Diplomarbeit

Numerische Lösung von Eigenwertproblemen kompakter, symmetrischer Integraloperatoren

Angefertigt am
Institut für numerische Simulation

Vorgelegt der
Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

August 2010

von
Michael Peters

aus
Rheinbach

Inhaltsverzeichnis

Kapitel 1

Einleitung

„Begin at the beginning,“, the King said gravely,
„and go on till you come to the end: then stop.“
- Lewis Carroll

Die vorliegende Arbeit behandelt die numerische Berechnung von *großen* Eigenwertproblemen kompakter und symmetrischer Integraloperatoren. Hierbei nennen wir ein Eigenwertproblem *groß*, wenn die aus der Diskretisierung des Problems resultierende Matrix nicht im Arbeitsspeicher eines Computers gespeichert werden kann, ohne eine eventuelle, spezielle Struktur der Matrix auszunutzen. Ein nicht großes Eigenwertproblem nennen wir *klein* (vgl. [30]). Die früheste numerische Behandlung von Eigenwertproblemen, ist das Jacobi-Verfahren, das auf den deutschen Mathematiker Carl Gustav Jacobi zurückgeht. Er berechnete bereits 1846 Eigenwerte symmetrischer Matrizen, indem er sie mit Hilfe ebener Rotationen in stark diagonaldominante Matrizen transformierte (vgl. [10]). Aber wozu werden Eigenwertprobleme eigentlich gelöst? Eigenwertprobleme treten in vielen Gebieten, beispielsweise in der Physik, in der Elektrotechnik, im Maschinenbau, in der Statik, in der Biologie, in der Informatik und in den Wirtschaftswissenschaften auf (vgl. [21]). Das wohl bekannteste Beispiel stammt aus der Physik. Schwingungsfähige Systeme besitzen ausgezeichnete Frequenzen, so genannte Resonanzfrequenzen, welche mit Hilfe von Eigenvektoren beschrieben werden können (vgl. [13, 21]). Im Jahr 1940 wurde die Tacoma-Brücke durch Windturbulenzen in derart starke Schwingungen versetzt, dass sie einstürzte. Daher muss bei der Konstruktion von Bauwerken eine entsprechende Spektralanalyse der zugehörigen Steifigkeitsmatrix durchgeführt und berücksichtigt werden, dass Erregerschwingungen und Resonanzbereiche der Steifigkeitsmatrix weit ge-

nug auseinander liegen (vgl. [13]). Eine weitere und die von uns betrachtete Anwendung für die Lösung von Eigenwertproblemen ist die Singulärwertzerlegung von Integraloperatoren. So verwendet man zur Lösung stochastischer Differentialgleichungen die so genannte Karhunen-Loève-Entwicklung, die gerade einer Singulärwertzerlegung in $L^2(\Omega_1 \times \Omega_2)$ entspricht, wobei Ω_1 und Ω_2 Gebiete sind (vgl. [15]). In der Praxis ist eine vollständige Singulärwertzerlegung großer Matrizen nicht realisierbar. Aus diesem Grunde verwendet man die abgebrochene Singulärwertzerlegung. Um dabei einen möglichst geringen Fehler zu produzieren, ist es notwendig, die größten Eigenwerte des Operators zu berechnen (vgl. [15]). Hierzu werden in dieser Arbeit insgesamt vier numerische Verfahren zur numerischen Lösung dieses Problems betrachtet. Das ältestete darunter ist die, auf die von Mises-Potenzmethode zurückgehende, Unterraumiteration. Betrachtet man die Folge von Vektoren, die von der Potenzmethode erzeugt werden, so gelangt man zu Krylov-Räumen wachsender Dimension und somit schließlich zum bekannten Lanczos-Verfahren (vgl. [10]). So wie die beiden Verfahren in dieser Arbeit Verwendung finden, dienen sie zur Projektion des Eigenwertproblems auf einen Unterraum. Das resultierende, kleine Eigenwertproblem wird dann mit Hilfe des QR-Verfahrens gelöst. Die beiden so gewonnenen Verfahren besitzen eine Komplexität von $\mathcal{O}(n^2 m)$, wenn n die Dimension und m die Anzahl der gesuchten Eigenwerte bezeichnet. Abschließend wird noch ein neues Verfahren zur Berechnung der größten Eigenwerte vorgestellt. Im Fall, dass die Eigenwerte hinreichend schnell abklingen, wie dies beispielsweise bei glatten Integralkernen der Fall ist (vgl. [27]), liefert das Cholesky-Verfahren eine sehr schnelle Konvergenz bei einem Aufwand von lediglich $\mathcal{O}(nm^2)$ für die Lösung des Eigenwertproblems (vgl. [16]).

Ich möchte diese Stelle nutzen, um mich bei Herrn Prof. Dr. Harbrecht für die Überlassung des Themas, die vielen hilfreichen Gespräche und die immer konstruktive Kritik zu bedanken. Ich danke Herrn Prof. Dr. Chernov für die Übernahme des Zweitgutachtens. Mein besonderer Dank gebührt Markus Siebenmorgen für die gegenseitige Unterstützung während unseres Studiums der Mathematik.

Diese Arbeit ist weiter wie folgt gegliedert.

In Kaptitel 2 werden die funktionalanalytischen Grundlagen geschaffen. Lineare-, kompakte- und Integraloperatoren werden eingeführt. Dabei wird besonders auf Projektionen eingegangen, da diese für die Analyse des Diskretisierungsfehlers essenziell sind. Abschließend wird die Singulärwertzerlegung für L^2-Operatoren und ihre Bestapproximationseigenschaft bewiesen.

In Kapitel 3 wird die Diskretisierung von Eigenwertproblemen kompakter Operatoren und die Abschätzung des dabei entstehenden Approximationsfehler behandelt.

In Kapitel 4 werden dann rein algebraische Eigenwertprobleme betrachtet. Hierzu wird zunächst die Kondition des Eigenwertproblems diskutiert und Konvergenzabschätzungen geliefert. Danach werden die oben angeführten Verfahren und ihr Zusammenhang erläutert.
In Kapitel 5 werden Beispiele betrachtet und die numerischen Resultate präsentiert. Hierbei wird unterschieden zwischen dem Cholesky-Verfahren und der Unterraumiteration, sowie dem Lanczos-Verfahren.

Bezeichnungen

*	adjungiert (bei komplexen Vektoren und Matrizen)
$\langle \cdot, \cdot \rangle$	Dualitätspaarung
$\|\cdot\|$	Norm
$\mathbb{C}$	Körper der komplexen Zahlen
$B_R(x)$	$\{y \in X \mid \|y - x\|_X < R\}$
$C, C(\cdot)$	Konstanten
C_0	Menge der stetigen Funktionen mit kompaktem Träger
$d(\cdot, \cdot)$	Metrik
$\delta_{i,j}$	Kronecker-Delta
$e_1, \ldots, e_n$	kanonische Basis, insbesondere des $\mathbb{R}^n$
H^s	Sobolev-Raum der Ordnung s (vgl. [29])
I	Identität (Operator)
$\mathbf{1}$	Indikatorfunktion
$\mathbb{K}$	$\mathbb{R}$ oder $\mathbb{C}$
L^1_{loc}	Menge der lokal integrierbaren Funktionen
L^2	Menge der quadratintegrierbaren Funktionen
L^p	Menge der Funktionen, deren p-te Potenz integrierbar ist
meas	Lebesgue-Maß
$\mathcal{N}(T)$	Kern eines Operators T
Ω	Gebiet im $\mathbb{R}^n$
$\oplus$	direkte Summe
$\otimes$	Tensorprodukt
P	Projektion
p'	dualer Exponent
$\mathbb{R}$	Körper der reellen Zahlen
Λ	Rayleigh-Quotient
rank	Rang eines Operators
$\mathcal{R}(T)$	Bild eines Operators T
sign	Vorzeichenfunktion
supp	Träger einer Funktion
T	linearer Operator

$^\intercal$	transponiert (bei reellen Vektoren und Matrizen)
V	Einfachschichtpotential
X, Y	meistens Hilbert-Räume
X'	Dualraum zu X
(X, d)	metrischer Raum
$(x_n)_{n\in\mathbb{N}}$	Folge
Y_n^m	Kugelflächenfunktion

Wichtige Formeln

Hölder'sche Ungleichung

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \qquad \text{mit } \frac{1}{p} + \frac{1}{q} = 1.$$

Für die Wahl $p = q = 2$ ergibt sich die Cauchy-Schwarz'sche Ungleichung (CSU).

Parallelogrammidentität

$$\|x + y\|_X^2 + \|x - y\|_X^2 = 2(\|x\|_X^2 + \|y\|_X^2)$$

Chebyshev-Polynome

$$c_0(x) := 1, c_1(x) := x, \qquad c_{n+1}(x) := 2xc_n(x) - c_{n-1}(x).$$

Legendre-Polynome

$$P_0(x) := 1, P_1(x) := (x), \qquad P_{n+1}(x) := \frac{1}{n+1}((2n+1)xP_n(x) - nP_{n-1}(x)).$$

Kapitel 2

Grundlagen

In dieser Diplomarbeit werden Hilbert-Schmidt-Operatoren betrachtet. Diese sind insbesondere kompakt. Daher werden in diesem Kapitel einige relevante Eigenschaften kompakter Operatoren zusammengetragen. Wesentlich ist hierbei der Satz über die Kompaktheit der Hilbert-Schmidt-Operatoren. Der Beweis wird für allgemeine L^p Räume geführt, später wird nur noch der Fall für den Hilbert-Raum L^2 betrachtet. Der Spektralsatz für kompakte Operatoren im Allgemeinen und der Spektralsatz für kompakte, normale Operatoren werden ohne Beweis angegeben. Die Abschnitte 1 bis 4 sind weitestgehend aus den entsprechenden Kapiteln in [1] übernommen und wurden für diese Arbeit adaptiert. Im Folgenden seien, sofern nicht explizit anders vorausgesetzt, $(X, (\cdot,\cdot)_X)$ und $(Y, (\cdot,\cdot)_Y)$ Hilbert-Räume über $\mathbb{R}$ oder $\mathbb{C}$.

2.1 Lineare Operatoren

2.1 Definition. Wir bezeichnen mit

$$L(X;Y) := \{T :\ X \to Y \mid T \text{ ist stetig und linear}\}$$

die Menge der *linearen Operatoren* von X nach Y und für $X = Y$ schreiben wir

$$L(X) := L(X;X).$$

Für jeden Operator $T \in L(X;Y)$ setzen wir

$$\|T\|_{L(X;Y)} := \sup_{\|x\|_X \leq 1} \|Tx\|_Y .$$

Nach dem folgenden Lemma 2.2 existiert das Supremum für jedes $T \in L(X;Y)$. Der Index der Operatornorm wird meistens weggelassen und nur

$\|T\|$ statt $\|T\|_{L(X;Y)}$ geschrieben. Weiter definieren wir für jeden Operator $T \in L(X;Y)$ durch

$$\mathcal{N}(T) := \{x \in X \mid Tx = 0\}$$

den *Nullraum* oder *Kern* von T, sowie durch

$$\mathcal{R}(T) := \{Tx \in Y \mid x \in X\}$$

den *Bildraum* von T.

Der Kern $\mathcal{N}(T)$ ist abgeschlossen in X. Im Allgemeinen ist aber der Bildraum $\mathcal{R}(T)$ nicht abgeschlossen in Y (vgl. [1]). Die Stetigkeit linearer Operatoren wird durch den folgenden Satz charakterisiert.

2.2 Lemma ([1])**.** *Ist $T: X \to Y$ linear und $x_0 \in X$, so sind äquivalent:*

1. $\sup_{\|x\|_X \leq 1} \|Tx\|_Y < \infty$
2. *Es existiert eine Konstante $c_2^T > 0$ mit $\|Tx\|_Y \leq c_2^T \|x\|_X$.*
3. *T ist stetig.*
4. *T ist stetig in x_0.*

Falls X endlich dimensional ist, so ist jede lineare Abbildung $T: X \to Y$ stetig (vgl. [1]). Es ist möglich $L(X;Y)$ als Vektorraum über $\mathbb{R}$ aufzufassen. Dieser besitzt unter Anderem die folgenden Eigenschaften.

2.3 Satz ([1])**.** *Es gilt:*

1. $(L(X;Y), \|\cdot\|_{L(X;Y)})$ *ist ein normierter Vektorraum.*
2. $L(X;Y)$ *ist ein Banach-Raum, falls Y ein Banach-Raum ist.*
3. *Sei Z ein $\mathbb{K}$-Vektorraum, sowie $T \in L(X;Y)$ und $S \in L(Y;Z)$, dann ist $ST \in L(X;Z)$ mit $\|ST\|_{L(X;Z)} \leq \|S\|_{L(Y;Z)} \|T\|_{L(X;Y)}$.*

2.4 Definition. Sei $T \in L(X;Y)$. Der durch

$$(x, T^*y)_X = (Tx, y)_Y \qquad \text{für alle } x \in X, y \in Y$$

definierte Operator $T^* \in L(Y;X)$ heißt *adjungierter Operator* zu T.
Falls $T \in L(X)$ und $T^* = T$ gilt, heißt T *selbstadjungiert*.

Eine wichtige Klasse von Operatoren sind jene, die mit ihrem adjungierten Operator kommutieren.

2.5 Definition. Ein Operator $X \in L(X)$ heißt *normal*, wenn

$$TT^* = T^*T$$

gilt, also falls er mit seinem adjungierten Operator kommutiert.

Bemerkung.

1. Jeder selbstadjungierte Operator ist normal.
2. T normal $\Leftrightarrow \ \|Tx\|_X = \|T^*x\|_X$ für alle $x \in X$

2.6 Definition. Sei X ein Hilbert-Raum über $\mathbb{R}$. Ein Operator $T \in L(X)$ heißt *symmetrisch*, falls T selbstadjungiert ist, also falls

$$(x, Ty)_X = (y, Tx)_X \qquad \text{für alle } x, y \in X.$$

2.7 Definition. Wer definieren die Menge

$$L^+(X) := \{T : \ X \longrightarrow X \mid T \in L(X) \text{ ist symmetrisch und } \textit{positiv definit}\}.$$

Hierbei nennen wir $T \in L(X)$ *gleichmäßig elliptisch* oder *positiv definit*, falls ein $c_1^T > 0$ existiert mit

$$T \geq c_1^T I \quad :\Longleftrightarrow \quad (u, Tu)_X \geq c_1^T (u, u)_X.$$

$T \in L(X)$ mit $T = T^*$ heißt *positiv* oder *elliptisch*, falls

$$(u, Tu)_X = (Tu, u)_X > 0 \qquad \text{für alle } u \in X.$$

2.8 Definition. Für einen symmetrischen Operator $T \in L(X)$ definieren wir die Bilinearform

$$b_T(\cdot, \cdot) := (T\cdot, \cdot)_X = (\cdot, T\cdot)_X.$$

Um später die Analytizität der Resolventenfunktion zeigen zu können, benötigen wir einen Satz über die Invertierbarkeit linearer Operatoren.

2.9 Satz. *Sei X ein Banach-Raum und $T \in L(X)$ mit*

$$\limsup_{i \to \infty} \left\|T^i\right\|^{\frac{1}{i}} < 1.$$

Dann ist $(I - T)^{-1} \in L(X)$ und besitzt die Darstellung

$$(I - T)^{-1} = \sum_{i=0}^{\infty} T^i.$$

Beweis. Wir definieren die Partialsumme $S_k := \sum_{i=0}^{k} T^i$. Sei $m \in \mathbb{N}$ und $r < 1$ mit $\|T^i\| \leq r^i$ für alle $i \geq m$. Also können wir für $m \leq k < l$ abschätzen

$$\|S_l - S_k\| = \left\| \sum_{i=k+1}^{l} T^i \right\| \leq \sum_{i=k+1}^{l} \left\|T^i\right\| \leq \sum_{i=k+1}^{l} r^i \to 0 \qquad \text{für } k \to \infty.$$

Die Partialsummen bilden also eine Cauchy-Folge und nach Satz 2.3 ist $L(X)$ vollständig, daher gilt

$$\lim_{k\to\infty} S_k =: S \in L(X).$$

Ferner erhalten wir

$$(I - T)S_k x = \sum_{i=0}^{k} (T^i - T^{i+1})x = x - T^{k+1}x \to x \qquad \text{für } k \to \infty,$$

da für $k \geq m$ immer $\left\|T^{k+1}x\right\| \leq r^{k+1} \|x\|_X$ gilt. Analog zeigt man $S(I - T) = I$. Damit ist $S = (I - T)^{-1}$ bewiesen. □

2.2 Projektionen

Essenziell für die Fehlerabschätzungen im nächsten Kapitel sind orthogonale Projektoren. Daher werden hier die grundlegenden Begriffe definiert und Eigenschaften zusammengetragen.

2.10 Definition. Ein linearer Operator $P \in L(X)$ heißt *(stetiger, linearer) Projektor*, falls

$$PP = P$$

gilt.

2.11 Satz. *Ist P ein Projektor, so auch $I - P$.*

Beweis. Es gilt

$$(I - P)^2 = I^2 - 2IP + P^2 = I - 2P + P = I - P.$$

□

2.12 Satz ([1])**.** *Sei $P \in L(X)$ ein Projektor. P ist genau dann kompakt, wenn* $\dim(\mathcal{R}(P)) < \infty$ *gilt.*

Der folgende Satz charakterisiert orthogonale Projektionen auf reellen Hilbert-Räumen.

2.13 Satz. *Sei $A \subset X$ nichtleer, abgeschlossen und konvex. Dann existiert genau eine Abbildung $P: X \longrightarrow A$ mit*

$$\|x - P(x)\|_X = \operatorname{dist}(x, A) := \inf_{y \in A} \|x - y\|_X \qquad \text{für alle } x \in X.$$

Diese Abbildung P heißt orthogonale Projektion *von X auf A.*

Beweis. Für jedes $x \in X$ existiert eine Folge $(a_i)_{i \in \mathbb{N}}$ in A mit

$$\|x - a_i\|_X \to \operatorname{dist}(x, A) =: d.$$

Die Folge $(a_i)_{i \in \mathbb{N}}$ heißt auch *Minimalfolge*. Wir zeigen nun, dass $(a_i)_{i \in \mathbb{N}}$ eine Cauchy-Folge in A ist und somit konvergiert. Nach der Parallelogrammidentität gilt

$$\|(x - a_j) - (x - a_i)\|_X^2 + \|(x - a_j) + (x - a_i)\|_X^2 = 2(\|x - a_i\|_X^2 + \|x - a_j\|_X^2).$$

Somit ist

$$\begin{aligned} \|a_i - a_j\|_X^2 &= 2\left(\|x - a_i\|_X^2 + \|x - a_j\|_X^2 - 2\left\|x - \frac{a_i + a_j}{2}\right\|_X^2\right) \\ &\leq 2(\|x - a_i\|_X^2 + \|x - a_j\|_X^2 - 2d^2) \to 0 \qquad \text{für } i, j \to \infty, \end{aligned}$$

da A konvex ist und somit $\frac{1}{2}(a_i + a_j) \in A$ gilt. A ist abgeschlossen und vollständig, daher existiert der Grenzwert

$$x_1 := \lim_{i \to \infty} a_i \in A$$

und da die Norm stetig ist, folgt

$$\|x - x_1\|_X^2 = d.$$

Die Eindeutigkeit sieht man mit dem selben Argument ein, denn ist $x_2 \in A$ mit $\|x - x_1\|_X^2 = d$, folgt sofort

$$\|x_1 - x_2\|_X^2 = 2(\|x - x_1\|_X^2 + \|x - x_2\|_X^2 - 2d^2) = 0.$$

Damit ist $P(x) := x_1$ eindeutig bestimmt. □

2.14 Satz. *Ist A zusätzlich zu den Voraussetzungen von Satz 2.13 ein Unterraum von X, so ist P ein Projektor und es gilt*

$$\mathcal{N}(P) \perp \mathcal{R}(P).$$

Beweis. Wir zeigen $(x - P(x), a)_X = 0$ für alle $a \in A$. Angenommen es gäbe ein $a \in A$ mit $(x - P(x), a)_X \neq 0$. Sei $\alpha := (x - P(x), a)_X$. Wir betrachten $y := P(x) + \frac{\alpha}{\|a\|_X^2} a$. Dann gilt

$$\begin{aligned} \|x - y\|_X^2 &= \left\| x - P(x) - \frac{\alpha}{\|a\|_X^2} a \right\|_X^2 \\ &= \|x - P(x)\|_X^2 - 2\frac{\alpha}{\|a\|_X^2}(x - P(x), a)_X + \frac{\alpha^2}{\|a\|_X^2} \\ &= \|x - P(x)\|_X^2 - \frac{\alpha^2}{\|a\|_X^2} \leq \|x - P(x)\|_X^2 \end{aligned}$$

im Widerspruch zur Optimalität von $P(x)$. Damit ist

$$x - P(x) \perp A \qquad \text{für alle } x \in X.$$

Nun betrachten wir die beiden Gleichungen

$$\begin{aligned} (x + y - P(x + y), a)_X &= 0 \qquad \text{für alle } a \in A \\ (x + y - (P(x) + P(y)), a)_X &= 0 \qquad \text{für alle } a \in A. \end{aligned}$$

Die Subtraktion der beiden Gleichungen liefert

$$(\underbrace{P(x + y) - (P(x) + P(y))}_{\in A}, a)_X = 0 \qquad \text{für alle } a \in A.$$

Daher muss $P(x + y) - (P(x) + P(y)) = 0 \in A$ gelten.
Ferner gilt für jedes $\alpha \in \mathbb{R}$, dass $\alpha x - \alpha P(x) \perp A$. Damit erhalten wir $\alpha P(x) = P(\alpha x)$. Hiermit ist die Linearität von P gezeigt. Offenbar gilt auch $\mathcal{R}(P) = A$. Für jedes $x \in \mathcal{N}(P)$ erhalten wir

$$(x, a)_X = (x - P(x), a)_X = 0 \qquad \text{für alle } a \in A.$$

Das bedeutet $\mathcal{N}(P) = A^\perp := \{x \in X \mid (x, a)_X = 0 \ \forall a \in A\}$, was genau

$$\mathcal{N}(P) \perp \mathcal{R}(P)$$

besagt. □

Bemerkung. Ist $A \subset X$ ein Unterraum, so ist A per Definition invariant unter P, insbesondere gilt

$$P(x) = x \qquad \text{für alle } x \in A.$$

2.3 Kompakte Operatoren

Für die Beweise in diesem Kapitel benötigen wir die äquivalenten Definitionen von Kompaktheit des nächsten Satzes.

2.15 Satz ([1]). *Für jede Teilmenge A eines metrischen Raumes (X, d) sind folgende Aussagen äquivalent:*

1. *jede offene Überdeckung von A enthält eine endliche Teilüberdeckung.* (überdeckungskompakt)
2. *jede Folge in A besitzt eine Teilfolge mit Grenzwert in A.* (folgenkompakt)
3. *(A, d) ist vollständig und für $\varepsilon > 0$ besitzt A eine endliche Überdeckung aus ε-Kugeln.* (präkompakt)

2.16 Definition. Sei (X, d) ein metrischer Raum. Dann heißt $A \subset X$ *kompakt*, falls A eine der äquivalenten Bedingungen aus Satz 2.15 erfüllt.

Um später zu zeigen, dass Hilbert-Schmidt-Operatoren kompakt sind, verwenden wir den folgenden Satz.

2.17 Satz (M. Riesz, [1]). *Für $1 \leq p < \infty$ ist $A \subset L^p(\mathbb{R}^n)$ genau dann präkompakt, wenn die folgenden drei Bedingungen erfüllt sind:*

(i) $\sup\limits_{f \in A} \|f\|_{L^p(\mathbb{R}^n)} < \infty$

(ii) $\sup\limits_{f \in A} \|f(\cdot + h) - f\|_{L^p(\mathbb{R}^n)} \to 0 \qquad$ *für* $|h| \to 0$

(iii) $\sup\limits_{f \in A} \|f\|_{L^p(\mathbb{R}^n \setminus B_R(0))} \to 0 \qquad$ *für* $R \to \infty$

2.18 Definition. Die Menge der *kompakten (linearen) Operatoren* von X nach Y ist definiert als

$$K(X;Y) := \{T \in L(X;Y) \mid \overline{T(B_1(0))} \text{ ist kompakt}\}.$$

Alternativ ist es möglich, kompakte Operatoren durch eine der äquivalenten Eigenschaften des nächsten Satzes zu definieren.

2.19 Satz ([1]). *Für einen kompakten Operator $T \in K(X;Y)$ sind äquivalent:*

1. $\overline{T(B_1(0))}$ *ist kompakt in* Y.

2. $A \subset X$ *beschränkt* $\Longrightarrow T(A)$ *ist präkompakt in* Y.

3. *Für jede beschränkte Folge* $(x_n)_{n\in\mathbb{N}} \subset X$ *besitzt* $(Tx_n)_{n\in\mathbb{N}}$ *eine konvergente Teilfolge.*

Für den nächsten Satz, der einige Eigenschaften kompakter Operatoren liefert, benötigen wir eine Definition.

2.20 Definition. Sei $T \in L(X;Y)$. Dann heißt T *vollstetig*, falls für jede Folge $(x_n)_{n\in\mathbb{N}} \subset X$ mit $\langle x_n, x'\rangle \to \langle x, x'\rangle$ für $n \to \infty$ und für jedes $x' \in X'$ folgt, dass $\|Tx_n - Tx\|_Y \to 0$ für $n \to \infty$.

2.21 Satz ([1])**.**

1. $T \in L(X;Y)$ *ist genau dann kompakt, wenn* T *vollstetig ist.*

2. $K(X;Y)$ *ist ein abgeschlossener Unterraum von* $L(X;Y)$.

3. *Für jeden linearen Operator* T *folgt aus* $\dim(\mathcal{R}(T)) < \infty$, *dass* $T \in K(X;Y)$.

4. $T \in L(X;Y)$ *ist genau dann kompakt, wenn es eine Folge von Operatoren* $(T_i)_{i\in\mathbb{N}} \subset L(X;Y)$ *gibt mit* $\dim(\mathcal{R}(T_i)) < \infty$ *für jedes* $i \in \mathbb{N}$, *so dass* $\|T - T_i\|_Y \to 0$ *für* $i \to \infty$.

2.22 Satz. *Sei* $T_1 \in L(X;Y)$ *und* $T_2 \in L(Y;Z)$, *dann gilt*

$$T_1 \text{ oder } T_2 \text{ kompakt } \Longrightarrow T_2T_1 \text{ ist kompakt.}$$

Beweis. Wir zeigen 3. aus Satz 2.19. Sei $(x_i)_{i\in\mathbb{N}}$ eine beschränkte Folge in X. T_1 ist beschränkt, also ist $(T_1x_i)_{i\in\mathbb{N}}$ eine beschränkte Folge in Y. Ist T_2 kompakt, so besitzt $(T_2T_1x_i)_{i\in\mathbb{N}}$ eine konvergente Teilfolge. Umgekehrt ist T_1 kompakt, besitzt $(T_1x_i)_{i\in\mathbb{N}}$ eine konvergente Teilfolge $(T_1x_{i_k})_{k\in\mathbb{N}}$. Wegen der Stetigkeit von T_2 konvergiert dann auch $(T_2T_1x_{i_k})_{k\in\mathbb{N}}$. □

Wir führen nun die Klasse von Operatoren ein, die in dieser Arbeit vorwiegend betrachtet werden.

2.23 Definition (Hilbert-Schmidt-Operatoren)**.** Seien $\Omega_1 \subset \mathbb{R}^n$ und $\Omega_2 \subset \mathbb{R}^m$ offen, sowie $1 < p, q < \infty$ und sei $K:\ \Omega_1 \times \Omega_2 \to \mathbb{R}$ messbar mit

$$\|K\| := \left(\int_{\Omega_1} \left[\int_{\Omega_2} |K(x,y)|^{p'} \,\mathrm{d}y \right]^{\frac{q}{p'}} \mathrm{d}x \right)^{\frac{1}{q}} < \infty,$$

wobei p' den zu p dualen Exponenten bezeichne, das heißt $\frac{1}{p} + \frac{1}{p'} = 1$. Dann definieren wir

$$Tf(x) := \int_{\Omega_2} K(x,y) f(y) \mathrm{d}y.$$

Der nächste Satz zeigt, dass diese Abbildung ein Operator ist. Operatoren dieser Art werden *Hilbert-Schmidt-Integraloperatoren* genannt. K heißt *Integralkern* zum Operator T.

2.24 Satz. *Unter den Voraussetzungen von Definition 2.23 gilt:*

$$T \in L(L^p(\Omega_2); L^q(\Omega_1)) \quad mit \quad \|T\| \leq \|K\|.$$

Beweis. Wegen $\|K\| < \infty$ folgt $K(x,\cdot) \in L^{p'}(\Omega_2)$ für fast alle x. Damit ist $Tf(x)$ nach der Hölder'schen Ungleichung für solche x definiert. Wir zeigen nun, dass $Tf(x)$ messbar ist: Der Beweis erfolgt via maßtheoretischer Induktion. Zunächst liefert zweimaliges Anwenden der Hölder'schen Ungleichung, dass $K \in L^1_{loc}(\Omega_1 \times \Omega_2)$ gilt. Mit dem Satz von Fubini folgt, dass

$$\int_D K(x,y) \mathrm{d}y$$

für jede beschränkte Teilmenge $D \subset \Omega_2$ messbar als Funktion von x ist. Daher ist Tf messbar für Treppenfunktionen f. Jede messbare Funktion $f \in L^p(\Omega_2)$ kann durch eine isotone Folge von Treppenfunktionen $(f_i)_{i\in\mathbb{N}}$ mit $f_i \nearrow f$ approximiert werden. Nach dem Satz von der monotonen Konvergenz konvergiert $Tf_i \longrightarrow Tf$ für fast alle x, also ist Tf messbar. Ferner gilt

$$\begin{aligned}
&\int_{\Omega_1} |Tf(x)|^q \,\mathrm{d}x = \int_{\Omega_1} \left| \int_{\Omega_2} K(x,y) f(y) \mathrm{d}y \right|^q \mathrm{d}x \\
&\leq \int_{\Omega_1} \left(\int_{\Omega_2} |K(x,y)|^{p'} \,\mathrm{d}y \right)^{\frac{q}{p'}} \left(\int_{\Omega_2} |f(y)|^p \right)^{\frac{q}{p}} \mathrm{d}x = \|K\|^q \cdot \|f\|^q_{L^p(\Omega_2)},
\end{aligned}$$

womit auch die Behauptung über die Norm gezeigt ist. □

2.25 Satz. *Unter den Voraussetzungen von Definition 2.23 gilt*

$$T \in K(L^p(\Omega_2); L^q(\Omega_1)).$$

Zum Beweis des Satzes benötigen wir folgendes Lemma.

2.26 Lemma. *Sei $1 \leq p < \infty$ und $f \in L^p(\mathbb{R}^n)$, dann gilt*

$$\|f(\cdot + h) - f\|_{L^p(\mathbb{R}^n)} \to 0 \qquad \text{für } |h| \to 0.$$

Beweis. Sei $(f_j)_{j\in\mathbb{N}} \subset C_0(\mathbb{R}^n)$ mit $\|f - f_j\|_{L^p(\mathbb{R}^n)} \to 0$ für $j \to \infty$. Zum Beweis der Existenz einer solchen Folge sei auf Lemma 1.22.2 in [1] verwiesen. Nun gilt

$$\begin{aligned} & \|f(\cdot + h) - f\|_{L^p(\mathbb{R}^n)} \\ \leq\ & \|f(\cdot + h) - f_j(\cdot + h)\|_{L^p(\mathbb{R}^n)} + \|f_j(\cdot + h) - f_j\|_{L^p(\mathbb{R}^n)} + \|f_j - f\|_{L^p(\mathbb{R}^n)} \\ \leq\ & 2\,\|f - f_j\|_{L^p(\mathbb{R}^n)} + \mathrm{meas}(\mathrm{supp}(f_j(\cdot + h) - f_j))^{\frac{1}{p}}\,\|f_j(\cdot + h) - f_j\|_{\infty}\,. \end{aligned}$$

Hierbei bezeichnet meas$(\cdot)$ das n-dimensionale Lebesgue-Maß und supp$(\cdot)$ den Träger der Funktion. Der erste Summand konvergiert nach Voraussetzung für $j \to \infty$. Da alle f_j einen kompakten Träger haben, also gleichmäßig stetig sind und $\mathrm{supp}(f_j(\cdot+h)-f_j)$ beschränkt ist, konvergiert auch der zweite Summand für $|h| \to 0$ gegen 0. □

Beweis von Satz 2.25. Nach Satz 2.24 ist T stetig mit $\|T\| \leq \|K\|$. Um die Kompaktheit von T zu zeigen, setzen wir T außerhalb von $\Omega_1 \times \Omega_2$ durch $K(x, y) = 0$ für $(x, y) \notin \Omega_1 \times \Omega_2$ fort. Dann gilt für $h \in \mathbb{R}^n$ und $f \in L^p(\Omega_2)$ mit $\|f\|_{L^p(\Omega_2)} \leq 1$ analog zum Beweis von Satz 2.24

$$\int_{\mathbb{R}^n} |Tf(x + h) - Tf(x)|^q \,\mathrm{d}x \leq \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^m} |K(x + h, y) - K(x, y)|^{p'} \,\mathrm{d}y \right)^{\frac{q}{p'}} \mathrm{d}x$$

und

$$\int_{\mathbb{R}^n \setminus B_R(0)} |Tf(x)|^q \,\mathrm{d}x \leq \int_{\mathbb{R}^n \setminus B_R(0)} \left(\int_{\mathbb{R}^m} |K(x, y)|^{p'} \,\mathrm{d}y \right)^{\frac{q}{p'}} \mathrm{d}x.$$

Wegen $\|K\| < \infty$ geht der letzte Ausdruck gegen 0 für $R \to \infty$.
Wir zeigen nun, dass für $h \to 0$ auch der erste Ausdruck gegen 0 konvergiert, womit die Kompaktheit von T dann aus Satz 2.17 folgt. Sei hierzu $K_h(x, y) := K(x + h, y)$. Wir betrachten $\|K_h - K\|$, allerdings bilden wir die Norm nun über $\mathbb{R}^n \times \mathbb{R}^m$, und approximieren K durch beschränkte Kerne mit kompaktem Träger

$$K_R(x, y) := \begin{cases} K(x, y) & \text{falls } |x|, |y| \leq R,\ |K(x, y)| \leq R, \\ 0 & \text{sonst.} \end{cases}$$

Sei $E_R := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid K(x, y) \neq K_R(x, y)\}$, dann gilt

$$|K_h - K| \leq |K_{Rh} - K_R| + |(\mathbf{1}_{E_R} \cdot K)_h| + |\mathbf{1}_{E_R} \cdot K|$$

und

$$\|K_h - K\| \leq C(\|K_{Rh} - K_R\| + \|\mathbf{1}_{E_R} \cdot K\|).$$

Es gilt $E_{R'} \subset E_R$ für $R' > R$ und $\text{meas}(\cap_{R>0} E_R) = 0$. Damit konvergiert $\|\mathbf{1}_{E_R} \cdot K\|$ gegen 0 für $R \to \infty$ nach dem Satz von der monotonen Konvergenz. Wir schätzen nun den ersten Summanden ab. Da K_R beschränkt ist und einen kompakten Träger hat, gilt für den Fall $\frac{q}{p'} \geq 1$

$$\|K_{Rh} - K_R\|^q \leq C(R) \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} |(K_{Rh} - K_R)(x,y)|^{p'} \, \mathrm{d}y \, \mathrm{d}x.$$

Ist $r := \frac{p'}{q} > 1$, dann gilt nach der Hölder'schen Ungleichung mit Exponent r

$$\begin{aligned} \|K_{Rh} - K_R\|^{p'} &= \left(\int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^m} |K_{Rh} - K_R|^{p'} (x,y) \mathrm{d}y \right)^{\frac{1}{r}} \mathrm{d}x \right)^r \\ &\leq C(R) \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} |(K_{Rh} - K_R)(x,y)|^{p'} \, \mathrm{d}y \, \mathrm{d}x. \end{aligned}$$

Das zuvor bewiesene Lemma liefert $K_{Rh} \to K_R$ in $L^{p'}(\mathbb{R}^n \times \mathbb{R}^m)$ für $|h| \to 0$. Damit sind die Voraussetzungen von Satz 2.17 erfüllt, also ist T kompakt. □

2.4 Das Spektrum kompakter Operatoren

In diesem Abschnitt sei X ein Hilbert-Raum über $\mathbb{C}$. Analog zu $L(X)$ setzen wir $K(X) := K(X;X)$.

