Lösung des symmetrischen Eigenwertproblems mit algebraischen Mehrgittermethoden

Dissertation

zur

Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität Rostock

vorgelegt von

Marcel Krüger, geb. am 21.09.1976 in Greifswald

aus Rostock

Rostock, 21.10.2011

urn:nbn:de:gbv:28-diss2012-0091-5

Gutachter:

Prof. Dr. Klaus Neymeyr (Universität Rostock)
Prof. Dr. Dirk Langemann (TU Braunschweig)

Abgabedatum:

21.10.2011

Datum der Verteidigung:

17.02.2012

Inhaltsverzeichnis

1. Einleitung

Die computergestützte Simulation ist heutzutage eine etablierte und unverzichtbare Methode, die in den verschiedensten technischen und theoretischen Wissenschaftsdisziplinen angewendet wird. So spielt sie eine wesentliche Rolle in der Physik oder den Ingenieurswissenschaften, beispielsweise bei der Konzeption, Voruntersuchung oder Fehlersuche im Bauwesen, der Elektrotechnik oder der Strömungsmechanik. Nicht selten stehen dabei Schwingungsprobleme im Mittelpunkt, was aus mathematischer Sicht häufig in Eigenwertproblemen elliptischer Differentialoperatoren mündet. Die analytische Lösung dieser Fragestellungen ist meist nur im Zusammenhang mit einfachen Gebieten zum Beispiel mittels Separationsansätzen lösbar, in den meisten Fällen ist sie aber nicht explizit anzugeben. Dies erfordert die numerische Behandlung dieser Aufgaben. Dazu wird das Operatoreigenwertproblem mittels Diskretisierung in ein diskretes verallgemeinertes Matrixeigenwertproblem, siehe Abschnitt 1.1, der Gestalt

$$Au = \lambda Mu, \quad A, M \in \mathbb{R}^{n\times n}, u \in \mathbb{R}^n, \lambda \in \mathbb{R}, \tag{1.1}$$

für große und dünnbesetzte Matrizen A und M überführt. Die resultierenden Matrizen sind dabei, bedingt durch den technischen Hintergrund, zumeist symmetrisch und positiv definit. Gleichermaßen ist bei technisch relevanten Untersuchungen nicht das gesamte Spektrum, sondern nur eine gewisse Anzahl der kleinsten Eigenwerte interessant, also die Lösung eines partiellen Eigenwertproblems erforderlich.
Bekanntermaßen existieren nach der Galois Theorie, speziell dem Satz von Abel-Ruffini, keine direkten Methoden zur Berechnung der Eigenwerte λ einer Matrix der Dimension $n > 4$. Darum stellen alle numerischen Eigenwertlöser (oder kurz Eigenlöser) iterative Verfahren dar und die Erforschung und Entwicklung zuverlässiger und schneller Algorithmen ist bis heute ein sehr aktives Forschungsgebiet der numerischen Mathematik, [2, 26, 64].
Historisch gesehen, beruhten die ersten Eigenlöser auf einer geeigneten Transformation der Ausgangsmatrix in eine Gestalt, die es erlaubt, Eigenwerte und Eigenvektoren explizit abzulesen. Das wohl älteste Verfahren stammt aus dem 19. Jahrhundert, vorgeschlagen von Jacobi. Das nach ihm benannte *Jacobi-Verfahren* basiert auf Ähnlichkeitstransformationen, mit dem Ziel, die Ausgangsmatrix in Diagonalgestalt zu überführen, was die Eigenwerte offenbart. Mitte der 50-iger Jahre des letzten Jahrhunderts ist, ausgehend von der gleichen zu Grunde liegenden Idee, der *QR-Algorithmus* entwickelt worden (beziehungsweise der *QZ-Algorithmus*, [53], für den Fall eines verallgemeinerten Eigenwertproblems). Beide Verfahren stellen (im numerischen Sinne) äußerst stabile Algorithmen dar. Jedoch ist ihre Anwendbarkeit auf Grund des etwa kubisch und damit überproportional wachsenden Aufwands für große Probleme der Dimension $n \approx 10^5$ oder $n \approx 10^6$ nicht mehr möglich. Solche hochdimensionalen Probleme sind heutzutage aber keine Seltenheit, da diese gerade vor dem Hintergrund technischer Untersuchungen auf natürliche Weise auf Grund hoher Approximationsanforderungen entstehen. Zur Behandlung dieser Probleme sind demzufolge Algorithmen wie Lanczos-Verfahren oder Arnoldi-Verfahren, als Varianten der Klasse von Krylovraum-Verfahren, entwickelt worden, die diesen Anforderungen gerecht werden. Zudem existieren weiterhin Methoden, die von der *Teile und Herrsche*-Strategie Gebrauch machen, [28]. Heutzutage sind diese Algorithmen Mittel der Wahl und liegen als implementierte Methoden in Software-Paketen, wie *LAPACK*, *ScaLAPACK* oder *ARPACK* vor. Der Grund hierfür liegt in ihrer universellen Verwendbarkeit, da sie das Eigenwertproblem aus dem Blickwinkel der linearen Algebra betrachten, also abgesehen von Eigenschaften der Matrizen A und M, wie Symmetrie oder komplexwertige Einträge, nur auf der

Basis der algebraischen Gleichung (1.1) agieren. Hierin liegt aber andererseits auch ein leichter Nachteil, da sie damit nicht optimal skalieren, das heißt, der Aufwand zur Lösung des Eigenwertproblems wächst, ähnlich der oben genannten klassischen Verfahren, deutlich schneller als die Problemdimension.
Für die im Vorfeld angesprochenen Eigenwertprobleme mit dünnbesetzten und positiv def niten Matrizen A, M (für die nur einige der kleinsten Eigenwerte bestimmt werden sollen) stellt es sich heraus, dass die in den sechziger bis achtziger Jahre vornehmlich durch russische Autoren vorgestellten und analysierten *vorkonditionierten Eigenlöser* (auch *vorkonditionierte Iterationen* beziehungsweise *preconditioned iterations*), [19, 20, 21, 25, 38, 65], als Lösungsansatz eine sehr effektive Methode für die oben beschriebenen Probleme darstellen. Als gradientenbasierte Verfahren prof tieren sie dabei von zwei Aspekten. Einerseits stellen sie eine inversenfreie Methode, das heißt sowohl A^{-1} als auch M^{-1} werden zur Umsetzung nicht benötigt, dar. Andererseits vermeiden sie Modif kationen der Matrizen A und M, wie beispielsweise Transformationen oder Zerlegungen, ein Vorteil, falls die Matrizen nicht explizit gegeben sind, sondern nur in Form einer Routine zur Realisierung des Matrix-Vektor-Produkts $y \mapsto Ay$ und $y \mapsto My$ vorliegen. Ohne tiefer ins Detail zu gehen, eine ausführliche Betrachtung f ndet in Kapitel 2 statt, ist als ein Bestandteil zur Umsetzung der vorkonditionierten Iterationen die Berechnung geeigneter Korrekturen in Form von vorkonditionierten Residuen nötig. Dies besteht im Wesentlichen aus der approximativen Lösung eines linearen Gleichungssystems. An dieser Stelle kommt ein weiterer Vorteil der Methode ins Spiel. Gerade vor dem hier erwähnten Hintergrund, der Lösung einer partiellen Differentialgleichung beziehungsweise eines Operatoreigenwertproblems und der damit gleichzeitig gegebenen zu Grunde liegenden Geometrie, bieten sich zur Lösung des auftretenden Gleichungssystems die geometrischen Mehrgitterverfahren an. Durch geschickte Nutzung der gegebenen geometrischen Informationen können sie eine fast optimale Komplexität, also einen fast linearen Zuwachs des Rechenaufwandes mit der Problemdimension, für die Berechnung der vorkonditionierten Residuen erreichen und damit die Komplexität des Eigenlösers niedrig halten. Daraus ergibt sich die Frage, ob es möglich ist, diese Komplexitätseigenschaften auch ohne Nutzung der Geometrie zu erzielen und damit, ähnlich zu den im Vorfeld angegebenen Eigenlösern, wie das Arnoldi-Verfahren, eine Behandlung aus „rein algebraischer Sicht“ zu ermöglichen. Prinzipiell kann diese Frage mit den algebraischen Mehrgitterverfahren positiv beantwortet werden. Diese nutzen die Idee der geometrischen Mehrgitterverfahren, bedienen sich im Gegensatz zu ihnen allerdings keiner Informationen über die Geometrie des Problems. Der Preis dafür liegt allerdings in einer (moderat) wachsenden Komplexität des Mehrgitterverfahrens und damit des Eigenlösers. Infolgedessen können die vorkonditionierten Iterationen unter Verwendung des algebraischen Mehrgitterverfahrens als „hybride“ Methode angesehen werden, da sie einerseits über gute Komplexitätseigenschaften verfügen und andererseits den Hintergrund oder die Struktur des zu Grunde liegenden Problems nicht benötigen, also eine algebraische Sichtweise ermöglichen. Dass diese Eigenlöser durchaus als Alternative zu den oben genannten etablierten Methoden, wie Lanczos- oder Jacobi-Davidson-Verfahren, betrachtet werden können, wird beispielsweise in [1] gezeigt.
Ziel der Arbeit ist es, die Leistungsfähigkeit der vorkonditionierten Iterationen unter Verwendung von algebraischer Mehrgittervorkonditionierung herauszustellen. Dabei soll im Gegensatz zu den Arbeiten von Arbenz et al., [1], beziehungsweise Borzi und Borzi, [5], eine deutlich umfangreichere Untersuchung stattf nden. So werden hier auch anisotrope und gitterfreie Probleme betrachtet. Die Basis dieser numerischen Untersuchungen am Ende der Arbeit bilden die im Verlauf der Arbeit angeführten Erörterungen zu den wesentlichen Komponenten, den vorkonditionierten Iterationen selbst und den algebraischen Mehrgitterverfahren zur Berechnung der benötigten Residuen. Dementsprechend gliedert sich die Arbeit folgendermaßen.
Kapitel 1 dient einer kurzen Einführung und stellt in Abschnitt 1.1 den Zusammenhang zwischen kontinuierlichem Operatoreigenwertproblem und diskretem Matrixeigenwertproblem unter Nutzung der Variationsformulierung dar.

Das Kapitel 2 widmet sich den vorkonditionierten Eigenlösern. Im Vordergrund steht hier die Konvergenzanalyse der gradientenbasierten Iteration PINVIT. Dabei wird eine ausführliche Darstellung des Konvergenzbeweises von Neymeyr und Knyazev, [41], präsentiert. Den Abschluss bilden die Abschnitte 2.3 und 2.4, in denen basierend auf PINVIT eine Hierarchie von Eigenlösern beziehungsweise deren Umsetzung als Unterraumiterationen zur simultanen Berechnung mehrerer der kleinsten Eigenwerte vorgestellt wird.
Kapitel 3 rückt die Mehrgitterverfahren zur Realisierung der Operation des Vorkonditionierers $y \mapsto B^{-1}y$ in den Mittelpunkt. Dabei wird in Abschnitt 3.1 das Mehrgitterverfahren motiviert und anschließend in Abschnitt 3.2 die Struktur erläutert. Abschnitt 3.3 geht dann ausführlich auf die algebraischen Mehrgitterverfahren und deren Umsetzung ein. Den Abschluss bilden in Abschnitt 3.4 klassische Konvergenzaussagen.
Das sich anschließende Kapitel 4 gibt einen Überblick zum Einsatz der Mehrgitterverfahren im Kontext der Eigenwertlöser. Im Zuge dieser Erörterungen wird ebenfalls die Implementierung der vorkonditionierten Iterationen beleuchtet.
Kapitel 5 dient dann umfangreichen Untersuchungen der im Vorfeld vorgestellten Eigenlöser. Dazu werden die Algorithmen auf mehrere Testprobleme angewendet und die Leistungsparameter angegeben.

1.1. Diskretisierung elliptischer Differentialoperatoren

Die folgenden Betrachtungen sollen den Zusammenhang zwischen einem Operatoreigenwertproblem mit einem selbstadjungierten, elliptischen, koerziven Differentialoperator zweiter Ordnung $\mathcal{L}$ und dem zu lösenden Matrixeigenwertproblem darlegen, [6, 34, 44]. Auch wenn die im Fokus der Arbeit stehenden vorkonditionierten Eigenlöser generell zur Lösung diskretisierter Operatoreigenwertprobleme mit symmetrisch positiv def niten Matrizen geeignet sind, wird mit Blick auf die numerischen Untersuchungen ein spezieller Typ dieser Operatorgleichungen im Vordergrund stehen. Dieser führt auf verallgemeinerte Eigenwertprobleme mit den im Vorfeld erwähnten großen, dünnbesetzten Matrizen von deren Struktur die hier studierten Eigenlöser prof tieren können.
Sei $\Omega \subset \mathbb{R}^d$, $d = 1,2,3$ ein beschränktes, offenes und zusammenhängendes Gebiet mit dem Rand $\Gamma = \Gamma_1 \cup \Gamma_2$ und $\Gamma_1 \cap \Gamma_2 = \varnothing$. Aufgabe ist es, Eigenwerte λ und zugehörige Eigenfunktionen u so zu bestimmen, dass sie die Operatorgleichung

$$-\nabla \cdot (\varepsilon(x)\nabla u) = \lambda u, \quad x \in \Omega, \tag{1.2}$$

$$u = 0, \quad x \in \Gamma_1, \tag{1.3}$$

$$n(x) \cdot \varepsilon(x)\nabla u = 0, \quad x \in \Gamma_2 \tag{1.4}$$

erfüllen. Die matrixwertige symmetrisch positiv def nite Funktion $\varepsilon(x)$ sei dabei stückweise stetig und $n(x)$ sei der Normalenvektor im Punkt $x \in \Gamma_2$. Eine Lösung $u \in C^2(\Omega) \cap C(\bar{\Omega})$ heißt *klassische Lösung* von (1.2). Die Bestimmung einer solchen Lösung ist auf Grund der starken Regularitätsforderung in den meisten Fällen nicht möglich. Man behilft sich daher mit der Suche in einem größeren Funktionenraum $\mathcal{H}^1 = \mathcal{H}^1(\Omega)$. Dieser enthält alle Funktionen $v \in L_2$, also der quadratisch Lebesque-integrierbaren Funktionen, deren schwache Ableitungen existieren und die ebenfalls L_2-Funktionen sind. Genaueres hierzu entnehme man beispielsweise [6] oder [44]. Die schwache Formulierung erhält man, indem die Operatorgleichung (1.2) mit Testfunktionen $v \in \mathcal{H}^1$ multipliziert und anschließend über das Gebiet Ω integriert wird. Dies liefert

$$\int_\Omega -\nabla \cdot (\varepsilon(x)\nabla u)v\,\mathrm{d}x = \int_\Omega \lambda uv\,\mathrm{d}x, \tag{1.5}$$

wobei unter Anwendung der Greenschen Formel eine Umformung zu

$$\int_\Omega \varepsilon(x)\nabla u \cdot \nabla v \mathrm{x} - \left[\int_{\Gamma_1} n(x)\cdot\varepsilon(x)\nabla u v \mathrm{x} + \int_{\Gamma_2} n(x)\cdot\varepsilon(x)\nabla u v \mathrm{x}\right] = \int_\Omega \lambda u v \mathrm{x} \tag{1.6}$$

möglich ist. Mit geeigneter Wahl von Testfunktionen v, welche die Randbedingungen erfüllen, also $v(x) = 0,\ x \in \Gamma_1$, und $v(x) \neq 0,\ x \in \Gamma_2$, erhält man auf Grund der geforderten Randbedingungen

$$\int_\Omega \varepsilon(x)\nabla u \cdot \nabla v \mathrm{x} = \int_\Omega \lambda u v \mathrm{x}. \tag{1.7}$$

Die schwache Formulierung oder auch Variationsformulierung für die Operatorgleichung (1.2) lautet demnach, f nde $(u, \lambda) \in \mathcal{H}^1 \times \mathbb{R}$, so dass

$$a(u, v) = \lambda(u, v), \quad \forall v \in \mathcal{H}^1 \tag{1.8}$$

gilt. Hierbei sind

$$a(u, v) := \int_\Omega \varepsilon(x)\nabla u \cdot \nabla v \mathrm{x} \tag{1.9}$$

eine symmetrische, koerzive und positive Bilinearform und

$$(u, v) := \int_\Omega u v \mathrm{x} \tag{1.10}$$

das innere Produkt auf $\mathcal{H}^1$. Funktionen $u \in \mathcal{H}^1$, die die Bedingung (1.8) erfüllen, nennt man schwache Lösungen der Operatorgleichung (1.2). Die schwache Formulierung ist aber immer noch ein kontinuierliches Problem. Zur numerischen (approximativen) Lösung von (1.8) wird ein diskretes Problem mittels der Methode der f niten Elemente (FEM) beziehungsweise der f niten Differenzen (FD) formuliert. Die Idee ist es dabei, diese approximative Lösung in einem endlichdimensionalen Teilraum

$$S \subset \mathcal{H}^1 \tag{1.11}$$

zu berechnen. Basis hierfür bildet die Triangulierung (oder Diskretisierung), also eine geeignete Partitionierung des Gebietes Ω in endlich viele Teilgebiete T. Diese werden auch Elemente genannt und sind im Fall $\Omega \subset \mathbb{R}$ Intervalle sowie für $\Omega \subset \mathbb{R}^2$ Dreiecke oder Vierecke und für $\Omega \subset \mathbb{R}^3$ Tetraeder, Würfel oder Quader. Eine Triangulierung $\mathcal{T}$ muss dabei folgende Bedingungen erfüllen:

(I) $\mathrm{vol}(T) > 0$, für alle $T \in \mathcal{T}$,

(II) $\bigcup_{T\in\mathcal{T}} = \bar{\Omega}$ und

(III) $\mathrm{Int}(T_i) \cup \mathrm{Int}(T_j) = \varnothing$ für alle $T_i, T_j \in \mathcal{T}$ mit $i \neq j$.

Dabei bezeichnet $\mathrm{vol}(T)$ das Volumen und $\mathrm{Int}(T)$ das Innere des Elementes T. Basierend auf einer solchen Triangulierung kann nun der oben erwähnte Teilraum S def niert werden. Unter Zuhilfenahme von linear unabhängigen Ansatzfunktionen $\{\phi_1, \ldots, \phi_n\} \in \mathcal{H}^1$ ist dieser durch

$$S := \mathrm{span}\{\phi_1, \ldots, \phi_n\} \tag{1.12}$$

gegeben und erfüllt somit die Bedingung (1.11) mit $\dim(S) = n$. Mithilfe dieser Basis kann nun die Aufgabenstellung des diskreten Eigenwertproblems formuliert werden. Dazu seien mit $u = \sum_{i=1}^n c_i\phi_i$ und $v = \sum_{i=1}^n \tilde{c}_i\phi_i$ die Basisdarstellungen von $u, v \in S$ gegeben. Mit (1.8) heißt dies

$$\sum_{i=1}^n \hat{c}_i\, a(\phi_i, \phi_j) = \lambda \sum_{i=1}^n \hat{c}_i\, (\phi_i, \phi_j), \quad (j = 1, \ldots, n). \tag{1.13}$$

Dies erlaubt eine Formulierung als diskretes (Matrix-)Eigenwertproblem der Gestalt

$$Au = \lambda M u \tag{1.14}$$

mit Matrizen $(A)_{ij} = a(\phi_i, \phi_j)$ und $(M)_{ij} = (\phi_i, \phi_j) \in \mathbb{R}^{n \times n}$ und der Umbenennung des Koeffizientenvektors $\hat{c}$ aus (1.13) in u. Dabei ist $A \in \mathbb{R}^{n \times n}$, die Steifigkeitsmatrix, symmetrisch positiv definit und $M \in \mathbb{R}^{n \times n}$, die Massematrix, ebenfalls symmetrisch positiv (semi)definit. Weiterhin ist zu beachten, dass $a(\phi_i, \phi_j) = (\phi_i, \phi_j) = 0$, falls $\text{supp}(\phi_i) \cap \text{supp}(\phi_j) = \varnothing$, also die Träger, als Teilgebiete von Ω, von ϕ_i und ϕ_j disjunkt sind. Wählt man nun Test- und Ansatzfunktionen mit sehr kleinem Träger, beispielsweise Hutfunktionen (siehe Kapitel 5), erzeugt dies dünnbesetzte Matrizen A und M, wie sie im Vorfeld beschrieben wurden.

2. Vorkonditionierte Iterationen

Dieser erste Abschnitt der Arbeit widmet sich der Konvergenzanalyse vorkonditionierter Iterationen zur Lösung des verallgemeinerten Eigenwertproblems

$$Au = \lambda Mu. \tag{2.1}$$

Dabei sind sowohl $A \in \mathbb{R}^{n\times n}$ als auch $M \in \mathbb{R}^{n\times n}$ symmetrisch positiv def nite Matrizen, wie sie beispielsweise bei der Diskretisierung von elliptischen Differentialoperatoren mittels Finiter-Elemente-Methode (FEM) auftreten. Zudem ist $\lambda \in \mathbb{R}$ ein Eigenwert und $u \in \mathbb{R}^n$ der zugehörige Eigenvektor. Zusammen bilden diese das Eigenpaar (λ, u). Das Spektrum des Matrixpaares (A, M) ist die Menge aller Eigenwerte, $\sigma(A, M) = \{\lambda \in \mathbb{R} : Au = \lambda Mu\}$. Da sowohl A als auch M symmetrisch positiv def nite Matrizen sind, besitzt (A, M) nur reelle und positive Eigenwerte. Sie seien aufsteigend angeordnet, das heißt

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

Ein elementares Verfahren zur Berechnung des kleinsten Eigenpaares (λ_1, u_1) stellt die inverse Vektoriteration nach Wielandt dar. Diese konstruiert eine Folge von Iterierten $u^{(j)}, j = 0, 1, 2, \ldots$ aus der Lösung des Gleichungssystems

$$Au^{(j+1)} = cMu^{(j)}, \quad c \neq 0, \tag{2.2}$$

und stellt daher die Potenzmethode für $A^{-1}M$

$$u^{(j+1)} = cA^{-1}Mu^{(j)}$$

dar. Wählt man $c \in \mathbb{R}$ als den Rayleigh-Quotienten von $u^{(j)}$, der durch

$$\lambda(u^{(j)}) = \frac{(u^{(j)}, Au^{(j)})_2}{(u^{(j)}, Mu^{(j)})_2}, \tag{2.3}$$

mit dem euklidischen Skalarprodukt $(\cdot, \cdot)_2$, gegeben ist, wird Stationarität in einem Eigenvektor u_i erreicht, da dann für $u^{(j)} = u_i$

$$Au^{(j+1)} = \lambda(u_i)Mu_i$$

mit

$$\lambda(u_i) = \frac{(u_i, Au_i)_2}{(u_i, Mu_i)_2} = \frac{(u_i, \lambda_i Mu_i)_2}{(u_i, Mu_i)_2} = \lambda_i$$

und somit $u^{(j+1)} = u_i$ gilt. Gleichzeitig ist damit auch der Eigenwert λ_i berechnet. Bekanntermaßen konvergiert dieses Verfahren, wenn die Startiterierte $u^{(0)}$ nicht senkrecht auf dem Unterraum, der vom kleinsten Eigenvektor aufgespannt wird, gewählt ist, gegen das kleinste Eigenpaar (λ_1, u_1). Der Konvergenzfaktor σ ergibt sich aus dem Verhältnis der kleinsten voneinander verschiedenen Eigenwerte, [64], lautet also

$$\sigma = \frac{\lambda_1}{\lambda_i} \quad \text{mit} \quad i = \min_j\{j : \lambda_1 < \lambda_j\}. \tag{2.4}$$

Mittels eines zusätzlichen Shift-Parameters τ kann die Iteration (2.2) so modif ziert werden, dass beliebige Eigenwerte aus dem Spektrum des Matrixpaares (A, M) berechnet werden können. Die zugehörige Iteration hat die Gestalt

$$(A - \tau M)u^{(j+1)} = \lambda(u^{(j)})Mu^{(j)}.$$

Ergebnis ist das Eigenpaar (λ_k, u_k) mit $\lambda_k = \min_{k=1,\dots,n} |\tau - \lambda_k|$. Eine Konvergenzbeschleunigung kann mittels der Rayleigh-Quotienten-Iteration, bei welcher der Shift τ in jedem Iterationsschritt j aktualisiert wird, erreicht werden. Durch die Wahl des Shifts τ_j als Rayleigh-Quotient der aktuellen Iterierten $u^{(j)}$ ist dann der betragsmäßig kleinste Eigenwert nahe Null und sichert somit nach Gleichung (2.4) eine schnelle Konvergenz.
Grundsätzlich ist es jedoch in jedem der vorgestellten Verfahren notwendig, ein lineares Gleichungssystem der Form $Ax = b$ zu lösen. Im Falle der Rayleigh-Quotienten-Iteration wird zudem noch die Systemmatrix in jedem Schritt modif ziert und erfordert die Lösung eines fast singulären Gleichungssystems. Auch wenn die aus Anwendungsproblemen resultierenden Matrizen häuf g nur dünnbesetzt sind, ist die Berechnung von $u^{(j+1)}$ im Falle hochdimensionaler Probleme ($n \approx 10^5 - 10^6$) aus numerischer Sicht viel zu aufwändig und benötigt im Allgemeinen mehr Speicherplatz als A selbst. Ebenso kann der Fall eintreten, dass die Matrix A nicht explizit bekannt ist, sondern nur eine Routine zur Realisierung des Matrix-Vektor-Produktes Ax zur Verfügung steht, was beispielsweise die Anwendung einer LU-Zerlegung verbietet.

2.1. Gradientenverfahren für den Rayleigh-Quotienten

Die im Vorfeld beschriebenen Aspekte motivieren Verfahren, bei denen die Inverse der Matrix A nicht benötigt wird, beziehungsweise die exakte Lösung des Gleichungssystems nicht notwendig ist. Ähnlich der Methodik bei iterativen Verfahren zum Lösen linearer Gleichungssysteme (vgl. auch Kapitel 3) besteht eine Methode darin, vorkonditionierte Residuen einzusetzen. Eines dieser Verfahren ist die *vorkonditionierte inverse Vektoriteration* (engl.: preconditioned inverse iteration oder kurz: *PINVIT*), deren Vorschrift

$$\hat{u} = u - B^{-1}(Au - \lambda(u)Mu) \tag{2.5}$$

lautet, [56], wobei u beziehungsweise $\hat{u}$ die vorher genutzten Bezeichnungen $u^{(j)}$ und $u^{(j+1)}$ ersetzen. Hierbei ist B^{-1} eine symmetrisch positiv def nite Matrix, der *Vorkonditionierer*. Sie stellt eine Näherung an die Inverse der Matrix A dar und wird im Weiteren ausführlicher diskutiert. Auch hier erkennt man, dass die Iteration (2.5) in einem Eigenpaar (λ_i, u_i) stationär ist. Festzuhalten ist weiterhin, dass PINVIT ein *gradientenbasiertes Abstiegsverfahren* darstellt, wie eine kurze Überlegung zeigt. Differenziert man den Rayleigh-Quotienten $\lambda(u)$ aus (2.3) bezüglich u ergibt sich

$$\nabla\lambda(u) = \frac{2Au(u, Mu)_2 - 2Mu(u, Au)_2}{(u, Mu)_2^2} = \frac{2}{(u, Mu)_2}\left(Au - \lambda(u)Mu\right).$$

Weiterhin sei der B-Gradient ∇_B mittels des euklidischen Gradienten durch

$$(\nabla\lambda(u), h)_2 = (\nabla_B\lambda(u), h)_B, \quad \forall h \in \mathbb{R}^n,$$

def niert. Hierbei ist $(\cdot,\cdot)_2$ das euklidische und $(\cdot,\cdot)_B$ das von B induzierte Skalarprodukt. Durch elementares Umstellen erhält man die Beziehung

$$\nabla_B\lambda(u) = B^{-1}\nabla\lambda(u) = \frac{2}{(u, Mu)_2}B^{-1}\left(Au - \lambda(u)Mu\right)$$

und damit eine Kollinearität der Korrekturrichtung in Gleichung (2.5) zu $\nabla_B \lambda(u)$. Eine einfache Umformung der Iteration (2.5) liefert weiterhin

$$\hat{u} = \lambda(u)A^{-1}Mu + (I - B^{-1}A)(u - \lambda(u)A^{-1}Mu) \tag{2.6}$$

und damit einen direkten Zusammenhang zur inversen Vektoriteration. $\hat{u}$ kann offensichtlich als Ergebnis einer „gestörten“ inversen Vektoriteration aufgefasst werden. Offensichtlich kann die vorkonditionierte Vektoriteration daher auf zwei Wegen motiviert werden. Einerseits aus dem Zugang über die inverse Vektoriteration und andererseits als gradientenbasiertes Abstiegsverfahren.

Bemerkung 2.1.1. *Im Gegensatz zur Anwendung eines Vorkonditionierers bei linearen Gleichungssystemen führt die Wahl $B = A$ nicht zur Einschrittkonvergenz, also dem Erhalt der Lösung in einem Schritt. Man erhält in diesem Fall die ursprüngliche inverse Vektoriteration.*

Die Subtraktion des Terms $\lambda(u)A^{-1}Mu$ in (2.6) ergibt weiterhin

$$\hat{u} - \lambda(u)A^{-1}Mu = (I - B^{-1}A)(u - \lambda(u)A^{-1}Mu) \tag{2.7}$$

und damit eine Fehlerfortpf anzungsgleichung, die eine notwendige Eigenschaft des Vorkonditionierers offenbart. Damit die Iteration (2.5) konvergiert, muss die Fehlerfortpf anzungsmatrix $I - B^{-1}A$ die Eigenschaft einer Kontraktion aufweisen. Dieses ist erfüllt, falls sie der Bedingung

$$\|I - B^{-1}A\|_A \leq \gamma < 1, \tag{2.8}$$

wobei $\|\cdot\|_A$ die von der A-Vektornorm induzierte A-Operatornorm bezeichnet, genügt. Ausführliche Betrachtungen zum Vorkonditionierer, wie etwa die Kontraktionseigenschaft, seine Existenz und die Konstruktion, sind Gegenstand des Abschnitts 2.2.
Es ist weiterhin möglich, ein zum Abstiegsverfahren äquivalentes Verfahren zu formulieren, welches ebenfalls auf dem Einsatz der vorkonditionierten Residuen basiert. Es resultiert aus folgendem Zusammenhang.

Bemerkung 2.1.2. *Seien $A, M \in \mathbb{R}^{n\times n}$ symmetrisch positiv def nite Matrizen. Dann ist die Berechnung des kleinsten Eigenwertes $0 < \lambda_1$ des Matrixpaares (A, M) äquivalent zur Berechnung des größten Eigenwertes $\mu_1 = \frac{1}{\lambda_1}$ des Matrixpaares (M, A).*

Zur Lösung des resultierenden verallgemeinerten Eigenwertproblems

$$Mu = \mu Au \tag{2.9}$$

muss folglich das größte Eigenpaar berechnet werden. Naheliegend ist daher ein *gradientenbasiertes Anstiegsverfahren*. Dieses ergibt sich aus (2.5) mit $\lambda(u) = \frac{1}{\mu(u)}$ als

$$\hat{u} = u + \frac{1}{\mu(u)} B^{-1}(Mu - \mu(u)Au) \tag{2.10}$$

mit dem Rayleigh-Quotienten

$$\mu(u) = \frac{(u, Mu)_2}{(u, Au)_2}.$$

In Anlehnung an die Darstellung (2.6) kann dieses Verfahren auch in der Gestalt

$$\mu(u)\hat{u} = A^{-1}Mu - (I - B^{-1}A)(A^{-1}Mu - \mu(u)u) \tag{2.11}$$

formuliert werden. Wiederum erkennt man die Struktur einer Fehlerfortpf anzungsgleichung. Demzufolge tritt auch hier Konvergenz auf, falls durch $(I - B^{-1}A)$ eine kontraktive Abbildung def niert ist.
Es bleibt festzuhalten, dass das Adaptieren der Methode zur iterativen Lösung linearer Gleichungssysteme mittels vorkonditionierter Residuen auf die inverse Vektoriteration auf natürliche Art und Weise im iterativen Eigenlöser (2.6) mündet. Diese auf dem Gradienten des Rayleigh-Quotienten basierende Iteration kann weiterhin in ein äquivalentes Anstiegsverfahren umformuliert werden. Beide hängen durch den in Bemerkung 2.1.2 beschriebenen Sachverhalt zusammen. Der wesentliche Unterschied besteht darin, dass die Anwendung des Abstiegsverfahrens den kleinsten und die des Anstiegsverfahrens den größten Eigenwert liefert. Dass beide Verfahren konvergieren, ist Kern dieses Kapitels und Gegenstand des folgenden Abschnitts.

2.2. Konvergenz gradientenbasierter Iterationen

Grundlage der in diesem Kapitel angeführten Konvergenzanalyse bilden die Arbeiten von Neymeyr und Knyazev, [40, 41, 56, 57]. Der Fokus liegt dabei auf der Arbeit [41], in der eine vergleichsweise kompakte Herleitung der Konvergenzrate gegeben ist. Diese unterscheidet sich von den älteren Arbeiten [56, 57] auch in der Darstellung der Abschätzung, welche bereits in [40] angedeutet wird. Dabei gibt [41] auch die Motivation zur Betrachtung des zu PINVIT äquivalenten Anstiegsverfahrens, welches im Vorfeld eingeführt wurde. Wie sich zeigen wird, vergleiche Satz 2.2.33, können die erhaltenen Ergebnisse dann unmittelbar auf das Abstiegsverfahren übertragen werden.
Ohne die Problemklasse einzugrenzen, soll vorerst das verallgemeinerte Eigenwertproblem für das Matrixpaar (M, A) in eine äquivalente Formulierung überführt werden. Dazu wird auf folgende Matrix zurückgegriffen.

Def nition 2.2.1. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrisch positiv def nite Matrix. Dann heißt eine symmetrisch positiv def nite Matrix $T \in \mathbb{R}^{n\times n}$ mit $T^2 = A$ die* **Wurzel** *von A.*

Lemma 2.2.2. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrisch positiv def nite Matrix. Dann existiert genau eine symmetrisch positiv def nite Matrix T mit $T^2 = A$.*

Beweis. A ist diagonalisierbar, das heißt, es existiert eine orthogonale Matrix V, mit $V^TV = I$, deren Spalten die Eigenvektoren von A enthalten. Somit gilt

$$A = VDV^T$$

mit einer Diagonalmatrix D, für deren Einträge $d_{ii} > 0$ gilt. Mit $D^{1/2} = \text{diag}(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}})$ ist dies äquivalent zu

$$A = VD^{1/2}V^TVD^{1/2}V^T$$

und man def niert damit

$$T := A^{1/2} = VD^{1/2}V^T.$$

Diese Matrix T ist symmetrisch und auf Grund der Ähnlichkeit zur positiv def niten Matrix $D^{1/2}$ ebenfalls positiv def nit. □

Es kann nun die Überführung des verallgemeinerten Eigenwertproblems in die eines Standardeigenwertproblems erfolgen.

Satz 2.2.3. *Gegeben sei das verallgemeinerte Eigenwertproblem* (2.9) *zum Matrixpaar (M, A) mit symmetrisch positiv def niten Matrizen M und A. Dann kann* (2.9) *auf das Standardeigenwertproblem*

(2.12) $$\tilde{M}z = \mu z$$

mit symmetrisch positiv definiter Matrix $\tilde{M}$ transformiert werden. Die resultierende Iterationsvorschrift des Anstiegsverfahrens lautet dabei

$$\hat{z} = z + \frac{1}{\tilde{\mu}(z)} \tilde{B}^{-1}(\tilde{M}z - \tilde{\mu}(z)z) \tag{2.13}$$

mit dem Rayleigh-Quotienten

$$\tilde{\mu}(z) = \frac{(z, \tilde{M}z)_2}{(z,z)_2}.$$

Beweis. Mit der Zerlegung $A = TT$ entsprechend Lemma 2.2.2 ist (2.9) äquivalent zu

$$Mu = \mu TTu.$$

Multiplikation mit T^{-1} und anschließende Substitution $z = Tu$ liefert

$$T^{-1}MT^{-1}z = \mu z$$

und für $\tilde{M} := T^{-1}MT^{-1}$ damit die Transformation auf das Standardeigenwertproblem

$$\tilde{M}z = \mu z. \tag{2.14}$$

Offensichtlich bleibt die Symmetrie erhalten, die positive Definitheit ebenso, da die Eigenwerte unter der Transformation unangetastet bleiben. Man erhält das in der Behauptung angegebene, zu (2.9) äquivalente Standardeigenwertproblem. Für den Rayleigh-Quotienten ergibt sich

$$\mu(u) = \frac{(u,Mu)_2}{(u,Au)_2} = \frac{(u,Mu)_2}{(u,TTu)_2} = \frac{(T^{-1}z, MT^{-1}z)_2}{(z,z)_2} = \frac{(z,\tilde{M}z)_2}{(z,z)_2} =: \tilde{\mu}(z). \tag{2.15}$$

Ausgehend von der Iteration (2.10) führt oben angegebene Transformation und Substitution auf

$$\begin{aligned} \hat{u} &= u + \frac{1}{\mu(u)} B^{-1}(Mu - \mu(u)Au) \\ T\hat{u} &= Tu + \frac{1}{\mu(u)} TB^{-1}(Mu - \mu(u)TTu) \\ \hat{z} &= z + \frac{1}{\tilde{\mu}(z)} TB^{-1}T(T^{-1}MT^{-1}z - \tilde{\mu}(z)z). \end{aligned}$$

Mit $\tilde{B}^{-1} := TB^{-1}T$ und $\tilde{M} := T^{-1}MT^{-1}$ erhält man schließlich die gewünschte Gestalt (2.13). □

Bemerkung 2.2.4. *Für das Matrixpaar (A, M) kann mittels der Wurzel der Matrix $M = \bar{T}\bar{T}$ auf die gleiche Weise das verallgemeinerte Eigenwertproblem $Au = \lambda Mu$ auf das Standardeigenwertproblem $\tilde{A}u = \lambda u$ transformiert werden. Das resultierende Abstiegsverfahren (PINVIT) lautet*

$$\hat{z} = z - \bar{B}^{-1}(\tilde{A}z - \tilde{\lambda}(z)z) \tag{2.16}$$

mit entsprechend modifiziertem Vorkonditionierer $\bar{B}^{-1}$.

Damit genügt es, die analytischen Betrachtungen auf das Standardeigenwertproblem einzugrenzen. Weiterhin soll für die Beweisführung eine vorläufige Einschränkung vorgenommen werden. Dabei geht man vorerst von einfachen Eigenwerten μ_i aus, das heißt

$$0 < \mu_n < \mu_{n-1} < \ldots < \mu_1.$$

Der Fall mehrfacher Eigenwerte wird am Ende des Kapitels in Lemma 2.2.35 erörtert.
Eine einfache Überlegung zeigt, dass für den Rayleigh-Quotienten

$$\min_{u\in\mathbb{R}^n} \mu(u) = \mu_n \quad \text{und} \quad \max_{u\in\mathbb{R}^n} \mu(u) = \mu_1 \tag{2.17}$$

und damit

$$\mu(u) \in [\mu_n, \mu_1]$$

gilt. Für eine Iterierte u sei zudem

$$\mu_{i+1} \leq \mu(u) \leq \mu_i. \tag{2.18}$$

Dann kann der Rayleigh-Quotient der Iterierten $\hat{u}$ folgendermaßen klassif ziert werden.

Bemerkung 2.2.5. *Sei $u \in \mathbb{R}^n$ mit $\mu(u) \in (\mu_{i+1}, \mu_i)$. Dann tritt für den Rayleigh-Quotienten der mittels Iteration* (2.13) *erhaltenen Folgeiterierten $\hat{u}$ genau einer der drei Fälle*

(I) $\mu(\hat{u}) = \mu_j$,

(II) $\mu(\hat{u}) \in (\mu_{j+1}, \mu_j)$, $j \neq i$, *oder*

(III) $\mu(\hat{u}) \in (\mu_{i+1}, \mu_i)$

ein.

Der erste Fall beschreibt die Berechnung des Eigenwertes μ_j und damit eine Lösung des Eigenwertproblems, konkret $M\hat{u} = \mu_j\hat{u}$. Dabei ist nicht sichergestellt, dass dies der größte Eigenwert ($j = 1$) ist. Soll dieser berechnet werden, so kann die Iteration mit einem $u' \perp \hat{u}$ (und nachfolgender Beibehaltung dieser Orthogonalitätsbedingung für alle Folgeiterierten) neu gestartet werden.
Der zweite Fall beschreibt den Sprung des Rayleigh-Quotienten in ein vom Ausgangsintervall verschiedenes, wobei in Lemma 2.2.17 gezeigt wird, dass hierbei $j < i$ gelten muss, also ein Sprung nur in Intervalle, deren Grenzen von größeren Eigenwerten gebildet werden, erfolgen kann. Ein solcher Sprung kann bei wiederholter Anwendung von (2.13) aber nicht beliebig oft auftreten, da der Rayleigh-Quotient nach Bedingung (2.17) durch μ_1 beschränkt bleibt.
Somit beschreibt der dritte Fall die letzte Möglichkeit, nämlich dass sowohl $\mu(u)$ als auch $\mu(\hat{u})$ im selben (offenen) Intervall liegen und keine Stationarität in einem Eigenwert gegeben ist. Für diese Situation kann folgende Konvergenzaussage getroffen werden.

Satz 2.2.6. *Zur Lösung des Standardeigenwertproblems* (2.12) *sei das Anstiegsverfahren aus* (2.13) *verwendet. Der Vorkonditionierer B^{-1} sei derart, dass die aus ihm resultierende Fehlerfortpf anzungsmatrix eine Kontraktion entsprechend Gleichung* (2.8) *ist. Für den Rayleigh-Quotienten der Iterierten $u \in \mathbb{R}^n$ gelte zudem $\mu_{i+1} < \mu(u) < \mu_i$. Dann gilt für den Rayleigh-Quotienten der Folgeiterierten $\mu(\hat{u})$ im Falle $\mu_{i+1} < \mu(\hat{u}) < \mu_i$ die Abschätzung*

$$\frac{\mu_i - \mu(\hat{u})}{\mu(\hat{u}) - \mu_{i+1}} \leq \sigma^2 \frac{\mu_i - \mu(u)}{\mu(u) - \mu_{i+1}} \tag{2.19}$$

mit dem Konvergenzfaktor

$$\sigma = 1 - (1-\gamma)\frac{\mu_i - \mu_{i+1}}{\mu_i} = \gamma + (1-\gamma)\frac{\mu_{i+1}}{\mu_i}. \tag{2.20}$$

Der Beweis des Satzes wird am Ende des Abschnitts 2.2.3 angegeben.
Bei der Betrachtung der Iterationsvorschrift (2.11) zeigt sich, dass das schlechtmöglichste Konvergenzverhalten in Abhängigkeit zweier Freiheitsgrade vorliegt. Zum einen ist dies der Vorkonditionierer B^{-1} und zum anderen der Rayleigh-Quotient $\mu(u)$. Dies führt zu einem zweistuf gen Extremalproblem, welches es zu analysieren gilt.
Nach einigen Vorbetrachtungen sowie Bemerkungen zum Vorkonditionierer beschäftigt sich der hier als „$\mathcal{B}_\gamma$-Analyse“ bezeichnete Teil mit der Lokalisierung der, im Sinne einer Konvergenzabschätzung, schlechtmöglichsten Folgeiterierten $\hat{u}$ in Abhängigkeit des gewählten Vorkonditionierers B^{-1}. Der sich anschließende, als „$\mathcal{L}_\rho$-Analyse“ bezeichnete Part, zeigt dann, dass für alle Iterierten mit demselben Rayleigh-Quotienten diese schlechtmöglichste Konvergenz in einem speziellen Unterraum eintritt, was die Betrachtungen auf eine niedrigdimensionale Analyse reduziert. Mittels dieser kann abschließend der Beweis zu Satz 2.2.6 formuliert werden. Formal lautet daher das Extremalproblem zur Bestimmung des schlechtmöglichsten Konvergenzverhaltens, also minimalem Zuwachs im Rayleigh-Quotienten,

$$\min_{u\in\mathcal{L}(\rho)} \min_{B^{-1}\in B_\gamma} \mu(\hat{u}), \tag{2.21}$$

wobei $\mathcal{L}(\rho)$ beziehungsweise B_γ im Zuge der Betrachtungen zu spezif zierende Mengen sind.

2.2.1. Der Vorkonditionierer

Bereits in der anfangs angegebenen Herleitung der gradientenbasierten Iterationen (2.5) und (2.10) wurde die im Wesentlichen einzige Forderung an den jeweiligen Vorkonditionierer formuliert, nämlich die kontraktive Eigenschaft der Fehlerfortpf anzungsmatrix $I - B^{-1}A$. Die Bedingung der Kontraktion ist erfüllt, falls der Spektralradius ϱ, der Betrag des betragsmäßig größten Eigenwertes, durch $\gamma < 1$ beschränkt ist, also

$$\varrho(I - B^{-1}A) \leq \gamma < 1$$

gilt. Da die Matrix $I - B^{-1}A$ im Allgemeinen unsymmetrisch ist, gilt dann $\|(I - B^{-1}A)\|_2 \neq \varrho(I - B^{-1}A)$. Dennoch ist der Spektralradius zugänglich und zwar mittels der A-Operatornorm.

Lemma 2.2.7. *Seien A, B symmetrisch positiv def nite Matrizen. Dann gilt*

$$\varrho(I - B^{-1}A) = \|(I - B^{-1}A)\|_A$$

Beweis. Aus der Def nition der A-Norm erhält man

$$\begin{aligned}
\|(I - B^{-1}A)\|_A &= \max_{0\neq v\in\mathbb{R}^n} \frac{\|(I - B^{-1}A)v\|_A}{\|v\|_A} \\
&= \max_{0\neq v\in\mathbb{R}^n} \frac{\|A^{1/2}(I - B^{-1}A)v\|_2}{\|A^{1/2}v\|_2} \\
&\overset{z=A^{1/2}v}{=} \max_{0\neq z\in\mathbb{R}^n} \frac{\|A^{1/2}(I - B^{-1}A)A^{-1/2}z\|_2}{\|z\|_2} \\
&= \max_{0\neq z\in\mathbb{R}^n} \frac{\|(I - A^{1/2}B^{-1}A^{1/2})z\|_2}{\|z\|_2} \\
&= \varrho(I - A^{1/2}B^{-1}A^{1/2}).
\end{aligned}$$

Die letzte Gleichung ergibt sich dabei aus der nun vorliegenden Symmetrie der Matrix $I - A^{1/2}B^{-1}A^{1/2}$ und der spektralen Äquivalenz von $B^{-1}A$ und $A^{1/2}B^{-1}A^{1/2}$. □

Die Iterationen (2.5) und (2.10) sind demnach konvergent, falls

$$\|I - B^{-1}A\|_A \leq \gamma < 1,$$

also die am Anfang unter (2.8) angegebenen Abschätzung, gilt. Alternativ kann dies auch mittels der spektralen Äquivalenz der Matrizen A und B zum Ausdruck gebracht werden. Diese ist allgemein gegeben, wenn sie den Abschätzungen

$$c_0(v, Bv) \leq (v, Av) \leq c_1(v, Bv), \quad c_0, c_1 \in \mathbb{R}, \tag{2.22}$$

genügen. Im vorliegenden Fall erhält man aus der Kontraktionsbedingung eine Ungleichungskette der Form

$$-\gamma \leq \sigma(I - B^{-1}A) \leq \gamma$$

und damit

$$1 - \gamma \leq \sigma(B^{-1}A) \leq 1 + \gamma.$$

Das Spektrum der Matrix $B^{-1}A$ kann mittels der Eigenschaft des Rayleigh-Quotienten für das Matrixpaar (A, B), dargestellt in (2.17), ersetzt werden. Man erhält

$$1 - \gamma \leq \frac{(v, Av)_2}{(v, Bv)_2} \leq 1 + \gamma$$

und daher

$$(1 - \gamma)(v, Bv)_2 \leq (v, Av)_2 \leq (1 + \gamma)(v, Bv)_2.$$

Die Bedingung der Kontraktion ist damit erfüllt, falls die Abschätzung der spektralen Äquivalenz mit den Konstanten

$$c_0 = 1 - \gamma \quad \text{und} \quad c_1 = 1 + \gamma \tag{2.23}$$

gilt. Ein beliebig gewählter Vorkonditionierer wird die Bedingung (2.22) mit den Konstanten aus (2.23) im Allgemeinen nicht erfüllen. Dieses kann aber durch geeignete Skalierung, vorausgesetzt die Konstanten c_0 und c_1 sind zugänglich, mittels des Faktors

$$\omega = \frac{2}{c_0 + c_1}$$

erreicht werden. Dies führt auf die Abschätzung

$$\|I - \omega B^{-1}A\|_A \leq \frac{c_1 - c_0}{c_0 + c_1} < 1.$$

Jedoch sind die Konstanten c_0 und c_1 im Allgemeinen aber nicht zugänglich. Dass auch ohne Kenntnis von c_0 und c_1 eine optimale Skalierung vorgenommen werden kann, wird in Abschnitt 2.3.1 erläutert. Wesentlich für die sich anschließende B_γ-Analyse ist eine Menge von Vorkonditionierern, deren Konstruktion im folgenden Lemma beschrieben wird.

