-1$ und $H^{(0)} := H$. Dieses Vorgehen veranschaulichen wir für $m = 4$ in folgendem Diagramm, indem wir die Matrixelemente, die in einem Schritt verändert werden, mit $*$ kennzeichnen und die übrigen von Null verschiedenen Elemente mit $+$:

$$H = \begin{pmatrix} ++++ \\ ++++ \\ \;+++ \\ \;\;++ \end{pmatrix} \xrightarrow{k=1} \begin{pmatrix} * & * & * & * \\ & * & * & * \\ & + & + & + \\ & & + & + \end{pmatrix} \xrightarrow{k=2} \begin{pmatrix} ++++ \\ \;\;*** \\ \;\;\;** \\ \;\;\;++ \end{pmatrix} \xrightarrow{k=3} \begin{pmatrix} ++++ \\ \;\;+++ \\ \;\;\;** \\ \;\;\;\;* \end{pmatrix} = R$$

In Alg. A.7 ist dieses Vorgehen zusammengefasst. Die Anzahl der Multiplikationen zur Durchführung des Verfahrens beträgt etwa

$$\sum_{k=1}^{m-1} 4(m-k+1) = \sum_{k=2}^{m} 4k \approx 2m^2$$

und ist damit um einen Faktor $m/3$ geringer als bei einer QR-Zerlegung mit Householder-Matrizen.

Algorithmus A.7 QR-Zerlegung einer Hessenberg-Matrix mit Givens-Rotationen

Input: $H = (h_{jk}) \in \mathbb{K}^{m \times m}$ Hessenberg-Matrix

1: **for** $k = 1, \ldots, m-1$ **do**
2: $\quad (c_k, s_k)^\top := \mathbf{rot}(h_{kk}, h_{k+1,k})$
3: $\quad$ **for** $l = l, \ldots, m$ **do**
4: $\quad\quad \begin{pmatrix} a_{k,l} \\ h_{k+1,l} \end{pmatrix} := \begin{pmatrix} \overline{c_k} & \overline{s_k} \\ -s_k & c_k \end{pmatrix} \begin{pmatrix} h_{k,l} \\ h_{k+1,l} \end{pmatrix}$
5: $\quad$ **end for**
6: **end for**

Output: H wird überschrieben mit der oberen Dreiecksmatrix $R = G(m-1, m, c_{m-1}, s_{m-1}) \cdots G(1, 2, c_1, s_1)H$

B. Riesz Theorie

Wir betrachen in diesem Abschnitt sogenannte Operator-Gleichungen 2.Art der Form

$$\lambda\varphi - A\varphi = f \tag{B.1}$$

mit einem kompakten linearen Operator $A : X \to X$ eines normierten linearen Raums X und $\lambda \in \mathbb{C} \setminus \{0\}$. Dabei richten wir uns nach [Kre99].

Sei

$$L := \mathrm{id} - A, \tag{B.2}$$

wobei id die Identität auf X bezeichnet.

Satz B.1 (1.Riesz)**.** *Der Nullraum des Operators L*

$$\ker(L) := \{\varphi \in X \mid L\varphi = 0\}$$

ist ein endlich-dimensionaler Teilraum von X.

Beweis. Der Nullraum des beschränkten linearen Operators L ist ein abgeschlossener Unterraum von X, denn für jede Folge $(\varphi_n)_n$ mit $\varphi_n \to \varphi$ und $L\varphi_n = 0$ gilt infolge der Stetigkeit $L\varphi = 0$. Für alle $\varphi \in \ker(L)$ ist $A\varphi = \varphi$ und dahher stimmt die Einschränkung von A auf $\ker(L)$ überein mit der Identität $A|_{\ker(L)} = \mathrm{id} : \ker(L) \to \ker(L)$. Der Operator A ist kompakt auf X und daher auch kompakt von $\ker(L)$ nach $\ker(L)$, da $\ker(L)$ abgeschlossen ist. Folglich ist $\ker(L)$ endlich-dimensional, da die Identität genau auf endlich-dimensionalen Räumen kompakt ist. □

Satz B.2 (2. Riesz)**.** *Der Bildraum*

$$L(X) := \{L\varphi \mid \varphi \in X\} \tag{B.3}$$

des Operators L ist ein abgeschlossener Unterraum.

