Über Gradientenverfahren zur Lösung von Eigenwertproblemen elliptischer Differentialoperatoren

Dissertation

zur

Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität Rostock

vorgelegt von

Ming Zhou, geb. am 09. Nov. 1978 in Kaiyuan, Provinz Liaoning, China

Wohnort: Rostock

Rostock, 03. Nov. 2011

urn:nbn:de:gbv:28-diss2012-0043-5

Gutachter: Prof. Dr. Klaus Neymeyr, Universität Rostock
Prof. Dr. Arnd Meyer, Technische Universität Chemnitz

Verteidigungsdatum: 10. Feb. 2012

Inhaltsverzeichnis

Danksagung

Mein besonderer Dank gilt meinen Eltern und meiner Frau für ihre stetige Unterstützung.

Zugleich danke ich meinen Kollegen der Arbeitsgruppe Numerische Mathematik für die schöne Arbeitsatmosphäre.

Bei meinem Doktorvater Prof. K. Neymeyr bedanke ich mich ganz herzlich für zahlreiche Anregungen und die angenehme Zusammenarbeit.

1 Einleitung

百物去其所與異，而從其所與同，故氣同則會，聲比則應，其驗皦然也。試調琴瑟而錯之，鼓其宮則他宮應之，鼓其商而他商應之，五音比而自鳴，非有神，其數然也。

……

天有陰陽，人亦有陰陽。天地之陰氣起，而人之陰氣應之而起，人之陰氣起，而天地之陰氣亦宜應之而起，其道一也。

……

其動以聲而無形，人不見其動之形，則謂之自鳴也。又相動無形，則謂之自然，其實非自然也，有使之然者矣。物故有實使之，其使之無形。

董仲舒【春秋繁露】

Dass die Dinge trotz ihrer Unterschiede *harmonisch* vereint werden können, ist wohlbekannt, zum Beispiel bei der Festlegung der *Tonhöhen von Saiteninstrumenten*. ... Die Natur besitzt die Zustände Yin und Yang, so auch der Mensch. Sie befinden sich oft im gleichen Zustand. ... Die sogenannte Spontaneität bezieht sich auf unersichtliche Bewegungen und Beziehungen wie *Schwingungen im Einklang*.

Dong Zhongshu, in „Chun-Qiu Fan-Lu“
(frei übersetzt)

Bereits vor über 2000 Jahren wurden von dem konfuzianischen Gelehrten Dong Zhongshu manche Schwingungszustände in der Natur als wesentliche Elemente beschrieben. Auch die vorliegende Arbeit widmet sich Schwingungen, allerdings in der Sprache der Mathematik als Eigenwertprobleme.

Heutzutage sind viele Eigenwertprobleme der selbstadjungierten und elliptischen Differentialoperatoren zweiter Ordnung in zahlreichen wissenschaftlichen und ingenieurtechnischen Anwendungen zu finden. Eigenwerte und Eigenfunktionen können dabei das oszillatorische Verhalten mechanischer und quantenmechanischer Systeme beschreiben. Häufig sind die Grundfrequenzen und die zugehörigen Schwingungsmoden von besonderem Interesse für die Analyse des Systemverhaltens. Solche Operatoreigenwertprobleme lassen sich nur unter besonderen Voraussetzungen analytisch lösen. In den meisten Fällen ist man auf numerische Verfahren zur approximativen Berechnung einiger der kleinsten Eigenwerte zusammen mit den entsprechenden Eigenfunktionen

angewiesen. Dabei werden typischerweise durch Finite-Differenzen-Diskretisierungen und Finite-Elemente-Diskretisierungen Matrixeigenwertprobleme für Matrizen beziehungsweise Matrixbüschel aufgestellt, welche weiter mit Verfahren der numerischen linearen Algebra zu lösen sind.

Numerische Lösungsverfahren für Matrixeigenwertprobleme stellen ein aktives und breites Forschungsgebiet dar. Das erste dazugehörende Verfahren wurde bereits im Jahr 1846 von Jacobi [37] vorgestellt. Dieses heute nach Jacobi benannte und immer noch aktuelle Verfahren beruht auf einer Folge von Ähnlichkeitstransformationen in Form planarer Drehungen, so dass eine Matrix sukzessiv auf Diagonalgestalt transformiert wird. Die gesuchten Eigenwerte können schließlich auf der Diagonalen abgelesen werden und die zugehörigen Eigenvektoren ergeben sich aus dem Produkt der Transformationen. Ab etwa Mitte des 20. Jahrhunderts mit der zunehmenden Verfügbarkeit elektronischer Rechenanlagen wurden zahlreiche weitere Verfahren für Matrixeigenwertprobleme entwickelt. Dabei sind die Variationen der Potenzmethode von Wielandt [93]–[98], das auf QR-Faktorisierungen basierende Verfahren von Francis [24, 25] und Kublanovskaya [52] und die von Lanczos [53] und Arnoldi [5] vorgestellten krylovraumartigen Verfahren besonders zu nennen. Einen schönen Überblick über die Entwicklungshistorie der Eigenwertproblemlöser im 20. Jahrhundert gibt die Arbeit von Golub und van der Vorst [30].

Numerische Lösungsverfahren für Matrixeigenwertprobleme sind oft auf bestimmte Problemstellungen zugeschnitten. So gibt es reelle und symmetrische sowie komplexe und nicht hermitesche Eigenwertprobleme, solche für kleine und dicht besetzte Matrizen sowie solche für große und dünn besetzte Matrizen. Eine große Auswahl an verfügbaren Verfahren findet man in „Templates for the solution of algebraic eigenvalue problems“ [6]. Wegen der Vielfältigkeit der Problemstellungen ist die Wahl der Verfahren sehr wichtig für eine effiziente Behandlung. Bei den Problemstellungen aus früheren Jahren handelt es sich oft um kleine und dicht besetzte Matrizen, und es sind sämtliche Eigenwerte beziehungsweise eine Normalform zu bestimmen. Solche Eigenwertprobleme lassen sich mit dem QR-Verfahren beziehungsweise dessen modifizierten Versionen in stabiler und schneller Weise numerisch lösen, siehe Wilkinson [99] sowie Golub und Uhlig [29]. Bei den oben genannten diskretisierten Eigenwertproblemen der Differentialoperatoren sind die Matrizen beziehungsweise die Matrizen der Matrixbüschel reell, symmetrisch, positiv definit sowie oft groß und dünn besetzt, und es sind nur einige der kleinsten Eigenwerte und zugehörige (reellwertige) Eigenvektoren, oder in anderen Worten, ein invarianter Unterraum zu den kleinsten Eigenwerten, gesucht. Man kann also mit Vektoriterationen einzelne Eigenvektoren oder mit Unterraumiterationen den invarianten Unterraum approximieren. Grundsätzlich sind krylovraumartige Verfahren für solche partiellen Eigenwertprobleme geeignet, denn ihre Anwendung ändert nicht die dünne Besetzungsstruktur der Matrizen und der Speicherbedarf bleibt damit kontrollierbar. Zu den bekannten Implementierungen zählen die Lanczos-Programme von Cullum und Willoughby [18, 19] und das Arnoldi-Programmpaket von Lehoucq, Sorensen und Yang [54]. Allerdings ist die Effizienz der krylovraumartigen Verfahren sehr problemabhängig, also abhängig von der Struktur des Spektrums. Gemäß den Theorien von Kaniel [38], Paige [77] und Saad [83] haben oft die inversen Krylovräume zur Lösung von obigen partiellen Eigenwertproblemen deutlich bessere Qualitäten als die Standard-Krylovräume. Diese benötigen jedoch exakte Lösungen von schlecht konditionierten Gleichungssystemen. Einige relativ moderne Verfahren, zum Beispiel das Jacobi-Davidson-Verfahren von Sleijpen und van der Vorst [90] und das vorkonditionierte CG-ähnliche Verfahren von Knyazev [42, 44], erlauben grobe Lösungen von Gleichungssystemen und verringern damit den gesamten Aufwand.

Im Gegensatz zur Verfahrensentwicklung ist die entsprechende Konvergenzanalyse ein bisher wenig entwickeltes Forschungsgebiet. Für viele moderne und effiziente Verfahren zur Lösung von Eigenwertproblemen elliptischer Differentialoperatoren, insbesondere gewisse vorkonditionierte Gradientenverfahren, können die vorhandenen Abschätzungen manches in der Praxis beobachtete Konvergenzverhalten nicht vernünftig reflektieren. Besonders wünschenswert sind die sogenannten scharfen Abschätzungen, wobei es Beispiele oder Grenzfälle gibt, mit denen die behaupteten Ungleichungen Gleichungen werden. Solche Abschätzungen sind bisher nur für einige relativ einfache Verfahren bekannt. Als neue Ergebnisse wurden kürzlich scharfe Abschätzungen für eine

Krylovraum-Iteration und ein vorkonditioniertes Gradientenverfahren mit optimaler Schrittweite aus der Untersuchung des Autors sowie der Zusammenarbeit mit Neymeyr und Ovtchinnikov gewonnen. Neu eingesetzte Beweistechniken, zum Beispiel eine Kurven-Differentiation und eine Ellipsen(Ellipsoide)-Analyse, ermöglichen einen klaren Beweisweg mit geometrischen Argumenten.

1.1 Das Eigenwertproblem und seine numerische Lösung

Für eine genaue Problemstellung der Arbeit betrachten wir zunächst bezüglich eines beschränkten und zusammenhängenden Gebiets Ω, welches eine Teilmenge von $\mathbb{R}^2$ oder $\mathbb{R}^3$ ist, das Eigenwertproblem eines selbstadjungierten und elliptischen Differentialoperators $\mathcal{L}$, welcher durch

$$\mathcal{L}\mathsf{u} := -\nabla \cdot \big(\mathsf{A}\nabla\mathsf{u}\big) + \mathsf{b}$$

mit den stetigen und reellwertigen Koeffizientenfunktionen A (Matrix) und b (Skalar) definiert wird. Mit homogener Dirichlet-Randbedingung oder einer Kombination von homogener Dirichlet-Randbedingung und homogener Neumann-Randbedingung sind einige der kleinsten Eigenwerte und zugehörige Eigenfunktionen zu ermitteln. Da Finite-Elemente-Diskretisierungen viel flexibler als Finite-Differenzen-Diskretisierungen sind, und Eigenwertprobleme für Matrixbüschel allgemeiner als Eigenwertprobleme für Matrizen sind, betrachten wir weiter eine Finite-Elemente-Diskretisierung (siehe Braess [14] für theoretische Grundlage und Modellprobleme).

Dabei lassen sich die gewünschten Eigenfunktionen gemäß einer Variationsformulierung durch stückweise stetig differenzierbare Funktionen aus einem Ansatzfunktionenraum $\mathcal{V}$, welcher ein endlichdimensionaler Unterraum des Sobolev-Raums $H^1(\Omega)$ ist, annähern. Jede Eigenfunktionsnäherung $\mathsf{v} \in \mathcal{V}$ soll mit der entsprechenden Eigenwertnäherung λ eine schwach formulierte Eigengleichung

$$\mathcal{B}(\mathsf{v}, \tilde{\mathsf{v}}) = \lambda(\mathsf{v}, \tilde{\mathsf{v}})_{L^2} \quad \forall\, \tilde{\mathsf{v}} \in \mathcal{V}$$

erfüllen, wobei $\mathcal{B}(\cdot,\cdot)$ die Bilinearform zum Operator $\mathcal{L}$ und $(\cdot,\cdot)_{L^2}$ das L^2-Skalarprodukt auf Ω bezeichnet. Gemäß der Eigenschaft von $\mathcal{L}$ und der gegebenen homogenen Randbedingung ist $\mathcal{B}(\cdot,\cdot)$ wie $(\cdot,\cdot)_{L^2}$ eine symmetrische und koerzive Bilinearform. Mittels einer beliebigen Basis von $\mathcal{V}$ lässt sich anschließend ein diskretes Eigenwertproblem bezüglich eines Matrixbüschels (A, M) mit symmetrischen und positiv definiten Matrizen $A, M \in \mathbb{R}^{n\times n}$ ($n := \dim \mathcal{V}$) aufstellen. (In manchen speziellen Fällen ist M nicht positiv definit. Dafür kann man das Eigenwertproblem durch eine Verschiebung des reziproken Spektrums umformulieren, so dass die zugehörigen Matrizen beide positiv definit sind.)

Mit den kleinsten Eigenwerten und zugehörigen Eigenvektoren von (A, M) lassen sich also die gewünschten Operatoreigenwerte und Eigenfunktionen näherungsweise ermitteln. Um gute Approximationen zu erhalten, wählt man oft einen Ansatzfunktionenraum mit sehr hoher Dimension, so dass die resultierenden Matrizen sehr groß sind. Die Besetzungsstruktur der Matrizen hängt von den gewählten Basisfunktionen ab. Typische Basisfunktionen sind zum Beispiel stückweise definierte lineare Polynome bezüglich einer Triangulierung des Gebiets. Damit werden dünn besetzte Matrizen erzeugt. In jeder Zeile wird die Anzahl der Nichtnull-Außerdiagonalkomponenten durch die Anzahl der benachbarten Gitterkanten des dieser Zeile entsprechenden Gitterpunkts von oben beschränkt. Außerdem ist der größte Eigenwert von (A, M) wegen der hohen Dimension des Ansatzfunktionenraums viel größer als die kleinsten Eigenwerte.

Für solche diskreten Eigenwertprobleme sind manche klassische Verfahren wenig geeignet. Zum Beispiel benötigen die Matrix-Transformationen wie das Jacobi-Verfahren und das QR-Verfahren eventuell eine Umformulierung in Eigenwertprobleme für Matrizen und können im Laufe des Verfahrens viele Nichtnull-Komponenten ergänzen, so dass die gegebene Besetzungsstruktur verschlechtert wird und mehr Speicherplatz benötigt wird. Außerdem ist es nicht nötig, alle Eigenwerte zu berechnen. Ein leicht durchführbares klassisches Verfahren ist die inverse Vektoriteration der Form $u' = A^{-1}Mu$. Dabei bleiben die Matrizen erhalten (damit sind matrixfreie Implementierungen

auch möglich) und die alte Iterierte u wird durch die neue Iterierte u' ersetzt, so dass kein zunehmender Speicherplatz nötig ist. Die Berechnung von $u' = A^{-1}Mu$ lässt sich durch das Lösen des linearen Gleichungssystems $Au' = Mu$ realisieren. Da A oft groß und dünn besetzt ist und eine große Konditionszahl hat, sind iterative und vorkonditionierte Verfahren dafür besonders geeignet. Betrachten wir zum Beispiel die Iteration $u'_{\text{neu}} = u'_{\text{alt}} - B^{-1}\big(Au'_{\text{alt}} - Mu\big)$ mit einem Vorkonditionierer B^{-1}, welcher als eine Näherungsinverse von A gilt. Man nennt B^{-1} Vorkonditionierer, denn mit einem guten B^{-1} hat die Matrix $B^{-1}A$ eine kleine Konditionszahl.

Für die Konstruktion des Vorkonditionierers stehen mehrere Verfahren zur Verfügung, zum Beispiel Mehrgitter-Verfahren, Multilevel-Verfahren und Verfahren der unvollständigen Matrix-Faktorisierungen, siehe Hackbusch [32], Yserentant [102], Bramble, Pasciak und Xu [16], Saad [85], Bollhöfer und Saad [12] für Details. Eine Historie der Vorkonditionierungstechnik zeigt die Arbeit von Benzi [9]. Solche Vorkonditionierungsverfahren lassen sich auch direkt mit einigen Lösungsverfahren für Eigenwertproblem kombinieren, siehe [33, 34, 55, 43, 48, 3, 87, 13] für Beispiele. Für diese effizienten Eigenwertproblemlöser gibt es jedoch kaum scharfe Abschätzungen zur Interpretation des Konvergenzverhaltens.

Als ein klassisches Verfahren besitzt die inverse Iteration schon seit langem eine im Wesentlichen geschlossene scharfe Konvergenztheorie. Ein bekannter Konvergenzfaktor ist der Quotient von zwei benachbarten Eigenwerten, siehe Parlett [79]. Damit kann man erklären, warum die inverse Vektoriteration viele Schritte benötigt, wenn die relevanten Eigenwerte dicht liegen. Natürlich kann man in diesem Fall eine Blockversion der Form $\mathcal{U}' = A^{-1}M\mathcal{U}$ (inverse Unterraumiteration) verwenden, welche „clusterrobust“ ist und deutlich weniger Schritte benötigen kann. Der zugehörige Konvergenzfaktor wurde ebenfalls in [79] vorgestellt, nämlich der Quotient von zwei möglicherweise gut getrennten Eigenwerten.

Diese klassischen Resultate bilden ein Muster der scharfen Konvergenzanalyse. Es wird darauf gehofft, dass gewisse mit der inversen Iteration verwandte vorkonditionierte Verfahren in ähnlicher Weise analysiert werden können. Zu bemerken ist, dass manche einfache Modifizierungen der inversen Vektoriteration keine vernünftigen vorkonditionierten Verfahren bringen. Betrachten wir zum Beispiel die oben erwähnte Implementierung der inversen Vektoriteration, wobei die Gleichungssysteme durch eine vorkonditionierte Iteration gelöst werden und jeder Schritt damit oft aus mehreren inneren Schritten besteht. Eine einfache Idee zur Beschleunigung ist, den Vorkonditionierer direkt in die äußere Schicht einzusetzen. Jedoch wird die resultierende Implementierung gegen einen Eigenvektor von $B^{-1}M$ konvergieren, welcher normalerweise kein Eigenvektor von $A^{-1}M$ beziehungsweise (A, M) ist.

Auch zu den verwandten Verfahren der inversen Vektoriteration gehört ein Typ von Gradientenverfahren, wobei es möglich ist, dass das einfache Ersetzen von A^{-1} durch B^{-1} zu einem gegen einen Eigenvektor von (A, M) konvergenten Verfahren führt. Im Folgenden werden die Nullvektoren und die Einheitsmatrizen unabhängig von deren Dimensionen mit $\mathbf{0}$ beziehungsweise I bezeichnet.

1.2 Gradientenverfahren für den Rayleigh-Quotienten

Die oben genannten Gradientenverfahren beruhen darauf, dass der kleinste Eigenwert von (A, M) mit symmetrischen und positiv definiten Matrizen $A, M \in \mathbb{R}^{n\times n}$ identisch zum Minimum der Funktion

$$\rho \;:\; \mathbb{R}^n\backslash\{\mathbf{0}\} \to \mathbb{R}, \qquad u \mapsto \frac{u^T Au}{u^T Mu} \tag{1.1}$$

namens Rayleigh-Quotient ist. Damit lässt sich die Berechnung des kleinsten Eigenwerts durch die Minimierung des Rayleigh-Quotienten realisieren, wofür mehrere Gradientenverfahren zur Verfügung stehen. Bezüglich verschiedener Skalarprodukte lassen sich verschiedene Gradienten für den Rayleigh-Quotienten definieren. Der Standard-Gradient

$$\nabla\rho \;:\; \mathbb{R}^n\backslash\{\mathbf{0}\} \to \mathbb{R}^n, \qquad u \mapsto \frac{2}{u^T Mu}\big(Au - \rho(u)Mu\big)$$

bezieht sich auf das Euklidische Skalarprodukt. Mit dem durch eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ induzierten Skalarprodukt lässt sich ein „H-Gradient“ durch $H^{-1}\nabla\rho$ definieren. Sofern $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ kein Eigenvektor von (A, M) ist, sind $\nabla\rho(u)$ und das Residuum

$$r(u) := Au - \rho(u)Mu \tag{1.2}$$

kollineare Nichtnull-Vektoren. Damit lässt sich ein „H-Gradientenverfahren“ mit $-H^{-1}r(u)$ als Suchrichtung konstruieren. Dazu kann man auf dem durch u und $H^{-1}r(u)$ aufgespannten Unterraum eine Rayleigh-Ritz-Prozedur durchführen, so dass der Rayleigh-Quotient $\rho(u')$ der neuen Iterierten u' möglichst klein ist. Diese Vorgehensweise lässt sich durch die Vorschrift

$$u' = \mathrm{RR}_{\rho,\min}\big(\mathrm{span}\{u,\ H^{-1}r(u)\}\big) \tag{1.3}$$

beschreiben, wobei die Rayleigh-Ritz-Prozedur $\mathrm{RR}_{\rho,\min}$ zum Rayleigh-Quotienten $\rho(\cdot)$ einen Ritzvektor zum kleinsten Ritzwert des Matrixbüschels (A, M) bezüglich des betrachteten Unterraums liefert.

Gemäß unserer Problemstellung in Abschnitt 1.1 ist der größte Eigenwert von (A, M) viel größer als die kleinsten Eigenwerte. In diesem Fall führt der Gradient zum durch A induzierten Skalarprodukt im Sinne der Anzahl der (äußeren) Schritte zu relativ schnell konvergenten Verfahren. Solche A-Gradientenverfahren lassen sich als Modifizierungen der inversen Vektoriteration interpretieren. Betrachten wir insbesondere das Verfahren (1.3) mit $H = A$. Der zugehörige Unterraum $\mathrm{span}\{u,\ A^{-1}r(u)\}$ ist identisch zum Krylovraum $\mathrm{span}\{u,\ A^{-1}Mu\}$ und hat eine bessere Qualität als $A^{-1}Mu$ zur Verkleinerung des Rayleigh-Quotienten. Das entsprechende vorkonditionierte Gradientenverfahren, also (1.3) mit $H = B$, lässt sich als eine gestörte Krylovraum-Iteration interpretieren und konvergiert mit einem hinreichend guten Vorkonditionierer trotz der Störung gegen einen Eigenvektor von (A, M). Es benötigt im Vergleich zu (1.3) mit $H = A$ oft einen geringeren gesamten Aufwand und in manchen speziellen Fällen sogar weniger äußere Schritte (1-Schritt-Konvergenz möglich). Die Qualität des Vorkonditionierers lässt sich (eventuell nach einer Umskalierung) gemäß einer bezüglich des Spektralradius $\mathscr{R}(\cdot)$ gestellten Bedingung

$$\mathscr{R}(I - B^{-1}A) \le \gamma < 1 \tag{1.4}$$

kontrollieren, welche auch eine nützliche Information für die Konvergenzanalyse anbietet.

Die Forschung über die Konvergenz der vorkonditionierten Gradientenverfahren sowie deren Blockversionen (Unterraumiterationen) zur Lösung von Matrixeigenwertproblemen begann vor etwa 50 Jahren mit der Arbeit von Samokish [86] im funktionalanalytischen Kontext, siehe D'yakonov [22] und Meyer [57] für einen Überblick über klassische Resultate sowie Knyazev [41], Bramble, Pasciak und Knyazev [15], Neymeyr [62] und Ovtchinnikov [73] für weitere Resultate. Obwohl die vorkonditionierten Gradientenverfahren bereits in [86, 22] in Bezug auf eine „B-Geometrie“ interpretiert wurden, konnte man mit solcher Interpretation nicht sehr klar begründen, warum diese Verfahren oft viel bessere Leistungen als die entsprechenden Euklidischen Gradientenverfahren haben. Eine Schwierigkeit dabei ist die geometrische Darstellung der Niveaumengen des Rayleigh-Quotienten. Nur für spezielle und niedrigdimensionale Fälle kann man diese sämtlich durch elliptische Kegel darstellen. Damit merkt man, dass ein solcher Kegel unter der durch einen guten Vorkonditionierer induzierten Geometrie fast ein Kreiskegel werden kann und die B-Gradientenvektoren als Normalvektoren die eine kleine Umgebung der Kegelachse erreichenden Richtungen besitzen können. Da der Ziel-Eigenvektor kollinear zur Kegelachse ist, ermöglichen die B-Gradientenvektoren eine schnellere Konvergenz. Eine Analogie ist die geometrische Betrachtung der vorkonditionierten Gradientenverfahren für die aus Randwertproblemen der elliptischen Differentialoperatoren hergeleiteten linearen Gleichungssysteme. Natürlich handelt es sich dabei um ein relativ einfaches Thema, denn die Niveaumengen der zugehörigen quadratischen Funktion lassen sich auch im Allgemeinen durch ähnliche Ellipsoide darstellen, welche gleiche Verhältnisse der Halbachsen besitzen.

Wir betrachten weiter vorkonditionierte Gradientenverfahren für Matrixeigenwertprobleme und interessieren uns für scharfe Konvergenzabschätzungen, welche auch für allgemeinere Fälle verwendbar sind und genaue analytische Formen besitzen, sowie keine unnatürlichen beziehungsweise

zugunsten der Analyse zusätzlich konstruierten Voraussetzungen benötigen. Der Ausgangspunkt für dieses Thema ist die Abschätzung von Neymeyr [60, 61] für das Verfahren

$$u' = u - B^{-1}r(u) \tag{1.5}$$

namens „Preconditioned inverse vector iteration“ (PINVIT). Dem Namen entsprechend wurde das Verfahren als eine gestörte inverse Vektoriteration interpretiert. Damit lässt sich eine Kugel definieren, welche alle möglichen u' unter der Bedingung (1.4) umfasst. Die Konvergenzanalyse basiert auf einem zweistufigen Optimierungsproblem für den Rayleigh-Quotienten. Die resultierenden Konvergenzfaktoren sind unabhängig von dem größten Eigenwert beziehungsweise von der Dimension der Diskretisierung und reflektieren eine „Diskretisierungsrobustheit“ des Verfahrens.

Durch Neymeyr [62] wurden weitere vorkonditionierte Gradientenverfahren als Verallgemeinerungen von (1.5) betrachtet und damit eine PINVIT-Hierarchie aufgestellt.

- PINVIT(1) ist das grundlegende Verfahren (1.5) und im Wesentlichen ein B-Gradientenverfahren mit fester Schrittweite.
- PINVIT(2) benutzt die PINVIT(1)-Iterierte, um einen zweidimensionalen Unterraum aufzustellen und darauf eine Rayleigh-Ritz-Prozedur durchzuführen

  $$u' = \mathrm{RR}_{\rho,\min}\big(\mathrm{span}\{u,\, u - B^{-1}r(u)\}\big).$$

  PINVIT(2) ist identisch zum Gradientenverfahren (1.3) mit $H = B$.
- PINVIT(k), $k > 2$, benutzt die PINVIT(1)-Iterierte und die weiteren $k-2$ vorherigen Iterierten $u_{2-k}, \ldots, u_{-1}$, um einen k-dimensionalen Unterraum aufzustellen und darauf eine Rayleigh-Ritz-Prozedur durchzuführen

  $$u' = \mathrm{RR}_{\rho,\min}\big(\mathrm{span}\{u_{2-k},\, \ldots,\, u_{-1},\, u,\, u - B^{-1}r(u)\}\big).$$

  Für $k > 2$ ist PINVIT(k) ein „Mehrschritt-Verfahren“ (vergleiche Mehrschritt-Verfahren für Anfangswertprobleme) und benötigt andere Verfahren, zum Beispiel PINVIT($k-1$), um die Anfangsiterierten zu erzeugen.

Zur Konvergenzanalyse für PINVIT-Verfahren können wir zunächst die „exakt vorkonditionierte Versionen“, das heißt PINVIT-Verfahren mit $B^{-1} = A^{-1}$, untersuchen, um gewisse Anregungen zu erhalten. Diese Spezialversionen gelten als Verallgemeinerungen der inversen Vektoriteration und bilden daher eine „Inverse vector iteration“(INVIT)-Hierarchie.

- INVIT(1) ist eine skalierte inverse Vektoriteration

  $$u' = \rho(u)A^{-1}Mu.$$

- INVIT(2) ist ein Krylovraum-Verfahren

  $$u' = \mathrm{RR}_{\rho,\min}\big(\mathrm{span}\{u,\, A^{-1}Mu\}\big)$$

  und identisch zum Gradientenverfahren (1.3) mit $H = A$.
- INVIT(k), $k > 2$, ist ein Mehrschritt-Verfahren

  $$u' = \mathrm{RR}_{\rho,\min}\big(\mathrm{span}\{u_{2-k},\, \ldots,\, u_{-1},\, u,\, A^{-1}Mu\}\big).$$

1.3 Inhaltsüberblick

In dieser Arbeit werden einige neue Ergebnisse aus der Untersuchung des Autors sowie der Zusammenarbeit mit Neymeyr und Ovtchinnikov über scharfe Konvergenzabschätzungen für Gradientenverfahren im Rahmen der (P)INVIT-Hierarchie präsentiert.

Als Vorbereitung werden in Kapitel 2 einige grundlegende Argumente und anschließend einige einfache Konvergenzaussagen für die betrachteten Gradientenverfahren vorgestellt. Für INVIT(1) lässt sich eine klassische Konvergenzabschätzung für die inverse Vektoriteration verwenden. Darauf basierend formulieren wir eine Musteranalyse und definieren damit die Maße der Konvergenzanalyse für die weiteren Verfahren.

Für INVIT(2) wurde in [62] eine klassische Konvergenzabschätzung mit der Beweistechnik von Knyazev und Skorokhodov [51] gezeigt. In der gemeinsamen Arbeit des Autors mit Neymeyr und Ovtchinnikov [68] wurden weitere Problemvarianten betrachtet und gleichartige Konvergenzabschätzungen vorgestellt. Dabei wurden außerdem eine Abschätzungsverschärfung und eine Beweisvereinfachung präsentiert. In Kapitel 3 wird in Bezug auf [68] eine weiter modifizierte Konvergenzanalyse für INVIT(2) und anschließend eine neue Konvergenzanalyse für ein allgemeineres Verfahren, nämlich eine Iteration der Krylovräume niedriger Dimensionen, formuliert.

Für PINVIT(1) wurde eine klassische Konvergenzabschätzung, wie oben erwähnt, in [60, 61] präsentiert. Durch Knyazev und Neymeyr [49, 50] und Argentati, Knyazev, Neymeyr und Ovtchinnikov [4] wurde diese umformuliert beziehungsweise in einfacherer Weise bewiesen. Für PINVIT(2) wurden in [62] einige geometrische Argumente und Vermutungen als Vorbereitungen der Analyse vorgestellt. In Kapitel 4 wird gemäß [50, 4] eine weiter modifizierte Konvergenzanalyse für PINVIT(1) und anschließend eine neue Konvergenzanalyse für PINVIT(2) formuliert.