Lemma 2.2.8. *Sei H eine Householderspiegelung, das heißt $H = I - 2vv^T$ mit $v \in \mathbb{R}^n$, $v^T v = 1$, und sei $\tilde{\gamma} \in [0, 1)$. Dann ist die Matrix*

$$\hat{B}^{-1} = A^{-1} + \tilde{\gamma} A^{-1/2} H A^{-1/2} \tag{2.24}$$

symmetrisch positiv definit und es gilt

$$\|I - \hat{B}^{-1}A\|_A = \tilde{\gamma}$$

Beweis. Die Symmetrie folgt, da sowohl A, $A^{-1/2}$ und H symmetrisch sind. Zum Nachweis der positiven Def nitheit sei $y \in \mathbb{R}^n$ und $z = A^{-1/2}y$. Dann gilt

$$\begin{aligned}(y, \hat{B}^{-1}y)_2 &= (y, A^{-1}y)_2 + \tilde{\gamma}(y, A^{-1/2}HA^{-1/2}y)_2 = (z,z)_2 + \tilde{\gamma}(z, Hz)_2 \\ &\geq \|z\|_2^2 - \tilde{\gamma}\|z\|_2\|Hz\|_2 = (1-\tilde{\gamma})\|z\|_2^2 > 0\end{aligned}$$

und weiterhin

$$\|(I - \hat{B}^{-1}A)y\|_A = \tilde{\gamma}\|(A^{-1/2}HA^{1/2})y\|_A = \tilde{\gamma}\|HA^{1/2}y\|_2 = \tilde{\gamma}\|y\|_A.$$

□

Im Abschnitt 2.2 wurde die Transformation des verallgemeinerten Eigenwertproblems auf das Standardeigenwertproblem als Grundlage für den Konvergenzbeweis diskutiert und vollzogen. Die unter (2.13) erhaltene Iterationsvorschrift beinhaltet nun allerdings nicht mehr den eigentlichen Vorkonditionierer B^{-1}, sondern einen aus der Umformung resultierenden mit $\tilde{B}^{-1}$ bezeichneten. Beide hängen durch die Beziehung

$$\tilde{B}^{-1} = TB^{-1}T$$

zusammen. Betrachtet man nun die Iterationsvorschrift (2.13), so ist es möglich, auch hier eine Fehlerfortpf anzungsgleichung zu formulieren,

$$\begin{aligned}\hat{u} &= u + \frac{1}{\mu(u)}\tilde{B}^{-1}(Mu - \mu(u)u) \\ \mu(u)\hat{u} &= \mu(u)u + \tilde{B}^{-1}(Mu - \mu(u)u) \\ \mu(u)\hat{u} - Mu &= (I - \tilde{B}^{-1})(\mu(u)u - Mu).\end{aligned}$$

Im Gegensatz zu den Gleichungen (2.7) und (2.11) führt die Transformation aus Satz 2.2.3 zur symmetrischen Fehlerfortpf anzungsmatrix $I - \tilde{B}^{-1}$. Daher gilt nun $\|I - \tilde{B}^{-1}\|_2 = \varrho(I - \tilde{B}^{-1})$ und die resultierende Bedingung an den Vorkonditionierer lautet

$$\|I - \tilde{B}^{-1}\|_2 \leq \gamma < 1. \tag{2.25}$$

In Anlehnung an Lemma 2.2.8 kann auch für diese Bedingung eine spezielle Menge von Vorkonditionierern konstruiert werden.

Lemma 2.2.9. *Sei H eine Householderspiegelung, das heißt $H = I - 2vv^T$ mit $v \in \mathbb{R}^n$, $v^Tv = 1$, und sei $\tilde{\gamma} \in [0, 1)$. Dann ist die Matrix*

$$\tilde{B}^{-1} = I + \tilde{\gamma}H \tag{2.26}$$

symmetrisch positiv def nit und es gilt

$$\|I - \tilde{B}^{-1}\|_2 = \tilde{\gamma}$$

Beweis. Der Beweis ergibt sich unter Verwendung derselben Argumentation wie in Lemma 2.2.8. □

Die Menge aller Vorkonditionierer $\tilde{B}^{-1}$, welche zu einem gegebenen γ die Bedingung (2.25) erfüllen, seien in der Menge

$$\mathcal{B}_\gamma = \{\tilde{B}^{-1} \in \mathbb{R}^{n\times n}; \tilde{B}^{-1} \text{ symmetrisch positiv def nit}; \|I - \tilde{B}^{-1}\|_2 \leq \gamma\}$$

zusammengefasst. Sie bildet die Grundlage der nun folgenden $\mathcal{B}_\gamma$-Analyse.

2.2.2. $\mathcal{B}_\gamma$-Analyse

Die $\mathcal{B}_\gamma$-Analyse bildet den ersten Teil der zur Konvergenzaussage führenden Betrachtungen des in Abschnitt 2.2 angeführten zweistufigen Extremalproblems. Innerhalb dieser Analyse sollen ausgehend von einer festen Iterierten u und ihres Rayleigh-Quotienten $\mu(u)$ unter Nutzung aller möglichen durch die Menge $\mathcal{B}_\gamma$ beschriebenen Vorkonditionierer die Eigenschaften der Folgeiterierten $\hat{u}$ und deren zugehörigen Rayleigh-Quotienten $\mu(\hat{u})$ untersucht werden. Dabei liegt der Schwerpunkt darin, die Folgeiterierte mit schlechtmöglichstem Konvergenzverhalten, also minimalem Zuwachs im Rayleigh-Quotienten, entsprechend Gleichung (2.21), zu charakterisieren.
Um die Analyse handlicher zu gestalten, soll eine weitere Transformation des Standardeigenwertproblems (2.12) vorgenommen werden.

Satz 2.2.10. *Mittels eines Basiswechsels kann die Konvergenzanalyse des Standardeigenwertproblems* (2.12) *auf die eines äquivalenten Standardeigenwertproblems mit Diagonalmatrix $\bar{M}$ überführt werden. Der zugehörige Rayleigh-Quotient lautet*

$$\bar{\mu}(d) = \frac{(d, \bar{M}d)_2}{(d,d)_2}$$

und die zugehörige Iterationsvorschrift des Abstiegsverfahrens ist

$$\hat{d} = d + \frac{1}{\bar{\mu}(d)} \bar{B}^{-1}(\bar{M}d - \bar{\mu}(d)d).$$

Beweis. Seien $z_1, \ldots, z_n$ die normierten Eigenvektoren von $\tilde{M}$ aus (2.14) und Z die orthogonale Matrix, deren Spalten die Vektoren z_i derart enthält, dass $\bar{M} := Z^T \tilde{M} Z = \operatorname{diag}(\mu_1, \ldots, \mu_n)$ gilt. Sei zudem d der Koeffizientenvektor von z bezüglich dieser Basis, das heißt $z = Zd$. Für den Rayleigh-Quotienten erhält man daher ausgehend von (2.15)

$$\tilde{\mu}(z) = \frac{(z, \tilde{M}z)_2}{(z,z)_2} = \frac{(Zd, \tilde{M}Zd)_2}{(Zd, Zd)_2} = \frac{(d, Z^T\tilde{M}Zd)_2}{(d,d)_2} = \frac{(d, \bar{M}d)_2}{(d,d)_2} =: \bar{\mu}(d).$$

Weiterhin liefert die Substitution $z = Zd$ in (2.14)

$$\begin{aligned} \tilde{M}Zd &= \tilde{\mu}(z)Zd \\ Z^T\tilde{M}Zd &= \tilde{\mu}(z)d \\ \bar{M}d &= \bar{\mu}(z)d. \end{aligned}$$

Die Iterationsvorschrift (2.13) wird zu

$$\begin{aligned} \hat{z} &= z + \frac{1}{\mu(z)} \tilde{B}^{-1}(\tilde{M}z - \mu(z)z) \\ Z\hat{d} &= Zd + \frac{1}{\tilde{\mu}(z)} \tilde{B}^{-1}(\tilde{M}Zd - \tilde{\mu}(z)Zd) \\ \hat{d} &= d + \frac{1}{\tilde{\mu}(z)} Z^T\tilde{B}^{-1}Z(Z^T\tilde{M}Zd - \tilde{\mu}(z)d) \end{aligned}$$

und mit $\bar{B}^{-1} := Z^T\tilde{B}^{-1}Z$ und $\bar{M} := Z^T\tilde{M}Z$ erhält man

$$\hat{d} = d + \frac{1}{\bar{\mu}(z)} \bar{B}^{-1}(\bar{M}d - \bar{\mu}(z)d),$$

die gewünschte Gestalt. □

Ergänzend sei bemerkt, dass diese Transformation zur Umsetzung der Iteration nicht realisiert werden muss, sie bildet an dieser Stelle ausschließlich ein analytisches Hilfsmittel.
Die im vorangegangenen Satz 2.2.10 dargestellte Transformation stellt im Wesentlichen einen Geometriewechsel dar, der keinen Einf uss auf die Konvergenzeigenschaften hat, jedoch im Weiteren erhebliche Vorteile bringt. Ebenso zieht die erneute Transformation des Vorkonditionierers keine Veränderung der Kontraktionsbedingung aus Gleichung (2.25) auf Grund der spektralen Äquivalenz nach sich.
Da die Eigenwerte durch die Transformationen in Satz 2.2.3 und Satz 2.2.10 nicht modif ziert werden, soll nun wieder mit den am Anfang genutzten Bezeichnungen fortgefahren werden. Das bedeutet, der Vorkonditionierer $\bar{B}^{-1}$ wird formal zu B^{-1}, die Matrix $\bar{M}$ zu M sowie der Rayleigh-Quotient $\bar{\mu}(\cdot)$ zu $\mu(\cdot)$. Auch der Koeff zientenvektor d soll wiederum mit u bezeichnet werden.
Die Basis für die analytischen Untersuchungen zur Lösung des Eigenwertproblems

$$Mu = \mu u \tag{2.27}$$

bildet somit die gradientenbasierte Iteration

$$\hat{u} = u + \frac{1}{\mu(u)} B^{-1}(Mu - \mu(u)u) \tag{2.28}$$

mit dem Rayleigh-Quotienten

$$\mu(u) = \frac{(u, Mu)_2}{(u,u)_2}, \tag{2.29}$$

wobei M eine Diagonalmatrix und B^{-1} ein Vorkonditionierer aus der Menge $\mathcal{B}_\gamma$ ist.
Die Beweisidee ist stark durch die dem Verfahren zu Grunde liegende Geometrie motiviert. Demzufolge soll diese genauer erörtert werden. Dazu sei folgende Menge def niert.

Def nition 2.2.11. *Zu einem gegebenen $u \in \mathbb{R}^n$ und einem Vorkonditionierer B, welcher der Abschätzung* (2.25) *genügt, sei mit*

$$E_\gamma(u) := \{Mu - (I - B^{-1})(Mu - \mu(u)u);\ B^{-1} \in \mathcal{B}_\gamma\}. \tag{2.30}$$

die Menge aller möglichen Iterierten $\hat{u}$ bezeichnet, welche bei der Durchführung eines Schrittes des Anstiegsverfahrens (2.28) *ausgehend von einer festen Iterierten u erhalten werden können.*

Diese Menge weist eine sehr einfache Geometrie auf.

Lemma 2.2.12. *Die Menge $E_\gamma(u)$ ist eine Kugel mit Mittelpunkt Mu und Radius $\gamma\|Mu - \mu(u)u\|_2$.*

Beweis. $E_\gamma(u)$ ist offensichtlich Teilmenge einer Kugel. Bleibt zu zeigen, dass zu gegebenen u ein $B^{-1} \in \mathcal{B}_\gamma$ existiert, so dass u auf jedes Element in $E_\gamma(u)$ abgebildet werden kann. Sei dazu $Mu + y$, $y \in \mathbb{R}^n$ ein beliebiger Punkt im Kugelinneren. Dann ist es möglich, $\tilde{\gamma}$ mit $0 \le \tilde{\gamma} \le \gamma$ aus der Beziehung

$$\|y\|_2 = \tilde{\gamma}\|Mu - \mu(u)u\|_2$$

zu bestimmen. Sei nun H eine Householderspiegelung derart, dass

$$y = -\tilde{\gamma}H(Mu - \mu(u)u)$$

gilt. Dann ist $y \in E_\gamma(u)$, da

$$\begin{aligned} y &= -\tilde{\gamma}H(Mu - \mu(u)u) \\ &= Mu - \mu(u)u - (I + \tilde{\gamma}H)Mu + \mu(u)(I + \tilde{\gamma}H)u \\ &= (I - \tilde{B}^{-1})(Mu - \mu(u)u), \end{aligned}$$

wobei $\tilde{B}^{-1}$ ein Vorkonditionierer entsprechend Gleichung (2.26) ist. □

Weiterhin ist es möglich, eine orthogonale Zerlegung der Kugel $E_\gamma(u)$ mittels ihrer generierenden Größen vorzunehmen und den Nullvektor als Folgeiterierte auszuschließen.

Lemma 2.2.13. *Sei $u \in \mathbb{R}^n$. Dann gilt*

(a) $(u, Mu - \mu(u)u)_2 = 0$

(b) $\|Mu\|_2^2 = \|\mu(u)u\|_2^2 + \|Mu - \mu(u)u\|_2^2$

(c) $0 \notin E_\gamma(u)$.

Beweis. Eine Umformung von (a) liefert

$$(u, Mu - \mu(u)u)_2 = (u, Mu)_2 - (u, \mu(u)u)_2 = (u, Mu)_2 - \frac{(u, Mu)_2}{(u, u)_2}(u, u)_2 = 0.$$

Für (b) ergibt sich unter Nutzung von (a)

$$\begin{aligned}
\|Mu\|_2^2 &= (Mu, Mu)_2 - 2(\underbrace{Mu - \mu(u)u, \mu(u)u}_{=0})_2 \\
&= (\mu(u)u, \mu(u)u)_2 + (Mu, Mu)_2 - 2(Mu, \mu(u)u)_2 + (\mu(u)u, \mu(u)u)_2 \\
&= (\mu(u)u, \mu(u)u)_2 + (Mu - \mu(u)u, Mu - \mu(u)u)_2 = \|\mu(u)u\|_2^2 + \|Mu - \mu(u)u\|_2^2.
\end{aligned}$$

Mit der Dreiecksungleichung sowie der Abschätzung $\|I - B^{-1}\|_2 < 1$ und der Beziehung

$$\|\mu(u)u\|_2^2 = \|Mu\|_2^2 - \|Mu - \mu(u)u\|_2^2 = (\|Mu\|_2 - \|Mu - \mu(u)u\|_2)(\|Mu\|_2 + \|Mu - \mu(u)u\|_2)$$

resultierend aus (b) erhält man für (c) mit $u \neq 0$

$$\begin{aligned}
\|\hat{u}\|_2 &= \|Mu + (I - B^{-1})(Mu - \mu(u)u)\|_2 \\
&\geq \|Mu\|_2 - \|Mu - \mu(u)u\|_2 \\
&= (\|Mu\|_2 + \|Mu + \mu(u)u\|_2)^{-1}\|\mu(u)u\|_2^2 > 0.
\end{aligned}$$

□

Basierend auf dieser Zerlegung können nun folgende geometrische Größen def niert werden.

Def nition 2.2.14. *Es sei zu einem $u \in \mathbb{R}^n$ und einem Vorkonditionierer B^{-1}, welcher der Abschätzung* (2.25) *genügt, der Winkel $\phi_\gamma(u)$ durch*

$$\phi_\gamma(u) = \arcsin\left(\gamma \frac{\|Mu - \mu(u)u\|_2}{\|Mu\|_2}\right)$$

gegeben. Weiterhin sei mit $\cos \sphericalangle\{y, z\} \in [0, \frac{\pi}{2}]$ *der Winkel zwischen zwei Vektoren durch*

$$\cos \sphericalangle\{y, z\} = \frac{(y, z)_2}{\|y\|_2 \, \|z\|_2}$$

def niert. Zudem kann der spezielle Winkel $\phi_1(u)$ auch als

$$\phi_1(u) = \arccos\left(\frac{(u, Mu)_2}{\|u\|_2 \, \|Mu\|_2}\right)$$

dargestellt werden.

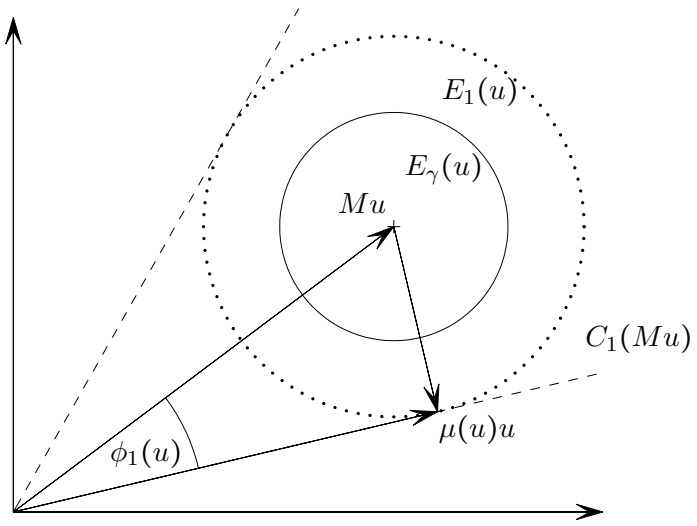

Abbildung 2.1.: Darstellung geometrischer Größen im Zusammenhang mit der Menge $E_\gamma(u)$.

Mit Hilfe dieses Winkels kann eine weitere geometrische Struktur, auf die später zurückgegriffen wird, konstruiert werden.

Def nition 2.2.15. *Zu einem gegebenen $u \in \mathbb{R}^n$ und γ aus Abschätzung* (2.8) *sei mit*

$$C_\gamma(Mu) := \{z : \sphericalangle\{z, Mu\} \leq \phi_\gamma(u)\}$$

der Kreiskegel um Mu mit Öffnungswinkel $\phi_\gamma(u)$ bezeichnet.

Eine schematische Darstellung der Geometrie und der def nierten Größen ist in Abbildung 2.1 zu erkennen.
Im nächsten Lemma sollen wesentliche Eigenschaften des Winkels $\phi_\gamma(u)$ in Abhängigkeit von γ herausgestellt und die Winkel der Iterierten u und $\hat{u}$ in Beziehung gesetzt werden.

Lemma 2.2.16. *Sei $\phi_\gamma(u)$ wie in Def nition 2.2.14. Dann gilt $\phi_\gamma(u) \leq \pi/2$ und $\sphericalangle\{\hat{u}, Mu\} \leq \phi_\gamma(u)$.*

Beweis. Die orthogonale Zerlegung aus Lemma 2.2.13 liefert

$$\|Mu\|_2^2 = \|\mu(u)u\|_2^2 + \|Mu - \mu(u)u\|_2^2.$$

Dies bedeutet $\|Mu - \mu(u)u\|_2 < \|Mu\|_2$ und daher

$$\sin \sphericalangle\{u, Mu\} = \sin \phi_1(u) = \frac{\|Mu - \mu(u)u\|_2}{\|Mu\|_2} \leq 1.$$

Die Kugel $E_\gamma(u)$ mit Mittelpunkt Mu und Radius $\gamma\|Mu - \mu(u)u\|_2$ enthält nach (2.30) $\mu(u)\hat{u}$, da $\gamma\|Mu - \mu(u)u\|_2 \geq \|I - B^{-1}\|_2\|Mu - \mu(u)u\|_2$. Das heißt aber, dass

$$\sin \sphericalangle\{\hat{u}, Mu\} \leq \gamma\frac{\|Mu - \mu(u)u\|_2}{\|Mu\|_2} < \gamma < 1.$$

□

Die gezeigte Reduktion des Öffnungswinkels mit jedem Iterationsschritt schafft aber noch keine Aussage über das Verhalten des zugehörigen Rayleigh-Quotienten. Dieser Zusammenhang soll nachfolgend dargestellt werden.

Lemma 2.2.17. *Sei $w \in \mathbb{R}^n$ mit $w \neq 0$ ein Vektor, dessen Rayleigh-Quotient $\mu(w)$ minimal unter allen Iterierten $\hat{u} \in E_\gamma(u)$ ist, also*

$$\mu(w) = \min_{\hat{u} \in E_\gamma(u)} \mu(\hat{u}).$$

Dann gilt

$$\mu(u) < \mu(w) \leq \mu(\hat{u}). \tag{2.31}$$

Beweis. Die Abschätzung $\mu(w) \leq \mu(\hat{u})$ ist nach Wahl von w als Minimierer klar. Weiterhin gilt nach Lemma 2.2.16 für den Winkel $\measuredangle\{w, Mu\} \leq \phi_\gamma(u)$ und damit $\cos \measuredangle\{w, Mu\} \geq \cos \phi_\gamma(u) > \cos \phi_1(u)$. Dies liefert die Abschätzung

$$\begin{aligned} 0 < \frac{(u, Mu)_2}{\|u\|_2 \|Mu\|_2} = \cos \phi_1(u) &< \cos \measuredangle\{w, Mu\} \\ &= \frac{(w, Mu)_2}{\|w\|_2 \|Mu\|_2} \\ &\leq \frac{(w, Mw)_2^{1/2} (u, Mu)_2^{1/2}}{\|w\|_2 \|Mu\|_2}. \end{aligned}$$

Somit ist

$$\frac{(u, Mu)_2^{1/2}}{(u, u)_2^{1/2}} < \frac{(w, Mw)_2^{1/2}}{(w, w)_2^{1/2}}$$

oder äquivalent

$$\sqrt{\mu(u)} < \sqrt{\mu(w)},$$

was wiederum

$$\mu(u) < \mu(w)$$

impliziert und damit die Behauptung zeigt. □

In Anlehnung an die unter Bemerkung 2.2.5 getroffenen Aussagen über den Verbleib des Rayleigh-Quotienten $\mu(\hat{u})$ im Intervall (μ_{i+1}, μ_i) und denen aus dem vorangegangenen Lemma 2.2.17 sei im Weiteren von der Situation

$$\mu_{i+1} < \mu(u) < \mu(w) \leq \mu(\hat{u}) < \mu_i \tag{2.32}$$

ausgegangen, vgl. Abbildung 2.2. Dies zeigt, dass selbst im Fall schlechtmöglichster Konvergenz, verkörpert durch die Folgeiterierte w, mit jedem Iterationsschritt zumindest qualitativ ein Zuwachs des Rayleigh-Quotienten stattfindet. Um quantitative Aussagen zu erhalten, sollen in den weiteren Betrachtungen die Lokalisierung und die Darstellung der oben spezifizierten Iterierten w genauer untersucht werden.

Satz 2.2.18. *Sei w der Minimierer aus 2.2.17. Dann ist w Element des Randes des unter Definition 2.2.15 angegebenen Kreiskegels $C_\gamma(Mu)$, also*

$$w \in \partial C_\gamma(Mu).$$

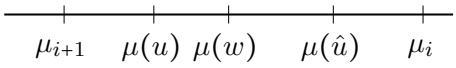

Abbildung 2.2.: Lage der Eigenwerte im Intervall $[\mu_{i+1}, \mu_i]$.

Beweis. Sei w im Inneren des Kegels. Damit wäre $\nabla\mu(w) = 0$, da w so gewählt wurde, dass $\mu(w)$ zumindest ein lokales Minimum des Rayleigh-Quotienten ist. Dies hieße gleichzeitig, dass $Mw - \mu(w)w = 0$. Diese Bedingung wird aber nur erfüllt, falls w ein Eigenvektor ist, was gleichzeitig impliziert, dass $\mu(w) = \mu_i$. Dies ist aber durch die Annahme in (2.32) ausgeschlossen. Damit muss w Element des Randes sein, also $w \in \partial C_\gamma(Mu)$. □

Diese Lokalisierung erlaubt es nun, den expliziten Zusammenhang zwischen der Iterierten u und w herzustellen.

Lemma 2.2.19. *Es existiert ein $\alpha := \alpha_\gamma(u) > -\mu_i$ und ein $\beta := \beta(\alpha)$, so dass sich die Folgeiterierte w als $(M + \alpha I)w = \beta Mu$ darstellen lässt.*

Beweis. Der Rayleigh-Quotient $\mu(\cdot) : \mathbb{R}^n \to \mathbb{R}$ ist eine glatte Funktion. Weiterhin besitzt der Kreiskegel $C_\gamma(Mu)$ eine glatte Oberfläche. Bekanntermaßen ist eine notwendige Bedingung für das Vorliegen eines Minimums, dass der Gradient $\nabla_w\mu(w)$ auf der Oberfläche des Kegels orthogonal steht und nach innen gerichtet ist. Damit ist der Gradient positiv proportional zu $Mw - \mu(w)w$. Da der Rayleigh-Quotient invariant bezüglich Skalierungen ist, sei w so gewählt, dass $(Mu - w, w) = 0$, etwa durch eine Skalierung mit dem Faktor $\delta = \frac{(Mu,w)}{(w,w)}$, gilt. Sei nun $\beta > 0$, so dass

$$\beta(Mu - w) = Mw - \mu(w)w.$$

Dies liefert

$$\beta Mu = (M + \alpha I)w,$$

wobei $\alpha = \beta - \mu(w) > -\mu(w) > -\mu_i$. Der Fall $\beta = 0$ kann ausgeschlossen werden, da w sonst ein Eigenvektor wäre, was nach Bemerkung 2.2.5 ausgeschlossen ist. □

Als ein letzter Aspekt der B_γ-Analyse soll ein weiterer wichtiger Zusammenhang zwischen der Iterierten u und dem Minimierer w herausgestellt werden.

Lemma 2.2.20. *Sei w wie in Lemma 2.2.17. Dann gilt*

$$u \in \text{span}\{u_i, u_{i+1}\} \Rightarrow w \in \text{span}\{u_i, u_{i+1}\}.$$

Beweis. Nach Lemma 2.2.19 ergibt sich w als

$$\beta Mu = (M + \alpha I)w.$$

Dies heißt, w ist Lösung des linearen Gleichungssystems

$$\beta(I - \alpha M^{-1})w = u.$$

Da M nach eingangs vollzogener Transformation, vergleiche Satz 2.2.10, Diagonalgestalt besitzt, hat dieses Gleichungssystem die Form

$$\underbrace{\beta \begin{pmatrix} 1 - \frac{\alpha}{\mu_1} & & \\ & \ddots & \\ & & 1 - \frac{\alpha}{\mu_n} \end{pmatrix}}_{D} w = u.$$

Damit ist D regulär, falls $\alpha \neq \mu_j$. Man erhält für $u = a_i u_i + a_{i+1} u_{i+1}$, mit $a_i, a_{i+1} \neq 0$, da u kein Eigenvektor oder der Nullvektor ist, die Darstellung

$$w = \beta^{-1}\left[\left(1 - \frac{\alpha}{\mu_i}\right) a_i u_i + \left(1 - \frac{\alpha}{\mu_{i+1}}\right) a_{i+1} u_{i+1}\right]$$

und damit

$$w \in \text{span}\{u_i, u_{i+1}\},$$

die Behauptung.
Für den Fall $\alpha = \mu_j$, wobei nach Lemma 2.2.19 $j > i+1$ gilt, ist D singulär. Daher hat $w = w_{|\alpha=\mu_j}$ die Form

$$w_{|\alpha=\mu_j} = t_i a_i u_i + t_{i+1} a_{i+1} u_{i+1} + \underbrace{a_j u_j}_{\in Ker(D)} \tag{2.33}$$

mit

$$t_i = \beta^{-1}\left(\frac{\mu_i}{\mu_i - \mu_j}\right) > 0 \quad \text{und} \quad t_{i+1} = \beta^{-1}\left(\frac{\mu_{i+1}}{\mu_{i+1} - \mu_j}\right) > 0,$$

da $\beta > 0$ und $0 < \mu_j < \mu_{i+1} < \mu_i$. Dies zeigt auch $t_i < t_{i+1}$, denn

$$t_i - t_{i+1} = \beta^{-1}\left[\left(\frac{\mu_i}{\mu_i - \mu_j}\right) - \left(\frac{\mu_{i+1}}{\mu_{i+1} - \mu_j}\right)\right] = \frac{(\mu_{i+1} - \mu_i)\mu_j}{(\mu_i - \mu_j)(\mu_{i+1} - \mu_j)} < 0.$$

Ein Vergleich der Rayleigh-Quotienten von $u = a_i u_i + a_{i+1} u_{i+1}$ und $w_{|\alpha=\mu_j}$ mit $a_j = 0$ in (2.33) liefert

$$\begin{aligned} \mu(u) - \mu(w_{|\alpha=\mu_j}) &= \frac{\sum_{k=i,i+1} \mu_k u_k^2}{\sum_{k=i,i+1} u_k^2} - \frac{\sum_{k=i,i+1} \mu_k t_k^2 u_k^2}{\sum_{k=i,i+1} t_k^2 u_k^2} \\ &= \frac{1 - \frac{\mu_i}{\mu_{i+1}}(t_i^2 - t_{i+1}^2)}{(\sum_{k=i,i+1} u_k^2)(\sum_{k=i,i+1} t_k^2 u_k^2)} > 0, \end{aligned}$$

da sowohl $(1 - \frac{\mu_i}{\mu_{i+1}}) < 0$ als auch $t_i^2 - t_{i+1}^2 < 0$. Dies führt auf die Abschätzung

$$\mu(u) > \mu(w_{|\alpha=\mu_j}). \tag{2.34}$$

Vergleicht man nun weiterhin die Rayleigh-Quotienten von $w_{|\alpha=\mu_j}$ und $\bar{w}_{|\alpha=\mu_j} = t_i a_i u_i + t_{i+1} a_{i+1} u_{i+1} + a_j u_j$ mit $a_j \neq 0$ erhält man

$$\begin{aligned} \mu(w_{|\alpha=\mu_j}) - \mu(\bar{w}_{|\alpha=\mu_j}) &= \frac{\sum_{k=i,i+1} \mu_k t_k^2 u_k^2}{\sum_{k=i,i+1} u_k^2} - \frac{\sum_{k=i,i+1} \mu_k t_k^2 u_k^2 + \mu_j a_j^2 u_j^2}{\sum_{k=i,i+1} t_k^2 u_k^2 + a_j^2 u_j^2} \\ &= \frac{(\mu_i - \mu_j) t_k^2 u_k^2 + (\mu_{i+1} - \mu j) t_{i+1}^2 u_{i+1}^2}{(\sum_{k=i,i+1} t_k^2 u_k^2)(\sum_{k=i,i+1} t_k^2 u_k^2 + a_j^2 u_j^2)} > 0, \end{aligned}$$

da $(\mu_i - \mu_j)$ und $(\mu_{i+1} - \mu_j)$ jeweils positiv sind. Das heißt

$$\mu(w_{|\alpha=\mu_j}) > \mu(\bar{w}_{|\alpha=\mu_j}).$$

Mit der Ungleichung (2.34) ergibt sich

$$\mu(u) > \mu(w_{|\alpha=\mu_j}) > \mu(\bar{w}_{|\alpha=\mu_j})$$

und damit ein Widerspruch zu Lemma 2.2.17, welches besagt, dass $\mu(u) < \mu(w)$. Daher kann der Fall $\alpha = \mu_j$ nicht eintreten und w besitzt nicht die Gestalt aus (2.33). Somit ist $w \in \text{span}\{u_i, u_{i+1}\}$. □

Dies bildet den Abschluß der B_γ-Analyse. Als Resultat bleibt festzuhalten, dass zu einem gegebenen Vorkonditionierer B^{-1} und der daraus resultierenden Kontraktionskonstanten γ der Fall der schlechtmöglichsten Konvergenz ausgehend von einer festen Iterierten u eintritt, falls die resultierende Iterierte $\hat{u} = w$ mit minimalem Zuwachs im Rayleigh-Quotienten auf dem Rand der Menge $E_\gamma(u)$ liegt. Weiterhin zeigt sich, dass bei der speziellen Lage der Iterierten u im Unterraum aufgespannt von den Eigenvektoren u_i und u_{i+1} auch die Folgeiterierte w mit schlechtmöglichster Konvergenz, also minimalem Zuwachs im Rayleigh-Quotienten, Element dieses Unterraums ist.

2.2.3. $\mathcal{L}_\rho$-Analyse

Bei den Betrachtungen innerhalb der B_γ-Analyse wurde ausgehend von einer festen Iterierten u argumentiert und resultierende Eigenschaften der Folgeiterierten herausgestellt. Es existiert aber nicht zwangsläuf g nur ein Vektor $u \in \mathbb{R}^n$ mit dem Rayleigh-Quotienten $\mu(u)$, sondern mehrere die denselben Rayleigh-Quotienten $\mu(u)$ aufweisen. Die $\mathcal{L}_\rho$-Analyse als zweiter Teil des anfangs angesprochenen zweistuf gen Extremalproblems (2.21) verfolgt nun das Ziel, unter all diesen Iterierten u mit festem Rayleigh-Quotienten $\mu(u) = \rho$ diejenigen zu klassif zieren, welche unter Anwendung der Iteration (2.28) zum geringsten Zuwachs im Rayleigh-Quotienten der Folgeiterierten $\hat{u} = w$ führen. Mittels der Untersuchung des Gradientenf usses wird sich dabei herausstellen, dass die schlechtmöglichste Konvergenz eintritt, falls dabei sowohl u als auch w in einem bestimmten Unterraum liegen. Dies führt zu einer niedrigdimensionalen Konvergenzanalyse, an deren Ende der Beweis zu Satz 2.2.6 formuliert werden kann. Für die weiteren Erörterungen seien die oben beschriebenen Iterierten in folgender Niveaumenge zusammengefasst.

Def nition 2.2.21. *Sei $\rho \in (\mu_{i+1}, \mu_i)$. Dann enthält die Menge $\mathcal{L}(\rho)$ alle Iterierten $u \in \mathbb{R}^n$ mit dem festen Rayleigh-Quotienten ρ, also*

$$\mathcal{L}(\rho) := \{u \in \mathbb{R}^n;\ u \neq 0;\ \mu(u) = \rho\}. \tag{2.35}$$

Hiervon ausgehend soll ein Zusammenhang zwischen der Lokalisierung einer Iterierten $u \in \mathcal{L}(\rho)$ und der Minimalität des Gradienten $\|\nabla\mu(u)\|_2$ hergestellt werden. Dabei ist die Templesche Ungleichung hilfreich, [70].

Lemma 2.2.22 (Templesche Ungleichung)**.** *Sei $H \in \mathbb{R}^{n\times n}$ eine symmetrisch positiv def nite Matrix. Weiterhin seien μ_1 und μ_2 die kleinsten Eigenwerte von H. Ist zudem für $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$ die Ungleichung $(x, Hx)_2 < \tilde{\mu}_2 < \mu_2$ $(\tilde{\mu}_2 \in \mathbb{R})$ erfüllt, so gilt*

$$\mu_1 \geq (x, Hx)_2 - \frac{(x, H^2x)_2 - (x, Hx)_2^2}{\tilde{\mu}_2 - (x, Hx)_2}.$$

Beweis. Da μ_1 ein Eigenwert von H und $\mu_1 < \tilde{\mu}_2$ gilt, ist die Matrix $(H - \mu_1 I)(H - \tilde{\mu}_2 I)$ positiv semidef nit. Das heißt,

$$\begin{aligned} 0 &\leq (x, (H - \mu_1 I)(H - \tilde{\mu}_2 I)x)_2 \\ &= (x, H^2x)_2 - \mu_1(x, Hx)_2 - \tilde{\mu}_2(x, Hx)_2 + \mu_1\tilde{\mu}_2(x, x)_2 \end{aligned}$$

und damit

$$\mu_1(x, (H - \tilde{\mu}_2 I)x)_2 \leq (x, (H - \tilde{\mu}_2 I)Hx)_2.$$

Nach Voraussetzung gilt $(x, (H - \tilde{\mu}_2 I)x)_2 < 0$ und daher

$$\mu_1 \geq \frac{\tilde{\mu}_2(x, Hx)_2 - (x, H^2x)_2}{\tilde{\mu}_2 - (x, Hx)_2} = (x, Hx)_2 - \frac{(x, H^2x)_2 - (x, Hx)_2^2}{\tilde{\mu}_2 - (x, Hx)_2}.$$

□

Mit Hilfe der Templeschen Ungleichung können nun die Minima lokalisiert werden.

Lemma 2.2.23. *Sei $\rho \in (\mu_{i+1}, \mu_i)$ fest und $\mathcal{L}(\rho)$ wie in Definition 2.2.21. Dann wird für $u \in \mathcal{L}(\rho)$ die Größe*

$$\|\nabla\mu(u)\|_2 \, \|u\|_2$$

minimal, wenn $u \in \text{span}\{u_i, u_{i+1}\}$.

Beweis. Der Gradient des Rayleigh-Quotienten lautet

$$\nabla\mu(u) = \frac{2}{(u,u)_2}\,(Mu - \mu(u)u)$$

und damit gilt für seine Norm

$$\|\nabla\mu(u)\|_2 = \frac{2}{\|u\|_2^2}\|Mu - \mu(u)u\|_2.$$

Da $u \in \mathcal{L}(\rho)$ heißt dies

$$\|\nabla\mu(u)\|_2 \, \|u\|_2 = \frac{2\|Mu - \rho u\|_2}{\|u\|_2}.$$

Die rechte Seite der Gleichung kann auf die Gestalt

$$\begin{aligned} \frac{\|Mu-\rho u\|_2^2}{\|u\|_2^2} &= \frac{(u,M^2u)_2 + \rho^2(u,u)_2 - 2\rho(u,Mu)_2}{(u,u)_2} \\ &= \rho^2 - 2\rho\frac{(u,Mu)_2}{(u,u)_2} + \frac{(u,M^2u)_2}{(u,u)_2} = \frac{(u,M^2u)_2}{(u,u)_2} - \rho^2 \end{aligned} \tag{2.36}$$

umgeformt werden. Aus der Templeschen Ungleichung, siehe Lemma 2.2.22, wobei μ_{i+1} und μ_i dort μ_1 und $\tilde{\mu}_2$ ersetzen, ergibt sich

$$(u, (M - \mu_i I)Mu)_2 \geq \mu_{i+1}(u, (M - \mu_i I)u)_2$$

beziehungsweise

$$(u, M^2u)_2 - (u, \mu_i Mu)_2 \geq \mu_{i+1}(u, Mu)_2 - \mu_{i+1}(u, \mu_i u)_2$$

und nach Division durch $(u,u)_2$ sowie mit dem Rayleigh-Quotienten der Niveaumenge $\rho = \mu(u) = \frac{(u,Mu)_2}{(u,u)_2}$ erhält man

$$\begin{aligned} \frac{(u,M^2u)_2}{(u,u)_2} &\geq \mu_{i+1}\rho - \mu_{i+1}\mu_i + \mu_i\rho - \rho^2 + \rho^2 \\ \frac{(u,M^2u)_2}{(u,u)_2} - \rho^2 &\geq (\mu_i - \rho)(\rho - \mu_{i+1}). \end{aligned}$$

Mit Gleichung (2.36) heißt dies

$$\frac{\|Mu - \rho u\|_2^2}{\|u\|_2^2} \geq (\mu_i - \rho)(\rho - \mu_{i+1}).$$

Dabei wird die linke Seite und damit ebenso $\|\nabla\mu(u)\| \, \|u\|$ minimal, falls Gleichheit angenommen wird. Die Templesche Ungleichung zeigt, dass dies äquivalent zum Fall

$$(M - \mu_i I)(M - \mu_{i+1}I)u = 0$$

ist. Auf Grund der Diagonalgestalt von M ist dies aber genau dann der Fall, wenn $u \in \text{span}\{u_i, u_{i+1}\}$. □

Lemma 2.2.24. *Mit den Voraussetzungen von Lemma 2.2.23 wird der Ausdruck*

$$\phi_1(u) - \phi_\gamma(u)$$

minimal, falls $u \in \operatorname{span}\{u_i, u_{i+1}\}$.

Beweis. Nach Def nition 2.2.14 gilt für den Öffnungswinkel

$$\phi_\gamma(u) = \arcsin\left(\gamma \frac{\|Mu - \mu(u)u\|_2}{\|Mu\|_2}\right).$$

Mit der orthogonalen Zerlegung aus Lemma 2.2.16, gegeben durch

$$\|Mu\|_2^2 = \|\mu(u)u\|_2^2 + \|Mu - \mu(u)u\|_2^2,$$

kann der Nenner des Arguments substituiert werden und man erhält unter Nutzung, dass $\mu(u) = \rho$

$$\phi_\gamma(u) = \arcsin\left(\gamma \frac{\|Mu - \rho u\|_2^2/\|u\|_2^2}{\rho^2 + \|Mu - \rho u\|_2^2/\|u\|_2^2}\right).$$

Da $\rho > 0$ und $\gamma \in (0,1)$ gilt für das Argument

$$t^2(u) := \gamma \frac{\|Mu - \rho u\|_2^2/\|u\|_2^2}{\rho^2 + \|Mu - \rho u\|_2^2/\|u\|_2^2} \in (0,1).$$

Die Funktion

$$f_\gamma(t(u)) := \phi_1(u) - \phi_\gamma(u) = \arcsin(t^2(u)) - \arcsin(\gamma t^2(u))$$

kann mit Hilfe der Additionstheoreme und der verkürzten Schreibweise $t = t(u)$ in

$$f_\gamma(t) = \arcsin \underbrace{\left(t^2\sqrt{1-\gamma t^2} - \gamma t^2\sqrt{1-t^2}\right)}_{g_\gamma(t)}$$

umgeformt werden. Das Argument

$$g_\gamma(t) = t^2\sqrt{1-\gamma t^2} - \gamma t^2\sqrt{1-t^2}$$

ist eine monoton steigende Funktion, da für $t \in (0,1)$ und $\gamma \in (0,1)$ die Abschätzung

$$\begin{aligned} g_\gamma'(t) &= -\frac{(2\gamma^2t^2-1)\sqrt{1-t^2} + (\gamma - 2\gamma t^2)\sqrt{1-\gamma^2t^2}}{\sqrt{1-\gamma^2t^2}\sqrt{1-t^2}} \\ &\geq -\frac{2\gamma^2t^2 - 1 + \gamma - 2\gamma t^2}{\sqrt{1-\gamma^2t^2}\sqrt{1-t^2}} \\ &= -\frac{(\gamma-1)(1+2\gamma t^2)}{\sqrt{1-\gamma^2t^2}\sqrt{1-t^2}} > 0 \end{aligned}$$

gilt. Damit ist auch $f_\gamma(t)$ monoton steigend und wird minimal, falls der Zähler $\|Mu-\rho u\|_2^2/\|u\|_2^2$ minimal wird. Dies ist nach vorangegangenen Lemma 2.2.23 aber der Fall, falls $u \in \operatorname{span}\{u_i, u_{i+1}\}$. □

Basierend auf den bisher erhaltenen Ergebnissen kann nun mittels des negativen normierten Gradientenf usses eine Lokalisierung der Iterierten $u \in \mathcal{L}(\rho)$, für die schlechtmöglichste Konvergenz auftritt, vorgenommen werden. Dies soll in mehreren Schritten erfolgen. Zunächst wird die Differentialgleichung des Gradientenf usses angegeben und einige ihrer Eigenschaften aufgezeigt. Im Anschluss daran wird ein Maß def niert, welches abschließend hilft, die Lokalisierung der Folgeiterierten w mit schlechtmöglichster Konvergenz vorzunehmen. Es wird sich zeigen, dass dies im Unterraum span$\{u_i, u_{i+1}\}$ der Fall ist, was eine niedrigdimensionale Analyse zulässt. Dazu sei folgende Menge charakterisiert.

Def nition 2.2.25. *Basierend auf der Menge $\mathcal{L}(\rho)$ entsprechend Def nition 2.2.21 und der des Kreiskegels $C_\gamma(Mu)$ aus 2.2.15 sei mit*

$$I_\gamma(\rho) = \{w \in \mathbb{R}^n;\ w \in \arg\min \mu(C_\gamma(Mu));\ u \in \mathcal{L}(\rho)\} \tag{2.37}$$

die Menge von Minimierern w entsprechend Lemma 2.2.17 bezeichnet.

Es kann nun der Gradientenf uss angegeben und untersucht werden.

Lemma 2.2.26. *Gegeben sei das Anfangswertproblem für den negativen Gradientenf uss $y(t)$ durch*

$$\begin{aligned} y'(t) &= -\frac{\nabla\mu(y(t))}{\|\nabla\mu(y(t))\|_2}, \quad t \geq 0,\ y(t) \in \mathbb{R}^n, \\ y(0) &= w, \quad w \in I_\gamma(\rho). \end{aligned} \tag{2.38}$$

Dann existiert ein eindeutiges $\bar{t} > 0$ mit

$$\mu(y(\bar{t})) = \rho,$$

wobei ρ die Niveaumenge generierende Größe aus Def nition 2.2.21 ist. Weiterhin gilt $\|y(t)\|_2 = \|w\|_2$ für $t \in [0, \bar{t}]$.

Beweis. Sei $f(t) : \mathbb{R} \to \mathbb{R}$, gegeben durch

$$f(t) = \mu(y(t)),$$

die Rayleigh-Quotienten Funktion. Dann ist f monoton fallend, da

$$\begin{aligned} \frac{d}{dt}\mu(y(t)) &= \frac{d}{dt}\frac{(y(t), My(t))_2}{\|y(t)\|_2^2} \\ &= \frac{2(My(t), y'(t))_2\|y(t)\|_2^2 - 2(y(t), y'(t))_2(y(t), My(t))_2}{\|y(t)\|_2^4} \\ &= \frac{2}{\|y(t)\|_2^2}(My(t) - \mu(y(t)), y'(t))_2 \\ &= (\nabla\mu(y(t)), y'(t))_2 \\ &= \left(\nabla\mu(y(t)), -\frac{\nabla\mu(y(t))}{\|\nabla\mu(y(t))\|_2}\right)_2 = -\|\nabla\mu(y(t))\|_2 \leq 0, \end{aligned}$$

wobei die Def nition des Rayleigh-Quotienten sowie die gegebenen Differentialgleichung (2.38) genutzt wurden. Weiterhin ist f sogar streng monoton fallend, denn der Fall

$$-\|\nabla\mu(y(t))\|_2 = 0$$

tritt nur ein, falls der Gradient verschwindet. Da dies aber nur für die Eigenvektoren gilt, in denen das Anstiegsverfahren nach Konstruktion stationär ist, wäre $\mu(y(t))$ ein Eigenwert. Da aber $[\rho, \mu(y(0))] \subset$

(μ_{i+1}, μ_i) ist dies nach Voraussetzung, dass das Intervall (μ_{i+1}, μ_i) keinen weiteren Eigenwert enthält, ausgeschlossen. Demzufolge existiert ein eindeutiges $\bar{t}$ mit $\mu(y(\bar{t})) = \rho$, was die Behauptung zeigt.
Für die Norm $\|y(t)\|_2$ gilt weiterhin

$$\frac{d}{dt}\|y(t)\|_2 = \frac{d}{dt}\sum_{i=1}^{n} y_i^2(t) = 2\sum_{i=1}^{n} y_i(t)y_i'(t) = 2(y(t), y'(t))_2.$$

Entsprechend der Differentialgleichung folgt hierbei für die rechte Seite

$$\begin{aligned}
(y(t), y'(t))_2 &= \frac{1}{\|\nabla\mu(y(t))\|_2}(y(t), \nabla\mu(y(t)))_2 \\
&= \frac{1}{\|\nabla\mu(y(t))\|_2}\left(y(t), \frac{2}{(y(t), y(t))_2}(My(t) - \mu(y(t))y(t))\right) \\
&= \frac{2}{\|\nabla\mu(y(t))\|_2\|y(t)\|_2^2}\underbrace{\left((y(t), My(t))_2 - \left(y(t), \frac{(y(t), My(t))_2}{(y(t), y(t))_2}(y(t))\right)_2\right)_2}_{=0} = 0.
\end{aligned}$$

Daher bleibt die Norm erhalten und es gilt $\|y(t)\|_2 = \|w\|_2$ für $t \in [0, \bar{t}]$. □

Lemma 2.2.27. *Sei $y(t)$ Lösung des Anfangswertproblems (2.38) mit $y(\bar{t}) = \rho$. Dann gilt für die Kurve L, def niert durch*

$$L := \{y(t), 0 \le t \le \bar{t}\},$$

dass ihre Länge durch $\bar{t}$ gegeben ist, also

$$\text{Länge}(L) = \bar{t}.$$

Beweis. Die Kurve L ist rektif zierbar, da $y(t)$ mit der durch die Differentialgleichung gegebenen stetigen Ableitung $y'(t)$ stetig differenzierbar ist. Daher ist die Länge der Kurve L durch

$$\text{Länge}(L) = \int_0^{\bar{t}} \|y'(t)\|dt = \int_0^{\bar{t}} \left\|\frac{\nabla\mu(y(t))}{\|\nabla\mu(y(t))\|}\right\| = \int_0^{\bar{t}} 1dt = \bar{t}$$

gegeben. □

Bemerkung 2.2.28. *Ohne Beschränkung der Allgemeinheit wird im Weiteren von $\|y(t)\|_2 = 1$ ausgegangen, das heißt, die Kurve L liegt auf Grund der gezeigten Norminvarianz auf der Oberf äche der Einheitskugel. Zudem seien im Folgenden alle zum Unterraum aufgespannt von u_i und u_{i+1} korrespondierenden Größen als spezielle Wahl mit * gekennzeichnet.*

Der Nachweis, dass der Zuwachs des Rayleigh-Quotienten im Unterraum $\text{span}\{u_i, u_{i+1}\}$ minimal wird, soll in zwei Schritten erfolgen. Zunächst betrachtet man die Längen der Kurven L und L^*, anschließend werden die Normen der Gradienten verglichen. Ziel ist es, den Rayleigh-Quotienten $\mu(w^*) = \mu(y^*(0))$ durch $\mu(w) = \mu(y(0))$ von oben zu beschränken, um die Minimalität im speziell gewählten Unterraum zu zeigen.