Beweis. Der Bildraum des linearen Operators L ist ein Unterraum. Sei f ein Element aus der abgeschlossenen Hülle $\overline{L(X)}$. Dann gibt es eine Folge $(\varphi_n)_n$ aus X mit $L\varphi_n \to f$ für $n \to \infty$. Da $\ker(L)$ endlich-dimensional und damit abgeschlossen ist, gibt es zu jedem φ_n eine beste Approximation ψ_n in $\ker(L)$, d.h.

$$\|\varphi_n - \psi_n\| = \inf_{\psi \in \ker(L)} \|\varphi_n - \psi\|.$$

Dann ist die durch

$$\widetilde{\varphi}_n := \varphi_n - \psi_n$$

definierte Folge beschränkt. Wir beweisen dies indirekt und nehmen an, die Folge sei nicht beschränkt. Dann gibt es eine Teilfolge $(\widetilde{\varphi}_{n(k)})_k$ mit $\|\widetilde{\varphi}_{n(k)}\| \geq k$ für alle $k \in \mathbb{N}$. Nun setzen wir

$$\chi_k := \frac{\widetilde{\varphi}_{n(k)}}{\|\widetilde{\varphi}_{n(k)}\|}, \qquad k \in \mathbb{N}.$$

Da $\|\chi_k\| = 1$, gibt es infolge der Kompaktheit von A eine Teilfolge $(\chi_{k(j)})_j$ mit

$$A\chi_{k(j)} \to \chi \in X, \qquad j \to \infty.$$

Ferner ist

$$\|L\chi_k\| = \frac{\|L\widetilde{\varphi}_{n(k)}\|}{\|\widetilde{\varphi}_{n(k)}\|} \leq \frac{\|L\varphi_{n(k)}\|}{k} \to 0, \qquad k \to \infty,$$

da die Folge $(L\varphi_n)_n$ konvergent und daher beschränkt ist (beachte $L\psi_n = 0$). Folglich gilt

$$L\chi_{k(j)} \to 0, \qquad j \to \infty,$$

und damit erhalten wir

$$\chi_{k(j)} = L\chi_{k(j)} + A\chi_{k(j)} \to \chi, \qquad j \to \infty.$$

Da L stetig ist, können wir aus den beiden vorangegangenen Gleichungen schließen, dass $L\chi = 0$. Weil $\psi_{n(k)} + \|\widetilde{\varphi}_{n(k)}\|\chi \in \ker(L)$ für alle k, erhalten wir nun

$$\begin{aligned}\|\chi_k - \chi\| =& \frac{1}{\|\widetilde{\varphi}_{n(k)}\|} \left\|\varphi_{n(k)} - \left(\psi_{n(k)} + \|\widetilde{\varphi}_{n(k)}\|\chi\right)\right\| \\ \geq& \frac{1}{\|\widetilde{\varphi}_{n(k)}\|} \inf_{\psi \in \ker(L)} \|\varphi_{n(k)} - \psi\| = \frac{1}{\|\widetilde{\varphi}_{n(k)}\|}\|\varphi_{(n(k)} - \psi_{n(k)}\| = 1,\end{aligned}$$

und dies steht im Widerspruch zu $\chi_{k(j)} \to \chi$.

Somit ist die Folge $(\widetilde{\varphi}_n)_n$ beschränkt, und wegen der Kompaktheit von A gibt es eine Teilfolge $(\widetilde{\varphi}_{n(k)})_k$ so, dass $(A\widetilde{\varphi}_{n(k)})_k$ für $k \to \infty$ konvergiert. Aus

$$\widetilde{\varphi}_{n(k)} = L\widetilde{\varphi}_{n(k)} + A\widetilde{\varphi}_{n(k)}$$

folgt dann $\widetilde{\varphi}_{n(k)} \to \varphi \in X$ für $k \to \infty$ und somit $L\widetilde{\varphi}_{n(k)} \to L\varphi$. Also ist $f = L\varphi \in L(X)$, d.h. $\overline{L(X)} = L(X)$. □