2.27 Definition. Die Menge

$$\rho(T) := \{\lambda \in \mathbb{C} \mid \mathcal{R}(\lambda I - T) = X \wedge \mathcal{N}(\lambda I - T) = \{0\}\}.$$

heißt *Resolventenmenge* von T. Das *Spektrum* von T ist dann definiert als

$$\sigma(T) := \mathbb{C} \setminus \rho(T).$$

Es kann zerlegt werden in das *Punktspektrum*

$$\sigma_p(T) := \{\lambda \in \sigma(T) \mid \mathcal{N}(\lambda I - T) \neq \{0\}\},$$

in das *kontinuierliche Spektrum*

$$\sigma_c(T) := \{\lambda \in \sigma(T) \mid \mathcal{N}(\lambda I - T) = \{0\} \wedge \mathcal{R}(\lambda I - T) \neq X \wedge \overline{\mathcal{R}(\lambda I - T)} = X\},$$

und das *Residualspektrum*

$$\sigma_r(T) := \{\lambda \in \sigma(T) \mid \mathcal{N}(\lambda I - T) = \{0\} \wedge \overline{\mathcal{R}(\lambda I - T)} \neq X\}.$$

Bemerkung ([1])**.** Es gilt $\lambda \in \sigma_p(T)$ genau dann, wenn ein $x \neq 0$ existiert mit

$$Tx = \lambda x.$$

Dann heißt λ *Eigenwert* zum *Eigenvektor* x. Handelt es sich bei X um einen Funktionenraum, so nennt man x auch *Eigenfunktion*. Der Unterraum

$$\mathcal{N}(\lambda I - N)$$

heißt *Eigenraum* von T zum Eigenwert λ. Ist $\dim X = \infty$ und $T \in K(X)$, so gilt $0 \in \sigma(T)$. Im Allgemeinen ist 0 aber kein Eigenwert. Jeder Eigenraum von T ist ein *T-invarianter Unterraum*.

2.28 Definition. Ein Unterraum $Y \subset X$ heißt *T-invariant*, falls $T(Y) \subset Y$ gilt.

Um später eine Aussage über die Kondition des Eigenwertproblems treffen zu können, benötigen wir die Resolventenfunktion.

2.29 Definition. Für $\lambda \in \rho(T)$ ist $\lambda I - T : X \to X$ bijektiv und wir können die Inverse definieren. Sei

$$R(\lambda, T) := (\lambda I - T)^{-1} \qquad \text{in } L(X).$$

$R(\lambda, T)$ heißt *Resolvente* von T in λ. Als Funktion von λ heißt $R(\lambda, T)$ *Resolventenfunktion*.

2.30 Satz. *$\rho(T)$ ist offen und die Resolventenfunktion ist eine komplex analytische Abbildung $R(\cdot, T) : \rho(T) \to L(X)$. Dies bedeutet, für jedes $\lambda_0 \in \rho(T)$ existiert ein $r_0 > 0$, so dass $B_{r_0}(\lambda_0) \subset \rho(T)$ gilt und $R(\lambda, T)$ sich für $\lambda \in B_{r_0}(\lambda_0)$ in einer Potenzreihe in $(\lambda - \lambda_0)$ entwickeln lässt mit Koeffizienten in $L(X)$. Insbesondere ist $R(\lambda, T)$ dann holomorph und es gilt*

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

Beweis. Sei $\lambda \in \rho(T)$. Für beliebiges $\mu \in \mathbb{C}$ erhalten wir

$$(\lambda - \mu)I - T = (\lambda I - T) \underbrace{\left(I - \mu R(\lambda, T)\right)}_{=:S(\mu)}.$$

Nach Satz 2.9 ist $S(\mu)$ invertierbar, wenn

$$|\mu| \cdot \|R(\lambda, T)\| < 1.$$

Dann ist auch $\lambda - \mu \in \rho(T)$ und es gilt

$$R(\lambda - \mu, T) = S(\mu)^{-1} R(\lambda, T) = \sum_{k=0}^{\infty} \mu^k R(\lambda, T)^{k+1}.$$

Für $r := \|R(\lambda, T)\|^{-1}$ ist dann $B_r(\lambda) \subset \rho(T)$, also ist $\operatorname{dist}(\lambda, \sigma(T)) \geq r$ und somit $\|R(\lambda, T)\|^{-1} \leq \operatorname{dist}(\lambda, \sigma(T))$. □

2.31 Satz ([1]). *Falls $T \in L(X) \setminus \{0\}$, ist $\sigma(T)$ kompakt und nichtleer. Ferner gilt*

$$\sup_{\lambda \in \sigma(T)} |\lambda| = \lim_{i \to \infty} \left\|T^i\right\|^{\frac{1}{i}} \leq \|T\|.$$

Dieses Supremum heißt Spektralradius *von T.*

Im weiteren Verlauf dieser Arbeit interessieren wir uns für $\sigma_p(T)$, also das Punktspektrum. Wir suchen Vektoren $x \in X \setminus \{0\}$ mit

$$Tx = \lambda x.$$

Es gilt der Spektralsatz für kompakte Operatoren aus [1].

2.32 Satz (Riesz-Schauder, [1]). *Für jeden Operator $T \in K(X)$ gilt:*

1. *$\sigma(T) \setminus \{0\}$ besteht aus abzählbar (endlich oder unendlich) vielen Eigenwerten mit 0 als einzig möglichem Häufungspunkt. Falls $\sigma(T)$ unendlich viele Elemente enthält, ist also $\overline{\sigma(T)} = \sigma_p(T) \cup \{0\}$.*

2. *Für $\lambda \in \sigma(T) \setminus \{0\}$ ist*

   $$1 \leq n_\lambda := \max\{n \in \mathbb{N} \mid \mathcal{N}((\lambda I - T)^{n-1}) \neq \mathcal{N}((\lambda I - T)^n)\} < \infty.$$

   n_λ heißt Index *von λ und* $\dim(\mathcal{N}(\lambda I - T))$ *heißt* Vielfachheit *von λ.*

3. (Riesz-Zerlegung) *Für $\lambda \in \sigma(T) \setminus \{0\}$ gilt*

   $$X = \mathcal{N}((\lambda I - T)^{n_\lambda}) \oplus \mathcal{R}((\lambda I - T)^{n_\lambda}).$$

   Beide Unterräume sind abgeschlossen, T invariant und $\mathcal{N}((\lambda I - T)^{n_\lambda})$ ist endlich-dimensional.

4. *$\sigma(T\big|_{\mathcal{R}((\lambda I - T)^{n_\lambda})}) = \sigma(T) \setminus \{\lambda\}$.*

5. *Ist P_λ für $\lambda \in \sigma(T) \setminus \{0\}$ die Projektion auf $\mathcal{N}((\lambda I - T)^{n_\lambda})$ bezüglich der Darstellung in 3. , dann gilt*

   $$P_\lambda P_\mu = \delta_{\lambda,\mu} P_\lambda \qquad \text{für } \lambda, \mu \in \sigma(T) \setminus \{0\}.$$

2.33 Satz ([1]). *Sei $T \in L(X)$ normal und X ein Hilbert-Raum über $\mathbb{C}$, dann gilt*

$$\sup_{\lambda \in \sigma(T)} |\lambda| = \|T\| .$$

Für kompakte, normale Operatoren gilt der folgende Spektralsatz aus [1].

2.34 Satz ([1]). *Sei X ein Hilbert-Raum über $\mathbb{C}$. Ist $T \in K(X) \setminus \{0\}$ normal, dann gilt:*

1. *Es existiert ein Orthonormalsystem $(e_k)_{k \in \mathcal{J}}$ mit $\mathcal{J} \subset \mathbb{N}$ und $(e_i, e_j)_X = \delta_{i,j}$, sowie eine zugehörige Menge $\{\lambda_k \mid k \in \mathcal{J}\} \subset \mathbb{C} \setminus \{0\}$, so dass*

   $$Te_k = \lambda_k e_k \qquad \text{für } k \in \mathcal{J}, \sigma(T) \setminus \{0\} = \{\lambda_k \mid k \in \mathcal{J}\}.$$

   Hierbei ist zu beachten, dass die Eigenwerte λ_k für verschiedene k gleich sein dürfen. Ist $\mathcal{J}$ unendlich, gilt $\lambda_k \to 0$ für $k \to \infty$.

2. *Für die Indizes gilt $n_{\lambda_k} = 1$ für jedes k.*

3. *$X = \mathcal{N}(T) \oplus \overline{\operatorname{span}\{e_k \mid k \in \mathcal{J}\}}$. Zusätzlich gilt hierbei $(x, y)_X = 0$ für $x \in \mathcal{N}(T)$ und $y \in \overline{\operatorname{span}\{e_k \mid k \in \mathcal{J}\}}$.*

4. *$Tx = \sum_{k \in \mathcal{J}} \lambda_k (x, e_k)_X e_k \qquad$ für $x \in X$.*

Bemerkung ([1]). Unter den Voraussetzungen des Satzes bemerken wir:

1. Falls T selbstadjungiert ist, gilt $\sigma_p(T) \subset [-\|T\|, \|T\|] \subset \mathbb{R}$ und $\|T\|$ oder $-\|T\|$ ist Eigenwert von T.

2. Ist T zusätzlich positiv semidefinit, also $(x, Tx)_X \geq 0$ für jedes $x \in X$, gilt $\sigma_p(T) \subset [0, \|T\|]$. Dann ist $\|T\|$ ein Eigenwert von T.

2.5 Die Singulärwertzerlegung

Im Folgenden sei $\Omega \subset \mathbb{R}^n$ ein Gebiet, das heißt offen, zusammenhängend und nicht leer. Wir betrachten nun den speziellen Fall $X = L^2(\Omega) := L^2(\Omega, \mathbb{R})$.

2.35 Definition. Der Integralkern $K \in L^2(\Omega \times \Omega)$ heißt *symmetrisch*, falls

$$K(x, y) = K(y, x) \qquad \text{für alle } x, y \in \Omega$$

gilt.

2.36 Lemma. *Sei $T \in L(L^2(\Omega))$ ein Hilbert-Schmidt-Operator. T ist symmetrisch, falls der Integralkern symmetrisch ist.*

Beweis. Die Behauptung folgt aus

$$\begin{aligned}(u,Tv)_{L^2(\Omega)} &= \int_\Omega u(x)\int_\Omega K(x,y)v(y)\mathrm{d}y\ \mathrm{d}x\\ &\overset{\text{Fubini}}{=} \int_\Omega\int_\Omega K(x,y)u(x)\mathrm{d}x\ v(y)\mathrm{d}y\\ &= (Tu,v)_{L^2(\Omega)} = (v,Tu)_{L^2(\Omega)}.\end{aligned}$$

□

2.37 Satz. *Seien $\Omega_1,\Omega_2\subset\mathbb{R}^n$ zwei Gebiete und $f\in L^2(\Omega_1\times\Omega_2)$. Dann kann f dargestellt werden als*

$$f(x,y)=\sum_{i=1}^{\infty}\alpha_i\phi_i(x)\psi_i(y)$$

mit $\alpha_i\in\mathbb{R}$ und Funktionen $\phi_i\in L^2(\Omega_1)$ und $\psi_i\in L^2(\Omega_2)$ für jedes $i=1,2,\ldots$.

Beweis. Wir definieren den Hilbert-Schmidt-Operator

$$T:\ L^2(\Omega_2)\longrightarrow L^2(\Omega_1),\qquad Tu(x):=\int_{\Omega_2}f(x,y)u(y)\mathrm{d}y.$$

Der zu T adjungierte Operator T^* ist dann gegeben durch

$$T^*:\ L^2(\Omega_1)\longrightarrow L^2(\Omega_2),\qquad T^*u(y):=\int_{\Omega_1}f(x,y)u(x)\mathrm{d}x.$$

Damit definiert $TT^*:\ L^2(\Omega_1)\to L^2(\Omega_1)$ einen linearen Operator. Dieser besitzt die Darstellung

$$\begin{aligned}TT^*u(x)&=\left(T\left[\int_{\Omega_1}f(y,\cdot)u(y)\mathrm{d}y\right]\right)(x)=\int_{\Omega_2}f(x,z)\int_{\Omega_1}f(y,z)u(y)\mathrm{d}y\ \mathrm{d}z\\ &\overset{\text{Fubini}}{=}\int_{\Omega_1}\underbrace{\left[\int_{\Omega_2}f(x,z)f(y,z)\mathrm{d}z\right]}_{=:K(x,y)}u(y)\mathrm{d}y.\end{aligned}$$

Wir zeigen nun, dass $K(x,y)\in L^2(\Omega_1\times\Omega_1)$ gilt.

$$\begin{aligned}\|K\|^2 &= \int_{\Omega_1}\int_{\Omega_1}|K(x,y)|^2\,\mathrm{d}y\ \mathrm{d}x=\int_{\Omega_1}\int_{\Omega_1}\left|\int_{\Omega_2}f(x,z)f(y,z)\mathrm{d}z\right|^2\mathrm{d}y\ \mathrm{d}x\\ &\overset{\text{CSU}}{\le} \int_{\Omega_1}\int_{\Omega_1}\int_{\Omega_2}f(x,z)^2\mathrm{d}z\int_{\Omega_2}f(y,z)^2\mathrm{d}z\ \mathrm{d}y\ \mathrm{d}x\\ &= \left(\int_{\Omega_1}\int_{\Omega_2}f(x,z)^2\mathrm{d}z\ \mathrm{d}x\right)\left(\int_{\Omega_1}\int_{\Omega_2}f(y,z)^2\mathrm{d}z\ \mathrm{d}y\right)\\ &= \|f\|^4_{L^2(\Omega_1\times\Omega_2)}<\infty.\end{aligned}$$

$TT^* \in L(\Omega_1)$ ist also ein Hilbert-Schmidt-Operator und daher kompakt. Weiterhin ist TT^* selbstadjungiert und folglich normal.
Sei $T^*u =: v \in L^2(\Omega_2)$, dann ist offensichtlich

$$(u, TT^*u)_{L^2(\Omega_1)} = (T^*u, T^*u)_{L^2(\Omega_2)} = (v, v)_{L^2(\Omega_2)} = \|v\|^2_{L^2(\Omega_2)} \geq 0,$$

somit ist TT^* auch positiv semidefinit.
Nach dem Spektralsatz für kompakte, normale Operatoren existiert nun ein Orthonormalsystem $(\phi_i)_{i\in\mathcal{J}}$, mit $\mathcal{J}$ endlich oder abzählbar unendlich, so dass

$$TT^*\phi_i = \lambda_i\phi_i$$

gilt. Dabei ist $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_i \to 0$, nach der Bemerkung im Anschluss des Spektralsatzes für kompakte, normale Operatoren.
Wir definieren die Funktionen $\psi_i \in L^2(\Omega_2)$ als

$$\psi_i(y) = \frac{1}{\sqrt{\lambda_i}}T^*\phi_i(y) = \frac{1}{\sqrt{\lambda_i}}\int_{\Omega_1} f(x,y)\phi_i(x)\mathrm{d}x.$$

Die $(\psi_i)_{i\in\mathcal{J}}$ bilden ebenfalls ein Orthonormalsystem. Es gilt nämlich

$$\begin{aligned}(\psi_i, \psi_j)_{\Omega_2} &= \frac{1}{\sqrt{\lambda_i\lambda_j}}(\frac{1}{\sqrt{\lambda_i}}T^*\phi_i, \frac{1}{\sqrt{\lambda_j}}T^*\phi_j)_{L^2(\Omega_2)} = \frac{1}{\sqrt{\lambda_i\lambda_j}}(TT^*\phi_i, \phi_j)_{L^2(\Omega_1)} \\ &= \frac{1}{\sqrt{\lambda_i\lambda_j}}(\lambda_i\phi_i, \phi_j)_{L^2(\Omega_1)} = \frac{\sqrt{\lambda_i}}{\sqrt{\lambda_j}}(\phi_i, \phi_j)_{L^2(\Omega_1)} = \delta_{i,j}.\end{aligned}$$

Andererseits ergibt sich die Beziehung

$$T\psi_i(x) = \frac{1}{\sqrt{\lambda_i}}TT^*\phi_i(x) = \frac{1}{\sqrt{\lambda_i}}\lambda_i\phi_i(x) = \sqrt{\lambda_i}\phi_i(x).$$

Wir ergänzen nun die Orthonormalsysteme $(\phi_i)_{i\in\mathcal{J}}$ und $(\psi_i)_{i\in\mathcal{J}}$ zu Orthonormalbasen von $L^2(\Omega_1)$, respektive $L^2(\Omega_2)$. Wir können f dann darstellen als

$$f(x,y) = \sum_{i,j=1}^{\infty}(f, \phi_i\otimes\psi_j)_{L^2(\Omega_1\times\Omega_2)}\phi_i\otimes\psi_j.$$

Dann gilt

$$\begin{aligned}(f, \phi_i\otimes\psi_j)_{L^2(\Omega_1\times\Omega_2)} &\overset{\text{Fubini}}{=} \int_{\Omega_1}\int_{\Omega_2} f(x,y)\phi_i(x)\psi_j(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{\Omega_1}\phi_i(x)T\psi_j(x)\mathrm{d}x \\ &= (\phi_i, T\psi_j)_{L^2(\Omega_1)}.\end{aligned}$$

Die $(\phi_i)_{i\in\mathcal{J}}$ bilden eine Basis von $\mathcal{R}(TT^*)$. Damit gilt $TT^*\phi_j = 0$ falls $j \notin \mathcal{J}$. Das bedeutet

$$T\psi_i(x) = \frac{1}{\sqrt{\lambda_i}} TT^*\phi_i(x) = \begin{cases} \sqrt{\lambda_i}\phi_i(x) & i \in \mathcal{J}, \\ 0 & i \notin \mathcal{J}. \end{cases}$$

Somit erhalten wir

$$(\phi_i, T\psi_j)_{L^2(\Omega_1)} = \sqrt{\lambda_i}\delta_{i,j}.$$

Durch Einsetzen der Koeffizienten erhalten wir für f die Darstellung

$$f(x,y) = \sum_{i\in\mathcal{J}} \sqrt{\lambda_i}\phi_i(x)\psi_i(y).$$

□

Bemerkung ([15])**.** Analog zu $TT^* \in L(\Omega_1)$ betrachtet man $T^*T \in L(\Omega_2)$. Der Operator T^*T besitzt dieselben Eigenwerte wie TT^*. Man erhält dann Eigenfunktionen $\tilde{\psi}_i \in L^2(\Omega_2)$ und zugehörige orthonormale Funktionen $\tilde{\phi}_i \in L^2(\Omega_1)$. Es gelten die Beziehungen

$$\psi_i = \tilde{\phi}_i, \quad \phi_i = \tilde{\psi}_i, \quad i \in \mathcal{J}.$$

Insgesamt ergibt sich also dieselbe Darstellung für f, nur die Rollen von Ω_1 und Ω_2 sind vertauscht.

Bricht man die Singulärwertzerlegung nach $m \in \mathbb{N}$ Termen ab, ergibt sich die folgende Bestapproximationseigenschaft.

2.38 Satz. *Die abgebrochene Singulärwertzerlegung ist optimal in $L^2(\Omega_1 \times \Omega_2)$ im folgenden Sinne. Seien $m \in \mathbb{N}$ und $(u_i)_{i=1}^m$,$(v_i)_{i=1}^m$ normiert,*

$$f(x,y) \approx \sum_{i=1}^m \alpha_i u_i(x) v_i(y),$$

dann gilt

$$\left\| f - \sum_{i=1}^m \sqrt{\lambda_i}(\phi_i \otimes \psi_i) \right\|_{L^2(\Omega_1\times\Omega_2)} \leq \left\| f - \sum_{i=1}^m \alpha_i (u_i \otimes v_i) \right\|_{L^2(\Omega_1\times\Omega_2)}.$$

Beweis. Ohne Beschränkung der Allgemeinheit kann angenommen werden, dass die $(u_i)_{i=1}^m$ orthonormal sind. Sonst orthonormalisiert man die Vektoren in $U_m := \operatorname{span}\{u_1, \ldots, u_m\}$. Dies führt lediglich zu veränderten Koeffizienten

α_i und veränderten Funktionen v_i, die Approximation bleibt aber die gleiche. Wir definieren durch

$$f_m(x,y) := \sum_{i=1}^{m} \underbrace{(f(\cdot,y),u_i)_{L^2(\Omega_1)}}_{=:\alpha_i} u_i(x),$$

die orthogonale Projektion von f auf U_m. Dies ist die Bestapproximation von $f(\cdot,y)$ in U_m bezüglich der $L^2(\Omega_1)$-Norm, also

$$f_m(\cdot,y) = \operatorname*{argmin}_{g\in U_m} \|f(\cdot,y)-g\|_{L^2(\Omega_1)}.$$

Hieraus folgt

$$v_i(y) = \frac{1}{\alpha_i}\int_{\Omega_1} f(x,y)u_i(x)\mathrm{d}x \qquad \text{für } i=1,\dots,m.$$

Nun ergänzen wir die Folge $(u_i)_{i=1}^m$ zu einer Orthonormalbasis von $L^2(\Omega_1)$ und definieren α_i und v_i für $i>m$ analog. Dann gilt

$$f(x,y) = \sum_{i=1}^{\infty} \alpha_i u_i(x) v_i(y).$$

Die ürsprüngliche Approximation f_m erhalten wir, wenn wir die Reihe nach m Termen abbrechen. Sei nun $\varepsilon_m(x,y)$ der Abschneidefehler, das heißt

$$\varepsilon_m(x,y) := f(x,y) - f_m(x,y) = \sum_{i=m+1}^{\infty} \alpha_i u_i(x) v_i(y).$$

Setzen wir nun die Darstellung der v_i in die von f_m ein, so ergibt sich für die $L^2(\Omega_2)$-Norm des Fehlers

$$\begin{aligned}
&\int_{\Omega_2} \varepsilon_m^2(x,y)\mathrm{d}y \\
=\ &\int_{\Omega_2} \left(\sum_{i=m+1}^{\infty} u_i(x) \int_{\Omega_1} f(x,y)u_i(x)\mathrm{d}x \right)^2 \mathrm{d}y \\
=\ &\sum_{i,j=m+1}^{\infty} u_i(x)u_j(x) \int_{\Omega_1}\int_{\Omega_1} \Bigg(\underbrace{\int_{\Omega_2} f(x_1,z)f(x_2,z)\mathrm{d}z}_{=K(x_1,x_2)} \Bigg) \mathrm{d}x_1\ \mathrm{d}x_2 \\
=\ &\sum_{i,j=m+1}^{\infty} u_i(x)u_j(x)(K, u_i\otimes u_j)_{L^2(\Omega_1\times\Omega_1)}.
\end{aligned}$$

Integrieren wir nun noch x über Ω_2, erhalten wir unter Berücksichtigung der Orthonormalität der u_i

$$\|\varepsilon_m\|^2_{L^2(\Omega_1\times\Omega_2)} = \int_{\Omega_1}\int_{\Omega_2} \varepsilon_m^2(x,y)\mathrm{d}y\ \mathrm{d}x = \sum_{i=m+1}^{\infty} (K, u_i\otimes u_i)_{L^2(\Omega_1\times\Omega_1)}.$$

Um die Optimalität der Singulärwertzerlegung zu zeigen, minimieren wir nun $\|\varepsilon_m\|^2_{L^2(\Omega_1\times\Omega_2)}$ unter der Nebenbedingung, dass die $(u_i)_i$ orthonormal sind. Dann muss für beliebiges $i > m$ die Funktion u_i das Sattelpunktproblem

$$F(u_i,\lambda_i) \longrightarrow \inf_{u_i\in L^2(\Omega_1)} \sup_{\lambda_i\in\mathbb{R}}$$

mit

$$F(u_i,\lambda_i) := \sum_{j=m+1}^{\infty} \left[(K, u_j\otimes u_j)_{L^2(\Omega_1\times\Omega_1)} - \lambda_j((u_j,u_j)_{L^2(\Omega_1)} - 1)\right]$$

lösen. Es gilt

$$\begin{aligned} F(u_i+tv,\lambda_i) &= F(u_i,\lambda_i) + 2t[(K, v\otimes u_i)_{L^2(\Omega_1\times\Omega_1)} - \lambda_i(u_i,v)_{L^2(\Omega_1)}] \\ &+ t^2[(K, v\otimes v)_{L^2(\Omega_1\times\Omega_1)} - \lambda_i(v,v)_{L^2(\Omega_1)}]. \end{aligned}$$

Damit muss die Lösung u_i des Sattelpunktproblems also

$$\frac{\partial}{\partial t}F(u_i+tv,\lambda_i)\bigg|_{t=0} = 2[(K, v\otimes u_i)_{L^2(\Omega_1\times\Omega_1)} - \lambda_i(u_i,v)_{L^2(\Omega_1)}] \stackrel{!}{=} 0$$

für alle $v\in L^2(\Omega_1)$ erfüllen. Das ist äquivalent zu

$$\int_{\Omega_1} K(x,y)u_i(y)\mathrm{d}y = \lambda_i u_i(x).$$

Somit ist gezeigt, dass keine Approximation mit m Summanden einen geringeren Fehler in $L^2(\Omega_1\times\Omega_2)$ besitzen kann, als die abgebrochene Singulärwertzerlegung. □

Kapitel 3

Das Eigenwertproblem kompakter Operatoren

In diesem Kapitel sei $(X, (\cdot,\cdot)_X)$ ein Hilbert-Raum über $\mathbb{R}$, im Zusammenhang mit der Singulärwertzerlegung speziell $X = L^2(\Omega)$ mit $\Omega \subset \mathbb{R}^n$ Gebiet.

Wir betrachten das verallgemeinerte Eigenwertproblem

$$Ku = \lambda Mu \tag{3.1}$$

für einen kompakten, symmetrischen Operator K und einen symmetrischen, gleichmäßig elliptischen Operator M.
Da X ein Hilbert-Raum ist, ist die Lösung der Operatorgleichung

$$Ku = \lambda Mu$$

äquivalent zur Lösung der Variationsformulierung

$$b_K(u, v) = b_M(\lambda u, v) \qquad \text{für alle } v \in X$$

(vgl. [29]). Wegen $M \in L^+(X)$ ist das allgemeine Eigenwertproblem (3.1) äquivalent zu

$$Au = \lambda u, \tag{3.2}$$

wobei wir $A := M^{-1}K$ setzen.

3.1 Lemma. *Der lineare Operator A ist symmetrisch in $X_M := (X, (\cdot,\cdot)_M)$. Hierbei ist $(\cdot,\cdot)_M := (\cdot, M\cdot)_X = (M\cdot,\cdot)_X$.*

Beweis. Es gilt

$$(x, A, y)_M = (x, MM^{-1}Ky)_X = (Kx, y)_X = (MM^{-1}Kx, y)_X = (Ax, y)_M.$$

□

Nach dem Satz von der inversen Abbildung (vgl. [1]) ist M^{-1} stetig. Hinreichend für die Kompaktheit von A ist damit nach Satz 2.22, dass K kompakt ist, was im Folgenden stets angenommen sei. Eine wichtige Größe bei der Bestimmung der Eigenwerte ist der *Rayleigh-Quotient*.

3.2 Definition. Seien A, M, K wie oben definiert, dann ist für jedes $v \in X \setminus \{0\}$

$$\Lambda(v) := \frac{(Kv, v)_X}{(Mv, v)_X} = \frac{(Av, v)_M}{(v, v)_M}$$

der *Rayleigh-Quotient* von v definiert.

Wir entwickeln $v \in X$ in der Eigenbasis, also

$$v = v_0 + \sum_{i \in I} c_i u_i \qquad \text{mit } v_0 \in \ker A,$$

wobei die Eigenwerte absteigend geordnet seien, das heißt $\lambda_1 \geq \lambda_2 \geq \ldots$. Da wir den allgemeinen Fall kompakter, symmetrischer Eigenwertprobleme betrachten, sind hier auch negative Eigenwerte explizit zugelassen. Die Koeffizienten c_i ergeben sich zu $c_i = (v, u_i)_M$. Die obige Zerlegung ist eben die aus dem Spektralsatz für kompakte, normale Operatoren. Aus dieser Darstellung sieht man leicht den folgenden Satz ein (vgl. [8]).

3.3 Satz ([8])**.** *Sei* $W_{i-1} := \operatorname{span}\{u_1, \ldots, u_{i-1}\}$ *für* $i \geq 2$ *und sei* $W_{i-1}^{\perp}$ *das orthogonale Komplement bezüglich* $(\cdot, \cdot)_M$*, das heißt* $W_{i-1}^{\perp} \cap W_{i-1} = \{0\}$ *und* $(w, w^{\perp})_M = 0$ *für jedes Paar* $w \in W_{i-1}$ *und* $w^{\perp} \in W_{i-1}^{\perp}$*. Dann gilt*

$$\lambda_1 = \max_{v \neq 0} \Lambda(v) \qquad \text{und} \qquad \lambda_i = \max_{v \in W_{i-1}^{\perp} \setminus \{0\}} \Lambda(v).$$

Eine Verallgemeinerung dieses Satzes ist der Satz von Courant-Fisher, wonach sich alle Eigenwerte als Sattelpunkte des Rayleigh-Quotienten beschreiben lassen.

3.4 Satz (Courant-Fisher)**.** *Seien* $M \in L^+(X)$ *und* $K \in K(X)$ *symmetrisch so, dass* $A = M^{-1}K$ *kompakt ist. Dann gilt*

$$\lambda_i = \min_{V_{i-1}} \max_{v \in V_{i-1}^{\perp} \setminus \{0\}} \Lambda(v) \qquad \text{und}$$

$$\lambda_i = \max_{V_i} \min_{v \in V_i \setminus \{0\}} \Lambda(v) \qquad \text{für } i = 1, 2, \ldots.$$

Hierbei bezeichnet V_i *einen beliebigen* $(i-1)$*-dimensionalen Unterraum und das orthogonale Komplement wird gebildet bezüglich des* $(\cdot, \cdot)_M$*-Innenproduktes.*

Beweis. Die Räume W_i seien so, wie im vorangegangenen Satz definiert. Ein einfaches Dimensionsargument liefert dann, dass ein $v \in V_{i-1}^{\perp} \cap W_i$ mit $v \neq 0$ existiert. Würde $V_{i-1}^{\perp} \cap W_i = \emptyset$ gelten, folgte daraus $W_i \subset V_{i-1}$, was wegen $\dim(W_i) > \dim(V_{i-1})$ nicht sein kann. Daher ist $\Lambda(v) \geq \lambda_i$ und somit $\sup_{v \in V_{i-1}^{\perp}} \Lambda(v) \geq \lambda_i$. Andererseits ist für die Wahl $V_{i-1} = W_{i-1}$ dann $u_i \in V_{i-1}^{\perp}$ und der Wert $\Lambda(v) = \lambda_i$ kann angenommen werden. Damit ist die erste Gleichheit gezeigt.
Analog existiert ein $v \in V_i \cap W_{i-1}^{\perp}$ mit $v \neq 0$. Also ist $\Lambda(v) \leq \lambda_i$ und somit auch $\inf_{v \in V_i \setminus \{0\}} \Lambda(v) \leq \lambda_i$. Für die Wahl $V_i = W_i$ wird das Minimum angenommen, $\lambda_i = \min_{v \in V_i \setminus \{0\}} \Lambda(v)$. Damit ist auch die zweite Behauptung gezeigt. □

Eine erste Fehlerbschätzung liefert der folgende Satz von Weyl. Dieser besagt, dass der Fehler in den Eigenwerten zweier Operatoren höchstens so groß ist, wie die Norm ihrer Differenz. Man kann ihn mit Hilfe des Satzes von Courant-Fisher zeigen (vgl. [8]).

3.5 Satz (Weyl, [8])**.** *Sei X ein Hilbert-Raum und seien $T, \hat{T} \in L(X)$. Die Eigenwerte von T beziehungsweise $\hat{T}$ seien absteigend geordnet, beginnend mit λ_1 beziehungsweise $\hat{\lambda}_1$. Dann gilt*

$$\left|\lambda_i - \hat{\lambda}_i\right| \leq \left\|T - \hat{T}\right\| \qquad \text{für } i = 1, 2, \ldots.$$

3.1 Die Rayleigh-Ritz-Approximation

Das Ausgangsproblem ist, den größten Eigenwert λ_1 von

$$Kx = \lambda Mx$$

zu finden. Das heißt, wir suchen die Lösung des Maximierungsproblems

$$\lambda_1 = \max_{v \neq 0} \Lambda(v).$$

Die Idee bei der Rayleigh-Ritz-Approximation ist, dieses Maximierungsproblem nur in einem endlichdimensionalen Teilraum $X_n \subset X$ mit $\dim(X_n) = n$ zu lösen. Es gilt

$$\frac{\mathrm{d}}{\mathrm{d}t} \Lambda(u + tv)\big|_{t=0} = \frac{2}{\|u\|_M}(Ku - \Lambda(u)Mu, v)_X,$$

denn

$$\begin{aligned}
& \frac{\mathrm{d}}{\mathrm{d}t}\Lambda(u+tv)\Big|_{t=0} \\
= & \frac{\mathrm{d}}{\mathrm{d}t}\frac{(u+tv,K(u+tv))_X}{(u+tv,u+tv)_M}\Big|_{t=0} \\
= & \frac{2((u,Kv)_X+t(v,Kv)_X)((u,u)_M+2t(u,v)_M+t^2(v,v)_M)}{(u+tv,u+tv)_M^2}\Big|_{t=0} \\
& -\frac{2((u,Ku)_X+2t(u,Kv)_X+t^2(v,Kv)_X)((u,v)_M+t(v,v)_M)}{(u+tv,u+tv)_M^2}\Big|_{t=0} \\
= & \frac{2(u,Kv)_X(u,u)_M-2(u,Ku)_X(u,v)_M}{(u,u)_M^2} \\
= & \frac{2}{\|u\|_M^2}((Ku,v)_X-\Lambda(u)(u,v)_M) \\
= & \frac{2}{\|u\|_M^2}(Ku-\Lambda(u)Mu,v)_X,
\end{aligned}$$

wobei $\|\cdot\|_M$ die Energienorm bezüglich M bezeichne. Hier ergibt sich noch einmal genau die Aussage des Satzes von Courant-Fisher: Die Eigenwerte sind die Extrema (Sattelpunkte) des Rayleigh-Quotienten. Damit ist das gesuchte λ_1 die größte unter allen Zahlen λ, so dass ein $u \in X \setminus \{0\}$ existiert mit

$$b_K(u,v) = \lambda b_M(u,v) \qquad \text{für alle } v \in X.$$

Statt dieses Problem zu lösen, wird das folgende endlichdimensionale Problem gelöst.
Suche $\hat{\lambda}_1$, die größte unter allen Zahlen $\hat{\lambda}$, sodass ein $\hat{u} \in X_n \setminus \{0\}$ existiert, mit

$$b_K(\hat{u},\hat{v}) = \hat{\lambda} b_M(\hat{u},\hat{v}) \qquad \text{für alle } \hat{v} \in X_n. \tag{3.3}$$

Definiert $P: X \to X_n$ die orthogonale Projektion von X auf X_n und setzt man

$$\hat{K} = PKP \qquad \text{und} \qquad \hat{K} = PKP,$$

so ist die Lösung der Variationsformulierung (3.3) äquivalent zu der Lösung des verallgemeinerten Eigenwertproblemes

$$\hat{K}\hat{u} = \hat{\lambda}\hat{M}\hat{u}, \tag{3.4}$$

wobei hier $\hat{u} \in X_n$ gilt. Führen wir nun eine Basis $\psi_1, \ldots, \psi_n$ auf X_n ein, so erhalten wir wie beim Ritz-Galerkin-Verfahren (vgl. [4]) das algebraische Eigenwertproblem

$$\overline{K}x = \hat{\lambda}\overline{M}x$$

mit $\overline{K} := (b_K(\psi_i, \psi_j))_{i,j=1}^n$, $\overline{M} := (b_M(\psi_i, \psi_j))_{i,j=1}^n$ und $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, dem Koeffizientenvektor von u bezüglich der Basis von X_n (vgl. [8]). Wie dieses algebraische Eigenwertproblem numerisch behandelt werden kann, wird im nächsten Kapitel beschrieben.