Lemma 2.2.29. *Sei $y(t)$ Lösung der Differentialgleichung (2.38) zum Anfangswert $y(0) = w$, $w \in I_\gamma(\rho)$ und $y^*(t^*)$ die Lösung zum Anfangswert $y^*(0) = w^*$, $w^* \in I_\gamma(\rho)$. Dann gilt für die Längen der resultierenden Kurven L und L^**

$$\bar{t}^* \le \bar{t}$$

Beweis. Nach Lemma 2.2.26 ist mit gegebenen Anfangswert $y(0) = w$ die eindeutige Lösung $y(\bar{t}) = u$ der Differentialgleichung bestimmt. Auf Grund der Eindeutigkeit ist ebenso jedem u ein eindeutiger Anfangswert w zugeordnet. Sei nun $u^* \in \text{span}\{u_i, u_{i+1}\}$. Dann ist, nach Lemma 2.2.20, auch $w^* \in \text{span}\{u_i, u_{i+1}\}$. Des Weiteren liegt ebenso die gesamte Lösungskurve L^* aus Lemma 2.2.27 in $\text{span}\{u_i, u_{i+1}\}$, da dieser ein invarianter Unterraum für den Gradienten des Rayleigh-Quotienten ist. An den Endpunkten gilt nach Def nition $\mu(y(\bar{t})) = \mu(u) = \rho = \mu(u^*) = \mu(y^*(\bar{t}^*))$.
Die Abschätzung der Längen kann nun mittels der Öffnungswinkel hergeleitet werden. Nach Lemma 2.2.24 gilt für $u^* \in \text{span}\{u_i, u_{i+1}\}$

$$\phi_1(u^*) - \phi_\gamma(u^*) \leq \phi_1(u) - \phi_\gamma(u). \tag{2.39}$$

Die rechte Seite kann dabei durch

$$\phi_1(u) - \phi_\gamma(u) \leq \sphericalangle\{y(0), y(\bar{t})\}$$

abgeschätzt werden, denn nach Voraussetzung gilt $y(0) = w \in \partial C_\gamma(Mu)$ und $y(\bar{t}) \notin C_\gamma(Mu)$, da $\mu(w) > \rho = \mu(y(\bar{t}))$. Hier ist wiederum

$$\sphericalangle\{y(0), y(\bar{t})\} \leq \text{Länge}(L) = \bar{t},$$

da nach der Annahme aus Bemerkung 2.2.28 $\|y(t)\|_2 = 1$ gilt und die kürzeste Verbindung zweier Punkte auf der Einheitskugel gerade dem Bogen zwischen diesen entspricht. Somit besitzt jede Kurve mindestens die Länge L.
Für die linke Seite der Gleichung (2.39), also für die spezielle Wahl der Größen $u, w \in \text{span}\{u_i, u_{i+1}\}$, gelten alle Abschätzungen mit Gleichheit, da die Kurve L^* im Schnitt der Oberf äche der Einheitskugel mit der von u_i und u_{i+1} aufgespannten Ebene liegt. Dies heißt aber, dass L^* der kürzeste Pfad von u^* zu w^* ist. Zusammengefasst erhält man die Ungleichungskette

$$\begin{aligned} \bar{t}^* = \text{Länge}(L^*) &= \sphericalangle\{y^*(0), y^*(\bar{t}^*)\} = \sphericalangle\{w^*, u^*\} = \phi_1(u^*) - \phi_\gamma(u^*) \\ &\leq \phi_1(u) - \phi_\gamma(u) \\ &\leq \sphericalangle\{y(0), y(\bar{t})\} \leq \text{Länge}(L) = \bar{t}. \end{aligned}$$

Somit ist die Länge der Kurve L am geringsten, falls $u, w \in \text{span}\{u_i, u_{i+1}\}$, also $u = u^*$ und damit $w = w^*$. □

Nach dieser Abschätzung für die Längen ist es möglich, auch die Norm der Gradienten in ein Verhältnis zu setzen. Dazu benötigt man zusätzlich folgendes Lemma.

Lemma 2.2.30. *Seien $f, g : [0, b] \to \mathbb{R}$, $b > 0$ streng monoton wachsende und stetig differenzierbare Funktionen. Zudem existiere ein $a \in [0, b]$ mit $f(a) = g(b)$. Falls für alle $\alpha, \beta \in [0, b]$ mit $f(\alpha) = g(\beta)$ gilt*

$$f'(\alpha) \leq g'(\beta),$$

dann ist für alle $\xi \in [0, a]$ auch

$$f(a - \xi) \geq g(b - \xi),$$

vergleiche auch Abbildung 2.3.

Beweis. Da f und g streng monoton wachsende und stetig differenzierbare Funktionen sind, existieren die Umkehrfunktionen g^{-1} beziehungsweise f^{-1} und deren Ableitungen $(g^{-1})'$, $(f^{-1})'$. Unter Zuhilfenahme dieser und des Fundamentalsatzes der Differential- und Integralrechnung kann ξ in der Form

$$\xi = b - (b - \xi) = g^{-1}(g(b)) - g^{-1}(g(b - \xi)) = \int_{g(b-\xi)}^{g(b)} (g^{-1})'(z)dz$$

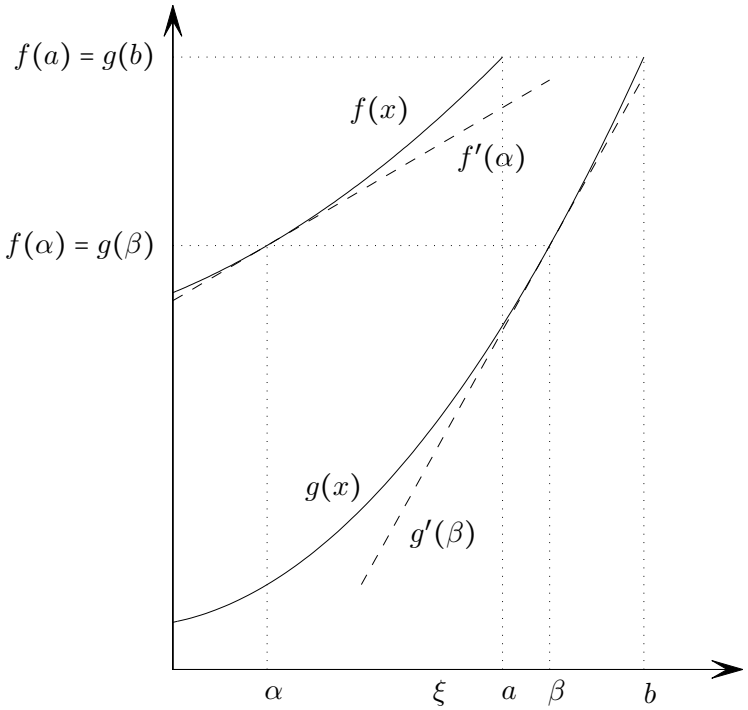

Abbildung 2.3.: Skizze zur Integration inverser Funktionen (Lemma 2.2.30).

geschrieben werden. Ebenso kann mit der Funktion f für $\xi = a - (a - \xi)$ verfahren werden. Daher ist mit $f(a) = g(b)$

$$\xi = \int_{g(b-\xi)}^{g(b)} (g^{-1})'(z)dz = \int_{f(a-\xi)}^{g(b)} (f^{-1})'(z)dz. \tag{2.40}$$

Sei nun $z = f(\alpha) = g(\beta)$, dann gilt für die Ableitungen $(g^{-1})'(z) \leq (f^{-1})'(z)$. Da f und g nach Voraussetzung streng monoton wachsende Funktionen sind, sind die Integranden jeweils positiv. Nach Konstruktion ist weiterhin $g(b - \xi) < g(b)$ und $f(a - \xi) < f(a) = g(b)$. Abschließend erhält man die Behauptung $f(a - \xi) \geq g(b - \xi)$ aus folgender Überlegung. Sei $f(a - \xi) \leq g(b - \xi) \leq g(b)$. Dann kann Gleichung (2.40) als

$$\int_{g(b-\xi)}^{g(b)} (g^{-1})'(z)dz = \int_{f(a-\xi)}^{g(b-\xi)} (f^{-1})'(z)dz + \int_{g(b-\xi)}^{g(b)} (f^{-1})'(z)dz$$

formuliert werden. Eine Umsortierung liefert

$$-\int_{f(a-\xi)}^{g(b-\xi)} (f^{-1})'(z)dz = \int_{g(b-\xi)}^{g(b)} ((f^{-1})' - (g^{-1})')(z)dz.$$

Dabei sind die Integranden wiederum positive Funktionen. Damit führt die Annahme $f(a-\xi) \leq g(b-\xi)$ zum Widerspruch, es gilt also $f(a - \xi) \geq g(b - \xi)$. □

Die im Vorfeld angegebenen Eigenschaften der Differentialgleichung und der Kurve L zusammen mit der Aussage des Lemmas 2.2.30 erlauben es nun, eine Lokalisierung des Minimierers w vorzunehmen.

Satz 2.2.31. *Sei, wie in Satz 2.2.18 gezeigt, $w \in \partial C_\gamma(Mu)$ mit der Darstellung aus Lemma 2.2.19 und $\mathcal{L}(\rho)$ beziehungsweise $I_\gamma(\rho)$ wie in Definition 2.2.21 und 2.2.25. Dann gilt*

$$\arg\min I_\gamma(\rho) \in \operatorname{span}\{u_i, u_{i+1}\}, \tag{2.41}$$

also der Fall schlechtester Konvergenz tritt für $w \in \operatorname{span}\{u_i, u_{i+1}\}$ ein.

Beweis. Für die Normen der Gradienten gilt nach Lemma 2.2.23

$$\|\nabla\mu(y^*(t^*))\|\,\|u^*\| \le \|\nabla\mu(y(t))\|\,\|u\|.$$

Auf Grund der in Lemma 2.2.26 gezeigten Norminvarianz der Lösung $y(t)$ und der Normierungsbedingung aus Bemerkung 2.2.28 folgt für die Norm des negativen Gradienten

$$-\|\nabla\mu(y^*(t^*))\| \ge -\|\nabla\mu(y(t))\|.$$

Der Gradientenfuß $y(t)$ aus Gleichung (2.38) ist eine stetig differenzierbare Funktion. Daher existieren Werte $0 \le t^* \le \bar{t}^*$ und $0 \le t \le \bar{t}$, in denen

$$\mu(y^*(t^*)) = \mu(y(t))$$

gilt. Dadurch sind mit $f = \mu(y^*(t))$ und $g = \mu(y(t))$ die Voraussetzungen des Lemmas 2.2.30 erfüllt. Wählt man dort $b = \bar{t}$ und $a = \xi = \bar{t}^*$ und beachtet, dass die Rayleigh-Quotienten Funktion streng monoton fallend ist, gilt

$$\mu(w^*) = \mu(y^*(0)) \le \mu(y(\bar{t} - \bar{t}^*)).$$

Gleichzeitig gilt aber nach Lemma 2.2.29, dass $\bar{t} - \bar{t}^* \ge 0$ und somit

$$\mu(y(\bar{t} - \bar{t}^*)) \le \mu(y(0)) = \mu(w).$$

Zusammengefasst heißt dies

$$\mu(w^*) = \mu(y^*(0)) \le \mu(y(\bar{t} - \bar{t}^*)) \le \mu(y(0)) = \mu(w).$$

Damit wird der geringste Zuwachs im Rayleigh-Quotienten und damit schlechtmöglichste Konvergenz erhalten, wenn für $w \in I_\gamma(\rho)$ auch $w \in \operatorname{span}\{u_i, u_{i+1}\}$ und damit $w = w^*$ gilt. □

Essentiell bleibt die Erkenntnis, dass der Fall schlechtmöglichster Konvergenz eintritt, falls für die betrachteten Vektoren $u, w \in \operatorname{span}\{u_i, u_{i+1}\}$ gilt. Dies bedeutet aber, dass die weitere Analyse auf diesen zweidimensionalen Unterraum beschränkt werden kann. Gleichzeitig liefert dies den Beweis zum am Anfang angegebenen Satz 2.2.6. Dieser sei nochmals formuliert.

Satz 2.2.32. *Zur Lösung des Standardeigenwertproblems* (2.27) *sei das Anstiegsverfahren aus* (2.28) *für $u \in \mathbb{R}^n$ verwendet. Der Vorkonditionierer B^{-1} sei derart, dass die aus ihm resultierende Fehlerfortpflanzungsmatrix eine Kontraktion entsprechend Gleichung* (2.25) *ist. Für den Rayleigh-Quotienten der Iterierten $u \in \mathbb{R}^n$ gelte zudem $\mu_{i+1} < \mu(u) < \mu_i$. Dann gilt für den Rayleigh-Quotienten der Folgeiterierten $\mu(\hat{u})$ im Falle $\mu_{i+1} < \mu(\hat{u}) < \mu_i$ die Abschätzung*

$$\frac{\mu_i - \mu(\hat{u})}{\mu(\hat{u}) - \mu_{i+1}} \le \sigma^2 \frac{\mu_i - \mu(u)}{\mu(u) - \mu_{i+1}} \tag{2.42}$$

mit dem Konvergenzfaktor

$$\sigma = 1 - (1-\gamma)\frac{\mu_i - \mu_{i+1}}{\mu_i} = \gamma + (1-\gamma)\frac{\mu_{i+1}}{\mu_i}. \tag{2.43}$$

Beweis. Es seien ohne Beschränkung der Allgemeinheit der Vektor u sowie die Eigenvektoren u_i und u_{i+1} normiert, also $\|u\|_2 = 1$ und $\|u_i\|_2 = \|u_{i+1}\|_2 = 1$. Weiterhin sei $u \in \text{span}\{u_i, u_{i+1}\}$ und kann daher als $u = t_i u_i + t_{i+1} u_{i+1}$ dargestellt werden. Die Koeffizienten t_i und t_{i+1} erhält man dabei aus der Gleichung der Normierungsbedingung

$$t_i^2 + t_{i+1}^2 = 1$$

und der Gleichung für den Rayleigh-Quotienten

$$\mu(u) = (u, Mu)_2 = \mu_i t_i^2 + \mu_{i+1} t_{i+1}^2.$$

Sie sind somit durch

$$t_i^2 = (u, u_i)_2^2 = \frac{\mu(u) - \mu_{i+1}}{\mu_i - \mu_{i+1}} \quad \text{und} \quad t_{i+1}^2 = (u, u_{i+1})_2^2 = \frac{\mu_i - \mu(u)}{\mu_i - \mu_{i+1}} \tag{2.44}$$

gegeben. Aus dem in Lemma 2.2.19 gezeigten Zusammenhang $(M + \alpha I)w = \beta M u$ folgt für die Koeffizienten von $w = \bar{t}_i u_i + \bar{t}_{i+1} u_{i+1}$, dass

$$\bar{t}_i^2 = (w, u_i)_2^2 = \frac{1}{\beta^2} \left(\frac{\mu_i}{\mu_i + \alpha} \right)^2 \frac{\mu(u) - \mu_{i+1}}{\mu_i - \mu_{i+1}} \quad \text{und} \quad \bar{t}_{i+1}^2 = (w, u_{i+1})_2^2 = \frac{1}{\beta^2} \left(\frac{\mu_{i+1}}{\mu_{i+1} + \alpha} \right)^2 \frac{\mu_i - \mu(u)}{\mu_i - \mu_{i+1}}.$$

Damit sind $\bar{t}_i$ und $\bar{t}_{i+1}$, abgesehen von den durch das Problem vorgegebenen Größen, nur vom Parameter α abhängig. Ebenfalls in Lemma 2.2.19 wurden die Beziehungen $\alpha > \mu_i$ sowie $\alpha \neq \mu_j$, $j > i + 1$ hergeleitet. Allerdings sind weitere Aussagen bezüglich α notwendig. Dazu betrachtet man den Winkel zwischen w und dem kegelgenerierenden Vektor Mu aus Definition 2.2.15, für den

$$0 < (w, \beta^{-1}(M + \alpha I)w)_2^2 = (w, Mu)_2^2$$

gilt. Dabei kann die rechte Seite mittels

$$(w, Mu)_2^2 = \|w\|_2^2 \|Mu\|_2^2 \cos^2 \phi_\gamma(u)$$

substituiert werden, wobei weiterführend nach Definition 2.2.14 und der trigonometrischen Beziehung $1 = \sin^2(\phi_\gamma(u)) + \cos^2(\phi_\gamma(u))$

$$\|Mu\|_2^2 \cos^2 \phi_\gamma(u) = \|Mu\|_2^2 - \gamma^2 \|Mu - \rho u\|_2^2$$

gilt. Ersetzt man w zudem durch $w = \beta(M + \alpha I)^{-1} Mu$ erhält man die Gleichung

$$(\beta(M + \alpha I)^{-1} Mu, Mu)_2^2 = \|\beta(M + \alpha I)^{-1} Mu\|_2^2 \left(\|Mu\|_2^2 - \gamma^2 \|Mu - \rho u\|_2^2 \right).$$

Unter Zuhilfenahme der Darstellung von u aus (2.44) resultiert diese Gleichung in einem Polynom $p(\alpha)$ zweiten Grades der Form $p(\alpha) = a\alpha^2 + b\alpha + c$. Die dabei auftretenden Koeffizienten lauten

$$\begin{aligned} 0 < a &= \gamma^2(\rho(\mu_i + \mu_{i+1}) - \mu_i \mu_{i+1}) \\ 0 < b &= 2\gamma^2 \rho \mu_i \mu_{i+1} \\ 0 > c &= -(1 - \gamma^2)\mu_i^2 \mu_{i+1}^2. \end{aligned}$$

Eine kurze Überlegung zeigt, dass die gesuchten Nullstellen des Polynoms $p(\alpha)$, bezeichnet mit α_1 und α_2, der Abschätzung $\alpha_1 < 0 < \alpha_2$ genügen. Mit den Aussagen über die Orientierung des Gradienten des Rayleigh-Quotienten aus Lemma 2.2.19 zeigt sich weiterhin, dass α_1 und α_2 jeweils mit dem Minimum

beziehungsweise Maximum des Rayleigh-Quotienten auf der Oberf äche des Kreiskegels korrespondieren. Ebenso wurde gezeigt, dass für $\beta > 0$ und $\alpha > \mu(w)$ das Minimum erhalten wird. Es kann daher geschlussfolgert werden, dass das zum Minimum gehörende α positiv ist. Da zudem a und b bezüglich $\rho \in (\mu_{i+1}, \mu_i)$ monoton steigende Funktionen sind, heißt dies, dass $p(\alpha)$ streng monoton fallend in ρ ist, sodass für $\rho \to \mu_i$ das Polynom $p(\alpha)$ den kleinsten Wert in $\alpha = \mu_{i+1}(1-\gamma)/\gamma$ annimmt.
Mit der Darstellung $(M + \alpha I)w = \beta M u$, wobei nun $\alpha > 0$, kann der Konvergenzfaktor formuliert werden. Dazu betrachtet man den Tangens des Winkels zwischen einer Iterierten und u_i in der von u_i und u_{i+1} aufgespannten Ebene und stellt diese ins Verhältnis. Für die hier betrachteten Vektoren u und w ergibt sich

$$\frac{(w, u_{i+1})_2^2}{(w, u_i)_2^2} = \sigma^2(\alpha)\frac{(u, u_{i+1})_2^2}{(u, u_i)_2^2}$$

und mit den Darstellungen der Koeff zienten $t_i, t_{i+1}, \bar{t}_i$ sowie $\bar{t}_{i+1}$ erhält man

$$\sigma^2(\alpha) = \frac{(w, u_{i+1})_2^2}{(w, u_i)_2^2}\frac{(u, u_i)_2^2}{(u, u_{i+1})_2^2} = \frac{\mu_i - \mu(w)}{\mu(w) - \mu_{i+1}}\frac{\mu(u) - \mu_{i+1}}{\mu_i - \mu(u)} = \left(\frac{\mu_{i+1}}{\mu_i}\frac{\mu_i + \alpha}{\mu_{i+1} + \alpha}\right)^2.$$

Hierbei ist $\sigma(\alpha) < 1$ eine streng monoton fallende Funktion bezüglich $\alpha > 0$ und wird somit maximal, falls α minimal ist. Dies ist nach den vorherigen Betrachtungen für $\alpha = \mu_{i+1}(1-\gamma)/\gamma$ der Fall und resultiert in einer oberen Schranke für $\sigma(\alpha)$ gegeben durch

$$\sigma(\alpha) \le \sigma = \gamma + (1-\gamma)\frac{\mu_{i+1}}{\mu_i} \tag{2.45}$$

und damit insgesamt der Abschätzung

$$\frac{\mu_i - \mu(w)}{\mu(w) - \mu_{i+1}} \le \sigma^2\frac{\mu_i - \mu(u)}{\mu(u) - \mu_{i+1}}. \tag{2.46}$$

Nach Wahl von w als Minimierer entsprechend 2.2.17 und der dort angegebenen Relation der Rayleigh-Quotienten gilt

$$\frac{\mu_i - \mu(\hat{u})}{\mu(\hat{u}) - \mu_{i+1}} \le \frac{\mu_i - \mu(w)}{\mu(w) - \mu_{i+1}}$$

und man erhält die gewünschte Abschätzung (2.42). □

Dies schließt die Analyse für das gradientenbasierte Anstiegsverfahren ab. Dabei zeigt der hergeleitete Konvergenzfaktor σ lediglich eine Abhängigkeit von den Eigenwerten μ_i und μ_{i+1}, welche den Rayleigh-Quotienten der Iterierten $\mu(u)$ einschließen, auf.
Gegenstand der Betrachtungen stellt aber das gradientenbasierte Abstiegsverfahren PINVIT dar. Jedoch lässt sich nun problemlos die hergeleitete Abschätzung des bisher betrachteten Anstiegsverfahrens auf das eigentlich zu untersuchende Problem übertragen.

Satz 2.2.33. *Gegeben sei das Standardeigenwertproblem*

$$Au = \lambda u \tag{2.47}$$

mit symmetrisch positiv def niter Matrix A. Zur Lösung sei die Iteration PINVIT aus (2.16) *angewendet. Dabei genüge der Vorkonditionierer B^{-1} der spektralen Abschätzung* (2.8) *und für den Rayleigh-Quotienten der Iterierten u gelte $\lambda_i < \lambda(u) < \lambda_{i+1}$. Dann gilt für den Rayleigh-Quotienten der Folgeiterierten $\hat{u}$ im Fall $\lambda_i < \lambda(\hat{u}) < \lambda_{i+1}$ die Abschätzung*

$$\frac{\lambda(\hat{u}) - \lambda_i}{\lambda_{i+1} - \lambda(\hat{u})} \le \sigma^2\frac{\lambda(u) - \lambda_i}{\lambda_{i+1} - \lambda(u)} \tag{2.48}$$

mit dem Konvergenzfaktor

$$\sigma = 1 - (1-\gamma)\frac{\lambda_{i+1}-\lambda_i}{\lambda_{i+1}} = \gamma + (1-\gamma)\frac{\lambda_i}{\lambda_{i+1}}. \tag{2.49}$$

Beweis. Die Eigenwerte der Matrixpaare (A, M) und (M, A) verhalten sich nach Bemerkung 2.1.2 gerade reziprok zueinander. Gleiches gilt für den Rayleigh-Quotienten. Die Substitution $\mu_j = \frac{1}{\lambda_j}$ und $\mu(\hat{u}) = \frac{1}{\lambda(\hat{u})}$ liefert damit für die linke Seite der Ungleichung (2.42)

$$\frac{\frac{1}{\lambda_i} - \frac{1}{\lambda(\hat{u})}}{\frac{1}{\lambda(\hat{u})} - \frac{1}{\lambda_{i+1}}} = \frac{(\lambda(\hat{u}) - \lambda_i)\lambda_{i+1}}{(\lambda_{i+1} - \lambda(\hat{u}))\lambda_i}.$$

Substitution von $\mu(u) = \frac{1}{\lambda(u)}$ ergibt für die rechte Seite derselben Ungleichung

$$\frac{\frac{1}{\lambda_i} - \frac{1}{\lambda(u)}}{\frac{1}{\lambda(u)} - \frac{1}{\lambda_{i+1}}} = \frac{(\lambda(u) - \lambda_i)\lambda_{i+1}}{(\lambda_{i+1} - \lambda(u))\lambda_i}.$$

Für den Konvergenzfaktor σ erhält man

$$\sigma = 1 - (1-\gamma)(1 - \frac{\lambda_i}{\lambda_{i+1}}) = \gamma + (1-\gamma)\frac{\lambda_i}{\lambda_{i+1}}.$$

Zusammen führt dies auf die gewünschte Abschätzung (2.48). □

Damit ist die Konvergenz der vorkonditionierten inversen Iteration (PINVIT) als gradientenbasiertes Abstiegsverfahren nachgewiesen. Es ist leicht zu erkennen, dass diese nur von der Güte des Vorkonditionierers B^{-1} sowie dem Verhältnis der Eigenwerte λ_i und λ_{i+1}, welche den Rayleigh-Quotienten einschließen, abhängt. Gleichzeitig f ndet man die Konvergenzrate der inversen Iteration wieder.

Korollar 2.2.34. *Für den Vorkonditionierer $B = A$ ist die Kontraktionsbedingung* (2.8) *mit $\gamma = 0$ erfüllt. Sei nun $\lambda(u) \in (\lambda_1, \lambda_2)$. Dann gilt für den Konvergenzfaktor*

$$\sigma = \gamma + (1-\gamma)\frac{\lambda_1}{\lambda_2} = \frac{\lambda_1}{\lambda_2}$$

und man erhält damit die Konvergenzrate der inversen Vektoriteration, siehe Gleichung (2.4)*.*

Eingangs wurde die Analyse auf den Fall einfacher Eigenwerte eingeschränkt. Dass mehrfache Eigenwerte die Konvergenzrate nicht beeinf ußen wird nun gezeigt, vergleiche auch [56].

Lemma 2.2.35. *Sei M symmetrisch positiv def nit mit mehrfachen Eigenwerten. Dann behält die Konvergenzabschätzung aus Satz 2.2.32 und damit auch die aus Satz 2.2.33 ihre Gültigkeit.*

Beweis. Sei $\bar{M} = \mathrm{diag}(\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_n, \ldots, \mu_n) \in \mathbb{R}^{m\times m}$ die aus M durch Ähnlichkeitstransformation erhaltene Diagonalmatrix. Dabei sind μ_k die paarweise verschiedenen Eigenwerte mit Mehrfachheit l_k. Zudem def niere mit $M' = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^{n\times n}$ die Matrix mit den gleichen aber einfachen Eigenwerten wie $\bar{M}$. Weiterhin sei $P : \mathbb{R}^{m\times m} \to \mathbb{R}^{n\times n}$ die Abbildung def niert durch

$$(Pu)_k = u'_k := \left(\sum_{j=1}^{l_k} u_{k,j}^2\right)^{\frac{1}{2}},$$

wobei $u = u_{k,j} \in \mathbb{R}^m$, mit $\|u\|_2 = 1$, die j-te Eigenkomponente zum Eigenwert μ_k enthält. Dann gilt für den Rayleigh-Quotienten von u

$$(u, \bar{M}u)_2 = \sum_{k=1}^{n} \sum_{j=1}^{l_k} \mu_k u_{k,j}^2 = \sum_{k=1}^{n} \mu_k \sum_{j=1}^{l_k} \left(u_{k,j}^2\right) = \sum_{k=1}^{n} \mu_k (Pu)_k^2 = (u', M'u')_2.$$

Damit besitzt das reduzierte Problem mit M' und das Ausgangsproblem mit $\bar{M}$ für u beziehungsweise u' den gleichen Rayleigh-Quotient und die Konvergenzabschätzungen bleiben damit unverändert. □

Diese Betrachtungen schließen die Konvergenzanalyse der gradientenbasierten Iterationsverfahren. Es ist damit sichergestellt, dass mit ihrer Hilfe die Berechnung der Eigenwerte eines verallgemeinerten Eigenwertproblems mit symmetrisch positiv def niten Matrizen A und M möglich ist.
Wie bereits in Abschnitt 2.2.1 herausgestellt, weist das Abstiegsverfahren (2.5) im Allgemeinen nicht die bestmögliche Konvergenz auf. So kann bereits durch eine geeignete Skalierung des Vorkonditionierers ein besseres Konvergenzverhalten erreicht werden. Wie dies umgesetzt werden kann, soll im folgenden Abschnitt dargestellt werden. Anschließend wird die simultane Berechnung mehrerer der kleinsten Eigenwerte betrachtet.

2.3. Eigenwertapproximationen in Unterräumen

Bereits bei den Betrachtungen im Abschnitt 2.2.1 wurde eine optimale Skalierung des Vorkonditionierers B^{-1} zum Erhalt der bestmöglichen Konstanten γ angesprochen. Der Nutzen einer Skalierung besteht offensichtlich darin, zu berechneten vorkonditionierten Residuen die bestmögliche Folgeiterierte $\hat{u}$ zu erhalten. Unter Einbeziehung eines solchen Skalierungsparameters ω nimmt das Abstiegsverfahren (2.5) folglich die Form

$$u^{(j+1)} = u^{(j)} - \omega^{(j)} B^{-1}(Au^{(j)} - \lambda(u^{(j)})Mu^{(j)}) \tag{2.50}$$

an, wobei hier der obere Index wieder den Iterationsschritt kennzeichnet. Dabei ist die optimale Skalierung (oder Schrittweite) durch

$$\omega^{(j)} = \arg\min_{\omega \in \mathbb{R}} \lambda(u^{(j)} - \omega B^{-1}(Au^{(j)} - \lambda(u^{(j)})Mu^{(j)}))$$

gegeben. Augenscheinlich ist die Bestimmung dieser Skalierung $\omega^{(j)}$ nur unter Kenntnis der Konstanten c_0 und c_1, welche im Allgemeinen nicht verfügbar sind, aus der spektralen Abschätzung (2.22) möglich. Jedoch existiert eine alternative Möglichkeit, um dieses optimale $\omega^{(j)}$ zu ermitteln und zwar mittels des Rayleigh-Ritz-Verfahrens.

2.3.1. Rayleigh-Ritz-Approximationen

Motiviert durch die Suche nach einer optimalen Schrittweite für die Iteration (2.50) soll nun die Möglichkeit zur Bestimmung von Bestapproximationen an die Eigenwerte und zugehörige Eigenvektoren in speziellen Unterräumen untersucht werden.

Def nition 2.3.1. *Seien $A, M \in \mathbb{R}^{n\times n}$ und sei $\mathcal{V}$ ein Unterraum des $\mathbb{R}^n$, kurz $\mathcal{V} \leq \mathbb{R}^n$. Erfüllt das Paar (θ, v) mit $v \neq 0$ die Beziehung*

$$Av - \theta Mv \perp \mathcal{V}$$

heißt (θ, v) **Rayleigh-Ritz-Approximation** *oder* **Ritz-Paar**. *Dabei ist θ der* **Ritzwert** *und v der zugehörige* **Ritzvektor**.

Bemerkung 2.3.2. *Im Folgenden sei wiederum nur das Standardeigenwertproblem $Au = \lambda u$ betrachtet. Die angeführten Aussagen lassen sich problemlos auf das verallgemeinerte Eigenwertproblem (für eine symmetrisch positiv definite Matrix M) übertragen, indem formal das Eigenwertproblem $M^{-1}Au = \lambda u$ betrachtet wird beziehungsweise die später beschriebene Matrix V der Bedingung $V^TMV = I$ genügt.*

Die in Definition 2.3.1 beschriebenen Rayleigh-Ritz-Approximationen können auf folgende Weise ermittelt werden.

Lemma 2.3.3. *Sei $\mathcal{V} \leq \mathbb{R}^n$ mit $\dim(\mathcal{V}) = s$. Weiterhin sei $V \in \mathbb{R}^{n\times s}$ eine Matrix derart, dass die Spalten $v_1, \ldots, v_s$ eine orthonormale Basis des Unterraumes $\mathcal{V}$ bilden, also $\operatorname{span}\{v_1, \ldots, v_s\} = \mathcal{V}$ und $V^TV = I$. Dann ist $(\theta_i, v_i),\ (i = 1, \ldots, s)$, genau dann Ritzpaar von $A \in \mathbb{R}^{n\times n}$, wenn θ_i ein Eigenwert des projizierten Eigenwertproblems*

$$V^TAVy_i = \theta_i y_i \tag{2.51}$$

zum Eigenvektor y_i ist. Für den Ritzvektor v_i gilt

$$v_i = Vy_i.$$

Beweis. Nach Definition 2.3.1 gilt, bedingt durch die Orthogonalitätsbedingung,

$$V^T(Av - \theta v) = 0.$$

Mit der Substitution $v = Vy$ und der Voraussetzung $V^TV = I$ führt dies auf

$$V^TAVy = \theta V^TVy = \theta y.$$

Damit existieren s Ritzwerte θ_i und zugehörige Ritzvektoren v_i als Lösung des Eigenwertproblems (2.51). □

Es soll nun gezeigt werden, dass die Ritzwerte die bestmöglichen Approximationen an die Eigenwerte der Matrix A bezüglich eines Unterraumes $\mathcal{V}$ sind. Dabei hilft folgende Darstellung der Eigenwerte einer Matrix.

Satz 2.3.4 (Courant-Fischer-Prinzip)**.** *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit und sei $\mathcal{W}_i,\ 1 \leq i \leq n$ ein i-dimensionaler Unterraum des $\mathbb{R}^n$. Mit $v \in \mathbb{R}^n\backslash\{0\}$ und dem Rayleigh-Quotienten $\lambda(v) = \frac{(v,Av)_2}{(v,v)_2}$ gilt für den i-ten Eigenwert die Darstellung*

$$\lambda_i = \min_{\mathcal{W}_i \leq \mathbb{R}^n} \max_{\substack{v\in\mathcal{W}_i,\\ v\neq 0}} \lambda(v), \quad (i = 1, \ldots, n), \tag{2.52}$$

wobei das Minimum über die Menge aller i-dimensionaler Teilräume des $\mathbb{R}^n$ genommen werden muss.

Beweis. Man betrachtet vorerst zwei spezielle Unterräume E_i und $E_{i-1}^{\perp}$. E_i sei dabei der i-dimensionale Unterraum, der von den ersten i normierten Eigenvektoren $u_1, \ldots, u_i$ aufgespannt wird. Der Rayleigh-Quotient von $0 \neq v = \sum_{j=1}^{i} \alpha_j u_j \in E_i$ genügt dann der Abschätzung

$$\lambda(v) = \frac{(v, Av)_2}{(v, v)_2} = \frac{\sum_{j=1}^{i} \lambda_j \alpha_j^2}{\sum_{j=1}^{i} \alpha_j^2} \leq \lambda_i$$

wobei Gleichheit für $v = u_i$ angenommen wird. Daher ist

$$\lambda_i = \max_{\substack{v\in\mathcal{W}_i,\\ v\neq 0}} \lambda(v).$$

Zieht man nun alle i-dimensionalen Unterräume E_i, die das Erzeugnis einer beliebigen Wahl von i Eigenvektoren sind, in Betracht, so erhält man

$$\lambda_i \geq \min_{E_i} \max_{\substack{v \in \mathscr{W}_i, \\ v \neq 0}} \lambda(v). \tag{2.53}$$

Der zweite Unterraum $E_{i-1}^{\perp}$ sei derart konstruiert, dass er das orthogonale Komplement zu E_{i-1} bildet, also $E_{i-1} \oplus E_{i-1}^{\perp} = \mathbb{R}^n$ gilt. Jedes Element $0 \neq z \in E_{i-1}^{\perp}$ genügt daher der Darstellung $z = \sum_{j=i}^{n} \tilde{\alpha}_j u_j$. Betrachtet man den Rayleigh-Quotienten, erhält man nach adaptierter Argumentation aus dem ersten Fall

$$\lambda_i = \min_{\substack{z \in E_{i-1}^{\perp}, \\ z \neq 0}} \lambda(z). \tag{2.54}$$

Sei nun $\mathscr{W}_i \neq E_i$ ein beliebiger i-dimensionaler Unterraum. Dann gilt, dass stets ein $w \in (\mathscr{W}_i \cap E_{i-1}^{\perp})$ existiert, da sonst $\dim(\mathscr{W}_i \oplus E_{i-1}^{\perp}) = n + 1 > n$ wäre. Für ein solches w gilt unter Nutzung von (2.54)

$$\max_{\substack{v \in \mathscr{W}_i, \\ v \neq 0}} \lambda(v) \geq \lambda(w) \geq \min_{\substack{z \in E_{i-1}^{\perp}, \\ z \neq 0}} \lambda(z) = \lambda_i.$$

Dies zeigt, für jeden beliebigen i-dimensionalen Unterraum ist λ_i eine untere Schranke für die Rayleigh-Quotienten der enthaltenen Vektoren $v \in \mathscr{W}_i$. Gleichzeitig ergab (2.53) eine obere Schranke für den speziellen Fall $\mathscr{W}_i = E_i$. Da E_i in der Menge aller i-dimensionalen Unterräume enthalten ist, liefern beide Abschätzungen zusammengenommen

$$\lambda_i = \min_{\mathscr{W}_i \leq \mathbb{R}^n} \max_{\substack{v \in \mathscr{W}_i, \\ v \neq 0}} \lambda(v),$$

also die Behauptung. □

Mit Hilfe des Satzes 2.3.4 kann nun eine Aussage zur Optimalität der Ritzwerte getroffen werden.

Lemma 2.3.5. *Die Eigenwerte θ_i des projizierte Eigenwertproblems* (2.51) *(Ritzwerte) sind die bestmöglichen Approximationen an die Eigenwerte von A im s-dimensionalen Unterraum $\mathscr{V} = \operatorname{span}\{v_1, \ldots, v_s\}$.*

Beweis. Die bestmöglichen Approximationen an die Eigenwerte von A sind im Unterraum $\mathscr{V}$ nach Satz 2.3.4 durch

$$\beta_i = \min_{\mathscr{U}_i \leq \mathscr{V}} \max_{\substack{v \in \mathscr{U}_i, \\ v \neq 0}} \lambda(v) \tag{2.55}$$

gegeben, wobei $\mathscr{U}_i$ ein i-dimensionaler Teilraum von $\mathscr{V}$ ist. Weiterhin kann jedes Element aus $v \in \mathscr{V}$ mittels eines Koeffizientenvektors $y \in \mathbb{R}^s$ und der Matrix $V \in \mathbb{R}^{n \times s}$, deren Spalten $\{v_1, \ldots, v_s\}$ eine Orthonormalbasis von $\mathscr{V}$ bilden, dargestellt werden, namentlich $v = Vy$. Damit erhält man aus (2.55) unter entsprechender Anpassung der Räume

$$\beta_i = \min_{W_i \leq \mathbb{R}^s} \max_{\substack{y \in W_i, \\ y \neq 0}} \lambda(Vy).$$

Für den Rayleigh-Quotienten auf der rechten Seite ergibt sich

$$\lambda(Vy) = \frac{(Vy, AVy)_2}{(Vy, Vy)_2} = \frac{(y, V^T AVy)_2}{(y, y)_2} =: \bar{\lambda}(y)$$

und damit nach 2.3.4

$$\beta_i = \min_{W_i \leq \mathbb{R}^s} \max_{\substack{y \in W_i, \\ y \neq 0}} \bar{\lambda}(y) = \theta_i,$$

da die θ_i die Eigenwerte der Matrix $V^T AV \in \mathbb{R}^{s \times s}$ sind. □

Die Berechnung der Ritzpaare entsprechend Lemma 2.3.3 def nieren somit das *Rayleigh-Ritz-Verfahren*. Die algorithmische Umsetzung soll im Kapitel 4 genauer erläutert werden. Augenscheinlich ersetzt man hierdurch ein Eigenwertproblem für die Matrix $A \in \mathbb{R}^{n\times n}$ nur durch eines für die Matrix $V^T AV \in \mathbb{R}^{s\times s}$. Dabei ist aber zu beachten, dass in der Regel $s \ll n$ gilt. Deutlich wird dies, wenn man sich nochmals die Motivation zur Einführung der Rayleigh-Ritz-Approximationen am Anfang des Abschnitts anschaut. Dabei wurde die Frage nach der Berechnung einer optimalen Schrittweite $\omega^{(j)}$ für die Iteration (2.50) aufgeworfen. Diese Frage wird implizit beantwortet, indem mittels des Rayleigh-Ritz-Verfahrens die bestmögliche Folgeiterierte $u^{(j+1)}$ im Unterraum

$$\mathcal{V}_2^{(j)} = \text{span}\{u^{(j)}, B^{-1}(Au^{(j)} - \lambda(u^{(j)})u^{(j)})\}$$

ermittelt wird. Da mittels dieser Iteration der kleinste Eigenwert von A berechnet werden soll, ist die beste Approximation folglich der Ritzvektor zum kleinsten Ritzwert des projizierten Eigenwertproblems (2.51). Die Dimension ist damit $s = 2$, was im Wesentlichen nur das Lösen einer quadratischen Gleichung erfordert.
Basierend auf der Berechnung von Bestapproximationen in speziellen Unterräumen mittels des Rayleigh-Ritz-Verfahrens ergibt sich die Möglichkeit, eine Hierarchie von Eigenlösern, welche dem gradientenbasierten Abstiegsverfahren (2.16) entlehnt sind, zu konstruieren.

2.4. Eine Klasse von Eigenlösern - Das (k)-Schema

Mit Blick auf die numerischen Untersuchungen im Kapitel 5 soll eine Klasse von Eigenlösern basierend auf PINVIT vorgestellt werden. Bereits im vorangegangenen Abschnitt wurde mit der Verwendung einer optimalen Schrittweite $\omega^{(j)}$ ein weiteres Verfahren hergeleitet. Essentiell liegt der Unterschied darin, dass im Falle von PINVIT eine einfache Korrektur der Iterierten $u^{(j)}$ durch das vorkonditionierte Residuum

$$d^{(j)} = B^{-1}(Au^{(j)} - \lambda(u^{(j)})Mu^{(j)}) \tag{2.56}$$

vorgenommen wurde, wohingegen die Bestimmung der optimalen Schrittweite $\omega^{(j)}$ die Bestapproximation an den gesuchten Eigenvektor im Unterraum

$$\mathcal{V}_2^{(j)} = \text{span}\{u^{(j)}, d^{(j)}\}$$

ermittelt. Das unter Nutzung der Bestapproximation beschriebene Verfahren heißt *preconditioned steepest descent*, kurz PSD.
Hat man nun bereits $j \geq 2$ Iterationen durchgeführt, so kann die Dimension des Unterraums $\mathcal{V}$ mit vorangegangenen Eigenvektorapproximationen erweitert werden. Nimmt man die Iterierte $u^{(j-1)}$ hinzu, so liefert dies das Verfahren *locally optimal preconditioned conjugate gradient*, kurz LOPCG, bei dem das Rayleigh-Ritz-Verfahren demnach auf den Unterraum

$$\mathcal{V}_3^{(j)} = \text{span}\{u^{(j-1)}, u^{(j)}, d^{(j)}\}$$

angewendet wird, [37, 39]. Auf diese Weise kann eine Hierarchie von Eigenlösern konstruiert werden, die als (k)-*Schema* bezeichnet werden, [59]. Dabei besitzt der jeweilige Unterraum, welcher im Rayleigh-Ritz-Verfahren Anwendung f ndet, die Dimension k und ist durch

$$\mathcal{V}_k^{(j)} = \text{span}\{u^{(j-k+2)}, \ldots, u^{(j-1)}, u^{(j)}, d^{(j)}\}$$

k	Eigenlöser	Unterraum
$k = 1$	preconditioned inverse iteration (PINVIT)	$u^{(j)} - d^{(j)}$
$k = 2$	precondition steepest descent (PSD)	$\mathcal{V}_2^{(j)}$
$k = 3$	locally optimal preconditioned conjugate gradient (LOPCG)	$\mathcal{V}_3^{(j)}$
$k \geq 4$	higher order schemes	$\mathcal{V}_k^{(j)}$

Tabelle 2.1.: Klasse von Eigenlösern (k-Schema)

gegeben. Eine Klassif zierung der (k)-Schemata ist in Tabelle 2.1 dargestellt.
Eine letzte Bemerkung betrifft die Konvergenz der hier neben PINVIT vorgestellten Iterationen. Für PSD ist neben PINVIT eine scharfe Konvergenzrate zugänglich, [55], wohingegen für LOPCG und die (k)-Schemata höherer Ordnung zurzeit keine Konvergenzanalyse vorliegt. Numerische Untersuchungen zeigen aber, und so auch für die hier im Kapitel 5 berechneten Probleme, dass der Namenszusatz *optimal* für LOPCG gerechtfertigt ist. Auch wenn die Berechnung der Folgeiterierten die Lösung eines (projizierten) Eigenwertproblems erfordert, wird dieser erhöhte Rechenaufwand durch die im Vergleich besseren Ritzwerte mehr als aufgewogen. Für die Schemata mit $k \geq 4$ ist dies dann aber nicht mehr der Fall, was auch bei den späteren numerischen Untersuchen festgestellt wird.

2.4.1. Unterraumiterationen

Bereits in der Einleitung wurde darauf hingewiesen, dass bei diskretisierten Eigenwertproblemen mit technischem Hintergrund häuf g nicht nur der kleinste Eigenwert, sondern einige der kleinsten Eigenwerte gesucht sind, so zum Beispiel die relevanten Eigenschwingungen eines mechanischen Systems. Um eine solche Berechnung zu realisieren, nutzt man Blockvarianten der oben vorgestellten Klasse von Eigenlösern, [58]. Diese sollen nun kurz vorgestellt werden, wobei für eine algorithmische Umsetzung auf Kapitel 4 verwiesen sei.
Seien $u^{(j,1)}, \ldots, u^{(j,s)}$ die s Eigenvektorapproximationen im j-ten Iterationsschritt, welche spaltenweise in der Matrix $V^{(j)} \in \mathbb{R}^{n \times s}$ zusammengefasst werden. Dann modif ziert sich die Iterationsvorschrift aus (2.5) für PINVIT (beziehungsweise dem (1)-Schema) zur blockweisen Variante in der Form

$$V^{(j+1)} = V^{(j)} - B^{-1}(AV^{(j)} - MV^{(j)}\Theta^{(j)}) \tag{2.57}$$

mit der Diagonalmatrix $\Theta^{(j)} \in \mathbb{R}^{s \times s}$ gegeben durch

$$\Theta^{(j)} = \begin{pmatrix} \theta_1^{(j)} & 0 & \ldots & 0 \\ 0 & \theta_2^{(j)} & \ldots & 0 \\ 0 & \ldots & \ddots & 0 \\ 0 & \ldots & 0 & \theta_s^{(j)} \end{pmatrix}.$$

Hierbei sind $\theta_i^{(j)}$, $(i = 1, \ldots, s)$, die Ritzwerte zu den Eigenvektorapproximationen $u^{(j,i)}$ im j-ten Iterationschritt. In kompakter Form kann die Iteration (2.57) auch als

$$V^{(j+1)} = V^{(j)} - D^{(j)} \tag{2.58}$$

mit der Matrix $D^{(j)} = (d^{(j,1)}, \ldots, d^{(j,s)}) \in \mathbb{R}^{n \times s}$, die spaltenweise die vorkonditionierten Residuen enthält, geschrieben werden.

Die Blockvarianten der (k)-Schemata mit $k \geq 2$ modif zieren sich im Wesentlichen nur durch den Unterraum, in dem die Bestapproximationen mittels des Rayleigh-Ritz-Verfahrens ermittelt werden. Im Fall $k = 2$ gestaltet sich dieser Unterraum als

$$\mathcal{V}_{2s}^{(j)} = \text{span}\{V^{(j)}, D^{(j)}\}, \tag{2.59}$$

mit $V^{(j)}$ und $D^{(j)}$ wie oben, und besitzt damit die Dimension $\dim(\mathcal{V}_{2s}^{(j)}) = 2s$. Für $k > 2$ hat der Raum die Gestalt

$$\mathcal{V}_{ks}^{(j)} = \text{span}\{V^{(j-k+2)}, \ldots, V^{(j-1)}, V^{(j)}, D^{(j)}\}. \tag{2.60}$$

Dies schließt die Betrachtungen zur vorkonditionierten Iteration PINVIT und ihrer Derivate des (k)-Schemas. Das folgenden Kapitel fokussiert nun die Berechnung der vorkonditionierten Residuen aus Gleichung (2.56) mittels Mehrgitterverfahren, also die Realisierung der Operation $y \mapsto B^{-1}y$.