Für den zentralen Satz dieses Abschnittes benötigen wir noch ein Lemma

Lemma B.3. *Sei X ein normierter Raum, $U \subset X$ ein abgeschlossener Unterraum mit $U \neq X$ und $\alpha \in (0,1)$. Dann existiert ein Element $\psi \in X$ mit $\|\psi\| = 1$ und $\|\psi - \varphi\| \geq \alpha$ für alle $\varphi \in U$.*

Beweis. Da $U \neq X$, gibt es ein $f \in X$ mit $f \notin U$. Da U abgeschlossen ist, gilt

$$\beta := \inf_{\varphi \in U} \|f - \varphi\| > 0, \qquad \varphi \in U.$$

Wir können ein $g \in U$ so finden, daß

$$\beta \leq \|f - g\| \leq \frac{\beta}{\alpha}$$

und definieren $\psi \in X$ durch

$$\psi := \frac{f - g}{\|f - g\|}.$$

$\|\psi\| = 1$ und für alle $\varphi \in U$ gilt

$$\|\psi - \varphi\| = \frac{1}{\|f - g\|} \|f - \underbrace{(g + \|f - g\|\varphi)}_{\in U}\| \geq \frac{\beta}{\|f - g\|} \geq \alpha.$$

□

Für $n \geq 1$ können die iterierten Operatoren L^n in der Form

$$L^n = (\mathrm{id} - A)^n = \mathrm{id} - A_n \tag{B.4}$$

dargestellt werden, wobei die Operatoren

$$A_n := \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} A^k \tag{B.5}$$

kompakt sind. Folglich sind nach den ersten beiden Riesz-Sätzen die Nullräume $\ker(L^n)$ endlich-dimensional und die Bildräume $L^n(X)$ abgeschlossen.

Satz B.4 (3. Riesz)**.** *Es gibt eine eindeutig bestimmte nicht-negative ganze Zahl r, genannt Rieszsche Zahl des Operators A, mit der die Eigenschaft*

$$\{0\} = \ker(L^0) \subsetneqq \ker(L^1) \subsetneqq \cdots \subsetneqq \ker(L^r) = \ker(L^{r+1}) = \cdots, \tag{B.6}$$

und

$$X = L^0(X) \supsetneqq L^1(X) \supsetneqq \cdots \supsetneqq L^r(X) = L^{r+1}(X) = \cdots. \tag{B.7}$$

Ferner gilt die direkte Summe

$$X = \ker(L^r) \oplus L^r(X), \tag{B.8}$$

d.h. zu jedem $\varphi \in X$ gibt es eindeutig bestimmte Elemente $\psi \in \ker(L^r)$ und $\chi \in L^r(X)$ mit $\varphi = \psi + \chi$.

Beweis. Der Beweis besteht aus vier Teilen.

1. Da für jedes φ mit $L^n\varphi = 0$ auch $L^{n+1}\varphi = 0$ gilt, haben wir trivialer Weise

$$\{0\} = \ker(L^0) \subset \ker(L^1) \subset \ker(L^2) \subset \cdots .$$

Wir nehmen nun an es gelte

$$\{0\} = \ker(L^0) \subsetneqq \ker(L^1) \subsetneqq \ker(L^2) \subsetneqq \cdots .$$

Da die Nullräume $\ker(L^n)$ nach Satz B.1 endlich-dimensional sind, folgt aus dem Lemma B.3 die Existenz einer Folge $(\varphi_n)_n$ mit $\varphi_n \in \ker(L^{n+1})$, $\|\varphi_n\| = 1$ und

$$\|\varphi_n - \varphi\| \geq \frac{1}{2}, \qquad \varphi \in \ker(L^n).$$

Für $n > m$ betrachten wir

$$A\varphi_n - A\varphi_m = \varphi_n - (\varphi_m + L\varphi_n - L\varphi_m)\,.$$

Hier gilt $\varphi_m + L\varphi_n - L\varphi_m \in \ker(L^n)$, da

$$L^n\,(\varphi_m + L\varphi_n - L\varphi_m) = L^{n-m-1}L^{m+1}\varphi_m + L^{n+1}\varphi_n - L^{n-m}L^{m+1}\varphi_m = 0.$$