Zur simultanen Berechnung von mehreren Eigenwerten und Eigenvektoren kann man Blockversionen der (P)INVIT-Verfahren verwenden. Für blockweise PINVIT(1) wurden durch Neymeyr [63] die Resultate für PINVIT(1) aus [60, 61] verallgemeinert. Jedoch kann diese verallgemeinerte Abschätzung zu grob sein, wenn die relevanten Eigenwerte zu dicht liegen. Das reflektiert nicht die beobachtete „Clusterrobustheit“ des Verfahrens. Eine andere Abschätzung von Bramble, Pasciak und Knyazev [15] hat ein solches Problem beseitigt, benötigt aber eine sehr technische Voraussetzung. Zu diesem Thema werden in Kapitel 5 einige Zwischenergebnisse der gewünschten Abschätzungen vorgestellt.

In Kapitel 6 werden die Ergebnisse zusammengefasst und zukünftige Aufgaben gestellt.

Im Anhang behandeln wir einige Modellprobleme mit den untersuchten Verfahren und vergleichen deren Konvergenzverhalten in Bezug auf Clusterrobustheit.

2 Theoretische Grundlagen

Dieses Kapitel ist eine Vorbereitung unserer Konvergenzanalyse für die Gradientenverfahren aus der (P)INVIT-Hierarchie. In Abschnitt 2.1 werden in Bezug auf die Problemstellung aus Abschnitt 1.1 einige häufig benutzte Begriffe und Argumente formuliert, womit wir gleich die einfache Konvergenz der betrachteten Verfahren im Sinne der Eigenwert-Approximation beziehungsweise der Eigenraum-Approximation zeigen können. In Abschnitt 2.2 wird die inverse Vektoriteration beziehungsweise INVIT(1) als Modell betrachtet, um die Struktur unserer Konvergenzanalyse vorzustellen. Es wird zunächst eine klassische Abschätzung zur Eigenraum-Approximation etwas allgemeiner formuliert. Eine Folgerung dieser Abschätzung wird weiter für eine Abschätzung zur Eigenwert-Approximation verwendet, nachdem eine niedrigdimensionale Problemstellung zur langsamsten Konvergenz durch eine Kurven-Differentiation formuliert worden ist.

2.1 Einfache Konvergenzaussagen

Gemäß der Problemstellung aus Abschnitt 1.1 formulieren wir zunächst eine allgemeine Voraussetzung.

Vorausetzung 2.1. *Betrachten wir zwei symmetrische und positiv definite Matrizen A, M aus $\mathbb{R}^{n\times n}$. Diese bilden ein Matrixbüschel (A, M), welches die Eigenwerte $\lambda_1 < \lambda_2 < \cdots < \lambda_m$ mit zugehörigen Vielfachheiten $q_1, q_2, \ldots, q_m$ und zugehörigen Eigenräumen $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m$ besitze.*

Zur Beschreibung der Orthogonalitätsrelation sowie der Approximation der Eigenräume verwenden wir einige Begriffe zu symmetrischen und positiv definiten Matrizen.

Definition 2.2 (mit einigen wohlbekannten Aussagen)**.** *Durch eine symmetrische und positiv definite Matrix H aus $\mathbb{R}^{n\times n}$ lässt sich ein Skalarprodukt auf $\mathbb{R}^n$ wie folgt induzieren*

$$(x, y)_H := x^T H y \quad \textit{für} \quad x, y \in \mathbb{R}^n.$$

Man nennt $(x, y)_H$ das H-Skalarprodukt von x und y. Basierend auf dem H-Skalarprodukt lassen sich verwandte Begriffe wie H-Orthogonalität, H-Vektornorm, H-Operatornorm und H-Orthonormalität definieren.

Weiter wird die H-orthogonale Projektion eines Vektors $x \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ auf einem Unterraum $\mathcal{Y}$ von $\mathbb{R}^n$, $\mathcal{Y} \neq \{\mathbf{0}\}$, durch die eindeutige Lösung des Minimierungsproblems

$$\|x - y\|_H \;\rightarrow\; \min \quad \textit{über} \quad y \in \mathcal{Y}$$

bezüglich der H-Vektornorm gegeben. Da im Folgenden keine nicht-orthogonale Projektionen untersucht werden, nennen wir H-orthogonale Projektionen der Einfachheit halber H-Projektionen. Mit beliebiger Basismatrix Y von $\mathcal{Y}$ lässt sich die H-Projektion von x auf $\mathcal{Y}$ durch

$$Y(Y^T H Y)^{-1} Y^T H x$$

darstellen. Die Matrix $Y(Y^T HY)^{-1}Y^T H$ ist unabhängig von der Wahl der Basismatrix und heißt die H-Projektionsmatrix von $\mathcal{Y}$. Ein Vektor $\tilde{y} \in \mathcal{Y}$ ist genau dann die H-Projektion von x auf $\mathcal{Y}$, wenn $x - \tilde{y}$ H-orthogonal zu $\mathcal{Y}$ ist.

Der H-Winkel zwischen zwei Vektoren $x, y \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ lässt sich durch

$$\angle_H(x,y) := \arccos\left(\frac{(x,y)_H}{\|x\|_H\,\|y\|_H}\right)$$

definieren. Für einen Vektor $x \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ und einen Unterraum $\mathcal{Y}$ von $\mathbb{R}^n$, $\mathcal{Y} \neq \{\mathbf{0}\}$, heißt

$$\angle_H(x,\,\mathcal{Y}) := \min_{y\in\mathcal{Y}\backslash\{\mathbf{0}\}} \angle_H(x,y)$$

der H-Winkel zwischen x und $\mathcal{Y}$. Mit der H-Projektion $\tilde{y}$ von x auf $\mathcal{Y}$ gilt

$$\sin\angle_H(x,\,\mathcal{Y}) = \frac{\|x-\tilde{y}\|_H}{\|x\|_H}, \quad \cos\angle_H(x,\,\mathcal{Y}) = \frac{\|\tilde{y}\|_H}{\|x\|_H}, \quad \tan\angle_H(x,\,\mathcal{Y}) = \frac{\|x-\tilde{y}\|_H}{\|\tilde{y}\|_H}.$$

Im Fall $H = I$ stimmen diese Begriffe mit den entsprechenden Euklidischen Begriffen überein. Dafür wird 2 statt I als der untere Index bei der Bezeichnung verwendet.

Gemäß Voraussetzung 2.1 und Definition 2.2 sind die Eigenräume $\mathcal{V}_i$ von (A, M) sowohl A-orthogonal als auch M-orthogonal zueinander, und deren direkte Summe ist identisch zu $\mathbb{R}^n$. Auf jedem Eigenraum stimmt die zugehörige A-Projektionsmatrix mit der zugehörigen M-Projektionsmatrix überein, denn mit einer beliebigen Basismatrix V_i gilt

$$V_i(V_i^TAV_i)^{-1}V_i^TA = V_i\big(V_i^T(AV_i)\big)^{-1}(AV_i)^T = V_i\big(V_i^T(\lambda_iMV_i)\big)^{-1}(\lambda_iMV_i)^T = V_i(V_i^TMV_i)^{-1}V_i^TM.$$

Daher lässt sich jedes $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ bezüglich der Eigenräume entwickeln, das heißt, es gibt Vektoren $v_i \in \mathcal{V}_i$ mit $u = \sum_{i=1}^m v_i$. Diese Vektoren v_i sind sowohl A-Projektionen als auch M-Projektionen und sind eindeutig bestimmt. Wenn eine Vektoriteration gegen einen Eigenvektor zu λ_1 konvergiert, können wir das Konvergenzverhalten der Vektoriterierten beispielsweise mittels des Sinuswerts

$$\sin\angle_M(u,\,\mathcal{V}_1) = \frac{\|\sum_{i=2}^m v_i\|_M}{\|u\|_M} = \frac{\sqrt{\sum_{i=2}^m \|v_i\|_M^2}}{\|u\|_M}$$

beschreiben. Alternativ können wir das Konvergenzverhalten der entsprechenden Eigenwertnäherungen mittels des in (1.1) definierten Rayleigh-Quotienten untersuchen. Für $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ mit der Darstellung $u = \sum_{i=1}^m v_i$ gilt

$$\rho(u) = \frac{u^TAu}{u^TMu} = \frac{\sum_{i=1}^m\sum_{j=1}^m v_i^TAv_j}{\sum_{i=1}^m\sum_{j=1}^m v_i^TMv_j} = \frac{\sum_{i=1}^m \lambda_i\|v_i\|_M^2}{\sum_{i=1}^m \|v_i\|_M^2},$$

das heißt, $\rho(u)$ ist eine Konvexkombination aller Eigenwerte. Außerdem ist $\rho(u)$ wegen der positiven Definitheit von A und M positiv. Daher gilt

$$0 < \min_{u\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \rho(u) = \lambda_1 < \lambda_2 < \cdots < \lambda_m = \max_{u\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \rho(u).$$

Mit der Einführung des Rayleigh-Quotienten lässt sich die Aufgabe der Berechnung des kleinsten Eigenwerts sowie eines zugehörigen Eigenvektors in ein Minimierungsproblem des Rayleigh-Quotienten umformulieren und anschließend durch verschiedene Gradientenverfahren behandeln. Betrachten wir in kompakter Weise die Gradientenverfahren zum durch eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ induzierten Skalarprodukt. Mit dem Standard-Gradienten $\nabla\rho$ des Rayleigh-Quotienten lässt sich der Gradient zum H-Skalarprodukt wegen

$$\big(H^{-1}\nabla\rho(u)\,,\,v\big)_H = \big(\nabla\rho(u)\,,\,v\big)_2 \qquad \forall\, v \in \mathbb{R}^n\backslash\{\mathbf{0}\}$$

durch $H^{-1}\nabla\rho$ definieren. Gemäß der Kollinearität zwischen dem Standard-Gradienten und dem in (1.2) definierten Residuum kann man mit $-H^{-1}r(u)$ als Suchrichtung ein einfaches Gradientenverfahren konstruieren. Um eine Schrittweite zu möglichst großer Reduktion des Rayleigh-Quotienten

zu ermitteln, lässt sich ein lokales Minimierungsproblem des Rayleigh-Quotienten auf dem Unterraum $\mathcal{Y} := \mathrm{span}\{u,\, H^{-1}r(u)\}$ aufstellen. Für eine beliebige Minimalstelle $y \in \mathcal{Y}$ ist dann (gemäß der Euler-Lagrange-Gleichung) mit den bezüglich beliebiger Vektoren $\tilde{y} \in \mathcal{Y}$ definierten Funktionen $f_{\tilde{y}} : \mathbb{R} \to \mathbb{R},\ \alpha \mapsto \rho(y + \alpha\tilde{y})$ eine notwendige Bedingung $(df_{\tilde{y}}/d\alpha)(0) = 0 \ \ \forall\ \tilde{y} \in \mathcal{Y}$, welche wegen $(df_{\tilde{y}}/d\alpha)(\alpha) = \big(\nabla\rho(y + \alpha\tilde{y}), \tilde{y}\big)_2$ äquivalent zu

$$\big(\nabla\rho(y), \tilde{y}\big)_2 = 0 \qquad \forall\ \tilde{y} \in \mathcal{Y}$$

ist. Mit einer beliebigen Basismatrix Y von $\mathcal{Y}$ und der Darstellung $y = Yc$ wird dann ein „lokales“ Eigenwertproblem

$$(Y^T AY)c = \rho(y)(Y^T MY)c$$

aufgestellt. Damit bilden $\rho(y)$ und y ein Ritzpaar von (A, M) bezüglich des Unterraums $\mathcal{Y}$.

Definition 2.3. *Betrachten wir eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$, eine symmetrische und positiv definite Matrix $M \in \mathbb{R}^{n\times n}$ und einen s-dimensionalen Unterraum $\mathcal{U}$ von $\mathbb{R}^n$, $\mathcal{U} \neq \{\mathbf{0}\}$. Mit einer beliebigen Basismatrix U von $\mathcal{U}$ heißt ein Vektor $u = Uc$ mit $c \in \mathbb{R}^s\backslash\{\mathbf{0}\}$ ein Ritzvektor zum Ritzwert $\theta \in \mathbb{R}$ des Matrixbüschels (A, M) bezüglich $\mathcal{U}$, falls c und θ die Gleichung*

$$(U^T AU)c = \theta(U^T MU)c$$

erfüllen. Für den in (1.1) definierten Rayleigh-Quotienten und das in (1.2) definierte Residuum gilt

$$\rho(u) = \theta, \qquad r(u) \perp_2 \mathcal{U}.$$

Im Fall $M = I$ heißen θ, u auch Ritzwert und Ritzvektor der Matrix A.

Bezüglich trivial angeordneter Ritzwerte (das heißt, mehrfache Ritzwerte werden mit ihrer Vielfachheit gezählt) lässt sich eine Verallgemeinerung des Courant-Fischer-Prinzips formulieren.

Satz 2.4. *Mit der Formulierung von Definition 2.3 gilt für die Ritzwerte $\theta_1 \le \theta_2 \le \cdots \le \theta_s$ von (A, M) bezüglich $\mathcal{U}$ und eine beliebige Basismatrix $U \in \mathbb{R}^{n\times s}$ von $\mathcal{U}$, dass*

$$\theta_i = \min_{\substack{C\in\mathbb{R}^{s\times i}\\ Rang(C)=i}} \max_{u\in\mathrm{span}\{UC\}\backslash\{\mathbf{0}\}} \rho(u), \qquad \theta_{s+1-i} = \max_{\substack{C\in\mathbb{R}^{s\times i}\\ Rang(C)=i}} \min_{u\in\mathrm{span}\{UC\}\backslash\{\mathbf{0}\}} \rho(u) \qquad \forall\ i \in \{1,\ldots,s\}.$$

Es gilt insbesondere

$$\theta_1 = \min_{c\in\mathbb{R}^s\backslash\{\mathbf{0}\}} \rho(Uc) = \min_{u\in\mathcal{U}\backslash\{\mathbf{0}\}} \rho(u), \qquad \theta_s = \max_{c\in\mathbb{R}^s\backslash\{\mathbf{0}\}} \rho(Uc) = \max_{u\in\mathcal{U}\backslash\{\mathbf{0}\}} \rho(u).$$

Beweis. Analog zum Beweis des Courant-Fischer-Prinzips, siehe zum Beispiel [79]. Der Schlüssel ist ein Dimensionsvergleich. □

Das obige Minimierungsproblem des Rayleigh-Quotienten auf $\mathcal{Y}$ lässt sich also durch einen beliebigen Ritzvektor zum kleinsten Ritzwert lösen. Durch die zugehörige Berechnung, genannt „Rayleigh-Ritz-Prozedur“, wird das optimierte H-Gradientenverfahren der Form (1.3) realisiert. (Es ist nicht für jede symmetrische und positiv definite Matrix H möglich, die Vorschrift (1.3) in $u' = u - \tau_{\mathrm{opt}}\, H^{-1}r(u)$ mit einer optimalen Schrittweite $\tau_{\mathrm{opt}} \in \mathbb{R}$ umzuschreiben. Man kann gewisse Beispiele konstruieren, in denen $H^{-1}r(u)$ ein Ritzvektor zum kleinsten Ritzwert ist und die entsprechende Schrittweite damit nicht durch eine reelle Zahl beschrieben werden kann.) Gemäß unserer Problemstellung betrachten wir (1.3) weiter insbesondere für $H = A$ und $H = B$, wobei B^{-1} ein Vorkonditionierer von A ist und die Qualitätsbedingung (1.4) erfüllt. Zur Konvergenzanalyse für diese zwei Varianten können wir zunächst im Vergleich zu (1.3) mit $H = I$ einige einfache Aussagen zeigen.

Lemma 2.5. *Betrachten wir (A, M) gemäß Voraussetzung 2.1. Falls $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ kein Eigenvektor von (A, M) ist, dann hat* $\operatorname{span}\{u, r(u)\}$ *die Dimension 2 und u ist kein Ritzvektor von (A, M) bezüglich* $\operatorname{span}\{u, r(u)\}$*. Außerdem liefert (1.3) mit $H = I$ eine Rayleigh-Quotient-Verkleinerung $\rho(u') < \rho(u)$.*

Beweis. Es gilt $r(u) \neq \mathbf{0}$ und $r(u)$ liegt wegen

$$\big(u,\, r(u)\big)_2 = u^T\big(Au - \rho(u)Mu\big) = u^TAu - \rho(u)u^TMu = 0$$

Euklidisch orthogonal zu u, so dass u und $r(u)$ linear unabhängig sind und $\operatorname{span}\{u, r(u)\}$ damit die Dimension 2 hat. Wäre u ein Ritzvektor, würde nach Definition 2.3 $r(u) \perp_2 \operatorname{span}\{u, r(u)\}$ gelten. Durch die Implikation

$$\big(r(u),\, r(u)\big)_2 = 0 \quad \Rightarrow \quad r(u) = \mathbf{0} \quad \Rightarrow \quad u \text{ ist Eigenvektor von } (A, M)$$

erhalten wir einen Widerspruch zur Voraussetzung. Daher ist u kein Ritzvektor.

Für (1.3) mit $H = I$ gilt zunächst $\rho(u') \leq \rho(u)$, denn u' ist eine Minimalstelle des Rayleigh-Quotienten über $\operatorname{span}\{u, r(u)\}$. Wäre $\rho(u') = \rho(u)$, dann wäre u ebenfalls eine Minimalstelle. Mittels der Euler-Lagrange-Gleichung (siehe Argumente vor Definition 2.3) würde dann gezeigt, dass u ein Ritzvektor ist. Dies widerspricht der vorherigen Aussage. □

Deshalb wird das Verfahren (1.3) mit $H = I$ ausgehend von einem Nichtnullvektor entweder bei einem Eigenvektor zu einem Zwischen-Eigenwert terminieren oder eine monoton fallende Folge des Rayleigh-Quotienten liefern, welche gegen den kleinsten Eigenwert λ_1 konvergiert. Im zweiten Fall lässt sich anschließend die Konvergenz der Vektoriterierten gegen einen Eigenvektor zu λ_1 zeigen.

Lemma 2.6. *Betrachten wir (A, M) gemäß Voraussetzung 2.1. Für beliebiges $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ gilt*

$$\sin^2 \angle_M(u,\, \mathcal{V}_1) \leq \frac{\rho(u) - \lambda_1}{\lambda_2 - \lambda_1}.$$

Im Fall $\|u\|_M = 1$ wird der quadratische Abstand von u zu $\mathcal{V}_1$ bezüglich der M-Vektornorm durch dieselbe Schranke von oben beschränkt.

Beweis. Mit einer Entwicklung $u = \sum_{i=1}^m v_i$ bezüglich der Eigenräume von (A, M) gilt

$$\sin^2 \angle_M(u,\, \mathcal{V}_1) = \frac{\sum_{i=2}^m \|v_i\|_M^2}{\|u\|_M^2} \leq \frac{\sum_{i=1}^m \frac{\lambda_i - \lambda_1}{\lambda_2 - \lambda_1}\|v_i\|_M^2}{\|u\|_M^2} = \frac{\|u\|_A^2 - \lambda_1\|u\|_M^2}{(\lambda_2 - \lambda_1)\|u\|_M^2} = \frac{\rho(u) - \lambda_1}{\lambda_2 - \lambda_1}.$$

Da der quadratische „M-Abstand“ durch $\sum_{i=2}^m \|v_i\|_M^2$ gegeben ist, wird dieser im Fall $\|u\|_M = 1$ ebenfalls durch $\big(\rho(u) - \lambda_1\big)/\big(\lambda_2 - \lambda_1\big)$ beschränkt. □

Falls die durch (1.3) mit $H = I$ konstruierten Vektoriterierten bezüglich der M-Vektornorm normiert werden, garantiert die Rayleigh-Quotient-Verkleinerung die Konvergenz der Vektoriterierten gegen einen Eigenvektor zu λ_1. Zur genaueren Beschreibung des Konvergenzverhaltens dieses Gradientenverfahrens kann man die Abschätzungen wie Theorem 2 von Longsine und McCormick [56] sowie Theorem 3.4 von Golub und Ye [31] und im Fall $M = I$ auch manche frühere Abschätzungen von Hestenes und Karush [35], Kantorovich [39], Kantorovich und Akilov [40] anwenden. Damit kann man zeigen, dass dieses Verfahren gemäß unserer Problemstellung eine schlechte Konvergenzgeschwindigkeit hat. Dass der größte Eigenwert viel größer als die kleinsten Eigenwerte ist, führt etwa zu einem nahe 1 liegenden Konvergenzfaktor, siehe auch [81, 80, 51, 101, 105, 106] für ähnliche Argumente. Mit der Inversen beziehungsweise einer geeigneten Näherungsinversen der Matrix A lässt sich dieses Verfahren beschleunigen. Die resultierenden Verfahren sind genau die Varianten von (1.3) mit $H = A$ beziehungsweise $H = B$.

Eine Analogie für solche Beschleunigungen bezieht sich auf das Gradientenverfahren mit optimaler Schrittweite zum Lösen eines linearen Gleichungssystems $A\bar{x} = b$ zu einer symmetrischen und positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$. Dieses Verfahren basiert auf dem Gradienten der quadratischen Funktion $f(x) := \frac{1}{2}\, x^T Ax - b^T x$. Für A mit einer großen spektralen Konditionszahl hat der Gradientenvektor $\nabla f(x) = Ax - b$ oft eine große Richtungsabweichung zum aktuellen Fehler $x - \bar{x}$ und führt zu einer langsamen Konvergenz. Die durch A^{-1} modifizierte Suchrichtung $A^{-1}\nabla f(x)$ und die dafür optimale Schrittweite 1 ermöglichen eine 1-Schritt-Konvergenz, denn $x - \bar{x} = A^{-1}(Ax - b)$. Diese einfache Tatsache lässt sich auch unter geometrischem Aspekt interpretieren. Beim obigen Gradientenverfahren ist die Niveaumenge $\{\tilde{x} \in \mathbb{R}^n\,;\, f(\tilde{x}) = f(x)\}$ ein Ellipsoid um $\bar{x}$ und ähnlich zur Einheitskugel der A-Vektornorm im Sinne der Halbachsen-Verhältnisse, so dass der Fehlervektor $x - \bar{x}$ A-orthogonal zur Tangentialhyperebene des Ellipsoids an x liegt. Gemäß den Eigenschaften der Niveaumenge liegt der Gradientenvektor $\nabla f(x)$ Euklidisch orthogonal zur Tangentialhyperebene. Deswegen sind $x - \bar{x}$ und $A^{-1}\nabla f(x)$ kollinear, so dass man entlang der Richtung $A^{-1}\nabla f(x)$ die exakte Lösung $\bar{x}$ finden kann. In anderen Worten sieht der Ellipsoid unter der A-Geometrie wie eine Kugel aus, so dass der A-Gradientenvektor $A^{-1}\nabla f(x)$ auf den Mittelpunkt, das heißt, die exakte Lösung $\bar{x}$, zeigt.

Kommen wir zurück zum Gradientenverfahren (1.3) für den Rayleigh-Quotienten. Die Variante zu $H = I$ lässt sich durch A^{-1} in die Variante zu $H = A$ modifizieren. Das führt normalerweise nicht so ideal zu einer 1-Schritt-Konvergenz, denn die Niveaumengen des Rayleigh-Quotienten sind nicht wie die Niveaumengen der obigen quadratischen Funktion ähnlich zur Einheitskugel der A-Vektornorm. Betrachten wir zum Beispiel die Niveaumengen des Rayleigh-Quotienten zu den Niveaus aus dem Intervall (λ_1, λ_2). Diese lassen sich durch Ellipsoide charakterisieren.

Lemma 2.7. *Für (A, M) gemäß Voraussetzung 2.1 besitze ein Vektor $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ die Entwicklung $u = \sum_{i=1}^m v_i$ bezüglich der Eigenräume von (A, M) und es gelte $\rho(u) \in (\lambda_1, \lambda_2)$, dann gilt die Ellipsoid-Gleichung $\sum_{i=2}^m \xi_i^2/\delta_i^2 = 1$ mit*

$$\xi_i := \frac{\|v_i\|_M}{\|v_1\|_M}, \quad \delta_i := \sqrt{\frac{\rho(u) - \lambda_1}{\lambda_i - \rho(u)}}, \quad i = 2, \ldots, m.$$

Beweis. Aus $\rho(u) = \big(\sum_{i=1}^m \lambda_i \|v_i\|_M^2\big)/\big(\sum_{i=1}^m \|v_i\|_M^2\big)$ folgt

$$\big(\rho(u) - \lambda_1\big)\|v_1\|_M^2 = \sum_{i=2}^m \big(\lambda_i - \rho(u)\big)\|v_i\|_M^2$$

und dann unmittelbar die Ellipsoid-Gleichung. □

Mittels solcher charakterisierenden Ellipsoide kann man die entsprechenden Niveaumengen als elliptische Kegel betrachten. Für die Niveaus aus $[\lambda_1, \lambda_m]\backslash(\lambda_1, \lambda_2)$ lassen sich die Niveaumengen analog durch Ellipsoide, Hyperboloide oder andere Hyperflächen zweiter Ordnung charakterisieren. Die charakterisierenden Hyperflächen können also zu unterschiedlichen Typen gehören, außerdem sind deren Halbachsen-Verhältnisse abhängig von den Niveaus und anders als diejenigen der Einheitskugel der A-Vektornorm. Trotzdem hat (1.3) mit $H = A$ für unsere Problemstellung oft ein besseres Konvergenzverhalten im Vergleich zu (1.3) mit $H = I$.

Beispiel 2.8. *Zur geometrischen Veranschaulichung des Gradientenverfahrens (1.3) benutzen wir die kleinen symmetrischen und positiv definiten Beispielmatrizen*

$$A := \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}, \quad M := \begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 5 \end{pmatrix}.$$

Die Eigenwerte von (A, M) sind sämtlich einfach: $\lambda_1 \approx 0.52308$, $\lambda_2 \approx 0.76156$, $\lambda_3 \approx 0.85095$. Mit dem Vektor $u = (1, -1, 3)^T$ wird jeweils ein Schritt der Varianten von (1.3) mit $H = I$ beziehungsweise $H = A$ durchgeführt. Gemäß Lemma 2.7 zeichnen wir in Abbildung 2.1 bezüglich einer durch

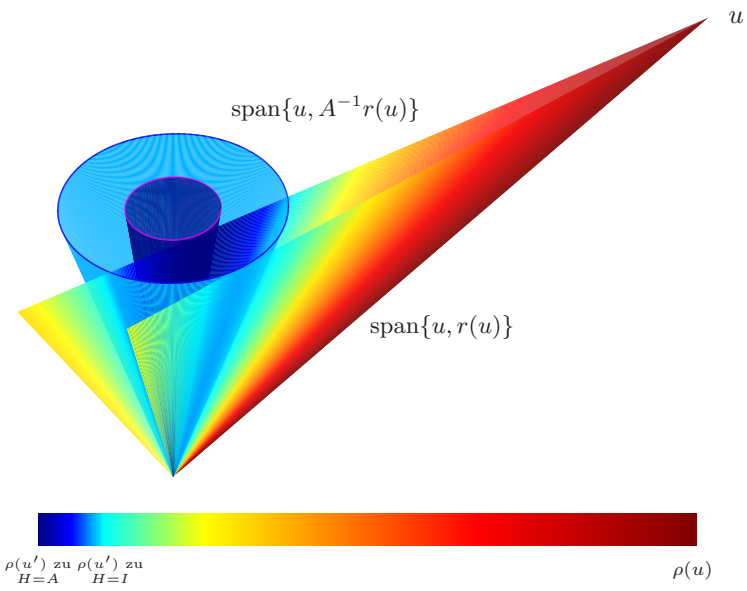


Abbildung 2.1: Einschritt-Vergleich der Varianten von (1.3)

Eigenvektoren von (A, M) gebildeten M-orthonormalen Basis von $\mathbb{R}^3$ die Koeffizientendarstellungen der relevanten Unterräume und der Rayleigh-Quotient-Niveaumengen zum jeweiligen kleinsten Ritzwert.

Dabei werden die Figuren zum Zweck einer besseren Veranschaulichung nur teilweise dargestellt und die zugehörigen Teilflächen (bis auf zwei kennzeichnende Ellipsen) werden gemäß Rayleigh-Quotienten eingefärbt. Die Unterräume $\mathrm{span}\{u, r(u)\}$, $\mathrm{span}\{u, A^{-1}r(u)\}$ können also als Tangentialebenen der entsprechenden Rayleigh-Quotient-Niveaumengen interpretiert werden. Der dem Unterraum $\mathrm{span}\{u, A^{-1}r(u)\}$ entsprechende Kegel hat offenbar einen kleineren Öffnungswinkel, so dass (1.3) mit $H = A$ zumindest in diesem Schritt schneller konvergiert. Die Halbachsen-Verhältnisse der charakterisierenden Ellipsen sind nicht weit von 1 entfernt, so dass auch (1.3) mit $H = I$ nicht sehr langsam konvergiert.

Dazu können wir beispielsweise die Komponente $A(3, 3)$ auf 36 vergrößern, so dass das geänderte Matrixbüschel (A, M) die Eigenwerte $\lambda_1 \approx 0.66442$, $\lambda_2 \approx 0.79952$, $\lambda_3 \approx 9.06149$ hat und die neuen charakterisierenden Ellipsen viel spitzer aussehen. Mit dem Startvektor $u^{(0)} = (1, 30, 1)^T$ werden jeweils 12 Schritte der Varianten von (1.3) mit $H = I$ beziehungsweise $H = A$ durchgeführt. Durch die Darstellung in Abbildung 2.2 beobachtet man die deutlich schnellere Konvergenz der zweiten Variante.

Wir untersuchen weiter die Variante von (1.3) mit $H = A$. Als eine einfache Konvergenzaussage lässt sich zeigen, dass diese Variante im Sinne der Rayleigh-Quotient-Verkleinerung nicht langsamer als die inverse Vektoriteration konvergiert. Zu bemerken ist, dass hier eine einfache Einschritt-Abschätzung betrachtet wird, welche nicht von bestimmten Eigenwerten und Eigenvektoren abhängt und Ergänzungen für weitere Abschätzungen in Satz 2.14 und Satz 3.12 liefert. Die Aussage, dass die inverse Vektoriteration bezüglich des Rayleigh-Quotienten konvergiert, lässt sich natürlich auch direkt in klassischer Weise zeigen.