3. Mehrgitterverfahren

Im vorangegangenen Kapitel lag der Untersuchungsschwerpunkt im Konvergenzverhalten der vorkonditionierten Iterationen zur Lösung des verallgemeinerten Eigenwertproblems für symmetrisch positiv definite Matrizen A und M. Die Namensgebung der Iteration ist dabei vom Einsatz eines Vorkonditionierers B^{-1} zur Berechnung der vorkonditionierten Residuen

$$d := B^{-1}(Au - \lambda(u)Mu), \tag{3.1}$$

die zur Korrektur der Näherungslösung u genutzt werden, abgeleitet. Die Charakterisierung des Vorkonditionierers erfolgte dabei unter analytischen Gesichtspunkten im Wesentlichen durch die Konstante γ aus der Kontraktionsbedingung

$$\|I - B^{-1}A\|_A \leq \gamma < 1, \tag{3.2}$$

welche dann explizit in die Konvergenzabschätzung (2.48) eingeht. Kern dieses Kapitels bilden Betrachtungen zur Berechnung dieser vorkonditionierten Residuen mittels algebraischer Mehrgitterverfahren, also der Realisierung der Operation $y \mapsto B^{-1}y$. Die nachfolgenden Ausführungen zu den Mehrgitterverfahren sollen (auch aus historischen Gründen) zunächst anhand der linearen Gleichungssysteme vorgenommen werden. Wie diese dann zur Lösung des Eigenwertproblems genutzt werden können, bildet den Schwerpunkt im sich anschließenden Kapitel 4.
Bevor die Lösung linearer Gleichungssysteme in den Mittelpunkt rückt, soll zunächst der Zusammenhang zwischen Randwertproblem und linearem Gleichungssystem dargestellt werden. Dieses Beispiel wird weiterhin bei der Motivation der Mehrgitterverfahren eine wichtige Rolle spielen.

Beispiel 1. Man betrachte die *Poisson-Gleichung*, also eine elliptische partielle Differentialgleichung, auf dem Gebiet $\Omega \subset \mathbb{R}^d$ gegeben durch

$$\begin{aligned} -\Delta u(x) &= f(x) \qquad x \in \Omega, \\ u(x) &= 0 \qquad x \in \partial\Omega. \end{aligned} \tag{3.3}$$

Zur numerischen Behandlung wird das Gebiet Ω diskretisiert. Dabei sei eine äquidistante Unterteilung mit Diskretisierungsparameter (oder auch der Schrittweite) $h := 1/(n+1)$ bezüglich jeder Raumdimension zu Grunde gelegt. Man erhält ein Gitter W_h, bestehend aus einer Menge von Knoten oder auch Variablen, gekennzeichnet mit $\mathcal{N}_h$. Weiterhin liefert die Diskretisierung des Differentialoperators $\Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$ mittels finiter Differenzen oder finiter Elemente (vergleiche Abschnitt 1.1) ein lineares Gleichungssystem $Ax = b$ mit der symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n^d \times n^d}$ und der rechten Seite $b \in \mathbb{R}^{n^d}$. Es sei bemerkt, dass ein möglichst geringer Diskretisierungsfehler und damit eine gute Näherung an die (analytische) Lösung für große n und daher kleine h erhalten wird.

Dieses elementare Beispiel verdeutlicht, ähnlich des in Abschnitt 1.1 erläuterten Zusammenhangs zwischen dem Operator- und Matrixeigenwertproblem, wie das Randwertproblem und das lineare Gleichungssystem korrespondieren. Im Folgenden soll ein kurzer Überblick zum Prinzip iterativer Verfahren

zur approximativen Lösung eines solchen Gleichungssystems gegeben werden.
Gegeben sei das lineare Gleichungssystem

$$Ax = b \tag{3.4}$$

mit der Matrix $A \in \mathbb{R}^{n\times n}$ sowie $x, b \in \mathbb{R}^n$. Iterative Verfahren beruhen darauf, das Residuum $r^{(k)}$ zu einer Näherungslösung $x^{(k)}$ gegeben durch

$$r^{(k)} = b - Ax^{(k)} \tag{3.5}$$

sukzessive in jedem Iterationsschritt $k = 0, 1, \ldots$ zu reduzieren. Die exakte Korrektur $e^{(k)}$, errechnet mittels

$$Ae^{(k)} = r^{(k)}, \tag{3.6}$$

liefert dabei die Lösung in einem Schritt. Diese Berechnung erfordert jedoch wiederum das Lösen eines linearen Gleichungssystems. Die Idee ist es nun, die exakte Lösung $e^{(k)}$ approximativ durch

$$B\tilde{e}^{(k)} = r^{(k)}$$

anzunähern, wobei $B \approx A$ eine leicht zu invertierende Näherung an A und daher B^{-1} den bereits aus Abschnitt 2.1 bekannten Vorkonditionierer darstellt. Aus der Korrekturgleichung erhält man somit

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \tilde{e}^{(k)} \\ &= x^{(k)} + B^{-1}r^{(k)} \\ &= x^{(k)} + B^{-1}(b - Ax^{(k)}) \end{aligned} \tag{3.7}$$

und def niert mit (3.7) die Iterationsvorschrift zum approximativen Lösen linearer Gleichungssysteme. Auch hier kann eine Fehlerfortpf anzungsgleichung formuliert werden. Für den Fehler $e^{(k+1)} := x^{(k+1)} - x^*$, wobei x^* die exakte Lösung beschreibt, gilt

$$\begin{aligned} e^{(k+1)} = x^{(k+1)} - x^* &= x^{(k)} - x^* + B^{-1}(b - Ax^{(k)}) \\ &= (I - B^{-1}A)(x^{(k)} - x^*) = (I - B^{-1}A)e^{(k)}. \end{aligned} \tag{3.8}$$

Das heißt, die Iteration (3.7) konvergiert, wenn für den Spektralradius der Matrix $\mathcal{S} := I - B^{-1}A$ die Abschätzung (3.2) gilt.
Der Zusammenhang zur vorkonditionierten Iteration aus Kapitel 2 ist damit klar zu erkennen. Ausgehend vom Eigenwertproblem wird das Residuum $b - Ax^{(k)}$ in Gleichung (3.7) durch $\lambda(u^{(k)})u^{(k)} - Au^{(k)}$, siehe Gleichung (2.5), ersetzt. Damit kann zur Berechnung der vorkonditionierten Residuen innerhalb der vorkonditionierten Iteration die Methode der iterativen Gleichungslöser verwendet werden. Es sei hier nochmals der Hinweis angeführt, dass die Wahl $B = A$ zwar zur Einschrittkonvergenz für das Lösen des Gleichungssystem $Ax = b$ führt, im Fall der vorkonditionierten Iteration aber nur die inverse Vektoriteration herbeiführt, vergleiche Bemerkung 2.1.1. Dennoch geht die Konstante γ als Spektralradius in die Konvergenzabschätzung der vorkonditionierten Iteration ein, was die Konstruktion möglichst „guter“ Vorkonditionierer B^{-1} erfordert.

Ausgehend von Betrachtungen klassischer Iterationsverfahren zum Lösen des Gleichungssystems (3.4) sollen im Folgenden die Mehrgitterverfahren motiviert und vorgestellt werden. Dazu wird in einem ersten Schritt das Prinzip anhand des Zweigitterverfahrens erörtert. Den Schwerpunkt bilden dann die algebraischen Mehrgitterverfahren und ihre Umsetzung. Den Abschluss bilden klassische Konvergenzaussagen für (geometrische) Mehrgitterverfahren.

3.1. Glättungseigenschaften klassischer Iterationsverfahren

Zur Motivation der Mehrgitterverfahren sollen an dieser Stelle einige Aspekte der klassischen Iterationsverfahren und ihrer Konvergenzeigenschaften angeführt werden, [33, 52, 73]. Mit Blick auf die weiteren Betrachtungen soll dies exemplarisch anhand des Jacobi- und des Gauß-Seidel-Verfahrens geschehen.

Def nition 3.1.1. *Sei $A \in \mathbb{R}^{n\times n}$. Dann ist, für $i = 1, \ldots, n$,*

- $D = \mathrm{diag}(A) = a_{ii}$ *der Diagonalanteil von A,*
- $L = \mathrm{tril}(A) = a_{ij},\ i > j$ *der strikte untere Dreiecksanteil von A und*
- $R = \mathrm{triu}(A) = a_{ij},\ i < j$ *der strikte obere Dreiecksanteil von A.*

Es gilt also $A = L + D + R$.

Basierend auf dieser Zerlegung können die oben genannten Iterationsverfahren def niert werden.

Def nition 3.1.2 (Klassische Iterationsverfahren)**.** *Sei $A \in \mathbb{R}^{n\times n}$ und $A = D + L + R$ eine additive Zerlegung entsprechend Def nition 3.1.1. Dann ist durch die Wahl des Vorkonditionierers $B = D$ in Gleichung* (3.7) *das* **Jacobi-Verfahren** *und durch die Wahl $B = D + L$ das* **Gauß-Seidel-Verfahren** *def niert. Die resultierenden Iterationsvorschriften lauten demnach*

$$x^{(k+1)} = x^{(k)} + D^{-1}(b - Ax^{(k)})$$

beziehungsweise

$$x^{(k+1)} = x^{(k)} + (D + L)^{-1}(b - Ax^{(k)}).$$

Der Vorteil dieser Verfahren ist die sehr einfache Umsetzung. Jedoch weisen sie auch einen entscheidenden Nachteil auf. Dieser soll anhand des im Vorfeld vorgestellten Beispiels 1 illustriert werden, siehe dazu auch [32, 33].

Beispiel 2. Im elementaren Fall mit $d = 1$ führt die äquidistante Diskretisierung von Gleichung (3.3) mittels f niter Differenzen, gegeben durch die Approximation

$$-u''(x_i) \approx \frac{-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})}{h^2} \quad (i = 1, \ldots, n), \tag{3.9}$$

bei lexikograf scher Nummerierung auf ein Gleichungssystem mit $A = \mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ und der rechten Seite $b = h^2 \cdot f(x_i)$. Die Eigenwerte der Matrix A sind durch $\lambda_i = 2 + 2\cos(\frac{i\pi}{n+1}),\ (i = 1, \ldots, n)$, gegeben. Für das Jacobi-Verfahren lautet der Vorkonditionierer $B = \mathrm{diag}(A) = D$. Der Spektralradius von B^{-1} ist somit $\beta = \frac{1}{2}$. Es gilt daher für die Eigenwerte μ_i der ebenfalls symmetrischen Fehlerfortpf anzungsmatrix $\mathcal{S} = I - B^{-1}A$

$$\mu_i = 1 - \frac{1}{2}\lambda_i = \cos\left(\frac{i\pi}{n+1}\right) = \cos(i\pi h).$$

Der Spektralradius ist demnach

$$\varrho(\mathcal{S}) = \max_{i=1,\ldots,n} \left|\cos\left(\frac{i\pi}{n+1}\right)\right| = \cos\left(\pi\frac{1}{n+1}\right).$$

Daher ist mit $h = 1/(n+1)$

$$\lim_{n\to\infty} \cos\left(\pi\frac{1}{n+1}\right) = \lim_{h\to 0} \cos(\pi h) = 1.$$

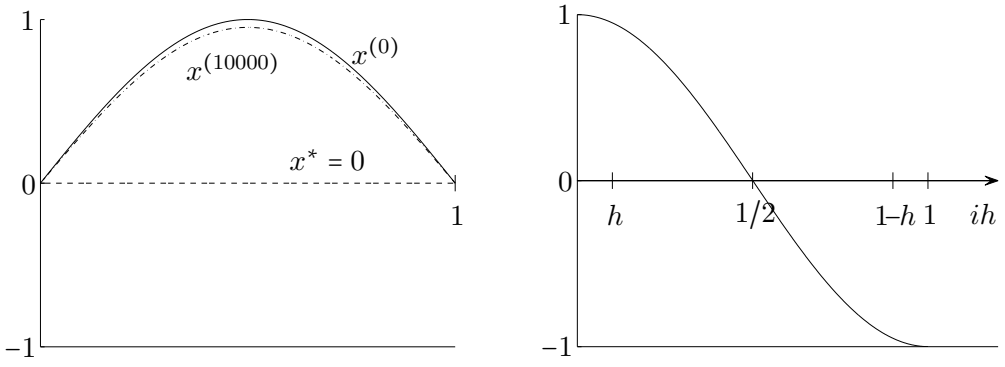

Abbildung 3.1.: Illustrationen des Beispiels 2. Links: Ergebnis des Jacobi-Verfahrens zur Berechnung des homogenen Gleichungssystems $Ax = 0$ mit Startiterierten $x^{(0)}$ als Eigenvektor zum größten Eigenwert. Rechts: Eigenwerte der Matrix $\mathcal{S}$ in Abhängigkeit von der Schrittweite h.

Somit verhält sich der Spektralradius der Matrix $\mathcal{S}$ im Falle des Jacobi-Verfahrens wie

$$1 - \mathcal{O}(h), \tag{3.10}$$

was bei feinerer Diskretisierung ($h \to 0$) zu einer langsamen Konvergenz führt.

Zur Verdeutlichung dieses Sachverhalts dient die Abbildung 3.1 links. Zu Grunde liegt das homogene Gleichungssystem $Ax = 0$ im Fall einer (verhältnismäßig geringen) Problemdimension $n = 1000$. Dargestellt sind die Startiterierte $x^{(0)}$ als (bezüglich der Maximumnorm normierter) Eigenvektor zum größtem Eigenwert $\mu_1 \approx 0.999995$ und die Iterierte $x^{(10000)}$. Dabei nähert sich $x^{(k)}$ nur langsam der gesuchten Lösung $x^* = 0$ an, im Beispiel verbleiben $99,5\%$ des Ausgangsfehlers $\|e^{(0)}\|_2 = \|x^{(0)}\|_2$. In der gleichen Abbildung sind auf der rechten Seite die Eigenwerte der Matrix $\mathcal{S}$ abgebildet. Man erkennt, dass das Spektrum für kleine h sehr nahe an 1 herankommt und damit das in Gleichung (3.10) geschilderte Verhalten zeigt. Im Falle des Gauß-Seidel-Verfahrens kann ein qualitativ ähnlicher Sachverhalt gezeigt werden.
Der Grund für das schlechte Konvergenzverhalten läßt sich dabei leicht begründen. Jeder Fehlervektor $e^{(k)}$ kann mittels der Basis aus Eigenvektoren $v_1, \dots v_n$ der Matrix $\mathcal{S} = I - B^{-1}A$ in der Form $e^{(k)} = \sum_{i=1}^{n} \alpha_i v_i$ entwickelt werden. Die Anwendung eines Iterationsschrittes bedeutet entsprechend Gleichung (3.8)

$$\begin{aligned} e^{(k+1)} = (I - B^{-1}A)e^{(k)} &= (I - B^{-1}A)\sum_{i=1}^{n} \alpha_i v_i \\ &= \sum_{i=1}^{n} \mu_i \alpha_i v_i, \end{aligned}$$

Ausgehend vom Fehler $e^{(0)}$ heißt dies für den $(k+1)$-ten Fehler

$$e^{(k+1)} = \sum_{i=1}^{n} \mu_i^{k+1} \alpha_i v_i.$$

In dieser Darstellung erkennt man, dass die Fehleranteile, welche mit kleinen Eigenwerten korrespondieren durch die Iteration sehr gut reduziert werden, die zu größeren Eigenwerten jedoch nur schlecht

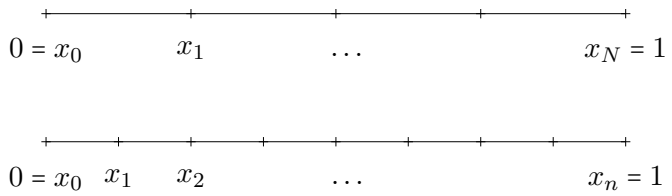

Abbildung 3.2.: Diskretisierungen des Gebietes $\Omega = [0,1]$ mit Schrittweite h (unten) beziehungsweise $H = 2h$ (oben).

behandelt werden können. Auf Grund dieser Feststellung ist es daher möglich, den Fehler $e^{(k+1)}$ in den hochfrequenten, beziehungsweise stark oszillierenden, Anteil und den niedrigfrequenten, beziehungsweise glatten, Anteil zu unterteilen. Diese Eigenschaft, dass die klassischen Iterationsverfahren den oszillierenden Anteil gut behandeln, führt dazu, dass diese auch als *Glätter* beziehungsweise *Glättungsoperation* bezeichnet werden. Um nun den durch den Glätter nur schlecht reduzierten, glatten Fehleranteil zu behandeln, bedient man sich der Methode der *Mehrgitterverfahren* (kurz *MG*). Diese bestehen aus einer Kombination zweier Iterationsverfahren, welche die in Gleichung (3.7) angegebene Struktur besitzen. Die erste Iteration behandelt dabei den hochfrequenten Fehleranteil und kann mittels der oben vorgestellten klassischen Iterationsverfahren realisiert werden. Diese werden daher nochmals Gegenstand der Betrachtungen in Abschnitt 3.3.1. Die zweite Iteration bildet die *Grobgitterkorrektur*. Dabei wird durch eine geeignete Projektion der glatte Fehleranteil in seinem Wesen verändert und in eine weniger glatte Grobgitterdarstellung überführt. Eine geeignete Behandlung des Fehlers in dieser Darstellung kann dann zur Korrektur der bereits geglätteten Approximation aus dem ersten Schritt genutzt werden. Dieses Vorgehen liefert eine drastische Konvergenzbeschleunigung (verglichen mit der bloßen Anwendung eines klassischen Iterationsverfahrens). Diese eher unspezif sche Beschreibung der Methode soll im Folgenden genauer erläutert und damit die Struktur eines Mehrgitterverfahrens hergeleitet werden.

3.2. Struktur der Mehrgitterverfahren

Der Begriff des Mehrgitterverfahrens resultiert aus dem Prinzip, nach dem das Verfahren konstruiert ist. Das einfachste Mehrgitterverfahren stellt dabei das Zweigitterverfahren dar. Daher soll dieses vorerst im Mittelpunkt stehen, um das prinzipielle Vorgehen zu erläutern. Anschließend wird gezeigt, wie aus diesem das Mehrgitterverfahren abgeleitet wird, [6, 33, 71].

Zweigitterverfahren

Bereits im Beispiel 1 wurde deutlich, dass mittels der Diskretisierung des betrachteten Differentialoperators die numerische Lösung der zu Grunde liegenden partiellen Differentialgleichung im Lösen eines linearen Gleichungssystems mündet. Eine wesentliche Größe bei der Konstruktion des linearen Gleichungssystems stellt der Diskretisierungsparameter $h = 1/(n+1)$, aus dem die Anzahl der Stützstellen n resultiert, dar. Betrachtet man nun für das gegebene Beispiel 1 zwei Diskretisierungen mit den Parametern h beziehungsweise $H := 2h$ (also einer Diskretisierung mit doppelter Schrittweite), so führt dies auf zwei Gitter, skizziert in Abbildung 3.2. Dabei seien im Folgenden die auftretenden Größen durch die Indizes h und H in ihrer Gitterzugehörigkeit gekennzeichnet. Durch die Wahl von Ansatzfunktionen $\phi_h^{(i)}, (i = 1, \ldots, n)$, beziehungsweise $\phi_H^{(i)}, (i = 1, \ldots, N)$, (vergleiche Abschnitt 1.1) mit

$W_h = \text{span}\{\phi_h^{(i)}\}$ und $W_H = \text{span}\{\phi_H^{(i)}\}$ erhält man zwei Funktionenräume, die der Beziehung

$$W_H \subset W_h \tag{3.11}$$

genügen. Beide Gitter bestehen aus den Knoten $\mathcal{N}_h$ beziehungsweise $\mathcal{N}_H$. Nach Konstruktion gilt für diese Mengen die Inklusion

$$\mathcal{N}_H \subset \mathcal{N}_h$$

oder äquivalent, die Menge der Grobgitterpunkte $\mathcal{N}_H$ ist in der Menge der Feingitterpunkte $\mathcal{N}_h$ enthalten. Nach Gleichung (3.9) können durch die Diskretisierung des Differentialoperators zwei Gleichungssysteme, gegeben durch

$$A_h x_h = b_h, \quad A_h \in \mathbb{R}^{n\times n}, \tag{3.12}$$

beziehungsweise

$$A_H x_H = b_H, \quad A_H \in \mathbb{R}^{N\times N}, \tag{3.13}$$

in Abhängigkeit vom Diskretisierungsparameter formuliert werden. Das Zweigitterverfahren involviert nun beide Diskretisierungen zur effektiven Lösung des linearen Gleichungssystems (3.12). Dabei geht man wie folgt vor.
Ausgehend von einer Approximation x_h (wobei der im Vorfeld genutzte Iterationsindex (k) vernachlässigt sei) werden ν_1 Vorglättungsschritte mit dem Glättungsoperator $\mathcal{S}$, beispielsweise einer klassische Iteration aus Abschnitt 3.1, vorgenommen. Dies resultiert in der geglätteten Approximation $\tilde{x}_h$. Ausgehend hiervon berechnet man das Residuum

$$r_h = b_h - A_h\tilde{x}_h.$$

Dieses wird mittels des *Restriktionsoperators* I_h^H auf das gröbere Gitter W_H abgebildet (die Transferoperatoren werden im Anschluss genauer betrachtet). Nun wird die Korrekturgleichung

$$A_H e_H = r_H$$

mit dem Grobgitteroperator A_H aus (3.13) und $r_H = I_h^H r_h$ gelöst. Der berechnete Fehler e_H wird nun wiederum auf das feine Gitter mittels des *Prolongationsoperators* I_H^h durch $e_h = I_H^h e_H$ transferiert. Im Anschluss f ndet die Korrektur von $\tilde{x}_h$ um e_h und üblicherweise eine Nachglättung durch ν_2 Schritte mittels $\mathcal{S}$ statt. Eine solche Zweigitterkorrektur ist in Algorithmus 1 sowie schematisch in Abbildung 3.3 dargestellt. Innerhalb des Algorithmus stellen die Schritte 1 und 7 die Glättungsoperation dar. Dabei bedeutet eine ν_i-malige Anwendung, genau ν_i Iterationsschritte mit der Iterationsvorschrift (3.7) durchzuführen. Hierbei wird für Vor- und Nachglättung normalerweise derselbe Glätter verwendet, was jedoch nicht zwingend erforderlich ist. Die Schritte 2 bis 6 dienen der Grobgitterkorrektur. Wesentlich ist dabei, dass die Problemdimension auf dem gröberen Gitter, auf dem das lineare Gleichungssystem $A_H e_H = r_H$ exakt gelöst wird, kleiner ist als die des Ausgangsproblems.

Algorithmus 1 Zweigitterkorrektur

Benötigt: A_h, A_H, I_h^H, I_H^h
(1) ν_1-malige Anwendung des Glätters $\mathcal{S}$ auf $A_h x_h = b_h$
(2) Berechnung des Residuums $r_h = b_h - A_h\tilde{x}_h$
(3) Restriktion von r_h auf das gröbere Gitter mittels I_h^H
(4) Lösen von $A_H e_H = r_H$
(5) Prolongieren von e_H auf das feinere Gitter mittels I_H^h
(6) Korrektur von $\tilde{x}_h$ durch e_h
(7) ν_2-malige Anwendung des Glätters $\mathcal{S}$ auf $A_h\tilde{x}_h = b_h$

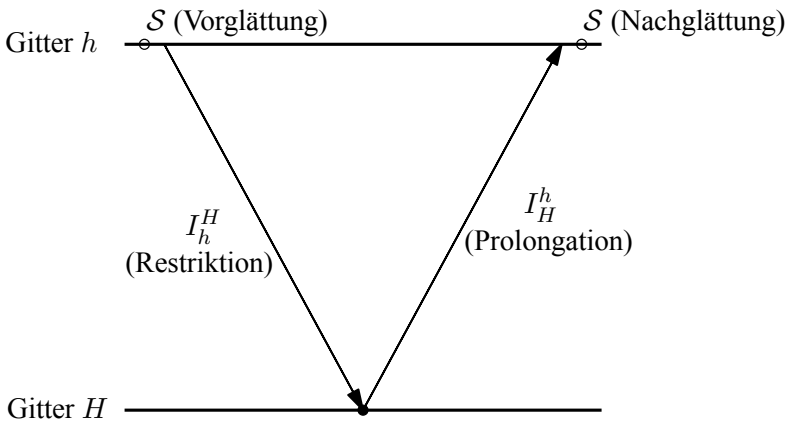

Abbildung 3.3.: Zweigittermethode (schematisch).

Anfangs wurde unterstellt, dass die Grobgitterkorrektur ebenfalls ein Iterationsverfahren der Form (3.7) darstellt. Dies ist nun leicht zu erkennen. Die explizite Gestalt der Korrekturgleichung in Schritt 6 lautet

$$\hat{x}_h = \tilde{x}_h + I_H^h e_H \tag{3.14}$$

und daher gilt mit $A_H e_H = r_H$ sowie $r_h = I_H^h r_H$ und (3.5)

$$\hat{x}_h = \tilde{x}_h + I_H^h A_H^{-1} I_h^H (b_h - A_h \tilde{x}_h).$$

Mit der Bezeichnung $B_h^{-1} := I_H^h A_H^{-1} I_h^H$ ergibt sich

$$\hat{x}_h = \tilde{x}_h + B_h^{-1}(b_h - A_h \tilde{x}_h) \tag{3.15}$$

und damit die Struktur eines iterativen Lösers entsprechend (3.7). Für den resultierenden Fehler $\hat{e}$ gilt

$$\begin{aligned} \hat{e}_h = \hat{x}_h - x^* &= \tilde{x}_h + I_H^h A_H^{-1} I_h^H (b_h - A_h \tilde{x}_h) - x^* \\ &= \tilde{x}_h - x^* - I_H^h A_H^{-1} I_h^H A_h (\tilde{x}_h - x^*) \\ \hat{e}_h &= \underbrace{(I - B_h^{-1} A_h)}_{\tilde{\mathcal{T}}_h^H} \tilde{e}_h, \end{aligned} \tag{3.16}$$

wobei $\tilde{\mathcal{T}}_h^H$ die Fehlerfortpf anzungsmatrix der Grobgitterkorrektur bezeichnet. Es sei bemerkt, dass eine bloße Anwendung der Grobgitterkorrektur nicht zwingend zu einer Reduktion des Fehlers führt, [71].

Lemma 3.2.1. *Die bloße Anwendung der Grobgitterkorrektur stellt kein konvergentes Verfahren dar. Es gilt*

$$\varrho\left(I_h - I_H^h A_H^{-1} I_h^H A_h\right) = \varrho\left(\tilde{\mathcal{T}}_h^H\right) \geq 1.$$

Beweis. Der Restriktionsoperator $I_h^H \in \mathbb{R}^{N \times n}$ mit $N < n$ besitzt einen nichttrivialen Kern. Daher existiert ein Fehler $e_h \neq 0$ mit $I_h^H e_h = 0$. Sei nun $\tilde{e}_h = A_h^{-1} e_h$. Dann gilt für die Fehlerfortpf anzung entsprechend Gleichung (3.16)

$$\hat{e}_h = \tilde{e}_h - I_H^h A_H^{-1} \underbrace{I_h^H \underbrace{A_h \tilde{e}_h}_{=e_h}}_{=0} = \tilde{e}_h$$

und damit $\varrho\left(\tilde{\mathcal{T}}_h^H\right) \geq 1$. □

Zusammengefasst ergibt sich im Falle der Zweigitterkorrektur unter Beachtung einer ν_1-maligen Vor- und ν_2-maligen Nachglättung mittels $\mathcal{S}$ sowie der Grobgitterkorrektur eine Fehlerfortpf anzungsgleichung der Gestalt

$$e_h^{(k+1)} = \mathcal{S}^{\nu_2}\tilde{\mathcal{T}}_h^H\mathcal{S}^{\nu_1}e_h^{(k)} = \mathcal{T}_h^H e_h^{(k)}. \tag{3.17}$$

Dabei nennt man $\mathcal{T}_h^H$ den *Zweigitteriterationsoperator*. Spektrale Eigenschaften dieses Operators bestimmen demzufolge die Konvergenzeigenschaften der Zweigitterkorrektur und er wird somit im Zuge der Konvergenzbetrachtungen im Abschnitt 3.4 nochmals eine Rolle spielen.
Den Abschluss der Betrachtungen zum Zweigitterverfahren sollen an dieser Stelle Bemerkungen zu den im Vorfeld erwähnten Transferoperatoren, der Restriktion beziehungsweise der Prolongation, bilden. Für die hier am Anfang konstruierten Gitter gilt, wie bereits bemerkt, die Inklusion $W_H \subset W_h$. Zur Berechnung der Korrektur muss das Residuum r_h mittels Restriktion auf das grobe Gitter abgebildet werden. Dabei kann dieser Restriktionsoperator auf verschiedene Weise def niert werden. Die einfachste Wahl stellt dabei der *Injektionsoperator* (*injection operator*) dar. Dieser weist jedem Grobgitterpunkt $x_i \in W_H$ den Wert des Residuums $r_h(x_i)$ des gleichen Punktes auf dem feineren Gitter $x_i \in W_h$ zu. Das heißt, für die Stützstellen $x_i \in W_H$ ist der zugehörige Fehlerwert durch

$$r_H(x_i) = r_h(x_i), \quad (i = 1, \ldots, N), \tag{3.18}$$

gegeben. Eine weitere Möglichkeit ist durch den *Mittelungsoperator* (*full weighting operator*) gegeben. Dieser bestimmt den Wert in Punkten des Grobgitters als gewichtetes Mittel aus der korrespondierenden Fehlergröße im Feingitterpunkt und den Fehlergrößen seiner direkten Nachbarn, vergleiche Abbildung 3.4 links. In Komponentenschreibweise gilt hierbei

$$r_H(x_i) = \frac{1}{4}[r_h(x_{i-1}) + 2r_h(x_i) + r_h(x_{i+1})], \quad (i = 1, \ldots, N). \tag{3.19}$$

Die resultierenden Restriktionsoperatoren haben in Matrixschreibweise die Gestalt

$$I_h^H = \begin{pmatrix} 1 & 0 & 0 & \cdots & & & \\ 0 & 0 & 1 & \cdots & & & \\ & & & \vdots & & & \\ & & & \cdots & 1 & 0 & 0 \\ & & & \cdots & 0 & 0 & 1 \end{pmatrix}$$

zur Umsetzung der in Gleichung (3.18) gegebenen Vorschrift beziehungsweise

$$I_h^H = \frac{1}{4}\begin{pmatrix} 1 & 2 & 1 & \cdots & & & \\ 0 & 0 & 1 & \cdots & & & \\ & & & \vdots & & & \\ & & & \cdots & 1 & 0 & 0 \\ & & & \cdots & 1 & 2 & 1 \end{pmatrix}$$

bei Anwendung der Restriktion entsprechend Gleichung (3.19).
Den zweiten Transferoperator bildet die Prolongation zur Abbildung der Grobgitterkorrektur e_H auf das feine Gitter. Auch hier existieren mehrere Varianten. Die am häuf gsten genutzte ist die *lineare Interpolation* (*linear interpolation*). Dabei wird für die in beiden Gittern enthaltenen Punkte $x_i \in W_h \cap W_H$ der Wert des Grobgitterpunktes übernommen, die Werte der Punkte, die im Grobgitter nicht enthalten sind, werden wiederum als gewichtetes Mittel der Nachbarpunkte berechnet, vergleiche Abbildung 3.4 rechts. Für die Komponenten gilt

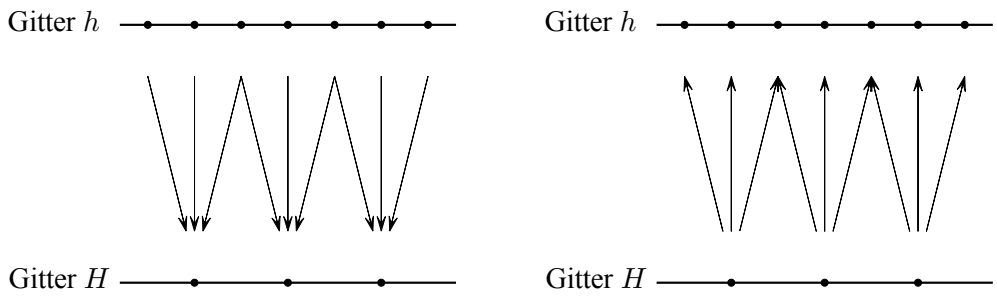

Abbildung 3.4.: Schematische Darstellung der Transferoperatoren. Links: Restriktion (full weighting operator). Rechts: Prolongation (linear interpolation).

$$e_h(x_i) = \begin{cases} e_H(x_i) & x_i \in W_H \cap W_h \\ \frac{1}{2}(e_H(x_{i-1}) + e_H(x_{i-1})) & x_i \in W_h \backslash W_H, \end{cases} \qquad (i = 2, \ldots, n-1).$$

Der zugehörige Prolongationsoperator besitzt daher unter Anpassung der Berechnung für die Randpunkte $i = 1$ und $i = n$ die Struktur

$$I_H^h = \frac{1}{2} \begin{pmatrix} 1 & 0 & \cdots & & \\ 2 & 0 & \cdots & & \\ 1 & 1 & \cdots & & \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ & & & 1 & 1 \\ & & & 0 & 2 \\ & & & 0 & 1 \end{pmatrix}.$$

Man beachte, dass sowohl die eben vorgestellten Restriktions- als auch die Prolongationsoperatoren für die im Vorfeld angegebene Vergröberungsstrategie mittels Verdoppelung der Schrittweite h, dem (geometrischen) *standard coarsening* für uniforme Gitter, gültig sind. Wählt man eine alternative Strategie zur Konstruktion des Grobgitters müssen diese Operatoren entsprechend angepasst werden.
Rückblickend bleibt festzuhalten, dass zur Umsetzung des Zweigitterverfahrens neben dem zu lösenden Gleichungssystem $A_h x_h = b_h$ folgende Komponenten notwendig sind:

- Das Feingitter W_h und das Grobgitter W_H,
- der Glätter $\mathcal{S}$,
- der Grobgitteroperator A_H sowie
- die Transferoperatoren (Restriktion und Prolongation).

Mehrgitterverfahren

Basierend auf dem Zweigitterverfahren kann nun das Mehrgitterverfahren konstruiert werden, [6, 33, 71]. Der Blick auf Algorithmus 1 zeigt, dass in Schritt 4 das lineare Gleichungssystem $A_H e_H = r_H$ auf dem Grobgitter gelöst wird. Damit weist dieser Schritt dieselbe Struktur des ursprünglich zu lösenden Problems $Ax = b$ auf. Ersetzt man nun diesen Schritt wiederum durch eine approximative Lösung mittels eines Zweigitterverfahrens, erhält man ein Dreigitterverfahren, wobei auch hier auf dem gröbsten Gitter ein Gleichungssystem zu lösen ist. Daher kann eine rekursive Anwendung des Zweigitterverfahrens erfolgen. Vorteil ist, dass die Problemdimension des jeweils gröberen Gitters geringer als die des feineren ist. Man wiederholt dieses rekursive Vorgehen solange, bis auf dem gröbsten Gitter die Anzahl

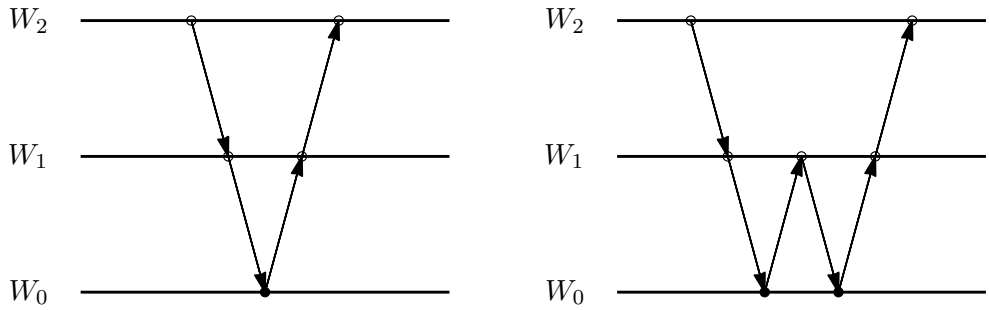

Abbildung 3.5.: Ausprägung des Mehrgitterverfahrens in Abhängigkeit von τ. Links: V-Zyklus ($\tau = 1$). Rechts: W-Zyklus ($\tau = 2$).

der Unbekannten und damit die Problemdimension so klein ist, dass der Aufwand zum exakten Lösen verhältnismäßig erscheint.
Basis dieses Vorgehens bildet eine „Familie“ von diskreten Operatoren A_l zu einer Gitterhierarchie mit den Gittern (oder Leveln) $l = 0, \ldots, L$. Dabei bezeichnet der Index $l = 0$ das gröbste und $l = L$ das feinste Level. Die Anzahl der Unbekannten auf den jeweiligen Leveln und damit die Problemdimension sei mit n_l gekennzeichnet, damit gilt $n_L = n$ und $A_L = A \in \mathbb{R}^{n \times n}$. Zur Abbildung der auftretenden Größen zwischen den Leveln sind L Operatoren für die Restriktion $I_l^{l-1} \in \mathbb{R}^{n_l \times n_{l-1}}$, $(l = 1, \ldots, L)$, sowie L Operatoren für die Prolongation $I_{l-1}^{l} \in \mathbb{R}^{n_{l-1} \times n_l}$, $(l = 0, \ldots, L-1)$, nötig.

Algorithmus 2 Mehrgitterverfahren (MG(A_l, b_l, x_l))

(1) Falls $l = 0$: Löse $A_0 x_0 = b_0$ exakt
(2) Sonst:
(2.1) ν_1-malige Anwendung des Glätters $\mathcal{S}$ auf $A_l x_l = b_l$
(2.2) Berechnung des Residuums $r_l = b_l - A_l x_l$
(2.3) Restriktion von r_l auf das gröbere Level (Gitter) mittels I_l^{l-1}
(2.4) τ-maliger Aufruf von MG$(A_{l-1}, r_{l-1}, 0)$
(2.5) Prolongieren von e_l auf das feinere Level (Gitter) mittels I_{l-1}^{l}
(2.6) Korrektur von x_l durch e_l
(2.7) ν_2-malige Anwendung des Glätters $\mathcal{S}$ auf $A_l x_l = b_l$

Die resultierende Struktur eines Mehrgitterverfahrens ist in Algorithmus 2 dargestellt. Der rekursive Aufruf des Mehrgitterverfahrens in Schritt 2.4 findet dabei für das Gleichungssystem $A_{l-1} e_{l-1} = r_{l-1}$ mit Startapproximation (*initial guess*) $e_{l-1} = 0$ statt. Der Parameter τ charakterisiert weiterhin die Ausprägung des Verfahrens. Im Fall $\tau = 1$ erhält man einen *V-Zyklus*, für $\tau > 1$ einen *W-Zyklus*. Dies ist schematisch in Abbildung 3.5 im Fall $L = 2$ dargestellt. Mit Blick auf die numerischen Untersuchungen im Kapitel 5 sowie der üblichen Notation in der Literatur wird ein V- beziehungsweise W-Zyklus unter Verwendung von ν_1 Vor- und ν_2 Nachglättungsschritten mit $V(\nu_1, \nu_2)$ beziehungsweise $W(\nu_1, \nu_2)$ gekennzeichnet.

3.3. Algebraische Mehrgitterverfahren

Im Wesentlichen existieren zwei Konzepte für Mehrgitterverfahren, die geometrischen und die algebraischen Mehrgitterverfahren, im Folgenden mit *MG*, *GMG* beziehungsweise *AMG* abgekürzt. Vorrangig

zur Lösung linearer Gleichungssysteme, die im Kontext der Lösung von Randwertaufgaben mit elliptischen Operatoren auftreten, kamen vorerst GMG zum Einsatz. Erste Ideen wurden dazu von den sowjetischen Mathematikern Federenko und Bakhvalov veröffentlicht, [3, 22, 23]. Mitte der siebziger Jahre wurden diese Ideen von Hackbusch beziehungsweise Brandt aufgegriffen und einer systematischen Analyse unterzogen, [11, 30]. Anfang der 1980-er Jahre erweiterte sich der Kreis der Autoren und so sind bis heute MG für eine Vielzahl von Problemen untersucht und erfolgreich auf diese angewendet worden. Eine umfangreiche Sammlung von Publikationen ist auf der Internetseite *MGNet* von Craig C. Douglas zu f nden, [18]. Im Zuge der Weiterentwicklung der MG entstand dann bereits Mitte der 1980-er Jahre die Idee der AMG. Diese als Ergänzung zu den GMG zu interpretierende Variante ist von Ruge und Stüben formuliert und durch das erschienene Programm *AMG1R5* realisiert worden, [51, 68]. Anfangs blieben die AMG im Hintergrund, aber seit Mitte der 1990-er Jahre ist die Weiterentwicklung voran getrieben worden. Heutzutage f ndet sich ein breites Anwendungsgebiet und verschiedenste Variationen von AMG, beispielsweise als (black-box-)Vorkonditionierer, wie es auch hier verwendet werden soll. Beispiele für die erfolgreiche Anwendung f nden sich unter anderem in [1, 5, 61, 62, 66, 67]. Es sei jedoch betont, dass GMG und AMG nicht als konkurrierende Methoden aufzufassen sind. So zeigen beide Verfahren je nach Problemstellung unterschiedliche Konvergenzeigenschaften auf, wie beispielsweise in [27] oder auch [76] verdeutlicht wird.

Wie bereits im Abschnitt 3.1 zur Motivation der MG angedeutet wurde, liegt die Ursache der Konvergenzbeschleunigung dieser Verfahren in einem effektivem Zusammenspiel von Glätter und Grobgitterapproximation zur Reduktion des Fehlers. Dabei weisen beide Zugänge die im vorherigen Abschnitt in Algorithmus 2 beschriebene Struktur auf. Der wesentliche Unterschied zwischen GMG und AMG liegt jedoch darin, wie Glätter und Grobgitter aufeinander abgestimmt werden, um die gewünschte Fehlerreduktion zu erreichen.

Wie die Bezeichnung suggeriert, sind GMG stark an die zu Grunde liegende Geometrie gekoppelt. Dies zeichnet sich bereits in den Erklärungen zum Zweigitterverfahren, speziell bei der Konstruktion der Gitter W_h und W_H und damit ebenso der Transferoperatoren in Abschnitt 3.2, ab. Der Einsatz der GMG erfolgt dabei meist auf adaptive Art. Das bedeutet, ausgehend von einer groben Diskretisierung, gegeben ist also das Level $l = 0$, wird vorerst eine angepasste Verfeinerung des Gitters, beispielsweise mittels geeigneter Fehlerschätzer, vorgenommen, [75]. Ist das Gitter fein genug, das heißt ist der Diskretisierungsfehler akzeptabel, so kann mittels der generierten Hierarchie nachiteriert und so die bis dahin vorliegende approximative Lösung verbessert werden. Dabei wird die Glättungsoperation so angepasst, dass eine effektive Dämpfung des oszillierenden Fehlers gesichert ist. Kurz, GMG basieren auf einer Hierarchie, die auf natürliche Weise aus der zu Grunde liegenden Geometrie abgeleitet wird, und wählen dazu einen effektiven Glätter.

AMG verfolgen im Gegensatz eine andere Strategie. Hier wird im Vorfeld ein meist einfacher Glätter, beispielsweise ein klassisches Iterationsverfahren, festgelegt und darauf basierend eine angepasste Vergröberungsstrategie zur Konstruktion der Hierarchie entwickelt. Dabei wird diese Strategie einzig auf Grundlage der gegebenen algebraischen Gleichungen, also des Gleichungssystems $Ax = b$, vorgenommen. Auf Grund der Verwandtschaft beider Methoden zeigen AMG daher viele der guten Eigenschaften der GMG, erlauben aber zudem eine von der Herkunft des Problems unabhängige Behandlung des linearen Gleichungssystems. Somit bieten AMG einen geometrie- beziehungsweise gitterfreien Zugang.

Ein wesentlicher Unterschied ergibt sich daher auch in der Umsetzung beider Methoden. Auf Grund des geometriefreien Ansatzes bei AMG kann die Konstruktion der Gitterhierarchie nur auf Basis des vorliegenden linearen Gleichungssystems vorgenommen werden. Damit ist eine (vorgeschaltete) adaptive Verfeinerung wie bei den GMG nicht möglich. Die Konstruktion der Hierarchie wird daher in einer dem eigentlichen Lösungsprozess vorgeschalteten *setup*-Phase realisiert. Die sich anschließende *solution*-Phase dient dann der eigentlichen approximativen Lösung des linearen Gleichungssystems, also

im vorliegenden Fall der Realisierung der Operation $y \mapsto B^{-1}y$.
Die nächsten Abschnitte geben einen Überblick zur Umsetzung der AMG. Ausgehend vom Glätter und seinen Eigenschaften soll zunächst der (algebraisch) glatte Fehler charakterisiert und daraus die Vergröberungsstrategie motiviert werden. Im Anschluss daran steht diese und damit die Konstruktion des Grobgitters beziehungsweise der Transferoperatoren im Mittelpunkt. Dabei soll einerseits das (klassische) *standard coarsening* für AMG als erprobte Variante für symmetrische reguläre M-Matrizen, eine spezielle Klasse der symmetrisch positiv def niten Matrizen, vorgestellt werden. Ergänzt wird dies durch einen alternativen Zugang, dem *smoothed aggregation*. Beide Methoden werden bei den numerischen Untersuchungen in Kapitel 5 Anwendung f nden. Den Abschluss des Kapitels bilden klassische Konvergenzabschätzungen zu Mehrgitterverfahren.

3.3.1. Glätter

Im Vorfeld wurde herausgestellt, dass sich der algebraische vom geometrischen Ansatz konzeptuell dadurch abgrenzt, dass zu einem gewählten Glätter eine passende Vergröberungsstrategie konstruiert wird. Dabei werden meist einfache Glätter, wie sie bereits in Abschnitt 3.1 mit dem Jacobi- und Gauß-Seidel-Verfahren vorgestellt wurden, involviert. Auf die Notwendigkeit einer Glättungsoperation wurde hingewiesen, da die reine Grobgitterkorrektur keine Konvergenz liefert, vergleiche Lemma 3.2.1. In diesem Abschnitt sollen dazu die Betrachtungen zum Glätter erweitert und der glatte Fehler aus algebraischer Sicht charakterisiert werden. Dabei gelten die folgenden Aussagen sowohl für den Einsatz der betrachteten Glätter im algebraischen als auch im geometrischen Mehrgitterverfahren. Ziel ist es an dieser Stelle, aus diesen Betrachtungen eine geeignete Strategie zur Konstruktion eines Grobgitters (und daraus einer Hierarchie) einzig auf Basis der algebraischen Gleichungen, also zur Umsetzung des algebraischen Mehrgitterverfahrens, zu motivieren.
Allgemein versteht man unter einem Glättungsschritt die einmalige Anwendung eines iterativen Verfahrens der Struktur (3.7). Der zugehörige Glätter beziehungsweise Glättungsoperator $\mathcal{S}$ entsteht aus der Betrachtung der Fehlerfortpf anzungsgleichung (3.8) und besitzt die Form

$$\mathcal{S} = I - B^{-1}A,$$

wobei B^{-1} der (symmetrisch positiv def nite) Vorkonditionierer ist und das jeweilige Iterationsverfahren spezif ziert. Dabei liegt Konvergenz der Glättungsoperation vor, wenn $\mathcal{S}$ eine Kontraktion ist, also der Spektralradius die Bedingung

$$\varrho(\mathcal{S}) = \|I - B^{-1}A\|_A \leq \gamma < 1$$

erfüllt. Der Glättungseffekt, beschrieben in Abschnitt 3.1, tritt durch ν-malige Anwendung des jeweiligen Glätters ein. Wie bereits beim Zweigitterverfahren erwähnt, f nden häuf g eine Vor- und Nachglättung mit jeweils ν_1 beziehungsweise ν_2 Schritten statt. Es wurde auch bemerkt, dass Vor- und Nachglätter nicht notwendigerweise dieselbe Glättungsoperation sein müssen. Jedoch ist dies (auch aus algorithmischer Sicht) üblicherweise der Fall. Dass die mehrmalige Anwendung eines Glätters wiederum als ein Glättungsschritt interpretiert werden kann, zeigt folgendes Lemma.