Daraus folgt

$$\|A\varphi_n - A\varphi_m\| \geq \frac{1}{2}, \qquad n > m,$$

und somit besitzt die Folge $(A\varphi_n)_n$ keine konvergente Teilfolge. Da $(\varphi_n)_n$ beschränkt ist, steht dies im Widerspruch zur Kompaktheit von A und wir wissen, dass es in der Folge $\ker(L^n)$ zwei aufeinanderfolgende Nullräume gibt, die übereinstimmen. Wir setzen

$$r := \min\{k \mid \ker(L^k) = \ker(L^{k+1})\}$$

und beweisen durch Induktion, dass

$$\ker(L^r) = \ker(L^{r+1}) = \ker(L^{r+2}) = \cdots .$$

Dazu nehmen wir an, es sei $\ker(L^k) = \ker(L^{k+1})$ bewiesen für ein $k \geq r$. Dann gilt für jedes $\varphi \in \ker(L^{k+2})$, dass $L^{k+1}L\varphi = L^{k+2}\varphi = 0$. Also folgt $L\varphi \in \ker(L^{k+1}) = \ker(L^k)$. Daher haben wir $L^{k+1}\varphi = L^kL\varphi = 0$ und folglich $\varphi \in \ker(L^{k+1})$. Dies impliziert $\ker(L^{k+2}) \subset \ker(L^{k+1})$.

Somit haben wir bewiesen, dass es eine nichtnegative ganze Zahl r gibt, für die (B.6) gilt.

2. Da wir jedes $\psi = L^{n+1}\varphi \in L^{n+1}(X)$ darstellen können in der Form $\psi = L^nL\varphi$, gilt in trivialer Weise

$$X = L^0(X) \supset L^1(X) \supset L^2(X) \supset \cdots .$$

Wir nehmen nun an es gilt

$$X = L^0(X) \supsetneqq L^1(X) \supsetneqq L^2(X) \supsetneqq \cdots .$$

Da die Bildräume $L^n(X)$ nach Satz B.2 abgeschlossene Unterräume sind, folgt aus dem Lemma B.3 die Existenz einer Folge $(\psi_n)_n$ mit $\psi_n \in L^n(X)$, $\|\psi_n\| = 1$ und

$$\|\psi_n - \psi\| \geq \frac{1}{2}, \qquad \psi \in L^{n+1}(X).$$

Wir schreiben $\psi_n = L^n\varphi_n$ und betrachten

$$A\psi_n - A\psi_m = \psi_n - (\psi_m + L\psi_n - L\psi_m), \qquad m > n.$$

Hier gilt $\psi_m + L\psi_n - L\psi_m \in L^{n+1}(X)$, da

$$\psi_m + L\psi_n - L\psi_m = L^{n+1}\left(L^{m-n-1}\varphi_m + \varphi_n - L^{m-n}\varphi_m\right).$$

Daraus folgt

$$\|A\psi_n - A\psi_m\| \geq \frac{1}{2}, \qquad m > n$$

und wir erhalten wir oben einen Widerspruch. Somit muss es in der Folge $L^n(X)$ zwei aufeinanderfolgende Bildräume geben, die übereinstimmen, und wir setzen

$$q := \min\{k \mid L^k(X) = L^{k+1}(X)\}.$$

Induktiv beweisen wir

$$L^q(X) = L^{q+1}(X) = L^{q+2}(X) = \cdots .$$

Dazu sei $L^k(X) = L^{k+1}(X)$ bewiesen für ein $k \geq q$. Dann gilt für jedes $\psi = L^{k+1}\varphi \in L^{k+1}(X)$ der Ausdruck $L^k\varphi = L^{k+1}\widetilde{\varphi}$ mit einem $\widetilde{\varphi} \in X$, da $L^k(X) = L^{k+1}(X)$. Folglich gilt $\psi = L^{k+2}\widetilde{\varphi} \in L^{k+2}(X)$ und somit $L^{k+1}(X) \subset L^{k+2}(X)$.

Damit haben wir bewiesen, dass es eine nichtnegative ganze Zahl q gibt, für die (B.7) gilt.