Lemma 2.9. *Für (A, M) gemäß Voraussetzung 2.1 und $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ gilt nach einem Schritt von*

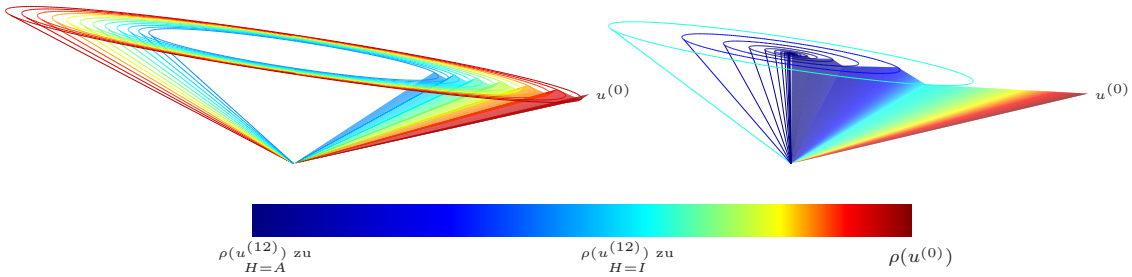


Abbildung 2.2: Mehrschritt-Vergleich der Varianten von (1.3)

(1.3) mit $H = A$, dass

$$\rho(u') \leq \rho\big(A^{-1}Mu\big) \leq \rho(u).$$

Falls u zusätzlich kein Eigenvektor ist, gilt $\rho(u') \leq \rho\big(A^{-1}Mu\big) < \rho(u)$.

Beweis. Wegen $A^{-1}r(u) = u - \rho(u)A^{-1}Mu$ und $\rho(u) \neq 0$ gehört $A^{-1}Mu$ zum Unterrraum $\operatorname{span}\{u,\, A^{-1}r(u)\}$, so dass $\rho(u') \leq \rho\big(A^{-1}Mu\big)$.
Um $\rho\big(A^{-1}Mu\big) \leq \rho(u)$ zu zeigen, können wir in einer Darstellung von $\rho\big(A^{-1}Mu\big)$ gewisse M-Skalarprodukte und A-Skalarprodukte konstruieren und Cauchy-Schwarz-Ungleichungen verwenden. Es gilt

$$\begin{aligned} 0 < \rho\big(A^{-1}Mu\big) &= \frac{(A^{-1}Mu)^T A(A^{-1}Mu)}{(A^{-1}Mu)^T M(A^{-1}Mu)} = \frac{(u,\, A^{-1}Mu)_M}{\|A^{-1}Mu\|_M^2} \\ &\leq \frac{\|u\|_M \, \|A^{-1}Mu\|_M}{\|A^{-1}Mu\|_M^2} = \frac{\|u\|_M^2}{\|u\|_M \, \|A^{-1}Mu\|_M} = \frac{(u,\, A^{-1}Mu)_A}{\|u\|_M \, \|A^{-1}Mu\|_M} \\ &\leq \frac{\|u\|_A \, \|A^{-1}Mu\|_A}{\|u\|_M \, \|A^{-1}Mu\|_M} = \sqrt{\rho(u)}\,\sqrt{\rho\big(A^{-1}Mu\big)}\,, \end{aligned}$$

so dass $\sqrt{\rho\big(A^{-1}Mu\big)} \leq \sqrt{\rho(u)}$ und weiter $\rho\big(A^{-1}Mu\big) \leq \rho(u)$ gilt.
Die Gleichheiten bei den oben eingesetzten Cauchy-Schwarz-Ungleichungen gelten genau dann, wenn u und $A^{-1}Mu$ kollinear sind, das heißt, wenn u ein Eigenvektor ist. Damit wird die Aussage zum zusätzlichen Fall gezeigt. □

Zusammen mit Lemma 2.6 lässt sich außerdem zeigen, dass (1.3) mit $H = A$ gegen einen Eigenvektor von (A, M) konvergiert. Im Vergleich zur inversen Vektoriteration hat dieses Verfahren außerdem den Vorteil, dass das einfache Ersetzen von A^{-1} durch einen Vorkonditionierer B^{-1} ein gegen einen gewünschten Eigenvektor konvergentes und im Sinne des gesamten Aufwands noch effizienteres Verfahren liefern kann, nämlich (1.3) mit $H = B$. Zur Konvergenzanalyse dafür kann man zunächst die vorkonditionierte Vektoriteration (1.5), also PINVIT(1), als eine „primitive“ Version betrachten. Gemäß der Umformulierung

$$u' = \rho(u)A^{-1}Mu - (I - B^{-1}A)\big(\rho(u)A^{-1}Mu - u\big). \tag{2.1}$$

gilt PINVIT(1) als eine gestörte inverse Vektoriteration. Darauf basierend lässt sich eine einfache Konvergenzaussage zeigen.

Lemma 2.10. *Betrachten wir (A, M) gemäß Voraussetzung 2.1. Falls $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ kein Eigenvektor von (A, M) ist, dann liefert PINVIT(1), siehe Vorschrift (1.5), unter der Qualitätsbedingung (1.4) des Vorkonditionierers eine Rayleigh-Quotient-Verkleinerung $\rho(u') < \rho(u)$.*

Beweis. Wegen der A-Orthogonalität

$$u^T A\big(\rho(u)A^{-1}Mu - u\big) = \rho(u)u^T Mu - u^T Au = 0$$

ist u die A-Projektion von $\rho(u)A^{-1}Mu$ auf $\operatorname{span}\{u\}$ und es gilt

$$\big\|\rho(u)A^{-1}Mu - u\big\|_A^2 = \big\|\rho(u)A^{-1}Mu\big\|_A^2 - \|u\|_A^2.$$

Für die A-Projektion $\tilde{u}$ von $\rho(u)A^{-1}Mu$ auf $\operatorname{span}\{u'\}$ gilt in ähnlicher Weise

$$\big\|\rho(u)A^{-1}Mu - \tilde{u}\big\|_A^2 = \big\|\rho(u)A^{-1}Mu\big\|_A^2 - \|\tilde{u}\|_A^2.$$

Aus der Minimierungseigenschaft der A-Projektion folgt

$$\big\|\rho(u)A^{-1}Mu - \tilde{u}\big\|_A^2 \leq \big\|\rho(u)A^{-1}Mu - u'\big\|_A^2.$$

Außerdem wird mit der Umformulierung (2.1) unter der Bedingung (1.4) gezeigt, dass

$$\big\|\rho(u)A^{-1}Mu - u'\big\|_A \leq \gamma\big\|\rho(u)A^{-1}Mu - u\big\|_A < \big\|\rho(u)A^{-1}Mu - u\big\|_A.$$

Zusammensetzen der obigen Ergebnisse liefert die einfache Ungleichung $\|\tilde{u}\|_A^2 > \|u\|_A^2$. Dabei lässt sich $\tilde{u}$ mittels einer A-Projektionsmatrix umschreiben, nämlich

$$\tilde{u} = \big(u'(u'^T Au')^{-1}u'^T A\big)\big(\rho(u)A^{-1}Mu\big) = \frac{\rho(u)(u', u)_M}{\|u'\|_A^2}u'.$$

Einsetzen in $\|\tilde{u}\|_A^2 > \|u\|_A^2$ liefert dann

$$\frac{\big(\rho(u)\big)^2(u', u)_M^2}{\|u'\|_A^4}\|u'\|_A^2 > \|u\|_A^2 \quad \Rightarrow \quad \big(\rho(u)\big)^2 > \frac{\|u'\|_A^2\,\|u\|_A^2}{(u', u)_M^2}.$$

Weiter gilt mit der Cauchy-Schwarz-Ungleichung $(u', u)_M^2 \leq \|u'\|_M^2\|u\|_M^2$, dass

$$\big(\rho(u)\big)^2 > \frac{\|u'\|_A^2\,\|u\|_A^2}{\|u'\|_M^2\|u\|_M^2} = \rho(u')\rho(u).$$

Wegen $\rho(u) > 0$ folgt dann $\rho(u') < \rho(u)$. □

Als eine verbesserte Version von PINVIT(1) liefert (1.3) mit $H = B$ ebenfalls eine Rayleigh-Quotient-Verkleinerung, und nach Lemma 2.6 lässt sich damit die Konvergenz der Vektoriterierten gegen einen Eigenvektor von (A, M) zeigen.

Zur genaueren Konvergenzanalyse für die Varianten von (1.3) mit $H = A$ beziehungsweise $H = B$, sowie für die jeweiligen dazu äquivalenten Verfahren INVIT(2) und PINVIT(2), kann man ebenfalls zunächst die einfacheren Verfahren INVIT(1) und PINVIT(1) untersuchen.

2.2 Musteranalyse und Konvergenzmaße

In diesem Abschnitt wird eine Musteranalyse für die inverse Vektoriteration vorgestellt. Die resultierenden scharfen Konvergenzabschätzungen sind natürlich auch für INVIT(1), nämlich eine

skalierte inverse Vektoriteration, verwendbar. Für die weiteren (P)INVIT-Verfahren wird mit einer ähnlichen Vorgehensweise und gleichartigen Konvergenzmaßen versucht, scharfe Konvergenzabschätzungen zu entwickeln.

Klassische Konvergenzanalysen für die inverse Vektoriteration (oder formal für die Potenzmethode) sind in Lehrbüchern wie [28, 79, 84] vorhanden und teilweise auch für unsymmetrische Eigenwertprobleme anwendbar. Gemäß unserer Problemstellung lässt sich zunächst zeigen, dass die M-Winkel der Vektoriterierten mit dem Ziel-Eigenraum eine monoton fallende Folge bilden.

Satz 2.11. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und einen Vektor $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ werde ein Schritt der inversen Vektoriteration durchgeführt: $u' = A^{-1}Mu$.*

Falls u kein Eigenvektor ist, dann besitzt u auf mindestens zwei Eigenräumen Nichtnull-Anteile. Seien die kleinsten zwei der entsprechenden Eigenwerte mit $\lambda_{\sigma(1)}$, $\lambda_{\sigma(2)}$, $1 \le \sigma(1) < \sigma(2) \le m$, bezeichnet, dann liegen $\tan \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)$ und $\tan \angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big)$ im Intervall $(0, \infty)$ und es gilt

$$\frac{\tan \angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big)}{\tan \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)} \le \frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}. \tag{2.2}$$

Gelte zusätzlich, dass u auf allen anderen Eigenräumen, das heißt auf den Eigenräumen mit Indizes aus $\{1, \ldots, m\} \backslash \{\sigma(1), \sigma(2)\}$, Null-Anteile besitzt, dann wird (2.2) eine Gleichung. (Damit ist (2.2) eine scharfe Abschätzung.)

Beweis. Der Vektor $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ lässt sich durch eine Entwicklung $u = \sum_{i=1}^m v_i$ bezüglich der Eigenräume von (A, M) darstellen. Falls u kein Eigenvektor ist, muss es mindestens zwei Nichnull-Vektoren v_i geben, denn sonst würde u zu einem Eigenraum gehören und wäre ein Eigenvektor. Mit den Indizes $\lambda_{\sigma(1)}$, $\lambda_{\sigma(2)}$ aus der Satz-Formulierung hat u eine genauere Darstellung $u = v_{\sigma(1)} + \sum_{i=\sigma(2)}^m v_i$ mit Nichtnullvektoren $v_{\sigma(1)}$, $v_{\sigma(2)}$, so dass

$$\tan \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big) = \frac{\big\| \sum_{i=\sigma(2)}^m v_i \big\|_M}{\big\| v_{\sigma(1)} \big\|_M} = \frac{\sqrt{\sum_{i=\sigma(2)}^m \|v_i\|_M^2}}{\|v_{\sigma(1)}\|_M} \in (0, \infty).$$

Dementsprechend hat u' eine Darstellung $u' = \lambda_{\sigma(1)}^{-1} v_{\sigma(1)} + \sum_{i=\sigma(2)}^m \lambda_i^{-1} v_i$ und es gilt

$$\tan \angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big) = \frac{\big\| \sum_{i=\sigma(2)}^m \lambda_i^{-1} v_i \big\|_M}{\big\| \lambda_{\sigma(1)}^{-1} v_{\sigma(1)} \big\|_M} = \frac{\sqrt{\sum_{i=\sigma(2)}^m \lambda_i^{-2} \|v_i\|_M^2}}{\lambda_{\sigma(1)}^{-1} \|v_{\sigma(1)}\|_M} \in (0, \infty).$$

Weiter erhalten wir

$$\frac{\tan \angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big)}{\tan \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)} = \frac{\sqrt{\sum_{i=\sigma(2)}^m \lambda_i^{-2} \|v_i\|_M^2}}{\lambda_{\sigma(1)}^{-1} \sqrt{\sum_{i=\sigma(2)}^m \|v_i\|_M^2}} \le \frac{\lambda_{\sigma(2)}^{-1}}{\lambda_{\sigma(1)}^{-1}} = \frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}.$$

Unter der zusätzlichen Bedingung besitzen die Summen $\sum_{i=\sigma(2)}^m \lambda_i^{-2} \|v_i\|_M^2$ und $\sum_{i=\sigma(2)}^m \|v_i\|_M^2$ jeweils nur einen Summanden, so dass eine Gleichung geliefert wird. □

Satz 2.11 lässt sich auch bezüglich der A-Orthogonalität formulieren. Die zugehörige Schranke wird wieder durch den Quotienten $\lambda_{\sigma(1)}/\lambda_{\sigma(2)}$ gegeben.

Da $\tan \angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big)$ bei einer Nichtnull-Umskalierung von u' unverändert bleibt, lässt sich Satz 2.11 auf INVIT(1) übertragen. Außerdem können wir nach Satz 2.11 eine Mehrschritt-Abschätzung bezüglich der praktischen Vorschrift

$$\begin{aligned} &u^{(0)} = u/\|u\|_M\,; \quad l = 0\,; \\ &\texttt{while true do} \quad u' = A^{-1}Mu^{(l)}\,; \quad l = l+1\,; \quad u^{(l)} = u'/\|u'\|_M\,; \quad \texttt{end do} \end{aligned} \tag{2.3}$$

der inversen Vektoriteration formulieren, nämlich

$$\frac{\tan \angle_M\big(u^{(l)},\, \mathcal{V}_{\sigma(1)}\big)}{\tan \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)} \leq \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)^l.$$

Weiter gilt für den Anteil $v^{(l)}_{\sigma(1)}$ von $u^{(l)}$ auf $\mathcal{V}_{\sigma(1)}$ wegen $\|u^{(l)}\|_M = 1$, dass

$$\big\|u^{(l)} - v^{(l)}_{\sigma(1)}\big\|_M = \sin \angle_M\big(u^{(l)},\, \mathcal{V}_{\sigma(1)}\big) \leq \tan \angle_M\big(u^{(l)},\, \mathcal{V}_{\sigma(1)}\big) \leq \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)^l \tan \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big).$$

In der Praxis wird die Anfangsiterierte normalerweise pseudozufällig erzeugt und die Orthogonalität könnte durch Rundungsfehler gestört werden, so dass (2.3) oft gegen einen Eigenvektor zu λ_1 konvergiert und der Konvergenzfaktor dafür durch $(\lambda_1/\lambda_2)^l$ gegeben wird.

Für Eigenwertnäherungen benutzen wir wie in [51, 62] eine relative Position als Konvergenzmaß. Falls der Rayleigh-Quotient eines Vektors $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ in einem durch Eigenwerte begrenzten offenen Intervall $\big(\lambda_{\sigma(1)},\, \lambda_{\sigma(2)}\big)$ liegt, dann wird mit dem Quotienten

$$\Delta(\alpha, \beta, \chi) := \frac{\chi - \alpha}{\beta - \chi} \tag{2.4}$$

eine relative Position durch $\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u)\big)$ gegeben, welche als eine Funktion für $\rho(u)$ betrachtet werden kann und monoton steigend auf $\big(\lambda_{\sigma(1)},\, \lambda_{\sigma(2)}\big)$ ist. Außerdem bezieht sich die Größe $\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u)\big)$ auf manche geometrische Argumente, siehe Lemma 2.7. Für dieses Konvergenzmaß betrachten wir zunächst einen Spezialfall.

Lemma 2.12. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und einen Vektor $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$, welcher auf den Eigenräumen mit Indizes aus $\{1, \ldots, m\}\backslash\{\sigma(1), \sigma(2)\}$, $1 \leq \sigma(1) < \sigma(2) \leq m$, Null-Anteile besitzt, werde ein Schritt der inversen Vektoriteration $u' = A^{-1}Mu$ durchgeführt. Falls u kein Eigenvektor ist, dann gehören $\rho(u)$ und $\rho(u')$ beide zum Intervall $\big(\lambda_{\sigma(1)},\, \lambda_{\sigma(2)}\big)$. Mit der Bezeichnung (2.4) gilt*

$$\frac{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u)\big)} = \frac{\tan^2 \angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big)}{\tan^2 \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)} = \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)^2. \tag{2.5}$$

Beweis. Nach Voraussetzungen lässt sich u durch eine Entwicklung $u = v_{\sigma(1)} + v_{\sigma(2)}$ bezüglich der Eigenräume $\mathcal{V}_{\sigma(1)}$, $\mathcal{V}_{\sigma(2)}$ darstellen. Falls u kein Eigenvektor ist, dann sind $v_{\sigma(1)}$ und $v_{\sigma(2)}$ keine Nullvektoren, so dass $u' = \lambda^{-1}_{\sigma(1)} v_{\sigma(1)} + \lambda^{-1}_{\sigma(2)} v_{\sigma(2)}$ kein Nullvektor und kein Eigenvektor ist und $\rho(u)$, $\rho(u')$ beide zu $\big(\lambda_{\sigma(1)},\, \lambda_{\sigma(2)}\big)$ gehören. Weiter erhalten wir

$$\begin{aligned}
\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u)\big) &= \frac{\rho(u) - \lambda_{\sigma(1)}}{\lambda_{\sigma(2)} - \rho(u)} = \frac{\frac{\lambda_{\sigma(1)}\|v_{\sigma(1)}\|^2_M + \lambda_{\sigma(2)}\|v_{\sigma(2)}\|^2_M}{\|v_{\sigma(1)}\|^2_M + \|v_{\sigma(2)}\|^2_M} - \lambda_{\sigma(1)}}{\lambda_{\sigma(2)} - \frac{\lambda_{\sigma(1)}\|v_{\sigma(1)}\|^2_M + \lambda_{\sigma(2)}\|v_{\sigma(2)}\|^2_M}{\|v_{\sigma(1)}\|^2_M + \|v_{\sigma(2)}\|^2_M}} \\
&= \frac{(\lambda_{\sigma(2)} - \lambda_{\sigma(1)})\|v_{\sigma(2)}\|^2_M}{(\lambda_{\sigma(2)} - \lambda_{\sigma(1)})\|v_{\sigma(1)}\|^2_M} = \frac{\|v_{\sigma(2)}\|^2_M}{\|v_{\sigma(1)}\|^2_M} = \tan^2 \angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)
\end{aligned}$$

und eine ähnliche Gleichung für u'. Schließlich folgt die Abschätzung aus Satz 2.11. □

Mit (2.5) erhalten wir eine Konvergenzabschätzung für Eigenwertnäherungen in einem niedrigdimensionalen Fall. Weiter lässt sich die zugehörige Voraussetzung aus einem Optimierungsproblem herleiten, so dass (2.5) zum Beweis einer allgemeineren Abschätzung einsetzbar ist. Im Vergleich zur Musteranalyse in [62] verwenden wir zu dieser Herleitung eine Kurven-Differentiation statt Lagrange-Multiplikatoren.

Lemma 2.13. *Betrachten wir das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und weiter eine Rayleigh-Quotient-Niveaumenge $\mathcal{N}(\tilde{\lambda}) := \{x \in \mathbb{R}^n \backslash \{\mathbf{0}\} \ ; \ \rho(x) = \tilde{\lambda}\}$, wobei $\tilde{\lambda} \in (\lambda_1, \lambda_m)$ und $\tilde{\lambda}$ kein Eigenwert von (A, M) ist. Falls u eine Extremstelle der auf $\mathcal{N}(\tilde{\lambda})$ definierten Funktion $f(x) := \rho\big(A^{-1}Mx\big)$ ist, gibt es genau zwei verschiedene Eigenwerte, auf deren Eigenräumen u Nichtnull-Anteile besitzt.*

Beweis. Auf der Niveaumenge $\mathcal{N}(\tilde{\lambda})$ kann man stetig differenzierbare Kurven (etwa wie Kegelschnitte) definieren. Für eine Extremstelle u von f auf $\mathcal{N}(\tilde{\lambda})$ betrachten wir eine beliebige durch u laufende stetig differenzierbare Kurve $\{g(t) \ ; \ t \in T \subseteq \mathbb{R}\} \subset \mathcal{N}(\tilde{\lambda})$ mit $g(t_g) = u$, dann ist u auch eine Extremstelle von f auf der Kurve, so dass t_g eine Extremstelle der Funktion $f \circ g$ ist und die notwendige Bedingung $\big(d\,(f \circ g)/dt\big)(t_g) = 0$ gilt. Daraus folgt mit

$$(f \circ g)(t) = \rho\big(A^{-1}Mg(t)\big) \quad \Rightarrow \quad \frac{d\,(f \circ g)}{dt}(t) = \Big(\nabla\rho\big(A^{-1}Mg(t)\big),\, A^{-1}M\dot{g}(t)\Big)_2,$$

wobei $\dot{g}(t)$ die Ableitung $(dg/dt)(t)$ bezeichnet, dass

$$0 = \Big(\nabla\rho\big(A^{-1}Mg(t_g)\big),\, A^{-1}M\dot{g}(t_g)\Big)_2 = \Big(MA^{-1}\nabla\rho\big(A^{-1}Mu\big),\, \dot{g}(t_g)\Big)_2.$$

Damit ist der Vektor $MA^{-1}\nabla\rho\big(A^{-1}Mu\big)$ Euklidisch orthogonal zum Tangentialvektor $\dot{g}(t_g)$. Wenn wir hinreichend viele auf $\mathcal{N}(\tilde{\lambda})$ und durch u laufende Kurven betrachten, dann lässt sich die Tangentialhyperebene $\dot{u}$ von $\mathcal{N}(\tilde{\lambda})$ an u durch mehrere solcher Tangentialvektoren bilden. Der Vektor $MA^{-1}\nabla\rho\big(A^{-1}Mu\big)$ liegt dann Euklidisch orthogonal zu $\dot{u}$.

Der Gradientenvektor $\nabla\rho(u)$ liegt ebenfalls Euklidisch orthogonal zu $\dot{u}$. Das lässt sich mittels der Eigenschaften der Niveaumenge zeigen. Für die Kurven der Form $\{g(t) \ ; \ t \in T \subseteq \mathbb{R}\} \subset \mathcal{N}(\tilde{\lambda})$ mit $g(t_g) = u$ ist die Funktion $\rho \circ g$ nach der Definition der Niveaumenge konstant, so dass

$$0 = \frac{d\,(\rho \circ g)}{dt}(t) = \Big(\nabla\rho\big(g(t)\big),\, \dot{g}(t)\Big)_2$$

für beliebiges $t \in T$ gilt. Insbesondere für $t = t_g$ erhalten wir $\big(\nabla\rho(u),\, \dot{g}(t_g)\big)_2 = 0$ und weiter die Orthogonalität $\nabla\rho(u) \perp_2 \dot{u}$.

Da die Tangentialhyperebene $\dot{u}$ eine Hyperebene ist, sind $MA^{-1}\nabla\rho\big(A^{-1}Mu\big)$ und $\nabla\rho(u)$ kollinear. Nach den Voraussetzungen ist u kein Nullvektor und kein Eigenvektor, so dass $\nabla\rho(u)$ kein Nullvektor ist. Daher gibt es ein $\tilde{\alpha} \in \mathbb{R}$ mit $MA^{-1}\nabla\rho\big(A^{-1}Mu\big) = \tilde{\alpha}\nabla\rho(u)$, das heißt

$$MA^{-1}\frac{2}{\|A^{-1}Mu\|_M^2}\Big(A\big(A^{-1}Mu\big) - \rho\big(A^{-1}Mu\big)M\big(A^{-1}Mu\big)\Big) = \frac{2\tilde{\alpha}}{\|u\|_M^2}\big(Au - \rho(u)Mu\big). \tag{2.6}$$

Mit $\alpha := \tilde{\alpha}\|A^{-1}Mu\|_M^2/\|u\|_M^2$, $\beta := \rho\big(A^{-1}Mu\big)$ sowie $\tilde{\lambda} = \rho(u)$ erhalten wir dann

$$M(A^{-1}M)u - \beta M(A^{-1}M)^2u = \alpha(Au - \tilde{\lambda}Mu).$$

Das lässt sich mit der Entwicklung $u = \sum_{i=1}^m v_i$ bezüglich der Eigenräume von (A, M) in

$$\sum_{i=1}^m \big(\lambda_i^{-1} - \beta\lambda_i^{-2}\big)Mv_i = \sum_{i=1}^m \alpha(\lambda_i - \tilde{\lambda})Mv_i$$

und weiter mit dem Polynom $p(\delta) := \alpha\delta^3 - \alpha\tilde{\lambda}\delta^2 - \delta + \beta$ in

$$\sum_{i=1}^m p(\lambda_i)\,\lambda_i^{-2}\,Mv_i = \mathbf{0}$$

umschreiben. Multiplizieren wir die beiden Seiten dieser Gleichung mit $\lambda_j^2 v_j^T$, $j = 1, \ldots, m$, von links, so erhalten wir wegen der M-Orthogonalität der Eigenräume die Gleichungen

$$p(\lambda_j)\, \|v_j\|_M^2 = \mathbf{0}, \qquad j = 1, \ldots, m.$$

Da u kein Nullvektor und kein Eigenvektor ist, sind mindestens zwei $\|v_j\|_M^2$ von Null verschieden. Damit hat $p(\delta)$ wegen $\lambda_j > 0 \;\forall\; j$ mindestens zwei positive Nullstellen. Außerdem ist $p(\delta)$ ein kubisches Polynom mit reellen Koeffizienten und hat höchstens drei positive Nullstellen. Weiter untersuchen wir die Koeffizienten von $p(\delta)$, um den Fall, dass $p(\delta)$ drei positive Nullstellen hat, auszuschließen.

Betrachten wir dafür die Darstellungen $p(\delta) = \alpha\delta^3 - \alpha\tilde{\lambda}\delta^2 - \delta + \beta$ und $p(\delta) = \alpha \prod_{i=1}^{3}(\delta - \delta_i)$, wobei $\delta_1, \delta_2, \delta_3$ die Nullstellen bezeichnen. Durch einen Koeffizientenvergleich wird $\prod_{i=1}^{3} \delta_i = -\beta/\alpha$ gezeigt. Wären alle drei Nullstellen positiv, dann müsste $-\beta/\alpha > 0$ gelten. Für β gilt bereits $\beta = \rho\big(A^{-1}Mu\big) > 0$. Um das Vorzeichen von α zu bestimmen, schreiben wir die Gleichung (2.6) mittels der Substitutionen $\alpha = \tilde{\alpha}\|A^{-1}Mu\|_M^2/\|u\|_M^2$, $u' = A^{-1}Mu$ sowie der Bezeichnung (1.2) (Rayleigh-Quotient-Residuum) als

$$MA^{-1}r(u') = \alpha r(u). \tag{2.7}$$

Da M^{-1} auch symmetrisch und positiv definit ist, lässt sich die M^{-1}-Vektornorm definieren. Multiplikation der beiden Seiten von (2.7) mit $\big(r(u)\big)^T M^{-1}$ von links liefert

$$\big(r(u)\big)^T A^{-1} r(u') = \alpha \|r(u)\|_{M^{-1}}^2.$$

Die neue linke Seite lässt sich weiter umschreiben zu

$$\begin{aligned}
\big(r(u)\big)^T A^{-1} r(u') &= \big(A^{-1}r(u)\big)^T r(u') = \Big(A^{-1}\big(Au - \rho(u)Mu\big)\Big)^T r(u') = \big(u - \rho(u)u'\big)^T r(u') \\
&= \big(u - \rho(u')u'\big)^T r(u') = \big(M^{-1}Au' - \rho(u')(M^{-1}M)u'\big)^T r(u') \\
&= \big(Au' - \rho(u')Mu'\big)^T M^{-1} r(u') = \|r(u')\|_{M^{-1}}^2
\end{aligned}$$

(am Anfang der zweiten Zeile wurde die Euklidische Orthogonalität zwischen $r(u')$ und u' benutzt). Damit gilt $\|r(u')\|_{M^{-1}}^2 = \alpha\|r(u)\|_{M^{-1}}^2$. Außerdem lässt sich u' durch $u' = \sum_{j=1}^{m} \lambda_i^{-1} v_i$ darstellen und hat dabei mindestens zwei Nichtnull-Summanden, so dass u' wie u kein Nullvektor und kein Eigenvektor ist. Daher sind die Residuen $r(u)$, $r(u')$ keine Nullvektoren, so dass $\|r(u)\|_{M^{-1}}^2$, $\|r(u')\|_{M^{-1}}^2$ positiv sind und $\alpha > 0$ sowie $-\beta/\alpha < 0$ gezeigt werden.

Damit hat $p(\delta)$ genau zwei positive Nullstellen und u genau zwei Nichtnull-Summanden bei der Entwicklung bezüglich der Eigenräume. □

Durch Kombination der niedrigdimensionalen Analyse aus Lemma 2.12 und der Extremstelleneinschränkung aus Lemma 2.13 lässt sich eine Muster-Konvergenzabschätzung für Eigenwertnäherungen zeigen.

Satz 2.14. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und einen Vektor $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ werde ein Schritt der inversen Vektoriteration durchgeführt: $u' = A^{-1}Mu$.*

Falls u kein Eigenvektor ist, dann gilt $\rho(u') < \rho(u)$. Falls $\rho(u)$ im Intervall $(\lambda_i, \lambda_{i+1})$ mit $i \in \{1, \ldots, m-1\}$ liegt, dann gilt mit der Bezeichnung (2.4), dass

$$\frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)} \leq \left(\frac{\lambda_i}{\lambda_{i+1}}\right)^2. \tag{2.8}$$

Gelte zusätzlich, dass u auf den Eigenräumen mit Indizes aus $\{1, \ldots, m\}\backslash\{i, i+1\}$ Null-Anteile besitzt, wird (2.8) eine Gleichung. (Damit ist (2.8) eine scharfe Abschätzung.)

Beweis. $\rho(u') < \rho(u)$ wurde bereits als $\rho\big(A^{-1}Mu\big) < \rho(u)$ in Lemma 2.9 bewiesen.