Lemma 3.3.1. *Sei $\mathcal{S} = I - B^{-1}A$ ein Glättungsoperator mit $\varrho(\mathcal{S}) < 1$. Dann ist die ν-malige Anwendung von $\mathcal{S}$ wiederum ein konvergenter Glätter der Form $\tilde{\mathcal{S}} = I - \tilde{B}^{-1}A$.*

Beweis. Sei $w \in \mathbb{R}^n$. Dann gilt

$$\begin{aligned}
\mathcal{S}^\nu w = (I - B^{-1}A)^\nu w &= \underbrace{(I - B^{-1}A) \cdot \ldots \cdot (I - B^{-1}A)}_{\nu\text{–mal}} w \\
&= \left(\sum_{i=0}^{\nu} I^{\nu-i} (-B^{-1}A)^i \right) w \\
&= \left(I - \binom{\nu}{1} B^{-1}A + \binom{\nu}{2} (B^{-1}A)^2 - \ldots \pm \binom{\nu}{\nu} (B^{-1}A)^\nu \right) w \\
&= \Bigg(I - \underbrace{\left(\binom{\nu}{1} B^{-1} + \binom{\nu}{2} B^{-1}AB^{-1} - \ldots \pm (B^{-1}A)^{\nu-1} B^{-1} \right)}_{\tilde{B}^{-1}} A \Bigg) w
\end{aligned}$$

und somit

$$\tilde{\mathcal{S}} = (I - \tilde{B}^{-1}A)w.$$

Sei nun (μ, v) ein Eigenpaar von $\mathcal{S}$ mit $1 > |\mu| = \varrho(\mathcal{S})$. Dann gilt

$$\mathcal{S}v = (I - B^{-1}A)v = \mu v$$

und daher

$$\mathcal{S}^\nu v = (I - B^{-1}A)^\nu v = \mu^\nu v = \tilde{\mathcal{S}} v.$$

Da $|\mu| < 1$ ist $|\mu^\nu| < 1$ und somit auch $\tilde{\mathcal{S}}$ eine Kontraktion. □

Ohne Beschränkung der Allgemeinheit kann also von einem einzelnen Vor- beziehungsweise Nachglättungsschritt ausgegangen werden.
In vielen Anwendungen für algebraische Mehrgitterverfahren, und so auch in den später verwendeten Algorithmen, f ndet das Gauß-Seidel-Verfahren Verwendung als Glättungsoperator. Für symmetrisch positiv def nite Matrizen ist die Konvergenz des Gauß-Seidel-Verfahrens sichergestellt. Um dies zu zeigen, benötigt man eine weitere Eigenschaft der Matrix A.

Lemma 3.3.2. *Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch positiv def nit. Dann ist $D = \operatorname{diag}(A)$ ebenfalls symmetrisch positiv def nit, insbesondere gilt also*

$$d_{ii} > 0, \quad i = 1, \ldots, n.$$

Beweis. Symmetrie von D als Diagonalmatrix ist offensichtlich. Da A symmetrisch positiv def nit ist, gilt für alle $x \in \mathbb{R}^n$ ungleich dem Nullvektor

$$x^T Ax > 0.$$

Insbesondere heißt dies für die Einheitsvektoren e_i

$$e_i^T Ae_i = a_{ii} > 0$$

und damit $d_{ii} = a_{ii} > 0$. Im Falle einer Diagonalmatrix sind die Eigenwerte aber gerade die Diagonalelemente. Somit ist D positiv def nit. □

Es kann nun die Konvergenz des Gauß-Seidel-Verfahrens für symmetrisch positiv def nite Matrizen gezeigt werden.

Satz 3.3.3 (Ostrowski, Reich)**.** *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv def nit und $A = D + L + L^T$ eine additive Zerlegung mit $L = \operatorname{tril}(A)$. Dann konvergiert das Gauß-Seidel-Verfahren mit $B = D + L$ für alle $x \in \mathbb{R}^n$. Insbesondere gilt also*

$$\varrho(\mathcal{S}) = \varrho(I - B^{-1}A) < 1.$$

Beweis. Es gilt

$$\mathcal{S} = I - B^{-1}A = (Q - I)(Q + I)^{-1} \tag{3.20}$$

mit

$$Q = A^{-1}(2B - A).$$

Sei $\mu \in \mathbb{C}$ ein Eigenwert von $\mathcal{S}$ und z der zugehörige Eigenvektor. Dann gilt

$$\mathcal{S}z = (Q - I)(Q + I)^{-1}z = \mu z$$

und weiter mit $y = (Q + I)^{-1}z$

$$\begin{aligned} (Q - I)y &= \mu(Q + I)y \\ Qy - y &= \mu Qy + \mu y \\ (1 - \mu)Qy &= (1 + \mu)y \\ Qy &= \tfrac{1+\mu}{1-\mu}y. \end{aligned}$$

Somit ist $\lambda = \frac{1+\mu}{1-\mu}$ ein Eigenwert zu $Q = A^{-1}(2B - A)$. Nach Konstruktion von $\mathcal{S}$ in (3.20) sind dessen Eigenwerte gegeben durch $\mu = \frac{\lambda-1}{\lambda+1}$ und daher ist

$$|\mu|^2 = \mu\bar{\mu} = \frac{\lambda - 1}{\lambda + 1} \cdot \frac{\bar{\lambda} - 1}{\bar{\lambda} + 1} = \frac{|\lambda|^2 - 1 + 2\Re(\lambda)}{|\lambda|^2 + 1 + 2\Re(\lambda)}, \tag{3.21}$$

wobei $\Re(\lambda)$ den Realteil von λ beschreibt. Demzufolge ist $|\mu| < 1$, falls $\Re(\lambda) > 0$ gilt. Damit bleibt zu zeigen, dass die Eigenwerte λ von Q in der rechten Halbebene liegen. Dazu betrachte man

$$A^{-1}(2B - A)x = \lambda x$$

und somit

$$x^T(2B - A)x = \lambda x^T Ax.$$

Transponieren liefert, mit $A = A^T$,

$$x^T(2B^T - A)x = \lambda x^T Ax.$$

Nach Addition erhält man

$$x^T(B + B^T - A)x = \lambda x^T Ax. \tag{3.22}$$

Nach Konstruktion von $B = D + L$ gilt für die linke Seite unter Nutzung der Symmetrieeigenschaften

$$\begin{aligned} B + B^T - A &= (D + L) + (D + L)^T - (D + L + L^T) \\ &= D + L + D^T + L^T - D - L - L^T \\ &= D. \end{aligned}$$

Daher ist auch diese Matrix nach Lemma 3.3.2 symmetrisch positiv def nit und mit (3.22) gilt $\Re(\lambda) > 0$. Man erhält für (3.21), dass

$$|\mu| < 1 \quad \text{und somit} \quad \varrho(\mathcal{S}) < 1.$$

□

Als Ergebnis bleibt festzuhalten, dass im Falle einer symmetrisch positiv def niten Matrix A die Anwendung des Gauß-Seidel-Verfahrens zur Reduktion des hochfrequenten Fehleranteils genutzt werden kann. Mit Blick auf die späteren Erörterungen und Berechnungen soll nun eine spezielle Klasse der symmetrisch positiv def niten Matrizen vorgestellt werden.

Def nition 3.3.4. *Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt* **M-Matrix***, falls sie die folgenden Eigenschaften erfüllt:*

(i) $a_{ii} > 0$,

(ii) $a_{ij} \leq 0,\ (i \neq j)$,

(iii) *A ist regulär und $A^{-1} \geq 0$, das heißt* $(A^{-1})_{ij} \geq 0,\ (i, j = 1, \ldots, n)$.

Bemerkung 3.3.5. *Äquivalent zur Def nition 3.3.4 ist die Existenz einer Darstellung von A in der Form*

$$A = sI - K$$

mit nichtnegativer Matrix K, das heißt $k_{ij} \geq 0$, und $s \geq \varrho(K)$, [4]. Falls zudem $s > \varrho(K)$ gilt, ist A eine positiv def nite M-Matrix.

Bemerkung 3.3.6. *Die Matrix $A = \text{tridiag}(-1, 2, -1)$ aus Beispiel 2 ist eine symmetrisch positiv def nite M-Matrix. Es gilt*

$$A = 2I - \text{tridiag}(1, 0, 1).$$

Auch wenn die Konvergenz des Glätters durch Satz 3.3.3 für die hier betrachtete Klasse von Matrizen sichergestellt ist, zeigt sich das in Abschnitt 3.1 beschriebene Konvergenzverhalten, dass nämlich nur hochfrequente Fehleranteile effektiv gedämpft werden. Wie bereits erwähnt, ist es Aufgabe der Grobgitterkorrektur, den verbleibenden glatten Anteil zu reduzieren. Bei algebraischen Mehrgitterverfahren ist dieser Fehler zudem die einzig verfügbare Größe, um eine geeignete Vergröberungsstrategie abzuleiten. Daher soll dieser Fehler im Folgenden genauer analysiert werden.

Def nition 3.3.7. *Ein Fehler $e = x - x^*$ heißt* **algebraisch glatt***, falls*

$$\mathcal{S}e \approx e \tag{3.23}$$

gilt. Für die Energienorm heißt dies gleichermaßen

$$\|\mathcal{S}e\|_A \approx \|e\|_A. \tag{3.24}$$

Dabei hilft diese Beschreibung des algebraisch glatten Fehlers jedoch noch nicht, um eine Strategie zur Vergröberung herzuleiten. Dazu muss der Glättungseffekt genauer untersucht werden. Zur Motivation dient folgendes Beispiel, [12, 69].

Beispiel 3. Bei der Druchführung eines Schrittes des Gauß-Seidel-Verfahrens berechnet sich die i-te Komponente der neuen Iterierten $x^{(k+1)}$ aus

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} + \frac{1}{a_{ii}}(f_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)}) \\ &= x_i^{(k)} + \frac{r_i^{(k)}}{a_{ii}}. \end{aligned}$$

Dabei ist $r_i^{(k)}$ das Residuum unmittelbar vor der Berechnung von $x_i^{(k+1)}$. Für den zugehörigen Fehler gilt folglich

$$x_i^{(k+1)} - x_i^* = e_i^{(k+1)} = e_i^{(k)} - \frac{r_i^{(k)}}{a_{ii}}. \tag{3.25}$$

Für einen algebraisch glatten Fehler $e^{(k+1)} \approx e^{(k)}$ heißt dies, dass der Korrekturterm $\left|\frac{r_i^{(k)}}{a_{ii}}\right|$ klein ist und damit insbesondere, dass $|r_i^{(k)}|$ klein ist.

Daraus resultiert zur weiteren Klassifikation des algebraisch glatten Fehlers folgende

Bemerkung 3.3.8. *Dass die während des Glättungsverfahrens auftretenden (skalierten) Residuen um vieles kleiner sind als der eigentliche Fehler, ist eine zu* (3.23) *äquivalente Eigenschaft des algebraisch glatten Fehlers.*

Ein Maß für den in Gleichung (3.25) beschriebenen Korrekturterm bildet der Ausdruck

$$\sum_{i=1}^{n} \frac{|r_i^{(k)}|^2}{a_{ii}} = (D^{-1}r^{(k)}, r^{(k)})_2 = (D^{-1}Ae^{(k)}, Ae^{(k)})_2 = \|e\|_{AD^{-1}A},$$

also die Operatornorm $\|\cdot\|_{AD^{-1}A}$. Aus Beispiel 3 und Bemerkung 3.3.8 kann somit geschlussfolgert werden, dass eine effektive Glättung vorliegt, solange $\|e\|^2_{AD^{-1}A}$ im Vergleich zu $\|e\|^2_A$ relativ groß ist. Liegt jedoch ein algebraisch glatter Fehler vor, so gilt

$$\|e\|^2_{AD^{-1}A} \ll \|e\|^2_A.$$

Diese Erkenntnisse sollen in einer Glättungseigenschaft für $\mathcal{S}$ festgehalten werden.

Definition 3.3.9. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit. Ein Glätter $\mathcal{S}$ erfüllt die* **(algebraische) Glättungseigenschaft** *oder auch* **smoothing property** *bezüglich einer Klasse $\mathcal{A}$ von Matrizen, falls die Abschätzung*

$$\|\mathcal{S}e\|^2_A \leq \|e\|^2_A - \vartheta\|e\|^2_{AD^{-1}A} \tag{3.26}$$

mit einer Konstanten $\vartheta > 0$ unabhängig von e für alle $A \in \mathcal{A}$ erfüllt ist. Ein **algebraisch glatter** *Fehler liegt vor, falls*

$$\|e\|^2_{AD^{-1}A} \ll \|e\|^2_A \tag{3.27}$$

gilt.

Für den hier im Vordergrund stehenden Glätter gilt

Bemerkung 3.3.10. *Das Gauß-Seidel-Verfahren erfüllt die Glättungseigenschaft* (3.26) *für die Klasse der symmetrisch positiv definiten M-Matrizen mit der Konstanten $\vartheta = 1/4$, vergleiche [12, 69].*

Auch wenn der exakte (algebraisch glatte) Fehler e nicht explizit zugänglich ist, kann, mittels seiner Klassifizierung über die Operatornormen in Gleichung (3.27), basierend auf e eine geeignete Strategie zur Konstruktion eines Grobgitters hergeleitet werden. Dies bildet den Schwerpunkt des folgenden Abschnitts.

3.3.2. Grobgitter-(korrektur)

Die zweite Komponente des Mehrgitterverfahrens bildet die Berechnung der Grobgitterkorrektur. Basis hierfür sind eine Gitterhierarchie aus $L + 1$ Leveln, wobei das feinste Level L durch das lineare Gleichungssystem gegeben ist, die Transferoperatoren der Restriktion und Prolongation sowie die Grobgitteroperatoren $A_l, (l = 0, \ldots, L - 1)$, vergleiche Abschnitt 3.2. Gegenstand dieses Abschnitts ist die Herleitung einer geeigneten Vergröberungsstrategie und damit die Konstruktion der angegebenen benötigten Komponenten für den Fall einer symmetrisch positiv def niten Matrix A. Als Eigenschaft der algebraischen Mehrgitterverfahren wurde mehrfach bemerkt, dass diese Vergröberungsstrategie entsprechend des zuerst festgelegten Glätters gewählt werden muss. Motivation für ein sinnvolles Vorgehen bildet daher der algebraisch glatte Fehler, charakterisiert im vorherigen Abschnitt. Ein weiteres wesentliches Merkmal, im Gegensatz zum geometrischen Zugang, ist es, dass dazu keinerlei Geometrie- beziehungsweise Gitterinformation notwendig sind. Wie nun ein solches Grobgitter (und damit eine Hierarchie) bestimmt werden kann, soll an zwei Varianten verdeutlicht werden. Einerseits ist dies das *standard coarsening* für algebraische Mehrgitterverfahren, beschrieben in [69], und andererseits die Methode *smoothed aggregation*, vorgestellt in [72]. Beide Methoden werden bei den numerischen Berechnungen zum Einsatz kommen und ihre Leistungsparameter untersucht.
Es sei bemerkt, dass obwohl sich der geometrische vom algebraischen Ansatz wesentlich unterscheidet, hier weiterhin mit den Begriffen des vorrangig im geometrischen Umfeld genutzten Vokabulars gearbeitet werden soll. Dies ist durchaus gerechtfertigt, denn identif ziert man die Diskretisierungspunkte im geometrischen Kontext mit den Knoten des gerichteten Graphen der Systemmatrix A, so kann die Indexmenge $\mathcal{N} = \{1, \ldots, n\}$, welche die Knotennummern enthält, als Menge von Gitterpunkten interpretiert werden. Geometrische Nachbarschaften bestehen demnach im algebraischen Kontext, falls zwei Knoten i und j durch eine Kante verbunden sind.
Die Struktur des Mehrgitterverfahrens, erläutert am Ende des Abschnitt 3.2, zeigt, dass dieses aus einer rekursiven Anwendung des Zweigitterverfahrens zu erhalten ist. Ebenso erfolgt die Konstruktion der Hierarchie, als Menge der Grobgitter $l = 0, \ldots, L - 1$, durch rekursive Anwendung der Vergröberungsstrategie. Ausgangspunkt bildet das feinste Level L, auf dem das zu lösende lineare Gleichungssystem gegeben ist, und damit die Indexmenge der Diskretisierungspunkte $\mathcal{N}_L = \{1, \ldots, n_L\}$. Beim *standard coarsening* ist es nun Ziel, diese Menge in zwei disjunkte Mengen C_l und F_l zu zerlegen. Eine solche Unterteilung führt auf ein *C/F-splitting* der Indexmenge, wobei C_l die auf dem gröberen Gitter (coarse grid) verbleibenden Knoten enthält und F_l somit diejenigen, welche im Zuge der Vergröberung nicht mehr betrachtet werden. Dabei sollte so vorgegangen werden, dass die berechneten Grobgitterkorrekturen möglichst gut auf dem Feingitter dargestellt werden und somit eine effektive Korrektur ermöglichen. Sich an den Ausführungen in [69] orientierend, soll daher aus dem Blickwinkel der Prolongation ein geeignetes Vorgehen hergeleitet werden. Dazu benötigt man folgende Abschätzung für die Energienorm.

Lemma 3.3.11. *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv def nit und $A = D + L + L^T$ eine additive Zerlegung entsprechend Def nition 3.1.1. Dann gilt für die Energienorm die Abschätzung*

$$\|x\|_A^2 \leq \|x\|_D \, \|x\|_{AD^{-1}A},$$

wobei $\|\cdot\|_D$ beziehungsweise $\|\cdot\|_{AD^{-1}A}$ die Operatornorm der jeweiligen symmetrisch positiv def niten Matrizen bezeichnen.

Beweis. Es gilt mit der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned}\|x\|_A^2 = (x, Ax)_2 &= (D^{1/2}x, D^{-1/2}Ax)_2 \\ &\leq (D^{1/2}x, D^{1/2}x)_2^{1/2}\,(D^{-1/2}Ax, D^{-1/2}Ax)_2^{1/2} \\ &= (x, Dx)_2^{1/2}\,(x, AD^{-1}Ax)_2^{1/2} \\ &= \|x\|_D\,\|x\|_{AD^{-1}A}.\end{aligned}$$

□

Wie erwähnt, bildet der algebraisch glatte Fehler e die Basis zur Konstruktion des *C/F-splittings*. Entsprechend Gleichung (3.27) ist dieser durch die Abschätzung

$$\|e\|_{AD^{-1}A}^2 \ll \|e\|_A^2$$

charakterisiert. Mit Hilfe der Abschätzung aus Lemma 3.3.11 folgt daraus ebenso die Beziehung

$$\|e\|_A^2 \ll \|e\|_D^2.$$

Eine Umformung der linken Seite, unter Berücksichtigung der Symmetrie von A, liefert

$$\begin{aligned}\|e\|_A^2 = (e, Ae)_2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}e_ie_j \\ &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} -a_{ij}e_i^2 + \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} -a_{ij}e_j^2 - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} -a_{ij}2e_ie_j + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}e_i^2 \\ &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} -a_{ij}(e_i - e_j)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}e_i^2.\end{aligned}$$

Mit der rechten Seite

$$\|e\|_D^2 = \sum_{i=1}^{n} a_{ii}e_i^2$$

gilt demnach insgesamt

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} -a_{ij}(e_i - e_j)^2 + \sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\right)e_i^2 \ll \sum_{i=1}^{n} a_{ii}e_i^2. \tag{3.28}$$

Ist nun A zusätzlich eine M-Matrix, so gilt einerseits nach Def nition 3.3.4, dass $a_{ij} \leq 0$ $(j \neq i)$. Andererseits gilt für die wichtigsten Fälle dieser Matrizenklasse $\sum_{j=1}^{n} a_{ij} \approx 0$, womit der zweite Term der linken Seite nahezu verschwindet. Betrachtet man nun unter diesen Annahmen die i-te Komponente der Summe des ersten Terms in Gleichung (3.28) erhält man

$$\frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{a_{ii}}\frac{(e_i - e_j)^2}{e_i^2} \ll 1, \quad (i = 1, \ldots, n). \tag{3.29}$$

Anhand dieser Beziehung erkennt man, dass der algebraisch glatte Fehler, der nur schlecht durch den Glätter $\mathcal{S}$ behandelt wird und für den also $\|e\|_{AD^{-1}A} \ll \|e\|_A$ gilt, sich nur geringfügig in solche Richtungen ändert, für die das Verhältnis $\frac{|a_{ij}|}{a_{ii}}$ groß ist. Man spricht in diesem Fall von einer *starken (negativen) Verbindung*, auch *strong (negative) connection*. Diese Verbindungen werden daher bei der Konstruktion des *C/F-splittings* am Ende der Betrachtungen zum *standard coarsening* eine wesentliche Rolle spielen. Vorerst soll der Blick auf den Prolongationsoperator gerichtet und damit auf eine geeignete Darstellung der Grobgitterkorrekturen auf dem Feingitter eingegangen werden.

Definition 3.3.12. *Sei $A \in \mathbb{R}^{n\times n}$. Dann ist die* **Nachbarschaftsmenge** *N_i des Punktes bzw. Knotens $i \in \mathcal{N}_l = \{1,\dots,n_l\}$ durch*

$$N_i = \{j : i \neq j,\ a_{ij} \neq 0\}$$

gegeben.

Es wird nun die im vorangegangenen Abschnitt beschriebene Eigenschaft des algebraisch glatten Fehlers erneut aufgegriffen, um eine geeignete (heuristische) Prolongation zu motivieren. Dazu betrachtet man die Fehlerkomponente $e_i,\ (i \in F_l)$ in einem Feingitterpunkt genauer. Obwohl diese im Vergleich zum im Punkt i vorliegenden Residuum r_i durchaus groß sein kann, vergleiche Bemerkung 3.3.8 (wobei der Iterationsindex k vernachlässigt sei), lässt sie sich approximativ als Linearkombination der Fehlerkomponenten in den Nachbarn mittels

$$(r_i \approx)\ a_{ii}e_i + \sum_{j\in N_i} a_{ij}e_j = 0$$

darstellen. Ausgehend hiervon gilt

$$e_i = -\frac{1}{a_{ii}} \sum_{j\in N_i} a_{ij}e_j.$$

Nun sind aber im Allgemeinen nicht alle Nachbarpunkte $j \in N_i$ Elemente des Grobgitters. Daher approximiert man diese Gleichung mittels

$$e_i = -\frac{1}{a_{ii}} \sum_{j\in N_i} a_{ij}e_j \approx \sum_{k\in P_i} w_{ik}e_k, \tag{3.30}$$

wobei $P_i \subseteq C_l \cap N_i$ eine Menge von Grobgitterpunkten ist und w_{ik} als geeignete Gewichte zu bestimmen sind. Eine sinnvolle Wahl für diese Gewichte ist durch

$$w_{ik} = -\alpha_i \frac{a_{ik}}{a_{ii}} \quad \text{mit} \quad \alpha_i = \frac{\sum_{j\in N_i} a_{ij}}{\sum_{k\in P_i} a_{ik}} \tag{3.31}$$

gegeben. Der Faktor α_i dient dazu, im Falle einer M-Matrix mit Zeilensumme Null konstante Fehlerfunktionen exakt zu interpolieren. Dass dies gilt, ist leicht ersichtlich, denn für $\sum_{j\in N_i} a_{ij} = 0$ ist

$$a_{ii}\left(1 - \sum_{k\in P_i} w_{ik}\right) = \sum_{j\in N_i} a_{ij} = 0$$

und daher $\sum_{k\in P_i} w_{ik} = 1$, da nach Lemma 3.3.2 $a_{ii} \neq 0$ gilt. Dabei beschränkten sich die bisherigen Betrachtungen auf die Feingitterpunkte $i \in F_l$. Für die verbleibenden Feingitterpunkte, die gleichzeitig im Grobgitter enthalten sind $(i \in C_l)$, erfolgt die Prolongation durch Verwendung der Werte des Grobgitters in diesen Punkten. Damit kann der Prolongationsoperator für das *standard coarsening* definiert werden. Er ist demnach durch

$$\left(I_{l-1}^{l}\right)_{ij} = \begin{cases} \delta_{ij} & \text{für } i \in C_l \\ w_{ij} & \text{für } i \in F_l \text{ und } j \in P_i \end{cases} \tag{3.32}$$

gegeben.

Bemerkung 3.3.13. *Der auf diese Weise gebildete Prolongationsoperator besitzt maximalen Rang, das heißt die Spalten von I_{l-1}^{l} sind paarweise linear unabhängig.*

Damit sind Abbildungen vom groben auf das feine Gitter möglich. Für den Transfer der Residuen vom feinen auf das grobe Gitter wird der Restriktionsoperator I_l^{l-1} benötigt. Im hier betrachteten Fall einer symmetrisch positiv definiten Matrix A_l wird dieser als Transponierte des Prolongationsoperators gewählt, also

$$I_l^{l-1} = \left(I_{l-1}^l\right)^T. \tag{3.33}$$

Um nun eine rekursive Anwendung der Vergröberungsstrategie vorzunehmen, benötigt man den Grobgitteroperator A_{l-1}. Im Fall der geometrischen Mehrgitterverfahren bestand die Möglichkeit diesen als Resultat einer Diskretisierung mit größerem Parameter h zu erhalten. Im algebraischen Fall kann A_{l-1} so nicht gewonnen werden, da keine geometrischen Informationen genutzt werden. Hier wird A_{l-1} mittels des *Galerkin-Prinzips* als

$$A_{l-1} = I_l^{l-1} A_l I_{l-1}^l \tag{3.34}$$

bestimmt. Der auf diese Weise gewonnene Operator A_{l-1} wird dabei als *Galerkin-Operator* bezeichnet. Unter der Voraussetzung, dass Gleichung (3.33) gilt und der Prolongationsoperator vollen Rang besitzt, erbt A_{l-1} die Symmetrie und positive Definitheit von A_l, denn

$$(A_{l-1}x_{l-1}, x_{l-1})_2 = (I_l^{l-1}A_l I_{l-1}^l x_{l-1}, x_{l-1})_2 = (A_l \underbrace{I_{l-1}^l x_{l-1}}_{y_l \neq 0}, I_{l-1}^l x_{l-1})_2 = (A_l y_l, A_l y_l)_2 > 0.$$

Damit ist ein rekursives Vorgehen zur Konstruktion weiterer Grobgitter nach beschriebenem Vorgehen sichergestellt. Über die eben genannten Eigenschaften des Operators A_{l-1} hinaus, genügt zudem der hieraus resultierende Zweigitteriterationsoperator $\tilde{\mathcal{T}}_l^{l-1}$ aus Gleichung (3.16) einem Variationsprinzip, [69].

Satz 3.3.14. *Sei $\tilde{\mathcal{T}}_l^{l-1} = I - I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l$ wie in Gleichung* (3.16)*. Dann ist $\tilde{\mathcal{T}}_l^{l-1}$ ein orthogonaler Projektor bezüglich der A-Norm, welche durch $A = A_l$ induziert wird. Weiterhin gilt*

$$\forall e_l : \|\tilde{\mathcal{T}}_l^{l-1} e_l\|_{A_l} = \min_{e_{l-1}} \|e_l - I_{l-1}^l e_{l-1}\|_{A_l},$$

das heißt, die Grobgitterkorrektur unter Einsatz des Galerkin-Operators liefert den kleinstmöglichen resultierenden Fehler $\tilde{e}_l = \tilde{\mathcal{T}}_l^{l-1} e_l$ bezüglich dieser A-Norm.

Beweis. Zum Nachweis der Projektionseigenschaft ist zu zeigen, dass $\tilde{\mathcal{T}}_l^{l-1}$ bezüglich des Skalarproduktes $(\cdot,\cdot)_{A_l}$ symmetrisch ist und zudem $(\tilde{\mathcal{T}}_l^{l-1})^2 = \tilde{\mathcal{T}}_l^{l-1}$ gilt. Die Symmetrie folgt, da

$$(\tilde{\mathcal{T}}_l^{l-1} y, y)_{A_l} = (A_l(I - I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l)y, y)_2 = (y, A_l(I - I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l)y)_2 = (y, \tilde{\mathcal{T}}_l^{l-1} y)_{A_l}$$

gilt. Weiterhin zeigt sich, dass

$$\begin{aligned}(I - I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l)^2 &= I - 2I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l + I_{l-1}^l A_{l-1}^{-1} \underbrace{I_l^{l-1} A_l I_{l-1}^l}_{A_{l-1}} A_{l-1}^{-1} I_l^{l-1} A_l \\ &= I - I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l\end{aligned}$$

erfüllt ist. Daher ist $\tilde{\mathcal{T}}_l^{l-1}$ ein orthogonaler Projektor.
Um die zweite Behauptung zu zeigen, benötigt man die Eigenschaften eines orthogonalen Projektors, hier mit Q bezeichnet. Diese seien hier der Vollständigkeit halber angeführt. Es sei dazu mit $\mathcal{R}(\cdot)$ das Bild einer Matrix gegeben. Dann gelten folgende Aussagen für Q.

i) $\mathcal{R}(Q) \perp_A \mathcal{R}(I-Q)$

ii) $\|u+v\|_A^2 = \|u\|_A^2 + \|v\|_A^2$ für $u \in \mathcal{R}(Q)$ und $v \in \mathcal{R}(I-Q)$

iii) $\forall u : \|Qu\|_A = \min_{v\in\mathcal{R}(I-Q)} \|u-v\|_A$

iv) $\|Q\|_A = 1$.

Zu i) Da Q symmetrisch und $Q^2 = Q$ gilt

$$(Qu,(I-Q)v)_A = (u,Q(I-Q)v)_A = (u,(Q-Q^2)v)_A = (u,0)_A = 0.$$

Zu ii) Eine Umformung der linken Seite ergibt

$$\|u+v\|_A^2 = ((u-v),(u-v))_A = (u,u)_A + 2(u,v)_A + (v,v)_A = (u,u)_A + (v,v)_A = \|u\|_A^2 + \|v\|_A^2,$$

da nach i) $(u,v) = 0$ gilt.

Zu iii) Mit der Zerlegung $u = Qu + (I-Q)u$ gilt

$$\begin{aligned}\min_{v\in\mathcal{R}(I-Q)} \|u-v\|_A^2 &= \min_{v\in\mathcal{R}(I-Q)} \|Qu + \underbrace{(I-Q)u - v}_{\in\mathcal{R}(I-Q)}\|_A^2 \\ &= \min_{\tilde{v}\in\mathcal{R}(I-Q)} \|Qu-\tilde{v}\|_A^2 = \min_{\tilde{v}\in\mathcal{R}(I-Q)} \left(\|Qu\|_A^2 - \|\tilde{v}\|_A^2\right) = \|Qu\|_A^2\end{aligned}$$

da ein $\tilde{v} \neq 0$ existiert, für das $(I-Q)v = 0$ gilt.

Zu iv) Mit der oben gegebenen Zerlegung von u erhält man

$$\begin{aligned}\|Q\|_A^2 &= \max_{u\neq 0} \frac{\|Qu\|_A^2}{\|u\|_A^2} \\ &= \max_{u\neq 0} \frac{\|Qu\|_A^2}{\|Qu\|_A^2 + \|(I-Q)u\|_A^2} \leq 1.\end{aligned}$$

Dabei wird Gleichheit angenommen, falls $u \in \mathcal{R}(Q)$ ist.

Es kann nun die zweite Behauptung des Satzes gezeigt werden. Sei dazu $Q = \tilde{\mathcal{T}}_l^{l-1}$. Dann ist $(I-Q) = I_{l-1}^l A_{l-1}^{-1} I_l^{l-1} A_l$ und damit $\mathcal{R}(I-Q) = \mathcal{R}(I_{l-1}^l)$. Aus der dritten der oben gezeigten Eigenschaften eines orthogonalen Projektors folgt damit

$$\|\tilde{\mathcal{T}}_l^{l-1} e_l\|_{A_l} = \min_{e_{l-1}} \|e_l - I_{l-1}^l e_{l-1}\|_{A_l}$$

und damit die Behauptung. □

Bemerkung 3.3.15. *Der vorangegangene Satz unterstreicht zudem, dass wie in Lemma 3.2.1 gezeigt, die reine Grobgitterkorrektur kein konvergentes Verfahren erzeugt, da $\|\tilde{\mathcal{T}}_l^{l-1}\|_A = 1$ gilt. Ist aber der Glätter $\mathcal{S}$ eine Kontraktion, stellt das Zwei- beziehungsweise Mehrgitterverfahren ein konvergentes Verfahren dar.*

Damit sind fast alle Komponenten zur Realisierung eines Zwei- beziehungsweise Mehrgitterverfahrens vorgestellt. Einzig die genaue Konstruktion eines geeigneten *C/F-splittings* wurde noch nicht betrachtet. Dies soll nun auf Basis der im Vorfeld gegebenen theoretischen Betrachtungen nachgeholt werden.

Die Konstruktion ist im Wesentlichen von zwei Gesichtspunkten geleitet. Einerseits resultiert aus einer geringen Menge verbleibender Grobgitterpunkte ein reduzierter Rechenaufwand zur Berechnung der Grobgitterkorrektur auf Grund der geringeren Anzahl Unbekannter. Andererseits wurde bei den Betrachtungen zum Interpolationsoperator bemerkt, dass zur effektiven Darstellung der berechneten Korrektur auf dem feinen Gitter eine ausreichende Menge Grobgitterpunkte nötig ist, vergleiche (3.30). Die folgenden Erörterungen leiten nun den ersten erwähnten Algorithmus, das *standard coarsening* für symmetrisch positiv def nite Matrizen, her.

Grobgitterkonstruktion mittels „standard coarsening“

Ziel des *standard coarsening* ist es, die Indexmenge $\mathcal{N}_l = \{1, \ldots, n_l\}$ in zwei disjunkte Mengen F_l und C_l zu unterteilen, sodass der unter Gleichung (3.32) beschriebene Prolongationsoperator gebildet werden kann. Basis hierfür bilden die starken negativen Verbindungen, welche im Zuge der Interpretation von Gleichung (3.29) beschrieben wurden. Dabei ist die Variable i stark negativ mit der Variablen j verbunden, falls

$$-a_{ij} \geq \epsilon \cdot \max_{a_{ik}} |a_{ik}| \tag{3.35}$$

mit $0 < \epsilon < 1$ gilt. Die Wahl von ϵ wird in der Literatur als unkritisch beschrieben, mit Blick auf ein effektives Mehrgitterverfahren stellt sich $\epsilon = 0.25$ in Anwendungen als vernünftiger Wert heraus. Weiterhin sei S_i die Menge aller zum Punkt $i \in \mathcal{N}_l$ stark negativ verbundenen Punkte, also

$$S_i = \{j \in N_i : -a_{ij} \geq \epsilon \max_{a_{ik}} |a_{ik}|\}, \quad (i = 1, \ldots, n_l),$$

mit der Nachbarschaftsmenge N_i aus Def nition 3.3.12. Da im Allgemeinen $S_i \neq S_j$ gilt, sei ferner mit S_i^T die Menge aller Punkte j, welche gleichzeitig stark negativ zum Punkt i verbunden sind, gekennzeichnet, also

$$S_i^T = \{j \in \mathcal{N}_l : i \in S_j\}.$$

Mithilfe dieser Menge kann ein Wichtungsmaß κ_i für Punkte einer Menge U_l in der Form

$$\kappa_i = |S_i^T \cap U_l| + 2|S_i^T \cap F_l|$$

def niert werden. Dabei bezeichnet $|\cdot|$ die Mächtigkeit einer Menge, also die Anzahl der enthaltenen Elemente, und U_l ist im weiteren Verlauf die Menge der noch nicht zu C_l beziehungsweise F_l zugeordneten Punkte. Anhand dieses Maßes kann nun eine geeignete Unterteilung der Punkte in U_l vorgenommen werden. Dazu werden im Vorfeld ($U_l = \mathcal{N}_l$) für alle Punkte $i \in U_l$ die Gewichte κ_i ermittelt. Der Punkt mit der höchsten Wichtung κ_i wird C-Variable, also Element von C_l. Alle zu diesem Punkt stark negativ verbundenen Punkte werden der Menge F_l zugeordnet und damit F-Variablen. Gleichzeitig werden die zugeordneten Punkte aus der Menge U_l entfernt. Anschließend aktualisiert man die Gewichte der verbleibenden Punkte und bestimmt wiederum die Variable mit höchstem κ_i als C-Variable. Dies wird solange wiederholt, bis keine unsortierten Punkte mehr in U_l verbleiben. Der Einsatz des Gewichtes sichert dabei ein solches Vorgehen, dass keine „zufälligen“ *C/F-splittings* entstehen. Es ist dazu so gewählt, dass Punkte, die viele starke negative Verbindungen besitzen auf dem Grobgitter verbleiben, dies sichert der erste Term in der Summe zur Bestimmung des Wertes von κ_i. Der zweite Term dieser Summe dient dazu, bei der Unterteilung möglichst zu verhindern, dass zwei stark negativ verbundene Punkte gleichzeitig C-Variablen werden. Der Algorithmus 3 skizziert das beschriebene Vorgehen. Abbildung 3.6 illustriert den Beginn einer solchen Unterteilung. Zu Grunde liegt dabei die äquidistante Diskretisierung der in Beispiel 1 gegebene Differentialgleichung im zweidimensionalen Fall auf dem Gebiet $\Omega = [a, b] \times [c, d]$. Die Buchstaben symbolisieren hier die Knoten mit ihrer Zugehörigkeit zu den (dort nicht indizierten) Mengen U, F und C im Zuge der Konstruktion des *C/F-splittings*.

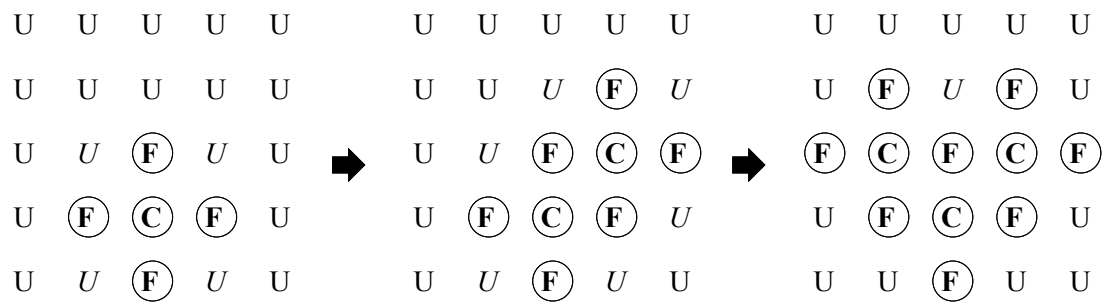

Abbildung 3.6.: Erste Schritte des *standard coarsening*. Kursive Buchstaben deuten auf Variablen mit höchstem Gewicht κ hin.

Algorithmus 3 Konstruktion eines C/F-splittings mittels *standard coarsening*

(1) Setze $U_l = \mathcal{N}_l$, $C_l = \varnothing$ und $F_l = \varnothing$
(2) Berechne κ_i für $i \in U_l$
(3) Solange $\kappa_i \neq 0,\ i \in U_l$
(3.1) Bestimme $\max_i \kappa_i$
(3.2) Bilde $C_l = C_l \cup \{i\}$ und $U_l = U_l \backslash \{i\}$
(3.3) Aktualisiere Mengen $\forall j \in S_i^T : F_l = F_l \cup \{j\}$ und $U_l = U_l \backslash \{j\}$
(3.4) Berechne neue Gewichte κ_i für $i \in U_l$

Damit ist der Algorithmus des *standard coarsening* für algebraische Mehrgitterverfahren in seinen einzelnen Komponenten erklärt. Dieses Verfahren wird bei den numerischen Untersuchungen in Form einer Implementierung basierend auf der *Algebraic MultiGrid-Toolbox* von Menno Verbeek, Jane Cullum und Wayne Joubert, [74], Anwendung f nden.

Bemerkung 3.3.16. *Die mittels standard coarsening erhaltene Menge von Grobgitterpunkten $\mathcal{N}_{l-1}$ genügt einer der in Abschnitt 3.2 beschriebenen, ähnlichen Inklusion, namentlich $\mathcal{N}_{l-1} \subset \mathcal{N}_l$. Dass dies, im Gegensatz zu den geometrischen, für die algebraischen Mehrgitterverfahren nicht zwangsläuf g auftreten muss, zeigt die folgende Betrachtung.*

Am Anfang des Abschnitts 3.3 ist neben dem *standard coarsening* eine weitere Variante zur Grobgitterkonstruktion erwähnt worden, die Methode *smoothed aggregation*. Auch diese soll nun, auf Grund ihres Einsatzes bei den numerischen Untersuchungen, genauer erläutert werden.

Grobgitterkonstruktion mittels „smoothed aggregation“

Auch diese Strategie zum Erhalt eines Grobgitters legt wie das *standard coarsening* die Konstruktion eines geeigneten Prolongationsoperators I_{l-1}^l zu Grunde, [72]. Ausgangspunkt ist wiederum der algebraisch glatte Fehler e_l auf dem Level l. Im Idealfall liegen alle diese Fehler im Bild des Prolongationsoperators $\mathcal{R}(I_{l-1}^l)$, das heißt,

$$\forall e_l \ \exists e_{l-1} : e_l = I_{l-1}^l e_{l-1}.$$

Damit wäre folglich eine exakte Korrektur der Näherungslösung auf dem feineren Gitter l möglich. Diese Forderung kann aber im Allgemeinen nicht erfüllt werden, es kann also kein solcher Prolongationsoperator konstruiert werden. Die Idee ist es nun, einen Prolongationsoperator derart zu formulieren, dass sein Bild $\mathcal{R}(I_{l-1}^l)$ möglichst nur algebraisch glatte Fehler enthält, um eine effektive Korrektur auf dem Level l sicherzustellen. Damit diese Eigenschaft des Bildes $\mathcal{R}(I_{l-1}^l)$ erhalten wird, betrachtet man die

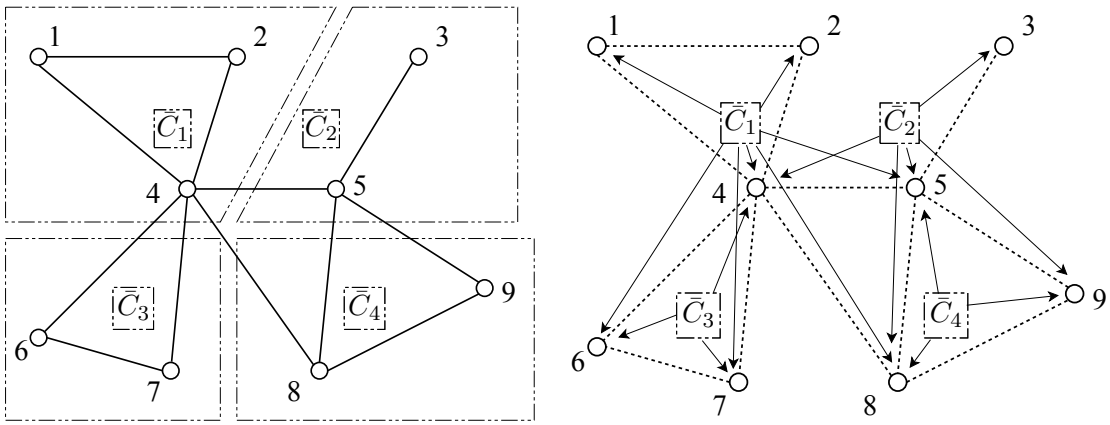

Abbildung 3.7.: Graph der starken negativen Verbindungen einer f ktiven Matrix A. Links: Unter Anwendung des Algorithmus 4 erhaltene Aggregate $\bar{C}_j$. Rechts: Resultierende Prolongation vom Grob- auf das Feingitter mittels der Werte $\bar{C}_j$ unter Verwendung des Jacobi-Verfahrens.

Charakterisierung des algebraisch glatten Fehlers in Def nition 3.3.7. Danach wird dieser auf dem Level l durch $\mathcal{S}e_l \approx e_l$ beschrieben. Um sicherzustellen, dass die geforderte Eigenschaft für einen (vorläuf gen) Prolongationsoperator $\tilde{I}_{l-1}^l$ gegeben ist, muss demnach

$$\mathcal{S}\underbrace{\tilde{I}_{l-1}^l e_{l-1}}_{e_l} \approx \underbrace{\tilde{I}_{l-1}^l e_{l-1}}_{e_l}$$

gelten. Wählt man nun den eigentlichen Prolongationsoperator in der Form

$$I_{l-1}^l = \mathcal{S}\tilde{I}_{l-1}^l, \tag{3.36}$$

so ist die Forderung, dass das Bild $\mathcal{R}(I_{l-1}^l)$ möglichst nur algebraisch glatte Fehler enthält, erfüllt. Folglich wird der Prolongationsoperator in zwei Schritten gebildet. Zuerst formuliert man den vorläuf gen Operator $\tilde{I}_{l-1}^l$. Anschließend erhält man die eigentliche Prolongation entsprechend Gleichung (3.36). Die Konstruktion von $\tilde{I}_{l-1}^l$ beruht dabei auf der *unknowns aggregation technique*, also einer geeigneten Aggregation von Punkten der Menge $\mathcal{N}_l$, [72]. Ergebnis ist eine Zerlegung der Menge $\mathcal{N}_l$ in disjunkte Teilmengen $\bar{C}_j$. Jede dieser Teilmengen (oder auch Aggregate) $\bar{C}_j$ wird dann durch einen „Grobgitterpunkt“ klassif ziert. Wesentliche Gemeinsamkeit zum *standard coarsening* ist wiederum die Verwendung der starken negativen Verbindungen eines Punktes $i \in \mathcal{N}_l$, siehe Gleichung (3.35), zur Bestimmung der Aggregate. Im Einzelnen gestaltet sich das Verfahren folgendermaßen, vergleiche auch Algorithmus 4. In einem ersten Durchlauf werden die Punkte sequentiell durchlaufen. Falls die gesamte Menge S_i der zu i stark negativ verbundenen Punkte noch nicht zugeordnet wurde, bildet diese das Aggregat $\bar{C}_j$. Im Allgemeinen werden nach Überprüfung aller Punkte $i \in U_l (= \mathcal{N}_l)$ noch nicht sämtliche Punkte zugeordnet sein. Dies geschieht im zweiten Durchlauf. Exemplarisch ist dies in der Abbildung 3.7 links zu erkennen. Es sind der Graph einer f ktiven Matrix $A \in \mathbb{R}^{9\times 9}$ und die aus Algorithmus 4 resultierenden Aggregate $\bar{C}_1$ bis $\bar{C}_4$ dargestellt.

Algorithmus 4 Konstruktion eines C/F-splittings mittels *smoothed aggregation*

(1) Setze $U_l = \mathcal{N}_l$ und $j = 0$
(2) Für $i = 1, \ldots, |U_l|$
(2.1) Falls $S_i \subset U_l$ setze $j = j + 1$, $\bar{C}_j = S_i$, $U_l = U_l \backslash \bar{C}_j$
(3) Für $i = 1, \ldots, |U_l|$
(3.1) Falls $i \in U_l$ setze $j = j + 1$, $\bar{C}_j = S_i \cup U_l$, $U_l = \backslash \bar{C}_j$

Basierend auf der durch Algorithmus 4 erhaltenen Zerlegung kann anschließend der vorläuf ge Prolongationsoperator konstruiert werden. Dieser ist durch

$$\tilde{I}_{l-1}^{l} = \begin{cases} 1 & \text{für } i \in \bar{C}_j \\ 0 & \text{für } i \notin \bar{C}_j \end{cases}, \forall i \in U_l, \forall j$$

gegeben. Entsprechend Gleichung (3.36) ergibt sich der eigentliche Prolongationsoperator dann in Abhängigkeit vom genutzten Glätter $\mathcal{S}$. Im Fall des Jacobi-Verfahrens ($\mathcal{S} = I - D^{-1}A$) ist exemplarisch auf der rechten Seite von Abbildung 3.7 die Prolongation im Fall der f ktiven Matrix A illustriert. Dabei werden die Fehlerkomponenten in den Feingitterpunkten nicht ausschließlich mittels der zugehörigen Grobgitterkomponente $\bar{C}_j$ ermittelt, sondern anteilig auch aus anderen Grobgitterkomponenten.
Auch beim *smoothed aggregation* verwendet man als Restriktionsoperator das Transponierte der Prolongation. Ebenso ergibt sich die Grobgitterapproximation A_{l-1} als Galerkin-Operator, wie es bereits bei den Betrachtungen zum *standard coarsening* erläutert wurde.

Bemerkung 3.3.17. *Es sei betont, dass die beschriebene Vergröberungsstrategie des smoothed aggregation einen rein algebraischen Zugang besitzt. Dies wird bei der Konstruktion des Grobgitters deutlich. Die ermittelten „Grobgitterpunkte“, oder besser die Knoten des Graphen der Matrix A_{l-1}, sind (im Gegensatz zum vorher erläuterten standard coarsening) selbst beim Vorliegen einer zum Problem zugehörigen Geometrie nicht mehr mit Gitterpunkten im geometrischen Sinne identif zierbar. Auch die Nachbarschaften def nieren sich einzig über die Kanten des Graphen der Matrix.*

Bei den numerischen Untersuchungen wird die Methode *smoothed aggregation* in Form des *Multilevel Preconditioning Package* des *Sandia National Laboratory* realisiert, [24].