Bemerkung. Unter der oben angegebenen Transformation sind $\mathbb{R}^n$ und X_n isometrisch isomorph. Insbesondere besitzen die Probleme also dieselben Eigenwerte (vgl. [8]). Ferner übertragen sich Symmetrie und Elliptizität auf das endlichdimensionale Problem und führen auf symmetrische beziehungsweise positiv definite Matrizen (vgl. [4]).

3.6 Satz. *Sei λ_i der i-te positive Eigenwert von (3.1) und analog $\hat{\lambda}_i$ derjenige von (3.4). Sind die Eigenwerte absteigend geordnet, so gilt*

$$\hat{\lambda}_i \leq \lambda_i.$$

Beweis. Nach dem Satz von Courant-Fisher gilt $\hat{\lambda}_i = \max\limits_{\overline{V}_i} \min\limits_{x \in \overline{V}_i} \overline{\Lambda}(x)$, wobei wir den Rayleigh-Quotienten

$$\overline{\Lambda}(x) := \frac{(\overline{K}x, x)_{\mathbb{R}^n}}{(\overline{M}x, x)_{\mathbb{R}^n}}$$

betrachten und $\overline{V}_i \subset \mathbb{R}^n$ mit $\dim(\overline{V}_i) = i$. Andererseits gilt aber $\overline{\Lambda}(x) = \Lambda(\hat{v})$, wobei $\hat{v}$ bezüglich der Basis $\psi_1, \ldots, \psi_n$ die Darstellung $\hat{v} = \sum_{i=1}^n x_i \psi_i$ besitzt. Daher gilt auch

$$\hat{\lambda}_i = \max_{\hat{V}_i \subset X_n} \min_{\hat{v} \in \hat{V}_i \setminus \{0\}} \Lambda(\hat{v}).$$

Wir vergleichen

$$\hat{\lambda}_i = \max_{\hat{V}_i \subset X_n} \min_{\hat{v} \in \hat{V}_i \setminus \{0\}} \Lambda(\hat{v}) \qquad \text{und} \qquad \lambda_i = \max_{V_i \subset X} \min_{v \in V_i \setminus \{0\}} \Lambda(v),$$

so folgt wegen $X_n \subset X$,

$$\max_{\hat{V}_i \subset X_n} \min_{\hat{v} \in \hat{V}_i \setminus \{0\}} \Lambda(\hat{v}) \leq \max_{V_i \subset X} \min_{v \in V_i \setminus \{0\}} \Lambda(v)$$

und damit die Behauptung. □

Betrachten wir nun noch einmal den Operator $A = M^{-1}K$ und seine Projektion $\hat{A} = PAP$ auf X_n bezüglich des $(\cdot, \cdot)_M$ Innenproduktes. Wir erhalten die Operatorgleichung

$$\hat{A}u = \hat{\lambda}u.$$

Dann ist X_n ein invarianter Unterraum von $\hat{A}$. Falls $\hat{\lambda} \neq 0$ kann man $u \in X$ offensichtlich durch $\hat{u} \in X_n$ ersetzen und gelangt zu der Variationsformulierung

$$(A\hat{u} - \hat{\lambda}\hat{u}, \hat{v})_M = 0 \qquad \text{für alle } \hat{v} \in X_n.$$

Dies entspricht dem verallgemeinerten Eigenwertproblem (3.4). Ferner gilt

$$\left\|(A - \hat{A})v\right\|_M^2 = \left\|P(A - \hat{A})v\right\|_M^2 + \left\|(I - P)(A - \hat{A})v\right\|_M^2 .$$

Setzen wir voraus, dass $\|v\|_M = 1$, dann erhalten wir einerseits

$$\left\|P(A - \hat{A})v\right\|_M \leq \|P\|_M \, \|(I - P)A\|_M = \|(I - P)A\|_M$$

und andererseits

$$\left\|(I - P)(A - \hat{A})v\right\|_M = \|(I - P)Av\|_M \leq \|(I - P)A\|_M .$$

Hierbei ist zu beachten, dass P der Orthoprojektor bezüglich des $(\cdot, \cdot)_M$-Innenproduktes ist. Diese Abschätzungen gelten für jedes $\|v\|_M = 1$. Daher gelangt man zusammen mit dem Satz von Weyl zu der Abschätzung

$$\left|\lambda_i - \hat{\lambda}_i\right| \leq \left\|A - \hat{A}\right\|_M \leq \sqrt{2} \, \|(I - P)A\|_M .$$

Sei $\overline{B_1^M(0)}$ die abgeschlossene Einheitskgel bezüglich des $(\cdot, \cdot)_M$-Innenproduktes. Da A kompakt ist, ist auch $\overline{AB_1^M(0)}$ kompakt. Existiert nun eine approximierende Folge von Unterräumen $\{X_n\}_{n \in \mathbb{N}}$, sodass für die zugehörigen Orthoprojektoren

$$\lim_{n \to \infty} (I - P_n)u = 0 \qquad \text{für alle } u \in X$$

gilt, erhalten wir zusammen mit der Kompaktheit von $\overline{AB_1^M(0)}$ auch die Konvergenz

$$\lim_{n \to \infty} \hat{\lambda}_i = \lambda_i \qquad \text{für } i = 1, 2, \ldots .$$

3.2 Gaps zwischen den Unterräumen

Seien $U_1, U_2 \subset X$ zwei Unterräume mit zugehörigen Orthoprojektoren P_1, P_2. Weiter sei $P_i^{\perp} := I - P_i$ für $i = 1, 2$. Wir definieren den Winkel zwischen zwei Vektoren $x, y \in X \setminus \{0\}$.

3.7 Definition. Für $x, y \in X \setminus \{0\}$ und X ein Hilbert-Raum über $\mathbb{R}$ sei

$$\gamma := \left(\frac{x}{\|x\|_X}, \frac{y}{\|y\|_X} \right)_X .$$

Dann existiert genau ein $\vartheta \in [0, \pi]$ mit $\gamma = \cos(\vartheta)$. ϑ heißt *Winkel* zwischen x und y, in Zeichen: $\angle(x, y) = \vartheta$.

3.8 Definition. Wir definieren für die Unterräume U_1 und U_2 das *Gap* zwischen den Unterräumen

$$\Theta(U_1; U_2) := \|P_1 - P_2\| = \left\|P_1^\perp - P_2^\perp\right\| .$$

Die Abbildung $\Theta(\cdot;\cdot)$ induziert eine Metrik auf dem Raum aller Unterräume von X.

Ferner setzen wir für diesen Abschnitt noch $S_i := S \cap U_i$ für $i = 1, 2$, wobei $S := \{u \in X \mid \|u\|_X = 1\}$, die Einheitssphäre bezeichne. Wir definieren weiter

$$\sigma_i := \min_{u \in S_i} \|P_j u\|_X , \qquad d_i := \max_{u \in S_i} \left\|P_j^\perp u\right\|_X \qquad \text{für } i \neq j,\ i = 1, 2.$$

3.9 Lemma. *Es gilt:*

1. $d_i = \sqrt{1 - \sigma_i^2}$ für $i = 1, 2$.
2. $\left\|P_i^\perp P_j\right\| \leq d_j$ für $i \neq j,\ i = 1, 2$.

Beweis.
Ad 1.

$$\begin{aligned} d_i^2 &= \max_{u \in S_i} \left\|P_j^\perp u\right\|_X^2 = \max_{u \in S_i} \|u - P_j u\|_X^2 = \max_{u \in S_i} [1 - 2(u, P_j u)_X + \|P_j u\|_X^2] \\ &= \max_{u \in S_i} [1 - \|P_j u\|_X^2] = 1 - \sigma_i^2. \end{aligned}$$

Ad 2.

$$\left\|P_i^\perp P_j\right\| = \max_{u \in S} \left\|P_i^\perp P_j u\right\| \leq \max_{u \in S_j} \left\|P_i^\perp u\right\| = d_j.$$

□

3.10 Lemma. *Sei P eine orthogonale Projektion und $x \in S$, dann gilt*

$$\|Pu\|_X = \cos(\vartheta) \qquad \text{und} \qquad \left\|P^\perp u\right\|_X = \sin(\vartheta).$$

wobei $\vartheta \in [0, \pi/2]$ genau der Winkel zwischen Pu und u ist, also $\vartheta = \angle(u, Pu)$.

Beweis. Nach Satz 2.14 gilt $(u, Pu)_X = (Pu, Pu)_X = \|Pu\|_X^2$ und wegen $u \in S$ gilt $\|u\|_X = 1$, somit ergibt sich

$$\cos(\vartheta) = \frac{(u, Pu)_X}{\|u\|_X \|Pu\|_X} = \frac{(Pu, Pu)_X}{\|Pu\|_X} = \|Pu\|_X \geq 0.$$

Insbesondere folgt hieraus, dass $\vartheta \in [0, \pi/2]$. Andererseits gilt

$$\left\|P^\perp u\right\|_X^2 = \|(I - P)u\|_X^2 = \|u\|_X - 2\|Pu\|_X^2 + \|Pu\|_X^2 = 1 - \|Pu\|_X^2,$$

dies bedeutet

$$\left\|P^\perp u\right\|_X = \sqrt{1 - \|Pu\|_X^2} = \sqrt{1 - \cos^2(\vartheta)} = \sin(\vartheta).$$

□

Gemäß dem Lemma kann man σ_i und d_i für $i = 1, 2$ wie folgt interpretieren.

3.11 Lemma.

$$\sigma_i = \min_{u \in S_i} \cos(\angle(u, P_j u)) \qquad \text{für } i \neq j,\ i = 1, 2.$$

$$d_i = \max_{u \in S_i} \sin(\angle(u, P_j u)) \qquad \text{für } i \neq j,\ i = 1, 2.$$

3.12 Lemma. *Es gelte*

$$\dim U_1 = \dim U_2 = n < \infty. \tag{3.5}$$

Dann folgt $d_1 = d_2$.

Beweis. Falls $d_1 = 1$ und $\sigma_1 = 0$ gilt, so existiert ein $u_1 \in U_1$ mit $\|u_1\|_X = 1$ und $P_2 u_1 = 0$. Also gilt $u_1 \perp U_2$. Dies impliziert zusammen mit (3.5), dass ein $u_2 \in U_2$ existiert, mit $u_2 \perp U_1$. Das sieht man wie folgt ein: Sei $\{\phi_1, \ldots, \phi_n\}$ eine Orthonormalbasis von U_1 mit $\phi_1 = u_1$. Und sei $\{\psi_1, \ldots, \psi_n\}$ eine Basis von U_2. Das gesuchte u_2 muss dann

$$(\phi_i, u_2)_X = (\phi_i, \sum_{j=1}^n \alpha_j \psi_j)_X = \sum_{j=1}^n \alpha_j (\phi_i, \psi_j)_X = 0 \qquad \text{für } i = 1, \ldots, n$$

mit gewissen Koeffizienten $\alpha_j \in \mathbb{R}$, die nicht alle verschwinden. Da aber $u_1 = \phi_1$ und $u_1 \perp U_2$ gilt, reduziert sich das Gleichungssystem zu

$$\sum_{j=1}^n \alpha_j (\phi_i, \psi_j)_X = 0 \qquad \text{für } i = 2, \ldots, n.$$

Die Matrix $[(\phi_i, \psi_j)_X]_{\substack{i=2,\dots,n \\ j=1,\dots,n}}$ hat höchstens den Rang $n-1$. Damit existiert mindestens eine nichttriviale Lösung, diese leistet das Gewünschte. Daher ist auch $\sigma_2 = 0$ und $d_2 = 1$. Seien nun $\sigma_1, \sigma_2 > 0$, $u_1 \in S_1$ mit $\sigma_1 = \|P_2 u_1\|_X$, das bedeutet

$$\sigma_1^2 = \min_{u_1 \in U_1} \frac{\|P_2 u_1\|_X^2}{\|u_1\|_X^2} \qquad \text{und} \qquad u_1 \in \operatorname*{argmin}_{u_1 \in U_1} \frac{\|P_2 u_1\|_X^2}{\|u_1\|_X^2}.$$

Dann gilt

$$P_2 u_1 - \sigma_1^2 u_1 \perp U_1.$$

Denn per Definition minimiert u_1 den Rayleigh-Quotienten bezüglich des Operators P_2 auf U_1. Das heißt für jedes $\tilde{u}_1 \in U_1$ ist

$$\frac{\mathrm{d}}{\mathrm{d}t} \Lambda(u_1 + t\tilde{u}_1)\big|_{t=0} = 0.$$

Andererseits gilt

$$\frac{\mathrm{d}}{\mathrm{d}t} \Lambda(u_1 + t\tilde{u}_1)\big|_{t=0} = \frac{2}{\|u_1\|_X^2}(P_2 u_1 - \Lambda(u_1) u_1, \tilde{u}_1)_X.$$

Beachtet man, dass $\Lambda(u_1) = (u_1, P_2 u_1)_X = \sigma_1^2$ gilt, erhält man

$$\frac{2}{\|u_1\|_X}(P_2 u_1 - \sigma_1^2 u_1, \tilde{u}_1)_X = 0 \qquad \text{für alle } \tilde{u}_1 \in U_1,$$

das ist die behauptete Orthogonalität. Ferner ergibt sich durch die Orthogonalität für $u_2 = \sigma_1^{-1} P_2 u_1$, dass $P_1 u_2 = \sigma_1 u_1$. Somit ist $\|P_1 u_2\|_X = \sigma_1$, damit ist dann wegen $u_2 \in S_2$ aber auch $\sigma_2 \leq \sigma_1$. Die umgekehrte Ungleichung zeigt man analog und erhält $\sigma_1 = \sigma_2$ woraus die Behauptung, $d_1 = d_2$ folgt. □

3.13 Satz. *Seien $U_1, U_2 \subset X$ Unterräume mit $\dim(U_1) = \dim(U_2) < \infty$. Dann gilt*

$$\Theta(U_1; U_2) = d_i = \sqrt{1 - \sigma_i^2} \leq 1 \qquad \text{für } i = 1, 2.$$

Beweis. Es gilt

$$\begin{aligned} (P_2 - P_1)u &= (P_2 - P_1)(P_1^\perp + P_1)u = P_2 P_1^\perp u + P_2 P_1 u - P_1 u \qquad \text{und} \\ (P_2 - P_1)u &= (P_2^\perp + P_2)(P_2 - P_1)u = -P_2^\perp P_1 u - P_2 P_1 u + P_2 u. \end{aligned}$$

Addition der beiden Gleichungen und Auflösen nach der linken Seite liefert

$$(P_2 - P_1)u = P_2 P_1^\perp u - P_2^\perp P_1 u.$$

Mit dieser Identität folgt

$$\|(P_2 - P_1)u\|_X^2 = \left\|P_2 P_1^\perp u\right\|_X^2 - \underbrace{2(P_2 P_1^\perp u, P_2^\perp P_1 u)_X}_{=0} + \left\|P_2^\perp P_1 u\right\|_X^2 .$$

Benutzen wir nun, dass $P_1^2 = P_1$ und $(P_1^\perp)^2 = P_1^\perp$, sowie $\left\|P_2 P_1^\perp\right\| = \left\|P_1^\perp P_2\right\|$ -letzteres sieht man leicht ein, indem man die Bilder der beiden Abbildungen vergleicht und feststellt, dass diese gleich sind-, erhalten wir

$$\begin{aligned} \|(P_2 - P_1)u\|_X^2 &\leq \left\|P_2 P_1^\perp\right\|^2 \left\|P_1^\perp u\right\|_X^2 + \left\|P_2^\perp P_1\right\|^2 \|P_1 u\|_X^2 \\ &\leq d_2^2 \left\|P_1^\perp u\right\|_X^2 + d_1^2 \|P_1 u\|_X^2 . \end{aligned}$$

Ist nun $u \in S$, so erhalten wir mit $d_1 = d_2$, dass

$$d_2^2 \left\|P_1^\perp u\right\|_X^2 + d_1^2 \|P_1 u\|_X^2 = d_1^2 \left\|(P_1 + P_1^\perp)u\right\|_X^2 = d_1^2 = d_2^2.$$

Also insgesamt

$$\|(P_2 - P_1)u\|_X^2 \leq d_i^2 \qquad \text{für } i = 1, 2.$$

Das Bilden des Supremums und Ziehen der Wurzel liefern die Ungleichheit. Andererseits gilt

$$\begin{aligned} \|P_1 - P_2\| &= \max_{u \in S} \|(P_1 - P_2)u\|_X \\ &\geq \max_{u \in S_i} \|(P_1 - P_2)u\|_X = \max_{u \in S_i} \left\|P_j^\perp u\right\|_X = d_i \end{aligned}$$

für $i \neq j$ und $i = 1, 2$. Zu beachten ist hierbei, dass die orthogonalen Projektoren kompakt sind, sofern das Bild endlichdimensional ist. Dies bedeutet, dass das Maximum immer existiert. □

Bemerkung. Der Satz sagt aus, dass das Gap $\Theta(U_1; U_2)$ unter diesen Voraussetzungen immer genau der maximale Sinus des Winkels ist, den jeder Vektor mit seiner Projektion einschließt. Das Gap ist gleich 1, wenn es einen Vektor gibt, der orthogonal zum jeweils anderen Unterraum ist.

3.3 Fehlerabschätzungen

In diesem Abschnitt wird die Fehleranalysis für den Fehler geliefert, der bei der Diskretisierung des Eigenwertproblems kompakter Operatoren in Eigenwerten und Eigenfunktionen entsteht.

3.14 Lemma. *Seien $M \in L^+(X)$ und $K \in K(X)$ symmetrisch. Sei (u_i, λ_i) ein Eigenpaar des allgemeinen Eigenwertproblemes, das heißt $Ku_i = \lambda_i M u_i$. Dann gilt für beliebiges $v \in X$ mit $\|v\|_M = 1$*

$$(K(v-u_i),(v-u_i))_X = \Lambda(v) - \lambda_i + \lambda_i \|v-u_i\|_M^2 .$$

Beweis. Es gilt

$$\begin{aligned}(K(v-u_i),(v-u_i))_X &= (Kv,v)_X - 2(Ku_i,v)_X + (Ku_i,u_i)_X \\ &= \Lambda(v) - 2\lambda_i(Mu_i,v)_X + \lambda_i(Mu_i,u_i)_X \\ &= \Lambda(v) - \lambda_i + \lambda_i\|v\|_M^2 - 2\lambda_i(u_i,v)_M + \lambda_i\|u_i\|_M^2 \\ &= \Lambda(v) - \lambda_i + \lambda_i\|v-u_i\|_M^2 .\end{aligned}$$

□

3.15 Lemma. *Sei $U = \operatorname{span}\{u_{i_1},\dots,u_{i_k}\}$, wobei $(u_{i_1},\dots,u_{i_k})$ beliebige Eigenfunktionen von (3.1) seien. Sei $P : (X,(\cdot,\cdot)_M) \to U$ der zugehörige Orthoprojektor auf U. Dann gilt*

$$((I-P)u, KPu)_X = 0 \qquad \text{für alle } u \in X$$

und

$$(Pu, KPu)_X + ((I-P)u, K(I-P)u)_X = (u, Mu) \qquad \text{für alle } u \in X.$$

Beweis. Ohne Beschränkung der Allgemeinheit können wir annehmen, dass die Eigenfunktionen bezüglich $(\cdot,\cdot)_M$ orthonormiert sind. Dann gilt in der Eigenbasis in X

$$Pu = \sum_{j=1}^{k} c_{i_r} u_{i_r}.$$

Damit ist offensichtlich, dass

$$(u - Pu, u_{i_r})_M = 0 \qquad \text{für alle } r = 1,\dots,k$$

gilt. Daher erhalten wir

$$(K(u-Pu),u_{i_r})_X = (u-Pu, Ku_{i_r})_X = \lambda_{i_r}(u-Pu,u_{i_r})_M = 0$$

für beliebiges $r = 1,\dots,k$, woraus unter Ausnutzung der Linearität des Innenproduktes die Behauptung folgt. □

3.16 Lemma. *Seien $(u_{i_1}, \ldots, u_{i_k})$ beliebige Eigenfunktionen von (3.1). Sei $U = \operatorname{span}\{u_{i_1}, \ldots, u_{i_k}\}$ und $P : (X, (\cdot,\cdot)_M) \to U$ der zugehörige Orthoprojektor. Sei $\hat{U} = \operatorname{span}\{\hat{u}_{i_1}, \ldots, \hat{u}_{i_k}\}$ eine Approximation an U in dem Sinne, dass $\Theta := \Theta_M(U;\hat{U}) < 1$ gilt. Dann folgt*

$$\Lambda(P\hat{u}) - \lambda_{\inf} \leq \frac{\Lambda(\hat{u}) - \lambda_{\inf}}{1 - \Theta^2} \qquad \text{für alle } \hat{u} \in \hat{U}.$$

Hierbei ist $\lambda_{\inf} \leq \inf_{u \in X} \Lambda(u)$ und darf also insbesondere negativ sein.

Beweis. Wir betrachten den Operator $A := K - \lambda_{\inf} M$. Dieser ist nach Konstruktion positiv semidefinit. Das Eigenwertproblem (3.1) kann damit umformuliert werden zu

$$Au = (\lambda - \lambda_{\inf})Mu.$$

Die Anwendung des vorangegangenen Lemmas liefert

$$(\hat{u}, A\hat{u})_X = (P\hat{u}, AP\hat{u})_X + ((I-P)\hat{u}, A(I-P)\hat{u})_X.$$

Damit erhalten wir

$$\begin{aligned}
\Lambda(P\hat{u}) - \lambda_{\inf} &= \frac{(P\hat{u}, KP\hat{u})_X}{(P\hat{u}, P\hat{u})_M} - \lambda_{\inf}\frac{(P\hat{u}, P\hat{u})_M}{(P\hat{u}, P\hat{u})_M} = \frac{(P\hat{u}, AP\hat{u})_X}{(P\hat{u}, P\hat{u})_M} \\
&\leq \frac{(\hat{u}, A\hat{u})_X}{(P\hat{u}, P\hat{u})_M} = \frac{(\hat{u}, A\hat{u})_X}{(\hat{u}, \hat{u})_M} \cdot \frac{(\hat{u}, \hat{u})_M}{(P\hat{u}, P\hat{u})_M} \\
&\leq (\Lambda(\hat{u}) - \lambda_{\inf}) \left(\min_{\hat{u} \in \hat{U}} \frac{(P\hat{u}, P\hat{u})_M}{(\hat{u}, \hat{u})_M} \right)^{-1} \\
&= \frac{\Lambda(\hat{u}) - \lambda_{\inf}}{1 - \Theta^2}
\end{aligned}$$

□

Wir betrachten das Eigenwertproblem (3.1), also

$$Ku = \lambda Mu$$

und beweisen nun das zentrale Resultat für die Approximation der Eigenwerte.
Sei hierzu $X_n \subset X$ ein endlichdimensionaler Unterraum mit dem zugehörigen Orthoprojektor $\hat{P} : X \to X_n$ bezüglich des $(\cdot,\cdot)_M$-Innenprodukts. Wir setzen $U_i := \operatorname{span}\{u_1, \ldots, u_i\}$ als den Unterraum, der von den ersten i Eigenvektoren aufgespannt wird. Hierbei fordern wir, dass $i \leq p < \dim(X_n)$ gilt.

Dabei sei p so gewählt, dass $\dim(\hat{P}\{U_p\}) = p$ erfüllt ist. Dies garantiert, dass für $\hat{U}_i' := \hat{P}\{U_i\}$ die Gleichheit $\dim(\hat{U}_i') = \dim(U_i)$ gilt. Schließlich sei

$$\vartheta(U_i) := \sup_{u \in U_i, \|u\|_M = 1} \left\| \hat{P}^{\perp} u \right\|_M = \Theta_M(U_i, \hat{U}_i') < 1.$$

Diese Forderung bedeutet, dass kein Eigenvektor orthogonal zu X_n sein soll. Mit Hilfe von $\vartheta(U_i)$ lässt sich nun der Fehler in den Eigenwerten angeben. Für den folgenden Satz verwenden wir die Darstellung der Vektoren in der Eigenbasis

$$v = v_0 + \sum_{i=1}^{\infty} c_i u_i + \sum_{i=-1}^{-\infty} c_i u_i \quad \text{mit } v_0 \in \mathcal{N}(K) \qquad \text{für alle } v \in X \tag{3.6}$$

mit $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ und $\lambda_{-1} \leq \lambda_{-2} \leq \ldots < 0$.

3.17 Satz. *Sei $\hat{U}_p = \operatorname{span}\{\hat{u}_1, \ldots, \hat{u}_p\}$ der Eigenraum zu den Eigenwerten $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p > 0$ des projezierten Eigenwertproblems (3.4). Die Vektoren $\hat{u}_1, \ldots, \hat{u}_p$ seien orthonormal im $(\cdot,\cdot)_M$-Innenprodukt. Ferner sei $\dim(\hat{U}_p' = p)$. Dann gilt*

$$0 \leq \lambda_i - \hat{\lambda}_i \leq (\lambda_i - \lambda_{-1})\vartheta(U_i)^2 \qquad \text{für alle } i = 1, \ldots, p.$$

Im Falle eines positv semidefiniten Eigenwertproblems erhalten wir sogar die Abschätzung

$$0 \leq \lambda_i - \hat{\lambda}_i \leq \lambda_i \vartheta(U_i)^2 \qquad \text{für alle } i = 1, \ldots, p.$$

Beweis. Wir betrachten die Rayleigh-Ritz-Approximation von (3.1) in dem Unterraum $\hat{U}_i' = \hat{P}\{U_i\} \subset X_n$. Seien $\hat{u}_1', \ldots \hat{u}_i'$ die $(\cdot,\cdot)_M$-orthonormierten Eigenvektoren dieser Approximation in $\hat{U}_i'$ und $\hat{\lambda}_1', \ldots, \hat{\lambda}_i'$ die zugehörigen Eigenwerte für $i \leq p$. Nach dem Satz von Courant-Fisher (Satz 3.4) gilt dann

$$\Lambda(\hat{u}_i') =: \hat{\lambda}_i' \leq \hat{\lambda}_i.$$

Zusammen mit Satz 3.6 folgern wir

$$0 \leq \lambda_i - \hat{\lambda}_i \leq \lambda_i - \hat{\lambda}_i'.$$

Nach Konstruktion ist $\dim(U_i) = \dim(\hat{U}_i)$ und $\vartheta(U_i) < 1$ war vorausgesetzt, also können wir Lemma 3.12 und Satz 3.13 anwenden. Hieraus folgt $\Theta_M(U_i; \hat{U}_i) < 1$. Nun verwenden wir Lemma 3.16. Sei hierzu $P : X \to U_i$, der Orthoprojektor von X auf U_i. Dann gilt

$$\Lambda(\underbrace{P\hat{u}_i'}_{\in U_i}) - \lambda_{-1} \leq \frac{\hat{\lambda}_i' - \lambda_{-1}}{1 - \vartheta(U_i)^2}$$

und wegen $\Lambda(P\hat{u}_i') \geq \lambda_i$ folgt

$$\lambda_i - \lambda_{-1} \leq \Lambda(P\hat{u}_i') - \lambda_{-1} \leq \frac{\hat{\lambda}_i' - \lambda_{-1}}{1 - \vartheta(U_i)^2}.$$

Dieser Ausdruck lässt sich einfach umformen zu

$$\lambda_i - \hat{\lambda}_i' \leq (\lambda_i - \lambda_{-1})\vartheta(U_i)^2.$$

Insgesamt erhalten wir somit

$$0 \leq \lambda_i - \hat{\lambda}_i \leq \lambda_i - \hat{\lambda}_i' \leq (\lambda_i - \lambda_{-1})\vartheta(U_i)^2,$$

das ist die Behauptung.
Wesentlicher Bestandteil des Beweises war Lemma 3.16. Im Falle eines positiv semidefiniten Operators kann das dort auftretende $\lambda_{\text{inf}} = 0$ gesetzt werden und der Zusatz folgt direkt. □

Bemerkung. Der soeben bewiesene Satz besagt, dass ein Fehler der Größenordnung $\vartheta(U_i) < \varepsilon$ bei der Approximation der Eigenräume mit einem Fehler der Größenordnung ε^2 in der Approximation der Eigenwerte einhergeht (vgl. [8]).

Kapitel 4

Algebraische Eigenwertprobleme

4.1 Die Kondition des Eigenwertproblems

Der intuitive Ansatz zur Bestimmung der Eigenwerte einer Matrix $A \in \mathbb{R}^{n\times n}$ ist derjenige, die Nullstellen des charakteristischen Polynoms

$$p_A(\lambda) = (-1)^n \det(A - \lambda I) \stackrel{!}{=} 0$$

zu bestimmen. Dieses Problem ist schlecht konditioniert, kleine Störungen in den Koeffizienten führen zu großen Veränderungen in den Nullstellen (vgl. [10, 13, 14]). Um die Kondition des Eigenwertproblems zu berechnen, benötigen wir das folgende Lemma.

4.1 Lemma ([14, 26])**.** *Seien $A, E \in \mathbb{R}^{n\times n}$ nicht notwendig symmetrisch und λ_i ein einfacher Eigenwert von A mit*

$$Av_i = \lambda_i v_i \qquad \text{und} \qquad u_i^* A = \lambda_i u_i.$$

Dann besitzt $A + \varepsilon E$ für genügend kleines $|\varepsilon| > 0$ einen einfachen Eigenwert $\lambda(\varepsilon)$ und es gilt

$$\lambda(\varepsilon) = \lambda_i + \varepsilon \frac{u_i^* E v_i}{u_i^* v_i} + \mathcal{O}(\varepsilon^2) \qquad \text{für } \varepsilon \to 0.$$

Ferner ist $\lambda(\varepsilon)$ für hinreichend kleines $|\varepsilon| > 0$ differenzierbar.

Mit dem Lemma lässt sich nun der folgende Satz über die Kondition der Eigenwertberechnung zeigen.

4.2 Satz. *Sei $A \in \mathbb{R}^{n\times n}$ und λ_i ein einfacher Eigenwert mit*

$$Av_i = \lambda_i v_i \qquad \text{und} \qquad u_i^* A = \lambda_i u_i.$$

Dann ist die absolute Kondition beziehungsweise relative Kondition der Berechnung des Eigenwerts gegeben als

$$\kappa_{\text{abs}} = \frac{1}{|\cos(\angle(u_i, v_i))|} \quad \text{beziehungsweise} \quad \kappa_{\text{rel}} = \frac{\|A\|_2}{|\lambda_i \cos(\angle(u_i, v_i))|}.$$

Beweis. Nach dem vorigen Lemma 4.1 ist $\lambda(\varepsilon)$ differenzierbar und es gilt

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\lambda(\varepsilon)\big|_{\varepsilon=0} = \frac{u_i^* E v_i}{u_i^* v_i}.$$

Damit ergibt sich die Kondition zu

$$\kappa_{\text{abs}} = \sup_{E\in\mathbb{R}^{N\times N}:\|E\|_2=1} \frac{u_i^* E v_i}{u_i^* v_i} = \frac{\|u_i\|_2 \|v_i\|_2}{\|u_i\|_2 \|v_i\|_2 |\cos(\angle(u_i, v_i))|} = \frac{1}{|\cos(\angle(u_i, v_i))|}.$$

Hieraus folgt auch sofort die Behauptung für die relative Kondition. □

Für hermitesche Matrizen, also für den Fall, der in dieser Arbeit behandelt wird, stimmen Rechts- und Linkseigenvektor überein. Somit hängt hierbei die Kondition nur von dem Verhältnis des größten zu dem zu berechnenden Eigenwert ab, in Zeichen

$$\kappa_{\text{rel}} = \frac{\max\limits_{\lambda\in\sigma(A)} |\lambda|}{|\lambda_i|}.$$

Damit ist insbesondere die Berechnung betraglich kleiner Eigenwerte schlecht konditioniert. Es bezeichne P den Projektor auf den Eigenraum eines Eigenwerts λ. Die Resolventenfunktion hat Singularitäten an den Eigenwerten, man kann sie aber auf $\mathcal{N}(P)$ eingeschränkt betrachten.

4.3 Definition. Die Funktion

$$R_{\text{red}}(\lambda) := \Big[(A - \lambda I)\big|_{\mathcal{N}(P)}\Big]^{-1}$$

heißt *reduzierte Resolvente*. Alternativ kann diese Funktion wie folgt charakterisiert werden:

$$R_{\text{red}}(\lambda)(A-\lambda I)x = R_{\text{red}}(\lambda)(A-\lambda I)(I-P)x = (I-P)x \qquad \text{für alle } x \in \mathbb{R}^n.$$

Bemerkung. Dass diese Inverse wohldefiniert ist, sieht man wie folgt ein. Man betrachte die Abbildung $(A-\lambda I)$ eingeschränkt auf $\mathcal{N}(P)$. Sei $x \in \mathcal{N}(P)$ und $(A - \lambda I)x = 0$. Dann gilt insbesondere $x \in \mathcal{N}(A - \lambda I) \subset \mathcal{R}(P)$. Damit muss bereits $x = 0$ gelten. Folglich ist die Abbildung injektiv, also auch bijektiv.

Für die Kondition der Berechnung von Eigenvektoren gilt der folgende Satz.