Mit $\lambda_i < \rho(u) < \lambda_{i+1}$ sind u und u' keine Eigenvektoren (siehe Beweis von Lemma 2.13), und es gilt entweder $\rho(u') \le \lambda_i$ oder $\lambda_i < \rho(u') < \rho(u)$. Für den Fall $\rho(u') \le \lambda_i$ ist der abzuschätzende Delta-Quotient nicht positiv, so dass die Ungleichung (2.8) trivial ist. Für den Fall $\lambda_i < \rho(u') < \rho(u)$ betrachten wir die Rayleigh-Quotient-Niveaumenge $\mathcal{N}\big(\rho(u)\big) := \{x \in \mathbb{R}^n \backslash \{\mathbf{0}\} \; ; \; \rho(x) = \rho(u)\}$ und eine Maximalstelle $\bar{u}$ der Funktion $f(x) := \rho\big(A^{-1}Mx\big)$ auf $\mathcal{N}\big(\rho(u)\big)$. Dann ist $\bar{u}$ kein Eigenvektor und es gilt mit $\bar{u}' := A^{-1}M\bar{u}$, dass $\lambda_i < \rho(u') \le \rho(\bar{u}') < \rho(\bar{u}) = \rho(u) < \lambda_{i+1}$. Weiter gilt

$$\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big) \le \Delta\big(\lambda_i, \lambda_{i+1}, \rho(\bar{u}')\big), \tag{2.9}$$

denn die auf dem Intervall $(\lambda_i, \lambda_{i+1})$ definierte Funktion $g(\alpha) := \Delta\big(\lambda_i, \lambda_{i+1}, \alpha\big)$ ist wegen

$$\frac{dg}{d\alpha}(\alpha) = (\lambda_{i+1} - \lambda_i)/(\lambda_{i+1} - \alpha)^2 > 0$$

streng monoton steigend.

Außerdem gilt nach Lemma 2.13, dass es genau zwei verschiedene Eigenwerte gibt, auf deren Eigenräumen $\bar{u}$ Nichtnull-Anteile besitzt. Seien $\sigma(1) < \sigma(2)$ die zugehörigen Indizes, dann liefert Lemma 2.12 angewandt auf $\bar{u}$ die Abschätzung

$$\frac{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\bar{u}')\big)}{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\bar{u})\big)} = \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)^2.$$

Da kein Eigenwert in $(\lambda_i, \lambda_{i+1})$ liegt, gilt $\lambda_{\sigma(1)} \le \lambda_i < \rho(\bar{u}') < \rho(\bar{u}) < \lambda_{i+1} \le \lambda_{\sigma(2)}$, so dass

$$\begin{aligned}\frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\bar{u}')\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\bar{u})\big)} &= \left(\frac{\rho(\bar{u}') - \lambda_i}{\rho(\bar{u}) - \lambda_i}\right)\left(\frac{\lambda_{i+1} - \rho(\bar{u})}{\lambda_{i+1} - \rho(\bar{u}')}\right) \le \left(\frac{\rho(\bar{u}') - \lambda_{\sigma(1)}}{\rho(\bar{u}) - \lambda_{\sigma(1)}}\right)\left(\frac{\lambda_{\sigma(2)} - \rho(\bar{u})}{\lambda_{\sigma(2)} - \rho(\bar{u}')}\right) \\ &= \frac{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\bar{u}')\big)}{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\bar{u})\big)} = \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)^2 \le \left(\frac{\lambda_i}{\lambda_{i+1}}\right)^2.\end{aligned}$$

Das liefert zusammen mit (2.9) und $\rho(\bar{u}) = \rho(u)$ die Ungleichung (2.8). Unter der zusätzlichen Bedingung lässt sich die zugehörige Gleichung wie Lemma 2.12 zeigen. □

Bei der Muster-Konvergenzabschätzung für Eigenvektornäherungen wurde bezüglich der praktischen Vorschrift (2.3) diskutiert, dass die inverse Vektoriteration oft gegen einen Eigenvektor zu λ_1 konvergiert. In anderen Worten konvergiert die zugehörige Rayleigh-Quotient-Folge gegen λ_1 und erreicht nach hinreichend vielen Schritten das Intervall (λ_1, λ_2). Das weitere Konvergenzverhalten lässt sich durch eine auf (2.8) basierende Mehrschritt-Abschätzung beschreiben (denn $\rho(u')$ bleibt erhalten, wenn u' bezüglich der M-Vektornorm normiert wird). Gelte im l-ten Schritt $\rho\big(u^{(l)}\big) \in (\lambda_1, \lambda_2)$, dann gilt nach weiteren $\bar{l}$ Schritten, dass

$$\frac{\Delta\big(\lambda_1, \lambda_2, \rho\big(u^{(l+\bar{l})}\big)\big)}{\Delta\big(\lambda_1, \lambda_2, \rho\big(u^{(l)}\big)\big)} \le \left(\frac{\lambda_1}{\lambda_2}\right)^{2\bar{l}}.$$

Eine andere Folgerung von Lemma 2.12 und Lemma 2.13 bezieht sich auf die schnellst mögliche Konvergenz der inversen Vektoriteration.

Satz 2.15. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und einen Vektor $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ werde ein Schritt der inversen Vektoriteration durchgeführt: $u' = A^{-1}Mu$.*

Falls $\rho(u)$ kein Eigenwert ist, gilt

$$\frac{\Delta\big(\lambda_1, \lambda_m, \rho(u')\big)}{\Delta\big(\lambda_1, \lambda_m, \rho(u)\big)} \ge \left(\frac{\lambda_1}{\lambda_m}\right)^2. \tag{2.10}$$

Gelte zusätzlich, dass u auf den Eigenräumen mit Indizes aus $\{2, \dots, m-1\}$ Null-Anteile besitzt, wird (2.10) eine Gleichung. (Damit ist (2.10) eine scharfe Abschätzung.)

Beweis. Analog zum Beweis von Satz 2.14. Zu betrachten ist also eine Minimalstelle der Funktion $f(x) := \rho\big(A^{-1}Mx\big)$ auf der Rayleigh-Quotient-Niveaumenge $\mathcal{N}\big(\rho(u)\big)$. □

Durch die Umstellung von (2.8), (2.10) nach $\rho(u')$ erhalten wir genau die Abschätzungen der Musteranalyse in [62]. Für weitere (P)INVIT-Verfahren können die direkten Abschätzungen für $\rho(u')$ sehr kompliziert sein. Mit den „Delta-Quotienten“ kann man aber immer noch klare und intuitive Abschätzungen formulieren.

Wenn eine untere Schranke $\bar{\lambda}$ von λ_1 bekannt ist, lässt sich die inverse Vektoriteration noch durch einen „Shift“ modifizieren:

$$u' = (A - \bar{\lambda}M)^{-1}Mu.$$

Es ist einfach zu zeigen, dass die obigen Konvergenzfaktoren der Form $\lambda_{\sigma(1)}/\lambda_{\sigma(2)}$ nach dieser Modifizierung durch $(\lambda_{\sigma(1)} - \bar{\lambda})/(\lambda_{\sigma(2)} - \bar{\lambda})$ gegeben werden, siehe [64]. Im Fall $\bar{\lambda} > 0$ handelt es sich um eine Beschleunigung im Sinne der langsamsten Konvergenz.

3 A-Gradientenverfahren - eine Analyse in Krylovräumen

Der Begriff Krylovraum gilt als eine theoretische Grundlage für viele numerische Verfahren, zum Beispiel für CG-Verfahren oder CG-ähnliche Verfahren zum Lösen eines linearen Gleichungssystems sowie für das Arnoldi-Verfahren zum Lösen eines Matrix-Eigenwertproblems. In diesem Kapitel untersuchen wir das Konvergenzverhalten einer Art der auf Krylovräumen niedriger Dimensionen basierenden Iterationsverfahren zur Berechnung der kleinsten Eigenwerte und zugehöriger Eigenvektoren von (A, M) gemäß Voraussetzung 2.1. In jedem Schritt wird die neue Iterierte durch einen Ritzvektor zum kleinsten Ritzwert bezüglich eines Krylovraums der alten Iterierten gegeben. Solche Verfahren können praktischerweise als „(inverse) Lanczos-Verfahren mit Neustart“ implementiert werden, vergleiche [23, 82, 17, 100]. Das Gradientenverfahren (1.3) mit $H = A$ beziehungsweise INVIT(2) gehört auch zu dieser Verfahrensart und bezieht sich auf zweidimensionale Krylovräume zur Matrix $A^{-1}M$.

Durch Knyazev und Skorokhodov [51] wurde eine klassische Konvergenzanalyse für ein Gradientenverfahren zur Berechnung der größten Eigenwerte und zugehöriger Eigenvektoren einer reellen und symmetrischen Matrix präsentiert. Als Konvergenzmaße wurden ähnliche Tangens-Quotienten und Delta-Quotienten wie in (2.2) und (2.8) verwendet. In der Arbeit von Neymeyr [62] wurde diese Konvergenzanalyse auf INVIT(2) übertragen. Die Grundidee in [51] beziehungsweise [62] ist, ein „Schatten-Verfahren“ des originalen Verfahrens zu konstruieren und damit eine niedrigdimensionale Analyse zu ermöglichen. Beim Beweis wurden interessante Eigenschaften zweidimensionaler Krylovräume, jedoch auch langatmige Berechnungen, benutzt. Bei unserer neuen Untersuchung wurden weitere Varianten der Gradientenverfahren ergänzt, und die klassische Konvergenzanalyse wurde modifiziert beziehungsweise verbessert. Insbesondere wurde die niedrigdimensionale Analyse mittels geometrischer Interpretation in kurzer Form umformuliert, siehe [68].

In diesem Kapitel ist eine ähnliche Konvergenzanalyse für Verfahren zu allgemeineren s-dimensionalen Krylovräumen ($s \geq 3$) zu entwickeln. (Diese ist keine einfache Verallgemeinerung der klassischen Resultate, denn zum Beispiel eine strikte Einschachtelung der Ritzwerte zu s-dimensionalen Krylovräumen lässt sich nicht wie in [51] einfach durch direkte Umrechnungen zeigen.) In Abschnitt 3.1 werden die zu untersuchenden Verfahren vorgestellt und die niedrigdimensionale Analyse begründet. Als Motivation der weiteren Analyse wird in Abschnitt 3.2 ein Verfahren zu zweidimensionalen Krylovräumen, welches identisch zu INVIT(2) ist, betrachtet. Dabei wird eine alternative Version der oben genannten Modifizierung der klassischen Konvergenzanalyse vorgestellt. Schließlich werden in Abschnitt 3.3 neue Konvergenzabschätzungen für Verfahren zu s-dimensionalen Krylovräumen präsentiert.

3.1 Auf Krylovräumen niedriger Dimensionen basierende Iterationsverfahren

Einfach gesagt, ist ein Krylovraum ein durch Vektoriterierte der Potenzmethode aufgespannter Unterraum. Für Krylovräume zu verschiedenen Matrizen benutzen wir folgende Bezeichnungen.

Definition 3.1. *Für eine Matrix $H \in \mathbb{R}^{n\times n}$ und einen Vektor $w \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ heißt der Unterraum*

$$\mathcal{K}_H^s(w) := \operatorname{span}\{H^{i-1}w \ ; \ i = 1, \ldots, s\} \tag{3.1}$$

der H-Krylovraum s-ter Stufe von w.

Zur Berechnung der extremalen Eigenwerte und zugehöriger Eigenvektoren einer Matrix kann man gewisse Krylovräume konstruieren und dazu Ritzwert-Folgen und Ritzvektor-Folgen berechnen. Das Gradientenverfahren (1.3) zu $H = A$ beziehungsweise INVIT(2) ist ein auf $A^{-1}M$-Krylovräumen basierendes Verfahren, denn

$$\operatorname{span}\{u,\, A^{-1}r(u)\} = \operatorname{span}\{u,\, A^{-1}Mu\} = \mathcal{K}^2_{A^{-1}M}(u).$$

Als eine Erweiterung von INVIT(1) beziehungsweise der inversen Vektoriteration ist INVIT(2) zumindest bezüglich der äußeren Schritte ein schnelleres Verfahren. In diesem Sinne haben die auf $A^{-1}M$-Krylovräumen mit größeren Dimensionszahlen basierenden Verfahren

$$u' = \mathrm{RR}_{\rho,\min}\big(\mathcal{K}^s_{A^{-1}M}(u)\big) \tag{3.2}$$

möglicherweise noch bessere Approximationseigenschaften. Natürlich soll die Dimensionszahl s relativ klein sein, so dass der Speicherbedarf noch angemessen ist.

Es gibt eine klassische Art von Abschätzungen, womit die Qualität des ganzen Krylovraums beschrieben wird, siehe [38, 77, 83, 79, 66]. Ein Nachteil dieser Art ist, dass die Abschätzung keine Mehrschritt-Abschätzung (a-priori Abschätzung) impliziert. Um die Leistung von (3.2) besser zu beschreiben, entwickeln wir im Folgenden gleichartige Konvergenzabschätzungen wie (2.2) und (2.8). Betrachten wir zunächst einige triviale Fälle und bestimmen damit Einschränkungen für nichttriviale Fälle.

Lemma 3.2. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und das Verfahren (3.2) mit $u \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ seien die $\bar{s}$ Eigenwerte, auf deren Eigenräumen u Nichtnull-Anteile besitzt, mit $\lambda_{\sigma(j)}$, $\sigma(j) \in \{1,\dots,m\}$, bezeichnet. Dabei bilden die Indizes $\sigma(j)$, $j = 1,\dots,\bar{s}$, eine steigende Folge. Dann wird im Fall $s \geq \bar{s}$ ein Eigenvektor zum Eigenwert $\lambda_{\sigma(1)}$ als u' geliefert. Im Fall $s < \bar{s}$ hat $\mathcal{K}^s_{A^{-1}M}(u)$ die Dimension s und enthält keinen Eigenvektor.*

Beweis. Benutzen wir eine Entwicklung $u = \sum_{j=1}^{\bar{s}} v_{\sigma(j)}$ bezüglich der Eigenräume von (A, M). Damit gehört u zum $A^{-1}M$-invarianten Unterraum $\mathcal{V} := \operatorname{span}\{v_{\sigma(1)},\dots,v_{\sigma(\bar{s})}\}$. $\mathcal{V}$ ist identisch zum Krylovraum $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u)$, denn

a) Es gilt offensichtlich $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u) \subseteq \mathcal{V}$.

b) Das erzeugende System $\{(A^{-1}M)^{i-1}u\,;\ i = 1,\dots,\bar{s}\}$ von $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u)$ ist ein linear unabhängiges System, so dass $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u)$ die gleiche Dimension $\bar{s}$ wie $\mathcal{V}$ hat.

   Beweis von b): Betrachten wir eine beliebige Linearkombination von $(A^{-1}M)^{i-1}u$ mit Koeffizienten $\alpha_i \in \mathbb{R}$. Es gilt

$$\sum_{i=1}^{\bar{s}} \alpha_i\big((A^{-1}M)^{i-1}u\big) = \sum_{i=1}^{\bar{s}} \alpha_i \sum_{j=1}^{\bar{s}} \lambda_{\sigma(j)}^{-(i-1)} v_{\sigma(j)}$$

$$= \sum_{j=1}^{\bar{s}}\sum_{i=1}^{\bar{s}} v_{\sigma(j)}\big(\lambda_{\sigma(j)}^{-1}\big)^{i-1}\alpha_i = \big[v_{\sigma(1)},\dots,v_{\sigma(\bar{s})}\big]\tilde{M}\big(\alpha_1,\dots,\alpha_{\bar{s}}\big)^T,$$

   wobei $\tilde{M} \in \mathbb{R}^{\bar{s}\times\bar{s}}$ mit $(\tilde{M})_{ji} = \big(\lambda_{\sigma(j)}^{-1}\big)^{i-1}$ eine Vandermonde-Matrix ist. Da $\lambda_{\sigma(j)}^{-1}$ paarweise verschieden sind, gilt $\det(\tilde{M}) \neq 0$, so dass $\tilde{M}$ regulär ist. Außerdem sind die Eigenvektoren $v_{\sigma(j)}$ wegen der A-Orthogonalität beziehungsweise der M-Orthogonalität der Eigenräume linear unabhängig. Damit gilt die Implikation

$$\sum_{i=1}^{\bar{s}} \alpha_i\big((A^{-1}M)^{i-1}u\big) = \mathbf{0} \quad\Rightarrow\quad \big[v_{\sigma(1)},\dots,v_{\sigma(\bar{s})}\big]\tilde{M}\big(\alpha_1,\dots,\alpha_{\bar{s}}\big)^T = \mathbf{0}$$

$$\Rightarrow\quad \tilde{M}\big(\alpha_1,\dots,\alpha_{\bar{s}}\big)^T = \mathbf{0} \quad\Rightarrow\quad \big(\alpha_1,\dots,\alpha_{\bar{s}}\big)^T = \mathbf{0},$$

so dass die Vektoren $(A^{-1}M)^{i-1}u$ linear unabhängig sind.

Für $s \geq \bar{s}$ gilt dann wegen $\mathcal{K}^s_{A^{-1}M}(u) \subseteq \mathcal{V} = \mathcal{K}^{\bar{s}}_{A^{-1}M}(u)$ und $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u) \subseteq \mathcal{K}^s_{A^{-1}M}(u)$ die Übereinstimmung $\mathcal{K}^s_{A^{-1}M}(u) = \mathcal{K}^{\bar{s}}_{A^{-1}M}(u) = \mathcal{V}$. Weiter sind die Eigenvektoren $v_{\sigma(j)}$ Ritzvektoren bezüglich $\mathcal{K}^s_{A^{-1}M}(u)$, denn die Residuen $r(v_{\sigma(j)})$ sind Nullvektoren und damit Euklidisch orthogonal zu $\mathcal{K}^s_{A^{-1}M}(u)$. Durch $\lambda_{\sigma(j)}$ sind dann sämtliche Ritzwerte gegeben, so dass (3.2) einen Ritzvektor zum kleinsten Ritzwert $\lambda_{\sigma(1)}$, das heißt einen Eigenvektor zum Eigenwert $\lambda_{\sigma(1)}$, als die neue Iterierte u' liefert.

Für die Aussage zu $s < \bar{s}$ zeigen wir per Induktion, dass der Krylovraum $\mathcal{K}^k_{A^{-1}M}(u)$ für beliebiges $k = 1, \ldots, \bar{s}-1$ die Dimension k hat und keinen Eigenvektor enthält.

Induktionsanfang: Für $k = 1$ ist $\bar{s}$ mindestens 2, so dass u kein Eigenvektor ist. Wegen $\mathcal{K}^1_{A^{-1}M}(u) = \mathrm{span}\{u\}$ hat $\mathcal{K}^1_{A^{-1}M}(u)$ die Dimensionszahl 1 und enthält keinen Eigenvektor.

Induktionsschritt: $k \to k+1$ für $k < \bar{s} - 1$. Es gilt zunächst $(A^{-1}M)^k u \notin \mathcal{K}^k_{A^{-1}M}(u)$, denn sonst würde $\mathcal{K}^{k+1}_{A^{-1}M}(u) = \mathcal{K}^k_{A^{-1}M}(u)$ gelten und weiter würde nacheinander gezeigt, dass die Krylovräume $\mathcal{K}^{k+l}_{A^{-1}M}(u)$ für $l = 1, 2, \ldots$ sämtlich identisch zu $\mathcal{K}^k_{A^{-1}M}(u)$ wären und die gleiche Dimension k hätten. Dann hätte insbesondere $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u)$ die Dimension $k < \bar{s}-1$, das ist ein Widerspruch zum oben bewiesenen Ergebnis $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u) = \mathcal{V}$. Daher gilt $(A^{-1}M)^k u \notin \mathcal{K}^k_{A^{-1}M}(u)$, so dass $\mathcal{K}^{k+1}_{A^{-1}M}(u)$ die Dimension $k+1$ hat.

Weiter nehmen wir an, dass $\mathcal{K}^{k+1}_{A^{-1}M}(u)$ einen Eigenvektor v enthält. So lässt sich v durch eine Linearkombination $\sum_{i=1}^{k+1} \alpha_i \big((A^{-1}M)^{i-1}u\big)$ darstellen. Dabei gilt $\alpha_{k+1} \neq 0$, denn sonst würde v zu $\mathcal{K}^k_{A^{-1}M}(u)$ gehören, das ist ein Widerspruch zur Induktionsvoraussetzung, dass $\mathcal{K}^k_{A^{-1}M}(u)$ keinen Eigenvektor enthält. Aus der Eigengleichung $Av = \lambda Mv$ folgt $A^{-1}Mv = \lambda^{-1}v$. Einsetzen der Darstellung $v = \sum_{i=1}^{k+1} \alpha_i \big((A^{-1}M)^{i-1}u\big)$ liefert

$$\sum_{i=1}^{k+1} \alpha_i \big((A^{-1}M)^{i}u\big) = \sum_{i=1}^{k+1} \alpha_i \lambda^{-1} \big((A^{-1}M)^{i-1}u\big).$$

Wegen $\alpha_{k+1} \neq 0$ lässt sich $(A^{-1}M)^{k+1}u$ durch eine Linearkombination von $(A^{-1}M)^{i-1}u$, $i = 1, \ldots, k+1$, darstellen und gehört damit zu $\mathcal{K}^{k+1}_{A^{-1}M}(u)$. Daraus folgt $\mathcal{K}^{k+1+l}_{A^{-1}M}(u) = \mathcal{K}^{k+1}_{A^{-1}M}(u)$ für $l = 1, 2, \ldots$ und weiter $\dim \mathcal{K}^{\bar{s}}_{A^{-1}M}(u) = k+1 < \bar{s}$, was $\mathcal{K}^{\bar{s}}_{A^{-1}M}(u) = \mathcal{V}$ widerspricht. Daher enthält $\mathcal{K}^{k+1}_{A^{-1}M}(u)$ keinen Eigenvektor. □

In der Praxis tritt meistens der nichttriviale Fall $s < \bar{s}$ auf. Außerdem lässt sich $\mathcal{K}^s_{A^{-1}M}(u)$ durch ein modifiziertes Lanczos-Verfahren konstruieren, wobei vollständige Orthogonalisierungen zur Stabilität eingesetzt werden.

$$\begin{aligned}
&u_1 = u/\|u\|_M\,;\\
&\texttt{for } i = 1 : s{-}1 \texttt{ do}\\
&\quad \bar{u} = A^{-1}r(u_i)\,; \quad \tilde{u} = \bar{u} - \textstyle\sum_{j=1}^{i} u_j \big(u_j^T M \bar{u}\big)\,;\\
&\quad \texttt{if } \|\tilde{u}\|_M = 0 \texttt{ then break; else } u_{i+1} = \tilde{u}/\|\tilde{u}\|_M\,; \texttt{ end if}\\
&\texttt{end do}
\end{aligned} \tag{3.3}$$

Durch (3.3) werden M-orthonormale Basisvektoren für $\mathcal{K}^s_{A^{-1}M}(u)$ berechnet (das lässt sich per Induktion einfach überprüfen), welche als eine Basismatrix U zusammengesetzt werden können. Dann ist $U^T MU$ eine Einheitsmatrix, so dass die Ritzwerte und zugehörige Ritzvektoren durch Diagonalisierung der Matrix $U^T AU$ ermittelt werden können. Beim Aufbau von $\mathcal{K}^s_{A^{-1}M}(u)$ werden tatsächlich skalierte A-Gradienten, nämlich $A^{-1}r(u_i)$, verwendet, daher zählt (3.3) auch zu einem A-Gradientenverfahren. Bei der Konvergenzanalyse für (3.3) müssen wir nicht die Lanczos-Basisvektoren betrachten. Für nichttriviale Fälle $s < \bar{s}$ in Lemma 3.2 ist $\{(A^{-1}M)^{i-1}u \;;\; i =$

$1, \ldots, \bar{s}\}$ bereits eine Basis von $\mathcal{K}^s_{A^{-1}M}(u)$. Im nächsten Schritt der Konvergenznalalyse können wir der Einfachheit halber einige zu (3.2) äquivalente Verfahren durch Basiswechsel formulieren.

3.1.1 Niedrigdimensionale Formulierung

Um Konvergenzabschätzungen wie (2.2) des Krylovraum-Verfahrens (3.2) zu entwickeln, können wir den Krylovraum durch einen Ziel-Eigenvektor erweitern und damit niedrigdimensionale Darstellungen nutzen. Diese Idee wurde in [51] für Verfahren zu zweidimensionalen Krylovräumen vorgestellt. Hier verallgemeinern wir die Argumente aus [51] auf einen s-dimensionalen Fall.

Lemma 3.3. *Betrachten wir den nichttrivialen Fall $s < \bar{s}$ in Lemma 3.2. Sei $v_{\sigma(1)}$ der Anteil von u auf $\mathcal{V}_{\sigma(1)}$, dann hat der Unterraum $\mathcal{W} := \operatorname{span}\{v_{\sigma(1)}\} + \mathcal{K}^s_{A^{-1}M}(u)$ die Dimension $s+1$. Mit einer beliebigen A-orthonormalen Basismatrix $W \in \mathbb{R}^{n\times(s+1)}$ von $\mathcal{W}$ lässt sich eine symmetrische und positiv definite Matrix $H := W^T M W$ aufstellen. Weiter ist durch $x := W^T A u$ der Koeffizientenvektor von u bezüglich W gegeben, das heißt $u = Wx$. Damit lässt sich $\mathcal{K}^s_{A^{-1}M}(u)$ durch $W\mathcal{K}^s_H(x)$ darstellen. Für beliebiges $\tilde{x} \in \mathbb{R}^{s+1}\backslash\{\mathbf{0}\}$ wird der Rayleigh-Quotient bezüglich H durch*

$$\mu(\tilde{x}) := \frac{\tilde{x}^T H \tilde{x}}{\tilde{x}^T \tilde{x}} \tag{3.4}$$

definiert. Für beliebiges $\tilde{u} \in \mathcal{W}\backslash\{\mathbf{0}\}$ und dessen Koeffizientenvektor $\tilde{x} := W^T A \tilde{u}$ bezüglich W gilt die Rayleigh-Quotient-Umrechnung

$$\rho(\tilde{u}) = \frac{1}{\mu(\tilde{x})}\,. \tag{3.5}$$

Weiter lassen sich Ritzwerte und Ritzvektoren für H definieren, siehe Definition 2.3.

$x' := W^T A u'$ ist genau dann ein Ritzvektor zum größten Ritzwert von H bezüglich $\mathcal{K}^s_H(x)$, wenn u' ein Ritzvektor zum kleinsten Ritzwert von (A, M) bezüglich $\mathcal{K}^s_{A^{-1}M}(u)$ ist.

Beweis. Für $s < \bar{s}$ enthält $\mathcal{K}^s_{A^{-1}M}(u)$ nach Lemma 3.2 keinen Eigenvektor, das heißt $v_{\sigma(1)} \notin \mathcal{K}^s_{A^{-1}M}(u)$. Damit gilt $\dim \mathcal{W} = s + 1$.

Mit einer beliebigen A-orthonormalen Basismatrix $W \in \mathbb{R}^{n\times(s+1)}$ von $\mathcal{W}$ ist $H := W^T M W$ offenbar symmtrisch. Außerdem gilt für jedes $\tilde{x} \in \mathbb{R}^{s+1}\backslash\{\mathbf{0}\}$, dass $W\tilde{x} \neq \mathbf{0}$, und weiter $\tilde{x}^T H \tilde{x} = (W\tilde{x})^T M (W\tilde{x}) > 0$. Damit ist H positiv definit.

Wegen der A-orthonormalität von W gilt $W^T A W = I$, so dass die Projektionsmatrix von $\mathcal{W}$ durch $W(W^T A W)^{-1} W^T A = W W^T A$ gegeben ist, siehe Definition 2.2. Wegen $u \in \mathcal{W}$ gilt $u = W W^T A u = Wx$.

Die $\mathcal{W}$-Darstellungen der Basisvektoren $(A^{-1}M)^{i-1}u$ von $\mathcal{K}^s_{A^{-1}M}(u)$ lassen sich induktiv bestimmen. Für $i = 1$ gilt bereits $W^T A u = x$ beziehungsweise $u = Wx$. Gelte für $i = k < s$, dass $\big(W^T A\big)\big((A^{-1}M)^{k-1}u\big) = H^{k-1}x$ beziehungsweise $(A^{-1}M)^{k-1}u = W H^{k-1} x$, dann gilt für $i = k{+}1$, dass

$$\big(W^T A\big)\big((A^{-1}M)^k u\big) = W^T M\big((A^{-1}M)^{k-1}u\big) = W^T M W H^{k-1} x = H^k x$$

beziehungsweise $(A^{-1}M)^k u = W H^k x$. Damit erhalten wir $\big(W^T A\big)\mathcal{K}^s_{A^{-1}M}(u) = \mathcal{K}^s_H(x)$ und $\mathcal{K}^s_{A^{-1}M}(u) = W\mathcal{K}^s_H(x)$.

Für beliebiges $\tilde{u} \in \mathcal{W}\backslash\{\mathbf{0}\}$ mit $\tilde{x} = W^T A \tilde{u}$ beziehungsweise $\tilde{u} = W\tilde{x}$ gilt unmittelbar nach entsprechenden Definitionen, dass

$$\rho(\tilde{u}) = \frac{\tilde{u}^T A \tilde{u}}{\tilde{u}^T M \tilde{u}} = \frac{\tilde{x}^T W^T A W \tilde{x}}{\tilde{x}^T W^T M W \tilde{x}} = \frac{\tilde{x}^T \tilde{x}}{\tilde{x}^T H \tilde{x}} = \frac{1}{\mu(\tilde{x})}.$$

x' ist genau dann ein Ritzvektor zum größten Ritzwert von H bezüglich $\mathcal{K}_H^s(x)$, wenn

$$\mu(x') = \max_{\tilde{x}\in\mathcal{K}_H^s(x)\setminus\{\mathbf{0}\}} \mu(\tilde{x}),$$

vergleiche Satz 2.4. Mit der Darstellung $x' = W^T A u'$ beziehungsweise $u' = Wx'$ folgt aus der Gleichheit $\mathcal{K}_{A^{-1}M}^s(u) = W\mathcal{K}_H^s(x)$ und der Rayleigh-Quotient-Umrechnung, dass

$$\mu(x') = \max_{\tilde{x}\in\mathcal{K}_H^s(x)\setminus\{\mathbf{0}\}} \mu(\tilde{x}) \quad\Leftrightarrow\quad \frac{1}{\mu(x')} = \min_{\tilde{x}\in\mathcal{K}_H^s(x)\setminus\{\mathbf{0}\}} \frac{1}{\mu(\tilde{x})} \quad\Leftrightarrow\quad \rho(u') = \min_{\tilde{u}\in\mathcal{K}_{A^{-1}M}^s(u)\setminus\{\mathbf{0}\}} \rho(\tilde{u}).$$

Damit wird die behauptete Äquivalenz-Eigenschaft zwischen x' und u' gezeigt. □

Für beliebiges $\tilde{u} \in \mathcal{W}$ mit $\tilde{x} = W^T A\tilde{u}$ gelten außerdem die folgenden Norm-Umrechnungen:

$$\|\tilde{u}\|_A = \sqrt{\tilde{u}^T A\tilde{u}} = \sqrt{\tilde{x}^T W^T AW\tilde{x}} = \sqrt{\tilde{x}^T\tilde{x}} = \|\tilde{x}\|_2, \qquad \text{analog} \quad \|\tilde{u}\|_M = \|\tilde{x}\|_H.$$

So erhalten wir nach Lemma 3.3 für den nichttrivialen Fall $s < \bar{s}$ eine zu (3.2) äquivalente Vorschrift

$$x' = \mathrm{RR}_{\mu,\max}\big(\mathcal{K}_H^s(x)\big), \tag{3.6}$$

wobei $\mathrm{RR}_{\mu,\max}$ einen Ritzvektor zum größten Ritzwert liefert. Zu bemerken ist, dass (3.6) nur zur abstrakten Analyse dient und in der Praxis nicht einsetzbar ist, denn zur Konstruktion von $\mathcal{W}$ beziehungsweise H wird ein Ziel-Eigenvektor $v_{\sigma(1)}$ benötigt.