Damit wurden zwei geeignete Methoden zur Konstruktion einer Gitterhierarchie, der zugehörigen Transferoperatoren und der Grobgitterapproximationen für den Fall symmetrisch positiv def niter Matrizen vorgestellt. Beide Verfahren f nden bei den numerischen Untersuchungen Anwendung. Der Vollständigkeit halber sei angemerkt, dass mittlerweile eine Vielzahl von algebraischen Mehrgitterverfahren, die auf unterschiedlichen Vergröberungsstrategien beruhen, konzipiert wurden. Dazu gehören unter anderem *AMGe*, vorgestellt in [15], eine Weiterentwicklung *spectral AMGe*, [17], *adaptive AMG*, [16] oder auch *bootstrap AMG*, [13, 14]. Einen weiteren Zugang stellt das *AGgregation-based algebraic MultiGrid* (AGMG) dar, welches auch als Programmcode verfügbar ist, [60, 63].

3.4. Konvergenz der Mehrgitterverfahren

Den Abschluss dieses Kapitels bildet eine kurze Betrachtung der Konvergenz von Mehrgitterverfahren zur Lösung des linearen Gleichungssystems $Ax = b$ mit symmetrisch positiv def niter Matrix A. Dabei sollen im Wesentlichen zwei Konvergenzbeweise Erwähnung f nden. Diese gehen zurück auf Braess und Hackbusch, [7], beziehungsweise Bramble, Pasciak und Xu, [9, 10, 77]. Es sei bemerkt, dass beide Beweise ihren Ursprung in den geometrischen Mehrgitterverfahren f nden. Da aber, wie am Anfang des Kapitels beziehungsweise im Abschnitt 3.2 dargestellt, sowohl der geometrische als auch der algebraische

Ansatz dieselbe Struktur aufweisen, sind die Aussagen auch auf das algebraische Mehrgitterverfahren übertragbar.
Bereits bei den Erläuterungen zur Motivation der Mehrgitterverfahren wurde erwähnt, dass diese aus einer Kombination von zwei Iterationsverfahren der Gestalt (3.7) entstehen. Dies sind einerseits der Glätter zur Reduktion der hochfrequenten Fehleranteile und andererseits die Grobgitterkorrektur zur Behandlung der glatten Fehleranteile. Basierend auf dieser Struktur formulierten Braess und Hackbusch die generelle Vorgehensweise zum Nachweis der Konvergenz. Dabei f ndet eine getrennte Analyse des Glätters und der Grobgitterkorrektur statt. Durch Zusammenführung der einzelnen Abschätzungen erhält man abschließend die Konvergenzrate des Mehrgitterverfahrens.
Bevor dies im Einzelnen dargelegt wird, soll, in Anlehnung an den Zweigitteriterationsoperator aus Gleichung (3.17), die Fehlerfortpf anzungsmatrix des Mehrgitterverfahrens hergeleitet werden, vergleiche auch [32, 71].

Satz 3.4.1. *Gegeben seien das lineare Gleichungssystem $Ax = b$, eine Hierarchie mit Leveln $l = 0, \dots, L$, die levelspezif schen Glättungsoperatoren $\tilde{\mathcal{S}}_l$ wie in Lemma 3.3.1 sowie die zugehörigen Prolongations- und Restriktionsoperatoren I_{l-1}^{l} bzw. I_l^{l-1}. Dann konvergiert das Mehrgitterverfahren, falls*

$$\varrho(\mathcal{T}_l) \le \gamma < 1, \quad (l = 0, \dots, L),$$

wobei die Operatoren $\mathcal{T}_l$ rekursiv durch

$$\begin{aligned} \mathcal{T}_0 &= 0 \\ \mathcal{T}_l &= \tilde{\mathcal{S}}_l(I_l - I_{l-1}^{l}(I_{l-1} - \mathcal{T}_{l-1}^{\tau})A_{l-1}^{-1} I_l^{l-1} A_l)\tilde{\mathcal{S}}_l, \end{aligned} \tag{3.37}$$

mit τ der Anzahl der Mehrgitteraufrufe und I_l beziehungsweise I_{l-1} den Einheitsmatrizen passender Dimension, auf dem jeweiligen Level l, gegeben sind.

Beweis. Sei $l = 1$. Dann ist

$$\begin{aligned} \mathcal{T}_1 &= \tilde{\mathcal{S}}_1(I_1 - I_0^1(I_0 - \mathcal{T}_0^{\tau})A_0^{-1} I_1^0 A_1)\tilde{\mathcal{S}}_1 \\ &= \tilde{\mathcal{S}}_1(I_1 - I_0^1 A_0^{-1} I_1^0 A_1)\tilde{\mathcal{S}}_1. \end{aligned}$$

Bleibt zu zeigen, dass für $l > 1$ die exakte Lösung auf dem gröberen Gitter mittels A_{l-1}^{-1}, da $l - 1$ nun nicht mehr das gröbste Level darstellt, durch die approximative Lösung mittels $B_{l-1}^{-1} := (I_{l-1} - \mathcal{T}_{l-1}^{\tau})A_{l-1}^{-1}$ ersetzt wird. Nach (3.15) gilt die Iterationsvorschrift für das gröbere Level $m = l - 1$

$$x_m^{(k+1)} = x_m^{(k)} + B_m^{-1}(b_m - A_m x_m^{(k)}) \tag{3.38}$$

oder äquivalent

$$x_m^{(k+1)} = (I_m - B_m^{-1} A_m)x_m^{(k)} + B_m^{-1} b_m.$$

Dabei ist B_m^{-1} also der Vorkonditionierer, dessen Wirkung aus einem Schritt einer Grobgitterkorrektur besteht. Bei der Durchführung von τ Schritten heißt dies mit $\mathcal{T}_m = (I_m - B_m^{-1} A_m)$

$$\begin{aligned} x_m^{(1)} &= \mathcal{T}_m x_m^{(0)} + B_m^{-1} b_m \\ x_m^{(2)} &= \mathcal{T}_m(\mathcal{T}_m x_m^{(0)} + B_m^{-1} b_m) + B_m^{-1} b_m = \mathcal{T}_m^2 x^{(0)} + \mathcal{T}_m B_m^{-1} b_m + B_m^{-1} b_m \\ &\vdots \\ x_m^{(\tau)} &= \mathcal{T}_m^{\tau} x_m^{(0)} + \mathcal{T}_m^{\tau-1} B_m^{-1} b_m + \dots + B_m^{-1} b_m. \end{aligned}$$

Mit dem Nullvektor als Startwert $x_m^{(0)}$ für die Iteration (3.38), vergleiche den rekursiven Aufruf in Algorithmus 2, und unter Verwendung der Neumann-Reihe, da $\varrho(\mathcal{T}_m) < 1$, ist

$$\begin{aligned} x_m^{(\tau)} &= \left(\sum_{i=0}^{\tau-1} \mathcal{T}_m^i\right) \cdot B_m^{-1} b_m \\ &= \left(\sum_{i=0}^{\tau-1} \mathcal{T}_m^i\right) \cdot (I_m - \mathcal{T}_m^\tau) A_m^{-1} b_m \\ &= (I_m - \mathcal{T}_m^\tau) A_m^{-1} b_m. \end{aligned}$$

Dies bedeutet, dass die Berechnung einer approximativen Lösung auf dem gröberen Level $l \neq 0$ im Mehrgitterfall, anstelle des exakten Lösens mittels A_0^{-1}, durch den Operator $(I_{l-1} - \mathcal{T}_{l-1}^\tau) A_{l-1}^{-1}$ (mit $m = l - 1$) ersetzt wird. Bezieht man nun noch die Vor- und Nachglättung auf dem Level l mit ein, ergibt dies den Operator $\mathcal{T}_l$ aus der Gleichung (3.37). □

Bemerkung 3.4.2. *Für $l = L$ heißt $\mathcal{T}_L$ der* **Mehrgitteriterationsoperator**. *Seine spektralen Eigenschaften bestimmen über die Konvergenz des Mehrgitterverfahrens. Die in Satz 3.4.1 hergeleitete Rekursion dient damit innerhalb der numerischen Untersuchungen im Kapitel 5 zur Bestimmung des Spektralradius von $\mathcal{T}_L$ und somit der Konvergenzrate γ.*

Konvergenzrate nach Hackbusch

Der folgende Konvergenzsatz von Hackbusch, [31, 32], beziehungsweise Braess und Hackbusch, [7], wird als die erste zufriedenstellende Konvergenzaussage für (geometrische) Mehrgitterverfahren zur Lösung linearer Gleichungssysteme resultierend aus der numerischen Behandlung von Randwertproblemen mit elliptischen Differentialoperatoren zitiert, [78]. Dabei wird jedoch gleichzeitig auf ein Problem vieler Beweise der 1980-er Jahre hingewiesen. Wie im hier skizzierten Beweis auch deutlich wird, basieren diese nämlich stark auf der zu Grunde liegende Geometrie und der Regularität des Randwertproblems. Der Konvergenzsatz lautet folgendermaßen.

Satz 3.4.3. *Die Energienorm des Mehrgitteriterationsoperators $\mathcal{T}_L$ genügt der Abschätzung*

$$\|\mathcal{T}_L\|_A \leq \frac{C}{C + 2\nu} < 1.$$

Dabei ist C eine von der Levelanzahl L und der Anzahl der Glättungsschritte ν unabhängige Konstante.

Auch hier erkennt man deutlich, dass eine bloße Grobgitterkorrektur $(\nu = 0)$ nicht zwingend ein konvergentes Verfahren nach sich zieht, vergleiche auch Lemma 3.2.1. Die Konvergenzgeschwindigkeit wird somit nach Satz 3.4.3 maßgeblich von der Konstanten C bestimmt. Daher sei diese etwas genauer beleuchtet, wobei auch die eingangs erwähnte Abhängigkeit vom zu Grunde liegenden Problem deutlich wird.
Basis der Konvergenzanalyse bilden zwei Komponenten. Einerseits ist dies die den Glätter $\mathcal{S}$ betreffende *Glättungseigenschaft* (*smoothing property*) und andererseits die *Approximationseigenschaft* (*approximation property*) der Grobgitterkorrektur. Die *Glättungseigenschaft* beschreibt dabei die bereits mehrfach erwähnte Kontraktionseigenschaft des Glätters, also die Bedingung

$$\|\mathcal{S}\|_A = \|I - B^{-1}A\|_A \leq \gamma < 1,$$

wobei nach Lemma 3.3.1 für die ν-malige Anwendung $\|\mathcal{S}^\nu\|_A \leq \gamma^\nu < 1$ gilt.
Die *Approximationseigenschaft* kann mittels einer A-orthogonalen Projektion Q beschrieben werden.

Wie am Anfang des Abschnitts 3.2 in Gleichung (3.11) beschrieben, führen zwei Diskretisierungen des Gebiets mit unterschiedlichen Diskretisierungsparametern, beispielsweise h und $H = 2h$, zu Funktionenräumen W_h beziehungsweise W_H, die der Inklusionseigenschaft $W_H \subset W_h$ genügen. Für W_h gilt daher eine Zerlegung der Form

$$W_h = W_H \oplus W_H^{\perp}.$$

Betrachtet man nun die Grobgitterkorrektur des Zweigitterverfahrens aus Gleichung (3.14), gegeben durch $x_h^{(k+1)} = x_h^{(k)} + I_H^h e_H$ mit $e_H = A_H^{-1} r_H$, kann diese bezüglich des Fehlers $\hat{e}_h = x_h^{(k)} - x_h^* \in W_h$ als

$$Q\hat{e}_h = \hat{e}_h + I_H^h e_H \tag{3.39}$$

mit einem Orthogonalprojektor $Q : W_h \to W_H^{\perp}$ (da $I_H^h e_H \in W_H$) formuliert werden. Für die Norm dieses Projektors Q kann eine Abschätzung der Form

$$\|Q\|_A \leq \min\{1, c\sqrt{1-\gamma}\} \tag{3.40}$$

gezeigt werden, wobei die Konstante c eine Abhängigkeit von der Regularität der zu Grunde liegenden Differentialgleichung besitzt. Im Fall des Mehrgitterverfahrens, also der approximativen Lösung von $A_H e_H = r_H$ durch $\tilde{e}_H$ auf dem Grobgitter, genügt die Grobgitterkorrektur der Abschätzung

$$\|e_H - \tilde{e}_H\|_A \leq \varepsilon \|e_H\|_A$$

mit $\varepsilon > 0$. Unter zusätzlicher Betrachtung der Nachglättung liefert die Kombination aller Betrachtungen für die Norm der Operatoren aus Gleichung (3.37) eine Abschätzung der Form

$$\|\mathcal{T}_l\|_A \leq \max_{0 \leq \gamma \leq 1} \gamma^{2\nu} \left[\varepsilon + (1-\varepsilon)\min\{1, c(1-\gamma)\}\right].$$

Durch eine rekursive Analyse der einzelnen Normabschätzungen für $l = 0, \ldots, L$ kann dann der im Satz 3.4.3 angegebene Konvergenzfaktor hergeleitet werden.

Konvergenzrate mittels Teilraumkorrekturverfahren

Die folgende Konvergenzabschätzung geht auf Bramble, Pasciak und Xu, [8, 9, 10, 77], zurück. Es wurde folgende Abschätzung zur Konvergenz der Mehrgitterverfahren hergeleitet.

Satz 3.4.4. *Für die Energienorm des Mehrgitteriterationsoperators $\mathcal{T}_L$ gilt*

$$\|\mathcal{T}_L\|_A^2 \leq 1 - \frac{2-\omega}{C_1(1+C_2)^2} < 1. \tag{3.41}$$

Die dabei auftretenden Konstanten C_1, C_2 sowie ω sollen auch hier genauer beleuchtet werden.
Basis für die Herleitung der Abschätzung bildet die Betrachtung der Mehrgitterverfahren als Teilraumkorrekturverfahren (*successive subspace correction method*). Dazu sei S der Vektorraum, welcher von den Spalten von A des zu lösenden linearen Gleichungssystems aus Gleichung (3.4) aufgespannt wird. Dieser Raum ist durch eine Summe von (nicht notwendigerweise disjunkten) Räumen $\mathcal{W}_i$ in der Form

$$S = \mathcal{W}_1 + \mathcal{W}_2 + \ldots + \mathcal{W}_J \tag{3.42}$$

darstellbar. Im Kontext der (geometrischen) Mehrgitterverfahren sind diese $\mathcal{W}_i$ gerade die Teilräume, in denen die Grobgitterkorrekturen e_i berechnet werden und der Inklusion

$$\mathcal{W}_1 \subset \mathcal{W}_2 \subset \ldots \subset \mathcal{W}_J = S \tag{3.43}$$

genügen. Wesentlich sind weiterhin die Projektionen $P_i : S \to \mathcal{W}_i$ und $Q_i : S \to \mathcal{W}_i$, gegeben durch

$$(P_i x_i, y_i)_A = (x, y_i)_A, \quad x \in S, y_i \in \mathcal{W}_i \tag{3.44}$$

beziehungsweise

$$(Q_i x_i, y_i)_2 = (x, y_i)_2, \quad x \in S, y_i \in \mathcal{W}_i \tag{3.45}$$

und die Einschränkung von A auf den Teilraum $\mathcal{W}_i$ durch

$$(A_i x_i, y_i)_2 = (A x_i, y_i)_2, \quad x_i, y_i \in \mathcal{W}_i. \tag{3.46}$$

Diese Operatoren unterliegen der Beziehung

$$A_i P_i = Q_i A. \tag{3.47}$$

Betrachtet man nun die Berechnung der Korrekturen $e_i^{(k)}, (i = 1, \ldots, J)$, im jeweiligen Teilraum $\mathcal{W}_i$, ergibt dies die neue Iterierte als

$$x^{(k+1)} = x^{(k)} + \sum_{i=1}^{J} e_i^{(k)}. \tag{3.48}$$

Jede einzelne Korrektur $e_i^{(k)}$ berechnet sich dabei als Lösung der auf den Teilraum $\mathcal{W}_i$ eingeschränkten Gleichung

$$A_i e_i^{(k)} = r_i^{(k)}$$

mit $r_i^{(k)} = b_i - A_i x_i^{(k)}$ und darin $x_i^{(k)} = P_i x^{(k)}$ und $b_i = Q_i b$. Äquivalent dazu ist die Formulierung mittels der Projektion P_i, die für die Korrektur bezüglich des i-ten Teilraumes

$$\tilde{x}^{(k)} = x^{(k)} + P_i(x^* - x^{(k)}),$$

mit der exakten Lösung x^*, lautet. Eine Umformung dieser Gleichung liefert unter Nutzung von (3.47)

$$\tilde{x}^{(k)} = x^{(k)} + A_i^{-1} Q_i A(x^* - x^{(k)}) = x^{(k)} + A_i^{-1} Q_i (b - A x^{(k)}) \tag{3.49}$$

und damit eine von der unbekannten Lösung x^* unabhängigen Darstellung. Um darin die Lösung eines linearen Gleichungssystems zu vermeiden (A_i kann durchaus eine ähnliche Dimension wie A selbst besitzen), wird die Korrektur nur approximativ mittels

$$e_i^{(k)} = A_i^{-1} Q_i (b - A x^{(k)}) \approx B_i^{-1} Q_i (b - A x^{(k)}) = \tilde{e}_i^{(k)} \tag{3.50}$$

mit symmetrisch positiv definitem Vorkonditionierer B_i^{-1} berechnet. Zusammengefasst (für jeden Teilraum i) ergibt sich als Korrektur entsprechend Gleichung (3.48)

$$x^{(k+1)} = x^{(k)} + B_i^{-1} Q_i (b - A x^{(k)}), \quad (i = 1, \ldots, J). \tag{3.51}$$

Dabei betrachtet man hier ein multiplikatives Verfahren, wobei die Korrektur für den i-ten Teilraum unter Einfluss der bereits berechneten Korrekturen $j < i$ erfolgt. In Abhängigkeit von dem in Gleichung (3.50) angegebenen Vorkonditionierer können nun die Konstanten C_1, C_2 und ω charakterisiert werden. Aus analytischen Gründen geht man hierzu von einer Zerlegung von S in eine direkte Summe der Form

$$S = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \ldots \oplus \mathcal{V}_J$$

aus. Die Konstante $C_1 > 0$ ergibt sich dann aus einer Stabilität der Zerlegung, gegeben durch die Abschätzung

$$\sum_{l=0}^{L}(B_l v_l, v_l) \leq C_1 \| \sum_{l=0}^{L} v_l \|^2, \quad \forall v_l \in \mathcal{V}_l.$$

Die Konstante $C_2 > 0$ erhält man aus einer Ungleichung vom Cauchy-Schwarz-Typ. Dazu sei die Existenz von Konstanten ξ_{ij} mit

$$a(w_i, v_j) \leq \xi_{ij} \left(B_i w_i, w_i\right)^{1/2} \left(B_j v_j, v_j\right)^{1/2}, \quad \forall w_i \in \mathcal{W}_i, v_j \in \mathcal{V}_j,$$

für $i \leq j$ vorausgesetzt, so dass für alle $x_i, y_j \in \mathbb{R}$

$$\sum_{i,j=0}^{L} \xi_{ij} x_i y_j \leq C_2 \left(\sum_{i=0}^{L} x_i^2 \right)^{1/2} \left(\sum_{j=0}^{L} y_j^2 \right)^{1/2}$$

gilt. Implizit bedeutet dies, dass der Spektralradius der Matrix $\Xi = \xi_{ij}$ durch $C_2 > 0$ beschränkt ist.
Für die Konstante ω soll die Beziehung $0 < \omega < 2$ gelten, was die Konvergenz der Iteration (3.51) ausdrückt und äquivalent zur Beziehung

$$(A_i w_i, w_i) \leq \omega (B_i w_i, w_i), \quad \forall w_i \in \mathcal{W}_i, \ (i = 1, \ldots, J)$$

ist.

Damit sind zwei klassische Konvergenzabschätzungen zu den (geometrischen) Mehrgitterverfahren vorgestellt worden. Eine weitere Abschätzung wurde von McCormick hergeleitet, [48, 49, 50]. Die hier angegebenen und die letztgenannte Abschätzung sind in einer Arbeit von Napov und Notay verglichen worden, [54]. Dabei zeigt sich, dass qualitativ gleichwertige Schranken vorliegen. Quantitativ zeigt sich, dass (unter den dort gemachten Annahmen) die Schranke von McCormick die schärfste Abschätzung liefert.

Dies schließt die Betrachtungen zu den Mehrgitterverfahren. Im folgenden Kapitel sollen nun Methoden zur Lösung des Eigenwertproblems basierend auf den Mehrgitterverfahren erläutert und dargestellt werden. Den Schwerpunkt bildet dabei die Anwendung der algebraischen Mehrgitterverfahren zur Realisierung der in Abschnitt 2.3 beschriebenen Klasse von Eigenlösern des (k)-Schemas.

4. Mehrgitterverfahren für Eigenwertprobleme

Im vorangegangenen Kapitel lag der Fokus auf der Anwendung von Mehrgitterverfahren zur Lösung linearer Gleichungssysteme, die aus der Diskretisierung elliptischer Differentialoperatoren vom Typ (3.3) resultieren. Kern der Arbeit stellen jedoch Eigenlöser für das verallgemeinerte Eigenwertproblem

$$Au = \lambda Mu \quad A, M \in \mathbb{R}^{n\times n},\ u \in \mathbb{R}^n,\ \lambda \in \mathbb{R} \tag{4.1}$$

mit symmetrisch positiv def niten Matrizen A und M dar, vergleiche Abschnitt 1.1. Dabei stellt sich häuf g die Frage nach den kleinsten s Eigenpaaren (λ_i, u_i), $(i = 1, \ldots, s)$, wie sie beispielsweise in der Strukturmechanik die Grundschwingungen eines mechanischen Systems repräsentieren. Bevor die vorkonditionierte Iteration und deren Derivate aus Abschnitt 2.3 in den Vordergrund rücken, soll an dieser Stelle ein Überblick zu den wichtigsten Methoden, die Mehrgitterverfahren zur Lösung des Eigenwertproblems involvieren, gegeben werden. Diese lassen sich dabei wie folgt klassif zieren.

(I) **Direkte Mehrgitter-Eigenlöser**,

(II) **Rayleigh-Quotient-Minimierungsalgorithmen unter Verwendung von Mehrgitterverfahren**,

(III) **Eigenlöser basierend auf Vorkonditionierung mittels Mehrgitterverfahren**.

Die sich nun anschließenden Erörterungen dienen der Vorstellung der jeweils zu Grunde liegenden Idee der Verfahren sowie deren algorithmischer Umsetzung.

4.1. Direkte Mehrgitter-Eigenlöser

Ein exemplarischer Algorithmus für diese Verfahrensklasse ist der Eigenlöser vorgestellt durch Hackbusch, [29, 32]. Dabei wird die Idee der Mehrgitterverfahren für Randwertprobleme aus Kapitel 3 direkt auf die Lösung des Eigenwertproblems übertragen. Ein Unterschied liegt jedoch in der zu Grunde liegenden linearen Gleichung. Diese lautet im Fall des Eigenwertproblems ausgehend von einer Eigenvektorapproximation $u^{(j)}$ und ihrem zugehörigem Rayleigh-Quotienten $\lambda(u^{(j)})$ als Eigenwertapproximation,

$$(A - \lambda(u^{(j)})M)u^{(j)} = 0, \tag{4.2}$$

wodurch Stationarität in einem Eigenpaar sichergestellt ist. Die Gleichung (4.2) ist somit ein homogenes Gleichungssystem mit geshifteter, also um die Matrix $\lambda(u^{(j)})M$ verschobener, Systemmatrix A. Durch die auftretende Abhängigkeit von $\lambda(u^{(j)})$ variiert der Glättungsoperator $\mathcal{S}$ und sei im Folgenden durch

$$\mathcal{S}(\lambda) := \mathcal{S}(\lambda(u^{(j)})) = I - B^{-1}(A - \lambda(u^{(j)})M) \tag{4.3}$$

def niert. Der Vorkonditionierer B^{-1} ist somit eine Näherung an $A - \lambda(u^{(j)})M$. Eine weitere Modif kation im Vergleich zum Zweigitterverfahren für lineare Gleichungssysteme, welches hier vorerst betrachtet werden soll, motiviert sich aus der Beobachtung, dass mit Annäherung des Rayleigh-Quotienten an den gesuchten Eigenwert, die geshiftete Matrix $A - \lambda(u^{(j)})M$ zunehmend singulär wird. Daher schlägt Hackbusch die Lösung des Gleichungssystems (4.2) im orthogonalen Komplement des Eigenraumes, welcher

vom zugehörigen Eigenvektor u aufgespannt wird, vor. Da dieser jedoch gesucht und somit nicht bekannt ist, behilft man sich mit der Eigenvektorapproximation $u^{(j)}$ und def niert damit die (approximative) orthogonale Projektion für einen Vektor v_h durch

$$Q_h v_h := v_h - (v_h, u_h^{(j)})_2 M_h u_h^{(j)}.$$

Die Struktur des resultierenden Zweigitterverfahrens ist in Algorithmus 5 dargestellt. Es kennzeichnet wiederum h das feinere und H das gröbere Level.

Algorithmus 5 Zweigitterverfahren für das Eigenwertproblem

(1) $\lambda = \lambda(u_h^{(j)})$
(2) $\tilde{u}_h = \mathcal{S}_h(\lambda) u_h^{(j)}$
(3) $d_H = I_h^H (A_h - \lambda M_h) \tilde{u}_h$
(4) $d_H^\perp = Q_H d_H$
(5) $v_H = (A_H - \lambda M_H)^{-1} d_H^\perp$
(6) $u_h^{(k+1)} = \tilde{u}_h - I_H^h Q_H v_H$

Ein rekursiver Aufruf, und damit die Konstruktion eines Mehrgitterverfahrens, ist nach genauerer Betrachtung von Schritt 5 nicht möglich, da die Defektgleichung nicht die Gestalt eines Eigenwertproblems aufweist. Dafür besitzt diese die Struktur eines linearen Gleichungssystems und es kann daher ein Mehrgitterverfahren zur Lösung linearer Gleichungssysteme, wie im vorherigen Kapitel erläutert, angewendet werden. Dies kann durch die in Algorithmus 6 angegebene Rekursion namens *SingularMultiGridMethod (SMGM)* realisiert werden.

Algorithmus 6 SMGM(l, u, f)

(1) Falls $l = 0$:
(1.1) $u = Q_0 (A_0 - \lambda_0 I)^{-1} Q_0 f$
(2) sonst:
(2.1) $u = \mathcal{S}_l(u, f, \lambda_l)$
(2.2) $d = Q_{l-1} I_l^{l-1} (A_l u - \lambda_l M_l u - f)$
(2.3) $v = 0$; für $j = 1$ bis τ: SMGM$(l-1, v, d)$
(2.4) $u = Q_l (u - I_{l-1}^l v)$

Hier kennzeichnet l wiederum das Level und $\mathcal{S}_l(u, f, \tilde{\lambda}_l)$ einen Glätter für lineare Gleichungssysteme mit rechter Seite f. In Schritt 2.3 erfolgt dabei der rekursive Aufruf und führt auf das Mehrgitterverfahren *EigenvalueMultiGridMethod (EMGM)*, dargestellt in Algorithmus 7.

Algorithmus 7 EMGM(l, u)

(1) $\lambda = \lambda(u_l^{(j)})$
(2) $\tilde{u}_l = \mathcal{S}_l(\lambda) u_l^{(j)}$
(3) $d = Q_{l-1} I_l^{l-1} (A_l - \lambda M_l) \tilde{u}_l$
(4) $v = 0$; für $j = 1$ bis τ: SMGM$(l-1, v, d)$
(5) $u_l^{k+1} = \tilde{u} - I_{l-1}^l v$

Der Vollständigkeit halber sei erwähnt, dass Hackbusch zusätzlich die Verwendung einer geschachtelten Iteration (*nested iteration*) vorschlägt, um möglichst gute Eigenvektorapproximationen zur Konstruktion der Projektionen Q_l und Eigenwertapproximationen λ_l für den Algorithmus SMGM zu erhalten, [32].

4.2. RQMG

Der Rayleigh-Quotient-Minimierungsalgorithmus unter Verwendung von Mehrgitterverfahren (*RQMG*), vorgestellt von Mandel und McCormick in [46], basiert, ähnlich wie die vorkonditionierten Iterationen, auf der Minimierung (beziehungsweise Optimierung) des Rayleigh-Quotienten $\lambda(u)$ (aus Platzgründen sei im Folgenden auf den Iterationsindex (j) verzichtet). Dazu wird dieser als Energiefunktional betrachtet und ausgehend von einer Eigenvektorapproximation u und einer Korrekturrichtung d, ein optimaler Wert s^* mit

$$\lambda(u - s^*d) = \min_{s\in\mathbb{R}} \lambda(u - sd)$$

bestimmt und entsprechend u durch $u = u - s^*d$ ersetzt. Es werden also die Ritzapproximation im Unterraum aufgespannt von u und d berechnet. Damit existiert ein enger Zusammenhang zwischen RQMG und der vorkonditionierten Iteration PSD, vergleiche Abschnitt 2.3.
Innerhalb eines Mehrgitterzyklus wird auf jedem Level l zu geeigneten Richtungen d_i^l, $(i = 1, \ldots, n_i)$, die Minimierung vorgenommen. Mandel und McCormick schlagen in ihrem Zugang die kanonische Basis vor, das heißt, d_i^l ist der i-te Einheitsvektor auf dem l-ten Level. (Man beachte, dass aus Gründen der Indizierung die Levelzugehörigkeit hier im Exponenten auftaucht.) Der Algorithmus für einen resultierenden V-Zyklus ist in Algorithmus 8 angeführt.

Algorithmus 8 Rayleigh-Quotient-Minimierung nach Mandel und McCormick (RQMG)

(1) Für $i = 1 : n_0$
(1.1) Minimiere $\lambda(u + sd_i^0)$
(1.2) $u = u + s^*d_i^0$
(2) Für $l = 1 : L$
(2.1) Für $i = 1 : n_l$
(2.1.1) Minimiere $\lambda(u + sI_0^1 \cdots I_{l-1}^l d_i^l)$
(2.1.2) $u = u + s^* I_0^1 \cdots I_{l-1}^l d_i^l$
(3) Für $l = L : 1$
(3.1) Für $i = 1 : n_l$
(3.1.1) Minimiere $\lambda(u + sI_0^1 \cdots I_{l-1}^l d_i^l)$
(3.1.2) $u = u + s^* I_0^1 \cdots I_{l-1}^l d_i^l$
(4) Für $i = 1 : n_0$
(4.1) Minimiere $\lambda(u + sd_i^0)$
(4.2) $u = u + s^*d_i^0$

Im Wesentlichen erkennt man, dass in dieser Form des Algorithmus jede Minimierung auf dem feinsten Level $(l = 0)$ stattfindet, da alle Korrekturrichtungen d_i^l auf dieses Level prolongiert werden. Dies bedeutet demnach, dass ein V-Zyklus darin besteht, für die aus Vektoren bestehende Menge $\mathcal{Z}$, definiert durch

$$\mathcal{Z} = \{P_0^1 \cdots P_{l-1}^l d_i^l,\ l = 0, \ldots, L,\ i = 1, \ldots, n_l\},$$

eine Korrektur in jede Richtung $d \in \mathcal{Z}$ vorzunehmen. Allerdings ist dies mit hohem Rechenaufwand verbunden. Alternativ ist eine Implementierung, welche nur Größen des aktuellen Levels l enthält, möglich. Diese ist in Algorithmus 9 angegeben.

Algorithmus 9 Rayleigh-Quotient-Minimierung nach Mandel und McCormick (RQMG)

(1) Für $i = 1 : n_0$
(1.1) Minimiere $\lambda(u^0 + sd_i^0)$
(1.2) $u^0 = u^0 + s^* d_i^0$
(2) Setze $q^0 = 0$; $r^0 = 0$; $a = (Au^0, u^0)_2$; $b = (Mu^0, u^0)_2$
(3) Für $l = 1 : L$
(3.1) Setze $q^l = I_{l-1}^l(q^{l-1} + A^{l-1}u^{l-1})$; $r^l = I_{l-1}^l(r^{l-1} + M^{l-1}u^{l-1})$; $u^l = 0$
(3.2) Für $i = 1 : n_l$
(3.2.1) Berechne optimales s^* für

$$\lambda^l(u^l - sd_i^l) = \frac{a - 2s(q_i^l + \alpha_i^l) + s^2 A_{ii}^2}{b - 2s(r_i^l + \beta_i^l) + s^2 M_{ii}^2}$$

mit

$$q_i^l = (q^l, d_i^l)_2, \quad r_i^l = (r^l, d_i^l)_2,$$
$$A_{ii}^l = (A^l d_i^l, d_i^l)_2, \quad M_{ii}^l = (M^l d_i^l, d_i^l)_2,$$
$$\alpha_i^l = (A^l u^l, d_i^l)_2 \text{ und } \beta_i^l = (M^l u^l, d_i^l)_2$$

.
(3.2.2) $u^l = u^l + s^* d_i^l$, $a = a - 2s^*(q_i^l + \alpha_i^l) + s^{*^2} A_{ii}^2$ und $b = b - 2s^*(r_i^l + \beta_i^l) + s^{*^2} M_{ii}^2$
(4) Für $l = L - 1 : 1$
(4.1) $u^l = u^l + I_{l+1}^l u^{l+1}$; $u^{l+1} = 0$
(4.2) Für $i = 1 : n_l$
(4.3) Minimierung und Korrektur wie in den Schritten (3.2.1) und (3.2.2)
(5) $u^0 = u^0 + I_1^0 u^1$
(6) Für $i = 1 : n_0$
(6.1) Minimiere $\lambda(u^0 + sd_i^0)$
(6.2) $u^0 = u^0 + s^* d_i^0$

Die Berechnung der Größe s^* stellt sich dabei als Lösung einer quadratischen Gleichung heraus, denn für den Rayleigh-Quotienten gilt

$$\begin{aligned} \lambda(u - sd_i) &= \frac{((u - sd_i), A(u - sd_i))_2}{((u - sd_i), M(u - sd_i))_2} \\ &= \frac{(u, Au)_2 - 2s(u, Ad_i)_2 + s^2(d_i, Ad_i)_2}{(u, Mu)_2 - 2s(u, Md_i)_2 + s^2(d_i, Md_i)_2}. \end{aligned}$$

Die Differentiation von $\lambda(u - sd_i)$ bezüglich u ergibt somit ein quadratisches Polynom in s dessen eine Nullstelle das Minimum des Rayleigh-Quotienten repräsentiert. Eine Berechnung des optimalen Parameters s in Algorithmus 9 innerhalb der Schritte 3.2.1 beziehungsweise 4.3 ist daher mittels

$$s^* = \frac{-2y}{y + \sqrt{y^2 - 4xz}}$$

mit den Größen

$$\begin{aligned} x &= (d_i, Md_i)_2 \left((q, d)_2 + (u, Ad_i)_2\right) - (d_i, Ad_i)_2 \left((r, d)_2 + (u, Md_i)_2\right) \\ y &= (u, Mu)_2 (d, Ad) - (u, Au)_2 (d, Md)_2 \\ z &= (u, Au)_2 \left((r, d)_2 + (u, Md)_2\right) - (u, Mu)_2 \left((q, d)_2 + (u, Ad)_2\right) \end{aligned}$$

möglich. Mandel und McCormick entwickelten diesen Algorithmus unter Verwendung des geometrischen Mehrgitterverfahrens. Eine Variante bei der algebraische Mehrgitterverfahren zum Einsatz kommen ist beispielsweise in der Arbeit von Hetmaniuk, [36], zu f nden.

4.3. Vorkonditionierte Iterationen

Die dritte und letzte Klasse der angeführten Eigenlöser stellen die vorkonditionierten Iterationen dar, welche das Kernstück dieser Arbeit bilden und im Kapitel 2 bereits vorgestellt wurden. Ein wesentlicher Unterschied zu den im Vorfeld genannten Eigenlösern besteht hier im Einsatz des Mehrgitterverfahrens. Die bisher angeführten Algorithmen wiesen eine starke „Verzahnung“ zwischen Mehrgitterverfahren und Berechnung der Eigenwert- und Eigenvektorapproximationen auf. Bei den vorkonditionierten Iterationen f ndet sich im Gegensatz dazu eine andere Struktur.
Hier dienen die Mehrgitterverfahren ausschließlich der Berechnung der vorkonditionierten Residuen, vergleiche Kapitel 3. Auf die Verbesserung von Eigenwert- und Eigenvektorapproximationen nehmen sie keinen direkten Einf uss. Strukturell ergibt sich daher für diese Algorithmen eine Unterteilung in eine innere und äußere Iteration, dargestellt im Algorithmus 10.

Algorithmus 10 Struktur der vorkonditionierten Iteration

(1) Innere Iteration
(1.1) Berechne vorkonditionierte Residuen
(2) Äußere Iteration
(2.1) Berechne neue Eigenwert- und Eigenvektorapproximationen

Konkret beinhaltet Schritt 1.1, also die innere Iteration, die approximative Lösung des linearen Gleichungssystems

$$Ad^{(j)} = Au^{(j)} - \lambda(u^{(j)})Mu^{(j)} \tag{4.4}$$

mittels des Vorkonditionierers B^{-1}, vergleiche Gleichung (2.56), hier realisiert durch das algebraische Mehrgitterverfahren. Dabei kann die Anwendung des algebraischen Mehrgitterverfahrens entsprechend den Erläuterungen in Kapitel 3 erfolgen. Da diese Berechnung vom eigentlichen Eigenwertproblem getrennt erfolgt, können hier Mehrgitterverfahren zudem zur Realisierung der Abbildung $y \mapsto B^{-1}y$ im Sinne einer *black-box*-Operation verwendet werden.
Die Verbesserung der Eigenwert- und Eigenvektorapproximationen f ndet dann innerhalb der äußeren Iteration mittels der berechneten vorkonditionierten Residuen statt. Entsprechend den Erläuterungen in Abschnitt 2.4 entscheidet die Art dieser Neuberechnung darüber, welches (k)-Schema Anwendung f ndet. Es soll hierbei, wie eingangs erwähnt, eine simultane Berechnung der s kleinsten Eigenwertapproximationen mittels der ebenfalls in Abschnitt 2.4 vorgestellten Unterraumiterationen erfolgen. Dabei bildet das Rayleigh-Ritz-Verfahren, welches in Abschnitt 2.3 betrachtet wurde, die Basis der Neuberechnung von Eigenwert- und Eigenvektorapproximationen. Die Implementierung ist in Algorithmus 11 dargestellt, siehe auch [64].

Algorithmus 11 Rayleigh-Ritz-Verfahren $(RR(A, M, V))$

(1) ORTHO(V)
(2) Bilde $\tilde{A} = V^T AV$ und $\tilde{M} = V^T MV$
(3) Berechne Eigenpaare (Θ, Y) des Matrixpaars $(\tilde{A}, \tilde{M})$
(4) Sortiere (Θ, Y) in nicht fallender Reihenfolge

Man beachte, dass das hier angegebene Rayleigh-Ritz-Verfahren $RR(A, M, V)$ keine Ritzpaare, sondern nur die Ritzwerte, in Form der Matrix Θ, und die zugehörigen Koeff zientenvektoren in Form der Spalten von Y liefert. Die Berechnung der Ritzvektoren mittels Y wird in den späteren Algorithmen gesondert betrachtet. Die wesentliche Größe des Rayleigh-Ritz-Verfahrens ist die Matrix V. Diese bestimmt den Unterraum, in dem die Ritzapproximationen ermittelt werden. Je nach Wahl des verwendeten (k)-Schemas ist dieser durch

$$\mathcal{V}_{ks}^{(j)} = \text{span}\{V^{(j-k+2)}, \ldots, V^{(j-1)}, V^{(j)}, D^{(j)}\} \tag{4.5}$$

in Abhängigkeit von k und s gegeben. Die dort auftretenden Matrizen $V^{(m)} \in \mathbb{R}^{n\times s}$ beinhalten jeweils spaltenweise die s Eigenvektorapproximationen aus den Iterationsschritten $m = j-k+2, \ldots, j$ und haben demzufolge die Struktur

$$V^{(m)} = [u^{(m,1)}, \ldots, u^{(m,s)}].$$

Die in der inneren Iteration berechneten vorkonditionierten Residuen f ießen, wie zu erkennen ist, in Form der Matrix

$$D^{(j)} = [d^{(j,1)}, \ldots, d^{(j,s)}]$$

in das Rayleigh-Ritz-Verfahren ein. Eine Ausnahme bildet dabei das (1)-Schema. Hier ist der Unterraum durch $\mathcal{V}_s^{(j)} = \text{span}\{V^{(j)} - D^{(j)}\}$ gegeben. Schritt 1 des Algorithmus 11 führt die notwendige Orthonormalisierung der Spalten von V durch, so dass $V^TV = I$ gilt. Hierzu kann beispielsweise das modif zierte Gram-Schmidt-Verfahren eingesetzt werden. Die Berechnung der Eigenpaare in Schritt 3 als Lösung des projizierten Eigenwertproblems $\tilde{A}y = \theta\tilde{M}y$ kann beispielsweise mittels des QR-Verfahrens erfolgen, da dies ein niedrigdimensionales Problem der Dimension $s \ll n$ darstellt. Die Sortierung in Schritt 4 ist derart, dass $\theta_{ii} \le \theta_{i+1\,i+1}$, $(i = 1, \ldots, s-1)$, gilt. Entsprechend einer etwaigen Umsortierung der Ritzwerte werden gleichermaßen die Koeff zientenvektoren in Y umgeordnet.
Mit diesen Vorbetrachtungen ist es nun möglich, die Algorithmen der (k)-Schemata zur Berechnung der s kleinsten Eigenwerte zu formulieren. Ausgangspunkt sind das Matrixpaar (A, M) (mit symmetrisch positiv def niten Matrizen) sowie die jeweils benötigten Startapproximationen $u^{(1,1)}, \ldots, u^{(1,s)}$ spaltenweise zusammengefasst in der Matrix $V^{(1)}$ und deren zugehörige Rayleigh-Quotienten

$$\theta_i^{(1)} = \lambda(u^{(1,i)}) = \frac{(u^{(1,i)}, Au^{(1,i)})_2}{(u^{(1,i)}, Mu^{(1,i)})_2}, \quad (i = 1, \ldots, s),$$

welche die Diagonalelemente der Marix $\Theta^{(1)} = \text{diag}(\theta_1^{(1)}, \ldots, \theta_s^{(1)})$ bilden. Für $k = 1$ erhält man die Iteration PINVIT, dargestellt in Algorithmus 12.

Algorithmus 12 PINVIT ((1)-Schema)

Initialisierung:
(1) Wähle $\hat{V}^{(1)} \in \mathbb{R}^{n\times s}$; Berechne $(\Theta^{(1)}, Y^{(1)}) = \text{RR}(A, M, \hat{V}^{(1)})$ und $V^{(1)} = \hat{V}^{(1)}Y^{(1)}$
Iteration $(j = 1, 2, 3, \ldots)$ **:**
(2) Berechne $R^{(j)} = AV^{(j)} - MV^{(j)}\Theta^{(j)}$
(3) Löse $BD^{(j)} = R^{(j)}$
(4) Setze $Z^{(j)} = V^{(j)} - D^{(j)}$
(5) Berechne $(\Theta^{(j+1)}, Y) = \text{RR}(A, M, Z^{(j)})$
(6) Bestimme $V^{(j+1)} = Z^{(j)}Y$
(7) Markiere konvergierte Ritzpaare $(\theta_i^{(j+1)}, u^{(j+1,i)})$

Die in Algorithmus 10 beschriebene innere Iteration wird durch die Schritte PINVIT-2 und PINVIT-3 umgesetzt. Dabei wird die Lösung des Gleichungssystems durch die Anwendung des algebraischen Mehrgitterverfahrens realisiert. In Schritt PINVIT-4 f ndet die Korrektur der aktuellen Iterierten $V^{(j)}$ um die vorkonditionierten Residuen $D^{(j)}$ statt. Anschließend werden in PINVIT-5 die neuen Ritzwerte mittels des Rayleigh-Ritz-Verfahrens bestimmt. In Schritt PINVIT-6 erfolgt die Berechnung der zugehörigen Ritzvektoren aus der Koeff zientenmatrix Y, vergleiche Lemma 2.3.3. Man beachte, dass $Z^{(j)}$ für diese Berechnung orthonormale Spalten besitzen muss. Der Schritt PINVIT-7 dient der Prüfung auf bereits konvergierte Ritzapproximationen.
Der nächste Algorithmus, das (2)-Schema (PSD), ist in Algorithmus 13 skizziert.

Algorithmus 13 PSD ((2)-Schema)

Initialisierung:
(1) Wähle $\hat{V}^{(1)} \in \mathbb{R}^{n\times s}$; Berechne $(\Theta^{(1)}, Y^{(1)}) = \mathrm{RR}(A, M, \hat{V}^{(1)})$ und $V^{(1)} = \hat{V}^{(1)}Y^{(1)}$
Iteration $(j = 1, 2, 3, \ldots)$ **:**
(2) Berechne $R^{(j)} = AV^{(j)} - MV^{(j)}\Theta^{(j)}$
(3) Löse $BD^{(j)} = R^{(j)}$
(4) Setze $Z^{(j)} = [V^{(j)}, D^{(j)}]$
(5) Berechne $(\Theta^{(j+1)}, Y) = RR(A, M, Z^{(j)})$
(6) Bestimme $V^{(j+1)} = Z^{(j)}Y^1$
(7) Markiere konvergierte Ritzpaare $(\theta_i^{(j+1)}, u^{(j+1,i)})$

Wie eingangs erwähnt, unterscheidet sich der Algorithmus PSD vom Algorithmus PNIVIT im Wesentlichen nur durch den Unterraum, auf den das Rayleigh-Ritz-Verfahren angewendet wird. Zudem wird deutlich, warum die Berechnung der Ritzvektoren dort ausgeklammert wurde. Um die Anzahl der Rechenoperationen zu reduzieren, werden in PSD-6 nicht alle Ritzvektoren bestimmt, sondern nur die, welche zu den geforderten s kleinsten Ritzwerten gehören. Dazu besteht die angegebene Matrix Y^1 gerade aus den ersten s Spalten von Y. Schematisch weist daher die Matrix Y, bei der Anwendung eines (k)-Schemas, die Gestalt

$$Y = [Y^1, Y^2, \ldots, Y^k] \quad \text{mit} \quad Y^p \in \mathbb{R}^{n\times s},\ (p = 1, \ldots, s), \tag{4.6}$$

auf. Abschließend wird auch hier die Prüfung auf Konvergenz in Schritt PSD-7 vorgenommen.
Als letzter in expliziter Form angegebener Algorithmus soll nun das (3)-Schema (LOBPCG) betrachtet werden, vergleiche Algorithmus 14.

Algorithmus 14 LOBPCG ((3)-Schema)

Initialisierung:
(1) Wähle $\hat{V}^{(1)} \in \mathbb{R}^{n\times s}$; Berechne $(\Theta^{(1)}, Y^{(1)}) = \mathrm{RR}(A, M, \hat{V}^{(1)})$ und $V^{(1)} = \hat{V}^{(1)}Y^{(1)}$
Iteration $(j = 1, 2, 3, \ldots)$ **:**
(2) Berechne $R^{(j)} = AV^{(j)} - MV^{(j)}\Theta^{(j)}$
(3) Löse $BD^{(j)} = R^{(j)}$
(4.1) Falls $j = 1$: Setze $Z^{(j)} = [V^{(j)}, D^{(j)}]$
(4.2) sonst: Setze $Z^{(j)} = [V^{(j)}, D^{(j)}, V^{(j-1)}]$
(5) Berechne $(\Theta^{(j+1)}, Y) = RR(A, M, Z^{(j)})$
(6.1) Falls $j = 1$: Berechne $V^{(j+1)} = Z^{(j)}Y^1$
(6.2) sonst: Berechne $V^{(j+1)} = Z^{(j)}Y^1$ und $V^{(j)} = [0, D^{(j)}, V^{(j-1)}]Y^1$
(6) Markiere konvergierte Ritzpaare $(\theta_i^{(j+1)}, u^{(j+1,i)})$

Der Algorithmus LOBPCG unterscheidet sich von den vorher genannten in zwei Punkten. Zum einen erkennt man, dass für die Durchführung des ersten Iterationsschrittes ($j = 1$) noch keine vorherigen Iterierten $V^{(j-1)} = V^{(0)}$ zur Verfügung stehen. Daher werden in den Schritten LOBPCG-4.1 und LOBPCG-6.1 die Berechnung der Folgeiterierten entsprechend der des Algorithmus PSD angewendet. Zum anderen werden in Schritt LOBPCG-6.2 nicht nur die Iterierten $V^{(j+1)}$ berechnet, sondern ebenso die Iterierten $V^{(j)}$ neu bestimmt. Der Grund hierfür liegt in einer Stabilisierung des Rayleigh-Ritz-Verfahrens denn mit zunehmender Konvergenz nimmt auch die Kollinearität der Ritzvektoren aus zwei aufeinander folgenden Iterationschritten, also die Kollinearität der Spalten der Matrizen $V^{(j)}$ und $V^{(j+1)}$, zu, wobei die angegebene Neuberechnung diesem Effekt entgegenwirkt. Trotzdem bleibt der Unterraum $\mathcal{V}_{3s}$ derselbe, vergleiche [39].
Die (k)-Schemata höherer Ordnung ($k \geq 4$) sollen hier nicht ausführlicher dargestellt werden. Sie resultieren auf natürliche Weise aus den im Vorfeld vorgestellten Algorithmen durch eine entsprechende Anpassung der auftretenden Matrix Z. Zu beachten ist jeweils, dass, wie bereits bei LOBPCG erwähnt, die Konstruktion des Unterraumes $\mathcal{V}_{ks}^{(j)}$ für $j < k$ zu Beginn der Iteration die Anwendung eines passenden Schemas erfordert.