4.4 Satz ([26])**.** *Sei $A \in \mathbb{R}^{n\times n}$ und λ ein Eigenwert von A. Die Kondition für die Berechnung eines zugehörigen Eigenvektors ist dann gegeben als*

$$\kappa_{\text{abs}} = \|R_{\text{red}}(\lambda)(I - P)\|_2 .$$

Ist A zusätzlich hermitesch, so gilt für die Kondition

$$\kappa_{\text{abs}} = \frac{1}{\text{dist}(\lambda, \sigma(A) \setminus \{\lambda\})}.$$

4.2 Numerische Methoden

Wie in Abschnitt 3.3 gezeigt wurde, hängt der Fehler, mit dem die Lösung des Eigenwertproblemes behaftet ist, lediglich von dem gewählten, endlichdimensionalen Teilraum ab. Wie gut die Lösung in diesem Raum approximierbar ist, ist letztlich die Aussage des Céa-Lemmas (vgl. [4]). In diesem Kapitel werden rein algebraische Probleme betrachtet, das heißt für $A, M \in \mathbb{R}^{n\times n}$ betrachten wir das Eigenwertproblem:

Finde $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$, $x \neq 0$, mit

$$Ax = \lambda Mx.$$

In diesem Abschnitt werden wir annehmen, dass die Eingabedaten A und M exakt vorliegen, ferner sei A immer symmetrisch und M symmetrisch und positiv definit, falls nicht anders gefordert.
Es bezeichne, sofern es nicht anders aus dem Kontext hervorgeht, $(\cdot, \cdot) = (\cdot, \cdot)_{\mathbb{R}^n}$ in diesem Kapitel immer das euklidische Skalarprodukt.
In dieser Arbeit habe ich mich besonderes mit zwei numerischen Verfahren zur Eigenpaarbestimmung auseinandergesetzt, der Unterraumiteration und dem Lanczos-Verfahren. Dabei habe ich die Unterraumiteration selbst implementiert. Für das Lanczos-Verfahren verwende ich die Softwarebibliothek ARPACK[1]. Ich habe mich für meine Diplomarbeit für eben diese beiden Methoden entschieden, da [26] folgend, die Unterraumiteration als Blockversion

[1] Online verfügbar unter `http://www.caam.rice.edu/software/ARPACK/`

der Potenzmethode einfach zu verstehen und zu implementieren sei und ARPACK als Stand der Technik im Bereich der Eigenwertberechnung gesehen werden kann. ARPACK wird beispielsweise als Eigenwertlöser[2] in der kommerziellen Numeriksoftware MATLAB[3] genutzt. In diesem Abschnitt werden zunächst Fehlerabschätzungen geliefert, um sinnvolle Abbruchskriterien für die iterativen Verfahren zu finden. Danach werden die einzelnen Verfahren besprochen. Begonnen wird mit der Unterraumiteration. Darauf folgend wird das QR-Verfahren vor- und der Zusammenhang zwischen beiden Verfahren hergestellt. Danach wird das Lanczos-Verfahren erklärt und schließlich der Zusammenhang zum QR-Verfahren und der Unterraumiteration hergestellt.

4.2.1 Fehlerabschätzungen

Wie zu Beginn des Kapitels erklärt, ist es wichtig a posteriori Abschätzungen für den Approximationsfehler, der durch iterative Verfahren entsteht, zu besitzen. In diesem Unterabschnitt werden hierzu einige Aussagen gezeigt.

4.5 Satz (Bauer-Fike)**.** *Sei $A \in \mathbb{R}^{n\times n}$ diagonalisierbar mit $A = VDV^{-1}$ und sei $(\tilde{\lambda}, \tilde{u})$ eine Approximation an ein Eigenpaar. Ferner sei $\|\tilde{u}\|_2 = 1$ und $r := A\tilde{u} - \tilde{\lambda}\tilde{u}$ das zugehörige Residuum. Dann existiert ein Eigenwert λ von A mit*

$$\left|\lambda - \tilde{\lambda}\right| \leq \operatorname{cond}_2(V) \|r\|_2 .$$

Beweis. Für $\lambda \in \sigma(A)$ ist die Behauptung trivial. Sei also $\lambda \notin \sigma(A)$. Dann ist $(A - \tilde{\lambda}I)$ invertierbar und es gilt

$$\tilde{u} = (A - \tilde{\lambda}I)^{-1} r = V(D - \tilde{\lambda}I)^{-1} V^{-1} r.$$

Damit folgt

$$\begin{aligned} 1 = \|\tilde{u}\|_2 &= \left\| V(D - \tilde{\lambda}I)V^{-1} r \right\|_2 \\ &\leq \|V\|_2 \left\|V^{-1}\right\|_2 \left\|(D - \tilde{\lambda}I)^{-1}\right\|_2 \|r\|_2 \\ &= \operatorname{cond}_2(V) \|r\|_2 \max_{\lambda_i \in \sigma(A)} \left|\lambda_i - \tilde{\lambda}\right|^{-1}, \end{aligned}$$

woraus sich die Behauptung ergibt. □

Sei A nun wieder als symmetrisch angenommen, da in diesem Fall V orthogonal ist, erhalten wir folgendes

[2] die MATLAB Funktion „eigs“ verwendet ARPACK

[3] Internetpräsenz unter `http://www.mathworks.de/`

4.6 Korollar. *Sei $(\tilde{\lambda}, \tilde{u})$ eine Approximation an ein Eigenpaar von A mit $\|\tilde{u}\|_2 = 1$. Dann existiert ein Eigenwert λ von A mit*

$$\left|\lambda - \tilde{\lambda}\right| \leq \|r\|_2 .$$

Für das zweite Resultat benötigen wir ein Lemma.

4.7 Lemma. *Sei $\tilde{u}$ normiert und Approximation an einen Eigenvektor von A. Ferner sei $\tilde{\lambda} = \Lambda(\tilde{u}) = (A\tilde{u}, \tilde{u})$ und (a, b) ein Intervall mit $\tilde{\lambda} \in (a, b)$ und $(a, b) \cap \sigma(A) = \emptyset$. Dann gilt*

$$(b - \tilde{\lambda})(\tilde{\lambda} - a) \leq \|r\|_2^2 .$$

Beweis. Es gilt

$$(r, \tilde{u}) = ((A - \tilde{\lambda}I)\tilde{u}, \tilde{u}) = \Lambda(\tilde{u}) - \tilde{\lambda}(\tilde{u}, \tilde{u}) = 0,$$

das bedeutet $\tilde{u} \perp r$. Damit ergibt sich

$$\begin{aligned} & ((A - aI)\tilde{u}, (A - bI)\tilde{u}) \\ = \; & ((A - \tilde{\lambda}I)\tilde{u} + (\tilde{\lambda} - a)\tilde{u}, (A - \tilde{\lambda}I)\tilde{u} + (\tilde{\lambda} - b)\tilde{u}) \\ = \; & \|r\|_2^2 + (\tilde{\lambda} - a)(\tilde{\lambda} - b). \end{aligned}$$

Sei $(u_i)_{i=1}^n$ die orthonormale Eigenbasis von A. Dann können wir $\tilde{u}$ in dieser Basis entwickeln und erhalten

$$\tilde{u} = \sum_{i=1}^{n} \alpha_i u_i$$

mit gewissen Koeffizienten $\alpha_i \in \mathbb{R}$. Daher ergibt sich für die linke Seite der Gleichung

$$((A - aI)\tilde{u}, (A - bI)\tilde{u}) = \sum_{i=1}^{N} |\alpha_i|^2 (\lambda_i - a)(\lambda_i - b) \geq 0$$

nach Voraussetzung. Daraus folgt direkt die Behauptung. □

4.8 Satz (Kato und Temple)**.** *Sei $\tilde{u}$ eine normierte Approximation an einen Eigenvektor von A. Ferner sei $\tilde{\lambda} = \Lambda(\tilde{u})$ und (α, β) ein Intervall mit $\tilde{\lambda} \in (\alpha, \beta)$, das zusätzlich genau einen Eigenwert λ von A enthält. Dann gilt*

$$-\frac{\|r\|_2^2}{\tilde{\lambda} - \alpha} \leq \tilde{\lambda} - \lambda \leq \frac{\|r\|_2^2}{\beta - \tilde{\lambda}}.$$

Beweis. Wir unterscheiden zwei Fälle. Sei zunächst $\lambda \leq \tilde{\lambda}$. Dann benutzen wir Lemma 4.7 mit $a = \lambda$ und $b = \beta$ und erhalten

$$0 \leq \tilde{\lambda} - \lambda \leq \frac{\|r\|_2^2}{\beta - \tilde{\lambda}}.$$

Ist umgekehrt $\lambda \geq \tilde{\lambda}$, so setzen wir $a = \alpha$ und $b = \lambda$. Lemma 4.7 liefert

$$0 \leq \lambda - \tilde{\lambda} \leq \frac{\|r\|_2^2}{\tilde{\lambda} - \alpha}.$$

Zusammen ergibt sich die Behauptung. □

Wählt man das Intervall (α, β) symmetrisch zu $\tilde{\lambda}$, ergibt sich die folgende Vereinfachung des Satzes.

4.9 Korollar. *Sei $\tilde{u}$ eine normierte Approximation an einen Eigenvektor von A. Ferner sei $\tilde{\lambda} = \Lambda(\tilde{u})$. Sei $\lambda \in \sigma(A)$ der Eigenwert mit minimalem Abstand zu $\tilde{\lambda}$ und sei $\delta = \min_i\{|\lambda_i - \tilde{\lambda}| \ : \ \lambda \neq \lambda_i\}$. Dann gilt*

$$\left|\tilde{\lambda} - \lambda\right| \leq \frac{\|r\|_2^2}{\delta}.$$

Beweis. Setzt man $\alpha = \tilde{\lambda} - \delta$ und $\beta = \tilde{\lambda} + \delta$, ergibt sich die Behauptung direkt aus Satz 4.8. □

Zum Schluss zeigen wir noch ein Resultat über den Fehler in den Eigenvektoren.

4.10 Satz. *Sei $\tilde{u}$ normiert und Approximation an einen Eigenvektor von $A \in \mathbb{R}^{n\times n}$. Ferner sei $\tilde{\lambda} = \Lambda(\tilde{u})$. Sei $\lambda \in \sigma(A)$ der Eigenwert mit minimalem Abstand zu $\tilde{\lambda}$ mit zugehörigem Eigenvektor u und sei $\delta = \min_i\{|\lambda_i - \tilde{\lambda}| \ : \ \lambda \neq \lambda_i\}$. Dann gilt*

$$\sin(\angle(\tilde{u}, u)) \leq \frac{\|r\|_2^2}{\delta}.$$

Beweis. Sei $\vartheta := \angle(\tilde{u}, u)$. Dann können wir $\tilde{u}$ darstellen als $\tilde{u} = u\cos(\vartheta) + z\sin(\vartheta)$ mit $z \perp u$. Es gilt

$$\begin{aligned} r = (A - \tilde{\lambda}I)\tilde{u} &= \cos(\vartheta)(A - \tilde{\lambda}I)u + \sin(\vartheta)(A - \tilde{\lambda}I)z \\ &= \cos(\vartheta)(\lambda - \tilde{\lambda})u + \sin(\vartheta)(A - \tilde{\lambda}I)z. \end{aligned}$$

Die beiden Vektoren auf der rechten Seite sind orthogonal, denn

$$(u, (A - \tilde{\lambda}I)z) = ((A - \tilde{\lambda}I)u, z) = (\lambda - \tilde{\lambda})(u, z) = 0.$$

Wir erhalten also

$$\|r\|_2^2 = \left\|(A - \tilde{\lambda}I)\tilde{u}\right\|_2^2 = \sin^2(\vartheta) \left\|(A - \tilde{\lambda}I)z\right\|_2^2 + \cos^2(\vartheta) \left|\lambda - \tilde{\lambda}\right|^2,$$

das heißt

$$\sin^2(\vartheta) \left\|(A - \tilde{\lambda}I)z\right\|_2^2 \leq \|r\|_2^2 .$$

Da $z \perp u$, gilt

$$\left\|(A - \tilde{\lambda}I)z\right\|_2 \geq \lambda_{\min}\big((A - \tilde{\lambda}I)\big|_{U^\perp}\big) = \delta$$

mit $U = \operatorname{span}\{u\}$. Damit folgt die Behauptung. □

Für hermitesche, beziehungsweise symmetrische Matrizen, wie wir sie betrachten, sind die gezeigten Abschätzungen scharf (vgl. [26]). Allerdings stellt sich die praktische Berechnung von δ als Problem heraus, denn hierbei handelt es sich um den zweitnächsten Eigenwert zu $\tilde{\lambda}$, der im Allgemeinen unbekannt ist (vgl. [26]). Man kann aber mit Hilfe des Satzes von Bauer-Fike eine verwendbare Abschätzung finden. Sei hierzu λ_j eben der Eigenwert, für den der Abstand zu $\tilde{\lambda}$ gleich δ wird. Sei $\tilde{\lambda}_j$ eine Approximation an λ_j mit zugehörigem Residuum r_j, dann gilt

$$\begin{aligned} \delta = \left|\tilde{\lambda} - \lambda_j\right| &\geq \left|(\tilde{\lambda} - \tilde{\lambda}_j) - (\lambda_j - \tilde{\lambda}_j)\right| \\ &\geq \left|\tilde{\lambda} - \tilde{\lambda}_j\right| - \left|\lambda_j - \tilde{\lambda}_j\right| \\ &\geq \left|\tilde{\lambda} - \tilde{\lambda}_j\right| - \|r_j\|_2 . \end{aligned}$$

Hierbei ist zu beachten, dass $\|r_j\|_2$ hinreichend klein ist, damit kein anderer Eigenwert näher an $\tilde{\lambda}$ ist als λ_j (vgl. [26]).

4.2.2 Projektionsverfahren

Die numerische Lösung großer Eigenwertprobleme erfolgt in der praktischen Anwendung nach dem folgenden Schema.
Sei $A \in \mathbb{R}^{n \times n}$ die Matrix, die durch die Diskretisierung eines symmetrischen Operators entsteht und sei U der Unterraum, der von den gesuchten Eigenvektoren von A aufgespannt wird.

Algorithmus (Projektionsverfahren zur Lösung von Eigenwertproblemen)**.**

1: Finde $\tilde{U}$ mit $\Theta(\tilde{U}; U) < 1$

2: Bestimme die orthogonale Projektion PAP des Operators A auf $\tilde{U}$
3: Bestimme alle Eigenwerte des projezierten Operators PAP

Der letzte Schritt erfolgt hierbei durch das QR-Verfahren, dem in der Praxis verwendeten Algorithmus, wenn man alle Eigenwerte einer Matrix benötigt (vgl. [13, 30, 31]). Das Lanczos-Verfahren und die Unterraumiteration unterscheiden sich nur in der Wahl des Unterraumes. Beim Lanczos-Verfahren verwendet man den Krylov-Raum, der von A und einem (zufälligen) Startvektor erzeugt wird. Bei der Unterraumiteration verwendet man nur die höchste Potenz des Operators A, die in dem Erzeuger des Krylov-Raums auftritt und wendet diese auf einen ganzen Unterraum X an. Beide Verfahren liefern einen Unterraum $\tilde{U}$, der den Eigenraum U approximiert.
Sofern nicht anders vorausgesetzt, betrachten wir in diesem Zusammenhang immer orthogonale Projektoren.

4.11 Satz. *Sei $U \subset \mathbb{R}^n$ ein Unterraum und sei $v_1, \ldots, v_m$ eine Orthonormalbasis von U. Ferner sei $V = [v_1, \ldots, v_m]$. Dann gilt für den Orthoprojektor $P: \mathbb{R}^n \longrightarrow U$ die Darstellung*

$$P = VV^\intercal.$$

Beweis. Sei $x \in \mathbb{R}^n$. Wir betrachten

$$VV^\intercal x = \sum_{i=1}^{m} v_i v_i^\intercal x = \sum_{i=1}^{m} (v_i, x) v_i =: y \in U.$$

Ergänzen wir $v_1, \ldots, v_m$ mit $\tilde{v}_{m+1}, \ldots, \tilde{v}_n$ zu einer Orthonormalbasis von $\mathbb{R}^n$, so besitzt x in dieser Basis die Darstellung

$$x = y + \sum_{m+1}^{n} (\tilde{v}_i, x) \tilde{v}_i$$

und schließlich

$$x - y = \sum_{m+1}^{n} (\tilde{v}_i, x) \tilde{v}_i \perp U.$$

Und daher

$$y = \operatorname*{argmin}_{z \in U} \|x - z\|_2 = Px.$$

□

Der Vektor $V^\intercal x$ ist genau der Koeffizientenvektor von Px bezüglich der Basis $v_1, \ldots, v_m$. Identifiziert man diese Basis mit der kanonischen Basis des $\mathbb{R}^m$, sind diese Räume isometrisch isomorph und wir können dann $V^\intercal x$ ebenfalls als Koeffizientenvektor bezüglich der kanonischen Basis des $\mathbb{R}^m$ auffassen. Hierdurch ergibt sich die Reduktion der Dimension des Problems von n auf m.

4.2.3 Das Gram-Schmidt-Verfahren

Zur Generierung der Orthonormalbasis des Unterraums verwenden wir das modifizierte Gram-Schmidt-Verfahren (MGS) mit dem Reorthogonalisierungskriterium aus [9].

Algorithmus. Das modifizierte Gram-Schmidt-Verfahren
input: Menge linear unabhängiger Vektoren $\{x_i\}_{i=1}^m \subset \mathbb{R}^n$, $L \in (0, \infty)$
output: Menge orthonormaler Vektoren $\{q_i\}_{i=1}^m \subset \mathbb{R}^n$

1: **for** $j := 1, \ldots, m$ **do**
2: **for** $k := 1, \ldots, j-1$ **do** ▷ Orthogonalisierung
3: $\beta_{k,j} := q_k^\intercal x_j$
4: $x_j := x_j - \beta_{k,j} q_k$
5: $L_j := \frac{1}{\|x_j\|_2} \sum_{k=1}^{j-1} |\beta_{k,j}|$
6: **if** $L_j > L$ **then**
7: **for** $k := 1, \ldots, j-1$ **do** ▷ Reorthogonalisierung
8: $\beta_{k,j} := q_k^\intercal x_j$
9: $x_j := x_j - \beta_{k,j} q_k$
10: $q_j := x_j / \|x_j\|_2$

Der Algorithmus liefert für $L < 1$ Vektoren q_i für $i = 1, \ldots, m$, die bis auf Maschinengenauigkeit orthogonal sind (vgl. [9]). Je kleiner L hierbei gewählt wird, desto mehr Reorthogonalisierungen werden durchgeführt. Für meine Implementierung habe ich mich an den Vorschlag aus [9] gehalten und immer $L = 0.99$ gewählt. In [9] wird gezeigt, dass nach der Reorthogonalisierung immer $L_j < 1$ gilt. Ist dies schon nach der Orthogonalisierung erfüllt, so erübrigt sich das Reorthogonalisieren.

4.2.4 Die Unterraumiteration

Der Unterraumiteration ist die Blockversion eines sehr einfachen Verfahrens, der Potenzmethode.

Algorithmus (Die Potenzmethode)**.**
input: $A \in \mathbb{R}^{n \times n}$ und $x_0 \in \mathbb{R}^n$
output: Folge von Iterierten $(x_i)_{i \geq 0}$

1: **for** $i := 0, 1, 2, \ldots$ **do**
2: $\tilde{x}_{i+1} = A x_i$
3: $x_{i+1} = \tilde{x}_{i+1} / \|\tilde{x}_{i+1}\|_2$

Die Normierung ist notwendig, um einen möglichen Underflow / Overflow zu vermeiden. Hierbei ist die Wahl der Norm beliebig (vgl. [14]). Um die

Konvergenzeigenschaften der Potenzmethode zu beweisen, sei $A \in \mathbb{R}^{n\times n}$ diagonalisierbar mit n Eigenwerten $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| \geq 0$. Die zugehörigen orthonormalen Eigenvektoren seien $u_1, \ldots, u_n$ und der Startvektor besitze in der Eigenbasis die Darstellung

$$x = \sum_{i=1}^{n} \alpha_i u_i.$$

Dann gilt der folgende Satz.

4.12 Satz. *Sei $\alpha_1 = x^\intercal u_1 \neq 0$ und $q := \frac{|\lambda_2|}{|\lambda_1|} < 1$, dann gilt:*

1. $$\|\tilde{x}_i\|_2 = |\lambda_1| + \mathcal{O}(q^i) \qquad \text{für } i \to \infty.$$

2. *Gilt $\lambda_1 > 0$, dann folgt*
$$\|x_i - \operatorname{sign}(\alpha_1) u_1\|_2 = \mathcal{O}(q^i) \qquad \text{für } i \to \infty.$$

3. *Gilt $\lambda_1 < 0$, dann folgt*
$$\left\|(-1)^{i+1} x_i - \operatorname{sign}(\alpha_1) u_1\right\|_2 = \mathcal{O}(q^i) \qquad \text{für } i \to \infty.$$

Beweis. Es gilt

$$A^i x = \sum_{k=1}^{n} \lambda_k^i u_k u_k^\intercal x = \sum_{k=1}^{n} \lambda_k^i u_k u_k^\intercal \sum_{l=1}^{n} \alpha_l u_l = \sum_{k=1}^{n} \lambda_k^i \alpha_k u_k.$$

Wir ziehen den ersten Summanden aus der Summe heraus und erhalten

$$A^i x = \lambda_1^i \alpha_1 \Bigg(u_1 + \underbrace{\sum_{k=2}^{n} \left[\frac{\lambda_k}{\lambda_1}\right]^i \frac{\alpha_k}{\alpha_1} u_k}_{=:r_i} \Bigg)$$

mit

$$\|r_i\|_2 = \left\| \sum_{k=2}^{n} \left[\frac{\lambda_k}{\lambda_1}\right]^i \frac{\alpha_k}{\alpha_1} u_k \right\|_2 \leq q^i \underbrace{\sum_{k=2}^{n} \left|\frac{\alpha_k}{\alpha_1}\right|}_{=:C} = C q^i.$$

Daher gilt

$$x_{i+1} = \frac{A^i x}{\|A^i x\|_2} = \operatorname{sign}(\lambda_1^i \alpha_1) \frac{u_1 + r_i}{\|u_1 + r_i\|_2}. \tag{4.1}$$

Mit

$$1 - Cq^i \leq \|u_1\|_2 - \|r_i\|_2 \leq \|u_1 + r_i\|_2 \leq \|u_1\|_2 + \|r_i\|_2 \leq 1 + Cq^i$$

folgt dann

$$x_i = \operatorname{sign}(\lambda_1^{i-1}\alpha_1)u_1 + \mathcal{O}(q^i) \qquad \text{für } i \to \infty,$$

woraus sich die drei Behauptungen ergeben. □

Bei der Potenzmethode konvergiert also $\|x_i\|_2 \to |\lambda_1|$ für $i \to \infty$ gegen den Spektralradius von A. Alternieren hierbei die Vorzeichen der Folge der x_i, so ist $\lambda_1 < 0$. Die Bedingung $x^\intercal u_1 \neq 0$ kann nicht a priori überprüft werden. In der Praxis wird aber aufgrund von Rundungsfehlern eine Komponente längs u_1 „eingeschleppt“ (vgl. [13]). Dass immer eine solche Komponente längs u_1 auftritt, wird zu einem Problem, wenn nicht nur der größte Eigenwert gesucht ist. Bei der Unterraumiteration ergibt sich die Notwendigkeit der regelmäßigen Orthogonalisierung des approximierenden Unterraumes, da dieser sonst zu einem eindimensionalen Unterraum in Richtung u_1 entarten würde.
Ist A, wie in unserem Fall, symmetrisch und verwendet man zur Normierung die $\|\cdot\|_2$-Norm, verdoppelt sich die Konvergenzordnung.

4.13 Satz. *Sei $A = A^\intercal$. Dann gilt*

$$|\lambda_1 - \Lambda(x_i)| = \mathcal{O}(q^{2i}) \qquad \text{für } i \to \infty.$$

Beweis. Es gilt $r_i \perp u_1$ für alle $i \geq 0$. Mit Gleichung (4.1) und

$$\gamma_i := \operatorname{sign}(\lambda_1^{i-1}\alpha_1)\,\|v_1 + r_{i-1}\|_2^{-1}$$

ergibt sich daher

$$(\lambda_1 I - A)x_i = \gamma_i(\lambda_1 I - A)r_{i-1} \perp u_1.$$

Ferner ist mit $x_i^\intercal = \gamma_i u_1^\intercal + \gamma_i r_{i-1}^\intercal$

$$|x_i^\intercal(\lambda_1 I - A)x_i| = \gamma_i^2 \left|r_{i-1}^\intercal(\lambda_1 I - A)r_{i-1}\right| \leq \underbrace{\|\lambda_1 I - A\|_2}_{\leq 2\|A\|_2} \|r_{i-1}\|_2^2 .$$

Andererseits gilt

$$\lambda_1 - x_i^\intercal \tilde{x}_{i+1} = \lambda_1 x_i^\intercal x_i - x_i^\intercal \tilde{x}_{i+1} = x_i^\intercal(\lambda_1 x_i - \tilde{x}_{i+1}) = x_i^\intercal(\lambda_1 I - A)x_i.$$

Zusammen ist also

$$|\lambda_1 - x_i^\intercal \tilde{x}_{i+1}| \leq 2\,\|A\|_2\,\|r_{i-1}\|_2^2 = \mathcal{O}(q^{2i}) \qquad \text{für } i \to \infty.$$

Verwendet man nun, dass $x_i^\intercal \tilde{x}_{i+1} = \Lambda(x_i)$ gilt, folgt die Behauptung. □

Eine direkte Verallgemeinerung der Potenzmethode ist die Unterraumiteration. Starten wir statt mit einem Startvektor $x_0 \in \mathbb{R}^n$ mit einem ganzen Unterraum $\operatorname{span}\{x_1, \ldots, x_m\}$ und $[x_1, \ldots, x_m] =: X_0 \in \mathbb{R}^{n \times m}$, gelangt man zu folgendem Algorithmus (vgl. [26]).

Algorithmus. Die Unterraumiteration
input: $A \in \mathbb{R}^{n \times n}$, $X_0 \in \mathbb{R}^{n \times m}$
output: Folge von Unterräumen $(X_i)_{i \geq 0}$
1: **for** $i := 0, 1, 2, \ldots$ **do**
2: $\quad \tilde{X}_{i+1} := AX_i$
3: $\quad \tilde{X}_{i+1} = QR$
4: $\quad X_{i+1} := Q$

Die Orthogonalisierung mittels QR-Zerlegung ist notwendig, da sonst im Laufe der Iteration, als Konsequenz aus Satz 4.12, die lineare Unabhängigkeit der Spalten von X_i verloren ginge. Die Orthogonalisierung wird durch ein Gram-Schmidt-Verfahren realisiert. Für die Unterraumiteration gilt der folgende

4.14 Satz ([26])**.** *Seien $\lambda_1, \ldots, \lambda_m$ mit $|\lambda_i| > |\lambda_{i+1}|$ die m dominanten Eigenwerte von A und seien $u_1, \ldots, u_m$ die zugehörigen Eigenvektoren. Ferner sei P_k der Orthoprojektor auf den Raum* $\operatorname{span}\{u_1, \ldots, u_k\}$*. Gilt dann*

$$\operatorname{rank}(P_k X_0) = k \qquad \text{für } k = 1, \ldots, m,$$

so konvergiert die k-te Spalte von X_i in Richtung u_k für alle $k = 1, \ldots, m$.

Ersetzen wir die zweite Zeile des Algorithmus durch $\tilde{X}_{i+1} = p(A)X_i$, wobei

$$p(A) = (A - \nu_1 I) \cdots (A - \nu_l I) \qquad \text{mit } \nu_1, \ldots, \nu_l \in \mathbb{R},\ l \in \mathbb{N}$$

gilt, so gelangen wir zur Unterraumiteration mit *Shifts*. Im symmetrischen Fall sind alle Eigenwerte reell, daher werden auch nur reelle Shifts verwendet. Im Allgemeinen verwendet man komplexe Zahlen als Shifts, dann wird aber darauf geachtet, dass diese in komplex-konjugierten Paaren $\nu_i, \overline{\nu}_i \in \mathbb{C}$ auftreten, damit die Berechnungen reell bleiben (vgl. [26, 30]).

4.15 Satz. *Sei p ein reelles Polynom vom Grad $l \in \mathbb{N}$. Sind $\{(\lambda_i, u_i)\}_{i=1}^n$ die Eigenpaare von A, so sind $\{(p(\lambda_i), u_i)\}_{i=1}^n$ die Eigenpaare von $p(A)$.*

Beweis. O.B.d.A. besitze p die Darstellung

$$p(x) = (x - \nu_1) \cdots (x - \nu_l) \qquad \text{mit } \nu_1, \ldots, \nu_l \in \mathbb{C}.$$

Wir zeigen die Behauptung mit vollständiger Induktion. Sei $l = 1$ und u_i ein Eigenvektor von A, dann gilt

$$p(A)u_i = (A - \nu_1 I)u_i = Au_i - \nu_1 u_i = (\lambda_i - \nu_1)u_i = p(\lambda_i)u_i.$$

Sei die Behauptung für $l - 1$ bereits gezeigt. Wir setzen $p_i(x) := (x - \nu_1) \cdots (x - \nu_i)$ für $1 \leq i \leq l$, dann gilt

$$\begin{aligned} p(A)u_i &= (A - \nu_1 I) \cdots (A - \nu_l I)u_i = p_{l-1}(A)(\lambda_i - \nu_l)u_i \\ &\overset{\text{Ind. Vor.}}{=} p_{l-1}(\lambda_i)(\lambda_i - \nu_l)u_i = p(\lambda_i)u_i. \end{aligned}$$

□

Für die Wahl der Shifts stellt man folgende Überlegung an. Wir zerlegen das Spektrum in m Eigenwerte, die gesucht sind, und $n - m$ andere Eigenwerte. Sortieren wir die Eigenwerte betraglich absteigend, also

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_m| > |\lambda_{m+1}| \geq \ldots \geq |\lambda_n|,$$

so sind, wie im Folgenden auch noch gezeigt wird, für die Konvergenz der Unterraumiteration die Quotienten $|\lambda_{m+1}/\lambda_i|$ für $i = 1, \ldots, m$ ausschlaggebend. Transformieren wir die Matrix A mit dem Polynom p, haben wir gezeigt, dass die Matrix $p(A)$ die Eigenwerte $p(\lambda_1), \ldots, p(\lambda_n)$ besitzt. Angenommen nach Transformation gilt

$$|p(\lambda_1)| \geq \ldots \geq |p(\lambda_n)|,$$

dann sind für die Konvergenz nunmehr die Quotienten $|p(\lambda_{m+1})/p(\lambda_i)|$ entscheidend. Ist es möglich, die Shifts $\nu_1, \ldots, \nu_l$ so zu wählen, dass

$$|p(\lambda_{m+1})/p(\lambda_i)| \ll 1$$

gilt, ergibt sich daraus eine schnelle Konvergenz des Verfahrens. Nach [26] ist es optimal für einfache Shifts, also $l = 1$ den Shift

$$\nu_1 = \frac{1}{2}(\lambda_{m+1} + \lambda_n)$$

zu wählen und so die Mitte des ungewollten Teils des Spektrums in den Ursprung des Koordinatensystems zu shiften.

Analog zur Rayleigh-Ritz-Approximation kompakter Operatoren kann man das Rayleigh-Ritz-Verfahren für algebraische Eigenwertprobleme verwenden.

Algorithmus (Das Rayleigh-Ritz-Verfahren)**.**
input: $A \in \mathbb{R}^{n \times n}$, Unterraum $X \subset \mathbb{R}^n$ mit $\dim(X) = m$
output: $\{\tilde{\lambda}_i, \tilde{u}_i\}_{i=1}^m$

1: Berechne eine Orthonormalbasis $(v_1)_{i=1}^m$ von X
2: $V := [v_1, \ldots, v_m]$
3: $A_m := V^\intercal A V$
4: Bestimme Eigenwerte $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ und Eigenvektoren $Y = [y_i, \ldots, y_m]$ von A_m
5: **for** $i := 1, \ldots, m$ **do**
6: $\quad \tilde{u}_i := V y_i$

Der Vorteil des Rayleigh-Ritz-Verfahrens wird offenbar, wenn $m \ll n$ gilt. In diesem Fall ist der Aufwand von $\mathcal{O}(m^3)$ für das Lösen des kleinen Eigenwertproblems wesentlich geringer als der von $\mathcal{O}(mn^2)$ für das Lösen des großen. Analog zur Rayleigh-Ritz-Approximation ist das Rayleigh-Ritz-Verfahren in dem Sinne optimal, dass das Residuum $A\tilde{u} - \tilde{\lambda}\tilde{u}$ orthogonal zum Unterraum X ist. Weiterhin gilt der folgende Satz.

4.16 Satz. *Falls X invariant unter A ist, so sind die Approximationen an die Eigenpaare exakt.*

Beweis. Ist $(\tilde{\lambda}, \tilde{u})$ die Approximation an ein Eigenpaar von A, so gilt

$$(A\tilde{u} - \tilde{\lambda}\tilde{u}, \tilde{v}) = 0 \qquad \text{für alle } \tilde{v} \in X.$$

Sei $P : \mathbb{R}^n \to X$ der Orthoprojektor auf X, dann folgt

$$P(A\tilde{u} - \tilde{\lambda}\tilde{u}) = 0.$$

Wegen $A\tilde{u} \in X$ ist dann $PA\tilde{u} = A\tilde{u}$ und die Gleichung wird zu

$$A\tilde{u} - \tilde{\lambda}\tilde{u} = 0.$$

Somit ist das Eigenpaar $(\tilde{\lambda}, \tilde{u})$ exakt. □

Die Idee des Rayleigh-Ritz-Verfahrens kann auf die Unterraumiteration angwendet werden. In diesem Fall dient dann die eigentliche Unterraumiteration dazu, den Unterraum für die Projektion zu generieren. Um die Anzahl der Orthogonalisierungs- und Projektionsschritte zu steuern, führen wir noch einen Parameter *it* ein.

Algorithmus. Die Unterraumiteration mit Projektion
input: $A \in \mathbb{R}^{n \times n}$, $X_0 \in \mathbb{R}^{n \times m}$
output: Folge von Unterräumen $(X_i)_{i \geq 0}$, Approximationen $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$
1: **for** $i := 0, 1, 2, \ldots$ **do**
2: $\quad \tilde{X}_{i+1} := A^{it} X_i$
3: $\quad \tilde{X}_{i+1} = QR$

4: $\tilde{X}_{i+1} := Q$
5: $A_m := \tilde{X}_{i+1}^{\mathsf{T}} A \tilde{X}_{i+1}$
6: Bestimme Eigenwerte $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ und Eigenvektoren $Y = [y_i, \ldots, y_m]$ von A_m
7: $X_{i+1} := \tilde{X}_{i+1} Y$
8: Bestimme ein neues *it*

Der Parameter *it* ist hierbei so zu wählen, dass die Vektoren aus $\tilde{X}_{i+1}$ noch linear unabhängig (in endlicher Genauigkeit) bleiben.
Sei S_i der Unterraum, der von den Spalten von X_i aufgespannt wird und sei P_i der Orthoprojektor auf S_i. Die Eigenwerte von A seien betraglich geordnet und es gelte

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_m| > |\lambda_{m+1}| \geq \ldots \geq |\lambda_n| .$$

Ferner sei P der Orthoprojektor auf den zu $\lambda_1, \ldots, \lambda_m$ gehörenden Eigenraum. Dann können wir den folgenden Satz zeigen.