Zur Konvergenzanalyse von (3.2) können wir zunächst die Iteration (3.6) untersuchen und schließlich die möglichen Abschätzungen für (3.6) in Abschätzungen für (3.2) transformieren. Eine alternative Krylovraum-Umformulierung ist durch eine beliebige M-orthonormale Basismatrix $\bar{W}$ von $\mathcal{W}$ möglich. Mit der Hilfsmatrix $\bar{H} := \bar{W}^T A\bar{W}$ lässt sich $\mathcal{K}_{A^{-1}M}^s(u)$ durch $\bar{W}\mathcal{K}_{\bar{H}^{-1}}^s(\bar{x})$ mit $\bar{x} := \bar{W}^T Mu$ darstellen. Im Folgenden verwenden wir hauptsächlich die erste Krylovraum-Umformulierung. Die zweite Krylovraum-Umformulierung wird erst bei einer Abschätzungsergänzung für INVIT(2) weiter betrachtet.

3.1.2 Eine strikte Einschachtelung der Ritzwerte

Gemäß der Krylovraum-Umformulierung in Lemma 3.3 können wir nun die Konvergenzmaße umformulieren beziehungsweise durch Zwischengrößen beschränken. Dabei ist eine strikte Einschachtelung der Ritzwerte nützlich. Für Verfahren zu zweidimensionalen Krylovräumen wurde eine solche strikte Einschachtelung in [51] durch direkte Umrechnung nachgewiesen. Für eine größere und abstrakte Dimensionszahl s ist diese Methode wenig geeignet. Mit Anregungen aus der allgemeinen Krylovraum-Analyse in [79] wird hier eine andere Beweismethode entwickelt.

Lemma 3.4. *Mit den Formulierungen in Lemma 3.2 und Lemma 3.3 sei im Fall $s < \bar{s}$ die Matrix H definiert, deren Eigenwerte durch $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{s+1}$ und deren Ritzwerte bezüglich $\mathcal{K}_H^s(x)$ durch $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_s$ bezeichnet und angeordnet werden. Es gilt dann $\mu_1 = \lambda_{\sigma(1)}^{-1}$ und durch*

$$\tilde{x}_j := \Bigg(\prod_{i=1,\ i\neq j}^{s} (H - \theta_i I)\Bigg)\, x$$

wird für die Ritzwerte θ_i jeweils ein Ritzvektor gegeben. Die Residuen dieser Ritzvektoren sind alle durch $\tilde{r} := \big(\prod_{i=1}^s (H - \theta_i I)\big)\, x$ darstellbar. Die Eigenwerte und die Ritzwerte werden strikt eingeschachtelt:

$$\mu_1 > \theta_1 > \mu_2 > \theta_2 > \cdots > \mu_s > \theta_s > \mu_{s+1}. \tag{3.7}$$

Damit sind alle Eigenwerte von H einfach. Seien $y_1, y_2, \ldots, y_{s+1}$ zugehörige Euklidisch normierte Eigenvektoren, dann ist der Anteil $v_{\sigma(1)}$ von u auf $\mathcal{V}_{\sigma(1)}$ kollinear zu Wy_1 und es gilt

$$\frac{\tan\angle_M\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_M\big(u,\,\mathcal{V}_{\sigma(1)}\big)} = \left|\frac{\tan\angle_M\big(u', v_{\sigma(1)}\big)}{\tan\angle_M\big(u, v_{\sigma(1)}\big)}\right| = \left|\frac{\tan\angle_H(x', y_1)}{\tan\angle_H(x, y_1)}\right|, \tag{3.8}$$

$$\frac{\tan\angle_A\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_A\big(u,\,\mathcal{V}_{\sigma(1)}\big)} = \left|\frac{\tan\angle_A\big(u', v_{\sigma(1)}\big)}{\tan\angle_A\big(u, v_{\sigma(1)}\big)}\right| = \left|\frac{\tan\angle_2(x', y_1)}{\tan\angle_2(x, y_1)}\right|. \tag{3.9}$$

Gelte zusätzlich $\rho(u) \in \big(\lambda_{\sigma(1)},\,\lambda_{\sigma(2)}\big)$, so gilt

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(u)\big)} \leq \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\Delta\big(\mu_1, \mu_2, \mu(x)\big)}. \tag{3.10}$$

Beweis. Der Unterraum $\mathcal{W} = \operatorname{span}\{v_{\sigma(1)}\} + \mathcal{K}^s_{A^{-1}M}(u)$ ist eine Teilmenge der direkten Summe der (A, M)-Eigenräume, auf denen u Nichtnull-Anteile besitzt. Unter den entsprechenden (A, M)-Eigenwerten ist $\lambda_{\sigma(1)}$ der kleinste, so dass der (A, M)-Rayleigh-Quotient jedes Nichtnullvektors aus $\mathcal{W}$ nicht kleiner als $\lambda_{\sigma(1)}$ ist. Da $\mathcal{W}$ einen Eigenvektor $v_{\sigma(1)}$ zu $\lambda_{\sigma(1)}$ enthält, ist $\lambda_{\sigma(1)}$ das Minimum des (A, M)-Rayleigh-Quotienten auf $\mathcal{W}$. Nach (3.5) ist dann der Kehrwert von $\lambda_{\sigma(1)}$ genau das Maximum des H-Rayleigh-Quotienten, also der größte Eigenwert von H, damit gilt $\mu_1 = \lambda^{-1}_{\sigma(1)}$.

Weiter zeigen wir für beliebiges $j \in \{1, \ldots, s\}$, dass $\tilde{x}_j$ ein Ritzvektor zu θ_j ist. Es sind also $\tilde{x}_j \in \mathcal{K}^s_H(x)$ und $(H - \theta_j I)\,\tilde{x}_j \perp_2 \mathcal{K}^s_H(x)$ zu zeigen.

a) Da $\prod_{i=1,\ i\neq j}^{s}(H - \theta_i I)$ ein Polynom $(s-1)$-ten Grades von H ist, gehört $\tilde{x}_j$ zu $\mathcal{K}^s_H(x)$.

b) Um die Orthogonalität zu zeigen, konstruieren wir für $\mathcal{K}^s_H(x)$ eine Ritzbasis, also eine aus Ritzvektoren bestehende Basis. Mit einer beliebigen Euklidisch orthonormalen Basis X von $\mathcal{K}^s_H(x)$ ist $X^T H X$ symmetrisch. Dann gibt es nach dem Spektralsatz eine orthogonale Matrix Q, womit $Q^T X^T H X Q = \operatorname{diag}(\theta_1, \theta_2, \ldots, \theta_s)$ gilt. So ist XQ eine Basis von $\mathcal{K}^s_H(x)$ und deren Spaltenvektoren $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_s$ sind jeweils zugehörige Ritzvektoren zu Ritzwerten $\theta_1, \theta_2, \ldots, \theta_s$. Damit sind die Residuen $(H - \theta_k I)\,\bar{x}_k$ Euklidisch orthogonal zu $\mathcal{K}^s_H(x)$ und insbesondere zu $\tilde{x}_k$, so dass

$$0 = \big((H - \theta_k I)\,\bar{x}_k,\,\tilde{x}_k\big)_2 = \left(\bar{x}_k,\,\Big(\prod_{i=1}^{s}(H - \theta_i I)\Big)\,x\right)_2 = \big(\bar{x}_k,\,(H - \theta_j I)\,\tilde{x}_j\big)_2.$$

Damit ist $(H - \theta_j I)\,\tilde{x}_j$ Euklidisch orthogonal zu jedem $\bar{x}_k$ und damit zu $\mathcal{K}^s_H(x)$.

Außerdem gilt offenbar $(H - \theta_j I)\,\tilde{x}_j = \tilde{r}$ für beliebiges $j \in \{1, \ldots, s\}$.

Zum Beweis der strikten Einschachtelung zeigen wir zunächst $\mu_1 > \theta_1$. Mittels der Euler-Lagrange-Gleichung (vergleiche Argumente vor Definition 2.3) lässt sich zeigen, dass jede Minimalstelle des (A, M)-Rayleigh-Quotienten auf $\mathcal{W}$ ein Ritzvektor zum kleinsten Ritzwert $\lambda_{\sigma(1)}$ ist. Da $\lambda_{\sigma(1)}$ ein Eigenwert ist, sind solche Minimalstellen auch Eigenvektoren. Im vorausgesetzten nichttrivialen Fall $s < \bar{s}$ enthält $\mathcal{K}^s_{A^{-1}M}(u)$ keinen Eigenvektor. Wegen $u' \in \mathcal{K}^s_{A^{-1}M}(u) \subset \mathcal{W}$ gilt also $\lambda_{\sigma(1)} < \rho(u')$. Mit $\mu_1 = \lambda^{-1}_{\sigma(1)}$ und der Rayleigh-Quotient-Umrechnung (3.5) gilt anschließend $\mu_1 > \mu(x') = \theta_1$. Nach dem verallgemeinerten Courant-Fischer-Prinzip aus Satz 2.4 angewandt auf H und $\mathcal{K}^s_H(x)$ gilt weiter die Einschachtelung

$$\mu_1 > \theta_1 \geq \mu_2 \geq \theta_2 \geq \cdots \geq \mu_s \geq \theta_s \geq \mu_{s+1}. \tag{3.11}$$

Wegen $\mathcal{K}^s_{A^{-1}M}(u) = W\mathcal{K}^s_H(x)$ hat $\mathcal{K}^s_H(x)$ wie $\mathcal{K}^s_{A^{-1}M}(u)$ die Dimension s, so dass es mindestens s verschiedene H-Eigenwerte gibt, auf deren Eigenräumen x Nichtnull-Anteile besitzt (vergleiche Lemma 3.2).

Würde x auf genau s Eigenräumen Nichtnull-Anteile besitzen, wäre dann $\mathcal{K}_H^s(x)$ ein H-invarianter Unterraum und alle Ritzwerte von H bezüglich $\mathcal{K}_H^s(x)$ wären Eigenwerte (dafür betrachten wir einen beliebigen Ritzwert θ mit einem zugehörigen Ritzvektor $\bar{x}$. Das Residuum $(H-\theta I)\,\bar{x}$ ist wegen der H-Invarianz von $\mathcal{K}_H^s(x)$ durch $\mathcal{K}_H^s(x)$ enthalten und damit Euklidisch orthogonal zu sich selbst, das heißt das Residuum ist ein Nullvektor, so dass θ und $\bar{x}$ die Eigengleichung $H\bar{x} = \theta\bar{x}$ erfüllen und θ ein Eigenwert wird). Außerdem gilt nach (3.11), dass alle Ritzwerte mit dem Eigenwert μ_1 unterschiedlich sind. Damit wäre jeder Ritzvektor auf $\mathcal{K}_H^s(x)$ ein Eigenvektor zu einem von μ_1 verschiedenen Eigenwert und damit Euklidisch orthogonal zu beliebigem Eigenvektor zu μ_1.

Um einen Widerspruch zu konstruieren, wählen wir einen Eigenvektor zu μ_1, nämlich den Koeffizientenvektor $W^TAv_{\sigma(1)}$. Die zugehörige Eigengleichung lässt sich wie folgt zeigen

$$H\big(W^TAv_{\sigma(1)}\big) = W^TMWW^TAv_{\sigma(1)} = W^TMv_{\sigma(1)} = W^T\lambda_{\sigma(1)}^{-1}Av_{\sigma(1)} = \mu_1\big(W^TAv_{\sigma(1)}\big). \tag{3.12}$$

Weiter würde also $W^TAv_{\sigma(1)} \perp_2 \mathcal{K}_H^s(x)$ gelten, was äquivalent zu $Av_{\sigma(1)} \perp_2 W\mathcal{K}_H^s(x)$ beziehungsweise $v_{\sigma(1)} \perp_A \mathcal{K}_{A^{-1}M}^s(u)$ ist. Insbesondere gilt $v_{\sigma(1)} \perp_A u$. Das ist ein Widerspruch dazu, dass u auf dem Eigenraum $\mathcal{V}_{\sigma(1)}$ einen Nichtnull-Anteil hat.

Daher muss es $s+1$ verschiedene H-Eigenwerte geben, zu deren Eigenräumen x nicht Euklidisch orthogonal liegt. Dann wird analog zum Beweis von Lemma 3.2 gezeigt, dass $\mathcal{K}_H^s(x)$ keinen Eigenvektor enthält (das zeigt nicht sofort die strikte Einschachtelung der Ritzwerte). Wäre ein Ritzwert θ_k zu einem Eigenwert μ identisch, betrachten wir dann einen beliebigen Eigenvektor y zu μ, damit $Hy = \mu y = \theta_k y$. Weiter nutzen wir die (durch Polynome definierten) Ritzvektoren $\tilde{x}_j$ und deren identische Residuum $\tilde{r}$. Es gilt

$$\big(\tilde{r},\, y\big)_2 = \Big(\Big(\prod_{i=1}^{s}(H-\theta_i I)\Big)\,x,\, y\Big)_2 = \Big(\Big(\prod_{i=1,\ i\neq k}^{s}(H-\theta_i I)\Big)\,x,\, \underbrace{(H-\theta_k I)\,y}_{=\mathbf{0}}\Big)_2 = 0.$$

Da $\mathcal{K}_H^s(x)$ keinen Eigenvektor enthält, ist das Residuum $\tilde{r}$ kein Nullvektor, so dass y zum Euklidisch orthogonalen Komplement von $\tilde{r}$ gehört. Jedoch ist dieses Komplement identisch zum Krylovraum $\mathcal{K}_H^s(x)$, der den Eigenvektor y nicht enthalten sollte. So entsteht ein Widerspruch und es gilt die strikte Einschachtelung (3.7).

Mit (3.12) wurde oben gezeigt, dass $W^TAv_{\sigma(1)}$ ein Eigenvektor zum Eigenwert μ_1 von H ist. Sei y_1 ein Euklidisch normailisierter Eigenvektor zu μ_1, dann sind $W^TAv_{\sigma(1)}$ und y_1 kollinear, so dass $v_{\sigma(1)}$ und Wy_1 auch kollinear sind. Es gibt also ein $\alpha \in \mathbb{R}\backslash\{0\}$ mit $v_{\sigma(1)} = \alpha W y_1$.

$v_{\sigma(1)}$ ist sowohl die M-Projektion von u auf $\mathcal{V}_{\sigma(1)}$ als auch die M-Projektion von u auf $\operatorname{span}\{v_{\sigma(1)}\}$, so dass $\tan\angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big) = \tan\angle_M\big(u,\, v_{\sigma(1)}\big)$ gilt und der Winkel spitz ist (siehe Definition 2.2). Weiter gilt mit $u = Wx$, $v_{\sigma(1)} = \alpha W y_1$ sowie $H = W^TMW$, dass

$$0 < \cos\angle_M\big(u, v_{\sigma(1)}\big) = \frac{\big(u, v_{\sigma(1)}\big)_M}{\|u\|_M\|v_{\sigma(1)}\|_M} = \frac{\big(Wx, \alpha W y_1\big)_M}{\|Wx\|_M\|\alpha W y_1\|_M} = \frac{\alpha\big(x, y_1\big)_H}{\|x\|_H|\alpha|\|y_1\|_H} = \operatorname{sgn}(\alpha)\cos\angle_H(x, y_1).$$

Damit gilt $\tan\angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big) = \tan\angle_M\big(u,\, v_{\sigma(1)}\big) = |\tan\angle_H(x, y_1)|$. Die M-Projektionen der Basisvektoren $(A^{-1}M)^{i-1}u$ von $\mathcal{K}_{A^{-1}M}^s(u)$ auf $\mathcal{V}_{\sigma(1)}$ sind kollinear zu $v_{\sigma(1)}$, so dass die M-Projektion $v'_{\sigma(1)}$ von u' auf $\mathcal{V}_{\sigma(1)}$ auch kollinear zu $v_{\sigma(1)}$ ist. Es gibt also ein $\beta \in \mathbb{R}$ mit $v'_{\sigma(1)} = \beta v_{\sigma(1)}$. Es gilt $\beta \neq 0$, denn sonst wäre u' A-orthogonal zu $v_{\sigma(1)}$ beziehungsweise x' Euklidisch orthogonal zu y_1, damit würde x' durch $\operatorname{span}\{y_2, \ldots, y_{s+1}\}$ enthalten, wobei $y_2, \ldots, y_{s+1}$ Euklidisch normierte Eigenvektoren zu $\mu_2, \ldots, \mu_{s+1}$ sind. Nach dem verallgemeinerten Courant-Fischer-Prinzip gilt dann $\mu_2 \geq \mu(x') = \theta_1$, was ein Widerspruch zur oben bewiesenen strikten Einschachtelung (3.7) ist. Weiter gilt mit $v'_{\sigma(1)} = \beta v_{\sigma(1)}$, $u' = Wx'$, $v_{\sigma(1)} = \alpha W y_1$ sowie $H = W^TMW$, dass $\tan\angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big) = |\tan\angle_M\big(u',\, v_{\sigma(1)}\big)| = |\tan\angle_H(x', y_1)|$. Damit wird (3.8) gezeigt. (3.9) lässt

sich analog zeigen. (Dabei können wir $Av_{\sigma(1)} = \lambda_{\sigma(1)} M v_{\sigma(1)},\ \lambda_{\sigma(1)} > 0$ nutzen, um M-Begriffe in A-Begriffe zu transformieren.)

Gelte zusätzlich $\rho(u) \in \big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}\big)$, dann liefert $0 < \lambda_{\sigma(1)} < \rho(u') \leq \rho(u) < \lambda_{\sigma(2)}$ die Ungleichungen

$$\lambda_{\sigma(1)}^{-1} > \rho(u')^{-1} \geq \rho(u)^{-1} > \lambda_{\sigma(2)}^{-1}.$$

Nach dem verallgemeinerten Courant-Fischer-Prinzip gilt außerdem

$$\mu_1 = \lambda_{\sigma(1)}^{-1} > \lambda_{\sigma(2)}^{-1} \geq \mu_2.$$

Zusammen mit $\rho(u')^{-1} = \mu(x')$ und $\rho(u)^{-1} = \mu(x)$ wird anschließend (3.10) gezeigt. □

Weiter können wir mit einer niedrigdimensionalen Anaylse gewünschte Abschätzungen bezüglich gewisser Hilfsgrößen konstruieren. Als Motivation betrachten wir zunächst den Fall $s = 2$, also für das INVIT(2)-Verfahren.

3.2 Konvergenzanalyse für Verfahren zu zweidimensionalen Krylovräumen

Gemäß Lemma 3.2, Lemma 3.3 und Lemma 3.4 ist für den nichttrivialen Fall ein s-dimensionaler Krylovraum in $\mathbb{R}^{s+1}$ zu untersuchen. Für $s = 2$ entsteht dann die folgende Voraussetzung für die niedrigdimensionale Analyse.

Vorausetzung 3.5. *Eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{3\times 3}$ besitze die Eigenwerte μ_1, μ_2, μ_3, und θ_1, θ_2 seien die Ritzwerte von H bezüglich eines Krylovraums $\mathcal{K}_H^2(x)$. Der Rayleigh-Quotient $\mu(\cdot)$ bezüglich H sei wie in (3.4) definiert und es gelte eine strikte Einschachtelung $\mu_1 > \theta_1 > \mu_2 > \theta_2 > \mu_3$. Weiter seien y_1, y_2, y_3 Euklidisch normierte Eigenvektoren zu μ_1, μ_2, μ_3 und x' ein Ritzvektor zu θ_1.*

3.2.1 Klassische Konvergenzresultate

Wir zeigen nun zwei klassische Abschätzungen mit der Beweistechnik aus [51] und einigen Modifizierungen.

Satz 3.6. *Unter Voraussetzung 3.5 gilt*

$$\left|\frac{\tan \angle_2(x', y_1)}{\tan \angle_2(x, y_1)}\right| \leq \kappa \quad \textit{mit} \quad \kappa := \frac{\mu_2 - \mu_3}{\mu_1 - \mu_3}. \tag{3.13}$$

Gelte zusätzlich $\mu(x) \in (\mu_2,\ \mu_1)$, dann gilt

$$0 < \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\Delta\big(\mu_1, \mu_2, \mu(x)\big)} \leq \left(\frac{\kappa}{2 - \kappa}\right)^2. \tag{3.14}$$

Die oberen Schranken lassen sich in Grenzfällen erreichen. (Damit sind (3.13) und (3.14) scharfe Abschätzungen.)

Beweis. Bezüglich der Eigenvektoren y_i lassen sich x und x' durch

$$x = \sum_{i=1}^{3} \xi_i\, y_i, \qquad x' = \sum_{i=1}^{3} \xi_i'\, y_i$$

darstellen. Gemäß Voraussetzung 3.5 sind ξ_1^2, ${\xi'}_1^2$ sowie $\xi_2^2+\xi_3^2$, ${\xi'}_2^2+{\xi'}_3^2$ von Null verschieden, denn sonst würde die strikte Einschachtelung nicht gelten. Weiter betrachten wir

$$\left(\frac{\tan\angle_2(x',y_1)}{\tan\angle_2(x,y_1)}\right)^2 = \frac{\left({\xi'}_2^2+{\xi'}_3^2\right)/{\xi'}_1^2}{\left(\xi_2^2+\xi_3^2\right)/\xi_1^2}. \tag{3.15}$$

Um diesen Quotienten mit Eigenwerten zu kombinieren, betrachten wir die Ritzvektoren (vergleiche Lemma 3.4)

$$\tilde{x}_1 = (H-\theta_2 I)x, \qquad \tilde{x}_2 = (H-\theta_1 I)x$$

und deren identisches Residuum $\tilde{r} = (H-\theta_1 I)(H-\theta_2 I)x$. Wegen der Euklidischen Orthogonalität $\tilde{r} \perp_2 \mathcal{K}_H^2(x)$ gilt $\tilde{r}^T\tilde{x}_1 = \tilde{r}^T\tilde{x}_2 = 0$, so dass

$$x^T(H-\theta_1 I)(H-\theta_2 I)^2 x = 0 \quad \text{und} \quad x^T(H-\theta_1 I)^2(H-\theta_2 I)x = 0.$$

Einsetzen der Basisdarstellung $x = \sum_{i=1}^3 \xi_i\, y_i$ liefert dann ein Gleichungssystem

$$\begin{cases} (\mu_1-\theta_1)(\mu_1-\theta_2)^2\xi_1^2 + (\mu_2-\theta_1)(\mu_2-\theta_2)^2\xi_2^2 + (\mu_3-\theta_1)(\mu_3-\theta_2)^2\xi_3^2 = 0 \\ (\mu_1-\theta_1)^2(\mu_1-\theta_2)\xi_1^2 + (\mu_2-\theta_1)^2(\mu_2-\theta_2)\xi_2^2 + (\mu_3-\theta_1)^2(\mu_3-\theta_2)\xi_3^2 = 0 \end{cases},$$

welches wegen $\xi_1^2 \neq 0$ und der strikten Einschachtelung der Ritzwerte nach ξ_2^2/ξ_1^2 und ξ_3^2/ξ_1^2 aufgelöst werden kann:

$$\frac{\xi_2^2}{\xi_1^2} = \frac{(\mu_1-\theta_1)(\mu_1-\theta_2)(\mu_1-\mu_3)}{(\theta_1-\mu_2)(\mu_2-\theta_2)(\mu_2-\mu_3)}, \qquad \frac{\xi_3^2}{\xi_1^2} = \frac{(\mu_1-\theta_1)(\mu_1-\theta_2)(\mu_1-\mu_2)}{(\theta_1-\mu_3)(\theta_2-\mu_3)(\mu_2-\mu_3)}.$$

x' ist ein Ritzvektor zu θ_1 und damit kollinear zu $\tilde{x}_1$. Es gibt also ein $\zeta \in \mathbb{R}\backslash\{0\}$ mit

$$x' = \zeta\,\tilde{x}_1 = \zeta\,(H-\theta_2 I)x = \zeta(H-\theta_2 I)\sum_{i=1}^3 \xi_i\, y_i = \sum_{i=1}^3 \zeta(\mu_i-\theta_2)\xi_i\, y_i.$$

Ein Koeffizientenvergleich mit der Darstellung $x' = \sum_{i=1}^3 \xi'_i\, y_i$ liefert

$$\frac{{\xi'}_2^2}{{\xi'}_1^2} = \left(\frac{\mu_2-\theta_2}{\mu_1-\theta_2}\right)^2 \frac{\xi_2^2}{\xi_1^2} = \frac{(\mu_1-\theta_1)(\mu_2-\theta_2)(\mu_1-\mu_3)}{(\theta_1-\mu_2)(\mu_1-\theta_2)(\mu_2-\mu_3)},$$

$$\frac{{\xi'}_3^2}{{\xi'}_1^2} = \left(\frac{\mu_3-\theta_2}{\mu_1-\theta_2}\right)^2 \frac{\xi_3^2}{\xi_1^2} = \frac{(\mu_1-\theta_1)(\theta_2-\mu_3)(\mu_1-\mu_2)}{(\theta_1-\mu_3)(\mu_1-\theta_2)(\mu_2-\mu_3)}.$$

Damit lässt sich der Quotient in (3.15) durch Eigenwerte und Ritzwerte darstellen. Jedoch hat eine solche Darstellung zu viele Parameter. Eine deutlich klarere Darstellung erhält man durch die Substitution

$$\kappa := \frac{\mu_2-\mu_3}{\mu_1-\mu_3}, \quad \nu_1 := \frac{\theta_1-\mu_3}{\mu_1-\mu_3}, \quad \nu_2 := \frac{\theta_2-\mu_3}{\mu_1-\mu_3}, \tag{3.16}$$

wofür die Bedingungen

$$\kappa \in (0,1), \quad \nu_1 \in (\kappa,1), \quad \nu_2 \in (0,\kappa) \tag{3.17}$$

wegen der strikten Einschachtelung der Ritzwerte gelten. Mit (3.16) erhalten wir neue Formen der Quotienten:

$$\begin{aligned} \frac{\xi_2^2}{\xi_1^2} &= \frac{(1-\nu_1)(1-\nu_2)}{(\nu_1-\kappa)(\kappa-\nu_2)\kappa}, & \frac{\xi_3^2}{\xi_1^2} &= \frac{(1-\nu_1)(1-\nu_2)(1-\kappa)}{\nu_1\nu_2\kappa}, \\ \frac{{\xi'}_2^2}{{\xi'}_1^2} &= \frac{(1-\nu_1)(\kappa-\nu_2)}{(\nu_1-\kappa)(1-\nu_2)\kappa}, & \frac{{\xi'}_3^2}{{\xi'}_1^2} &= \frac{(1-\nu_1)\nu_2(1-\kappa)}{\nu_1(1-\nu_2)\kappa}. \end{aligned} \tag{3.18}$$

Einsetzen in (3.15) liefert

$$\left(\frac{\tan\angle_2(x',y_1)}{\tan\angle_2(x,y_1)}\right)^2 = \frac{\nu_2(\kappa-\nu_2)(\nu_1-\nu_2-\nu_1\nu_2+\kappa\nu_2)}{(1-\nu_2)^2(\nu_1+\nu_2-\kappa+\nu_1\nu_2-\kappa\nu_1-\kappa\nu_2+\kappa^2)}.$$

Für festes κ können wir diese Darstellung als eine Funktion f von Variablen ν_1 und ν_2 betrachten und dann gemäß (3.17) das Supremum von f auf $(\kappa,1)\times(0,\kappa)$ beziehungsweise das Maximum von f auf $[\kappa,1]\times[0,\kappa]$ untersuchen. Dafür ist die partielle Ableitung von f nach ν_1 zu nutzen. Mit

$$\frac{\partial f}{\partial\nu_1}(\nu_1,\nu_2) = \frac{\nu_2(\kappa-\nu_2)(1-\kappa)(2\nu_2-\kappa)}{(1-\nu_2)^2(\nu_1+\nu_2-\kappa+\nu_1\nu_2-\kappa\nu_1-\kappa\nu_2+\kappa^2)^2} \begin{cases} \leq 0 & \text{für } \nu_2\in[0,\kappa/2) \\ \geq 0 & \text{für } \nu_2\in[\kappa/2,\kappa] \end{cases}$$

gilt

$$f(\nu_1,\nu_2) \begin{cases} \leq f(\kappa,\nu_2) =: g_1(\nu_2) & \text{für } \nu_2\in[0,\kappa/2), \\ \leq f(1,\nu_2) =: g_2(\nu_2) & \text{für } \nu_2\in[\kappa/2,\kappa]. \end{cases}$$

Weiter sind Maxima von g_1, g_2 auf deren Definitionsintervallen zu bestimmen. Für $\nu_2\in[0,\kappa/2)$ gilt

$$g_1(\nu_2) = f(\kappa,\nu_2) = \frac{(\kappa-\nu_2)^2}{(1-\nu_2)^2}, \qquad \frac{dg_1}{d\nu_2}(\nu_2) = \frac{2(1-\kappa)(\kappa-\nu_2)}{-(1-\nu_2)^3} < 0,$$

so dass $g_1(\nu_2)\leq g_1(0)=\kappa^2$. Für $\nu_2\in[\kappa/2,\kappa]$ gilt

$$g_2(\nu_2) = f(1,\nu_2) = \frac{\nu_2(\kappa-\nu_2)(1-2\nu_2+\kappa\nu_2)}{(1-\nu_2)^2(1+2\nu_2-2\kappa-\kappa\nu_2+\kappa^2)},$$

$$\frac{dg_2}{d\nu_2}(\nu_2) = \frac{(1-\kappa)p(\nu_2)}{(1-\nu_2)^3(1+2\nu_2-2\kappa-\kappa\nu_2+\kappa^2)^2},$$

wobei p ein kubisches Polynom ist:

$$p(\nu_2) = \kappa(1-\kappa)-(1-\kappa)(2-\kappa)(1+2\kappa)\nu_2+2(1-2\kappa)(2-\kappa)\nu_2^2+2(2-\kappa)\nu_2^3.$$

Es gilt $p(\nu_2)<0$ für $\nu_2\in[\kappa/2,\kappa]$, denn mit $\kappa\in(0,1)$ gilt

$$\lim_{\nu_2\to-\infty} p(\nu_2) = \lim_{\nu_2\to-\infty} p(\nu_2)/\nu_2^3 \cdot \lim_{\nu_2\to-\infty}\nu_2^3 = \underbrace{2(2-\kappa)}_{>0}\lim_{\nu_2\to-\infty}\nu_2^3 < 0,$$

$$p(0) = \kappa(1-\kappa) > 0,$$

$$p(\kappa/2) = -\frac{1}{4}\kappa^2(2-2\kappa+\kappa^2) < 0,$$

$$p(\kappa) = -\kappa(1-\kappa)^2 < 0,$$

$$\lim_{\nu_2\to\infty} p(\nu_2) = \lim_{\nu_2\to\infty} p(\nu_2)/\nu_2^3 \cdot \lim_{\nu_2\to\infty}\nu_2^3 = \underbrace{2(2-\kappa)}_{>0}\lim_{\nu_2\to\infty}\nu_2^3 > 0,$$

so dass p genau drei reelle Nullstellen hat, welche sich jeweils in Intervallen $(-\infty,0)$, $(0,\kappa/2)$ und (κ,∞) befinden. Daher hat p keinen Vorzeichenwechsel im Intervall $[\kappa/2,\kappa]$, das heißt, für $\nu_2\in[\kappa/2,\kappa]$ ist $p(\nu_2)$ wie $p(\kappa/2)$ negativ. Damit gilt für $\nu_2\in[\kappa/2,\kappa]$, dass

$$\frac{dg_2}{d\nu_2}(\nu_2) < 0 \quad \text{und damit} \quad g_2(\nu_2)\leq g_2(\kappa/2) = \left(\frac{\kappa}{2-\kappa}\right)^2.$$

Kombination der obigen Ergebnisse liefert $\left|\frac{\tan\angle_2(x',y_1)}{\tan\angle_2(x,y_1)}\right| = \sqrt{f(\nu_1,\nu_2)} \leq \kappa$ und zeigt (3.13).