In den folgenden Kapiteln sollen die hier vorgestellten Algorithmen an mehreren Problemstellungen getestet werden. Abgesehen von den numerischen Ergebnissen wird der Blick ebenso auf Aspekte der Konvergenz gerichtet.

5. Iterative Eigenlöser mit algebraischer Mehrgittervorkonditionierung

Nachdem die vorangegangenen Kapitel die ausführliche Betrachtung der theoretischen Grundlagen zu den vorkonditionierten Eigenlösern und algebraischen Mehrgitterverfahren beinhalteten, soll in diesem Kapitel die numerischen Anwendung der in Kapitel 4 vorgestellten vorkonditionierten Eigenlöser erfolgen. Bei den Untersuchungen steht dabei die Leistungsfähigkeit der einzelnen Algorithmen im Vordergrund. Um diese darzustellen, werden die vorgestellten Algorithmen PINVIT, PSD, LOBPCG und ebenso Verfahren höherer Ordnung auf verschiedene Problemstellungen angewendet. Speziell für PINVIT erfolgt dazu eine numerische Konvergenzanalyse. Weiterhin wird mit Blick auf das Mehrgitterverfahren zusätzlich das Augenmerk auf die Verwendung zweier Vergröberungsstrategien, der Methode des *standard coarsening* und des *smoothed aggregation*, erläutert in Abschnitt 3.3.2, gerichtet werden. Entsprechend den Ausführungen in der Einleitung werden zudem ein Differentialoperator mit variierenden Koeff zienten (anisotropes Problem) und geometriefreie Probleme untersucht. Obwohl die algebraischen Mehrgitterverfahren nicht von einer zu Grunde liegenden Geometrie abhängig sind, sollen dennoch, wenn möglich, graphische Interpretationen der betrachteten Probleme und ihrer Lösung angebracht werden.
Die Implementierung betreffend seien folgende, teilweise bereits erwähnten, Fakten festgehalten. Die Konstruktion der jeweiligen Steif gkeitsmatrizen A und zugehöriger Massematrix M in den folgenden Untersuchungen werden durch das Programm *FreeFem++*, [35], realisiert. Die Implementierung der vorgestellten Algorithmen wurde mittels Matlab, [47], vorgenommen. Weiterhin f ndet die Umsetzung des *standard coarsening* mittels einer auf der Algebraic MultiGrid-Toolbox, hier kurz *AMT*, [74], basierender und angepasster Implementierung statt. Die Methode *smoothed aggregation* wird durch das *Multilevel Preconditioning Package* (*ML-Package*) von *Sandia National Laboratories*, welches über eine Matlab-Schnittstelle verfügt, realisiert, [24].

5.1. Modellproblem I

Im folgenden Abschnitt steht vorerst die inverse Vektoriteration PINVIT, also das (1)-Schema, unter algebraischer Mehrgittervorkonditionierung im Mittelpunkt. Dabei sollen die Konvergenzraten des algebraischen Mehrgitterverfahrens (innere Iteration) und der vorkonditionierten Iteration (äußere Iteration) numerisch anhand eines ersten Modellproblems getestet werden. Hierbei ist es Aufgabe, den kleinsten Eigenwert $(s = 1)$ zu berechnen.
Als Modellproblem dient das Eigenwertproblem für den Laplace-Operator auf dem Einheitsquadrat mit Dirichlet-Randbedingungen. Das zu Grunde liegende kontinuierliche Operatoreigenwertproblem lautet somit

$$\begin{aligned} -\Delta u(x,y) &= \lambda u(x,y) \qquad \Omega = [0,1]\times[0,1] \\ u(x,y) &= 0 \qquad\qquad u(x,y) \in \partial\Omega. \end{aligned} \tag{5.1}$$

Die analytischen Lösungen können in diesem Fall explizit angegeben werden. Dabei sind die Eigenpaare $(\lambda_{kl}, u_{kl}(x,y))$ für $k,l = 1,2,\ldots$ durch die Eigenwerte

$$\lambda_{kl} = \pi^2(k^2 + l^2)$$

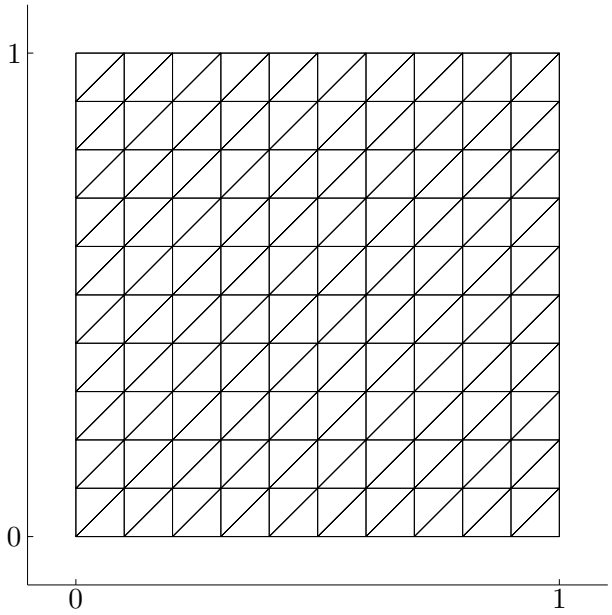

Abbildung 5.1.: Exemplarische Standardtriangulierung des Gebietes $\Omega = [0,1] \times [0,1]$ für den Diskretisierungsparameter $h = 0.1$ $(n = 81)$.

und die zugehörigen Eigenfunktionen durch

$$u_{kl}(x,y) = \sin(k\pi x)\sin(l\pi y)$$

gegeben.
Zur numerischen Behandlung der Operatorgleichung (5.1) wird entsprechend den Ausführungen innerhalb der Einleitung im Abschnitt 1.1 das Gebiet Ω diskretisiert und mittels der Methode der finiten Elemente in das verallgemeinerte Matrixeigenwertproblem

$$Au = \lambda Mu \quad A, M \in \mathbb{R}^{n\times n}, u \in \mathbb{R}^n, \lambda \in \mathbb{R}, \tag{5.2}$$

also dem Eigenwertproblem für das Matrixpaar (A, M) überführt.
Die folgenden Untersuchungen basieren auf einer äquidistanten Diskretisierung des Gebietes Ω bezüglich beider Raumkoordinaten, beispielhaft angedeutet in Abbildung 5.1. In Abhängigkeit vom Diskretisierungsparameter h ergibt sich $n = (1-h)^2/h^2$ und man erhält auf diese Weise die Menge der (inneren) Stützstellen $\mathcal{N}_h = \{x_i : i = 1,\ldots,n\}$. Die Test- und Ansatzfunktionen ϕ_i sind als stückweise lineare Funktionen der Art

$$\phi_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{5.3}$$

gewählt. Damit ergibt sich $S = \text{span}\{\phi_1,\ldots,\phi_n\}$ mit $\dim(S) = n$, siehe Gleichung (1.12).
Bei einer lexikographischen Nummerierung der Stützstellen ist die Steifigkeitsmatrix A eine Blockmatrix, gegeben durch $A = \text{tridiag}(-I, A_D, -I)$ mit Blöcken $A_D = \text{tridiag}(-1,4,-1)$ und der Einheitsmatrix I mit passender Dimension. Die Massematrix M hat dann die Blockgestalt $M = \frac{h^2}{12} \cdot \text{tridiag}(I_M^T, M_D, I_M)$ mit Blöcken $M_D = \text{tridiag}(1,6,1)$ und $I_M = \text{tridiag}(0,1,1)$. Für kleine Parameter h und damit großem n sind A und M somit symmetrische, dünnbesetzte und auf Grund der Elliptizität

des Operators in Gleichung (5.1) positiv def nite Matrizen.
Mithilfe des so formulierten diskreten Modellproblems (5.2) sollen in den folgenden Betrachtungen die Konvergenzraten numerisch untersucht werden. Dazu wird einerseits die Konvergenzrate des Mehrgitterverfahrens mittels des Mehrgitteriterationsoperators $\mathcal{T}_L$ aus den Erörterungen des Abschnitts 3.4 (Satz 3.4.1) und andererseits der Konvergenzfaktor der vorkonditionierten Iteration aus Abschnitt 2.2 (Satz 2.2.33) berechnet.

Konvergenzrate des algebraischen Mehrgitterverfahrens

Gegenstand dieses Abschnitts ist eine Betrachtung der Konvergenz der algebraischen Mehrgitterverfahren zur Berechnung der vorkonditionierten Residuen aus Gleichung (3.1) für das oben angegebene Eigenwertproblem (5.2), also einer Betrachtung des Konvergenzverhaltens der inneren Iteration aus Algorithmus 10. Wie bereits in Abschnitt 2.1 beziehungsweise Kapitel 3 herausgestellt, ist die wesentliche Größe hierbei die Konstante γ aus der Abschätzung für die Fehlerfortpf anzungsmatrix

$$\|I - B^{-1}A\|_A \leq \gamma < 1. \tag{5.4}$$

Entsprechend Satz 3.4.1 kann die Konstante aus dem Mehrgitteriterationsoperator $\mathcal{T}_L$ gewonnen werden. Der numerischen Bestimmung der Konstanten γ liegt dazu folgende Situation zu Grunde. Es wird eine Hierarchie mittels des in Abschnitt 3.3.2 beschriebenen *standard coarsening* konstruiert. Zur Berechnung der Konstanten γ werden dann zwei Ansätze gewählt. Einerseits wird der Spektralradius des Mehrgitteriterationsoperator, folgend gekennzeichnet mit $\varrho(\mathcal{T}_L^{\text{expl}})$, numerisch aus seiner expliziten Konstruktion berechnet. Andererseits wird γ als empirischer Konvergenzfaktor $\varrho(\mathcal{T}_L^{\text{num}})$ durch Anwendung des algebraischen Mehrgitterverfahrens für 1000 zufällige Startvektoren $u^{(0)}$ zur Lösung des homogenen linearen Gleichungssystems $Au = 0$ mit A aus (5.2) bestimmt. Die Berechnung des empirischen Faktors ergibt sich dann aus, vergleiche auch [71],

$$\varrho(\mathcal{T}_L^{\text{num}}) = \left(\frac{\|r^{(p)}\|_2}{\|r^{(m)}\|_2} \right)^{1/(p-m)} . \tag{5.5}$$

Hierbei beschreibt $r^{(i)} = -Au^{(i)}$ das Residuum im i-ten Iterationsschritt und p gibt die Anzahl der Iterationsschritte zur (approximativen) Lösung des homogenen Gleichungssystems an. Dabei wird $u^{(p)}$ als Lösung betrachtet, falls das zugehörige Residuum der Bedingung

$$\|Au^{(p)}\|_2 = \|r^{(p)}\|_2 \leq 10^{-14}$$

genügt. Weiterhin beschreibt $m \geq 1$ einen geeigneten Iterationsindex, dessen Wahl an späterer Stelle konkretisiert wird.
Die wesentlichen Parameter zur Umsetzung des algebraischen Mehrgitterverfahrens mittels *AMT* lauten

- Glätter: Gauß-Seidel-Verfahren
- $\epsilon = 0.25$ (zur Def nition der starken negativen Verbindungen)
- $n_0 \leq 10$ (maximale Anzahl der auf dem gröbstem Level verbleibenden Punkte)
- V$(2,2)$-Zyklus ($\tau = 1$, $\nu_1 = \nu_2 = 2$).

Eine hiermit konstruierte Hierarchie ist in Abbildung 5.2 für den Diskretisierungsparameter $h = 0.05$ dargestellt. Dabei sind neben dem feinsten Level $l = 4$ (der angedeuteten Triangulierung) mit $n_4 = 361$ inneren Stützstellen ebenso die Level $l = 3$ (durch × gekennzeichnet) mit $n_3 = 181$, $l = 2$ (o) mit $n_2 = 54$,

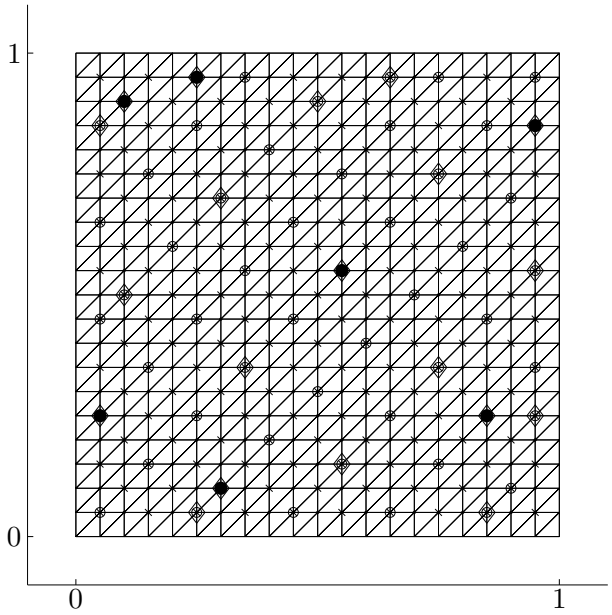

Abbildung 5.2.: Resultierende Hierarchie für das Modellproblem ($h = 0.05$). • kennzeichnet verbleibende Punkte auf dem gröbsten Level.

Level l	n_i	$\varrho(\mathcal{S}_i)$	$\varrho(\mathcal{T}_i^{\text{expl}})$
0	7	(*)	0
1	20	0.448	0.0031
2	54	0,783	0.0280
3	181	0.939	0.0501
4	361	0.976	0.1092

Tabelle 5.1.: Die zur Berechnung von $\varrho(\mathcal{T}_L^{\text{expl}})$ auftretenden Größen ($h = 0.05$).

$l = 1$ ($\diamond$) mit $n_1 = 20$ sowie das Level $l = 0$ ($\bullet$) mit $n_0 = 7$ Stützstellen abgebildet. Zu Anschauungszwecken soll sich zunächst auf niedrigdimensionale Probleme beschränkt werden. Die Anwendung der Eigenlöser auf hochdimensionale Eigenwertprobleme folgt dann in den sich anschließenden Abschnitten 5.2 und 5.3.
Die ersten Untersuchungen widmen sich der Berechnung von $\varrho(\mathcal{T}_L^{\text{expl}})$. Die Tabelle 5.1 stellt die dabei auftretenden Größen für die in Abbildung 5.2 dargestellte Diskretisierung zum Parameter $h = 0.05$ dar. Die erste und zweite Spalte zeigt die Levelnummer und zugehörige Problemdimension. Die letzten Spalten geben Auskunft über die Spektralradien des levelspezif schen Glättungsoperators beziehungsweise der intermediär auftretenden Matrizen $\mathcal{T}_l$, ($l = 0, 1, 2, 3$), welche zur rekursiven Berechnung von $\mathcal{T}_L^{\text{expl}}$ benötigt werden, siehe Satz 3.4.1. Die angegebenen Spektralradien werden dabei mittels *ARPACK* (in Form der Matlab-Routine *eig*) ermittelt. Der auf diese Weise berechnete Konvergenzfaktor beträgt dementsprechend

$$\varrho(\mathcal{T}_L^{\text{expl}}) = \gamma \approx 0.1092. \tag{5.6}$$

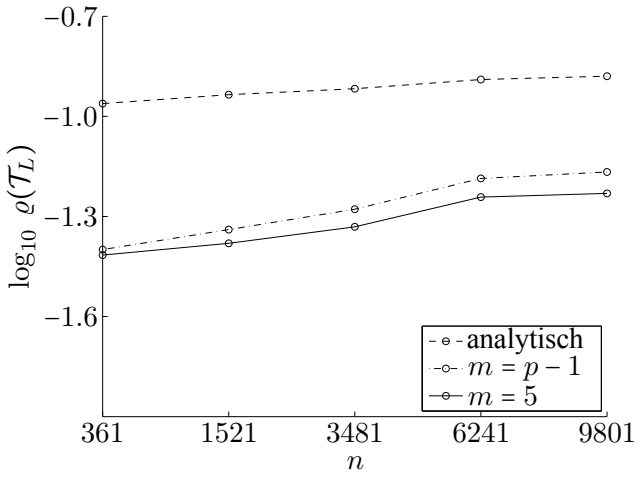

Abbildung 5.3.: Dekadischer Logarithmus der Spektralradien der empirisch und analytisch bestimmten Fehlerfortpf anzungsmatrizen.

h	n_L	L	n_L	$\varrho(\mathcal{T}_L^{\text{expl}})$	$\varrho(\mathcal{T}_L^{\text{num}})$	
					$m = 5$	$m = p - 1$
0.0500	361	4	7	0.1092	0.0384	0.0399
0.0250	1521	5	8	0.1161	0.0417	0.0458
0.0167	3481	6	6	0.1210	0.0467	0.0527
0.0125	6241	7	3	0.1289	0.0573	0.0651
0.0100	9801	7	2	0.1320	0.0588	0.0681

Tabelle 5.2.: Konvergenzraten zu unterschiedlichen Diskretisierungen.

In Anlehnung an die im Abschnitt 3.1 dargelegten nachlassenden Konvergenzeigenschaften der klassischen Iterationen hinsichtlich einer feineren Diskretisierung ($h \to 0$), soll auch das algebraische Mehrgitterverfahren daraufhin untersucht werden. Tabelle 5.2 beziehungsweise Abbildung 5.3 geben dazu Auskunft über das Konvergenzverhalten in Abhängigkeit vom Diskretisierungsparameter h. Die ersten Spalten zeigen die dazu gewählten Parameter h und die daraus resultierende Anzahl innerer Stützstellen n_L. Zusätzlich ist die Levelanzahl L und die Anzahl der verbleibenden Punkte auf dem gröbsten Gitter angegeben. Die fünfte Spalte gibt die Werte für γ unter expliziter Berechnung der Matrix $\mathcal{T}_L$ an. In den letzten Spalte stehen die aus Gleichung (5.5) erhaltenen numerischen Konvergenzraten. Dabei ist einerseits der asymptotische Konvergenzfaktor für $m = 5$ und andererseits für $m = p - 1$, also die Konvergenzrate im letzten Iterationsschritt, angegeben. Beide wurden jeweils als arithmetisches Mittel der einzelnen Konvergenzraten zu 1000 zufällig gewählten Startvektoren $u^{(0)}$ erhalten. Bei einer genaueren Betrachtung fällt auf, dass die Abschätzungen für $\varrho(\mathcal{T}_L^{\text{expl}})$ im Vergleich zu $\varrho(\mathcal{T}_L^{\text{num}})$ eine deutlich pessimistischere Konvergenz vorhersagen, die erstgenannten aber auch eine obere Abschätzung darstellen. Trotzdem erkennt man in beiden Fällen einen im Verhältnis zur steigenden Problemdimension nur moderat wachsenden Konvergenzfaktor. Dies unterstreicht die im Abschnitt 3.1 erwähnte drastische Konvergenzbeschleunigung durch den Einsatz von Mehrgitterverfahren verglichen mit der Verwendung eines klassischen Iterationsverfahrens.

i	1	2	3	4	5
λ_i	19.8611	49.8717	50.1680	80.8931	101.1

Tabelle 5.3.: Kleinste Eigenwerte des Matrixpaares (A, M) für $h = 0.05$.

Konvergenzrate der vorkonditionierten Iteration PINVIT

An dieser Stelle soll eine kurze Verif kation der in Kapitel 2 hergeleiteten Konvergenzabschätzung aus Satz 2.2.33 erfolgen. Diese lautet

$$\frac{\lambda(\hat{u}) - \lambda_i}{\lambda_{i+1} - \lambda(\hat{u})} \leq \sigma^2 \frac{\lambda(u) - \lambda_i}{\lambda_{i+1} - \lambda(u)} \tag{5.7}$$

mit dem Konvergenzfaktor

$$\sigma = 1 - (1-\gamma)\frac{\lambda_{i+1} - \lambda_i}{\lambda_{i+1}} = \gamma + (1-\gamma)\frac{\lambda_i}{\lambda_{i+1}}. \tag{5.8}$$

Wiederum soll die Basis eine Diskretisierung mit Parameter $h = 0.05$ bilden. Damit sei die Konstante γ, als Spektralradius der Fehlerfortpf anzungsmatrix $\mathcal{T}_L^{\text{expl}}$, durch die Berechnungen im vorangegangenen Abschnitt gegeben, siehe Gleichung (5.6) beziehungsweise Tabelle 5.1, das heißt

$$\gamma \approx 0.1092.$$

Die (numerisch ermittelten) kleinsten Eigenwerte des Matrixpaares (A, M) sind in Tabelle 5.3 angegeben. Daher berechnet sich für das Quadrat des Konvergenzfaktors σ aus Gleichung (5.8) für $i = 1$ aus quasianalytischer Sicht, da sowohl γ als auch λ_1 und λ_2 approximative Größen sind, als

$$\sigma^2 \approx \left(0.1092 + (1 - 0.1092)\frac{19.8611}{49.8717}\right)^2 = 0.2153. \tag{5.9}$$

Der numerische Test von σ^2 beruht auf der Berechnung empirischer Konstanten $\tilde{\sigma}_{ij}^2$, die ausgehend von $i = 1, \ldots, 1000$ zufällig gewählten Startvektoren $u^{(0)}$ ermittelt werden. Falls der Rayleigh-Quotient der im Iterationsverlauf erhaltenen Iterierten $u^{(j)}$ im i-ten Durchlauf die Bedingung $\lambda_1 < \lambda(u^{(j)}) < \lambda_2$ erfüllt, wird ein zugehöriges $\tilde{\sigma}_{ij}^2$ entsprechend Gleichung (5.7) ermittelt. Das schlechtmöglichste Konvergenzverhalten wird dabei durch das größte auftretende $\tilde{\sigma}_{ij}^2$ verkörpert. Unter der Verwendung der algebraischen Mehrgittervorkonditionierung mit den im Vorfeld angegebenen Parametern ergibt sich aus den Berechnungen

$$\bar{\sigma}^2 = \max_i \max_j \tilde{\sigma}_{ij}^2 \leq 0.1768 < 0.2153 = \sigma^2.$$

Damit liegt $\bar{\sigma}^2$, wie erwartet, unter der theoretischen Schranke aus Gleichung (5.9).
Selbst für die Wahl $\gamma = 0.0384$, als Ergebnis der empirischen Abschätzungen für $\varrho(\mathcal{T}_L^{\text{num}})$, vergleiche Tabelle 5.2, bleibt für den nach Gleichung (5.8) berechneten Konvergenzfaktor σ^2 die Konvergenzabschätzung gültig, denn es gilt für die zugehörige Konstante $0.1768 = \bar{\sigma}^2 < \sigma^2 = 0.1775$. Dies verif ziert, zumindest im betrachteten Szenario, die angegebene Konvergenzaussage des Satzes 2.2.33.

Numerische Lösung des Modellproblems

Abschließend sollen die Ergebnisse der numerischen Berechnungen mittels PINVIT für das unter Gleichung (5.2) gegebene Eigenwertproblem angegeben werden. Dabei wird das Ritzpaar $(\theta^{(j)}, u^{(j)})$ als

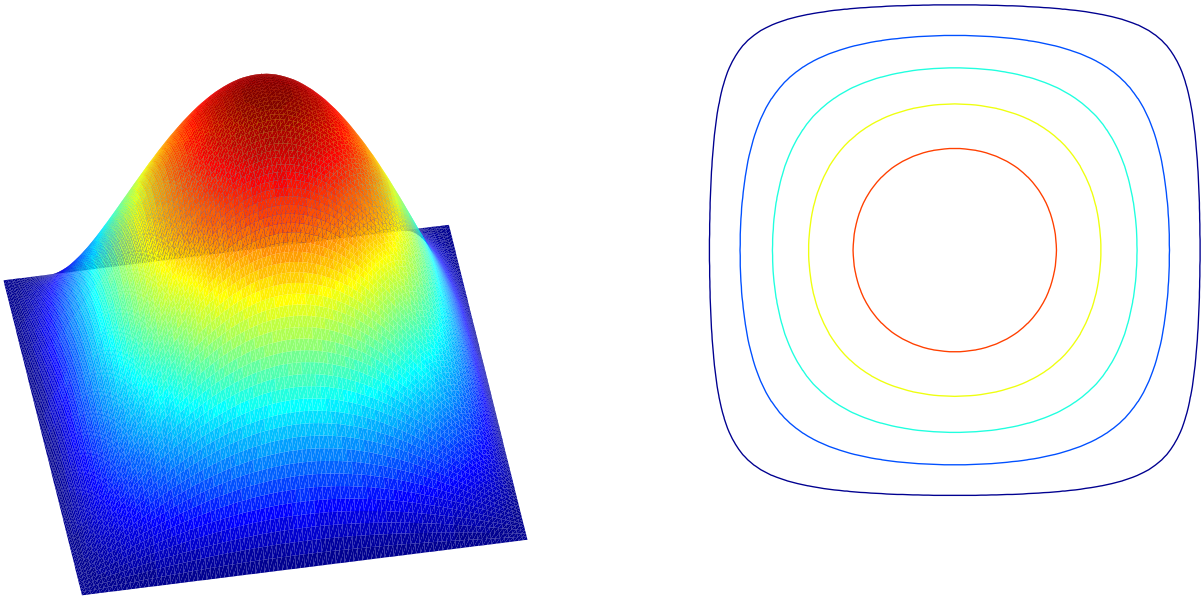

Abbildung 5.4.: Lösung zum Modellproblem I. Links: Eigenvektor u_1 zum kleinsten Eigenwert $\lambda_1 = 19.8611$. Rechts: Höhenliniendarstellung von u_1.

stationär betrachtet, falls die Eigenvektorapproximation die Bedingung

$$\text{err} = \|Au^{(j)} - \theta^{(j)} M u^{(j)}\| \leq 10^{-10}$$

erfüllt. Auf der linken Seite der Abbildung 5.4 ist eine dreidimensionale Darstellung des Eigenvektors u_1 zugehörig zur berechneten Eigenwertapproximation $\lambda_1 = 19.8611$ dargestellt. Auf der rechten Seite ist die zugehörige Höhenliniendarstellung abgebildet. Bei der Berechnung des vorkonditionierten Residuums $d^{(j)}$ kam hierbei ein einzelner $\mathrm{V}(2,2)$-Zyklus, also ein Schritt der inneren Iteration aus Algorithmus 10, zur Anwendung. Zum Erhalt der oben angegebenen Lösung sind insgesamt $N_{iter} = 28$ Iterationsschritte der vorkonditionierten Iteration PINVIT notwendig.
Die Anwendung der vorkonditionierten Iteration unter Verwendung vorkonditionierter Residuen, welche die Bedingung

$$\|r^{(j)} - Ad^{(j)}\|_2 \leq 10^{-14}$$

erfüllen, also die Umsetzung der inversen Vektoriteration (vergleiche Bemerkung 2.1.1), führt auf $N_{iter} = 27$ Schritte. Diese nur leicht verbesserte Konvergenzgeschwindigkeit steht allerdings nicht im Verhältnis zur rund fünfmal höheren Rechenzeit. Dies unterstreicht deutlich, dass eine rein approximative Berechnung der vorkonditionierten Residuen $d^{(j)}$ mittels eines einzigen V-Zyklus durchaus ausreichend ist.

Diese Untersuchungen zeigen, dass die vorkonditionierte Iteration PINVIT mit algebraischer Mehrgittervorkonditionierung zur Lösung dieses verallgemeinerten Eigenwertproblems geeignet ist und die im Vorfeld gezeigten Konvergenzeigenschaften am Beispiel verif ziert werden können. Die weitergehenden Berechnungen in den folgenden Abschnitten sollen nun unterstreichen, dass diese Anwendbarkeit nicht auf elementare und niedrigdimensionale Probleme mit einfacher Geometrie beschränkt ist.

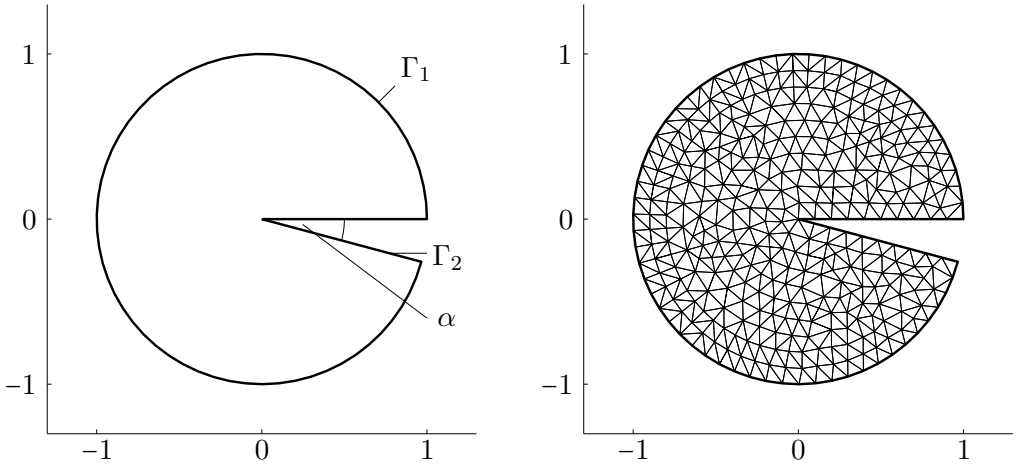

Abbildung 5.5.: Modellproblem II. Links: Gebiet Ω aus Gleichung (5.11). Rechts: Exemplarische Triangulierung mit $n = 363$ Stützstellen.

5.2. Modellproblem II - Unterraumiterationen

In Anlehnung an das in der Einführung unter Gleichung (1.2) formulierte Operatoreigenwertproblem sollen die Algorithmen auf ein weiteres zweidimensionales Testproblem def niert durch

$$\begin{aligned} -\nabla \cdot (\varepsilon_{x,y} \nabla u) &= \lambda u \\ u &= 0 \quad \text{auf } \Gamma_1 \\ n_{x,y} \cdot \varepsilon_{x,y} \nabla u &= 0 \quad \text{auf } \Gamma_2 \end{aligned} \tag{5.10}$$

angewendet werden, [42, 43]. Das Gebiet Ω sei (in Polarkoordinaten) durch

$$\Omega = \{(t, \varphi);\ t \in [0,1], \varphi \in [0, 2\pi - \alpha]\} \tag{5.11}$$

mit $\alpha > 0$ gegeben. Es stellt somit einen geöffneten Kreis dar und ist in Abbildung 5.5 für $\alpha = \frac{\pi}{12}$ dargestellt. Der Rand des Gebietes kann in Abhängigkeit vom Öffnungswinkel α durch

$$\begin{aligned} \Gamma_1 &= \{(t,\varphi) : t \in [0,1], \varphi = 0\} \text{ und } \{t = 1, \varphi \in [0, 2\pi - \alpha]\} \\ \Gamma_2 &= \{(t,\varphi) : t \in [0,1], \varphi = 2\pi - \alpha\} \end{aligned}$$

beschrieben werden. Für die folgenden Betrachtungen sei vorerst $\varepsilon_{x,y} \equiv 1$, das heißt, es wird wiederum das Operatoreigenwertproblem

$$-\Delta u = \lambda u, \tag{5.12}$$

nun aber mit den oben angegebenen gemischten Randbedingungen untersucht. Auch für diesen Fall können die analytischen Lösungen explizit angegeben werden. Diese sind durch

$$u_{k,l}(t,\varphi) = c \cdot \sin(\nu(k,\alpha)\ \varphi) \cdot J_{\nu(k,\alpha)}(\omega_{k,l}\, t), \quad (k, l = 0, 1, 2, \dots),$$

mit

$$\nu(k,\alpha) = \frac{k + \frac{1}{2}}{2 - \frac{\alpha}{\pi}}$$

	PINVIT, N_{iter} = 60		PSD, N_{iter} = 41		LOBPCG, N_{iter} = 20	
λ_{ex}	λ_i	$\|\text{err}\|$	λ_i	$\|\text{err}\|$	λ_i	$\|\text{err}\|$
7.82239	7.96429	6.507E-11	7.96429	4.845E-11	7.96429	4.572E-11
12.50257	12.50297	3.513E-11	12.50297	3.660E-11	12.50297	2.078E-11
17.95369	17.95399	5.969E-11	17.95399	4.634E-11	17.95399	7.338E-11
24.14822	24.14876	6.116E-11	24.14876	3.160E-11	24.14876	2.573E-11
31.06592	31.06684	7.867E-11	31.06684	3.265E-11	31.06684	3.280E-11
35.07851	35.53111	5.650E-11	35.53111	7.506E-11	35.53111	5.643E-11
38.69114	38.69259	5.895E-11	38.69259	4.195E-11	38.69259	9.424E-11
44.89436	44.89748	9.624E-11	44.89748	5.044E-11	44.89748	2.427E-11
47.01133	47.01348	9.739E-11	47.01348	4.060E-11	47.01348	3.375E-11
55.50095	55.50382	8.521E-11	55.50382	7.302E-11	55.50382	5.626E-11
56.01619	56.01922	9.958E-11	56.01922	5.734E-11	56.01922	7.351E-11
65.69705	65.70124	7.909E-11	65.70124	9.326E-11	65.70124	5.831E-11
66.88484	66.88902	9.114E-11	66.88902	9.652E-11	66.88902	7.820E-11
76.04650	76.05212	9.712E-11	76.05212	8.289E-11	76.05212	4.414E-11
79.03348	79.03937	8.248E-11	79.03937	9.372E-11	79.03937	7.894E-11

Tabelle 5.4.: Exakte und berechnete Eigenwerte des Operatoreigenwertproblems aus Gleichung (5.10).

und der Besselfunktion erster Art $J_{\nu(k,\alpha)}$ mit Ordnung $\nu(k,\alpha)$ gegeben. Die exakten Eigenwerte λ_{ex} ergeben sich als Quadrat der positiven Nullstellen $(\omega_{k,l})^2$, $(k,l = 0,1,2,\dots)$, von $J_{\nu(k,\alpha)}$.
Für den im Folgenden untersuchten Fall sei, wie in Abbildung 5.5 dargestellt, der Öffnungswinkel $\alpha = \frac{\pi}{12}$. Die exakten Eigenwerte sind in der ersten Spalte von Tabelle 5.4 (aufsteigend sortiert) angegeben.
Die numerischen Formulierung des verallgemeinerten Matrixeigenwertproblems mit Matrixpaar (A, M) basiert wiederum auf der schwachen Formulierung aus Abschnitt 1.1 unter Nutzung linearer Test- und Ansatzfunktionen ϕ_i, gegeben durch Gleichung (5.3). Für die folgenden Betrachtungen führt dies auf Matrizen A und M der Größe $n = 297078$.
Wie bei den Erörterungen zum Modellproblem I in Abschnitt 5.1 soll zunächst auch hier das algebraische Mehrgitterverfahren basierend auf dem *standard coarsening*, umgesetzt durch *AMT*, zum Einsatz kommen. Die zu Grunde liegenden Parameter gleichen ebenso denen des Modellproblems I. Ausgenommen die maximale Anzahl der auf dem gröbsten Level verbleibenden Punkte wird auf $n_0 \leq 100$ erhöht. Motiviert durch die Feststellung am Ende des Abschnitts 5.1 soll auch hier nur ein einzelner $\text{V}(2,2)$-Zyklus zur Berechnung der vorkonditionierten Residuen vorgenommen werden.
Im Gegensatz zum vorher betrachteten Modellproblem für das nur eine Untersuchung mittels PINVIT stattfand, sollen nun auch die (k)-Schemata mit $k > 1$ zum Einsatz kommen. Gleichzeitig soll nicht nur der kleinste Eigenwert des Matrixpaares (A, M) ermittelt, sondern einige der kleinsten, also das partielle Eigenwertproblem gelöst werden. Es kommen daher die im Abschnitt 2.3 vorgestellten Unterraumvarianten der vorkonditionierten Iterationen zum Einsatz. Für die äußere Iteration wird dazu $s = 20$ festgelegt und die Iteration mit zufälligem $\hat{V}^{(1)} \in \mathbb{R}^{n\times s}$ gestartet. Ziel ist es dabei, die 15 kleinsten Eigenwerte zu berechnen. Auch hier wird das i-te Ritzpaar $(\theta^{(j,i)}, u^{(j,i)})$ im j-ten Iterationsschritt als stationär betrachtet, falls es der Bedingung

$$\text{err}_i = \|Au^{(j,i)} - \theta^{(j,i)} Mu^{(j,i)}\| \leq 10^{-10} \tag{5.13}$$

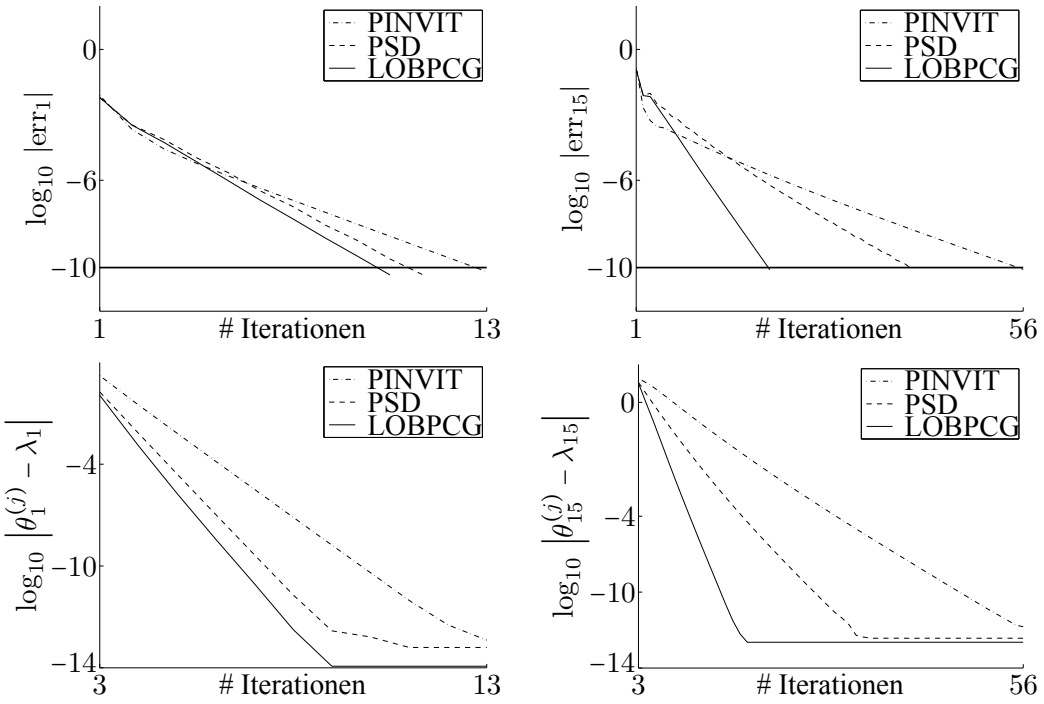

Abbildung 5.6.: Fehlerverläufe für die (k)-Schemata ($k = 1, 2, 3$). Oben: Fehlerverlauf der Eigenvektorapproximationen für $i = 1$ (links) und $i = 15$ (rechts). Unten: Fehlerverlauf der Eigenwertapproximationen für $i = 1$ (links) und $i = 15$ (rechts).

genügt. Bereits konvergierte Ritzpaare werden in den nachfolgenden Iterationsschritten nicht weiter berücksichtigt. Ein positiver Nebeneffekt ist die damit einhergehende Reduktion des Rechenaufwandes, da einerseits für diese Paare keine weiteren vorkonditionierten Residuen berechnet werden müssen. Andererseits führt die Verringerung ihrer Anzahl (für die Algorithmen des (k)-Schemas mit $k \geq 2$) auch zu einer Verkleinerung der Dimension des Unterraums für das Rayleigh-Ritz-Verfahren, also der Dimension des zu lösenden projizierten Eigenwertproblems. Einzig zur Orthonormalisierung der Matrix Z bei der Anwendung des Rayleigh-Ritz-Verfahrens, vergleiche Alorithmen 12, 13 und 14, sind sie wesentlich und können daher nicht vernachlässigt werden.
Der erste Blick soll auf die berechneten Eigenwertapproximationen gerichtet werden. Diese sind in Tabelle 5.4 für die Iterationen PINVIT, PSD und LOBPCG mit zugehörigem (Abbruch-)Fehler angegeben. Alle Algorithmen liefern Eigenwertapproximationen für jeden der gesuchten Eigenwerte $\lambda_i, (i = 1, \ldots, 15)$. Gleichzeitig sind im Kopf der Tabelle die benötigten Iterationsschritte der einzelnen (k)-Schemata angegeben. Die Ergebnisse für das (4)-Schema beziehungsweise (5)-Schema sind nicht explizit angeführt, sie verhalten sich jedoch ähnlich zu den dargestellten Ergebnissen. Einzig die Iterationsanzahl sinkt sowohl im Falle des (4)-Schemas als auch für das (5)-Schema auf $N_{iter} = 18$. Die berechneten Eigenfunktionen zu den kleinsten Eigenwertapproximationen λ_1 bis λ_6 sind in den Abbildungen 5.8 und 5.9 am Ende des Abschnitts dargestellt. Dabei bef nden sich auf der linken Seite der Abbildungen die dreidimensionalen Darstellungen und jeweils rechts die zugehörige Höhenliniendarstellung.
Es soll nun eine quantitative Untersuchung der (k)-Schemata erfolgen. Dazu sei einerseits das Konvergenzverhalten, gemessen an der Zahl der Iterationsschritte, und andererseits auch ein Vergleich der

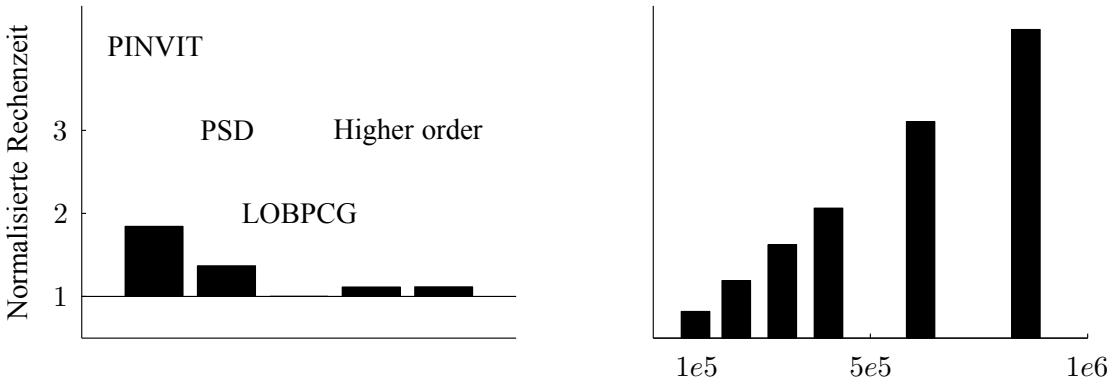

Abbildung 5.7.: Links: Normalisierte Zeiten der Lösungsphasen (LOBPCG = 1). Rechts: Zeitlicher Aufwand zur Lösung des verallgemeinerten Eigenwertproblems mittels LOBPCG.

Rechenzeiten als Kriterium herangezogen. Um dieses zu analysieren, betrachte man die Abbildung 5.6. Diese stellt auf der oberen linken Seite den logarithmierten Fehler aus Gleichung (5.13) für den kleinsten (λ_1) und oben rechts für den größten (λ_{15}) der gesuchten Eigenwerte für die (k)-Schemata mit $k = 1, 2, 3$ über den Iterationsverlauf dar. Dabei erkennt man deutlich, dass von diesen Verfahren für beide Eigenwerte LOBPCG das asymptotisch beste Konvergenzverhalten aufweist. Allerdings suggerieren die Graphiken für die ersten Iterationsschritte ein besseres Konvergenzverhalten von PINVIT. Auch wenn der Fehler err als Abbruchkriterium dient, ist er als Maß für die Konvergenz der vorkonditionierten Iterationen nur eingeschränkt zu verwenden. Hierzu muss der Blick auf den Verlauf der zugehörigen Ritzwerte gerichtet werden. Dementsprechend ist dieser in den unteren Graf ken der Abbildung 5.6 dargestellt. Hier erkennt man deutlich, dass sowohl PSD als auch LOBPCG im Vergleich zu PINVIT in jedem Iterationsschritt die besseren Approximationen an die gesuchten Eigenwerte λ_1 beziehungsweise λ_{15} liefern.
Wie bereits im Vorfeld angedeutet wurde und der Namenszusatz *optimal* suggeriert, stellt LOBPCG bezüglich der Rechenzeit das eff zienteste Verfahren der (k)-Schemata dar. Deutlich wird dies auf der linken Seite der Abbildung 5.7, in der die normalisierten Laufzeiten der reinen Lösungsphasen, also ohne Zeiten der *setup*-Phase zur Konstruktion der Hierarchie (diese sind bei allen Algorithmen gleich), angedeutet sind. Auch wenn für dieses optimale Verhalten kein analytischer Beweis verfügbar ist, kann diese Beobachtung heuristisch interpretiert werden. Durch die Hinzunahme der vorherigen Iterierten vergrößert sich mit steigendem k der Raum $\mathscr{V}_{ks}$ zur Bestimmung der Bestapproximationen innerhalb des Rayleigh-Ritz-Verfahrens. Dies spiegelt sich in der schnelleren Konvergenz der äußeren Iteration wieder. Daher sinkt zwar die Anzahl der Iterationsschritte N_{iter} bei einer Erhöhung der Ordnung des Schemas, allerdings geht dies auch mit einem erhöhten Rechenaufwand zur Lösung des projizierten Eigenwertproblems einher. Dabei scheint es, dass der Zugewinn an Informationen im Fall der (k)-Schemata mit $k > 3$ den Rechenaufwand nicht egalisieren kann und somit der zeitliche Gesamtbedarf steigt. Diese Optimalität f ndet sich auch in den Untersuchungen in [5] wieder. Auf Grund dieser Feststellung sollen sich die weiteren Untersuchungen nun auf das (3)-Schema, also LOBPCG, fokussieren.
Zu diesem Zweck wird der Algorithmus LOBPCG auf das in Gleichung (5.10) gegebene Problem zu unterschiedlichen Diskretisierungen angewendet. Dabei seien Eigenwertprobleme für das Matrixpaar (A, M) mit den Dimensionen $n = 97429$, $n = 191257$, $n = 297078$, $n = 403688$, $n = 615636$ und $n = 857811$ betrachtet. Wiederum sollen jeweils die 15 kleinsten Eigenwertapproximationen unter Ver-

n	AMT			ML		
	L	N_{iter}	λ_{15}	L	N_{iter}	λ_{15}
97429	7	20	79.05168	3	31	79.05168
191257	8	20	79.04268	3	34	79.04268
297078	8	20	79.03937	4	38	79.03937
403688	9	20	79.03792	4	39	79.03792
615636	9	21	79.03634	4	40	79.03634
857811	9	21	79.03557	4	43	79.03557

Tabelle 5.5.: Kenngrößen zur Berechnung der 15-ten Eigenwertapproximation mittels LOBPCG zu verschiedenen Diskretisierungen und Vorkonditionierern.

wendung eines Unterraumes $\mathcal{Y}_{ks} = \mathcal{Y}_{60}$, also für $s = 20$, berechnet werden. Tabelle 5.5 stellt einige Kenngrößen dieser Untersuchungen dar. Speziell für die Diskretisierung mit $n = 97429$ Stützstellen konstruiert das *standard coarsening* (*AMT*) eine Hierarchie mit $L = 7$ Leveln. Die jeweiligen Problemdimensionen $n_i, i = 6, \ldots, 0$, der gröberen Level sind $97429, 44051, 15465, 5637, 2046, 743, 270$ und 96. Die rechte Graf k in Abbildung 5.7 gibt Auskunft über die Zunahme des zeitlichen Gesamtbedarfs (ohne *setup*-Phase zur Konstruktion der Hierarchie) zur Lösung des verallgemeinerten Eigenwertproblems in Abhängigkeit von der Problemdimension. Es ist dabei eine lineare Zunahme zu vermuten, was ein Indiz für sehr gute Komplexitätseigenschaften, im Sinne eines (fast) linear zunehmenden Aufwandes mit steigender Problemdimension, ist.
Im Abschnitt 3.3.2 über die algebraischen Mehrgitterverfahren wurde neben dem *standard coarsening* als zweite Vergröberungsstrategie auch die Methode *smoothed aggregation* zur Konstruktion einer Hierarchie vorgestellt. Diese wird hier, wie bereits am Anfang des Kapitels erwähnt, durch das *ML-Package* realisiert. Ebenfalls in Tabelle 5.5 sind in der letzten Spalte einige Kenngrößen von LOBPCG unter Verwendung vorkonditionierter Residuen, welche durch das algebraische Mehrgitterverfahren, basierend auf dieser Vergröberungsstrategie, berechnet wurden, für die gleichen Matrixpaare (A, M) angegeben. Dabei fällt einerseits auf, dass die konstruierte Hierarchie einem viel stärkeren Vergröberungsprozess unterliegt, was sich in der geringeren Anzahl der Level L widerspiegelt. Andererseits benötigt LOBPCG in Folge eine höhere Anzahl an Iterationsschritten N_{iter} zur Lösung des Problems. Der Grund hierfür wurde bereits bei den Betrachtungen zur Herleitung einer geeigneten Vergröberungsstrategie in Abschnitt 3.3.2 deutlich. Durch die verhältnismäßig geringe Anzahl verbleibender Punkte im jeweils gröberen Gitter umfasst das Bild des Prolongationsoperators zur Abbildung der berechneten Grobgitterkorrekturen weniger darstellbare algebraisch glatte Fehler. Daher sind die erwähnten Beobachtungen eine logische Konsequenz aus der den Verfahren zu Grunde liegenden Vergröberungsstrategien. Abschließend sei festgehalten, dass sowohl unter Einsatz von *AMT* als auch des *ML-Package* alle 15 der geforderten, kleinsten Eigenwerte gefunden wurden. Dies unterstreicht den bereits erörterten Aspekt, dass die algebraischen Mehrgitterverfahren im Sinne einer *black-box*-Operation zur Realisierung der Operation $y \mapsto B^{-1}y$ verwendet werden können, da mit dem Algorithmus LOBPCG unabhängig vom verwendeten algebraischen Mehrgitterverfahren das gestellte Problem gelöst wurde.