4.17 Satz. *Sei $S_0 = \operatorname{span}\{x_1, \ldots, x_m\}$, wobei $[x_1, \ldots, x_m] = X_0$ gelte. Ferner seien $(Px_k)_{k=1}^m$ linear unabhängig. Dann existiert jeweils genau ein $s_k \in S_0$, so dass $Ps_k = u_k$ für die Eigenvektoren $u_1, \ldots, u_m$ von A gilt. Außerdem gilt für alle $k = 1, \ldots, m$ die folgende Ungleichung*

$$\left\| P_i^{\perp} u_k \right\|_2 \leq \|u_k - s_k\|_2 \left| \frac{\lambda_{m+1}}{\lambda_k} \right|^i .$$

Beweis. Nach Voraussetzung bilden die Vektoren $(Px_j)_{j=1}^m$ eine Basis von $P\mathbb{R}^n$. Daher können die Eigenvektoren dargestellt werden als

$$u_k = \sum_{j=1}^{m} \alpha_j P x_j = P \sum_{j=1}^{m} \alpha_j x_j = P s_k.$$

Wir zerlegen nun

$$s_k = P s_k + P^{\perp} s_k = u_k + w.$$

Dann gilt mit $y = \lambda_k^{-i} A^i s_k$ die Gleichheit

$$y - u_k = \lambda_k^{-i} A^i w.$$

Sei $W = P^{\perp}\mathbb{R}^n$. Dann ist nach dem Spektralsatz W auch ein invarianter Unterraum, es gilt nämlich $W = \operatorname{span}\{u_{m+1}, \ldots, u_n\}$. Mit $w \in W$ folgt dann

$$y - u_k = \lambda_k^{-i} \left[A\big|_W \right]^i w.$$

Also ist

$$\|y - u_k\|_2 \leq \left\| \lambda_k^{-i} \left[A\big|_W \right]^i \right\|_2 \|w\|_2 .$$

Der betraglich größte Eigenwert von $A\big|_W$ ist λ_{m+1}. Demnach ist der Spektralradius von $\lambda_k^{-1} A\big|_W$ gegeben durch $|\lambda_{m+1}/\lambda_k|$. Dies bedeutet, per Definition des Spektralradius

$$\|y - u_k\|_2 \leq \left| \frac{\lambda_{m+1}}{\lambda_k} \right|^i \|w\|_2 .$$

Benutzen wir, dass

$$\left\| P_i^{\perp} u_k \right\|_2 = \min_{y \in S_i} \|y - u_k\|_2$$

gilt, folgt die Behauptung. □

Bemerkung. Die Forderung der linearen Unabhängigkeit von $Px_1, \ldots, Px_m$ ist äquivalent zu der Bedingung, dass

$$\det(U^{\intercal} X_0) \neq 0$$

mit $U := [u_1, \ldots, u_m]$ gilt. Dies ist aber nichts anderes als eine Verallgemeinerung der Bedingung an den Startvektor aus der Potenzmethode (vgl. [26]).

Der obige Satz liefert, dass die Eigenvektoren unterschiedlich schnell konvergieren. Unter praktischen Gesichtspunkten ist es daher sinnvoll, bereits konvergierte Eigenvektoren nicht weiter mit der Matrix A zu multiplizieren. Dennoch muss eine *Deflation* (vgl. [26]) mit diesen Vektoren durchgeführt werden. Dies geschieht in der Praxis dadurch, dass die noch nicht konvergierten Vektoren gegen die bereits konvergierten orthogonalisiert werden, dies bezeichnet man als *explizite Deflation* (vgl. [10]). Ergänzen wir zusätzlich noch die Verwendung von Shifts, gelangen wir zu folgendem Algorithmus, der in dieser Form für die numerischen Beispiele in Abschnitt 4.3 implementiert wurde.

Algorithmus (Die Unterraumiteration mit Projektion, Deflation, Shifts)**.**
input: $A \in \mathbb{R}^{n \times n}$, $X_0 \in \mathbb{R}^{n \times m}$, $nev \leq m$
output: $\tilde{\lambda}_i$, $\tilde{u}_i$ für $i = 1, \ldots, nev$
1: $i := 1$
2: **while** $i \leq nev$ **do**
3: Wähle Shifts $\nu_1, \ldots, \nu_l$
4: $\tilde{Z} := [q_1, \ldots, q_{i-1}, p(A)X]$
5: Orthonormalisiere $\tilde{Z}$ beginnend bei Spalte i
Speichere das Resultat in Z

6: $A_m := Z^\intercal AZ$
7: Bestimme Eigenwerte $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ und Eigenvektoren $Y = [y_i, \ldots, y_m]$ von A_m
8: Überprüfe $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ auf Konvergenz
9: Füge die i_{conv} neu konvergierten Eigenwerte zu $Q = [q_1, \ldots, q_{i-1}]$ hinzu
10: $i := i + i_{conv}$
11: $X := Z[y_i, \ldots, y_m]$

4.2.5 Krylov-Räume

Das Lanczos-Verfahren verwendet einen *Krylov-Raum* zur Projektion. Daher werden diese hier eingeführt.

4.18 Definition. Sei $A \in \mathbb{R}^{n\times n}$ eine Matrix und $v \in \mathbb{R}^n$ ein Vektor, dann ist für $m \in \mathbb{N}$ der *Krylov-Raum* als

$$\mathcal{K}_m(A, v) := \operatorname{span}\{v, Av, A^2v, \ldots, A^{m-1}v\}$$

definiert.

Wenn Mehrdeutigkeiten ausgeschlossen sind, schreiben wir auch $\mathcal{K}_m$ statt $\mathcal{K}_m(A, v)$. Im Zusammenhang mit Krylov-Räumen ist die Definition des *Minimalpolynoms* eines Vektors von Bedeutung.

4.19 Definition. Sei $A \in \mathbb{R}^{n\times n}$ und $v \in \mathbb{R}^n$. Dann heißt das reelle monische Polynom p mit kleinstem Grad m, das ungleich dem Nullpolynom ist und

$$p(A)v = 0$$

erfüllt, das *Minimalpolynom* von v. In diesem Zusammenhang bezeichnet man m als den *Grad* von v bezüglich A. Offensichtlich kann der Grad eines Vektors nie größer als n sein.

Der nächste Satz ist eine sehr intuitive Charakterisierung von Krylov-Räumen. Der Beweis folgt sofort aus der Definition.

4.20 Satz. *Der Krylov-Raum $\mathcal{K}_m$ ist die Menge aller $x \in \mathbb{R}^n$, die als*

$$x = p(A)v,$$

mit einem reellen Polynom p vom Grad kleiner oder gleich $m-1$ dargestellt werden können.

4.21 Satz. *Sei μ der Grad des Minimalpolynoms von v. Dann ist $\mathcal{K}_\mu$ ein invarianter Unterraum von A und $\mathcal{K}_\mu = \mathcal{K}_m$ für jedes $m \geq \mu$.*

Beweis. Zu zeigen ist, dass $A\mathcal{K}_\mu \subset \mathcal{K}_\mu$ gilt. Darum sei $x \in \mathcal{K}_\mu$ beliebig aber fest. Dann gilt

$$x = \sum_{i=0}^{\mu-1} \alpha_i A^i v \qquad \text{mit } \alpha_0, \ldots, \alpha_{\mu-1} \in \mathbb{R}.$$

Daraus folgt

$$Ax = A \sum_{i=0}^{\mu-1} \alpha_i A^i v = \sum_{i=1}^{\mu} \alpha_{i-1} A^i v.$$

Sei $p(x) = \sum_{i=0}^{\mu} \gamma_i x^i$ das Minimalpolynom von v. Durch Umskalierung kann erreicht werden, dass $\gamma_\mu = \alpha_{\mu-1}$ gilt. Nutzen wir aus, dass $p(A)v = 0$, ergibt sich

$$\begin{aligned} Ax &= \sum_{i=1}^{\mu} \alpha_{i-1} A^i v - p(A)v = \sum_{i=1}^{\mu} \alpha_{i-1} A^i v - \sum_{i=0}^{\mu} \gamma_i A^i v \\ &= -\gamma_0 v + \sum_{i=1}^{\mu} (\alpha_{i-1} - \gamma_i) A^i v = -\gamma_0 v + \sum_{i=1}^{\mu-1} (\alpha_{i-1} - \gamma_i) A^i v. \end{aligned}$$

Das bedeutet $Ax \in \operatorname{span}\{v, Av, \ldots, A^{\mu-1}v\} = \mathcal{K}_\mu$. Insbesondere gilt dann auch $A^\mu v \in \operatorname{span}\{v, Av, \ldots, A^{\mu-1}v\} = \mathcal{K}_\mu$, womit auch der zweite Teil des Satzes gezeigt ist. □

4.22 Satz. *Es gilt $\dim(\mathcal{K}_m(A, v)) = m$ genau dann, wenn das Minimalpolynom von v bezüglich A mindestens den Grad m hat.*

Beweis. Sei $\dim(\mathcal{K}_m(A, v)) = m$. Dann bilden die Vektoren $v, Av, \ldots, A^{m-1}v$ eine Basis von $\mathcal{K}_m$. Also folgt für jedes beliebige Polynom p vom Grad $m-1$ aus

$$p(A)v = \sum_{i=0}^{m-1} \alpha_i A^i v = 0$$

bereits, dass $\alpha_0 = \alpha_1 = \ldots = \alpha_{m-1}$ gelten muss. Damit ist p aber das Nullpolynom. Also muss das Minimalpolynom von v einen Grad größer oder gleich m haben.
Ist umgekehrt der Grad des Minimalpolynoms von v größer oder gleich $m-1$, bedeutet das, es existieren keine Koeffizienten $\alpha_0, \ldots, \alpha_{m-1}$, die nicht alle verschwinden, sodass

$$\sum_{i=0}^{m-1} \alpha_i A^i v = 0$$

gilt. Damit sind die m Vektoren $v, Av, \ldots, A^{m-1}v$ linear unabhängig. Folglich ist die Dimension von $\mathcal{K}_m$ gleich m. □

4.23 Satz. *Sei $P_m : \mathbb{R}^n \to \mathcal{K}_m$ ein nicht notwendig orthogonaler Projektor auf $\mathcal{K}_m$ und sei $A_m = P_m A\big|_{\mathcal{K}_m}$. Dann gilt für jedes reelle Polynom p vom Grad kleiner oder gleich $m-1$, dass*

$$p(A)v = p(A_m)v.$$

Ferner gilt für jedes reelle Polynom p vom Grad kleiner oder gleich m, dass

$$P_m p(A)v = p(A_m)v.$$

Beweis. Es genügt die Behauptung für die Monome $p_i(x) = x^i$ für $i = 0, \ldots, m-1$ zu zeigen. Der Beweis erfolgt mit vollständiger Induktion. Die Behauptung ist offensichtlich für p_0 erfüllt. Ist die Behauptung für $i > 0$ erfüllt, so ergibt sich

$$p_{i+1}(A)v = Ap_i(A)v \overset{\text{Ind. Vor.}}{=} Ap_i(A_m)v.$$

Für $i+1 \leq m-1$ haben wir $p_{i+1}(A)v \in \mathcal{K}_m$ und somit

$$p_{i+1}(A)v = P_m p_{i+1}(A)v = P_m A p_i(A_m)v.$$

Andererseits ist $p_i(A_m)v \in \mathcal{K}_m$ und daher

$$P_m A p_i(A_m)v = P_m A\big|_{\mathcal{K}_m} p_i(A_m)v = p_{i+1}(A_m)v.$$

Damit ist die Behauptung für den Fall $i+1 \leq m-1$ gezeigt. Für den Fall $i+1 = m$ bleibt $P_m p(A)v = p(A_m)v$ zu zeigen. Dies folgt aber sofort, wenn man beide Seiten von $p_{m-1}(A)v = p_{m-1}(A_m)v$ mit $P_m A$ multipliziert. □

Für das folgende Theorem benötigen wir eine Definiton.

4.24 Definition. Sei $v_1, \ldots, v_m$ eine Orthonormalbasis des Krylov-Raums $\mathcal{K}_m(A, v)$. Ferner sei $V_m = [v_1, \ldots, v_m]$. Dann definieren wir für $A_m = V_m^\intercal A V^m$ das charakteristische Polynom

$$\overline{p}_m(\lambda) := \det(A_m - \lambda I).$$

Bemerkung. Das charakteristische Polynom $\overline{p}_m(\lambda)$ ist wohldefiniert, das heißt, es ist unabhängig von der Wahl der Orthonormalbasis von $\mathcal{K}_m$ eindeutig bestimmt.

Beweis. Sei $\tilde{v}_1, \ldots, \tilde{v}_m$ eine weitere Orthonormalbasis von $\mathcal{K}_m$ und sei $\tilde{V}_m = [\tilde{v}_1, \ldots, \tilde{v}_m]$. $\mathcal{K}_m$ ist als m-dimensionaler $\mathbb{R}$-Vektorraum isometrisch isomorph zu $\mathbb{R}^m$. Somit existiert eine orthogonale Matrix Q_m, die die Basistransformation von V_m nach $\tilde{V}_m$ bezüglich der Koeffizientenvektoren in $\mathbb{R}^m$ beschreibt, das bedeutet

$$\tilde{V}_m = V_m Q_m \qquad \text{und} \qquad V_m = \tilde{V}_m Q_m^\intercal.$$

Daher gilt

$$\begin{aligned} \overline{p}_m(\lambda) &= \det(A_m - \lambda I) = \det(V_m^\intercal A V_m - \lambda I) \\ &= \det(Q_m^\intercal V_m^\intercal A V_m Q_m - \lambda Q_m^\intercal Q_m) = \det(\tilde{V}_m^\intercal A \tilde{V}_m - \lambda I), \end{aligned}$$

womit die Wohldefiniertheit gezeigt ist. □

4.25 Theorem. *Sei $\overline{p}_m(\lambda)$ das charakteristische Polynom, resultierend aus der orthogonalen Projektion von A auf $\mathcal{K}_m(A, v)$. Dann minimiert $\overline{p}_m(\lambda)$ das Funktional*

$$J(p) := \|p(A)v\|_2$$

über alle monischen Polynome vom Grad m.

Beweis. Sei $P : \mathbb{R}^n \to \mathcal{K}_m$ der Orthoprojektor auf $\mathcal{K}_m$ und sei $A_m = PAP$. Nach dem Satz von Cayley-Hamilton (vgl. [3]) gilt $\overline{p}_m(A_m) = 0$ und damit

$$(\overline{p}_m(A_m)v, w) = 0 \qquad \text{für alle } w \in \mathcal{K}_m.$$

Nach Satz 4.23 gilt $\overline{p}_m(A_m) = P\overline{p}_m(A)$ und daher

$$(P\overline{p}_m(A)v, w) = 0 \qquad \text{für alle } w \in \mathcal{K}_m.$$

Da orthogonale Projektoren selbstadjungiert sind, folgt weiter

$$(\overline{p}_m(A)v, w) = (\overline{p}_m(A)v, Pw) = (P\overline{p}_m(A)v, w) = 0 \qquad \text{für alle } w \in \mathcal{K}_m.$$

Dies ist äquivalent zu

$$(\overline{p}_m(A)v, A^i v) = 0 \qquad \text{für } i = 0, \ldots, m-1.$$

Setzen wir nun $\overline{p}_m(x) = x^m - q(x)$ wobei q ein Polynom vom Grad $m-1$ ist, erhalten wir das Gleichungssystem

$$(A^m v - q(A)v, A^j v) = 0 \qquad \text{für } i = 0, \ldots, m-1.$$

Dieses entspricht genau den Gauß'schen Normalengleichungen des Minimierungsproblems

$$\min_{s \in \Pi_{m-1}} \|A^m v - s(A)v\|_2 .$$

Hierbei bezeichnet Π_{m-1} den Raum aller reellen Polynome vom Grad kleiner oder gleich $m-1$. Damit ist die Behauptung gezeigt. □

Im Zusammenhang mit dem QR-Verfahren sind noch die folgenden Sätze von Bedeutung.

4.26 Definition. Die Matrix $H \in \mathbb{R}^{n\times n}$ heißt *obere Hessenberg-Matrix* genau dann, wenn $h_{i,j} = 0$ für $i > j+1$ gilt. Wir nennen H *echte obere Hessenberg-Matrix*, wenn zusätzlich $h_{i,j} \neq 0$ für $i = j+1$ gilt. Falls $H^\intercal = H$ ist H offensichtlich tridiagonal. Gilt dann $h_{i,j} \neq 0$ für $i = j+1$, nennen wir H *echte Tridiagonalmatrix*.

Bemerkung. Im symmetrischen Fall sind die oberen Hessenberg-Matrizen genau die Tridiagonalmatrizen. Auf diese werden wir uns im Folgenden beschränken. Die Annahme einer echten oberen Hessenberg-Matrix bedeutet nichts weiter, als dass das Eigenwertproblem nicht entkoppelt und in zwei kleinere Eigenwertprobleme zerfällt. Sei

$$H = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix},$$

so partitioniert, dass $H_{2,1}$ die Nullmatrix ist, dann sind $H_{1,1}$ und $H_{2,2}$ wieder obere Hessenberg-Matrizen und das Eigenwertproblem für H zerfällt in die Eigenwertprobleme für $H_{1,1}$ und $H_{2,2}$ (vgl. [26, 31]).

4.27 Satz. *Sei $H \in \mathbb{R}^{n\times n}$ eine echte obere Hessenberg-Matrix. Dann gilt*

$$\mathcal{K}_m(H, e_1) = \operatorname{span}\{e_1, \ldots, e_m\} \qquad \text{für } m = 1, \ldots, n.$$

Beweis. Der Beweis erfolgt per vollständiger Induktion. Für $m = 1$ ist die Behauptung klar. Sei die Behauptung für $m < n$ erfüllt. Dann gilt

$$\begin{aligned} \mathcal{K}_{m+1}(H, e_1) &= \operatorname{span}\{e_1, He_1, \ldots, H^m e_1\} \\ &\overset{\text{Ind. Vor.}}{=} \operatorname{span}\{e_1, e_2, \ldots, e_{m-1}, H^m e_1\}. \end{aligned}$$

Für den Vektor $H^m e_1$ ergibt sich

$$\begin{aligned} H^m e_1 &= H^{m-1}[h_{1,1}, h_{2,1}, 0, \ldots, 0]^\intercal \\ &= H^{m-2}[h_{1,1}^2 + h_{1,1}h_{1,2}, h_{2,1}^2 + h_{2,1}h_{2,2}, h_{2,1}h_{3,2}, 0, \ldots, 0]^\intercal \\ &\vdots \\ &= [\tilde{h}_1, \ldots, \tilde{h}_m, 0, \ldots, 0]. \end{aligned}$$

Wegen der echten oberen Hessenberg-Form von H gilt nun

$$\tilde{h}_m = h_{2,1}h_{3,2}\cdots h_{m+1,m} \neq 0.$$

Orthonormalisieren von $H^m e_1$ gegen $e_1, \ldots, e_{m-1}$ liefert dann die Behauptung. □

4.28 Satz. *Gilt $x = p(A)e_1$ für ein beliebiges reelles Polynom, dann folgt*

$$p(A)\mathcal{K}_m(A, e_1) = \mathcal{K}_m(A, x) \qquad \text{für } m = 1, \ldots, n.$$

Beweis. Sei $y \in p(A)\mathcal{K}_m(A, e_1)$. Dann besitzt y die Darstellung

$$y = \sum_{i=0}^{m-1} \alpha_i p(A) A^i e_1,$$

mit $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{R}$. Offensichtlich kommutiert $p(A)$ mit A^i für alle $i = 0, \ldots, m-1$. Daher folgt

$$y = \sum_{i=0}^{m-1} \alpha_i p(A) A^i e_1 = \sum_{i=0}^{m-1} \alpha_i A^i p(A) e_1 = \underbrace{\sum_{i=0}^{m-1} \alpha_i A^i x}_{\in \mathcal{K}_m(A,x)}.$$

Die umgekehrte Inklusion folgt analog. □

Zu jeder Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$, so dass $Q^\intercal AQ$ obere Hessenberg-Form hat (vgl. [11]). Für den folgenden Satz benötigen wir diese Aussage.

4.29 Satz. *Sei $H = Q^\intercal AQ$ eine echte obere Hessenberg-Matrix und seien $q_1, \ldots, q_n$ die Spalten von Q. Dann gilt*

$$\operatorname{span}\{q_1, \ldots, q_m\} = \mathcal{K}_m(A, q_1) \qquad \text{für } m = 1, \ldots, n.$$

Beweis. Der Beweis erfolgt wieder per vollständiger Induktion über m. Für $m = 1$ ist die Behauptung klar. Sei die Behauptung für m erfüllt. Aus der Gleichung $AQ = QH$ erhält man dann

$$Aq_m = \sum_{i=1}^{m+1} q_i h_{i,j}.$$

Durch Umstellen der Gleichung ergibt sich

$$q_{m+1} b_{m+1,m} = Aq_m - \sum_{i=1}^{m} q_i h_{i,j}. \tag{4.2}$$

Nach Induktionsvoraussetzung gilt

$$Aq_m \in A\mathcal{K}_m(A, q_1) \subseteq \mathcal{K}_{m+1}(A, q_1).$$

Daraus folgt mit Gleichung (4.2) und $h_{m+1,m} \neq 0$, wegen der echten Hessenberg-Form von H, dass

$$q_{m+1} \in \mathcal{K}_{m+1}(A, q_1)$$

gilt. Damit ist die Inklusion $\operatorname{span}\{q_1, \ldots, q_{m+1}\} \subseteq \mathcal{K}_{m+1}(A, q_1)$ gezeigt. Umgekehrt gilt $\dim(\operatorname{span}\{q_1, \ldots, q_{m+1}\}) = m+1$ und $\dim(\mathcal{K}_{m+1}(A, q_1)) \leq m+1$, woraus auch die umgekehrte Inklusion folgt. □

4.2.6 Die Rayleigh-Quotient-Iteration

Ein effizientes Verfahren zur Bestimmung der Eigenvektoren, wenn Näherungen an die Eigenwerte bekannt sind, ist die Rayleigh-Quotient-Iteration (RQI). Insbesondere lassen sich mit diesem Verfahren auch Eigenwerte im Inneren des Spektrums bestimmen (vgl. [13, 25]).

Algorithmus. Die Rayleigh-Quotient-Iteration
input: $A \in \mathbb{R}^{n\times n}$ und (μ_0, z_0) mit $\|z_0\|_2 = 1$
output: Folge von Iterierten $\{(\mu_i, z_i)\}_{i>0}$
1: **for** $i := 0, 1, 2 \ldots$ **do**
2: $\quad \mu_{i+1} := \Lambda(z_i)$
3: $\quad \tilde{z}_{i+1} := (\mu_{i+1} I - A)^{-1} z_i$
4: $\quad z_{i+1} = \tilde{z}_{i+1} / \|\tilde{z}_{i+1}\|_2$

Man bricht die Iteration ab, wenn $\|\tilde{z}_{i+1}\|_2$ sehr groß wird. In diesem Fall ist das Gleichungssystem fast singulär, man hat also eine gute Approximation an ein Eigenpaar gefunden. Man kann zeigen, dass falls A symmetrisch ist, die Konvergenz dieses Verfahrens lokal kubisch ist.

4.30 Satz ([14]). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und (μ_0, z_0) eine hinreichend gute Approximation an ein Eigenpaar (λ, u) von A mit $\|z_0\|_2 = \|u\|_2 = 1$. Dann existiert ein $C > 0$, sodass für die Iterierten der RQI*

$$|\lambda - \mu_{i+1}| \leq C\,|\lambda - \mu_i|^3 \qquad \text{für } i = 0, 1, 2, \ldots$$

gilt.

4.2.7 Das QR-Verfahren

Seit 50 Jahren ist das QR-Verfahren der Standard-Algorithmus, wenn es darum geht, die Eigenwerte und Eigenvektoren kleiner Matrizen zu berechnen (vgl. [30]). Symmetrische Matrizen lassen sich immer mittels orthogonaler Matrizen tridiagonalisieren. Der Aufwand hierfür ist $\mathcal{O}(n^3)$, wenn man die Anzahl der benötigten Multiplikationen zu Grunde legt. Die Eigenwerte sind unter dieser Ähnlichkeitstransformation invariant.

4.31 Satz ([11]). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Dann existiert eine Matrix $Q \in \mathbb{R}^{n\times n}$ mit $Q^\intercal Q = I$, so dass*

$$Q^\intercal A Q = T$$

mit T tridiagonal gilt.

Die Tridiagonalisierung lässt sich effizient und numerisch stabil mit Householder-Reflexionen implementieren, der Aufwand ist dennoch $\mathcal{O}(n^3)$ (vgl. [11, 31]).
Im Folgenden sei immer

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & 0 & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

Nachdem man $Q^\intercal AQ = T$ tridiagonalisert hat, wendet man auf T den folgenden Algorithmus an.

Algorithmus. Der QR-Algorithmus
input: $T \in \mathbb{R}^{n\times n}$ Folge von Shifts $(\nu_i)_{i\geq 0} \subset \mathbb{R}$
output: Folge von Iterierten $(T_i)_{i\geq 0}$
1: $T_0 := T$
2: **for** $i := 0, 1, 2, \dots$ **do**
3: $\quad T_i - \nu_i I = Q_i R_i$
4: $\quad T_{i+1} := R_i Q_i + \nu_i I$

Die benötigte QR-Zerlegung kann hierbei sehr effizient in $\mathcal{O}(n)$ Operationen mittels Givens-Rotationen umgesetzt werden. Jede Iterierte ist dann wieder tridiagonal und symmetrisch (vgl. [13]). Man kann ferner zeigen, dass die Konvergenz des QR-Verfahrens, wie auch beim RQI, lokal kubisch ist (vgl. [13]). Außerdem besitzen die Iterierten des QR-Verfahrens die folgenden Eigenschaften.

4.32 Lemma ([13])**.** *Für die Iterierten des QR-Verfahrens gilt*

1. $T_{i+1} = Q_i^\intercal T_i Q_i$,
2. $T_{i+1} = (Q_0 Q_1 \cdots Q_k)^\intercal T (Q_0 Q_1 \cdots Q_k)$,
3. $\prod_{j=0}^{i} (T - \mu_j I) = (Q_0 Q_1 \cdots Q_k)(R_k R_{k-1} \cdots R_0)$.

Das Lemma impliziert, dass alle Iterierten T_i dieselben Eigenwerte besitzen, da sie aus Ähnlichkeitstransformationen auseinander hervorgehen.
Für meine Diplomarbeit verwende ich die Softwarebibliothek LAPACK[4] zur Berechnung der Eigenwerte. Für Probleme der Dimension $n \leq 50$ wird hier

[4] `http://www.netlib.org/lapack/`

ein *Doppel-Shift*-QR-Algorithums und für größere Probleme ein *Multi-Shift*-QR-Algorithmus verwendet (vgl. [30]). Um das genauer zu erläutern, modifizieren wir den QR-Algorithmus durch Einführung einer Transformationsfunktion p_i wie bei der Unterraumiteration (vgl. [26, 30]).

Algorithmus. QR-Algorithmus mit expliziten Shifts
input: $T \in \mathbb{R}^{n\times n}$
output: Folge von Iterierten $(T_i)_{i\geq 0}$
1: $T_0 := T$
2: **for** $i := 0, 1, 2, \ldots$ **do**
3: $\quad p_i(T_i) = Q_i R_i$
4: $\quad T_{i+1} := Q_i^\intercal T_i Q_i$

Wählen wir in dem Algorithmus $p_i(x) = p(x) = (x - \nu_1)\cdots(x - \nu_l)$, so entspricht, gemäß Lemma 4.32, ein Schritt des QR-Algorithmus mit expliziten Shifts genau l Schritten des QR-Algorithmus. Wählen wir

$$p(x) = (x - \mu)(x - \nu),$$

erhalten wir den *Doppel-Shift*-QR-Algorithmus. Für die Wahl von Funktionen

$$p(x) = (x - \nu_1)\cdots(x - \nu_l)$$

erhalten wir den *Multi-Shift*-QR-Algorithmus. Hierbei kann l theoretisch beliebig groß gewählt werden. In der Praxis wählt man aber, wegen auftretender Rundungsfehler, l nicht wesentlich größer als sechs (vgl. [30]). Der Doppel-Shift-QR-Algorithmus bietet sich bei Eigenwertproblemen mit komplexen Shifts an. Wählt man, wie schon bei der Unterraumiteration erläutert, komplex konjugierte Paare $\nu, \overline{\nu}$ als Shifts, bleiben alle Rechnungen reell. Für eine detailierte Erklärung der Shift-Strategien sei auf [26, 30, 31] verwiesen. Wie wir bereits gesehen haben, kann die Konvergenz durch eine geeignete Wahl der Funktionen p_i beschleunigt werden. Dieses Verhalten überträgt sich von der Unterraumiteration auch auf das QR-Verfahren.
Es ist möglich das QR-Verfahren als geschachtelte Unterraumiteration zu interpretieren, was nun erläutert werden soll. Wir beginnen mit dem Zusammenhang zwischen dem QR-Verfahren mit expliziten Shifts und dem QR-Verfahren mit impliziten Shifts. Hierzu benötigen wir folgenden

4.33 Satz. *Sei $A \in \mathbb{R}^{n\times n}$ und $x \in \mathbb{R}^n \setminus \{0\}$. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$, so dass $Qe_1 = \beta x$ für ein $\beta \neq 0$ und $T = Q^\intercal AQ$ ist tridiagonal.*

Beweis. Zur Tridiagonalisierung der Matrix A benötigt man $n-2$ Householder-Reflexionen Q_i für $i = 1, \ldots, n-2$ (vgl. [11, 31]). Für jede dieser Matrizen gilt $Q_i e_1 = e_1$ für $i = 1, \ldots, n-2$. Daher ist es möglich die Matrix A zunächst mit einer Matrix Q_0 zu transformieren, die $Q_0 e_1 = \beta x$ erfüllt. Wählen wir dann die Householder-Reflexionen $\tilde{Q}_i e_1 = e_1$ für $i = 1, \ldots, n-2$ bezüglich $Q_0^\intercal A Q_0$, so gilt mit $Q := Q_0 \tilde{Q}_1 \cdots \tilde{Q}_{n-2}$, dass

$$Q e_1 = Q_0 e_1 = \beta x.$$

Also ist die erste Spalte von Q proportional zu x. □

Algorithmus. QR-Verfahren mit impliziten Shifts

input: $T \in \mathbb{R}^{n \times n}$

output: Folge von Iterierten $(T_i)_{i \geq 0}$

1: $T_0 := T$
2: **for** $i := 0, 1, 2, \ldots$ **do**
3: Wähle Shifts $\nu_1, \ldots, \nu_l$
4: $x := p(T_i)e_1 = \alpha(T_i - \nu_1 I) \cdots (T_i - \nu_l I)e_1$
5: $x = [\underbrace{*, \ldots, *}_{=:\tilde{x}}, \underbrace{0, \ldots, 0}_{n-l-1}]^\intercal$ mit $\tilde{x} \in \mathbb{R}^{l+1}$
6: Wähle $\tilde{Q}_0 \in \mathbb{R}^{(l+1)\times(l+1)}$ mit $\tilde{Q}_0 e_1 = \beta \tilde{x}$ für $\beta \neq 0$
7: $Q_0 := \begin{bmatrix} \tilde{Q}_0 & 0 \\ 0 & I_{n-l-1} \end{bmatrix}$
8: $B_0 := Q_0^\intercal T_i Q_0$
9: **for** $j := 1, \ldots, n-2$ **do**
10: $B_j := Q_j^\intercal B_{j-1} Q_j$
11: $T_{i+1} := B_{n-2}$

Wir wollen den Algorithmus erläutern.
In Zeile 4 wird das Produkt $x := p(T_i)e_1 = \alpha(T_i - \nu_1 I) \cdots (T_i - \nu_l I)e_1$ gebildet. Da T_i tridiagonal ist, sind nur die ersten $l+1$ Einträge von x ungleich 0. Diese Einträge werden in $\tilde{x}$ gespeichert (Zeile 5). Die in Zeile 7 aufgestellte Matrix Q_0 erfüllt $Q_0 e_1 = \beta x$. Die beidseitige Multiplikation von T_i mit Q_0 in Zeile 8 stört die Tridiagonalform von T_i und es entsteht eine Beule (engl. bulge) mit Spitzen in Position $(l+2, 1)$ beziehungsweise in Position $(1, l+2)$. In der For-Schleife in Zeile 9 wird diese Beule durch Ähnlichkeitstransformationen Q_i für $i = 1, \ldots, n-2$ eliminiert. Dies bezeichnet man als *bulge chase* und dementsprechend derartige Algorithmen als *bulge-chasing-Algorithmen* (vgl.[11, 13, 31]).
Den Zusammenhang zwischen impliziten und expliziten QR-Verfahren liefert nun das *Implicit-Q-Theorem.*

4.34 Satz (Implicit-Q-Theorem, [11]). *Seien $Q = [q_1, \ldots, q_n]$ und $V = [v_1, \ldots, v_n]$ orthogonale Matrizen mit $Q^\intercal AQ = T$ und $V^\intercal AV = S$, wobei S und T beide tridiagonal sind. Sei k der kleinste Index mit $t_{k+1,k} = 0$. Wir setzen $k = n$, falls T echte obere Hessenberg-Form besitzt. Gilt dann $v_1 = q_1$, so folgt*

$$v_i = \pm q_i \qquad \text{und} \qquad |t_{i-1,i}| = |t_{i-1,i}| \qquad \text{für } i = 2, \ldots, k.$$

Ist ferner $k < n$, so ist auch $s_{k+1,k} = 0$.

Wählen wir $v_1 = \beta p(T)e_1$, was nach Satz 4.33 möglich ist, und betrachten $Q = [q_1, \ldots, q_n]$ als resultierende Matrix des QR-Verfahrens mit expliziten Shifts und $V = [v_1, \ldots, v_n]$ als diejenige aus dem QR-Verfahren mit impliziten Shifts, so gilt $q_1 = v_1$. Wir nehmen an, dass $Q^\intercal TQ$ echte Tridiagonalform besitzt, sonst zerfällt das Problem in zwei entkoppelte Teilprobleme (vgl. [11]). Das Implicit-Q-Theorem liefert die Aussage, dass die Zerlegungen bis auf Vorzeichen übereinstimmen. $p(T_i)$ muss also niemals explizit aufgestellt werden, es genügt die erste Spalte $p(T_i)e_1$ zu berechnen.