3. Wir zeigen nun $r = q$. Hierfür nehmen wir an, es sei $r > q$ und $\varphi \in \ker(L^r)$. Da $L^{r-1}\varphi \in L^{r-1}(X) = L^r(X)$, können wir dann darstellen $L^{r-1}\varphi = L^r\widetilde{\varphi}$ mit einem $\widetilde{\varphi} \in X$. Weil $L^{r+1}\widetilde{\varphi} = L^r\varphi = 0$, gilt $\widetilde{\varphi} \in \ker(L^{r+1}) = \ker(L^r)$, d.h. $L^{r-1}\varphi = L^r\widetilde{\varphi} = 0$. Somit haben wir $\varphi \in N(L^{r-1})$ und folglich $N(L^{r-1}) = N(L^r)$ im Widerspruch zur Definition von r.

Nun nehmen wir an, es sei $r < q$ und $\psi = L^{q-1}\varphi \in L^{q-1}(X)$. Da $L\psi = L^q\varphi \in L^q(X) = L^{q+1}(X)$, können wir darstellen $L\psi = L^{q+1}\widetilde{\varphi}$ mit einem $\widetilde{\varphi} \in X$. Daher gilt $L^q(\varphi - L\widetilde{\varphi}) = L\psi - L^{q+1}\widetilde{\varphi} = 0$, und da $N(L^{q-1}) = N(L^q)$, können wir schließen, dass $L^{q-1}(\varphi - L\widetilde{\varphi}) = 0$ und somit $\psi = L^q\widetilde{\varphi} \in L^q(X)$. Also ist $L^{q-1}(X) = L^q(X)$ im Widerspruch zur Definition von q.

4. Sei $\psi \in \ker(L^r) \cap L^r(X)$. Dann ist $\psi = L^r\varphi$ mit einem $\varphi \in X$ und $L^r\psi = 0$. Folglich gilt $L^{2r}\varphi = 0$, und dies bedeutet $\varphi \in \ker(L^{2r}) = \ker(L^r)$. Daher haben wir

$\psi = L^r\varphi = 0$. Sei nun $\varphi \in X$ beliebig. Dann ist $L^r\varphi \in L^r(X) = L^{2r}(X)$, also gilt $L^r\varphi = L^{2r}\widetilde{\varphi}$ mit einem $\widetilde{\varphi} \in X$. Nun setzen wir $\psi := L^r\widetilde{\varphi} \in L^r(X)$ und $\chi := \varphi - \psi$. Dann ist $L^r\chi = L^r\varphi - L^{2r}\widetilde{\varphi}$ und daher $\chi \in \ker(L^r)$. Somit beweist die Darstellung $\varphi = \chi + \psi$ schließlich die direkte Summe $X = \ker(L^r) \oplus L^r(X)$. □

Theorem B.5. *Sei $A : X \to X$ ein kompakter linearer Operator in einem normierten Raum X. Dann ist $I-A$ genau dann injektiv, wenn es surjektiv ist. Falls $I-A$ injektiv ist (und daher bijektiv), so ist der inverse Operator $(I-A)^{-1} : X \to X$ beschränkt.*

Beweis. Nach (B.6) ist die Injektivität von $L := I - A$ äquivalent zu $r = 0$ und nach (B.7) ist die Surjektivität ebenfalls äquivalent zu $r = 0$. Somit sind Injektivität und Surjektivität von L äquivalent.

Es bleibt zu zeigen, dass L^{-1} beschränkt ist. Hierzu nehmen wir an, L^{-1} sei nicht beschränkt. Dann gibt es eine Folge $(f_n)_n$ mit $\|f_n\| = 1$ und $\|\varphi_n\| \geq n$ für $\varphi_n := L^{-1}f_n$. Wir setzen

$$g_n := \frac{f_n}{\|\varphi_n\|}, \qquad \psi_n := \frac{\varphi_n}{\|\varphi_n\|}, \qquad n \in \mathbb{N}.$$

Dann gilt $g_n \to 0$ für $n \to \infty$ und $\|\psi_n\| = 1$. Da A kompakt ist, gibt es eine Teilfolge $(\psi_{n(k)})_k$ mit $A\psi_{n(k)} \to \psi \in X$ für $k \to \infty$. Aus

$$\psi_n - A\psi_n = g_n$$

entnehmen wir $\psi_{n(k)} \to \psi$ und $\psi \in \ker(L)$. Folglich ist $\psi = 0$ im Widerspruch zu $\|\psi\| = 1$ für alle $n \in \mathbb{N}$. □

Wir können den letzten Satz auf die Lösbarkeit von Operatorgleichungen 2.Art umformulieren.