Die Schranke κ ist der Grenzwert von $\sqrt{f(\nu_1,\nu_2)}$ für $\nu_1 \to \kappa$ und $\nu_2 \to 0$, das heißt der Grenzwert des Betrags des Tangens-Quotienten für $\theta_1 \to \mu_2$ und $\theta_2 \to \mu_3$.

Gelte zusätzlich $\mu(x) \in (\mu_2, \mu_1)$, dann gilt $\Delta\big(\mu_1,\mu_2,\mu(x)\big) = \big(\mu(x)-\mu_1\big)/\big(\mu_2-\mu(x)\big) > 0$. Außerdem gilt $\Delta\big(\mu_1,\mu_2,\mu(x')\big) > 0$ wegen $\mu(x') = \theta_1$ und der strikten Einschachtelung. Der Delta-Quotient ist also positiv.

Zur Konstruktion einer oberen Schranke können wir den Delta-Quotienten wieder mittels der Parameter aus (3.16) umschreiben. Mit $x = \sum_{i=1}^{3} \xi_i\, y_i$ erhalten wir

$$\mu(x) = \frac{\mu_1\xi_1^2 + \mu_2\xi_1^2 + \mu_3\xi_3^2}{\xi_1^2+\xi_2^2+\xi_3^2} \quad \Rightarrow \quad \Delta\big(\mu_1,\mu_2,\mu(x)\big) = \frac{(\mu_1-\mu_2)\xi_2^2/\xi_1^2 + (\mu_1-\mu_3)\xi_3^2/\xi_1^2}{(\mu_1-\mu_2) - (\mu_2-\mu_3)\xi_3^2/\xi_1^2}$$

und weiter

$$\begin{aligned}\frac{\Delta\big(\mu_1,\mu_2,\mu(x')\big)}{\Delta\big(\mu_1,\mu_2,\mu(x)\big)} &= \frac{\theta_1-\mu_1}{\mu_2-\theta_1} \cdot \frac{(\mu_1-\mu_2) - (\mu_2-\mu_3)\xi_3^2/\xi_1^2}{(\mu_1-\mu_2)\xi_2^2/\xi_1^2 + (\mu_1-\mu_3)\xi_3^2/\xi_1^2} \\ &= \frac{1-\nu_1}{\nu_1-\kappa} \cdot \frac{1-\left(\dfrac{\kappa}{1-\kappa}\right)\xi_3^2/\xi_1^2}{\xi_2^2/\xi_1^2 + \left(\dfrac{1}{1-\kappa}\right)\xi_3^2/\xi_1^2} = \frac{(\kappa-\nu_2)(\nu_1+\nu_2-1)}{(1-\nu_2)(\nu_1+\nu_2-\kappa)}.\end{aligned}$$

Betrachten wir diese Darstellung als eine Funktion $\tilde{f}$ von Variablen ν_1 und ν_2 und bestimmen analog zum Beweis der ersten Abschätzung das Maximum von $\tilde{f}$ auf $[\kappa,1]\times[0,\kappa]$. Für $\nu_2 \in [0,\kappa]$ ist die partielle Ableitung von $\tilde{f}$ nach ν_1 nicht negativ:

$$\frac{\partial \tilde{f}}{\partial \nu_1}(\nu_1,\nu_2) = \frac{(\kappa-\nu_2)(1-\kappa)}{(1-\nu_2)(\nu_1+\nu_2-\kappa)^2} \geq 0.$$

Damit gilt

$$\tilde{f}(\nu_1,\nu_2) \leq \tilde{f}(1,\nu_2) = \frac{(\kappa-\nu_2)\nu_2}{(1-\nu_2)(1+\nu_2-\kappa)} =: \tilde{g}(\nu_2).$$

Nach Betrachtung der Ableitung

$$\frac{d\tilde{g}}{d\nu_2}(\nu_2) = \frac{(1-\kappa)(\kappa-2\nu_2)}{(1-\nu_2)^2(1+\nu_2-\kappa)^2}$$

wird das Maximum von $\tilde{g}$ auf dem Intervall $[0,\kappa]$ an der Stelle $\nu_2 = \kappa/2$ angenommen. Damit gilt

$$\frac{\Delta\big(\mu_1,\mu_2,\mu(x')\big)}{\Delta\big(\mu_1,\mu_2,\mu(x)\big)} = \tilde{f}(\nu_1,\nu_2) \leq \tilde{f}(1,\nu_2) = \tilde{g}(\nu_2) \leq \tilde{g}(\kappa/2) = \left(\frac{\kappa}{2-\kappa}\right)^2.$$

Diese Schranke ist der Grenzwert von $\tilde{f}(\nu_1,\nu_2)$ für $\nu_1 \to 1$ und $\nu_2 \to \kappa/2$, das heißt der Grenzwert des Delta-Quotienten für $\theta_1 \to \mu_2$ und $\theta_2 \to (\mu_2+\mu_3)/2$. □

3.2.2 Verschärfte Konvergenzabschätzung für Ritzvektoren

In numerischen Experimenten haben wir beobachtet, dass der Betrag des Tangens-Quotienten aus (3.13) für x mit $\mu(x) \in (\mu_2,\mu_1)$ eine lokal kleinere obere Schranke haben soll. Die von uns vermutete Schranke lässt sich interessanterweise durch eine Konvexkombination von $\kappa/(2-\kappa)$ und κ darstellen, nämlich

$$\frac{\kappa(1+\kappa-\kappa^2)}{2-\kappa} = \frac{\kappa}{2-\kappa}\cdot(1-\kappa) + \kappa\cdot\kappa.$$

Zum Beweis dieser Vermutung sind die Nullstellen von einem Polynom vierten Grades zu untersuchen, wofür die folgende Diskriminantenformel nützlich ist.

Satz 3.7. *Ein Polynom $q(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4$ vierten Grades und reellen Koeffizienten besitzt zwei einfache reelle Nullstellen und zwei nichtreelle Nullstellen genau dann, wenn die Diskriminante*

$$\begin{aligned}\mathcal{D}(q) := (a_1a_2a_3)^2 - 4(a_1a_3)^3 - 4a_1^2a_2^3a_4 + 18a_1^3a_2a_3a_4 - 27a_1^4a_4^2 + 256a_0^3a_4^3\\ +a_0(-4a_2^3a_3^2 + 18a_1a_2a_3^3 + 16a_2^4a_4 - 80a_1a_2^2a_3a_4 - 6a_1^2a_3^2a_4 + 144a_1^2a_2a_4^2)\\ +a_0^2(-27a_3^4 + 144a_2a_3^2a_4 - 128a_2^2a_4^2 - 192a_1a_3a_4^2)\end{aligned}$$

negativ ist.

Beweis. Diese Diskriminantenformel wurde in [8] vorgestellt. Es ist einfach (aber aufwendig) zu überprüfen, dass $\mathcal{D}(q)$ auch durch

$$a_4^6 \prod_{j=2}^{4} \prod_{i=1}^{j-1} (t_i - t_j)^2$$

darstellbar ist, wobei $t_1, \dots, t_4 \in \mathbb{C}$ die vier Nullstellen von $q(t)$ sind. (Setze dafür $q = a_4 \prod_{i=1}^{4}(t - t_i)$, dann sind die Koeffizienten $a_0, \dots, a_3$ durch a_4 und die Nullstellen $t_1, \dots, t_4$ darstellbar. Weiter ist nur ein aufwendiger Vergleich der zwei Formen von $\mathcal{D}(q)$ nötig.)

Da $q(t)$ reelle Koeffizienten besitzt, treten die nichtreellen Nullstellen in konjugierten Paaren auf (denn $q(t) = 0 \Rightarrow q(\bar{t}) = \overline{q(t)} = \bar{0} = 0$). Betrachten wir damit die folgende Fallunterscheidung.

1) Habe $q(t)$ ein Paar nichtreeller Nullstellen und zwei verschiedene reelle Nullstellen, dann gilt $\mathcal{D}(q) < 0$. Zum Beweis seien o.B.d.A. $t_1 = \alpha + \mathrm{i}\beta$, $t_2 = \alpha - \mathrm{i}\beta$, $\alpha, \beta, t_3, t_4 \in \mathbb{R}$, $\beta \neq 0$, $t_3 \neq t_4$, damit gilt

$$(t_1 - t_2)^2 = -4\beta^2 < 0, \qquad (t_3 - t_4)^2 > 0,$$

$$(t_1-t_3)^2(t_2-t_3)^2 = \big((t_1-t_3)(t_2-t_3)\big)^2 = \big((\alpha-t_3+\mathrm{i}\beta)(\alpha-t_3-\mathrm{i}\beta)\big)^2 = \big((\alpha-t_3)^2+\beta^2\big)^2 > 0,$$

und analog $(t_1 - t_4)^2(t_2 - t_4)^2 > 0$, so dass $\mathcal{D}(q) = a_4^6 \prod_{j=2}^{4} \prod_{i=1}^{j-1}(t_i - t_j)^2 < 0$.

2) Habe $q(t)$ ein Paar nichtreeller Nullstellen und eine reelle Doppelnullstelle, dann besitzt $\mathcal{D}(q)$ Nullfaktoren in der Darstellung mit Nullstellen, so dass $\mathcal{D}(q) = 0$.

3) Habe $q(t)$ zwei Paare nichtreeller Nullstellen, dann gilt $\mathcal{D}(q) \geq 0$. Zum Beweis seien o.B.d.A. $t_1 = \alpha + \mathrm{i}\beta$, $t_2 = \alpha - \mathrm{i}\beta$, $t_3 = \gamma + \mathrm{i}\delta$, $t_4 = \gamma - \mathrm{i}\delta$, $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, $\beta, \delta \neq 0$, damit gilt

$$(t_1 - t_2)^2 = -4\beta^2 < 0, \qquad (t_3 - t_4)^2 = -4\delta^2 < 0,$$

$$(t_1 - t_3)^2(t_2 - t_4)^2 = \big((t_1 - t_3)(t_2 - t_4)\big)^2 = \big((t_1 - t_3)\overline{(t_1 - t_3)}\big)^2 = \big(|t_1 - t_3|^2\big)^2 \geq 0,$$

und analog $(t_1 - t_4)^2(t_2 - t_3)^2 \geq 0$, so dass $\mathcal{D}(q) = a_4^6 \prod_{j=2}^{4} \prod_{i=1}^{j-1}(t_i - t_j)^2 \geq 0$.

4) Habe $q(t)$ vier reelle Nullstellen, dann sind die Faktoren $(t_i - t_j)^2$ sämtlich nicht negativ, also $\mathcal{D}(q) \geq 0$.

Daher ist die Voraussetzung von a) äquivalent zu $\mathcal{D}(q) < 0$. □

Weiter können wir die oben genannte Verschärfung zeigen.

Satz 3.8. *Unter Voraussetzung 3.5 und der Bedingung $\mu(x) \in (\mu_2, \mu_1)$ gilt*

$$\left|\frac{\tan \angle_2(x', y_1)}{\tan \angle_2(x, y_1)}\right| \leq \frac{\kappa(1 + \kappa - \kappa^2)}{2 - \kappa} \quad \textit{mit} \quad \kappa := \frac{\mu_2 - \mu_3}{\mu_1 - \mu_3}. \tag{3.19}$$

Beweis. Wie im Beweis von Satz 3.6 wird das Quadrat der linken Seite durch die Funktion

$$f(\nu_1,\nu_2) := \frac{\nu_2(\kappa-\nu_2)(\nu_1-\nu_2-\nu_1\nu_2+\kappa\nu_2)}{(1-\nu_2)^2(\nu_1+\nu_2-\kappa+\nu_1\nu_2-\kappa\nu_1-\kappa\nu_2+\kappa^2)}$$

dargestellt und es lässt sich zeigen, dass

$$f(\nu_1,\nu_2) \le \left(\frac{\kappa}{2-\kappa}\right)^2 \quad \text{für} \quad (\nu_1,\nu_2) \in (\kappa,1)\times[\kappa/2,\kappa). \tag{3.20}$$

Für $\nu_2 \in (0,\kappa/2)$ wird eine Einschränkung für ν_1 und ν_2 durch die Bedingung $\mu_1 > \mu(x) > \mu_2$ erzeugt

$$\mu_2 < \mu(x) = \frac{\sum_{i=1}^3 \mu_i\xi_i^2}{\sum_{i=1}^3 \xi_i^2} \quad \Leftrightarrow \quad \frac{\xi_3^2}{\xi_1^2} < \frac{\mu_1-\mu_2}{\mu_2-\mu_3} = \frac{1-\kappa}{\kappa}$$

$$\overset{(3.18)}{\Longleftrightarrow} \quad \frac{(1-\nu_1)(1-\nu_2)(1-\kappa)}{\nu_1\nu_2\kappa} < \frac{1-\kappa}{\kappa} \quad \Leftrightarrow \quad \nu_1+\nu_2 > 1. \tag{3.21}$$

Damit gilt $\nu_1 > \max\{\kappa, 1-\nu_2\}$. Außerdem ist $\dfrac{\partial}{\partial\nu_1} f(\nu_1,\nu_2) < 0$ für $\nu_2 \in (0,\kappa/2)$, so dass

1) Falls $\kappa/2 > 1-\kappa$, betrachten wir

   1a) Für $\nu_2 \in (0, 1-\kappa)$ gilt $1-\nu_2 > \kappa$, so dass $\nu_1 > \max\{\kappa, 1-\nu_2\} = 1-\nu_2$ und

   $$f(\nu_1,\nu_2) < f(1-\nu_2,\nu_2) =: g_3(\nu_2). \tag{3.22}$$

   1b) Für $\nu_2 \in [1-\kappa, \kappa/2)$ gilt $1-\nu_2 \le \kappa$, so dass $\nu_1 > \max\{\kappa, 1-\nu_2\} = \kappa$. Mit der Hilfsfunktion g_1 im Beweis von Satz 3.6 ergibt sich

   $$f(\nu_1,\nu_2) < f(\kappa,\nu_2) = g_1(\nu_2) \le g_1(1-\kappa) = f(\kappa,1-\kappa) \overset{(3.22)}{=\!=} g_3(1-\kappa).$$

   Zusammensetzen von 1a) und 1b) liefert $f(\nu_1,\nu_2) < \sup\limits_{0<\nu_2<1-\kappa} g_3(\nu_2)$.

2) Falls $\kappa/2 \le 1-\kappa$, gilt

   $$\nu_2 < \kappa/2 \le 1-\kappa \quad \Rightarrow \quad 1-\nu_2 > \kappa.$$

   So erhalten wir $\nu_1 > \max\{\kappa, 1-\nu_2\} = 1-\nu_2$ und $f(\nu_1,\nu_2) < g_3(\nu_2)$ wie in (3.22). Damit gilt

   $$f(\nu_1,\nu_2) < \sup_{0<\nu_2<\kappa/2} g_3(\nu_2).$$

Zusammensetzen von 1) und 2) liefert die folgende Aussage. Für alle $(\nu_1,\nu_2) \in (\kappa,1)\times(0,\kappa/2)$ mit $\nu_1+\nu_2 > 1$ gilt

$$f(\nu_1,\nu_2) < \sup_{0<\nu_2<\min\{\kappa/2,1-\kappa\}} g_3(\nu_2)$$

mit

$$g_3(\nu_2) := f(1-\nu_2,\nu_2) = \frac{\nu_2(\kappa-\nu_2)(1-3\nu_2+\kappa\nu_2+\nu_2^2)}{(1-\nu_2)^2(1-2\kappa+\kappa^2+\nu_2-\nu_2^2)}. \tag{3.23}$$

Außerdem gilt

$$\sup_{0<\nu_2<\min\{\kappa/2,1-\kappa\}} g_3(\nu_2) = \max_{0\le\nu_2\le\min\{\kappa/2,1-\kappa\}} g_3(\nu_2),$$

denn g_3 ist für $0 \le \nu_2 \le \min\{\kappa/2, 1-\kappa\}$ (dabei $0 < \kappa < 1$) eine rationale Funktion mit dem Nichtnull-Nenner

$$(1-\nu_2)^2(1-2\kappa+\kappa^2+\nu_2-\nu_2^2) = \underbrace{(1-\nu_2)}_{>0}^2\big(\underbrace{(1-\kappa)}_{>0}^2 + \underbrace{\nu_2(1-\nu_2)}_{\ge 0}\big) > 0, \tag{3.24}$$

und daher stetig im Intervall $[\,0,\ \min\{\kappa/2,\,1-\kappa\}\,]$.

Im Gegensatz zu g_1 ist g_3 nicht für alle zugelassenen ν_2 monoton steigend. Statt das lokale Maximum von g_3 zu bestimmen, zeigen wir für alle $\nu_2 \in [\,0,\ \min\{\kappa/2,\,1-\kappa\}\,]$ die Ungleichung

$$g_3(\nu_2) < \left(\frac{\kappa(1+\kappa-\kappa^2)}{2-\kappa}\right)^2, \tag{3.25}$$

die nach (3.23) und (3.24) äquivalent zu

$$\tilde{p}(\nu_2) := \nu_2(\kappa-\nu_2)(1-3\nu_2+\kappa\nu_2+\nu_2^2)(2-\kappa)^2-(1-\nu_2)^2(1-2\kappa+\kappa^2+\nu_2-\nu_2^2)\big(\kappa(1+\kappa-\kappa^2)\big)^2 < 0$$

ist. $\tilde{p}(\nu_2)$ besitzt einen positiven Faktor $(1-\kappa)$: Es gilt $\tilde{p}(\nu_2) = (1-\kappa)\,p(\nu_2)$ mit

$$\begin{aligned} p(\nu_2) = &-(1-\kappa)\kappa^2(1+\kappa-\kappa^2)^2+\kappa(4+\kappa-7\kappa^3-\kappa^4+6\kappa^5-2\kappa^6)\nu_2 \\ &-(4+12\kappa-5\kappa^2-4\kappa^3-6\kappa^4+2\kappa^5+3\kappa^6-\kappa^7)\nu_2^2 \\ &+3(2-\kappa^2)(2+\kappa^2-\kappa^3)\nu_2^3-(2-\kappa^2)(2+\kappa^2-\kappa^3)\nu_2^4. \end{aligned}$$

Damit ist (3.25) äquivalent zu $p(\nu_2) < 0$, was im Folgenden für alle $\nu_2 \in [\,0,\ \min\{\kappa/2,\,1-\kappa\}\,]$ gezeigt wird. Da $p(\nu_2)$ ein Polynom vierten Grades mit reellen Koeffizienten ist, hat $p(\nu_2)$ nach Satz 3.7 zwei einfache reelle Nullstellen und zwei nichtreelle Nullstellen genau dann, wenn die Diskriminante $\mathcal{D}(p)$ negativ ist. Diese Diskriminante lässt sich durch einige Polynome von κ darstellen, nämlich

$$\mathcal{D}(p) = \underbrace{\kappa^3(2-\kappa^2)(2-\kappa)^2(2+\kappa^2-\kappa^3)}_{>0 \text{ für } 0<\kappa<1}\, q(\kappa)$$

mit

$$\begin{aligned} q(\kappa) = &-2560+17472\kappa-65600\kappa^2+110960\kappa^3+9520\kappa^4-295884\kappa^5+334952\kappa^6+84028\kappa^7 \\ &-488800\kappa^8+489340\kappa^9-295988\kappa^{10}+85068\kappa^{11}+224298\kappa^{12}-479273\kappa^{13}+412725\kappa^{14} \\ &-128734\kappa^{15}-66680\kappa^{16}+91787\kappa^{17}-57755\kappa^{18}+34424\kappa^{19}-17792\kappa^{20}+3068\kappa^{21} \\ &+3580\kappa^{22}-3084\kappa^{23}+1124\kappa^{24}-208\kappa^{25}+16\kappa^{26}. \end{aligned}$$

Nach der Untersuchung der Nullstellen des konkreten Polynoms $q(\kappa)$ wird gezeigt, dass die Diskriminante $\mathcal{D}(p)$ für $0 < \kappa < 0.88$ negativ ist. Damit

I) Falls $0 < \kappa < 0.88$, dann besitzt $p(\nu_2)$ genau zwei reelle Nullstellen, und die beiden sind verschieden. Betrachte

$$p(\kappa/2) = -\frac{1}{16}(1+\kappa)(4-6\kappa+3\kappa^2)\kappa^3(2-\kappa)^3 < 0 \quad \text{für } 0<\kappa<0.88,$$

$$p(1-\kappa) = (2+\kappa)(2-5\kappa+\kappa^2-\kappa^3+\kappa^4)(1-\kappa)^3 < 0 \quad \text{für } 0.43<\kappa<0.88.$$

Setze $\nu_2^* := \min\{\kappa/2,\,1-\kappa\}$, dann gilt $p(\nu_2^*) < 0$:

$$\begin{cases} 0<\kappa<0.88, \quad \kappa/2 > 1-\kappa \quad \Rightarrow \quad 0.88 > \kappa > 2/3 > 0.43 \quad \Rightarrow \quad p(\nu_2^*) = p(1-\kappa) < 0, \\ 0<\kappa<0.88, \quad \kappa/2 \le 1-\kappa \quad \Rightarrow \quad p(\nu_2^*) = p(\kappa/2) < 0. \end{cases}$$

Zusammen mit $p(1) = (1-\kappa)(2-\kappa)^2 > 0$ und

$$\lim_{\nu_2\to\infty} p(\nu_2) = \lim_{\nu_2\to\infty} p(\nu_2)/\nu_2^4 \cdot \lim_{\nu_2\to\infty} \nu_2^4 = -(2-\kappa^2)(2+\kappa^2-\kappa^3) \lim_{\nu_2\to\infty} \nu_2^4 < 0$$

wird gezeigt, dass sich die zwei reellen Nullstellen von $p(\nu_2)$ jeweils in $(\nu_2^*, 1)$ und $(1,\infty)$ befinden. Damit hat $p(\nu_2)$ keinen Vorzeichenwechsel im Intervall $[\,0, \nu_2^*]$, also

$$p(\nu_2^*) < 0 \quad \Rightarrow \quad p(\nu_2) < 0 \quad \forall\ \nu_2 \in [\,0, \nu_2^*] = [\,0,\ \min\{\kappa/2,\,1-\kappa\}\,].$$

II) Falls $0.88 \leq \kappa < 1$, betrachten wir die Ableitung von $p(\nu_2)$:

$$\frac{dp}{d\nu_2}(\nu_2) = \kappa(4+\kappa-7\kappa^3-\kappa^4+6\kappa^5-2\kappa^6)-(8+24\kappa-10\kappa^2-8\kappa^3-12\kappa^4+4\kappa^5+6\kappa^6-2\kappa^7)\nu_2$$
$$+9(2-\kappa^2)(2+\kappa^2-\kappa^3)\nu_2^2-4(2-\kappa^2)(2+\kappa^2-\kappa^3)\nu_2^3.$$

Für einige Stellen ν_2 kann man das Vorzeichen von $(dp/d\nu_2)(\nu_2)$ unter der Bedingung $0.88 \leq \kappa < 1$ durch Untersuchung der Nullstellen der entsprechenden Polynome von κ bestimmen:

$$\frac{dp}{d\nu_2}(1-\kappa) = 12-36\kappa+23\kappa^2+4\kappa^3+4\kappa^4-\kappa^7-7\kappa^6+2\kappa^8 > 0,$$

$$\frac{dp}{d\nu_2}(\kappa/2) = -\frac{1}{4}\kappa^2(2-\kappa)(4-4\kappa+4\kappa^2+\kappa^3-5\kappa^4+2\kappa^5) < 0,$$

$$\frac{dp}{d\nu_2}(1) = (3-2\kappa)(2-\kappa)^2 > 0,$$

$$\lim_{\nu_2\to\infty}\frac{dp}{d\nu_2}(\nu_2) = \lim_{\nu_2\to\infty} p(\nu_2)/\nu_2^3 \cdot \lim_{\nu_2\to\infty}\nu_2^3 = -4(2-\kappa^2)(2+\kappa^2-\kappa^3)\lim_{\nu_2\to\infty}\nu_2^3 < 0.$$

Aus $0.88 \leq \kappa < 1$ folgt außerdem $1-\kappa < \kappa/2 < 1$, so dass $(dp/d\nu_2)(\mu_2)$ drei reelle Nullstellen besitzt, die sich jeweils in $(1-\kappa, \kappa/2)$, $(\kappa/2, 1)$ und $(1, \infty)$ befinden. Damit hat $(dp/d\nu_2)(\nu_2)$ keinen Vorzeichenwechsel im Intervall $[\,0, 1-\kappa\,]$, also

$$\frac{dp}{d\nu_2}(1-\kappa) > 0 \quad \Rightarrow \quad \frac{dp}{d\nu_2}(\nu_2) > 0 \quad \forall\ \nu_2 \in [\,0, 1-\kappa\,],$$

das heißt, $p(\nu_2)$ ist monoton steigend in diesem zugelassenen Intervall. Wegen

$$p(1-\kappa) = (2+\kappa)(2-5\kappa+\kappa^2-\kappa^3+\kappa^4)(1-\kappa)^3 < 0 \quad \text{für} \quad 0.88 \leq \kappa < 1$$

gilt schließlich

$$p(\nu_2) < 0 \quad \forall\ \nu_2 \in [\,0, 1-\kappa\,] = [\,0,\ \min\{\kappa/2,\ 1-\kappa\}\,].$$

Damit gilt $p(\nu_2) < 0 \quad \forall\ \nu_2 = [\,0,\ \min\{\kappa/2,\ 1-\kappa\}\,]$ unter der Bedingung $0 < \kappa < 1$, die nach der Definition von κ bereits erfüllt ist. Daraus folgt

$$f(\nu_1,\nu_2) < \max_{0\leq\nu_2\leq\min\{\kappa/2,1-\kappa\}} g_3(\nu_2) < \left(\frac{\kappa(1+\kappa-\kappa^2)}{2-\kappa}\right)^2. \tag{3.26}$$

für alle $(\nu_1, \nu_2) \in (\kappa, 1) \times (0, \kappa/2)$ mit $\nu_1 + \nu_2 > 1$.

Zusammensetzen von (3.20) und (3.26) zeigt

$$f(\nu_1,\nu_2) < \max\left\{\left(\frac{\kappa}{2-\kappa}\right)^2,\ \left(\frac{\kappa(1+\kappa-\kappa^2)}{2-\kappa}\right)^2\right\} = \left(\frac{\kappa(1+\kappa-\kappa^2)}{2-\kappa}\right)^2$$

für alle $(\nu_1, \nu_2) \in (\kappa, 1) \times (0, \kappa)$ mit $\nu_1 + \nu_2 > 1$ und weiter die Behauptung. □

3.2.3 Neue Beweistechnik mit geometrischen Argumenten

In der gemeinsamen Arbeit des Autors mit Neymeyr und Ovtchinnikov [68] wurden einige klassische Argumente aus [51] unter geometrischen Aspekten viel intuitiver beschrieben. Das ermöglicht eine Vereinfachung des Beweises der klassischen Resultate in Abschnitt 3.2.1. Der Ausgangspunkt der neuen Beweistechnik ist die Betrachtung der geometrischen Bedeutungen der Konvergenzmaße, vergleiche Lemma 2.7 zu einer ähnlichen Problemstellung.