Der Zusammenhang zur Unterraumiteration ergibt sich jetzt wie folgt. Wir führen einen Schritt des QR-Verfahrens mit impliziten Shifts aus. Ist T_1 nun keine echte Tridiagonalmatrix mehr, zerfällt das Problem in zwei Teilprobleme. Daher nehmen wir für die folgende Überlegung T_1 als echte Tridiagonalmatrix an. Aus Satz 4.29 ergibt sich, dass die m führenden Spalten von $Q := Q_0 \cdots Q_{n-2}$ einen Krylov-Raum aufspannen,

$$\operatorname{span}\{q_1, \ldots, q_m\} = \mathcal{K}_m(T, q_1) \qquad \text{für } m = 1, \ldots, n.$$

Da $q_1 = \beta x = \beta p(T_0)e_1$ gilt, liefert Satz 4.28

$$\mathcal{K}_m(T_0, q_1) = \mathcal{K}_m(T_0, x) = p(T_0)\mathcal{K}_m(T_0, e_1).$$

Weiter folgt aus Satz 4.27, dass $\mathcal{K}_m(T_0, e_1) = \operatorname{span}\{e_1, \ldots, e_m\}$ gilt. Insgesamt erhalten wir

$$p(T)\operatorname{span}\{e_1, \ldots, e_m\} = \operatorname{span}\{q_1, \ldots, q_m\} \qquad \text{für } m = 1, \ldots, n.$$

Dies bedeutet, dass die Spalten von Q das Ergebnis einer Unterraumiteration mit $p(T)$ und Startvektoren $e_1, \ldots, e_n$ sind.
Die Ähnlichkeitstransformation $T_1 = Q^\intercal TQ$ beschreibt einen Basiswechsel, der jeden Vektor x auf Qx abbildet. Die Vektoren $q_1, \ldots, q_n$ werden zurück auf die kanonische Basis $e_1, \ldots, e_n$ abgebildet, da $Q^\intercal q_i = e_i$ für $i = 1, \ldots, n$ gilt (vgl. [31]).
Insgesamt bewirkt also ein QR-Schritt eine geschachtelte Unterraumiteration mit $p(T_0)$ auf den Unterräumen $\operatorname{span}\{e_1, \ldots, e_m\}$ für $m = 1, \ldots, n$. Die resultierenden Unterräume werden dann über den Basiswechsel Q auf die Unterräume $\operatorname{span}\{e_1, \ldots, e_m\}$ für $m = 1, \ldots, n$ zurücktransformiert (vgl. [31]).

4.2.8 Das Lanczos-Verfahren

Das Lanczos-Verfahren ist im symmetrischen Fall das in der Praxis meist genutzte Verfahren, wenn es um die Lösung großer, dünn besetzter Eigenwertprobleme geht, wie sie häufig aus der Diskretisierung von partiellen Differenzialgleichungen entstehen. In diesem Fall ist der Aufwand zur Berechnung des Matrix-Vektor-Produktes $w := Av$ eher von der Ordnung $\mathcal{O}(n)$ als $\mathcal{O}(n^2)$ (vgl. [11, 23]). Wie in [11] führen wir das Lanczos-Verfahren als Optimierung des Rayleigh-Quotienten

$$\Lambda(x) = \frac{x^\intercal Ax}{x^\intercal x} \qquad x \neq 0.$$

ein. Sei $A \in \mathbb{R}^{n\times n}$ wieder symmetrisch. Wir haben bereits gezeigt, dass

$$\lambda_n = \min_{x\in\mathbb{R}^n} \Lambda(x) \qquad \text{und} \qquad \lambda_1 = \max_{x\in\mathbb{R}^n} \Lambda(x)$$

gilt, wobei $\lambda_1 \geq \ldots \geq \lambda_n$ die absteigend geordneten Eigenwerte von A sind.

4.35 Definition. Sei $(q_i)_{i=1}^m \subset \mathbb{R}^n$ ein System orthonormaler Vektoren mit $m \leq n$ und sei $Q_k = [q_1, \ldots, q_k]$ für $k \leq m$. Dann definieren wir

$$\begin{aligned} M_k &:= \lambda_1(Q_k^\intercal A Q_k) = \max_{\|y\|_2=1} \Lambda(Q_k y) \leq \lambda_1 \qquad \text{und} \\ m_k &:= \lambda_k(Q_k^\intercal A Q_k) = \min_{\|y\|_2=1} \Lambda(Q_k y) \geq \lambda_n. \end{aligned}$$

Die Gültigkeit dieser Ungleichungen haben wir bereits gezeigt. Die Idee des Lanczos-Algorithmus ist die Überlegung, wie es möglich ist, das System der q_i so aufzubauen, dass M_k und m_k sukzessive bessere Approximationen an λ_n, respektive λ_1 sind (vgl. [11]).
Sei $u_k \in \operatorname{span}\{q_1, \ldots, q_k\}$ mit $M_k = \Lambda(u_k)$. Da $\Lambda(x)$ am schnellsten in Richtung des Gradienten

$$\nabla\Lambda(x) = \frac{2}{x^\intercal x}(Ax - \Lambda(x)x)$$

wächst, können wir, unter der Voraussetzung $\nabla\Lambda(u_k) \neq 0$, garantieren, dass $M_{k+1} > M_k$ gilt, wenn q_{k+1} so gewählt wird, dass

$$\nabla\Lambda(u_k) \in \operatorname{span}\{q_1, \ldots, q_{k+1}\} \tag{4.3}$$

erfüllt ist. Analog fordern wir für $v_k \in \operatorname{span}\{q_1, \ldots, q_k\}$ mit $m_k = \Lambda(v_k)$, dass

$$\nabla\Lambda(v_k) \in \operatorname{span}\{q_1, \ldots, q_{k+1}\}, \tag{4.4}$$

da $\Lambda(x)$ am schnellsten in Richtung des negativen Gradienten $-\nabla\Lambda(x)$ abfällt. Es ist

$$\nabla\Lambda(x) \in \operatorname{span}\{x, Ax\} = \mathcal{K}_2(A, x).$$

Damit sind die Gleichungen (4.3) und (4.4) erfüllt, falls

$$\operatorname{span}\{q_1, \ldots, q_k\} = \mathcal{K}_k(A, q_1)$$

gilt. Die Gültigkeit dieser Identität haben wir bereits in Satz 4.29 gezeigt. Also muss q_{k+1} so gewählt werden, dass ferner

$$\operatorname{span}\{q_1, \ldots, q_{k+1}\} = \mathcal{K}_{k+1}(A, q_1)$$

gilt. Damit haben wir das Problem auf die Berechnung einer Orthonormalbasis der Krylov-Räume $\mathcal{K}_m(A, q_1)$ für $m = 1, 2, \ldots$ zurückgeführt. Nach Satz 4.29 kann eine Orthonormalbasis einfach durch eine Tridiagonalisierung der Matrix mit Hilfe der QR-Zerlegung erzeugt werden. Dass hierbei der Startvektor q_1 beliebig gewählt werden kann, liefert der Satz 4.33. Da allerdings durch eine QR-Zerlegung mittels Householder-Reflexionen oder Givens-Rotationen die Struktur der Matrix A zerstört werden kann, berechnet man in der Praxis die Einträge von $T = Q^\intercal AQ$ direkt. Hierzu setzt man

$$QT = AQ$$

an. Mit

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \ldots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \ldots & 0 & \beta_{n-1} & \alpha_n \end{bmatrix}$$

führt dies auf folgende Gleichungen:

$$Aq_k = \beta_{k-1}q_{k-1} + \alpha_k q_k + \beta_k q_{k+1} \qquad \text{mit } \beta_0 q_0 := 0 \tag{4.5}$$

für $k = 1, \ldots, n-1$.
Die Orthonormalität der q_k liefert

$$\alpha_k = q_k^\intercal A q_k.$$

Ferner gilt

$$q_{k+1} = \frac{1}{\beta_k} r_k \qquad \text{und } \beta_k = \pm \|r_k\|_2$$

mit

$$r_k = (A - \alpha_k I)q_k - \beta_{k-1}q_{k-1}$$

(vgl. [11]). Falls $r_k = 0$ gilt, bricht die Iteration ab. Dies ist aber, wie wir bereits gezeigt haben, äquivalent dazu, dass wir einen invarianten Unterraum von A gefunden haben. In diesem Falle sind alle Approximationen an die Eigenwerte exakt. Insgesamt gelangen wir zu folgendem Algorithmus.

Algorithmus (Der Lanczos-Prozess)**.**

1: Initialisiere $r_0 := q_1$, $\beta_0 := 1$, $q_0 := 0$, $i := 0$
2: **while** $\beta_i \neq 0$ **do**
3: $\quad q_{i+1} := r_i/\beta_i$
4: $\quad i := i + 1$
5: $\quad \alpha_i := q_i^\intercal A q_i$
6: $\quad r_i := (A - \alpha_i I)q_i - \beta_{i-1}q_{i-1}$
7: $\quad \beta_i := \|r_i\|_2$

Ohne Beschränkung der Allgemeinheit wird hierbei β_i als positiv vorausgesetzt. Die entstehenden Vektoren q_i für $i = 1, 2, \ldots$ heißen dann *Lanczos-Vektoren* (vgl. [11]).

Bemerkung. Es besteht ein direkter Zusammenhang zwischen Lanczos-Prozess und Orthogonal-Polynomen. Darauf wollen wir kurz eingehen. In exakter Arithmetik besitzt der Algorithmus die Form einer Dreitermrekursion

$$\beta_i q_{i+1} = Aq_i - \alpha_i q_i - \beta_{i-1}q_{i-1}.$$

Sei der Grad (vgl. Definition 4.19) von q_1 größer oder gleich m. Dann gilt $\dim(\mathcal{K}_m) = m$. Ferner haben wir festgestellt, dass $\mathcal{K}_m = \{x \in \mathbb{R}^n \mid x = p(A)q_1, p \in \Pi_{m-1}\}$. Folglich existiert ein Isomorphismus zwischen $\mathcal{K}_m$ und Π_{m-1}

$$\begin{aligned} \Pi_{m-1} &\longrightarrow \mathcal{K}_m, \\ p &\longmapsto x = p(A)q_1. \end{aligned}$$

Wir können Π_{m-1} mit dem Innenprodukt

$$(p, q)_{q_1} := (p(A)q_1, q(A)q_1)$$

versehen. Aufgrund der Annahme, dass m nicht größer als der Grad von q_1 ist, ist dies eine nicht entartete Bilinearform. Bemerken wir nun, dass die Vektoren q_i dargestellt werden können als

$$q_i = p_{i-1}(A)q_1,$$

überträgt sich die Orthogonalität der q_i auf die zugehörigen Polynome p_{i-1} bezüglich des Innenproduktes $(\cdot, \cdot)_{q_1}$ (vgl. [26]).

Bevor wir auf die Details der Implementierung des Lanczos-Verfahrens in ARPACK eingehen, halten wir noch eine Fehlerabschätzung und eine Konvergenzaussage fest.

4.36 Satz. *Sei $A \in \mathbb{R}^{n\times n}$ und sei $q_1 \in \mathbb{R}^n$ mit $\|q_1\|_2 = 1$. Dann bricht der Lanczos-Prozess bei $i = m$ ab, falls m der Grad von q_1 bezüglich A ist. Ferner gilt für $i = 1, \ldots, m$ die Darstellung*

$$AQ_i = Q_i T_i + r_i e_i^\intercal,$$

wobei

$$T_i = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \ldots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \beta_{i-1} \\ 0 & \ldots & 0 & \beta_{i-1} & \alpha_i \end{bmatrix}$$

und $Q_i = [q_1, \ldots, q_i]$ orthogonal ist, mit $\mathcal{R}(Q_i) = \mathcal{K}_i(A, q_1)$.

Beweis. Der erste Teil des Satzes wurde bereits im Abschnitt Krylov-Räume gezeigt. Der Beweis des zweiten Teils ist analog zum Beweis von Satz 4.29 und erfolgt per vollständiger Induktion.
Angenommen die Iteration hat bereits $Q_i = [q_1, \ldots, q_i]$ mit

$$\mathcal{R}(Q_i) = \mathcal{K}_i(A, q_1) \quad \text{und} \quad Q_i^\intercal Q_i = I_i$$

erzeugt. Aus Zeile 6 des Algorithmus ist leicht zu sehen, dass dann $AQ_i = Q_i T_i + r_i e_i^\intercal$ gilt. Das ist äquivalent zu

$$Q_i^\intercal A Q_i = T_i + Q_i^\intercal r_i e_i^\intercal.$$

Nun gilt aber $\alpha_k = q_k^\intercal A q_k$ für $k = 1, \ldots, i$ und

$$q_{k+1}^\intercal A q_k = q_{k+1}^\intercal (A q_k - \alpha_k q_k - \beta_{k-1} q_{k-1}) \overset{(4.5)}{=} q_{k+1}^\intercal \beta_k q_{k+1} = \beta_k$$

für $k = 1, \ldots, i-1$. Also haben wir $Q_i^\intercal A Q_i = T_i$ und damit $Q_i^\intercal r_i = 0$. Ist $r_i \neq 0$, dann ist $q_{i+1} = r_i / \|r_i\|_2$ orthogonal zu $q_1, \ldots, q_i$ und

$$q_{i+1} \in \operatorname{span}\{A q_i, q_i, q_{i-1}\} \subseteq \mathcal{K}_{i+1}(A, q_1).$$

□

Eine Aussage über die Konvergenz des Lanczos-Prozesses liefert der nächste Satz.

4.37 Satz ([11]). *Sei $A \in \mathbb{R}^{n\times n}$ mit Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und zugehörigen Eigenvektoren $u_1, \ldots, u_n$. Sind $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_i$ die Eigenwerte der Matrix T_i, die aus i Schritten des Lanczos-Prozesses resultieren, so gilt*

$$\lambda_1 \geq \tilde{\lambda}_1 \geq \lambda_1 - \frac{(\lambda_1 - \lambda_n)\tan^2(\angle(q_1, u_1))}{(c_{i-1}(1+2\rho_1))^2},$$

wobei $\rho_1 = (\lambda_1 - \lambda_2)/(\lambda_2 - \lambda_n)$ und c_{i-1} das Chebyshev-Polynom vom Grad $i-1$ ist.

Die Fehlerabschätzung bezüglich des kleinsten Eigenwertes ergibt sich hieraus als Korollar.

4.38 Korollar ([11]). *In der Situation des obigen Satzes gilt*

$$\lambda_n \leq \tilde{\lambda}_n \leq \lambda_1 + \frac{(\lambda_1 - \lambda_n)\tan^2(\angle(q_n, u_n))}{(c_{i-1}(1+2\rho_n))^2}$$

mit $\rho_n = (\lambda_{n-1} - \lambda_n)/(\lambda_1 - \lambda_{n-1})$.

Bemerkung. Da der Krylov-Raum $\mathcal{K}_m(A, z)$ auch den Vektor $A^{m-1}z$ enthält, ist durch den Lanczos-Prozess eine bessere Approximation an den größten Eigenwert gegeben, als durch die einfache Potenzmethode (vgl. [11, 14]). Trotzdem gibt es Situationen, in denen die Unterraumiteration noch konkurrieren kann. Beispielsweise, wenn der Arbeitsspeicher sehr begrenzt ist, oder der Abstand zwischen den gesuchten und den anderen Eigenwerten sehr groß ist (vgl. [10]).

Die Implementierung des Lanczos-Verfahrens erfolgt in der Softwarebibliothek ARPACK in Form der so genannten *implicit restarted Arnoldi method* (vgl. [23]). Auf deren Funktionsweise werden wir nun abschließend noch eingehen.
A priori ist nicht klar, wie viele Lanczos-Vektoren berechnet werden müssen, um akzeptable Approximationen an die Eigenwerte zu erhalten. Um dieses Problem zu lösen, kann die *implicit-restarting*-Technik verwendet werden. Zentrale Idee hierbei ist, den Startvektor der Iteration in jedem Schritt mit einem geigneten Polynom von A vorzukonditionieren und so ungewollte Anteile des Spektrums herauszudämpfen. Hierdurch beschränkt sich der Speicherbedarf von ARPACK zur Berechnung von k Eigenwerten auf $2nk + \mathcal{O}(k^2)$. Dies wird durch die Kombination von einem QR-Verfahren mit impliziten Shifts mit einem Lanczos-Prozess festgelegter Länge erreicht. (vgl. [23]).

Zur besseren Übersicht sei ab jetzt $V_m = [v_1, \ldots, v_m]$ die Matrix, deren Spalten eine Orthonormalbasis von $\mathcal{K}_m$ bilden und sei Q die orthogonale Transformation aus dem QR-Verfahren. Ferner setzen wir $m = k + p$. Sei nun

$$\begin{aligned} AV_m &= V_m T_m + r_m e_m^\intercal \\ &= [V_m, v_{m+1}] \begin{bmatrix} T_m \\ \beta_m e_m^\intercal \end{bmatrix} \end{aligned}$$

ein Lanczos-Prozess der Länge m. Wenden wir auf T_m einen QR-Schritt mit implizitem Shift ν an, so ist die zugehörige orthogonale Transformation $Q \in \mathbb{R}^{m \times m}$ eine obere Hessenberg-Matrix (vgl. [23, 28]) und wir gelangen zu

$$AV_m Q = [V_m Q, v_{m+1}] \begin{bmatrix} Q^\intercal T_m Q \\ \beta_m e_m^\intercal Q \end{bmatrix}.$$

Die Anwendung von p impliziten Shifts $\nu_1, \ldots, \nu_p$ führt deshalb auf

$$AV_m^+ = [V_m^+, v_{m+1}] \begin{bmatrix} T_m^+ \\ \beta_m e_m^\intercal \hat{Q} \end{bmatrix} \tag{4.6}$$

mit $V_m^+ = V_m \hat{Q}$, $T_m^+ = \hat{Q}^\intercal T_m \hat{Q}$ und $\hat{Q} = Q_1 \cdots Q_p$, wobei Q_j die orthogonale Matrix aus dem QR-Schritt mit Shift ν_j bezeichne. Wir partitionieren

$$V_m^+ = [V_k^+, \hat{V}_p], \quad T_m^+ = \begin{bmatrix} T_k^+ & \hat{\beta}_k e_k e_1^\intercal \\ \hat{\beta}_k e_1 e_k^\intercal & \hat{T}_p \end{bmatrix}.$$

Da jedes $Q_1, \ldots, Q_p$ eine obere Hessenberg-Matrix ist, gilt

$$\beta_m e_m^\intercal \hat{Q} = [\underbrace{0, \ldots, 0, \tilde{\beta}_m}_{k}, \underbrace{b^\intercal}_{p}]$$

(vgl. [23, 28]). Einsetzen in Gleichung (4.6) liefert

$$A[V_k^+, \hat{V}_p] = [V_k^+, \hat{V}_p, v_{m+1}] \begin{bmatrix} T_k^+ & \hat{\beta}_k e_k e_1^\intercal \\ \hat{\beta}_k e_1 e_k^\intercal & \hat{T}_p \\ \tilde{\beta}_m e_k^\intercal & b^\intercal \end{bmatrix}.$$

Gleichsetzen der ersten k Spalten dieser Gleichung ergibt nun

$$AV_k^+ = V_k^+ T_k^+ + r_k^+ e_k^\intercal \tag{4.7}$$

mit $r_k^+ = \hat{V}_p e_1 \hat{\beta}_k + v_{m+1} \tilde{\beta}_m$. Setzen wir

$$v_{k+1}^+ = \frac{1}{\left\| r_k^+ \right\|_2} r_k^+,$$

erhalten wir unter Beachtung von $(V_k^+)^\intercal \hat{V}_p e_1 = 0$ und $(V_k^+)^\intercal v_{m+1} = 0$, dass

$$(V_k^+)^\intercal v_{k+1}^+ = 0.$$

Damit besitzt Gleichung (4.7) die Darstellung aus Satz 4.36 und entspricht somit einem Lanczos-Prozess der Länge k. Benutzt man dies als Ausgangspunkt, kann man p weitere Schritte des Lanczos-Prozesses ausführen und gelangt wieder zu einer Faktorisierung der Länge m (vgl. [28]). Setzt man dieses Vorgehen iterativ fort, erhält man die *implicit restarted Arnoldi method*.
Ein Abbruchkriterium für die Iteration liefert hierbei die folgende Überlegung. Sei $T_m s = \tilde{\lambda} s$ mit $\|s\|_2 = 1$ und sei $\hat{x} = V_m s$. Dann gilt offensichtlich

$$\left\| A\hat{x} - \tilde{\lambda}\hat{x} \right\|_2 = \|AV_m s - V_m T_m s\|_2 = \|r_m\|_2 \left| e_m^\intercal s \right|.$$

Im hermiteschen Fall liefert dies eine präzise Abschätzung für den Fehler $\left|\tilde{\lambda} - \lambda\right|$, wobei λ der nächstgelegene Eigenwert von A ist. Die Softwarebibliothek ARPACK benutzt als Abbruchskriterium für die Iteration das Kriterium

$$\|r_m\|_2 \left| e_m^\intercal s \right| \leq \max(\varepsilon_M \|T_m\|, tol \cdot |\tilde{\lambda}|),$$

wobei ε_M die Maschinengenauigkeit und tol eine vom Benutzer definierte Genauigkeit ist (vgl. [23]).
Für die Feinheiten der Implementierung und eine detailierte Darstellung sei auf [28] und [23] verwiesen. Die Quintessenz ist, dass durch die im QR-Verfahren verwendeten impliziten Shifts $\nu_1, \ldots, \nu_p$ der Startvektor des Lanczos-Prozesses $Ve_1 =: v_1$ in jedem Zyklus der implicit restarted Arnoldi method mit einem Polynom vorkonditioniert wird,

$$v_1^{\text{neu}} = p(A) v_1^{\text{alt}}.$$

Hierbei ist $p(x) = \frac{1}{\tau}(x - \nu_1) \cdots (x - \nu_p)$ mit einem Normierungsfaktor $\tau \in \mathbb{R}$. Dadurch wird erreicht, dass ungewollte Anteile des Spektrums herausgedämpft werden (vgl. [23, 28]). Andererseits ist für den Fall, dass $m = n$ gewählt wird, das Residuum $r_m = 0$. Daher entspricht die implicit restarted Arnoldi method hier genau einem QR-Verfahren mit impliziten Shifts. Auch im Fall $m < n$ sind die ersten k Spalten von V_m und der obere rechte $k \times k$-Block von T_m mathematisch äquivalent zu denjenigen Matrizen, die in einem vollständigen QR-Verfahren auftreten würden. Hier sei auf den entsprechenden Zusammenhang zwischen der Unterraumiteration und dem QR-Verfahren verwiesen, auf den im Voraus eingegangen wurde. In diesem Sinne kann man daher die implicit restarted Arnoldi method als ein abgeschnittenes QR-Verfahren auffassen (vgl. [23]).

4.2.9 Das verallgemeinerte Eigenwertproblem

Das verallgemeinerte Eigenwertproblem

$$Ax = \lambda Mx \tag{4.8}$$

mit $A, M \in \mathbb{R}^{n\times n}$, beide symmetrisch und M zusätzlich positiv definit, kann auf zwei Arten behandelt werden. Die einfachere Möglichkeit ist, das Eigenwertproblem in ein Standard-Eigenwertproblem zu transformieren. Um die Symmetrie zu erhalten, verwendet man hierzu im Allgemeinen eine Wurzel der Matrix $M = X^2$ oder die Cholesky-Zerlegung $M = LL^\intercal$. Dann ist das verallgemeinerte Eigenwertproblem (4.8) äquivalent zu

$$Ax = \lambda LL^\intercal x.$$

Setzen wir $y = L^\intercal x$, so ergibt sich durch Multiplikation mit L^{-1} das Standard-Eigenwertproblem

$$L^{-1}AL^{-\intercal}y = \lambda y.$$

In der Praxis kann sich dieses Vorgehen jedoch als Nachteil erweisen. Der Aufwand für die Berechnung der Cholesky-Zerlegung von M ist $\mathcal{O}(n^3)$. Zusätzlich wird durch die Zerlegung die Struktur der Matrix womöglich zerstört (vgl. [11, 23, 25]). Besitzt die Matrix M eine Struktur, die es erlaubt, die Matrix günstig, das heißt mit einem Aufwand von beispielsweise $\mathcal{O}(n)$ oder $\mathcal{O}(n\log(n))$ zu invertieren, bietet sich eine analoge Vorgehensweise zu der bei verallgemeinerten Eigenwertproblemen kompakter Operatoren an (vgl. Kapitel 3). Da M invertierbar ist, kann das Eigenwertprobem (4.8) umgeschrieben werden zu

$$M^{-1}Ax = \lambda x.$$

Wählen wir nun den Raum $\mathbb{R}^n$ bezüglich des Innenproduktes $(\cdot,\cdot)_M := (\cdot, M\cdot)_{\mathbb{R}^n}$ und der Norm $\|\cdot\|_M = \sqrt{(\cdot,\cdot)_M}$, so ist $M^{-1}A$ in diesem Innenprodukt symmetrisch. Damit können alle Resultate für das Lanczos-Verfahren verwendet werden. Für die Unterraumiteration ergibt sich hingegen folgendes Vorgehen.

Algorithmus (Die Unterraumiteration für das verallgemeinerte Eigenwertproblem)**.**

1: $\tilde{X} := M^{-1}AX$
2: orthogonalisiere $\tilde{X}$ bzgl. $(\cdot,\cdot)_M$ in X
3: $A_m = X^\intercal AX$
4: bestimme alle Eigenwerten von A_m

Bestimmt man die Projektion von

$$Ax = \lambda Mx,$$

so ergibt sich insbesondere $X^\intercal MX = I$ (vgl. [12]) und das Eigenwertproblem reduziert sich in der Projektion zu

$$A_m x = \tilde{\lambda} x.$$

Dieses kann nun mittels des QR-Verfahrens gelöst werden.

4.2.10 Das Cholesky-Verfahren

Zum Abschluss des Kapitels Numerische Methoden wird hier noch ein interessantes Verfahren vorgestellt für den Fall, dass der betrachtete kompakte Operator positiv semidefinit ist und seine Eigenwerte ein gewisses, genauer zu definierendes Abklingverhalten aufweisen. Unter den gegebenen Voraussetzungen ist auch die Galerkin-Diskretisierung des Operators symmetrisch und positiv semidefinit. Daher sei in diesem Abschnitt $A \in \mathbb{R}^{n\times n}$ immer eine symmetrische, positiv semidefinite Matrix. Wir werden untersuchen, wann es möglich ist, mit Hilfe der Cholesky-Zerlegung eine *Niedrigrangapproximation* (vgl. [2, 16]) an die Matrix A zu berechnen. Dies bedeutet, wir suchen zu einem vorgegebenen $\varepsilon > 0$ ein $A_m \in \mathbb{R}^{n\times n}$ mit $\operatorname{rank}(A_m) = m$, das

$$\|A - A_m\|_2 \leq \varepsilon$$

erfüllt. Wir haben bereits gezeigt, dass die abgebrochene Singulärwertzerlegung in der Hinsicht optimal ist, dass sie unter allen Matrizen vom Rang m eben diesen Fehler minimiert. Bezeichnen wir mit $\lambda_1 \geq \ldots \geq \lambda_n$ die Eigenwerte von A, so gilt in diesem Fall

$$\|A - A_m\|_2 = \lambda_{m+1}$$

(vgl. [2]). Eine andere Möglichkeit den Approximationsfehler ε zu messen, bietet die *Spurnorm*.

4.39 Definition. Wir definieren

$$\|A\|_{\mathrm{tr}} := \operatorname{trace}(A) = \sum_{i=1}^{n} a_{i,i}.$$

Man sieht leicht ein, dass alle Diagonalelemete von A nichtnegativ sind. Aus der positiv Semidefinitheit folgt nämlich

$$a_{i,i} = (e_i, Ae_i) \geq 0.$$

Um im Folgenden die Darstellungen möglichst unkompliziert zu halten, bezeichnen wir für Matrizen A die i-te Spalte mit a_i. Essenziell für die Überlegungen in diesem Abschnitt ist die Abschätzung der Euklidischen- durch die Spurnorm.

4.40 Satz. *Auf der Menge der symmetrischen und positiv definiten Matrizen wird durch* $\|\cdot\|_{\mathrm{tr}}$ *eine Norm definiert und es gilt*

$$\frac{1}{n}\|A\|_{\mathrm{tr}} \leq \|A\|_2 \leq \|A\|_{\mathrm{tr}}.$$

Beweis. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit Eigenwerten $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$. Nach [3] gilt

$$\mathrm{trace}(A) = \sum_{i=1}^{n} \lambda_i.$$

Hieraus folgt bereits $\|A\|_{\mathrm{tr}} = 0 \Leftrightarrow A = 0$, da alle Eigenwerte nichtnegativ sind. Die anderen Normeigenschaften gehen direkt aus der Definition hervor. Ferner gilt

$$\frac{1}{n}\|A\|_{\mathrm{tr}} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i \leq \frac{1}{n}n\lambda_1 = \|A\|_2 \leq \sum_{i=1}^{n}\lambda_i = \mathrm{trace}(A) = \|A\|_{\mathrm{tr}}.$$

□

Nehmen wir im Folgenden stets an, dass $A \neq 0$, so existiert mindestens ein positives Diagonalelement. Ohne Einschränkung nehmen wir an, dass sich dieses Element in der $(1,1)$-Position befindet. Sonst existiert ein Index $1 \leq i \leq n$, so dass $a_{i,i} > 0$ gilt und wir können die erste und die i-te Zeile, sowie die erste und i-te Spalte vertauschen. Diese Transposition wird durch eine Permutationsmatrix $P \in \mathbb{R}^{n\times n}$ mit $P^2 = I$ beschrieben. Das Schur-Komplement bezüglich $a_{1,1}$ ist dann wieder positiv semidefinit.

4.41 Lemma. *Sei*

$$A = \left[\begin{array}{c|c} a_{1,1} & A_{2,1}^{\intercal} \\ \hline A_{2,1} & A_{2,2} \end{array}\right] \in \mathbb{R}^{n\times n}$$

mit $a_{1,1} > 0$. *Dann ist das Schur-Komplement*

$$S := A_{2,2} - \frac{1}{a_{1,1}} A_{2,1}A_{2,1}^{\intercal} \in \mathbb{R}^{(n-1)\times(n-1)}$$

wohldefiniert und ebenfalls positiv semidefinit.

Beweis. Da $a_{1,1} > 0$ gilt, ist das Schur-Komplement wohldefiniert. Die Symmetrie folgt aus

$$S^\intercal = A_{2,2}^\intercal - \frac{1}{a_{1,1}}\big(A_{2,1}A_{2,1}^\intercal\big)^\intercal = A_{2,2} - \frac{1}{a_{1,1}}A_{2,1}A_{2,1}^\intercal = S.$$

Sei $y \in \mathbb{R}^{n-1}$. Wir setzen $x := -\frac{1}{a_{1,1}}A_{2,1}^\intercal y \in \mathbb{R}$. Dann folgt

$$0 \le \begin{bmatrix} x \\ y \end{bmatrix}^\intercal A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}^\intercal \begin{bmatrix} a_{1,1}x + A_{2,1}^\intercal y \\ xA_{2,1} + A_{2,2}y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}^\intercal \begin{bmatrix} 0 \\ Sy \end{bmatrix} = y^\intercal S y,$$

das heißt, S ist positiv semidefinit. □

Für positiv semidefinite Matrizen ist die symmetrische Pivotisierung (vgl. [11]), wie wir sie in diesem Abschnitt verwenden, äquivalent zur totalen Pivotisierung.

4.42 Satz. *Für $A \in \mathbb{R}^{n\times n} \setminus \{0\}$ symmetrisch und positiv semidefinit gilt*

1. $|a_{i,j}| \le \sqrt{a_{i,i}a_{j,j}} \le (a_{i,i} + a_{j,j})/2$ *für* $i, j = 1, \ldots, n$,
2. $\max\limits_{i,j=1,\ldots,n} |a_{i,j}| = \max\limits_{i=1,\ldots,n} a_{i,i}$,
3. $a_{i,i} = 0 \implies a_{i,j} = a_{j,i} = 0$ *für* $j = 1, \ldots, n$.

Beweis. Da das Schur-Komplement S von A ebenfalls positiv semidefinit ist und somit $s_{i,i} \ge 0$ für $i = 1, \ldots, n-1$ gilt, folgt

$$0 \le s_{j,j} = a_{j,j} - a_{j,1}^2/a_{1,1}.$$

Da diese Beziehung für jede symmetrische Transposition PAP von Zeilen und Spalten von A gelten muss, ergibt sich unter der Voraussetzung $a_{i,i} \ne 0$

$$0 \le s_{j,j} = a_{j,j} - a_{j,i}^2/a_{i,i} \qquad \text{für } i, j = 1, \ldots, n.$$

Dies bedeutet

$$a_{i,j}^2 = a_{j,i}^2 \le a_{i,i}a_{j,j},$$

woraus sich die Behauptung über das geometrische Mittel ergibt. Andererseits lässt sich das geometrische Mittel immer durch das arithmetische Mittel abschätzen. Damit ist 1. gezeigt. 2. und 3. sind direkte Folgerungen aus 1. □

Wir stellen nun A dar als

$$PAP = \frac{1}{a_{1,1}} a_1 a_1^\intercal + \left[\begin{array}{c|c} 0 & 0^\intercal \\ \hline 0 & S \end{array}\right].$$

Falls $S \neq 0$ gilt, können wir eine analoge Zerlegung für S finden. Wir werden diese Zerlegung nun iterieren. Hierzu sei $S^{(i)}$ das Schur-Komplement im i-ten Schritt, mit der Darstellung

$$S^{(i)} = \left[\begin{array}{c|c} s_{1,1}^{(i)} & \left(S_{2,1}^{(i)}\right)^\intercal \\ \hline S_{2,1}^{(i)} & S_{2,2}^{(i)} \end{array}\right] \in \mathbb{R}^{(n-i)\times(n-i)},$$

wobei $S^{(0)} := A$ sei. Damit gelangen nach m Schritten zu einer Zerlegung der Form

$$P_m \cdots P_2 P_1 A P_1 P_2 \cdots P_m = \sum_{i=1}^{m} \hat{l}_i \hat{l}_i^\intercal + \left[\begin{array}{c|c} 0 & 0^\intercal \\ \hline 0 & S^{(m)} \end{array}\right] \tag{4.9}$$

mit

$$\hat{l}_i := \frac{1}{\sqrt{s_{1,1}^{(i-1)}}} P_m P_{m-1} \cdots P_{i+1} \big[\underbrace{0, \ldots, 0}_{i-1}, s_1^{(i-1)}\big]^\intercal$$

und

$$S^{(i)} := S^{(i-1)} - \frac{1}{s_{1,1}^{(i-1)}} s_1^{(i-1)} \big(s_1^{(i-1)}\big)^\intercal$$

für $i = 1, 2, \ldots, m$. Um zur endgültigen Darstellung für A zu gelangen, multiplizieren wir (4.9) von beiden Seiten mit den Permutationsmatrizen und erhalten

$$A = \sum_{i=1}^{m} l_i l_i^\intercal + E_m \tag{4.10}$$

mit

$$l_i := \frac{1}{\sqrt{s_{1,1}^{(i-1)}}} P_1 P_2 \cdots P_i \big[0, \ldots, 0, s_1^{(i-1)}\big]^\intercal \qquad \text{für } i = 1, \ldots, m$$

und

$$E_m := P_1 P_2 \cdots P_m \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & S^{(m)} \end{array}\right] P_m \cdots P_2 P_1.$$

Setzen wir zur Vereinfachung noch $L_m := [l_1, \ldots, l_m]$, dann besitzt A die Darstellung

$$A = L_m L_m^\intercal + E_m.$$

4.43 Satz. *Gilt* $\operatorname{rank}(A) = k < n$*, so bricht die Iteration nach* k *Schritten mit* $E_k = 0$ *ab und die Zerlegung ist exakt.*

Beweis. Wir verwenden die Darstellung (4.9) und setzen $\hat{L}_i := [\hat{l}_1, \ldots, \hat{l}_i]$, solange $s_{1,1}^{(i)} > 0$ gilt. Dann besitzt $\hat{L}_i$ untere Dreiecksgestalt mit positiven Diagonalelementen und somit vollen Spaltenrang i. Nach Konstruktion des Verfahrens wird der Rang von A in jedem Schritt um 1 reduziert. Das bedeutet $s_{i,i}^{(k)} = 0$ für $i = 1, \ldots, n-k$, wegen $\operatorname{rank}(A) = k$. Aufgrund der positiven Semidefinitheit muss dann $S^{(k)}$ bereits die Nullmatrix sein, woraus auch $E_k = 0$ folgt. □

Bemerkung. Besitzt A vollen Rang und werden die Permutationen $P_i = I$ für $i = 1, \ldots, n$ gewählt, liefert das Verfahren die gewöhnliche Cholesky-Zerlegung von A.