Korollar B.6. *Sei $A : X \to X$ ein kompakter linearer Operator in einem normierten Raum X. Falls die homogene Gleichung*

$$\varphi - A\varphi = 0 \tag{B.9}$$

nur die triviale Lösung $\varphi = 0$ besitzt, so besitzt für jedes $f \in X$ die inhomogene Gleichung

$$\varphi - A\varphi = f \tag{B.10}$$

genau eine Lösung $\varphi \in X$ und diese Lösung hängt stetig von f ab.

Falls die homogene Gleichung eine nichttriviale Lösung besitzt, dann gibt es nur endlich viele linear unabhängige Lösungen $\varphi_1, ..., \varphi_r$ und die inhomogene Gleichung ist entweder nicht lösbar oder ihre allgemeine Lösung ist von der Form

$$\varphi = \widetilde{\varphi} + \sum_{k=1}^{r} \alpha_k \varphi_k,$$

wobei α_1,...,α_r beliebige komplexe Zahlen sind und $\widetilde{\varphi}$ eine beliebige Lösung der inhomogenen Gleichung ist.

Die grundlegende Bedeutung der Riesz Theorie kompakter Operatoren verdankt sie dem Umstand, dass sie die Existenz einer Lösung des inhomogenen Problems zurückführt auf den Nachweiß, dass die homogene Gleichung nur die triviale Lösung besitzt.

Korollar B.7. *Satz B.5 und Korollar B.6 bleiben gültig, wenn* $\mathrm{id} - A$ *ersetzt wird durch* $S - A$*, wobei* $S : X \to Y$ *ein beschränkter linearer Operator mit einer beschränkten Inversen* $S^{-1} : Y \to X$ *ist und* $A : X \to Y$ *ein kompakter linearer Operator eines normierten Raums* X *in einen normierten Raum* Y *ist.*

Beweis. Dies folgt unmittelbar aus der äquivalenten Umformung von

$$S\varphi - A\varphi = f$$

in die Gestalt

$$\varphi - S^{-1}A\varphi = S^{-1}f,$$

wobei $S^{-1}A$ kompakt ist. □

Satz B.8. *Der durch die direkte Summe* $X = \ker(L^r) \oplus L^r(X)$ *definierte Projektionsoperator* $P : X \to \ker(L^r)$ *ist kompakt.*

Beweis. Wir nehmen an, P sei nicht beschränkt. Dann gibt es eine Folge $(\varphi_n)_n$ mit $\|\varphi_n\| = 1$ und $\|P\varphi_n\| \geq n$. Für

$$\psi_n := \frac{\varphi_n}{\|P\varphi_n\|}, \qquad n \in \mathbb{N},$$

gilt dann $\psi_n \to 0$ für $n \to \infty$, und $\|P\psi_n\| = 1$ für $n \in \mathbb{N}$. Da $(P\psi_n)_n$ im endlich-dimensionalen Raum $\ker(L^r)$ beschränkt ist, gibt es eine Teilfolge $(\psi_{n(k)})_k$ mit der Eigenschaft

$$P\psi_{n(k)} \overset{k\to\infty}{\longrightarrow} \psi \in \ker(L^r).$$

Da $L^r(X)$ abgeschlossen ist, folgt andererseits aus $P\psi_{n(k)} - \psi_{n(k)} \in L^r(X)$ Konvergenz

$$\psi = \lim_{k\to\infty} \left(P\psi_{n(k)} - \psi_{n(k)}\right) \in L^r(X).$$

Also ist $\psi = 0$ im Widerspruch zu $\|P\psi_{n(k)}\| = 1$ für $k \in \mathbb{N}$. Daher ist P beschänkt, und folglich, da der Bildraum von P endlich-dimensional ist, kompakt. □

Satz B.9. *Sei* $A : X \to X$ *ein kompakter linearer Operator in einem unendlich-dimensionalen normierten Raum* X*. Dann gehört* $\lambda = 0$ *zum Spektrum* $\sigma(A)$*, und* $\sigma(A) \setminus \{0\}$ *besteht aus höchstens abzählbar vielen Eigenwerten, die lediglich* $\lambda = 0$ *als Häufungspunkt haben können. Weiters sind die Eigenräume* $\ker(A - \lambda\,\mathrm{id})$ *und die verallgemeinerten Eigenräume* $\ker(A - \lambda\,\mathrm{id})^r$ *mit der Riesz-Zahl* r *endlich-dimensional.*

Beweis. Wir nehmen an, $\lambda = 0$ sei ein regulärer Wert von A, d.h. A^{-1} existiert und ist beschränkt. Dann ist $\mathrm{id} = A^{-1}A$ kompaktund damit einen Widerspruch dazu, dass X endlich-dimensional ist. Also gehört $\lambda = 0$ zum Spektrum $\sigma(A)$.