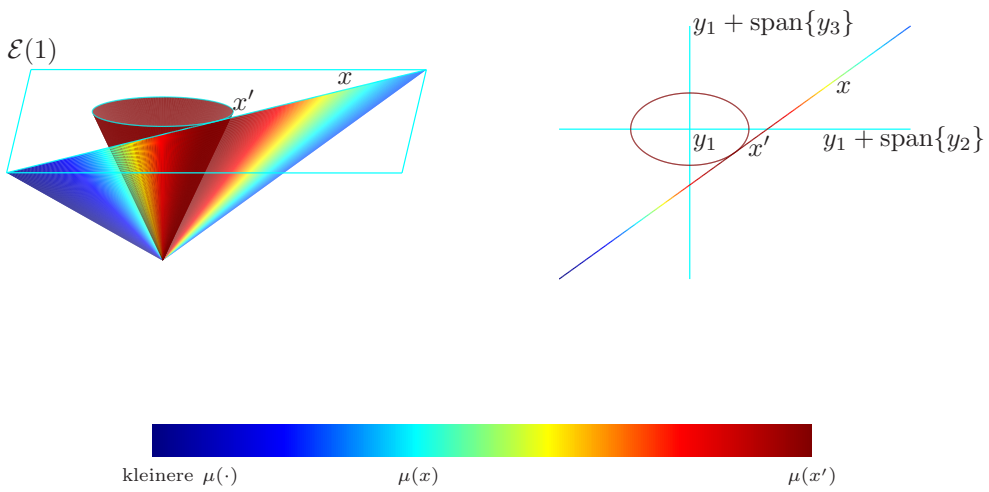


Abbildung 3.1: Darstellung für INVIT(2) in einem affinen Raum

Lemma 3.9. *Unter Voraussetzung 3.5 sei ein affiner Raum* $\mathcal{E}(\alpha_1) := \alpha_1 y_1 + \mathrm{span}\{y_2, y_3\}$ *für* $\alpha_1 \in \mathbb{R}\backslash\{0\}$ *definiert. Für ein* $\tilde{\mu} \in (\mu_2, \mu_1)$ *ist dann der Durchschnitt der Rayleigh-Quotient-Niveaumenge* $\mathcal{N}(\tilde{\mu}) := \{w \in \mathbb{R}^3\backslash\{\mathbf{0}\} \ ; \ \mu(w) = \tilde{\mu}\}$ *mit* $\mathcal{E}(\alpha_1)$ *eine Ellipse, deren Halbachsen durch* $|\alpha_1|\sqrt{\Delta\big(\mu_1, \mu_2, \tilde{\mu}\big)}$ *und* $|\alpha_1|\sqrt{\Delta\big(\mu_1, \mu_3, \tilde{\mu}\big)}$ *gegeben sind.*

Beweis. Ein beliebes $w \in \mathcal{E}(\alpha_1)$ lässt sich durch $w = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3$ darstellen. Wegen

$$\mu(w) = \frac{\mu_1\alpha_1^2 + \mu_2\alpha_2^2 + \mu_3\alpha_3^2}{\alpha_1^2 + \alpha_2^2 + \alpha_3^2} = \tilde{\mu} \quad \Leftrightarrow \quad \frac{\alpha_2^2}{\alpha_1^2\Delta\big(\mu_1, \mu_2, \tilde{\mu}\big)} + \frac{\alpha_3^2}{\alpha_1^2\Delta\big(\mu_1, \mu_3, \tilde{\mu}\big)} = 1$$

und $\Delta\big(\mu_1, \mu_2, \tilde{\mu}\big) > \Delta\big(\mu_1, \mu_3, \tilde{\mu}\big) > 0$ ist der Durchschnitt $\mathcal{E}(\alpha_1) \cap \mathcal{N}(\tilde{\mu})$ die behauptete Ellipse.
(Man kann auch zeigen, dass dieser Durchschnitt für $\tilde{\mu} = \mu_2$ zwei parallele Geraden und für $\tilde{\mu} \in (\mu_3, \mu_2)$ eine Hyperbel wird.) □

Die Niveaumenge $\mathcal{N}(\tilde{\mu})$ besteht also aus solchen Ellipsen und ist ein elliptischer Doppelkegel ohne den Nullpunkt $\mathbf{0}$. Betrachten wir nun den Kegel $\mathcal{N}\big(\mu(x')\big)$ mit einem beliebigen $\alpha_1 \in \mathbb{R}\backslash\{0\}$. Der Krylovraum $\mathcal{K}_H^2(x)$ ist eine Tangentialebene von $\mathcal{N}\big(\mu(x')\big)$, und die Schnittgerade $\mathcal{E}(\alpha_1) \cap \mathcal{K}_H^2(x)$ ist dementsprechend eine Tangente der Ellipse $\mathcal{E}(\alpha_1) \cap \mathcal{N}\big(\mu(x')\big)$. Der Berührungspunkt lässt sich durch eine Umskalierung von x' erhalten. Da die Abschätzungen (3.13) und (3.14) bei Umskalierungen von x und x' unverändert bleiben, können wir im Folgenden

$$x = y_1 + \beta_2 y_2 + \beta_3 y_3, \qquad x' = y_1 + \beta_2' y_2 + \beta_3' y_3$$

setzen und das Konvergenzverhalten weiter im affinen Raum $\mathcal{E}(1)$ untersuchen.
Zum Beweis von (3.13) gilt zunächst

$$\tan^2 \angle_2(x', y_1) = {\beta_2'}^2 + {\beta_3'}^2, \qquad \frac{{\beta_2'}^2}{\Delta\big(\mu_1, \mu_2, \mu(x')\big)} + \frac{{\beta_3'}^2}{\Delta\big(\mu_1, \mu_3, \mu(x')\big)} = 1,$$

so dass

$$\tan^2 \angle_2(x', y_1) \le {\beta_2'}^2 + \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\Delta\big(\mu_1, \mu_3, \mu(x')\big)}{\beta_3'}^2 = \Delta\big(\mu_1, \mu_2, \mu(x')\big). \tag{3.27}$$

In anderen Worten ist $|\tan \angle_2(x', y_1)|$ der Euklidischer Abstand von x' zum Mittelpunkt y_1 der Ellipse $\mathcal{E}(1) \cap \mathcal{N}\big(\mu(x')\big)$ und damit nicht größer als deren große Halbachse.

Eine obere Schranke der großen Halbachse kann durch den Euklidischen Abstand des Mittelpunkts y_1 zu einem Punkt, welcher auf der der großen Halbachse entsprechenden Geraden $y_1 + \operatorname{span}\{y_2\}$ und außerhalb der Ellipse liegt, gegeben werden. Der Schnittpunkt w der Ellipsentangente $\mathcal{E}(1) \cap \mathcal{K}_H^2(x)$ mit $y_1 + \operatorname{span}\{y_2\}$ ist genau ein solcher Punkt. Wegen $w \in \mathcal{K}_H^2(x)$ lässt sich w zunächst durch

$$w = \vartheta x + \zeta H x = (\vartheta + \zeta\mu_1)y_1 + (\vartheta + \zeta\mu_2)\beta_2 y_2 + (\vartheta + \zeta\mu_3)\beta_3 y_3$$

darstellen. Wegen $w \in \big(y_1 + \operatorname{span}\{y_2\}\big)$ soll dann

$$\vartheta + \zeta\mu_1 = 1, \qquad \vartheta + \zeta\mu_3 = 0$$

gelten. Damit erhalten wir $\zeta = 1/(\mu_1 - \mu_3)$, $\theta = -\mu_3/(\mu_1 - \mu_3)$ und weiter

$$w = y_1 + \frac{\mu_2 - \mu_3}{\mu_1 - \mu_3}\beta_2 y_2 = y_1 + \kappa\beta_2 y_2.$$

Durch den quadratischen Euklidischen Abstand $\|w - y_1\|_2^2 = \kappa^2\beta_2^2$ wird also eine obere Schranke von $\Delta\big(\mu_1, \mu_2, \mu(x')\big)$ gegeben. Zusammen mit (3.27) liefert das die Abschätzung

$$\tan^2 \angle_2(x', y_1) \le \kappa^2\beta_2^2 \le \kappa^2(\beta_2^2 + \beta_3^2) = \kappa^2 \tan^2 \angle_2(x, y_1)$$

und weiter (3.13). Als eine Folgerung erhalten wir auch eine Tangens-Abschätzung bezüglich des H-Skalarproduktes. Mit

$$\begin{aligned}\tan^2 \angle_H(x', y_1) &= \frac{1}{\mu_1}\big(\mu_2{\beta'_2}^2 + \mu_3{\beta'_3}^2\big) = \frac{\mu_2}{\mu_1}\left({\beta'_2}^2 + \frac{\mu_3}{\mu_2}{\beta'_3}^2\right) \le \frac{\mu_2}{\mu_1}\big({\beta'_2}^2 + {\beta'_3}^2\big) = \frac{\mu_2}{\mu_1}\tan^2 \angle_2(x', y_1) \\ &\le \frac{\mu_2}{\mu_1}\kappa^2\beta_2^2 \le \kappa^2\frac{\mu_2}{\mu_1}\left(\beta_2^2 + \frac{\mu_3}{\mu_2}\beta_3^2\right) = \kappa^2 \tan^2 \angle_H(x, y_1)\end{aligned}$$

gilt also

$$\left|\frac{\tan \angle_H(x', y_1)}{\tan \angle_H(x, y_1)}\right| \le \kappa. \tag{3.28}$$

Bei den Abschätzungen (3.13) und (3.28) ist die Schranke κ der Grenzwert des betragsmäßigen Tangens-Quotienten für $x \to y_2$. Dabei konvergieren β_2^2 und ${\beta'}_2^2$ beide gegen ∞.

Zur Bestimmung der oberen Schranke in (3.14) stellen wir die Vektoren $w = y_1 + \alpha_2 y_2 + \alpha_3 y_3$ aus $\mathcal{E}(1)$ durch die entsprechenden zweidimensionalen Koeffizientenvektoren $Pw = (\alpha_2, \alpha_3)^T$ dar. Die Richtung der Tangente $\mathcal{E}(1) \cap \mathcal{K}_H^2(x)$ wird dann durch $Px - Px'$ dargestellt. Betrachten wir weiter die Ellipse $\mathcal{E}(1) \cap \mathcal{N}\big(\mu(x')\big)$ als die Einheitskugel der Vektornorm $\|\cdot\|_D$ mit

$$D = \operatorname{diag}(\delta_2, \delta_3), \qquad \delta_2 := 1/\Delta\big(\mu_1, \mu_2, \mu(x')\big), \quad \delta_3 := 1/\Delta\big(\mu_1, \mu_3, \mu(x')\big).$$

Da x' der Berührungspunkt der Tangente ist, gilt eine elliptische Orthogonalität

$$Px - Px' \perp_D Px'.$$

Damit gilt bezüglich eines D-Winkels, dass

$$\begin{aligned}\sin^2 \angle_D(Px - Px', Px) &= \frac{\|Px'\|_D^2}{\|Px\|_D^2} = \frac{1}{\delta_2\beta_2^2 + \delta_3\beta_3^2} = \frac{\delta_2^{-1}}{\beta_2^2 + \frac{\delta_2^{-1}}{\delta_3^{-1}}\beta_3^2} = \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\beta_2^2 + \frac{\mu(x') - \mu_3}{\mu(x') - \mu_2}\beta_3^2} \\ &\ge \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\beta_2^2 + \frac{\mu(x) - \mu_3}{\mu(x) - \mu_2}\beta_3^2} = \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\beta_2^2 + \frac{\Delta\big(\mu_1, \mu_2, \mu(x)\big)}{\Delta\big(\mu_1, \mu_3, \mu(x)\big)}\beta_3^2} = \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\Delta\big(\mu_1, \mu_2, \mu(x)\big)}.\end{aligned} \tag{3.29}$$

Dabei ist zu bemerken, dass x zu einer Ellipse mit Halbachsen $\Delta\big(\mu_1,\mu_2,\mu(x)\big)$ und $\Delta\big(\mu_1,\mu_3,\mu(x)\big)$ gehört. Weiter lässt sich das Sinus-Quadrat nach oben abschätzen: Mit dem oben betrachteten Schnittpunkt w ist $Px-Pw$ kollinear zu $Px-Px'$, so dass

$$\cos^2\angle_D(Px-Px',Px)=\cos^2\angle_D(Px-Pw,Px)=\frac{(Px-Pw,Px)_D^2}{\|Px-Pw\|_D^2\|Px\|_D^2}.$$

Einsetzen von $Px=(\beta_2,\beta_3)^T$, $Pw=(\kappa\beta_2,0)^T$ liefert

$$\cos^2\angle_D(Px-Px',Px)=\frac{\big(\delta_2(1-\kappa)\beta_2^2+\delta_3\beta_3^2\big)^2}{\big(\delta_2(1-\kappa)^2\beta_2^2+\delta_3\beta_3^2\big)\big(\delta_2\beta_2^2+\delta_3\beta_3^2\big)}.$$

Mit $\vartheta:=\beta_2^2/\beta_3^2$ lässt sich $\cos^2\angle_D(Px-Px',Px)$ weiter durch

$$\frac{\big(\delta_2(1-\kappa)\vartheta+\delta_3\big)^2}{\big(\delta_2(1-\kappa)^2\vartheta+\delta_3\big)\big(\delta_2\vartheta+\delta_3\big)}$$

darstellen und als eine Funktion $f(\vartheta)$ für $\vartheta\in[0,\infty)$ betrachten. Wegen

$$\frac{df}{d\vartheta}(\vartheta)=\underbrace{\frac{\big(\delta_2(1-\kappa)\vartheta+\delta_3\big)\delta_2\delta_3\kappa^2}{\big(\delta_2(1-\kappa)^2\vartheta+\delta_3\big)^2\big(\delta_2\vartheta+\delta_3\big)^2}}_{>0}\big(\delta_2(1-\kappa)\vartheta-\delta_3\big)$$

ist $\vartheta=(\delta_3/\delta_2)/(1-\kappa)$ eine Extremstelle von $f(\vartheta)$ auf $[0,\infty)$. Nach Betrachtung der zweiten Ableitung ist diese eine Minimalstelle. Das Minimum lautet dann $4(1-\kappa)/(2-\kappa)^2$. Damit gilt

$$\sin^2\angle_D(Px-Px',Px)=1-\cos^2\angle_D(Px-Px',Px)\le 1-\frac{4(1-\kappa)}{(2-\kappa)^2}=\left(\frac{\kappa}{2-\kappa}\right)^2$$

und weiter mit (3.29) die Abschätzung (3.14). Die Schranke lässt sich im Grenzfall $\mu(x)\to\mu_1$ erreichen. Dabei gilt außerdem $\mu(x')\to\mu_1$.

Der Beweis der verschärften Abschätzung in Satz 3.8 lässt sich durch diese geometrischen Argumente umformulieren, jedoch nicht wesentlich vereinfachen. Dabei können wir zunächst die Koordinaten von Px durch die Koordinaten von Px' darstellen. Mit der D-Orthogonalität $Px-Pw\perp_D Px'$ gilt

$$\delta_2(\beta_2-\kappa\beta_2)\beta_2'+\delta_3(\beta_3-0)\beta_3'=0.$$

Mit der Kollinearität von $Px-Pw$ und $Px-Px'$ gilt

$$\frac{\beta_3-0}{\beta_2-\kappa\beta_2}=\frac{\beta_3-\beta_3'}{\beta_2-\beta_2'}.$$

Auflösen nach β_2 und β_3 liefert

$$\beta_2=\frac{\delta_2\beta'^2_2+\delta_3\beta'^2_3}{\delta_2\kappa\beta_2'},\qquad \beta_3=\frac{(\kappa-1)(\delta_2\beta'^2_2+\delta_3\beta'^2_3)}{\delta_3\kappa\beta_3'}. \tag{3.30}$$

Weiter können wir mit einem Einheitsvektor $e_2:=(1,0)^T$ die $\mathcal{E}(1)$-Darstellung $\operatorname{span}\{e_2\}$ der Geraden $y_1+\operatorname{span}\{y_2\}$ und den D-Winkel $\angle_D\big(Px',\operatorname{span}\{e_2\}\big)$ betrachten und das Quadrat von dessen Tangenswert als einen Parameter t benutzen. Es gilt mit $Px'=(\beta_2',\beta_3')^T$, dass $(\beta_2',0)^T$ und $(0,\beta_3')^T$ jeweils die D-Projektionen von Px' auf $\operatorname{span}\{e_2\}$ und dessen D-orthogonalen Komplement $\operatorname{span}\{e_3\}$ mit $e_3:=(0,1)^T$ sind. Das liefert

$$t:=\tan^2\angle_D\big(Px',\operatorname{span}\{e_2\}\big)=\frac{\|(0,\beta_3')^T\|_D^2}{\|(\beta_2',0)^T\|_D^2}=\frac{\delta_3\beta'^2_3}{\delta_2\beta'^2_2}.$$

Zusammen mit (3.30) und den Parametern $\nu = \delta_2/\delta_3,\ \xi = 1-\kappa$ gilt dann

$$\frac{\tan^2 \angle(x', e_j)}{\tan^2 \angle(x, e_j)} = \frac{\beta'^2_2 + \beta'^2_3}{\beta_2^2 + \beta_3^2} = \frac{(1-\xi)^2 t(1+\nu t)}{(1+t)^2(\xi^2\nu + t)}.$$

Das können wir als eine Funktion $f(\nu, t)$ für $\nu \in (0,1)$ und $t \in (0,\infty)$ betrachten. Die partielle Ableitung

$$\frac{\partial f}{\partial \nu}(\nu, t) = \frac{(1-\xi)^2 t(t+\xi)(t-\xi)}{(1+t)^2(\xi^2\nu + t)^2}$$

hat einen Vorzeichenwechsel an $t = \xi$, damit formulieren wir eine Fallunterscheidung:
1) Für $t \geq \xi$ gilt $(\partial f/\partial \nu)(\nu, t) \geq 0$, so dass $f(\nu, t)$ eine obere Schranke $f(\xi, t)$ hat, welche durch

$$f(\xi, \xi) = \frac{(1-\xi)^2}{(1+\xi)^2} = \left(\frac{\kappa}{2-\kappa}\right)^2$$

von oben beschränkt wird.
2) Für $0 < t < \xi$ folgt aus der Bedingung $\mu(x) > \mu_2$ eine Einschränkung der Parameter $\nu,\ t$.

$$\mu(x) = \frac{\mu_1 + \mu_2\beta_2^2 + \mu_3\beta_3^2}{1 + \beta_2^2 + \beta_3^2} > \mu_2 \quad \Rightarrow \quad \beta_3^2 < \frac{\mu_1 - \mu_2}{\mu_2 - \mu_3} = \frac{1-\kappa}{\kappa}.$$

Mit (3.30) und den Definitionen von $\nu,\ \xi, t$ lässt sich diese Ungleichung als

$$\frac{\xi - \nu}{1-\xi} < \frac{t}{1+t}$$

umschreiben. Damit formulieren wir die Einschränkung

$$\Rightarrow \quad \nu > \bar{\nu}(t) := \frac{\xi^2 + 2\xi t - t}{\xi + \xi t}. \tag{3.31}$$

Wegen $t < \xi$ gilt $(\partial f/\partial \nu)(\nu, t) < 0$. Mit (3.31) erhalten wir

$$f(\nu, t) \leq \sup_{0<t<\min\{\xi,\tilde{\xi}\}} g(t)$$

mit

$$g(t) := f(\bar{\nu}(t),\ t) = \frac{(\xi + \xi t + \xi^2 t - t^2 + 2\xi t^2)(1-\xi)^2 t}{\xi(\xi^3 - \xi t + 2\xi^2 t + t + t^2)(1+t)^2},$$

$$\tilde{\xi} := \begin{cases} \xi^2/(1-2\xi) & \text{für } \xi \in (0, 0.5), \\ \infty & \text{für } \xi \in [0.5, 1). \end{cases}$$

denn das Maximum von $f(\nu, t)$ wird in den Mengen

$$S_1 := \{(\nu, t)\ ;\ \nu = \bar{\nu}(t),\ 0 < \nu < \xi,\ 0 < t < \xi\},$$
$$S_2 := \{(0, t)\ ;\ \tilde{\xi} \leq t < \xi\}.$$

angenommen. Für $\tilde{\xi} \geq \xi$ gilt $S_2 = \emptyset$. Sonst wird das lokale Maximum von $f(\nu, t)$ innerhalb S_2 an $\left(0, \tilde{\xi}\right)$, das heißt $S_1 \cap S_2$, angenommen, denn $f(0, t) = (1-\xi)^2/(1+t)^2$ ist für $0 < t < \xi$ eine monoton fallende Funktion. Damit können wir weiter nur die Menge S_1 betrachten.

Für die resultierende Beweisaufgabe kann man wieder Diskriminante eines Polynoms vierten Grades verwenden. Ob es einen wesentlich einfacheren Beweis gibt, ist zur Zeit eine offene Frage.

Zur numerischen Auswertung der exakten Schranke lässt sich die Bedingung $t < \min\{\xi, \tilde{\xi}\}$ noch vereinfachen:

$$0 < t < \begin{cases} \xi^2/(1-2\xi) & \text{für } \xi \in (0, \frac{1}{3}), \\ \xi & \text{für } \xi \in [\frac{1}{3}, 1). \end{cases}$$

3.2.4 Rückwärtseinsetzen der Zwischenresultate

Basierend auf der obigen niedrigdimensionalen Analyse erhalten wir gewünschte Abschätzungen für originale Verfahrensformulierungen.

Satz 3.10. *Mit den Formulierungen in Lemma 3.2, Lemma 3.3 und Lemma 3.4 sei für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und das Verfahren (3.2) der nichttriviale Fall $s = 2 < \bar{s}$ betrachtet. Es gilt*

$$\frac{\tan\angle_A\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_A\big(u,\,\mathcal{V}_{\sigma(1)}\big)} \le \hat{\kappa}, \qquad \frac{\tan\angle_M\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_M\big(u,\,\mathcal{V}_{\sigma(1)}\big)} \le \hat{\kappa} \tag{3.32}$$

$$\textit{mit} \qquad \hat{\kappa} := \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)\left(\frac{\lambda_{\sigma(\bar{s})}-\lambda_{\sigma(2)}}{\lambda_{\sigma(\bar{s})}-\lambda_{\sigma(1)}}\right).$$

Gelte zusätzlich $\rho(u) \in \big(\lambda_{\sigma(1)},\,\lambda_{\sigma(2)}\big)$, dann gilt

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u)\big)} \le \left(\frac{\hat{\kappa}}{2-\hat{\kappa}}\right)^2, \tag{3.33}$$

$$\frac{\tan\angle_A\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_A\big(u,\,\mathcal{V}_{\sigma(1)}\big)} \le \frac{\hat{\kappa}(1+\hat{\kappa}-\hat{\kappa}^2)}{2-\hat{\kappa}}, \tag{3.34}$$

$$\frac{\tan\angle_M\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_M\big(u,\,\mathcal{V}_{\sigma(1)}\big)} \le \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)\left(\frac{\tilde{\kappa}(1+\tilde{\kappa}-\tilde{\kappa}^2)}{2-\tilde{\kappa}}\right) \quad \textit{mit} \quad \tilde{\kappa} := \frac{\lambda_{\sigma(\bar{s})}-\lambda_{\sigma(2)}}{\lambda_{\sigma(\bar{s})}-\lambda_{\sigma(1)}}. \tag{3.35}$$

Die oberen Schranken in (3.32) und (3.33) lassen sich in Grenzfällen erreichen. (Damit sind diese Abschätzungen scharf.)

Beweis. Mit den Formulierungen in Lemma 3.4 gilt nach dem Courant-Fischer-Prinzip, dass

$$\mu_1 = \lambda_{\sigma(1)}^{-1} > \lambda_{\sigma(2)}^{-1} \ge \mu_2 > \mu_3 \ge \lambda_{\sigma(\bar{s})}^{-1}.$$

Damit wird gezeigt, dass

$$\kappa = \frac{\mu_2-\mu_3}{\mu_1-\mu_3} \le \frac{\lambda_{\sigma(2)}^{-1}-\lambda_{\sigma(\bar{s})}^{-1}}{\lambda_{\sigma(1)}^{-1}-\lambda_{\sigma(\bar{s})}^{-1}} = \hat{\kappa}.$$

Tangensabschätzungen in (3.32) folgen dann jeweils aus (3.9), (3.13) und (3.8), (3.28):

$$\frac{\tan\angle_A\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_A\big(u,\,\mathcal{V}_{\sigma(1)}\big)} = \left|\frac{\tan\angle_2(x',y_1)}{\tan\angle_2(x,y_1)}\right| \le \kappa \le \hat{\kappa},$$

analog $\tan\angle_M\big(u',\,\mathcal{V}_{\sigma(1)}\big)/\tan\angle_M\big(u,\,\mathcal{V}_{\sigma(1)}\big) \le \hat{\kappa}$.
Gelte zusätzlich $\rho(u) \in \big(\lambda_{\sigma(1)},\,\lambda_{\sigma(2)}\big)$, entsteht zunächst die Bedingung $\mu(x) \in (\mu_2, \mu_1)$. Weiter folgt die Abschätzung (3.33) aus (3.10), (3.14) und einer Monotonie

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u)\big)} \le \frac{\Delta\big(\mu_1,\mu_2,\mu(x')\big)}{\Delta\big(\mu_1,\mu_2,\mu(x)\big)} \le \left(\frac{\kappa}{2-\kappa}\right)^2 \le \left(\frac{\hat{\kappa}}{2-\hat{\kappa}}\right)^2.$$

Die Abschätzung (3.34) folgt in ähnlicher Weise aus (3.9), (3.19) und einer Monotonie

$$\frac{\tan\angle_A\big(u',\,\mathcal{V}_{\sigma(1)}\big)}{\tan\angle_A\big(u,\,\mathcal{V}_{\sigma(1)}\big)} = \left|\frac{\tan\angle_2(x',y_1)}{\tan\angle_2(x,y_1)}\right| \le \frac{\kappa(1+\kappa-\kappa^2)}{2-\kappa} \le \frac{\hat{\kappa}(1+\hat{\kappa}-\hat{\kappa}^2)}{2-\hat{\kappa}}.$$

Da (3.19) bezüglich der H-Winkel im Allgemeinen nicht gilt, benutzen wir zum Beweis von (3.35) eine andere Krylovraum-Umformulierung. Mit den Formulierungen in Lemma 3.3 definieren wir $\bar{H} := \bar{W}^T A \bar{W}$ mit einer beliebigen M-orthonormalen Basismatrix $\bar{W}$ von $\mathcal{W}$, dann ist $\mathcal{K}^s_{A^{-1}M}(u)$ durch $\bar{W}\mathcal{K}^s_{\bar{H}^{-1}}(\bar{x})$ mit $\bar{x} := \bar{W}^T M u$ darstellbar. Mit dem Rayleigh-Quotienten $\bar{\mu}(\cdot)$ bezüglich $\bar{H}$ gilt für beliebiges $\tilde{u} \in \mathcal{W}\backslash\{\mathbf{0}\}$ und den entsprechenden Koeffizientenvektor $\tilde{x} := \bar{W}^T M \tilde{u}$, dass $\rho(\tilde{u}) = \bar{\mu}(\tilde{x})$, und $\bar{x}' := \bar{W}^T M u'$ ist genau dann ein Ritzvektor zum kleinsten Ritzwert von $\bar{H}$ bezüglich $\mathcal{K}^s_{\bar{H}^{-1}}(\bar{x})$, wenn u' ein Ritzvektor zum kleinsten Ritzwert von (A, M) bezüglich $\mathcal{K}^s_{A^{-1}M}(u)$ ist. Seien $\bar{\mu}_1 \le \bar{\mu}_2 \le \cdots \le \bar{\mu}_{s+1}$ die Eigenwerte von $\bar{H}$ und $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{s+1}$ zugehörige Euklidisch normierte Eigenvektoren. Dann gilt analog zu Lemma 3.4 eine strikte Einschachtelung der Ritzwerte beziehungsweise die Umrechnung

$$\frac{\tan\angle_M\big(u',\, \mathcal{V}_{\sigma(1)}\big)}{\tan\angle_M\big(u,\, \mathcal{V}_{\sigma(1)}\big)} = \left|\frac{\tan\angle_2(\bar{x}', \bar{y}_1)}{\tan\angle_2(\bar{x}, \bar{y}_1)}\right|.$$

Weiter betrachten wir den Fall $s = 2$, $\rho(u) \in \big(\lambda_{\sigma(1)},\, \lambda_{\sigma(2)}\big)$, und dabei den Krylovraum

$$\mathcal{K}^2_{\bar{H}^{-1}}(\bar{x}) = \operatorname{span}\{\bar{x}, \bar{H}^{-1}\bar{x}\} = \operatorname{span}\{w, \bar{H}w\} = \mathcal{K}^2_{\bar{H}}(w)$$

mit $w := \bar{H}^{-1}\bar{x}$. Wegen $\bar{\mu}(\bar{x}) = \rho(u) \in \big(\lambda_{\sigma(1)},\, \lambda_{\sigma(2)}\big)$, $\bar{\mu}_1 = \lambda_{\sigma(1)} < \lambda_{\sigma(2)} \le \bar{\mu}_2 < \bar{\mu}_3 \le \lambda_{\sigma(\bar{s})}$ und der Rayleigh-Quotient-Verkleinerung der inversen Vektoriteration (vergleiche Satz 2.14) gilt

$$\bar{\mu}_1 < \bar{\mu}(w) \le \bar{\mu}(\bar{x}) < \bar{\mu}_2.$$

Dann wird analog zu Satz 3.8 gezeigt, dass

$$\left|\frac{\tan\angle_2(\bar{x}', \bar{y}_1)}{\tan\angle_2(w, \bar{y}_1)}\right| \le \frac{\bar{\kappa}(1 + \bar{\kappa} - \bar{\kappa}^2)}{2 - \bar{\kappa}} \quad \text{mit} \quad \bar{\kappa} := \frac{\bar{\mu}_3 - \bar{\mu}_2}{\bar{\mu}_3 - \bar{\mu}_1}.$$

Nach der Tangensabschätzung der inversen Vektoriteration (vergleiche Satz 2.11) gilt außerdem

$$\left|\frac{\tan\angle_2(w, \bar{y}_1)}{\tan\angle_2(\bar{x}, \bar{y}_1)}\right| \le \frac{\bar{\mu}_1}{\bar{\mu}_2}.$$

Zusammensetzen dieser Zwischenabschätzungen liefert (3.35).