Wie eingangs erklärt, wollen wir die pivotisierte Cholesky-Zerlegung verwenden, um eine Niedrigrangapproximation der Matrix A zu berechnen. Der Fehler soll in der Spurnorm gemessen werden. Wegen

$$\operatorname{trace}(S) = \operatorname{trace}(A) - \frac{1}{a_{1,1}} \|a_1\|_2^2 ,$$

ist die optimale Pivotisierungsstrategie genau dasjenige Diagonalelement $a_{i,i}$ als Pivotelement zu verwenden mit

$$i = \operatorname*{argmax}_{i=1,\ldots,n} \frac{1}{a_{i,i}} \|a_i\|_2^2 .$$

Hierbei ist zu beachten, dass nur symmetrische Permutationen verwendet werden können, da sonst die Symmetrie der Matrix verloren ginge. Damit ist diese Pivotisierungsstrategie für die Cholesky-Zerlegung in der Tat optimal. Um das Maximierungsproblem zu lösen, müssen jedoch alle Einträge der Matrix A bekannt sein. Um dies zu vermeiden, wählen wir eine andere Strategie, wir eliminieren das betragsgrößte Element der Matrix. Nach Satz 4.42 befindet sich dieses immer auf der Diagonalen von A (vgl. [16]). Damit gelangen wir zu folgendem Algorithmus.

Algorithmus. Die pivotisierte Cholesky-Zerlegung
input: $A \in \mathbb{R}^{n \times n}$, $\varepsilon > 0$
output: $A_m = \sum_{i=1^m} l_i l_i^{\mathsf{T}}$ mit $\|A - A_m\|_{\mathrm{tr}} \leq \varepsilon$
1: initialisiere $\pi := [1, \ldots, n]$, $d = [a_{1,1}, \ldots, a_{n,n}]$, $tr := \|d\|_1$, $m := 1$
2: **while** $tr > \varepsilon$ **do**
3: $\quad i_{max} := \operatorname{argmax}\{d_{\pi_i} \mid i = m, \ldots, n\}$

4: vertausche π_m und $\pi_{i_{\max}}$
5: $l_{m,\pi_m} := \sqrt{d_{\pi_m}}$
6: **for** $i = m+1, \ldots, n$ **do**
7: $\quad l_{m,\pi_i} := a_{\pi_i,\pi_m} / l_{m,\pi_m}$
8: **for** $k = 1, \ldots, m$ **do**
9: $\quad$ **for** $i = m+1, \ldots, n$ **do**
10: $\quad\quad l_{m,\pi_i} := l_{m,\pi_i} - l_{k,\pi_i} * l_{k,\pi_m} / l_{m,\pi_m}$
11: $tr := 0$
12: **for** $i = m+1, \ldots, n$ **do**
13: $\quad d_{\pi_i} := d_{\pi_i} - (l_{i,\pi_i})^2$
14: $\quad tr := tr + d_{\pi_i}$
15: $m := m + 1$

Für die Komplexität des Algorithmus gilt der nächste Satz.

4.44 Satz ([16]). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, dann ist der Aufwand für die Berechnung von m Schritten der pivotisierten Cholesky-Zerlegung $\mathcal{O}(m^2 n)$.*

Falls die Eigenwerte von A hinreichend schnell exponentiell abklingen, ist es möglich mit Hilfe der pivotisierten Cholesky-Zerlegung eine Rang-m Approximation zu bestimmen, die exponentiell konvergiert (vgl. [16]).

4.45 Satz. *Gilt für die Eigenwerte $\lambda_1, \ldots, \lambda_n$ von A*

$$4^m \lambda_m \leq C e^{-bm} \qquad \text{für } m = 1, \ldots, n$$

mit Konstanten $b, C > 0$, dann liefert die pivotisierte Cholesky-Zerlegung eine Approximation A_m mit $\operatorname{rank}(A_m) = m \sim |\log(\varepsilon/n)|$ und $\|A - A_m\|_{\mathrm{tr}} \leq C\varepsilon$ für $\varepsilon \to 0$.

Beweis. Ohne Beschränkung der Allgemeinheit nehmen wir an, dass A so permutiert ist, dass das k-te Pivotelement sich in der (k,k)-Position befindet. Dann besitzt $L_m \in \mathbb{R}^{n\times m}$ untere Dreiecksgestalt. Wir setzen die folgenden Partitionen für die Matrizen $A = LL^\intercal$ und L_m an

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \qquad L_m = \begin{bmatrix} L_{1,1} & 0 \\ L_{2,1} & 0 \end{bmatrix}.$$

Wegen

$$A_m = L_m L_m^\intercal = \begin{bmatrix} L_{1,1}L_{1,1}^\intercal & L_{1,1}L_{2,1}^\intercal \\ L_{2,1}L_{1,1}^\intercal & L_{2,1}L_{2,1}^\intercal \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & L_{2,1}L_{2,1}^\intercal \end{bmatrix}$$

ist $L_1 L_1^\intercal \in \mathbb{R}^{m\times m}$ bereits die pivotisierte Cholesky-Zerlegung von $A_{1,1} \in \mathbb{R}^{m\times m}$. Dies bedeutet, dass

$$\frac{1}{\lambda_m(A_{1,1})} = \left\|A_{1,1}^{-1}\right\|_2 = \left\|L_{1,1}^{-1}\right\|_2^2,$$

wobei $\lambda_m(A_{1,1})$ den kleinsten Eigenwert von $A_{1,1}$ bezeichnet. Nach [18] gilt die Abschätzung

$$\left\|L_{1,1}^{-1}\right\|_2 \leq \frac{\sqrt{4^m + 6m - 1}}{3l_{m,m}}. \tag{4.11}$$

Demnach können wir abschätzen

$$\frac{1}{\lambda_m(A_{1,1})} \leq \frac{4^m + 6m - 1}{9l_{m,m}^2} \leq \frac{4^m}{l_{m,m}^2}.$$

Andererseits gilt offensichtlich

$$\|A - A_m\|_{\mathrm{tr}} \leq (n-m) l_{m,m}^2.$$

Das bedeutet

$$\|A - A_m\|_{\mathrm{tr}} \leq (n-m) 4^m \lambda_m(A_{1,1}) \leq n 4^m \lambda_m(A_{1,1}).$$

Da $A_{1,1}$ genau die orthogonale Projektion von A auf den Unterraum $\mathrm{span}\{e_1, \ldots, e_m\}$ ist, gilt $\lambda_m(A_{1,1}) \leq \lambda_m$ nach Satz 3.6 und daher ist

$$\|A - A_m\|_{\mathrm{tr}} \leq N 4^m \lambda_m \leq nCe^{-bm}.$$

Um einen vorgegebenen Fehler mit $\|A - A_m\|_{\mathrm{tr}} \leq \varepsilon$ zu erreichen, muss

$$\varepsilon \leq nCe^{-bm}$$

gelten, woraus direkt

$$\log\left(\frac{\varepsilon}{nC}\right) \leq -bm$$

folgt. Dies bedeutet

$$\left|\log\left(\frac{\varepsilon}{nC}\right)\right| \geq bm.$$

Folglich existiert ein $\tilde{C} > 0$ mit

$$\tilde{C}\left|\log\left(\frac{\varepsilon}{n}\right)\right| \geq m$$

und die Behauptung ist gezeigt. □

Nach [18] ist die Abschätzung (4.11) scharf und der Faktor 4^m kann nicht vermieden werden (vgl. [16]). Andererseits werden wir bei den numerischen Resultaten sehen, dass gerade im Fall des Beispieles aus [18], wo der Faktor 4^m auftaucht, die Konvergenz dennoch sehr schnell ist. Die numerischen Beispiele in [16], die auch hier noch einmal vorgestellt werden, lassen sogar vermuten, dass die Approximation mit Hilfe der pivotisierten Cholesky-Zerlegung immer in der Hinsicht optimal konvergiert, dass der Rang m der pivotisierten Cholesky-Zerlegung proportional zur Anzahl der Terme in der abgebrochenen Singulärwertzerlegung mit der selben Approximationsgüte ist (vgl. [16]). Die Essenz ist, dass der Approximationsfehler durch die Spurnorm kontrollierbar ist und die pivotisierte Cholesky-Zerlegung im Fall $m \ll n$, auch wenn dies erst aposteriori prüfbar ist, eine verwertbare Niedrigrangapproximation liefert.
Wir werden jetzt sehen, wie sich die Niedrigrangapproximation nutzen lässt, um eine schnelle Methode zur Eigenwertberechnung zu gewinnen. Wir nehmen dazu an, wir haben durch die pivotisierte Cholesky-Zerlegung eine Niedrigrangapproximation $A_m = L_m L_m^\intercal \in \mathbb{R}^{n\times n}$ mit $\operatorname{rank}(A_m) = m$ von A und

$$\|A - A_m\| \leq \varepsilon$$

gefunden. Wir betrachten nun das algebraische Eigenwertproblem

$$Ax = \lambda M x. \tag{4.12}$$

Einsetzen der Niedrigrangapproximation von A führt dann auf

$$A_m x = \tilde{\lambda} M x. \tag{4.13}$$

Da M symmetrisch und positiv definit ist, existiert die Wurzel $X^2 = M$. Multiplizieren wir (4.13) von rechts und links mit X liefert dies

$$X^{-1} L_m L_m^\intercal X^{-1} y = \tilde{\lambda} y, \qquad y = X^{-1} x. \tag{4.14}$$

Nun besitzt

$$(X^{-1}L_m)(X^{-1}L_m)^\intercal$$

dieselben nicht verschwindenden Eigenwerte, wie

$$(X^{-1}L_m)^\intercal (X^{-1}L_m).$$

Also können wir das Eigenwertproblem (4.14) ersetzen durch

$$L_m^\intercal M^{-1} L_m y = \tilde{\lambda} y, \qquad x = M^{-1} L_m y. \tag{4.15}$$

Das heißt, wir haben das ursprüngliche Eigenwertproblem (4.13) der Dimension n in ein äquivalentes Eigenwertproblem der Dimension m transformiert. Man beachte, dass die Berechnung der inversen Massenmatrix bei allen in dieser Arbeit vorgestellten Verfahren zur Eigenwertberechnung durchgeführt werden muss und über die iterative Lösung eines linearen Gleichungssystems erfolgen kann.
Mit Hilfe des Satzes von Bauer-Fike gelangen wir zu folgender Fehlerabschätzung bezüglich der Eigenwerte des ursprünglichen Problems (vgl. [16]). Es gilt

$$\begin{aligned} \left|\lambda_i - \tilde{\lambda}_i\right| &\leq \left\|X^{-1}(A - A_m)X^{-1}\right\|_2 \\ &\leq \left\|M^{-1}\right\|_2 \|A - A_m\|_2 \leq \left\|M^{-1}\right\|_2 \varepsilon \qquad \text{für } i = 1, \ldots, m. \end{aligned}$$

Dies bedeutet, der Fehler in den Eigenwerten hängt nur von dem Approximationsfehler durch die pivotisierte Cholesky-Zerlegung und der Kondition der Massenmatrix ab. Der Aufwand zur Bestimmung von m Eigenwerten hat sich dabei von $\mathcal{O}(n^2 m)$ auf $\mathcal{O}(nm^2)$ reduziert. Die Lösung des kleinen Eigenwertproblems (4.15) erfolgt mit Hilfe des QR-Verfahrens.

4.3 Numerische Resultate

4.3.1 Das Cholesky-Verfahren

Für das Cholesky-Verfahren wurden insgesamt fünf Beispiele gerechnet. Die ersten vier Beispiele entstammen hierbei aus [16]. Das fünfte Beispiel findet man in [19]. Für jede der ausgewählten Funktionen, beziehungsweise Matrizen ist zunächst das Abklingverhalten der Spur bei der pivotisierten Cholesky-Zerlegung tabellarisch dargestellt. Danach sind für die Beispiele noch die größten Eigenwerte der Matrix A angegeben. Dabei entspricht die Anzahl der berechneten Eigenwerte genau dem ermittelten Rang aus der pivotisierten Cholesky-Zerlegung. Die Matrix A entstammt in den ersten drei Beispielen jeweils genau den Auswertungen der Kernfunktion $f(x, y)$ an den n^2 äquidistanten Gitterpunkten des Einheitsquadrats $[0, 1] \times [0, 1]$. In den anderen beiden Beispielen ist die Matrix A explizit angegeben. Durch die, im Folgenden ausgeführte Galerkin-Diskretisierung für das Eigenwertproblem des Hilbert-Schmidt-Operators

$$T\psi(x) = \int_{[0,1]} f(x, y)\psi(y)\mathrm{d}y$$

mit $\psi \in L^2([0, 1])$, ergibt sich ein verallgemeinertes Eigenwertproblem. Dieses wird für die ersten vier Beispiele gelöst.

Für die Diskretisierung wählen wir n stückweise lineare Ansatzfunktionen mit kompaktem Träger auf dem Intervall $[0,1]$ mit $h := 1/(n-1)$ und $x_i := ih$, wie folgt:

$$\begin{aligned}
\phi_0(x) &:= -\frac{2}{h}(x - x_1), \qquad x_0 \le x \le x_1 \\
\phi_{n-1}(x) &:= \frac{2}{h}(x - x_{n-2}), \qquad x_{n-2} \le x \le x_{n-1} \\
\phi_i(x) &:= \begin{cases} \frac{1}{h}(x - x_{i-1}), & x_{i-1} \le x \le x_i \\ -\frac{1}{h}(x - x_{i+1}), & x_i < x \le x_{i+1} \end{cases}, \text{ für } i = 1, \ldots, n-2.
\end{aligned}$$

Hierbei sind ϕ_0 und ϕ_{n-1} doppelt so hoch wie die anderen Ansatzfunktionen gewählt, damit alle Funktionen die selbe L_1-Norm haben. Es gilt

$$\|\phi_i\|_{L^1([0,1])} = \int_{\text{supp}\,(\phi_i)} |\phi_i(x)|\,\mathrm{d}x = h, \qquad \text{für } i = 0, \ldots, n-1.$$

In Abbildung 4.1 sind die Funktionen für $n = 10$ dargestellt. Das Eigenwert-

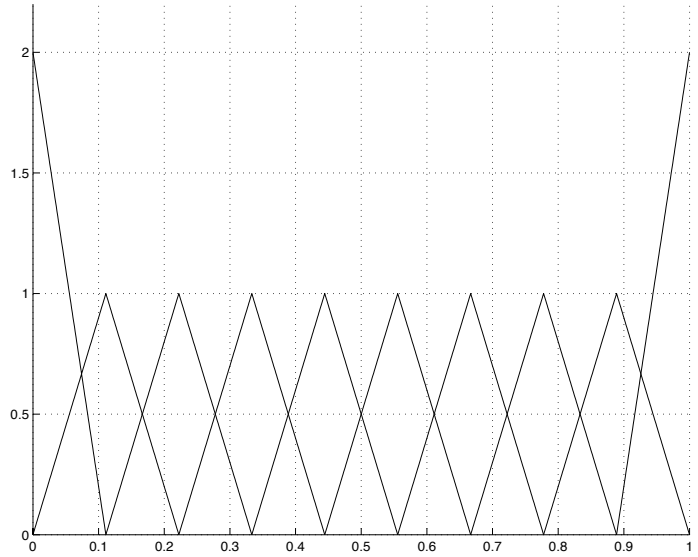

Abbildung 4.1: Stückweise lineare Ansatzfunktionen auf $[0,1]$

problem für den Operator T lautet

$$Tu = \lambda u.$$

Das ist in $L^2([0,1])$ äquivalent zu

$$(Tu, v)_{L^2([0,1])} = \lambda (u, v)_{L^2([0,1])}, \qquad \text{für alle } v \in L^2([0,1]).$$

Lösen wir dieses Problem nun im Teilraum $\operatorname{span}\{\phi_0, \ldots, \phi_{n-1}\}$ der Dimension n, gelangen wir auf das algebraische Eigenwertproblem

$$\left[(T\phi_i, \phi_j)_{L^2([0,1])}\right]_{i,j=0,\ldots,n-1} = \lambda \left[(\phi_i, \phi_j)_{L^2([0,1])}\right]_{i,j=0,\ldots,n-1}.$$

Für die nun folgenden Beispiele III und IV wurde die Dimension $n = 10^5$ und für die Beispiele I, II, und V die Dimension $n = 10^6$ gewählt.

Beispiel I: Der Gauß-Kern

Das erste Beispiel, das wir betrachten wollen, ist die Gauß'sche Kernfunktion. Diese ist gegeben durch

$$f(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-y)^2}{\sigma^2}\right).$$

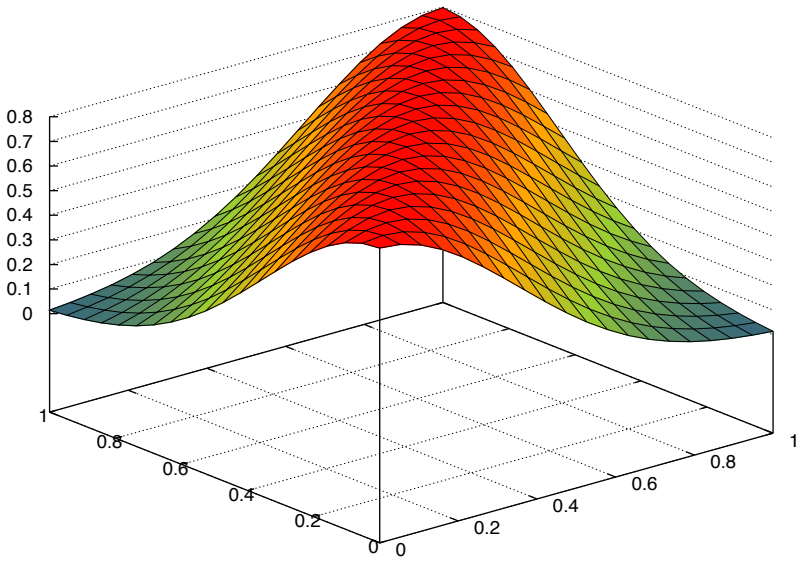

Abbildung 4.2: Gauß-Kern

Abbildung 4.2 zeigt einen Plot der Funktion f für $n = 20$. Da die Funktion auf ganz $(0,1)\times(0,1)$ analytisch ist, klingen nach [27] die Eigenwerte exponentiell ab. Zu beachten ist aber, dass die Voraussetzungen des Konvergenzsatzes nur im Fall $\sigma \geq 0,1$ erfüllt sind (vgl. [16]).

ε \ σ	1	0.5	0.1	0.05	0.01
10^{-1}	2	3	10	19	89
10^{-2}	3	5	15	28	137
10^{-3}	4	5	19	36	173
10^{-4}	5	6	21	39	187
10^{-5}	5	7	24	46	214
10^{-6}	5	8	27	50	238

Tabelle 4.1: Gauß-Kern: Rang mit $\|A - A_m\|_{\mathrm{tr}} \leq \varepsilon$ für $n = 10^6$

Tabelle 4.1 stellt die Anzahl der benötigten Cholesky-Vektoren und somit den Rang von L_m dar, um die in der ersten Spalte aufgetragene relative Genauigkeit der Approximation zu erreichen. Der Fehler ist hierbei in der Spurnorm gemessen.

Wie man sieht ist auch im Fall $\varepsilon = 10^{-6}$ noch $m \ll n$, so dass sich das Eigenwertproblem für den Hilbert-Schmidt-Operator effizient lösen lässt.

Abbildung 4.3 zeigt die Eigenwerte des diskretisierten Operators (blaue Kreise) und das zugehörige Abklingen der Spur (grüne Kreuze). Ferner ist das Verhältnis der Spur des jeweiligen Schur-Komplements zu den Eigenwerten dargestellt (rote Quadrate).

Der Fehler in den Eigenwerten wird durch die Spurnorm kontrolliert und diese beträgt hier $\varepsilon = 10^{-6}$. Daher stimmen die dargestellten Eigenwerte bis auf einen Fehler von 10^{-6} mit denen des Hilbert-Schmidt-Operators überein. Das bedeutet wiederum, dass die Spur hier dasselbe Abklingverhalten aufweist wie die Eigenwerte des Hilbert-Schmidt-Operators. Dies liegt an der Äquivalenz der Spurnorm und der ℓ_2-Norm im Falle exponentiell abklingender Eigenwerte. Dies lässt schließen, dass die pivotisierte Cholesky-Zerlegung hier optimal konvergiert (vgl. [16]).

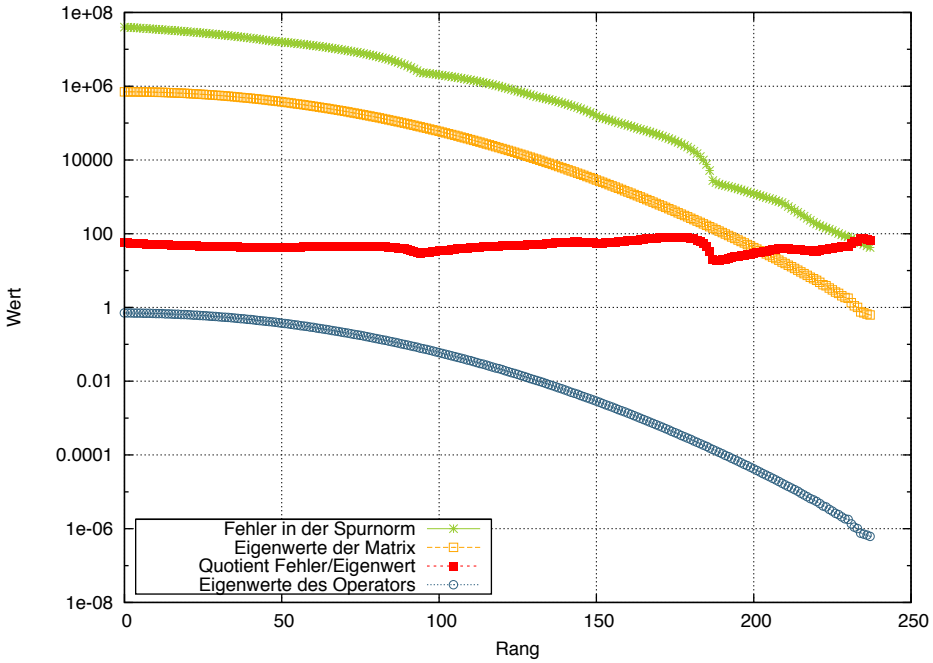

Abbildung 4.3: Eigenwerte

Beispiel II: Der unstetige Gauß-Kern

Als zweites Beispiel betrachten wir wieder den Gauß-Kern, der allerdings durch eine Skalierungsfunktion mit Sprüngen versehen wurde. Wir setzen also

$$
\begin{aligned}
f(x,y) &= \frac{1}{a(x,y)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-y)^2}{\sigma^2}\right), \\
a(x,y) &:= \begin{cases} 4, & x,y < 0,5 \\ 1, & x,y \geq 0,5 \\ -2, & \text{sonst.} \end{cases}
\end{aligned}
$$

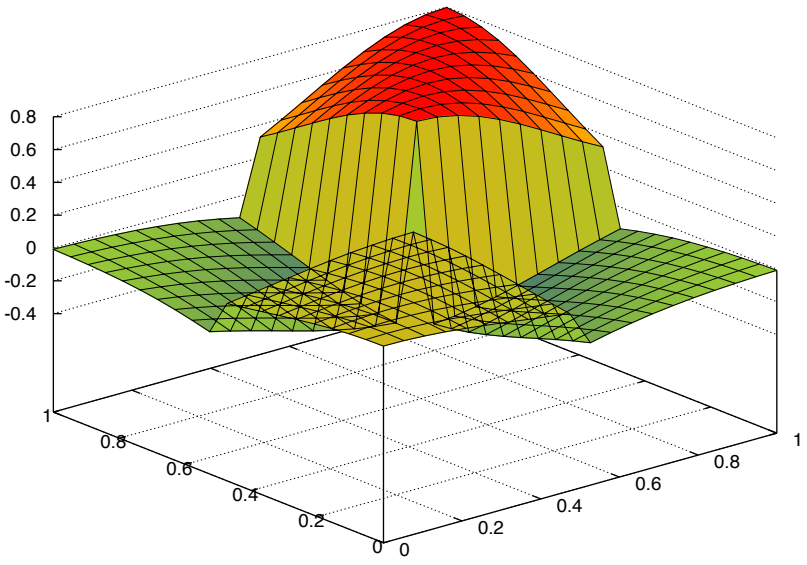

Abbildung 4.4: unstetiger Gauß-Kern

Die Abbildung 4.4 zeigt einen Plot der Kernfunktion für $n = 20$.
Nach [27] klingen auch hier die Eigenwerte exponentiell ab. In diesem Kontext ist die Tabelle 4.2 zu verstehen, die trotz der Sprünge, annähernd die gleichen Konvergenzraten wie im vorangegangenen Beispiel zeigt. (vgl. [16]).

ε \ σ	1	0.5	0.1	0.05	0.01
10^{-1}	2	3	10	17	81
10^{-2}	3	4	15	28	131
10^{-3}	4	5	18	34	168
10^{-4}	4	6	21	39	186
10^{-5}	5	7	24	45	211
10^{-6}	5	8	26	50	234

Tabelle 4.2: unstetiger Gauß-Kern: Rang mit $\|A - A_m\|_{\mathrm{tr}} \leq \varepsilon$ für $n = 10^6$

Auch bei den in Abbildung 4.5 dargestellten Eigenwerten ist das Abklingverhalten von Spur und Eigenwerten wie im vorigen Beispiel zu erkennen. Somit hängt die Approximationsgüte der Cholsky-Zerlegung nicht von der Glattheit der Kernfunktion ab (vgl. [16]).

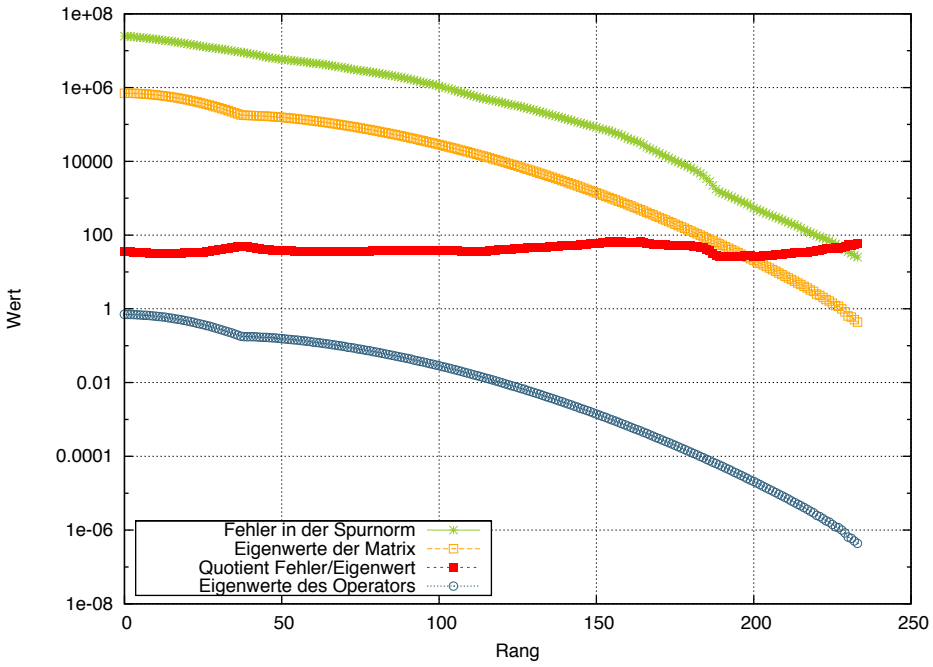

Abbildung 4.5: Eigenwerte

Beispiel III: Der Poisson-Kern

Für dieses Beispiel wählen wir den Poisson-Kern

$$f(x, y) = \sigma \exp(-\sigma |x - y|).$$

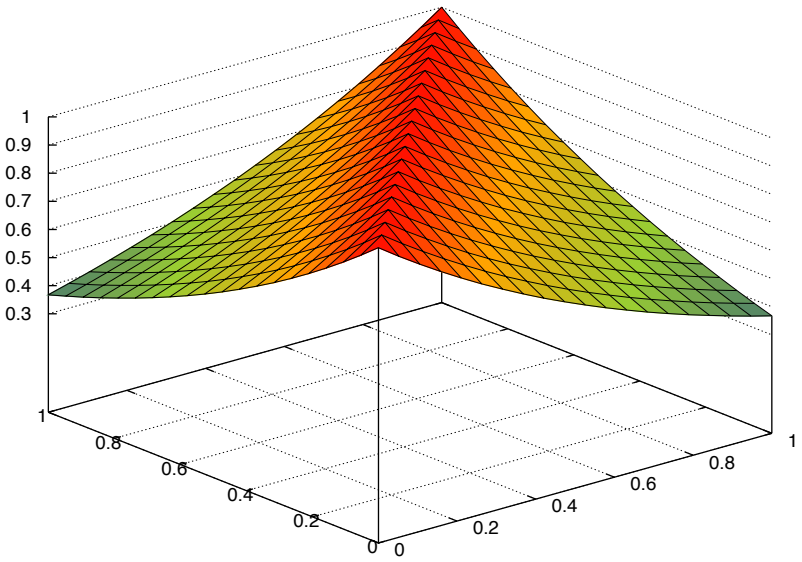

Abbildung 4.6: Poisson-Kern

In Abbildung 4.6 ist dieser für $n = 20$ geplottet. Wie man sieht, hat der Poisson-Kern einen Knick entlang der Diagonalen und ist somit nicht mehr analytisch. In der Tat verhalten sich die Eigenwerte hier nur noch wie $\lambda_k \sim k^{-2}$ (vgl. [16]). Damit sind die Voraussetzungen des Konvergenzsatzes nicht mehr erfüllt, dennoch ist aus Tabelle 4.3 ein gewisses Abklingverhalten abzulesen.

ε \ σ	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}
10^{-1}	5	1	1	1	1
10^{-2}	36	5	1	1	1
10^{-3}	376	36	5	1	1
10^{-4}	3616	376	36	5	1

Tabelle 4.3: Poisson-Kern: Rang mit $\|A - A_m\|_{\mathrm{tr}} \leq \varepsilon$ für $n = 10^5$

Wie bereits bemerkt, klingen die Eigenwerte in diesem Beispiel nur quadratisch ab. Das bedeutet, dass die Spur höchstens linear abklingen kann. Eben dieses Abklingverhalten lässt sich aus Abbildung 4.7 ablesen.

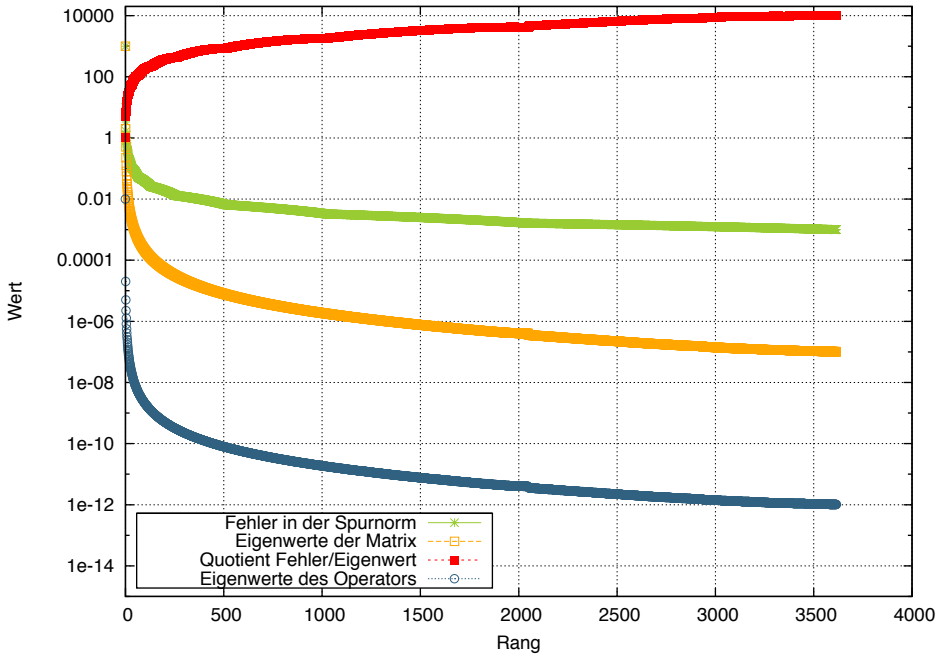

Abbildung 4.7: Eigenwerte

Beispiel IV: Die Zufallsmatrix

Für dieses Beispiel wird eine Matrix mit exponentiell abklingenden Eigenwerten zufällig generiert. Hierzu benutzen wir die Darstellung

$$A = \sum_{i=1}^{m} \lambda_i v_i v_i^\mathsf{T}, \quad \lambda_i = \exp(-\sigma i), \quad v_i^\mathsf{T} v_j = \delta_{i,j}.$$

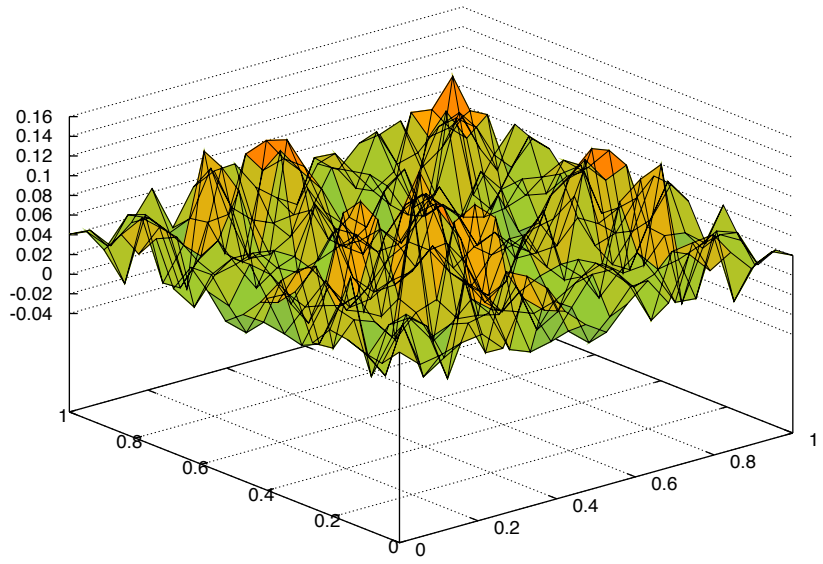

Abbildung 4.8: Zufalls-Kern

Die Vektoren v_i für $i = 1, \ldots, m$ werden mit einem Zufallsgenerator initialisiert und dann mit dem Gram-Schmidt-Verfahren aus Unterabschnitt 4.2.3 orthonormalisiert. Das Beispiel soll untermauern, dass die Konvergenz der pivotisierten Cholesky-Zerlegung nur vom Abklingen der Eigenwerte abhängt. Abbildung 4.8 stellt die Einträge der Matrix analog zu den vorigen Beispielen als Auswertung einer Kernfunktion an den Gitterpunkten dar. Die Funktionswerte sind hierbei durch die Matrixeinträge gegeben.
Das Beispiel wird mit $m = 2000$ Vektoren gerechnet. Dabei ergeben sich in Abhängigkeit von σ die in Tabelle 4.4 dargestellten Ränge für die Approximation.