Für $\lambda \neq 0$ können wir die Riesz Theorie auf den Operator $\lambda\,\mathrm{id} - A$ anwenden. Entweder ist $\ker(\lambda\,\mathrm{id} - A) = \{0\}$ und $(\lambda\,\mathrm{id} - A)^{-1}$ existiert und ist beschränkt nach Satz B.5. Oder $\ker(\lambda\,\mathrm{id} - A) \neq \{0\}$, d.h. λ ist ein Eigenwert. Also ist jedes $\lambda \neq 0$ entweder regulär oder ein Eigenwert von A. Es bleibt zu zeigen, das es zu jeden $R > 0$ höchstens endlich viele Eigenwerte λ mit $|\lambda| \geq R$ gibt. Hierzu nehmen wir an, es gibt eine Folge $(\lambda_n)_n$ paarweise verschiedener Eigenwerte mit $|\lambda_n| \geq R$. Dann wählen wir Eigenelemente φ_n gemäß $A\varphi_n = \lambda_n\varphi_n$ und erklären endlich-dimensionale Unterräume $U_n := \mathbf{span}\{\varphi_1, ..., \varphi_n\}$. Die Eigenelemente φ_k, $k = 1, ..., n$ sind linear unabhängig, denn wäre z.B.

$$\varphi_n = \sum_{k=1}^{n-1} \alpha_k \varphi_k$$

mit schon linear unabhändigen Vektoren $\varphi_1, ..., \varphi_{n-1}$, so folgt

$$0 = (\lambda_n I - A)\varphi_n = \sum_{k=1}^{n-1} \alpha_k \left(\lambda_n\varphi_k - A\varphi_k\right) = \sum_{k=1}^{n-1} \alpha_k \underbrace{(\lambda_n - \lambda_k)}_{\neq 0}\varphi_k,$$

also $\alpha_k = 0$, $k = 1, ..., n-1$. Daher gilt $U_{n-1} \subsetneqq U_n$ und nach Lemma B.3 können wir eine Folge $(\psi_n)_n$ von Elementen $\psi_n \in U_n$ auswählen mit $\|\psi_n\| = 1$ und

$$\|\psi_n - \psi\| \geq \frac{1}{2}, \qquad xi \in U_{n-1}.$$

Aus der Darstellung

$$\psi_n = \sum_{k=1}^{n} \alpha_{nk}\varphi_k$$

entnehmen wir, dass

$$\lambda_n\psi_n - A\psi_n = \sum_{k=1}^{n-1} (\lambda_n - \lambda_k)\alpha_{nk}\varphi_k \in U_{n-1}.$$

Folglich gilt

$$A\psi_n - A\psi_m = \lambda_n\psi_n - (\lambda_n\psi_n - A\psi_n + A\psi_m) = \lambda_n(\psi_n - \psi), \qquad m < n,$$

wobei $\psi = (\lambda_n\psi_n - A\psi_n + A\psi_m)/\lambda_n \in U_{n-1}$. Daher haben wir

$$\|A\psi_n - A\psi_m\| \geq \frac{|\lambda_n|}{2} \geq \frac{R}{2}, \qquad m < n,$$

und die Folge $(A\psi_n)_n$ besitzt keine konvergente Teilfolge im Widerspruch zur Kompaktheit von A. □

Literaturverzeichnis

[Alt99] Alt, Hans W.: *Lineare Funktionalanalysis: eine anwendungsorientierte Einführung*. Third. Berlin : Springer-Verlag, 1999