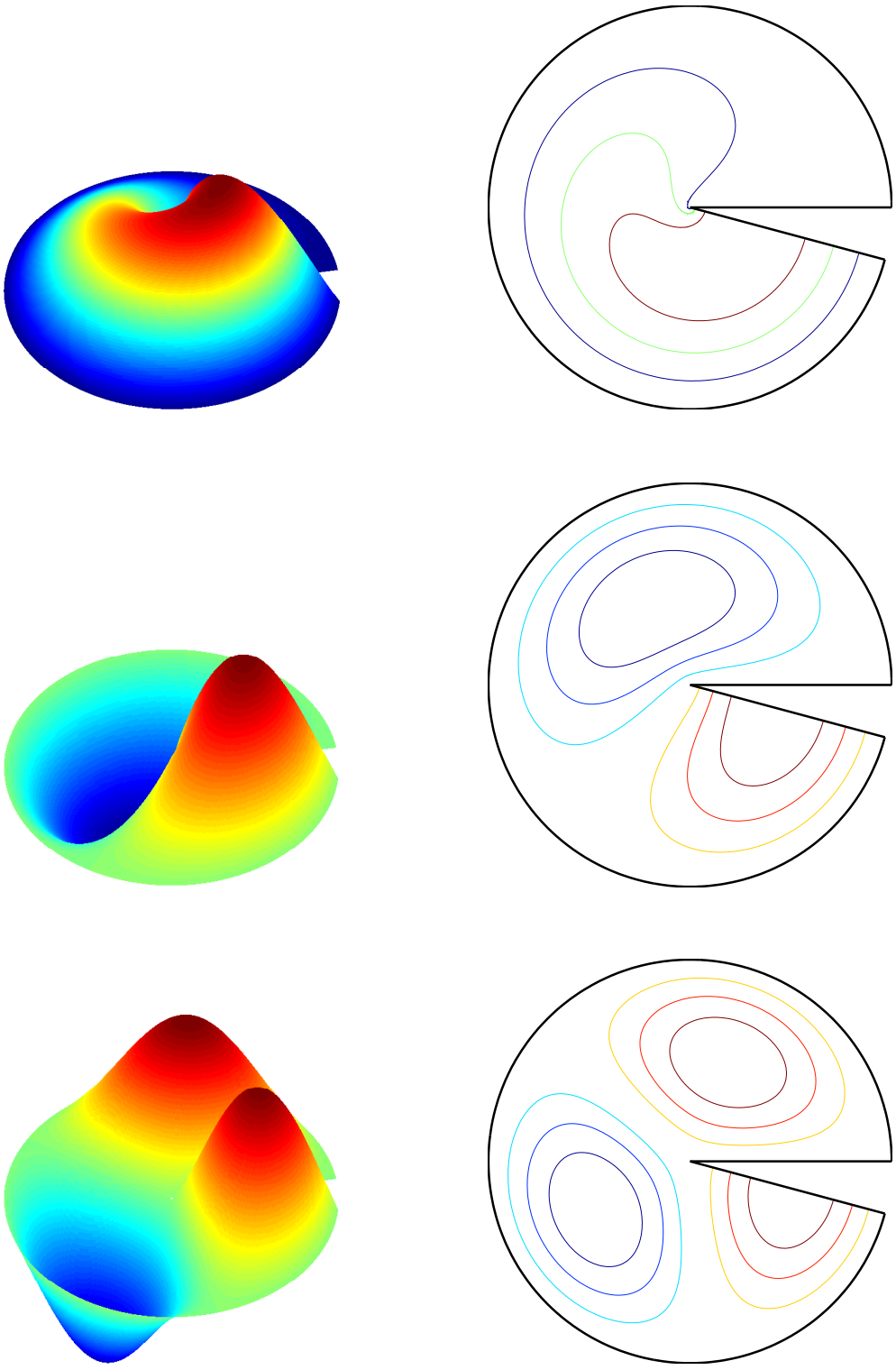

Abbildung 5.8.: Eigenfunktion zu den drei kleinsten berechneten Eigenwerten des Modellproblems II. Links: 3D-Darstellung. Rechts: Höhenliniendarstellung.

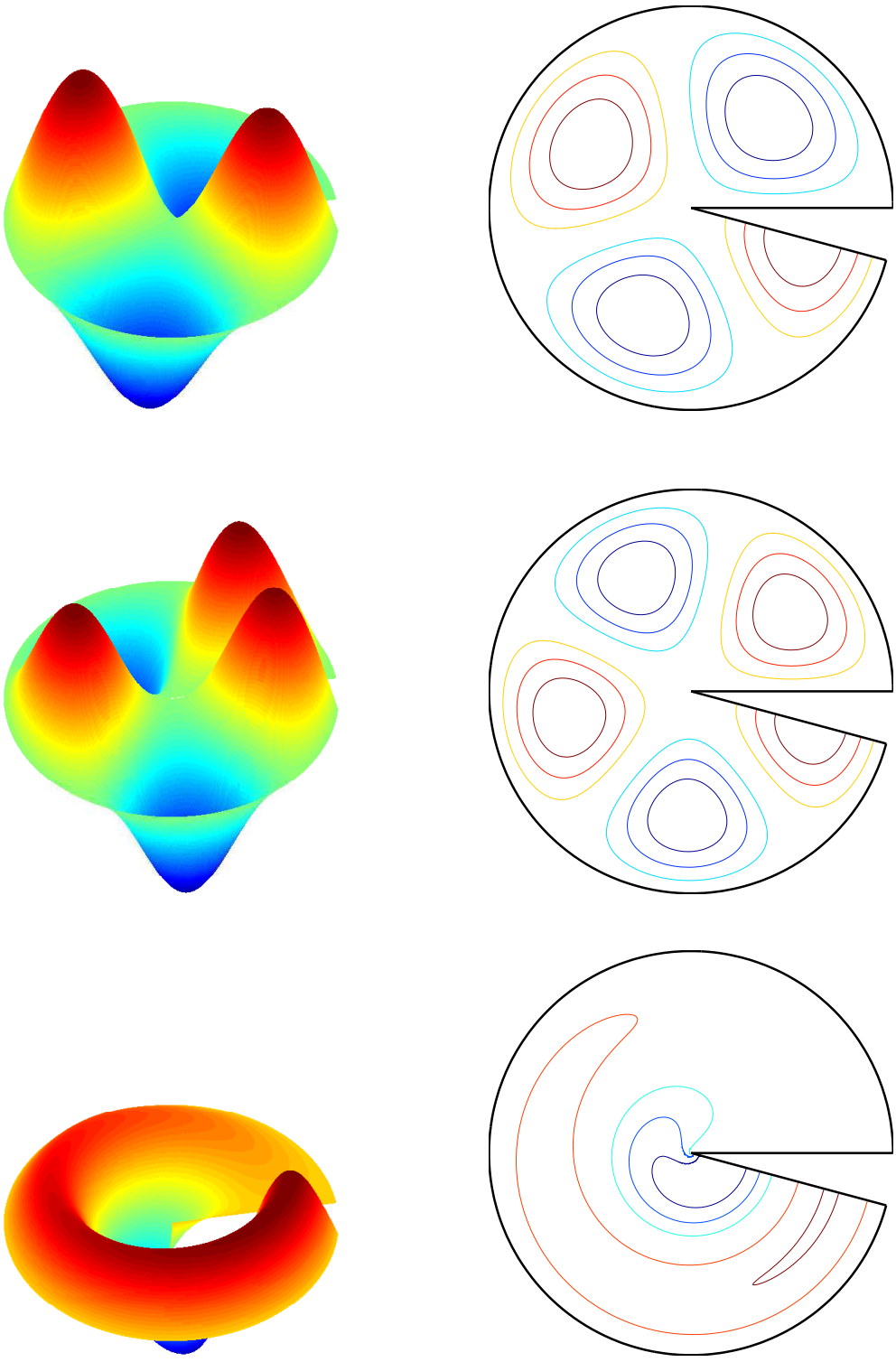

Abbildung 5.9.: Eigenfunktionen zum vierten bis sechsten berechneten Eigenwert des Modellproblems II. Links: 3D-Darstellung. Rechts: Höhenliniendarstellung.

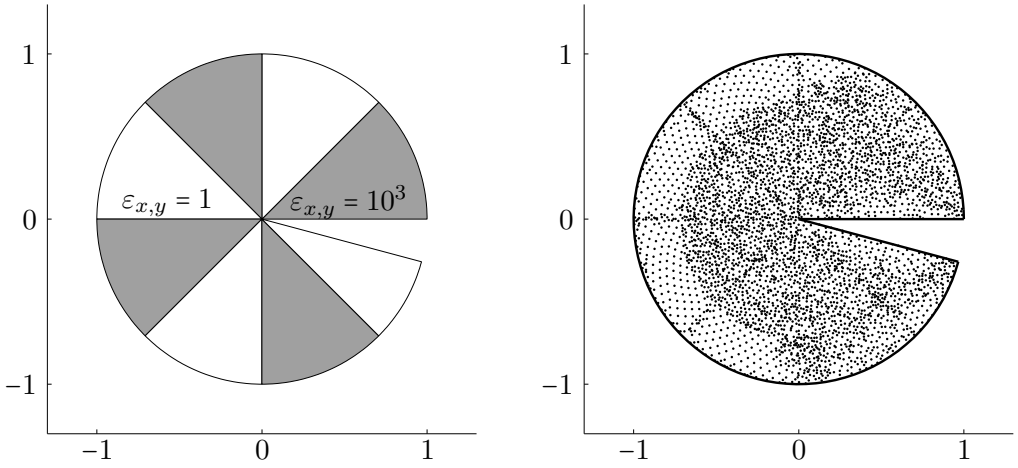

Abbildung 5.10.: Anisotropes Modellproblem III. Links: Verteilung der Anisotropie $\varepsilon_{x,y}$. Rechts: Gitterpunkte des Levels $l = 4$ im Fall einer Diskretisierung mit $n = 191257$ Unbekannten.

5.3. Modellproblem III - Ein stark anisotropes Modellproblem

Dieser Abschnitt soll die Anwendbarkeit der Eigenlöser, die auf einer Vorkonditionierung mittels algebraischen Mehrgitterverfahren beruhen, zur Lösung stark anisotroper Probleme aus der Klasse der betrachteten Operatoreigenwertprobleme verdeutlichen, wie sie zum Beispiel in den Arbeiten von Arbenz, [1], oder Borzi und Borzi, [5], nicht betrachtet werden. Vorrangig in der Literatur für algebraische Mehrgitterverfahren wird aber unterstrichen, dass diese für die Lösung von Fragestellungen aus der genannten Problemklasse sehr gut geeignet sind, [69]. Um dies zu verdeutlichen, wird wiederum das Operatoreigenwertproblem aus Gleichung (5.10) gegeben durch

$$\begin{aligned} -\nabla \cdot (\varepsilon_{x,y} \nabla u) &= \lambda u \\ u &= 0 \quad \text{auf } \Gamma_1 \\ n_{x,y} \cdot \varepsilon_{x,y} \nabla u &= 0 \quad \text{auf } \Gamma_2 \end{aligned} \tag{5.14}$$

betrachtet. Der Unterschied liegt in der Wahl der Funktion $\varepsilon_{x,y}$, welche die Anisotropie beschreibt. Diese ist im Folgenden durch

$$\varepsilon_{x,y} = \begin{cases} 1 & \text{für } \varphi \in [(2m+1)\frac{\pi}{4}, (2m+2)\frac{\pi}{4}),\ t \in [0,1] \\ 10^3 & \text{für } \varphi \in [(2m)\frac{\pi}{4}, (2m+1)\frac{\pi}{4}),\ t \in [0,1], \end{cases} \qquad (m = 0,1,2,3) \tag{5.15}$$

gegeben. Die Verteilung der Funktion $\varepsilon_{x,y}$ auf dem Gebiet Ω ist in Abbildung 5.10 auf der linken Seite dargestellt.
Auch hier soll sich auf die Berechnung der gesuchten Eigenwerte mittels LOBPCG, dem effektivsten (k)-Schema, konzentriert werden. Die Parameter seien wie im vorhergehenden Abschnitt gewählt, das heißt, es sollen für $s = 20$ die 15 kleinsten Eigenpaare unter Verwendung eines $\mathrm{V}(2,2)$-Zyklus, basierend auf einer durch *AMT* konstruierten Hierarchie, ermittelt werden.
Bevor die Ergebnisse der Berechnung präsentiert werden, soll ein kurzer Blick auf die Hierarchie gerichtet werden. Dazu ist in Abbildung 5.10 rechts die Menge der Punkte des Levels $l = 4$ im Falle einer Diskretisierung mit $n = 191257$ Unbekannten dargestellt. Auffällig ist, dass während des Vergröberungsprozesses keine „homogene“ Verteilung der Grobgitterpunkte erfolgt, sondern die Sprungstellen

λ_1	40.91474	λ_6	123.08373	λ_{11}	150.66492
λ_2	57.87177	λ_7	123.24433	λ_{12}	170.50737
λ_3	57.96446	λ_8	123.50582	λ_{13}	179.32855
λ_4	58.04346	λ_9	150.17197	λ_{14}	207.80323
λ_5	95.82098	λ_{10}	150.38864	λ_{15}	208.10978

Tabelle 5.6.: Berechnete kleinste Eigenwerte zum anisotropen Modellproblem III.

	1	10^3	10^6	10^9
γ	0.3142	0.3343	0.3353	0.3384
N_{iter}	20	22	22	22
L	8	9	9	9

Tabelle 5.7.: Parameter des algebraischen Mehrgitterverfahrens zu unterschiedlich def nierten Anisotropien.

der Funktion $\varepsilon_{x,y}$ automatisch während des Vergröberungsprozesses detektiert werden. Es sei nochmals unterstrichen, dass dies einzig auf Basis der zu Grunde liegenden algebraischen Gleichung geschieht und keine geometrischen Informationen genutzt werden.
Die Berechnung des verallgemeinerten Eigenwertproblems für das entsprechend formulierte Matrixpaar (A, M) der Dimension $n = 297078$ liefert die in Tabelle 5.6 angegebenen Eigenwerte. Eine Darstellung der berechneten Eigenfunktionen zu ausgewählten Eigenwerten sind in den Abbildungen 5.11 sowie 5.12 zu f nden. Man erkennt deutlich, dass die Verteilung der Anisotropie $\varepsilon_{x,y}$ zu einer starken Lokalisierung der Gradienten der Eigenfunktionen führt.
Auch für diese Berechnungen soll das algebraische Mehrgitterverfahren etwas genauer analysiert werden. Ein Vergleich zum isotropen Modellproblem II ermöglicht zudem die Betrachtung der Tabelle 5.7. Im Gegensatz zu diesem wird auf Grund der Anisotropie eine Hierarchie mit einem Level mehr, also $L = 9$ Leveln, konstruiert. Zudem steigt die Anzahl der benötigten Iterationsschritte leicht von $N_{iter} = 20$ auf $N_{iter} = 22$. Dabei wird weiterhin deutlich, dass nur ein geringer Qualitätsverlust des Vorkonditionierers, charakterisiert durch γ, berechnet als $\varrho(\mathcal{T}_L^{\text{num}})$ entsprechend den Erläuterungen in Abschnitt 5.1 mit $m = 5$, auftritt. Ebenso führt eine Erhöhung des Koeff zienten in der Def nition der Anisotropie, siehe Gleichung (5.15), von 10^3 auf 10^6 beziehungsweise 10^9 zu keiner wesentlichen Verschlechterung dieser Qualität. Insgesamt bleibt festzuhalten, dass, zumindest für das vorliegende Modellproblem, eine Berechnung der Eigenwerte mittels des genutzten iterativen Eigenlösers unter Verwendung der algebraischen Mehrgittervorkonditionierung möglich ist.

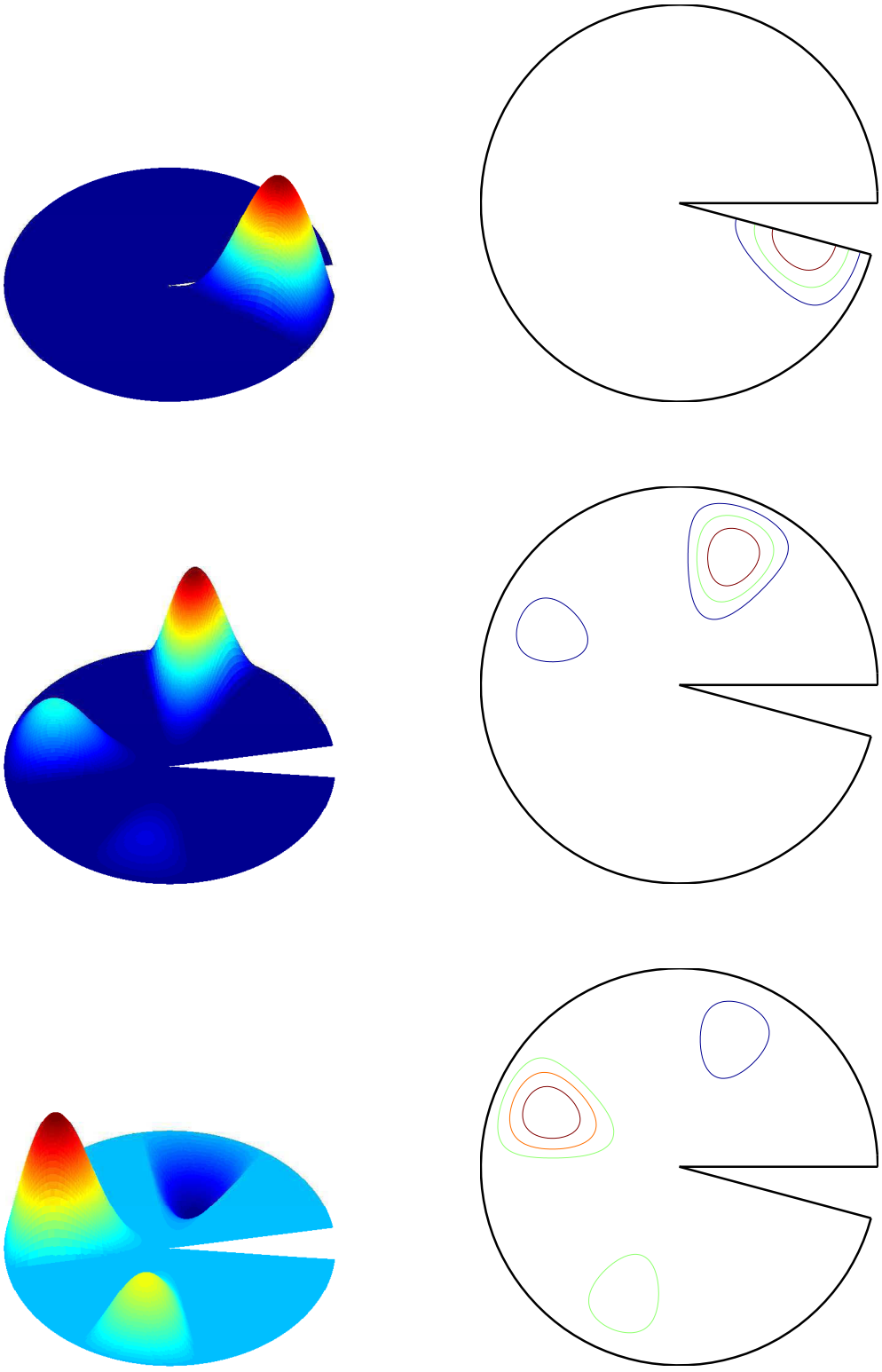

Abbildung 5.11.: Eigenfunktionen zu den kleinsten Eigenwerten des Modellproblems III. Links: 3D-Darstellung. Rechts: Höhenliniendarstellung.

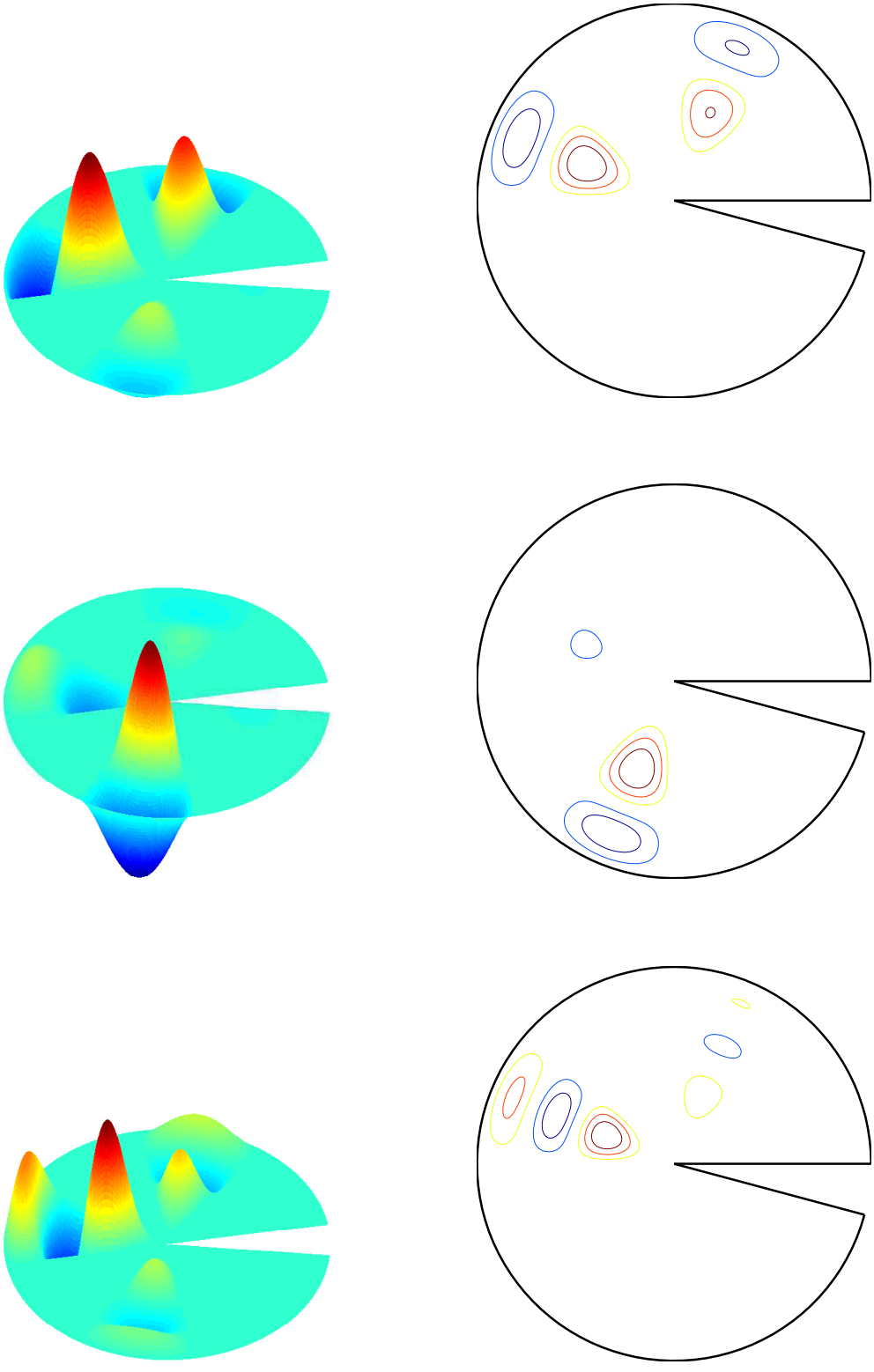

Abbildung 5.12.: Eigenfunktionen zu den berechneten Eigenwerten λ_6, λ_8 und λ_{14} des Modellproblems III. Links: 3D-Darstellung. Rechts: Höhenliniendarstellung.

5.4. Algebraische Eigenwertprobleme der Harwell-Boeing-Bibliothek

Anfang der 1980-er Jahre resultierten aus technischen Fragestellungen Probleme, die mit klassischen Methoden, wie *Jacobi-Verfahren* oder *QZ-Algorithmus*, nicht mehr zu bewältigen waren. Dies lag insbesondere an der steigenden Problemdimension und an den damit wachsenden Anforderungen an physikalische Speicherkapazitäten, da die benannten klassischen Verfahren angenehme Strukturen der auftretenden Matrizen zerstören. Heutzutage sind mit etablierten Verfahren, wie Jacobi-Davidson oder Lanzcos-Verfahren, solche Probleme sehr gut zu lösen. Ebenso eignen sich hierbei aber auch die vorkonditionierten Iterationen unter Verwendung algebraischer Mehrgittervorkonditionierung, wie nun gezeigt werden soll.
Im Zuge der wachsenden Herausforderungen ist von industrieller Seite die *Harwell-Boeing-Bibliothek* (*Harwell-Boeing-Collection*) veröffentlicht worden (zu finden unter [45]), welche unter anderem auch verallgemeinerte Eigenwertprobleme aus der Strukturmechanik beinhaltet. Auch wenn die dort angegebenen Probleme einem technischen Hintergrund besitzen, sind sie in einer rein algebraischen Form präsentiert, das heißt, es sind nur die jeweiligen Matrixpaare (A, M) gegeben. Dies verbietet den durchaus vorstellbare Einsatz einer geometrischen Mehrgittervorkonditionierung, kann aber mittels des algebraischen Ansatzes realisiert werden. Zur Demonstration der Anwendbarkeit seien hier drei Beispiele aus dem Set *BCSSTRUC1: BCS Structural Engineering Matrices*, genauer die verallgemeinerten Eigenwertprobleme des „TV-Studios“, einer eingespannten Platte und die Verformung eines heißen Dichtungsringes, betrachtet. Diese stellen zwar nur niedrigdimensionale Probleme dar, sollen aber dennoch zum Beweis der Leistungsfähigkeit herangezogen werden.
Alle Probleme werden durch die vorkonditionierten Iteration *LOBPCG* unter Verwendung algebraischer Mehrgittervorkonditionierung mittels eines $\mathrm{V}(2,2)$-Zyklus mit Parametern wie unter Abschnitt 5.1 und jeweils mit der Blockgröße $s = 10$ berechnet. Dabei sollen die acht kleinsten Eigenwerte ermittelt werden. Die berechneten Eigenwert- und Eigenvektorapproximation werden als stationär betrachtet, falls sie der Bedingung

$$\mathrm{err} = \|Au^{(j)} - \theta^{(j)} Mu^{(j)}\| \leq 10^{-6} \tag{5.16}$$

genügen.

Eingespannte Platte

Als erster Fall soll das Problem der eingespannten Platte untersucht werden. Das gestellte verallgemeinerte Eigenwertproblem besitzt die Dimension $n = 1083$. Das algebraische Mehrgitterverfahren konstruiert eine Hierarchie mit $L = 6$ Leveln. Diese besitzen die jeweiligen Dimensionen $n_0 = 5, n_1 = 13, n_2 = 35, n_3 = 123, n_4 = 243$ und $n_5 = 540$. Zur Berechnung werden $N_{iter} = 21$ Iterationen benötigt. Die fünf kleinsten berechneten Eigenwerte sind in Tabelle 5.8 in der ersten Zeile und darunter der zugehörige, nach Gleichung (5.16) erhaltene Abbruchfehler angegeben. Als zusätzliche Referenz dienen die mit *ARPACK* in Form der Matlab-Routine *eigs* errechneten Eigenwerte, zu finden in der dritten Zeile.

	λ_1	λ_2	λ_3	λ_4	λ_5
LOBPCG	29068634.21	120700725.42	120700725.42	259981418.80	387801377.44
err	4.020E-08	3.005E-08	2.630E-08	3.451E-08	2.305E-08
ARPACK	29068634.21	120700725.42	120700725.42	259981418.80	387801377.44

Tabelle 5.8.: Kleinste Eigenwerte zum Harwell-Boeing-Problem *Eingespannte Platte*.

Man erkennt, dass die vorkonditionierte Iteration zur Lösung des Problems geeignet ist.

Verformung eines heißen Dichtungsringes

Die zweite Problemstellung weist eine vergleichbare Dimension zum ersten auf und ist durch $n = 1086$ gegeben. Im Gegensatz wird durch *AMT* eine Hierarchie mit $L = 5$ Leveln generiert. Die einzelnen Level besitzen die Dimensionen $n_0 = 7, n_1 = 20, n_3 = 52, n_4 = 175$ und $n_5 = 509$. Es werden $N_{iter} = 165$ Schritte benötigt. In Tabelle 5.9 sind, wie in den vorherigen Betrachtungen, die fünf kleinsten Eigenwerte und Referenzwerte angegeben.

	λ_1	λ_2	λ_3	λ_4	λ_5
LOBPCG	0.078648	0.079390	0.079611	0.081514	0.082800
err	2.272E-08	3.654E-08	5.033E-08	3.085E-08	1.474E-08
ARPACK	0.078648	0.079390	0.079611	0.081514	0.082800

Tabelle 5.9.: Kleinste Eigenwerte zum Harwell-Boeing-Problem *Verformung eines Dichtungsringes*.

Auch hier bleibt festzuhalten, dass eine erfolgreiche Berechnung der gesuchten Größen stattf ndet.

TV-Studio

Das als letztes betrachtete Problem des „*TV-Studios*“ hat die Dimension $n = 1074$. Das algebraische Mehrgitterverfahren konstruiert hier eine Hierarchie mit $L = 4$ Leveln, die jeweiligen Problemdimensionen der Level sind mit $n_0 = 7, n_1 = 14, n_2 = 102$ und $n_3 = 455$. Zur Berechnung sind $N_{iter} = 102$ Iterationsschritte nötig. Tabelle 5.10 gibt wiederum die fünf kleinsten jeweils mit *LOBPCG* und zur Referenz mittels *ARPACK* berechneten Eigenwerte an.

	λ_1	λ_2	λ_3	λ_4	λ_5
LOBPCG	6.90070	18.14203	18.14237	18.14237	84.78616
err	1.555E-07	1.670E-07	5.910E-08	2.566E-07	1.725E-08
ARPACK	6.90070	18.14203	18.14237	18.14237	84.78616

Tabelle 5.10.: Kleinste Eigenwerte zum Harwell-Boeing-Problem *TV-Studio*.

Als kurzes Resümee bleibt festzuhalten, dass die vorkonditionierten Iterationen auf Basis algebraischer Mehrgittervorkonditionierung zur Lösung der exemplarisch gewählten Beispiele der *Harwell-Boeing-Bibliothek* erfolgreich eingesetzt werden können. Es muss aber auch erwähnt werden, dass ein Vergleich der Rechenzeiten zwischen *LOBPCG* und der Lösung mittels *ARPACK* deutlich zugunsten des Letztgenannten ausfällt. Dies liegt aber hauptsächlich darin begründet, dass die vorkonditionierten Iterationen ihre Vorteile vorrangig beim Einsatz zur Lösung hochdimensionaler Probleme ausspielen können.

6. Zusammenfassung

Die vorliegende Arbeit behandelt die Lösung des verallgemeinerten Eigenwertproblems

$$Au = \lambda Mu$$

für symmetrisch positiv def nite Matrizen A und (möglicherweise semidef nite Matrizen) M. Vor dem Hintergrund technischer Fragestellungen entstehen dabei große und dünnbesetzte Matrizen, nicht selten in der Größenordnung einer Million Unbekannter, bei denen klassische Methoden wie der QZ-Algorithmus nicht mehr angewendet werden können. Die hier thematisierte Methode der vorkonditionierten Iterationen ist dafür bekannt, eine adäquate Behandlung von Eigenwertproblemen aus dieser Problemklasse vorzunehmen. Die Konvergenzrate dieser Iterationsverfahren wird maßgeblich von der Güte des verwendeten Vorkonditionierers B^{-1} zur Berechnung der auftretenden Korrekturgrößen bestimmt. Eine eff ziente Methode für diese Berechnung stellen Mehrgitterverfahren dar. Im Gegensatz zum etablierten Einsatz geometrischer Mehrgitterverfahren ist in dieser Arbeit die Verwendung algebraischer Mehrgitterverfahren vorgeschlagen und untersucht worden. Ein wesentlicher Vorteil liegt dabei in der Verwendbarkeit für geometriefreie Probleme, da diese Methode einzig auf Basis der algebraischen Gleichungen umgesetzt wird.

Die Arbeit gliedert sich dabei in drei wesentliche Teile.
Der erste Teil widmet sich den gradientenbasierten Iterationen. Im Zuge der Betrachtungen ist neben einer ausführlichen Darstellung des Konvergenzbeweises der vorkonditionierten Iteration (PINVIT) von Knyazev und Neymeyr das (k)-Schema zur Konstruktion einer Klasse von Eigenlösern basierend auf PINVIT zur späteren Betrachtung im Anwendungsteil angeführt.
Der zweite Teil rückt die Mehrgitterverfahren in den Mittelpunkt. Diese werden in ihrem allgemeingültigen Aufbau motiviert und dargestellt. Anschließend richtet sich der Fokus auf die algebraischen Mehrgitterverfahren. Speziell die Methoden des *standard coarsening* nach Stüben und *smoothed aggregation* nach Vaněk werden, wiederum mit Blick auf die Anwendungen, genauer beleuchtet.
Der letzte Teil der Arbeit dient der umfangreichen Untersuchung der vorgestellten Algorithmen anhand mehrerer Modellprobleme. Im einzelnen werden dabei neben einem klassischen Beispiel ebenso gitterfreie und stark anisotrope Probleme, die in der Literatur erwähnt aber seltener untersucht werden, betrachtet. Begleitet werden alle Berechnungen von der Angabe aussagekräftiger Leistungsparameter für die jeweils verwendeten Algorithmen.

In der Arbeit konnte gezeigt werden, dass die vorkonditionierten Iterationen unter Verwendung algebraischer Mehrgitterverfahren zur Lösung der gestellten Probleme verwendet werden können. Gerade im Fall anisotroper Probleme, bei denen sie nur wenig in ihrer Leistungsfähigkeit gegenüber homogenen Problemen einbüßen, sollten sie einem Vergleich zu etablierten Methoden, beispielsweise in einem Szenario ähnlich des von Arbenz et al. untersuchten, standhalten. Zudem untermauern die durchgeführten Untersuchungen die den algebraischen Mehrgitterverfahren in der Literatur vielfach unterstellten guten Eigenschaften bei der approximativen Lösung linearer Gleichungssysteme. Zwar ist bekannt, dass die geometrischen Mehrgitterverfahren bei den betrachteten Fragestellungen durch Nutzung der zu Grunde liegenden Geometrie im Allgemeinen bessere Konvergenz- und Komplexitätseigenschaften aufweisen, doch zeigen die Ergebnisse, dass die hier betrachteten algebraischen Mehrgitterverfahren einen leistungs-

fähigen Ersatz insbesondere für den Fall gitterfreier Probleme darstellen. Auf Grund ihres mathematischen Konzepts, einzig auf Basis der linearen Gleichungen zu arbeiten, bilden algebraische Mehrgitterverfahren daher eine eff ziente Möglichkeit zur Realisierung der Abbildung $y \mapsto B^{-1}y$ in Sinne einer *black-box*-Operation.

Literaturverzeichnis

[1] P. Arbenz, U. L. Hetmaniuk, R. B. Lehoucq, and R. S. Tuminaro. A comparison of eigensolvers for large-scale 3D modal analysis using AMG-preconditioned iterative methods. *Int. J. Numer. Methods Eng.*, 64(2):204–236, 2005.

[2] Z. e. Bai, J. e. Demmel, J. e. Dongarra, A. e. Ruhe, and H. e. Van der Vorst. *Templates for the solution of algebraic eigenvalue problems. A practical guide.* Software - Environments - Tools, 11. Philadelphia: SIAM, 2000.

[3] N. Bakhvalov. On the convergence of a relaxation method under natural constraints on the elliptic operator. *Zh. Vychisl. Mat. Mat. Fiz.*, 6:861–883, 1966.

[4] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences.* Classics in Applied Mathematics, 9. Philadelphia: SIAM, 1994.

[5] A. Borzì and G. Borzì. Algebraic multigrid methods for solving generalized eigenvalue problems. *Int. J. Numer. Methods Eng.*, 65(8):1186–1196, 2006.

[6] D. Braess. *Finite elements. Theory, fast solvers and applications in elasticity theory.* Berlin: Springer, 2007.

[7] D. Braess and W. Hackbusch. A new convergence proof for the multigrid method including the V-cycle. *SIAM J. Numer. Anal.*, 20:967–975, 1983.

[8] J. H. Bramble and J. E. Pasciak. Uniform convergence estimates for multigrid V-cycle algorithms with less than full elliptic regularity. Quarteroni, Alf o (ed.) et al., Domain decomposition methods in science and engineering. The sixth international conference on domain decomposition, Como, Italy, June 15-19, 1992. Providence: American Mathematical Society. Contemp. Math. 157, 17-26, 1994.

[9] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comput.*, 57(195):1–21, 1991.

[10] J. H. Bramble, J. E. Pasciak, and J. Xu. The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms. *Math. Comput.*, 56(193):1–34, 1991.

[11] A. Brandt. Multi-Level adaptive solutions to boundary-value problems. *Math. Comput.*, 31(138):333–390, 1977.

[12] A. Brandt. Algebraic multigrid theory: The symmetric case. *Appl. Math. Comput.*, 19:23–56, 1986.

[13] A. Brandt. Multiscale scientif c computation: Review 2001. Barth, Timothy J. (ed.) et al., Multiscale and multiresolution methods. Theory and applications. Berlin: Springer. Lect. Notes Comput. Sci. Eng. 20, 3-95, 2002.

[14] A. Brandt, J. Brannick, K. Kahl, and I. Livshits. Bootstrap AMG. *SIAM J. Sci. Comput.*, 33(2):612–632, 2011.

[15] M. Brezina, A. Cleary, R. Falgout, V. Henson, and J. Jones. Algebraic multigrid based on element interpolation (AMGe). *SIAM J. Sci. Comput.*, 22(5):1570–1592, 2000.

[16] M. Brezina, R. Falgout, T. A. Manteuffel, C. MacLachlan, S. McCormick, and J. Ruge. Adaptive algebraic multigrid. *SIAM J. Sci. Comput.*, 27(4):1261–1286, 2006.

[17] T. Chartier, R. D. Falgout, V. E. Henson, J. Jones, T. Manteuffel, S. McCormick, J. Ruge, and P. S. Vassilevski. Spectral AMGe (ρAMGe). *SIAM J. Sci. Comput.*, 25(1):1–26, 2003.

[18] C. C. Douglas. MGNet Homepage. http://www.mgnet.org/.

[19] E. D'yakonov. Iteration methods in eigenvalue problems. *Math. Notes*, 34:945–953, 1983.

[20] E. D'yakonov and M. Orekhov. On the minimization of computational work in eigenvalue problems. *Dokl. Akad. Nauk SSSR*, 235(5):1005–1008, 1977.

[21] E. D'yakonov and M. Orekhov. Minimization of the computational labor in determining the first eigenvalues of differential operators. *Math. Notes*, 27:382–391, 1980.

[22] R. Fedorenko. A relaxation method for solving elliptic difference equations. *U.S.S.R. Comput. Math. Math. Phys.*, 1:1092–1096, 1961.

[23] R. Fedorenko. The speed of convergence of one iterative process. *U.S.S.R. Comput. Math. Math. Phys.*, 4:227–235, 1964.

[24] M. Gee, C. Siefert, J. Hu, R. Tuminaro, and M. Sala. ML 5.0 smoothed aggregation user's guide. Technical Report SAND2006-2649, Sandia National Laboratories, 2006.

[25] S. Godunov, V. Ogneva, and G. Prokopov. On the convergence of the modified method of steepest descent in the calculation of eigenvalues. *Am. Math. Soc., Translat., II. Ser.*, 105:111–116, 1976.

[26] G. H. Golub and H. A. van der Vorst. Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.*, 123(1-2):35–65, 2000.

[27] T. Grauschopf, M. Griebel, and H. Regler. Additive multilevel preconditioners based on bilinear interpolation, matrix-dependent geometric coarsening and algebraic multigrid coarsening for second-order elliptic PDEs. *Appl. Numer. Math.*, 23(1):63–95, 1997.

[28] M. Gu and S. C. Eisenstat. A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM J. Matrix Anal. Appl.*, 16(1):172–191, 1995.

[29] W. Hackbusch. On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method. *SIAM J. Numer. Anal.*, 16:201–215, 1979.

[30] W. Hackbusch. On the convergence of multi-grid iterations. *Beitr. Numer. Math.*, 9:213–239, 1981.

[31] W. Hackbusch. Multi-grid convergence theory. Multigrid methods, Proc. Conf., Köln-Porz 1981, Lect. Notes Math. 960, 177-219 (1982)., 1982.

[32] W. Hackbusch. *Multi-grid methods and applications.* Springer Series in Computational Mathematics, 4. Berlin: Springer, 1985.

[33] W. Hackbusch. *Iterative solution of large sparse systems of equations.* Applied Mathematical Sciences, 95. New York: Springer, 1994.

[34] W. Hackbusch. *Elliptic differential equations: theory and numerical treatment. Transl. from the German by Regine Fadiman and Patrick D. F. Ion.* Springer Series in Computational Mathematics, 18. Dordrecht: Springer, 2010.

[35] F. Hecht. FreeFem++. Software and documentation. http://www.freefem.org/.

[36] U. Hetmaniuk. A Rayleigh quotient minimization algorithm based on algebraic multigrid. *Numer. Linear Algebra Appl.*, 14(7):563–580, 2007.

[37] A. Knyazev. A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace. Numerical treatment of eigenvalue problems. Vol. 5, Proc. Workshop, Oberwolfach/Germ. 1990, ISNM 96, 143-154, 1991.

[38] A. V. Knyazev. Preconditioning eigensolvers – an Oxymoron? *ETNA, Electron. Trans. Numer. Anal.*, 7:104–123, 1998.

[39] A. V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comput.*, 23(2):517–541, 2001.

[40] A. V. Knyazev and K. Neymeyr. A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems. *Linear Algebra Appl.*, 358(1-3):95–114, 2003.

[41] A. V. Knyazev and K. Neymeyr. Gradient f ow approach to geometric convergence analysis of preconditioned eigensolvers. *SIAM J. Matrix Anal. Appl.*, 31(2):621–628, 2009.

[42] M. Krüger. Algebraic multigrid preconditioning for iterative eigensolvers. *PAMM*, 8:10817–10818, 2008.

[43] M. Krüger. Algebraic multigrid preconditioning for iterative eigensolvers. *Archives of Transport*, 22:97–108, 2010.

[44] S. Larsson and V. Thomée. *Partial differential equations with numerical methods.* Texts in Applied Mathematics, 45. Berlin: Springer, 2009.

[45] J. Lewis. Dynamic analyses in structural engineering. Boeing Computer Services, Seattle, Washington, USA, 1982. http://math.nist.gov/MatrixMarket/data/Harwell-Boeing/bcsstruc1/bcsstruc1.html.

[46] J. Mandel and S. McCormick. A multilevel variational method for $Au = \lambda Bu$ on composite grids. *J. Comput. Phys.*, 80(2):442–452, 1989.

[47] MATLAB. *Version 7.7.0 (R2008b)*. The MathWorks Inc., Natick, Massachusetts, 2008.

[48] S. McCormick. Multigrid methods for variational problems: Further results. *SIAM J. Numer. Anal.*, 21:255–263, 1984.

[49] S. McCormick. Multigrid methods for variational problems: General theory for the V- cycle. *SIAM J. Numer. Anal.*, 22:634–643, 1985.

[50] S. McCormick and J. Ruge. Multigrid methods for variational problems. *SIAM J. Numer. Anal.*, 19:924–929, 1982.

[51] S. F. e. McCormick. *Multigrid methods.* Frontiers in Applied Mathematics, 3. Philadelphia: SIAM, 1987.

[52] A. Meister. *Numerical methods for linear systems of equations. An introduction to modern methods. With MATLAB-implementations of C. Vömel.* Studium. Wiesbaden: Vieweg+Teubner, 2011.

[53] C. Moler and G. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.*, 10:241–256, 1973.

[54] A. Napov and Y. Notay. Comparison of bounds for V-cycle multigrid. *Appl. Numer. Math.*, 60:176–192, 2010.

[55] K. Neymeyr. A geometric convergence theory for the preconditioned steepest descent iteration. submitted to `arXiv:1108.2365v1 [math.NA]`.

[56] K. Neymeyr. A geometric theory for preconditioned inverse iteration. I: Extrema of Rayleigh quotient. *Linear Algebra Appl.*, 322(1-3):61–85, 2001.

[57] K. Neymeyr. A geometric theory for preconditioned inverse iteration. II: Convergence estimates. *Linear Algebra Appl.*, 322(1-3):87–104, 2001.

[58] K. Neymeyr. A geometric theory for preconditioned inverse iteration applied to a subspace. *Math. Comput.*, 71(237):197–216, 2002.

[59] K. Neymeyr. On preconditioned eigensolvers and invert-Lanczos processes. *Linear Algebra Appl.*, 430(4):1039–1056, 2009.

[60] Y. Notay. AGMG. Software and documentation. http://homepages.ulb.ac.be/~ynotay/.

[61] Y. Notay. A robust algebraic multilevel preconditioner for non-symmetric M-matrices. *Numer. Linear Algebra Appl.*, 7(5):243–267, 2000.

[62] Y. Notay. Robust parameter-free algebraic multilevel preconditioning. *Numer. Linear Algebra Appl.*, 9(6-7):409–428, 2002.

[63] Y. Notay. An aggregation-based algebraic multigrid method. *ETNA, Electron. Trans. Numer. Anal.*, 37:123–146, 2010.

[64] B. N. Parlett. *The symmetric eigenvalue problem.* Prentice-Hall Series in Computational Mathematics. New Jersey: Prentice-Hall, Inc., 1980.

[65] W. Petryshyn. On the eigenvalue problem $Tu - \lambda Su = 0$ with unbounded and nonsymmetric operators T and S. *Philos. Trans. R. Soc. Lond., Ser. A*, 262:413–458, 1968.

[66] A. Reusken. Convergence of the multilevel full approximation scheme including the V- cycle. *Numer. Math.*, 53(6):663–686, 1988.

[67] A. Reusken. On a robust multigrid solver. *Computing*, 56(3):303–322, 1996.

[68] K. Stüben. Algebraic multigrid (AMG): Experiences and comparisons. *Appl. Math. Comput.*, 13:419–451, 1983.

[69] K. Stüben. Algebraic multigrid (amg): An introduction with applications. *GMD report 53*, March 1999.

[70] G. Temple. The theory of Rayleigh's principle as applied to continuous systems. *Proceedings Royal Soc. London (A)*, 119:276–293, 1928.

[71] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid. With guest contributions by A. Brandt, P. Oswald, K. Stüben.* Orlando: Academic Press, 2001.

[72] P. Vaněk. Fast multigrid solver. *Appl. Math., Praha*, 40(1):1–20, 1995.

[73] R. S. Varga. *Matrix iterative analysis. 2nd revised and expanded ed.* Springer Series in Computational Mathematics, 27. Berlin: Springer, 2000.

[74] M. Verbeek, J. Cullum, and W. Joubert. Algebraic MultiGrid - Toolbox for Matlab. Developed at the University of California at Los Alamos National Laboratory (the University) under contract with the U.S. Department of Energy (DOE).

[75] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-ref nement techniques.* Wiley-Teubner Series Advances in Numerical Mathematics. Chichester: John Wiley & Sons. Stuttgart: Teubner, 1996.

[76] C.-T. Wu and H. C. Elman. Analysis and comparison of geometric and algebraic multigrid for convection-diffusion equations. *SIAM J. Sci. Comput.*, 28(6):2208–2228, 2006.

[77] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34(4):581–613, 1992.

[78] H. Yserentant. Old and new convergence proofs for multigrid methods. Iserles, A. (ed.), Acta Numerica 1993. Cambridge: Cambridge University Press, 285-326, 1993.