ε \ σ	1	0.5	0.1	0.05	0.01
10^{-1}	3	6	29	61	333
10^{-2}	6	11	56	115	610
10^{-3}	8	15	81	167	873
10^{-4}	10	21	106	216	1126
10^{-5}	13	25	130	266	1375
10^{-6}	15	30	154	315	1618

Tabelle 4.4: Zufallsmatrix: Rang mit $\|A - A_m\|_{\mathrm{tr}} \leq \varepsilon$ für $n = 10^5$

In Abbildung 4.9 erkennt man, dass die Spur der Matrix fast das gleiche Abklingverhalten aufweist wie die Eigenwerte.

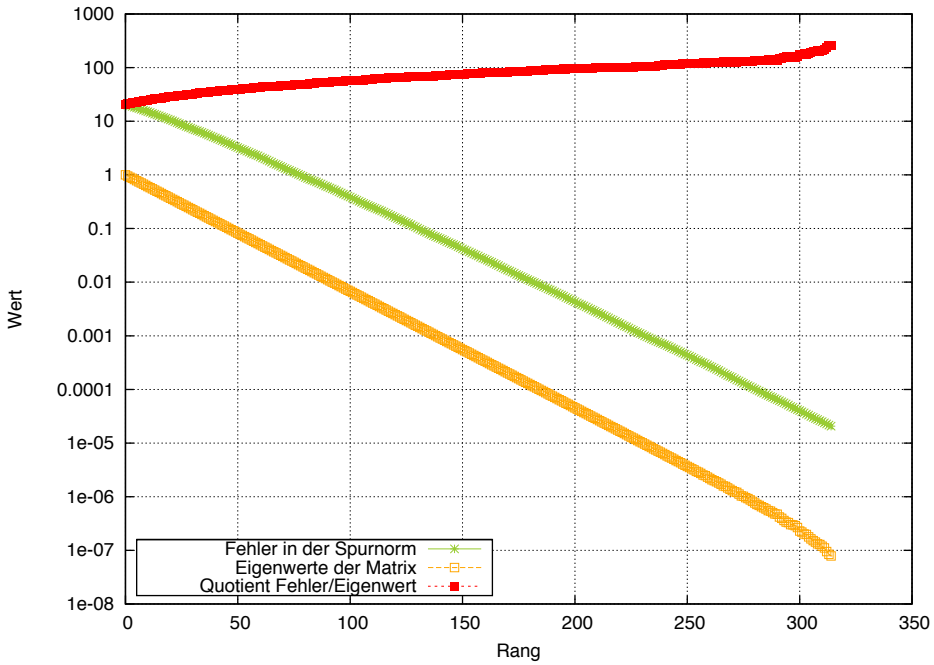

Abbildung 4.9: Eigenwerte

Beispiel V: Die Kahan-Matrix

Das folgende Beispiel stammt von W. Kahan (vgl. [18, 19]). Hier taucht der Wachstumsfaktor 4^m aus dem Konvergenzbeweis explizit auf (vgl. [18, 19]). Dennoch werden wir sehen, dass die Konvergenzgeschwindigkeit gut ist. Sei

$$\begin{aligned} A &= T(\sigma)^\intercal T(\sigma), \\ T(\sigma) &:= \operatorname{diag}[1, s, s^2, \ldots, s^{n-1}] \begin{bmatrix} 1 & -c & \ldots & -c \\ & \ddots & \ddots & \vdots \\ & & \ddots & -c \\ & & & 1 \end{bmatrix}, \\ & s := \sin(\sigma), \quad c := \cos(\sigma) \end{aligned}$$

(vgl. [18]).

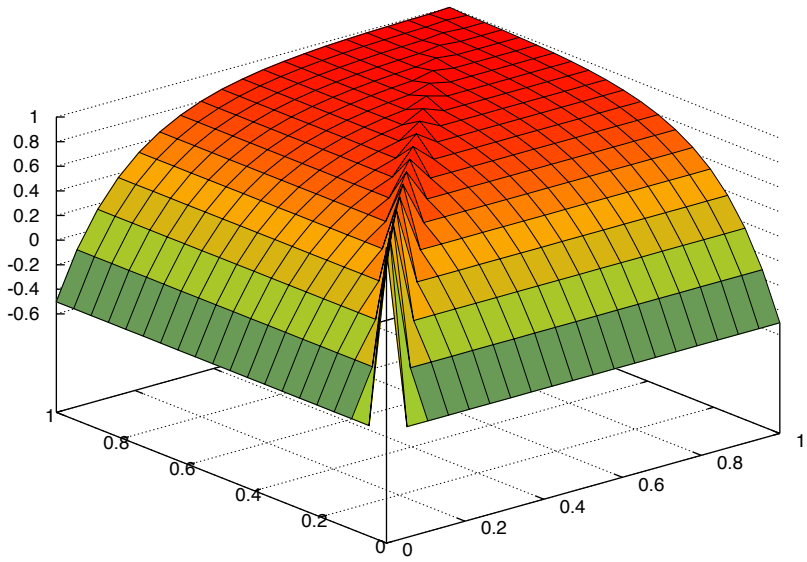

Abbildung 4.10: Matrix A für $\sigma = \pi/3$

Die Abbildung 4.10 zeigt den Plot der Matrix A, mit derselben Interpretation wie im vorangegangenen Beispiel.
Für die Implementierung dieses Beispiels benötigen wir einen

4.46 Satz. *Für die Einträge der Matrix A gilt*

$$a_{i,j} = \begin{cases} 1, & i = j \\ 1 - (1+c)s^{2(i-1)}, & i < j \\ a_{j,i}, & i > j. \end{cases}$$

Beweis. Der Beweis erfolgt induktiv. Für $i = 1$ ist die Aussage für das Diagonalelement klar. Sei die Aussage für $i > 1$ erfüllt. Es gilt

$$\begin{aligned} a_{i,i} &= [-c, -cs, \ldots, -cs^{i-2}, s^{i-1}][-c, -cs, \ldots, -cs^{i-2}, s^{i-1}]^\intercal \\ &= c^2 + c^2 s^2 + \ldots + c^2 s^{2(i-2)} + s^{2(i-1)}, \end{aligned}$$

wobei die $n - i$ Nulleinträge der Vektoren weggelassen sind. Dann folgt

$$\begin{aligned} a_{i+1,i+1} &= c^2 + c^2 s^2 + \ldots + c^2 s^{2(i-1)} + s^{2i} \\ &= c^2 + c^2 s^2 + \ldots + c^2 s^{2(i-2)} + s^{2(i-1)} \underbrace{(c^2 + s^2)}_{=1} \overset{\text{Ind. Vor.}}{=} a_{i,i}. \end{aligned}$$

Sei nun $i < j$. Es gilt

$$\begin{aligned} a_{i,j} &= [-c, -cs, \ldots, -cs^{i-2}, s^{i-1}, 0, \ldots, 0] \\ &\quad \cdot [-c, -cs, \ldots, -cs^{j-2}, s^{j-1}, 0, \ldots, 0]^\intercal \\ &= c^2 + c^2 s^2 + \ldots + c^2 s^{2(i-2)} - cs^{2(i-1)} \\ &= a_{i,i} - s^{2(i-1)} - cs^{2(i-1)} = 1 - (1+c)s^{2(i-1)}, \end{aligned}$$

Nach Konstruktion ist A symmetrisch, woraus schließlich $a_{i,j} = a_{j,i}$ für $i, j = 1, \ldots, n$ folgt. □

Mit Hilfe des Satzes ist es möglich die Einträge von A numerisch stabil zu berechnen. Ferner muss die Matrix nicht mehr explizit aufgestellt werden.

ε \ σ	0	$\pi/4$	$\pi/3$	$2\pi/5$	$\pi/2$
10^{-1}	1	4	9	23	900001
10^{-2}	1	7	17	46	990001
10^{-3}	1	10	25	76	999001
10^{-4}	1	14	35	99	999901
10^{-5}	1	17	44	122	999991
10^{-6}	1	21	52	145	1000000

Tabelle 4.5: $\|A\|_{\text{tr}} \leq \varepsilon$ für $n = 10^6$.

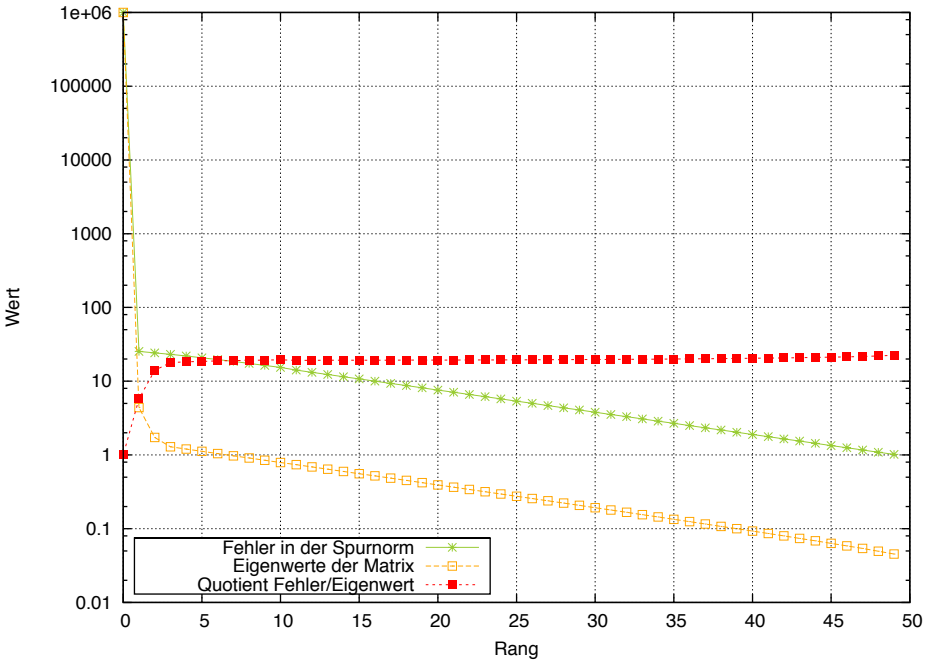

Abbildung 4.11: Eigenwerte

Man kann erkennen, dass auch hier die Spur optimal abklingt. Allerdings ist zu bemerken, dass für $\sigma \to \pi/2$ die Matrix gegen die Einheitsmatrix konvergiert. In diesem Grenzfall sind dann alle Eigenwerte gleich 1 und das Abklingverhalten ist nur noch linear mit einer Reduktion um 1 in jedem Schritt. Deshalb wird die volle Cholesky-Zerlegung berechnet. Bei meinen numerischen Tests ergeben sich Instabilitäten für Winkel $\sigma \in [5\pi/12, \pi/2)$. Hier bricht die pivotisierte Cholesky-Zerlegung mit einer negativen Spur ab. Eine Erklärung hierzu liefert der folgende

4.47 Satz. *Für die Einträge von $A^{(k)}$, resultierend aus k Schritten der pivotisierten Cholesky-Zerlegung, angewendet auf die Matrix A, gilt*

$$a_{i,j}^{(k)} = a_{i,j} - c^2 \sum_{l=0}^{k-1} s^{2l} \qquad \text{für } i,j = k+1,\dots n$$

und $a_{i,j}^{(k)} = 0$ sonst. Die Matrix $[a_{i+k,j+k}^{(k)}]_{i,j=1,\dots,n-k}$ entspricht hierbei genau dem Schur-Komplement im k-ten Schritt.

Beweis. Der Beweis erfolgt per vollständiger Induktion über k. Für $k = 0$ ist

die Aussage klar. Sei die Behauptung für $k > 0$ erfüllt. Dann gilt

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \frac{a_{i,k+1}^{(k)} a_{k+1,j}^{(k)}}{a_{k+1,k+1}^{(k)}} = a_{i,j}^{(k)} - \frac{\left(a_{k+1,j}^{(k)}\right)^2}{a_{k+1,k+1}^{(k)}}.$$

Einsetzen der Induktionsvoraussetzung liefert nun

$$a_{i,j}^{(k+1)} = a_{i,j} - c^2 \sum_{l=0}^{k-1} s^{2l} - \frac{\left(a_{k+1,j} - c^2 \sum_{l=0}^{k-1} s^{2l}\right)^2}{a_{k+1,k+1} - c^2 \sum_{l=0}^{k-1} s^{2l}}.$$

Es gilt $k + 1 < j$. Dann liefert das Einsetzen der Darstellungen aus dem Beweis von Satz 4.46

$$\frac{\left(a_{k+1,j} - c^2 \sum_{l=0}^{k-1} s^{2l}\right)^2}{a_{k+1,k+1} - c^2 \sum_{l=0}^{k-1} s^{2l}} = \frac{(-cs^{2k})^2}{s^{2k}} = c^2 s^{2k}$$

und damit die Behauptung. □

Der Satz impliziert zusammen mit dem Beweis von Satz 4.46, dass -in exakter Arithmetik- die Diagonalelemente einer Iterierten immer gleich groß sind. Dies haben auch meine numerischen Tests ergeben. Dadurch erfolgt keine Pivotisierung (vgl. [19]). Dieses Verhalten steht allerdings in keinem Widerspruch zu Satz 4.45.
Der Beweis von Satz 4.46 suggeriert, dass die letzte Zeile / Spalte von A die maximale euklidische Norm besitzt. Beginnt man die Cholesky-Zerlegung mit $a_{n,n}$ als Pivot, so vermeidet man in diesem Beispiel die auftretende Instabilität. Das Beispiel zeigt, dass es problematisch sein kann, eine Pivotsuche nur bezüglich des größten Matrixeintrages durchzuführen.

4.3.2 Unterraumiteration und ARPACK

Als Modellproblem wird das Eigenwertproblem für das Einfachschichtpotential auf der Oberfläche $S_2 := \{x \in \mathbb{R}^3 \mid \|x\|_2 = 1\}$ der Einheitskugel $D_3 := \{x \in \mathbb{R}^3 \mid \|x\|_2 < 1\}$ betrachtet. Ausgehend von der Laplace-Gleichung

$$\Delta u = 0 \qquad \text{für } x \in D_3 \tag{4.16}$$

kann man mit Hilfe der Fundamentallösung des Laplace-Operators

$$U^*(x,y) = \frac{1}{4\pi \|x-y\|_2} \tag{4.17}$$

und des Einfachschichtpotentialansatzes eine Lösung der Laplace-Gleichung angeben. Es gilt

$$u(\tilde{x}) = \int_{S_2} U^*(\tilde{x},y) w(y) \mathrm{d}\sigma_y \qquad \text{für } \tilde{x} \in D_3, \tag{4.18}$$

wobei die unbekannte Dichte $w \in H^{-1/2}(S_2)$ beispielsweise durch die Dirichlet-Randbedingung

$$u(x) = g(x) \qquad \text{für } x \in S_2 \tag{4.19}$$

mit $g \in H^{1/2}(S_2)$ berechnet werden kann. Dies führt dann auf die Randintegralgleichung

$$\int_{S_2} U^*(x,y) w(y) \mathrm{d}\sigma_y = g(x) \qquad \text{für } x \in S_2 \tag{4.20}$$

(vgl. [29]). Wir interessieren uns für die Eigenwerte des Einfachschichtpotentials

$$Vw(x) := \int_{S_2} U^*(x,y) w(y) \mathrm{d}\sigma_y \qquad \text{für } w \in H^{-1/2}(S_2).$$

Zunächst halten wir fest, dass das Einfachschichtpotential positiv definit ist.

4.48 Satz ([20, 29])**.** *Sei $w \in H^{-1/2}(S_2)$. Dann existiert $c_1^V > 0$ mit*

$$\langle Vw, w \rangle \geq \|w\|_{H^{-1/2}(S_2)}.$$

Für die Kompaktheit kann man entweder nachrechnen, dass der Intgralkern in $L^2(S_2 \times S_2)$ ist, oder man verwendet folgenden

4.49 Satz ([29])**.** *Sei $\Gamma := \partial\Omega$ der Rand eines Lipschitz-Gebietes Ω. Dann ist*

$$V : \; H^{-1/2-s}(\Gamma) \longrightarrow H^{1/2+s}(\Gamma)$$

für alle $s \in [-\frac{1}{2}, \frac{1}{2}]$ beschränkt.

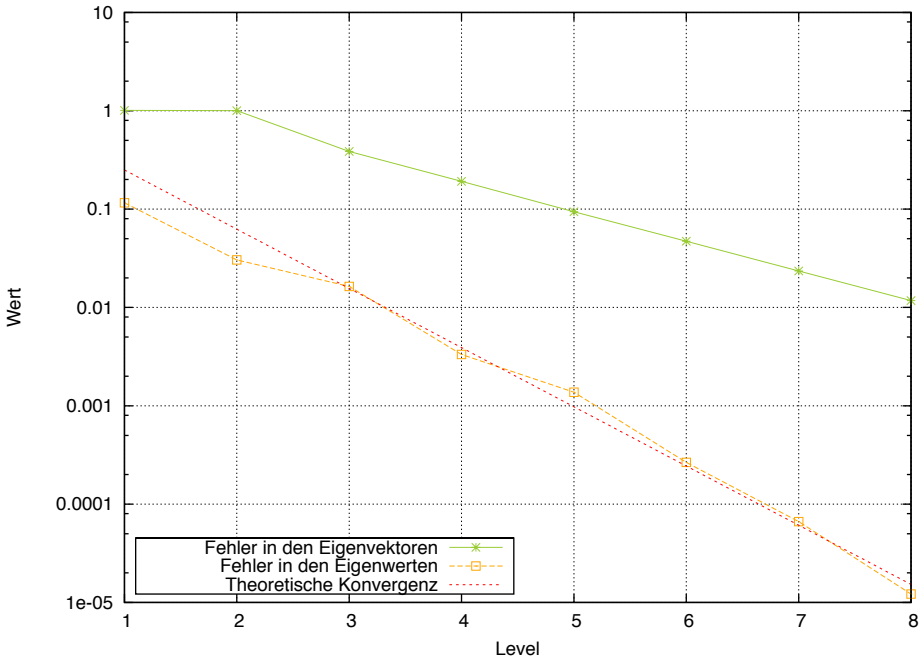

Abbildung 4.12: Konvergenz ARPACK

Wir verwenden diesen Satz mit $s = \frac{1}{2}$. Die Inklusion $\iota : H^1(S_2) \hookrightarrow L^2(S_2)$ ist kompakt (vgl. [1]). Damit ist $V\iota : L^2(S_2) \to L^2(S_2)$ kompakt nach Satz 2.22.
Die Eigenfunktionen des Einfachschichtpotentials auf S_2 sind nun genau die *Kugelflächenfunktionen* Y_n^m (vgl. [6, 22]).

4.50 Definition.

$$Y_n^m(\vartheta, \phi) := \sqrt{\frac{(2n+1)(n-|m|)!}{4\pi(n-|m|)!}} P_n^{|m|}(\cos(\vartheta)) e^{im\phi}$$

mit

$$P_n^m(t) := (1-t^2)^{m/2} \frac{\mathrm{d}^m}{\mathrm{d}t^m} P_n(t) \qquad \text{für } m = 0, 1, \ldots, n,$$

wobei $P_n(t)$ das Legendre-Polynom vom Grad n ist.

Für die Herleitung der Kugelflächenfunktionen sei ebenfalls auf [6] verwiesen. Der folgende Satz liefert nun die Eigenwerte des Einfachschichtpotentials (vgl. [5, 24]).

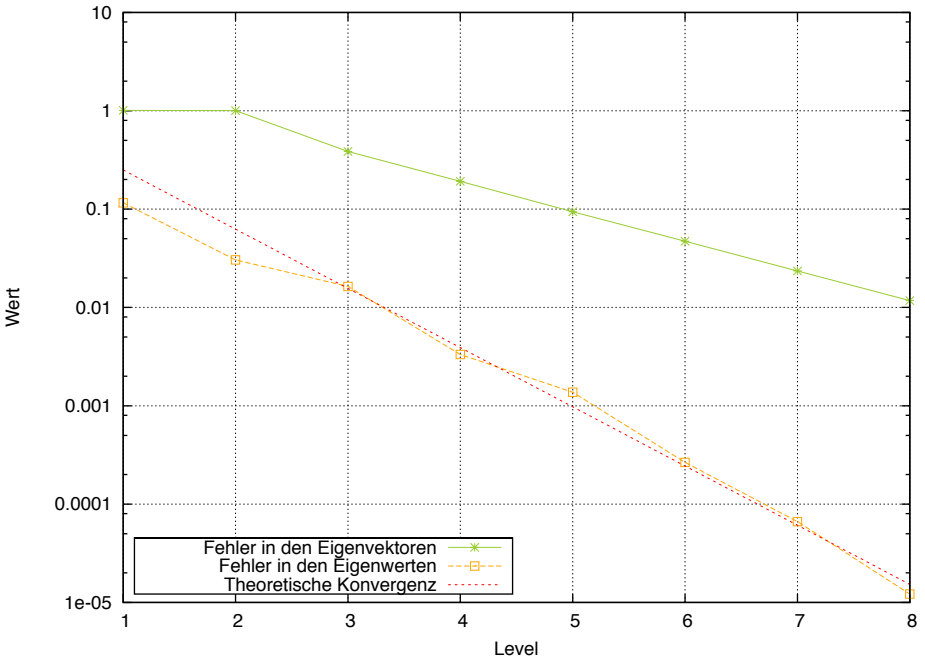

Abbildung 4.13: Konvergenz Unterraumiteration

4.51 Satz. *Es gilt*

$$V Y_n^m = \frac{1}{2n+1} Y_n^m$$

für $n = 0, 1, 2, \ldots$ *und* $m = -n, \ldots, n$.

Das Eigenwertproblem für das Einfachschichtpotential wird nun mit Hilfe eines *Wavelet-Galerkin-Verfahrens mit Matrix-Kompression* diskretisiert (vgl. [17]). Eine Implementierung wurde mir freundlicherweise von Herrn Prof. Dr. Harbrecht für diese Arbeit zur Verfügung gestellt. Aus der Diskretisierung ergibt sich ein verallgemeinertes Eigenwertproblem, das mittels ARPACK und der von mir implementierten Unterraumiteration gelöst wird. Da bei diesem Beispiel sowohl Eigenwerte als auch Eigenfunktionen bekannt sind, ist es möglich, den Approximationsfehler -wie in Kapitel 3 hergeleitet- explizit zu bestimmen. Nach [7] gilt für die Konvergenz des Wavelet-Galerkin-Verfahrens

$$\|u - u_h\|_{L^2(S_2)} \leq Ch \|u\|_{H^1(S_2)}$$

mit einer Konstanten $C > 0$. Dabei bezeichnet u die exakte Lösung, u_h die jeweilige Galerkin-Approximation und $h = 2^{-j}$ die Schrittweite der Diskretisierung mit den *Leveln* (vgl. [7, 17]) $j = 1, 2, \ldots$.

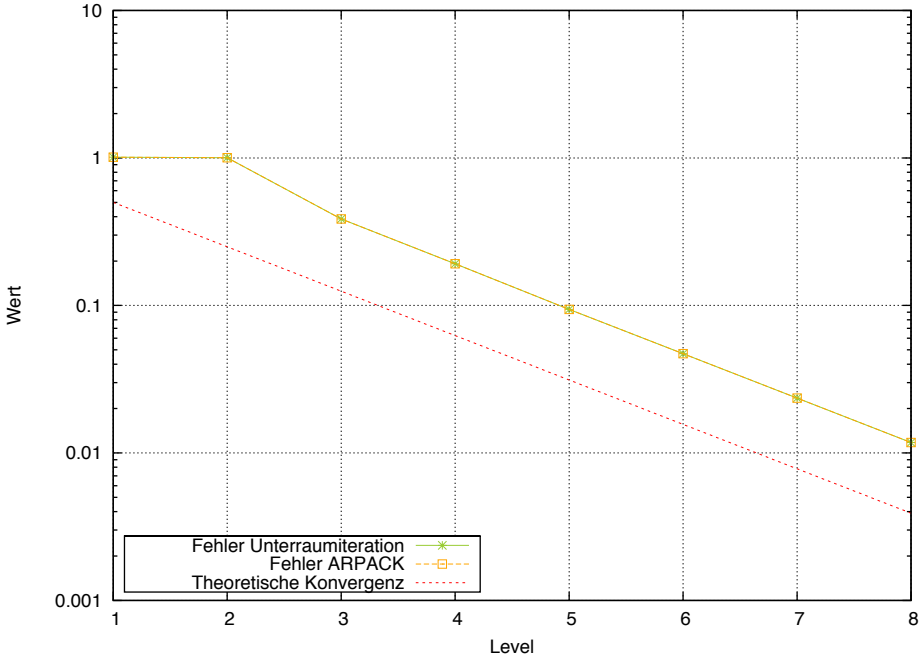

Abbildung 4.14: Konvergenz der Unterräume

Sowohl ARPACK als auch die Unterraumiteration wurden für die Level $j = 2, 3, \ldots, 8$ getestet[5]. Für zwei Level beträgt die Anzahl der Freiheitsgrade $n = 96$. Diese erhöht sich mit jedem Level um einen Faktor 4. Das bedeutet bei acht Leveln beträgt die Anzahl der Freiheitsgrade bereits $n = 393216$. (vgl. [17]).
Für jedes Level wurden jeweils die 49 größten Eigenwerte des diskretisierten Einfachschichtpotentials bestimmt. Zu diesen gehören genau sieben Eigenräume mit wachsender Dimensionen $2i - 1$ für $i = 1, \ldots, 7$. Hierbei ist i genau der Index der betraglich verschiedenen Eigenwerte

$$\lambda_i = \frac{1}{2i - 1}$$

für $i = 1, \ldots, 7$ (vgl. Satz 4.51). Somit sind in diesem Fall die Voraussetzungen von Satz 4.17 erfüllt und die Konvergenz der Eigenvektoren ist gewährleistet.
Als Fehlermaß für die numerisch bestimmten Eigenräume wird das in Kapitel 3 eingeführte Gap verwendet. Aus den Abbildungen 4.12 und 4.13 lässt sich

[5]Für $j = 1$ ist $n = 24$, hier wurden jeweils 23 Eigenwerte berechnet. Diese willkürlich erscheinende Zahl resultiert aus einer technischen Limitation des Softwarepakets ARPACK, wodurch nicht alle Eigenwerte einer Matrix bestimmt werden können (vgl. [23]).

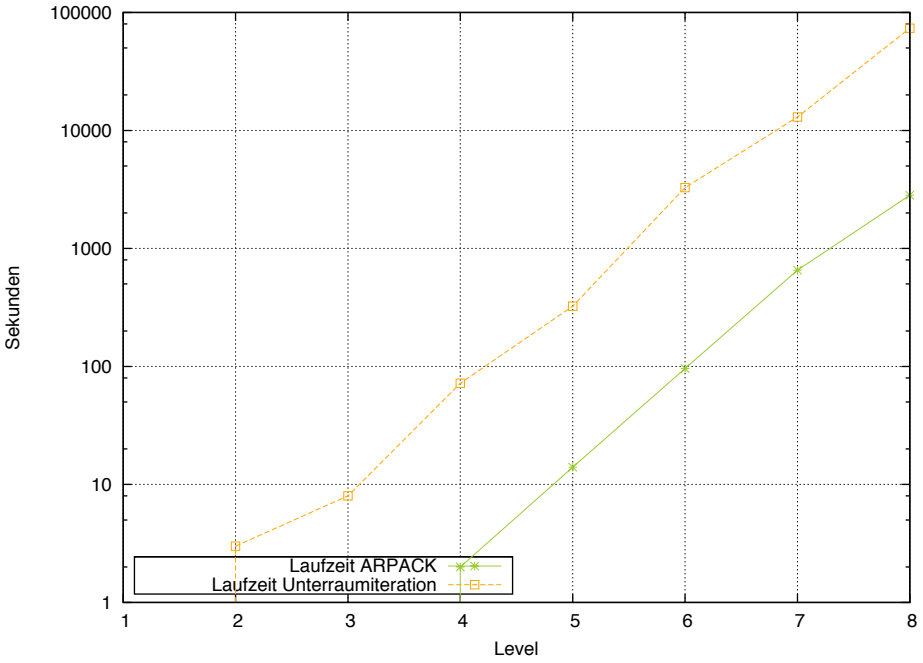

Abbildung 4.15: Vergleich der Laufzeiten

entnehmen, dass beide Verfahren die in Kapitel 3 hergeleitete, theoretische Konvergenzrate der Eigenwerte reflektieren.
Wir erinnern uns, dass ein Fehler der Größe ε in den Eigenräumen, zu einem Fehler der Größe ε^2 in den Eigenwerten führt. Genau dieses Verhalten ist sowohl bei der Unterraumiteration als auch bei ARPACK zu erkennen. Hierzu ist in den Plots jeweils auch die optimale Konvergenzrate dargestellt (rot).
Abbildung 4.14 zeigt das Konvergenzverhalten der Eigenräume bei der Unterraumiteration (grün) und bei ARPACK (orange) in einem gemeinsamen Plot, zusammen mit der Konvergenzrate der Galerkin-Diskretisierung (rot). Wie man sehen kann, ist die Konvergenz beider Verfahren identisch und genau um eine Konstante schlechter als 2^{-j} für $j = 1, 2, \ldots$. In den numerischen Tests benötigt die Unterraumiteration für die Konvergenz eine große Zahl von Matrix-Vektor-Multiplikationen und Orthogonalisierungsschritten. Dies führt dazu, dass die Unterraumiteration eine wesentlich längere Laufzeit hat als ARPACK. Abbildung 4.15 zeigt die Laufzeiten beider Verfahren im Vergleich.

Literaturverzeichnis

[1] Alt, H. W.: *Lineare Funktionalanalysis*. 4. Auflage. Springer, 2002

[2] Bebendorf, M.: Approximation of boundary element matrices. In: *Numer. Math.* 86 (2000), Nr. 4, S. 565–589

[3] Bosch, S.: *Lineare Algebra*. 2. Auflage. Springer, 2003

[4] Braess, D.: *Finite Elemente*. 4. Auflage. Springer, 2007

[5] Buffa, A. ; Sauter, S.: On the Acoustic Single Layer Potential: Stabilization and Fourier Analysis. In: *SIAM J. Sci. Comput.* 28 (2006), Nr. 5, S. 1974–1999

[6] Colton, D. ; Kress, R.: *Inverse Acoustic and Electromagnetic Scattering Theory*. 1. Auflage. Springer, 1992

[7] Dahmen, W. ; Harbrecht, H. ; Schneider, R.: Compression Techniques for Boundary Integral Equations—Asymptotically Optimal Complexity Estimates. In: *SIAM J. Numer. Anal.* 43 (2006), Nr. 6, S. 2251–2271

[8] D'Yakonov, E. G.: *Optimization in Solving Elliptic Problems*. 1. Auflage. CRC-Press, 1995

[9] Giraud, L. ; Langou, J.: A Robust Criterion for the Modified Gram-Schmidt Algorithm with Selective Reorthogonalization. In: *SIAM J. Sci. Comput.* 25 (2003), Nr. 2, S. 417–441

[10] Golub, G. H. ; van der Vorst, H. A.: Eigenvalue computation in the 20th century. In: *J. Comput. Appl. Math.* 123 (2000), Nr. 1-2, S. 35–65

[11] Golub, G. H. ; van Loan, C. F.: *Matrix Computations*. 3. Auflage. John Hopkins University Press, 1996

[12] Golub, G. H. ; Ye, Q.: An Inverse Free Preconditioned Krylov Subspace Method for Symmetric Generalized Eigenvalue Problems. In: *SIAM J. Sci. Comput.* 24 (2002), Nr. 1, S. 312–334

[13] Hanke-Bourgeois, M.: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. 3. Auflage. Vieweg+Teubner, 2008

[14] Harbrecht, H.: *Skript: Numerische Mathematik*. 2009 `http://harbrecht.ins.uni-bonn.de/teaching/numath2/skript.pdf`

[15] Harbrecht, H.: *Skript: Numerische Methoden für stochastische PDEs*. 2010

[16] Harbrecht, H. ; Peters, M. ; Schneider, R.: On the low rank approximation bei the pivoted Cholesky decomposition. (2010). `http://www.simtech.uni-stuttgart.de/publikationen/prints.php?ID=166`

[17] Harbrecht, H. ; Schneider, R.: Wavelet Galerkin Schemes for Boundary Integral Equations—Implementation and Quadrature. In: *SIAM J. Sci. Comput.* 27 (2006), Nr. 4, S. 1347–1370

[18] Higham, N. J.: A Survey of Condition Number Estimation for Triangular Matrices. In: *SIAM Review* 29 (1987), S. 575–596

[19] Higham, N. J.: *Accuracy and Stability of Numerical Algorithms*. 1. Auflage. Society for Industrial and Applied Mathematics, 1996

[20] Hsiao, G. C. ; Wendland, W. L.: A finite element method for some integral equations of the first kind. In: *Journal of Mathematical Analysis and Applications* 58 (1977), Nr. 3, S. 449–481

[21] J. Weickert: *Skript: Mathematik für Informatiker II*. 2007 `http://www.mia.uni-saarland.de/Teaching/MFI07/mfi2.pdf`. – Stand: 3. Juli 2010

[22] Kress, R.: Minimizing the condition number of boundary integral operators in acoustic and electromagnetic scattering. In: *Q J Mechanics Appl Math* 38 (1985), Nr. 2, S. 323–341

[23] Lehoucq, R. B. ; Sorensen, D. C. ; Yang, C.: *ARPACK Users Guide*. 1. Auflage. Society for Industrial and Applied Mathematics, 1998

[24] Martensen, E. ; Ritter, S.: Potential theory. In: *Lecture Notes in Earth Sciences* 65 (1997), S. 17–66

[25] Peters, G. ; Wilkinson, J. H.: $Ax = \lambda Bx$ and the Generalized Eigenproblem. In: *SIAM J. Numer. Anal.* 7 (1970), Nr. 4, S. 479–492

[26] Saad, Y.: *Numerical Methods for Large Eigenvalue Problems*. Halsted Press, 1992 `http://www-users.cs.umn.edu/~saad/EIGBOOK.tar.gz`

[27] Schwab, Christoph ; Todor, Radu A.: Karhunen-Loève approximation of random fields by generalized fast multipole methods. In: *J. Comput. Phys.* 217 (2006), Nr. 1, S. 100–122

[28] Sorensen, D. C.: Implicit application of polynomial filters in a k-step Arnoldi method. In: *SIAM J. Matrix Anal. Appl.* 13 (1992), Nr. 1, S. 357–385

[29] Steinbach, O.: *Numerische Näherungsverfahren für elliptische Randwertprobleme*. 1. Auflage. Vieweg+Teubner, 2003

[30] Watkins, D. S.: QR-like algorithms for eigenvalue problems. In: *J. Comput. Appl. Math.* 123 (2000), Nr. 1-2, S. 67–83

[31] Watkins, David S.: The QR Algorithm Revisited. In: *SIAM Rev.* 50 (2008), Nr. 1, S. 133–145