[Bey12] Beyn, Wolf-Jürgen: An integral method for solving nonlinear eigenvalue problems. In: *Linear Algebra and its Applications* 436 (2012), Nr. 10, 3839 - 3863. http://dx.doi.org/https://doi.org/10.1016/j.laa.2011.03.030. – DOI https://doi.org/10.1016/j.laa.2011.03.030. – ISSN 0024–3795. – Special Issue dedicated to Heinrich Voss's 65th birthday

[BO91] Babuška, I. ; Osborn, J.: Eigenvalue problems. In: *Handbook of numerical analysis, Vol. II*. North-Holland, Amsterdam, 1991 (Handb. Numer. Anal., II), S. 641–787

[BS08] Brenner, Susanne C. ; Scott, L. R.: *Texts in Applied Mathematics*. Bd. 15: *The mathematical theory of finite element methods*. Third. Springer, New York, 2008. – xviii+397 S. http://dx.doi.org/10.1007/978-0-387-75934-0. http://dx.doi.org/10.1007/978-0-387-75934-0. – ISBN 978–0–387–75933–3

[DL90] Dautray, Robert ; Lions, Jacques-Louis: *Mathematical analysis and numerical methods for science and technology. Vol. 3*. Berlin : Springer-Verlag, 1990. – x+515 S. – Spectral theory and applications, With the collaboration of Michel Artola and Michel Cessenat, Translated from the French by John C. Amson

[HS96] Hislop, P. D. ; Sigal, I. M.: *Applied Mathematical Sciences*. Bd. 113: *Introduction to spectral theory*. New York : Springer-Verlag, 1996. – x+337 S. – With applications to Schrödinger operators

[Kal15] Kaltenbäck, Michael: *Fundament Analysis*. Heldermann, N, 2015 (Berliner Studienreihe zur Mathematik). – x+488 S. – ISBN 978–3–88538–126–6

[Kar96a] Karma, Otto: Approximation in eigenvalue problems for holomorphic Fredholm operator functions. I. In: *Numer. Funct. Anal. Optim.* 17 (1996), Nr. 3-4, 365–387. http://dx.doi.org/10.1080/01630569608816699. – DOI 10.1080/01630569608816699. – ISSN 0163–0563

[Kar96b] Karma, Otto: Approximation in eigenvalue problems for holomorphic Fredholm operator functions. II. (Convergence rate). In: *Numer. Funct. Anal. Optim.* 17 (1996), Nr. 3-4, 389–408. http://dx.doi.org/10.1080/

01630569608816700. – DOI 10.1080/01630569608816700. – ISSN 0163–0563

[Kat58] Kato, Tosio: Perturbation theory for nullity, deficiency and other quantities of linear operators. In: *J. Analyse Math.* 6 (1958), 261–322. http://dx.doi.org/10.1007/BF02790238. – DOI 10.1007/BF02790238. – ISSN 0021–7670

[Kat95] Kato, Tosio: *Perturbation theory for linear operators*. Springer-Verlag, Berlin, 1995 (Classics in Mathematics). – xxii+619 S. – ISBN 3–540–58661–X. – Reprint of the 1980 edition

[Kre99] Kress, Rainer: *Applied Mathematical Sciences*. Bd. 82: *Linear integral equations*. Second. New York : Springer-Verlag, 1999. – xiv+365 S. – ISBN 0–387–98700–2

[Kre02] Kress, R.: Chapter 1.2.1 - Specific Theoretical Tools. Version: 2002. http://dx.doi.org/http://dx.doi.org/10.1016/B978-012613760-6/50004-8. In: Sabatier, Roy P. (Hrsg.): *Scattering*. London : Academic Press, 2002. – DOI http://dx.doi.org/10.1016/B978–012613760–6/50004–8. – ISBN 978–0–12–613760–6, 37 - 51

[Saa11] Saad, Yousef: *Classics in Applied Mathematics*. Bd. 66: *Numerical methods for large eigenvalue problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. – xvi+276 S. http://dx.doi.org/10.1137/1.9781611970739.ch1. http://dx.doi.org/10.1137/1.9781611970739.ch1. – ISBN 978–1–611970–72–2. – Revised edition of the 1992 original [1177405]

[Yos74] Yosida, Kôsaku: *Functional analysis*. Fourth. Springer-Verlag, New York-Heidelberg, 1974. – xii+496 S. – Die Grundlehren der mathematischen Wissenschaften, Band 123