Die oberen Schranken in (3.32) und (3.33) lassen sich mit denjenigen u, die auf genau drei Eigenräumen Nichtnull-Anteile haben, also mit $\bar{s} = 3$, in ähnlichen Grenzfällen wie bei der niedrigdimensionalen Analyse erreichen. □

3.2.5 Konvergenzabschätzung für Ritzwerte unter allgemeineren Bedingungen

Ausgehend von den Bedingungen der Form $\rho(u) \in (\lambda_i, \lambda_{i+1})$ können wir eine gleichartige Konvergenzabschätzung wie (2.8) als Folgerung von (3.33) formulieren. Dafür betrachten wir wieder Rayleigh-Quotient-Niveaumengen und verwenden eine Kurven-Differentiation, vergleiche Theorem 2.1 in [68] zu einer ähnlichen Problemstellung.

Lemma 3.11. *Betrachten wir das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und eine Rayleigh-Quotient-Niveaumenge $\mathcal{N}(\tilde{\lambda}) := \{x \in \mathbb{R}^n\backslash\{\mathbf{0}\} \;;\; \rho(x) = \tilde{\lambda}\}$, wobei $\tilde{\lambda} \in (\lambda_1, \lambda_m)$ und $\tilde{\lambda}$ kein Eigenwert von (A, M) ist. Für jedes $x \in \mathcal{N}(\tilde{\lambda})$ ist dann $\mathcal{K}^2_{A^{-1}M}(x)$ zweidimensional und es gibt ein eindeutiges $\alpha(x) \in \mathbb{R}$, so dass $\alpha(x)\, x + A^{-1}Mx$ ein Ritzvektor zum kleinsten Ritzwert von (A, M) bezüglich $\mathcal{K}^2_{A^{-1}M}(x)$ ist. Damit ist auf $\mathcal{N}(\tilde{\lambda})$ eine Funktion*

$$f(x) := \rho\big(\alpha(x)\, x + A^{-1}Mx\big)$$

definierbar. Falls u eine Extremstelle von f ist, dann gibt es höchstens drei verschiedene Eigenwerte, auf deren Eigenräumen u Nichtnull-Anteile besitzt.

Beweis. Für jedes $x \in \mathcal{N}(\tilde{\lambda})$ ist $\rho(x) = \tilde{\lambda}$ kein Eigenwert von (A, M), so dass x kein Eigenvektor ist und $\mathcal{K}^2_{A^{-1}M}(x)$ zweidimensional ist. Nach Lemma 2.9 wird gezeigt, dass $\mathcal{K}^2_{A^{-1}M}(x)$ einen Vektor mit kleinerem Rayleigh-Quotienten als x besitzt, so dass x kein Ritzvektor zum kleinsten Ritzwert von (A, M) bezüglich $\mathcal{K}^2_{A^{-1}M}(x)$ ist und $\mathcal{K}^2_{A^{-1}M}(x)$ zwei eindimensionale Ritzräume besitzt. Mit diesen Ritzräumen hat die Gerade $\tilde{\alpha}\, x + A^{-1}Mx$ jeweils einen eindeutigen Schnittpunkt. Damit gibt es ein eindeutiges $\alpha(x) \in \mathbb{R}$, so dass $\alpha(x)\, x + A^{-1}Mx$ ein Ritzvektor zum kleinsten Ritzwert von (A, M) bezüglich $\mathcal{K}^2_{A^{-1}M}(x)$ ist.

Für eine Extremstelle u von f auf $\mathcal{N}(\tilde{\lambda})$ betrachten wir eine beliebige durch u laufende stetig differenzierbare Kurve $\{g(t)\ ;\ t \in T \subseteq \mathbb{R}\} \subset \mathcal{N}(\tilde{\lambda})$ mit $g(t_g) = u$, dann ist u auch eine Extremstelle von f auf dieser Kurve, so dass t_g eine Extremstelle der Funktion $f \circ g$ ist und die notwendige Bedingung $\big(d\,(f \circ g)/dt\big)(t_g) = 0$ gilt. Daraus folgt mit

$$(f \circ g)(t) = \rho\big(\alpha\big(g(t)\big)\, g(t) + A^{-1}Mg(t)\big),$$

$$\frac{d\,(f \circ g)}{dt}(t) = \Big(\nabla\rho\big(\alpha\big(g(t)\big)\, g(t) + A^{-1}Mg(t)\big)\Big)^T \Big(\alpha\big(g(t)\big)\, \dot{g}(t) + \frac{d\,(\alpha \circ g)}{dt}(t)g(t) + A^{-1}M\dot{g}(t)\Big),$$

wobei $\dot{g}(t)$ die Ableitung $(dg/dt)(t)$ bezeichnet, dass

$$\begin{aligned}
0 &= \Big(\nabla\rho\big(\alpha\big(g(t_g)\big)\, g(t_g) + A^{-1}Mg(t_g)\big)\Big)^T \Big(\alpha\big(g(t_g)\big)\, \dot{g}(t_g) + \frac{d\,(\alpha \circ g)}{dt}(t_g)g(t_g) + A^{-1}M\dot{g}(t_g)\Big) \\
&= \Big(\underbrace{\nabla\rho\big(\alpha(u)\, u + A^{-1}Mu\big)}_{\perp_2\ u}\Big)^T \Big(\alpha(u)\, \dot{g}(t_g) + \frac{d\,(\alpha \circ g)}{dt}(t_g)u + A^{-1}M\dot{g}(t_g)\Big) \\
&= \Big(\underbrace{\big(MA^{-1} + \alpha(u)I\big)\nabla\rho\big(\alpha(u)\, u + A^{-1}Mu\big)}_{=:\ w}\Big)^T \dot{g}(t_g).
\end{aligned}$$

Analog zum Beweis von Lemma 2.13 wird gezeigt, dass der Vektor w Euklidisch orthogonal zur Tangentialhyperebene $\dot{u}$ von $\mathcal{N}(\tilde{\lambda})$ an u liegt und damit kollinear zum Normalvektor $\nabla\rho(u)$ ist. Es gibt also ein $\tilde{\beta} \in \mathbb{R}$, so dass $w = \tilde{\beta}\nabla\rho(u)$. Mit den Bezeichnungen $\alpha := \alpha(u)$, $\tilde{u} := \alpha(u)\, u + A^{-1}Mu$ ergibt sich

$$\big(MA^{-1} + \alpha I\big)\frac{2}{\|\tilde{u}\|^2_M}\big(A\tilde{u} - \rho(\tilde{u})M\tilde{u}\big) = \frac{2\tilde{\beta}}{\|u\|^2_M}\big(Au - \rho(u)Mu\big).$$

Mit $\beta := \tilde{\beta}\|\tilde{u}\|^2_M/\|u\|^2_M$ erhalten wir weiter

$$\big(MA^{-1} + \alpha I\big)\big(A - \rho(\tilde{u})M\big)\big(\alpha u + A^{-1}Mu\big) = \beta\big(Au - \rho(u)Mu\big).$$

Das lässt sich mit der Entwicklung $u = \sum_{i=1}^m v_i$ bezüglich der Eigenräume von (A, M) in

$$\sum_{i=1}^m \big(\lambda_i - \rho(\tilde{u})\big)\big(\lambda_i^{-1} + \alpha\big)^2 Mv_i = \sum_{i=1}^m \beta\big(\lambda_i - \rho(u)\big)Mv_i$$

und weiter mit einem kubischen Polynom p in

$$\sum_{i=1}^m p(\lambda_i)\, \lambda_i^{-2}\, Mv_i = \mathbf{0}$$

umschreiben. Multiplizieren der beiden Seiten diesr Gleichung mit $\lambda_j^2 v_j^T$, $j = 1, \ldots, m$ liefert

$$p(\lambda_j)\, \|v_j\|^2_M = \mathbf{0}, \qquad j = 1, \ldots, m,$$

das heißt, für jeden Index j muss mindestens eine der folgenden zwei Aussagen gelten: (1) Der Eigenwert λ_j ist eine Nullstelle von p. (2) Der Vektor u hat einen Nichtnull-Anteil auf $\mathcal{V}_j$.

Hätte u auf mehr als drei Eigenräumen Nichtnull-Anteile, dann müsste p mehr als drei positive Nullstellen besitzen. Das ist ein Widerspruch dazu, dass p ein kubisches Polynom ist. □

Mit Lemma 3.11 kann man aus der Bedingung $\rho(u) \in (\lambda_i, \lambda_{i+1})$ Voraussetzungen von Satz 3.10 konstruieren und weiter die Abschätzung (3.33) modifizieren.

Satz 3.12. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und einen Vektor $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ werde ein Schritt des INVIT(2)-Verfahrens durchgeführt: $u' = \mathrm{RR}_{\rho,\min}\big(\mathcal{K}^2_{A^{-1}M}(u)\big)$.*

Falls u kein Eigenvektor ist, dann gilt $\rho(u') < \rho(u)$. Gelte zusätzlich $\rho(u) \in (\lambda_i, \lambda_{i+1})$ für ein $i \in \{1, \ldots, m-1\}$, dann gilt

$$\frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)} \leq \left(\frac{\check{\kappa}}{2 - \check{\kappa}}\right)^2 \quad \textit{mit} \quad \check{\kappa} := \left(\frac{\lambda_i}{\lambda_{i+1}}\right)\left(\frac{\lambda_m - \lambda_{i+1}}{\lambda_m - \lambda_i}\right). \tag{3.36}$$

Die obere Schranke in (3.36) lässt sich in einem Grenzfall erreichen. (Damit ist (3.36) eine scharfe Abschätzung.)

Beweis. $\rho(u') < \rho(u)$ wurde bereits in Lemma 2.9 für ein zu INVIT(2) äquivalentes Verfahren bewiesen. Für $\rho(u) \in (\lambda_i, \lambda_{i+1})$ gilt entweder $\rho(u') \leq \lambda_i$ oder $\lambda_i < \rho(u') < \rho(u)$. Für den Fall $\rho(u') \leq \lambda_i$ ist der abzuschätzende Delta-Quotient nicht positiv, so dass (3.36) trivial ist. Für den Fall $\lambda_i < \rho(u') < \rho(u)$ betrachten wir die Rayleigh-Quotient-Niveaumenge $\mathcal{N}\big(\rho(u)\big) := \{x \in \mathbb{R}^n \backslash \{\mathbf{0}\} \ ; \ \rho(x) = \rho(u)\}$ und eine Maximalstelle $\tilde{u}$ der Funktion $f(x) := \rho\big(\alpha(x)\, x + A^{-1}Mx\big)$ auf $\mathcal{N}\big(\rho(u)\big)$. Mit $\tilde{u}' := f(\tilde{u})$ gilt dann

$$\lambda_i < \rho(u') \leq \rho(\tilde{u}') < \rho(\tilde{u}) = \rho(u) < \lambda_{i+1}, \tag{3.37}$$

daher ist $\rho(\tilde{u}')$ kein Eigenwert.

Nach Lemma 3.11 hat $\tilde{u}$ auf höchstens drei Eigenräumen Nichtnull-Anteile. Da $\rho(\tilde{u}')$ kein Eigenwert ist, soll $\tilde{u}$ auf genau drei Eigenräumen Nichtnull-Anteile haben (denn sonst wäre $\tilde{u}'$ ein Eigenvektor und $\rho(\tilde{u}')$ ein Eigenwert, vergleiche Satz 3.2).

Nummerieren wir diese drei Eigenräume mit Indizes $\sigma(1) < \sigma(2) < \sigma(3)$, dann gilt entweder $\lambda_{\sigma(1)} < \lambda_{\sigma(2)} \leq \lambda_i < \rho(\tilde{u}) < \lambda_{i+1} \leq \lambda_{\sigma(3)}$ oder $\lambda_{\sigma(1)} \leq \lambda_i < \rho(\tilde{u}) < \lambda_{i+1} \leq \lambda_{\sigma(2)} < \lambda_{\sigma(3)}$. Im ersten Fall erhalten wir nach dem Courant-Fischer-Prinzip einen Widerspruch zu (3.37), nämlich $\rho(\tilde{u}') \leq \lambda_{\sigma(2)} \leq \lambda_i$. Im zweiten Fall gilt nach Satz 3.10 angewandt auf $\tilde{u}$ und $\bar{s} = 3$, dass

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\tilde{u}')\big)}{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\tilde{u})\big)} \leq \left(\frac{\hat{\kappa}}{2 - \hat{\kappa}}\right)^2 \quad \text{mit} \quad \hat{\kappa} := \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)\left(\frac{\lambda_{\sigma(3)} - \lambda_{\sigma(2)}}{\lambda_{\sigma(3)} - \lambda_{\sigma(1)}}\right).$$

Wegen $\lambda_{\sigma(1)} \leq \lambda_i < \lambda_{i+1} \leq \lambda_{\sigma(2)} < \lambda_{\sigma(3)} \leq \lambda_m$ gilt

$$\left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}}\right)\left(\frac{\lambda_{\sigma(3)} - \lambda_{\sigma(2)}}{\lambda_{\sigma(3)} - \lambda_{\sigma(1)}}\right) \leq \left(\frac{\lambda_i}{\lambda_{i+1}}\right)\left(\frac{\lambda_m - \lambda_{i+1}}{\lambda_m - \lambda_i}\right),$$

das heißt $\hat{\kappa} \leq \check{\kappa}$. Damit

$$\begin{aligned}\frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\tilde{u}')\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\tilde{u})\big)} &= \left(\frac{\rho(\tilde{u}') - \lambda_i}{\rho(\tilde{u}) - \lambda_i}\right)\left(\frac{\lambda_{i+1} - \rho(\tilde{u})}{\lambda_{i+1} - \rho(\tilde{u}')}\right) \leq \left(\frac{\rho(\tilde{u}') - \lambda_{\sigma(1)}}{\rho(\tilde{u}) - \lambda_{\sigma(1)}}\right)\left(\frac{\lambda_{\sigma(2)} - \rho(\tilde{u})}{\lambda_{\sigma(2)} - \rho(\tilde{u}')}\right) \\ &= \frac{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\tilde{u}')\big)}{\Delta\big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \rho(\tilde{u})\big)} \leq \left(\frac{\hat{\kappa}}{2 - \hat{\kappa}}\right)^2 \leq \left(\frac{\check{\kappa}}{2 - \check{\kappa}}\right)^2.\end{aligned}$$

Das liefert zusammen mit $\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big) \leq \Delta\big(\lambda_i, \lambda_{i+1}, \rho(\tilde{u}')\big)$, vergleiche (2.9), und $\rho(\tilde{u}) = \rho(u)$ die Abschätzung (3.36). Die Schranke von (3.36) lässt sich mit denjenigen u, die genau auf den Eigenräumen $\mathcal{V}_i$, $\mathcal{V}_{i+1}$ und $\mathcal{V}_m$ Nichtnull-Anteile haben, in einem ähnlichen Grenzfall wie bei der niedrigdimensionalen Analyse erreichen. □

Die schnellste Konvergenz von INVIT(2) unter der Bedingung $\rho(u) \in (\lambda_i, \lambda_{i+1})$ bezieht sich auf einen trivialen Fall. Im Unterraum $\mathcal{V}_1 \oplus \mathcal{V}_m$ kann man einen Vektor finden, dessen Rayleigh-Quotient den gegebenen Wert hat. Die damit erzeugte neue INVIT(2)-Iterierte ist dann ein Eigenvektor zum Eigenwert λ_1.
Durch die Abschätzungen in Abschnitt 3.2.4 und Abschnitt 3.2.5 kann man auch Mehrschritt-Abschätzungen formulieren. Außerdem lässt sich einfach zeigen, dass das $(A - \bar{\lambda}M)$-Gradientverfahren $u' = \mathrm{RR}_{\rho,\min}\left(\mathcal{K}^2_{(A-\bar{\lambda}M)^{-1}M}(u)\right)$ mit einer positiven unteren Schranke $\bar{\lambda}$ von λ_1 bessere Konvergenzfaktoren mittels Substitutionen der Form $\lambda_{\sigma(i)} \rightarrow \lambda_{\sigma(i)} - \bar{\lambda}$ in den Schranken besitzt.

3.3 Konvergenzanalyse für Verfahren zu s-dimensionalen Krylovräumen

Gemäß Lemma 3.2, Lemma 3.3 und Lemma 3.4 entsteht für den nichttrivialen Fall des Krylovraum-Verfahrens (3.2) folgende Voraussetzung zur niedrigdimensionalen Analyse.

Vorausetzung 3.13. *Eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{(s+1)\times(s+1)}$ besitze Eigenwerte $\mu_1, \mu_2, \ldots, \mu_{s+1}$, und $\theta_1, \theta_2, \ldots, \theta_s$ seien Ritzwerte von H bezüglich eines Krylovraums $\mathcal{K}^s_H(x)$. Der Rayleigh-Quotient $\mu(\cdot)$ bezüglich H sei wie in (3.4) definiert und es gelte eine strikte Einschachtelung (3.7). Weiter seien $y_1, y_2, \ldots, y_{s+1}$ Euklidisch normierte Eigenvektoren zu $\mu_1, \mu_2, \ldots, \mu_{s+1}$ und x' ein Ritzvektor zu θ_1.*

3.3.1 Konvergenzabschätzung für Ritzvektoren

Bei der Verallgemeinerung der klassischen Konvergenzanalyse aus Abschnitt 3.2.1 für INVIT(2) auf das Krylovraumverfahren (3.2) haben wir eine Tangens-Abschätzung mit Konvergenzfaktor

$$\prod_{j=3}^{s+1} \frac{\mu_2 - \mu_j}{\mu_1 - \mu_j}$$

entdeckt und relativ technisch bewiesen. Wir formulieren hier einen vereinfachten Beweis nach Anregung der neuen Beweistechnik aus Abschnitt 3.2.3. Zunächst werden die geometrischen Argumente in Lemma 3.9 verallgemeinert.

Lemma 3.14. *Unter Voraussetzung 3.13 sei ein affiner Raum*

$$\mathcal{E}(\alpha_1) := \alpha_1 y_1 + \mathrm{span}\{y_2, \ldots, y_{s+1}\}$$

für $\alpha_1 \in \mathbb{R}\backslash\{0\}$ definiert. Für ein $\tilde{\mu} \in (\mu_2, \mu_1)$ ist dann der Durchschnitt der Rayleigh-Quotient-Niveaumenge $\mathcal{N}(\tilde{\mu}) := \{w \in \mathbb{R}^{s+1}\backslash\{\mathbf{0}\} \ ; \ \mu(w) = \tilde{\mu}\}$ mit $\mathcal{E}(\alpha_1)$ ein Ellipsoid, dessen Halbachsen durch $|\alpha_1|\sqrt{\Delta(\mu_1, \mu_j, \tilde{\mu})}$, $j = 2, \ldots, s+1$ gegeben sind.

Beweis. Ein beliebes $w \in \mathcal{E}(\alpha_1)$ lässt sich durch $w = \sum_{j=1}^{s+1} \alpha_j y_j$ darstellen. Wegen

$$\mu(w) = \frac{\sum_{j=1}^{s+1} \mu_j \alpha_j^2}{\sum_{j=1}^{s+1} \alpha_j^2} = \tilde{\mu} \quad \Leftrightarrow \quad \sum_{j=2}^{s+1} \frac{\alpha_j^2}{\alpha_1^2 \Delta(\mu_1, \mu_j, \tilde{\mu})} = 1$$

und $\Delta(\mu_1, \mu_j, \tilde{\mu}) > 0$ für $j = 2, \ldots, s+1$ ist $\mathcal{E}(\alpha_1) \cap \mathcal{N}(\tilde{\mu})$ das behauptete Ellipsoid. □

Betrachten wir nun $\mathcal{N}\big(\mu(x')\big)$ mit einem beliebigen $\alpha_1 \in \mathbb{R}\backslash\{0\}$. Der Krylovraum $\mathcal{K}_H^s(x)$ ist eine Tangentialhyperebene von $\mathcal{N}\big(\mu(x')\big)$ in $\mathbb{R}^{s+1}$, so dass $\mathcal{E}(\alpha_1)\cap\mathcal{K}_H^2(x)$ eine Tangentialhyperebene des Ellipsoids $\mathcal{E}(\alpha_1)\cap\mathcal{N}\big(\mu(x')\big)$ in $\mathcal{E}(\alpha_1)$ ist. Der Berührungspunkt lässt sich durch eine Umskalierung von x' erhalten. Damit können wir die oben genannte Tangens-Abschätzung beweisen.

Satz 3.15. *Unter Voraussetzung 3.13 gilt*

$$\left|\frac{\tan\angle_2(x',y_1)}{\tan\angle_2(x,y_1)}\right| \le \kappa \quad \textit{mit} \quad \kappa := \prod_{j=3}^{s+1}\frac{\mu_2-\mu_j}{\mu_1-\mu_j}. \tag{3.38}$$

Die obere Schranke in (3.38) lässt sich in einem Grenzfall erreichen. (Damit ist (3.38) eine scharfe Abschätzung.)

Beweis. Da die Abschätzung (3.38) bei Umskalierungen von x und x' invariant ist, setzen wir

$$x = y_1 + \sum_{j=2}^{s+1}\beta_j y_j, \qquad x' = y_1 + \sum_{j=2}^{s+1}\beta_j' y_j$$

und untersuchen das Konvergenzverhalten im affinen Raum $\mathcal{E}(1)$. Es gilt also

$$\tan^2\angle_2(x',y_1) = \sum_{j=2}^{s+1}{\beta_j'}^2, \qquad \sum_{j=2}^{s+1}\frac{{\beta_j'}^2}{\Delta\big(\mu_1,\mu_j,\mu(x')\big)} = 1,$$

so dass

$$\tan^2\angle_2(x',y_1) \le {\beta_2'}^2 + \sum_{j=3}^{s+1}\frac{\Delta\big(\mu_1,\mu_2,\mu(x')\big)}{\Delta\big(\mu_1,\mu_j,\mu(x')\big)}{\beta_j'}^2 = \Delta\big(\mu_1,\mu_2,\mu(x')\big). \tag{3.39}$$

Damit ist $|\tan\angle_2(x',y_1)|$ nicht größer als die größte Halbachse des Ellipsoids $\mathcal{E}(\alpha_1)\cap\mathcal{N}\big(\mu(x')\big)$. Zur Konstruktion einer oberen Schranke der größten Halbachse wählen wir einen Punkt, welcher auf der dieser Halbachse entsprechenden Geraden $y_1+\operatorname{span}\{y_2\}$ und außerhalb der Ellipse liegt. Ein passender Punkt ist der Schnittpunkt w der Tangentialhyperebene $\mathcal{E}(\alpha_1)\cap\mathcal{K}_H^2(x)$ mit $y_1+\operatorname{span}\{y_2\}$. Wegen $w\in\mathcal{K}_H^s(x)$ stellen wir als

$$w = \sum_{i=1}^{s}\zeta_i H^{i-1}x = \sum_{i=1}^{s}\zeta_i\sum_{j=1}^{s+1}\beta_j\mu_j^{i-1}y_j = \sum_{j=1}^{s+1}\beta_j\sum_{i=1}^{s}\zeta_i\mu_j^{i-1}y_j = \sum_{j=1}^{s+1}\beta_j p_{s-1}(\mu_j)y_j$$

mit $\beta_1=1$ und einem Polynom $p_{s-1}(\delta) := \sum_{i=1}^{s}\zeta_i\delta^{i-1}$ dar.

Nach Voraussetzung 3.13 sind die Koeffizienten $\beta_2,\dots,\beta_{s+1}$ ebenfalls von Null verschieden. Wegen $w\neq\mathbf{0}$ und der Darstellung $w=\sum_{j=1}^{s+1}\beta_j p_{s-1}(\mu_j)y_j$ ist dann p_{s-1} keine Nullfunktion. Wegen $w\in\big(y_1+\operatorname{span}\{y_2\}\big)$ liegt w zu den Eigenvektoren $y_3,\dots,y_{s+1}$ Euklidisch orthogonal. Zusammen mit der Orthonormalitätseigenschaft der Eigenvektoren $y_1,\dots,y_{s+1}$ gilt also

$$0 = y_k^T w = \sum_{j=1}^{s+1}\beta_j p_{s-1}(\mu_j)y_k^T y_j = \beta_k p_{s-1}(\mu_k) \quad \text{für} \quad k=3,\dots,s{+}1.$$

Wegen $\beta_k\neq 0$ gilt $p_{s-1}(\mu_k)=0$ für $k=3,\dots,s{+}1$. In anderen Worten sind diese μ_k Nullstellen von p_{s-1}. Da μ_k paarweise verschieden sind, haben wir sämtliche Nullstellen von p_{s-1} gefunden, so dass es ein $\tilde\zeta\neq 0$ mit

$$p_{s-1}(\delta) = \tilde\zeta\prod_{k=3}^{s+1}(\delta-\mu_k)$$

gibt. Damit gilt

$$w = \sum_{j=1}^{s+1} \beta_j p_{s-1}(\mu_j) y_j = \sum_{j=1}^{s+1} \beta_j \tilde{\zeta} \left(\prod_{k=3}^{s+1} (\mu_j - \mu_k) \right) y_j.$$

Nutzen wir nochmal die Bedingung $w \in \big(y_1 + \operatorname{span}\{y_2\}\big)$, erhalten wir mit $\beta_1 = 1$ und einer Normierung bezüglich des Koeffizieten zu y_1, dass

$$w = y_1 + \beta_2 \left(\prod_{j=3}^{s+1} \frac{\mu_2 - \mu_j}{\mu_1 - \mu_j} \right) y_2 = y_1 + \beta_2 \kappa y_2.$$

Weiter wird durch den quadratischen Euklidischen Abstand $\|w - y_1\|_2^2 = \kappa^2 \beta_2^2$ eine obere Schranke von $\Delta\big(\mu_1, \mu_2, \mu(x')\big)$ gegeben. Das liefert zusammen mit (3.39), dass

$$\tan^2 \angle_2(x', y_1) \le \kappa^2 \beta_2^2 \le \kappa^2 \sum_{j=2}^{s+1} \beta_j^2 = \kappa^2 \tan^2 \angle_2(x, y_1)$$

und weiter (3.38). Die Schranke κ ist der Grenzwert des betragsmäßigen Tangens-Quotienten für $x \to y_2$. Dabei konvergieren β_2^2 und ${\beta'}_2^2$ beide gegen ∞. □

Als eine Folgerung gilt auch eine Tangens-Abschätzung bezüglich des H-Skalarproduktes. Mit

$$\begin{aligned} \tan^2 \angle_H(x', y_1) &= \frac{1}{\mu_1} \sum_{j=2}^{s+1} \mu_j {\beta'_j}^2 = \frac{\mu_2}{\mu_1} \left({\beta'_2}^2 + \sum_{j=3}^{s+1} \frac{\mu_j}{\mu_2} {\beta'_j}^2 \right) \le \frac{\mu_2}{\mu_1} \sum_{j=2}^{s+1} {\beta'_j}^2 = \frac{\mu_2}{\mu_1} \tan^2 \angle_2(x', y_1) \\ &\le \frac{\mu_2}{\mu_1} \kappa^2 \beta_2^2 \le \kappa^2 \frac{\mu_2}{\mu_1} \left(\beta_2^2 + \sum_{j=3}^{s+1} \frac{\mu_j}{\mu_2} \beta_j^2 \right) = \kappa^2 \tan^2 \angle_H(x, y_1) \end{aligned}$$

gilt also

$$\left| \frac{\tan \angle_H(x', y_1)}{\tan \angle_H(x, y_1)} \right| \le \kappa. \tag{3.40}$$

Die Schärfe der Schranke κ in der Abschätzung (3.40) lässt sich ebenfalls im Grenzfall $x \to y_2$ zeigen. Basierend auf (3.38) und (3.40) werden gewünschte Abschätzungen für originale Verfahrensformulierungen erhalten.

Satz 3.16. *Mit den Formulierungen in Lemma 3.2, Lemma 3.3 und Lemma 3.4 sei für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und das Verfahren (3.2) der nichttriviale Fall $s < \bar{s}$ betrachtet. Es gilt*

$$\frac{\tan \angle_A\big(u', \mathcal{V}_{\sigma(1)}\big)}{\tan \angle_A\big(u, \mathcal{V}_{\sigma(1)}\big)} \le \hat{\kappa}, \qquad \frac{\tan \angle_M\big(u', \mathcal{V}_{\sigma(1)}\big)}{\tan \angle_M\big(u, \mathcal{V}_{\sigma(1)}\big)} \le \hat{\kappa} \tag{3.41}$$

$$\textit{mit} \qquad \hat{\kappa} := \left(\frac{\lambda_{\sigma(1)}}{\lambda_{\sigma(2)}} \right)^{s-1} \prod_{j=1}^{s-1} \left(\frac{\lambda_{\sigma(\bar{s}+1-j)} - \lambda_{\sigma(2)}}{\lambda_{\sigma(\bar{s}+1-j)} - \lambda_{\sigma(1)}} \right).$$

Die obere Schranke $\hat{\kappa}$ lässt sich in einem Grenzfall erreichen. (Damit sind die Abschätzungen in (3.41) scharf.)

Beweis. Mit den Formulierungen in Lemma 3.4 gilt nach dem Courant-Fischer-Prinzip, dass

$$\mu_1 = \lambda_{\sigma(1)}^{-1} > \lambda_{\sigma(2)}^{-1} \ge \mu_2, \qquad \mu_j \ge \lambda_{\sigma(\bar{s}-(s+1)+j)}^{-1} \quad \text{für} \quad j = 3, \ldots, s{+}1.$$

Damit wird gezeigt, dass

$$\kappa = \prod_{j=3}^{s+1} \frac{\mu_2 - \mu_j}{\mu_1 - \mu_j} \leq \prod_{j=3}^{s+1} \frac{\lambda_{\sigma(2)}^{-1} - \lambda_{\sigma(\bar{s}-(s+1)+j))}^{-1}}{\lambda_{\sigma(1)}^{-1} - \lambda_{\sigma(\bar{s}-(s+1)+j))}^{-1}} = \hat{\kappa}.$$

Tangensabschätzungen in (3.41) folgen dann jeweils aus (3.9), (3.38) und (3.8), (3.40):

$$\frac{\tan \angle_A\big(u', \mathcal{V}_{\sigma(1)}\big)}{\tan \angle_A\big(u, \mathcal{V}_{\sigma(1)}\big)} = \left| \frac{\tan \angle_2(x', y_1)}{\tan \angle_2(x, y_1)} \right| \leq \kappa \leq \hat{\kappa},$$

analog $\tan \angle_M\big(u', \mathcal{V}_{\sigma(1)}\big) / \tan \angle_M\big(u, \mathcal{V}_{\sigma(1)}\big) \leq \hat{\kappa}$. Die obere Schranke $\hat{\kappa}$ lässt sich mit denjenigen u, die auf genau $s{+}1$ Eigenräumen Nichtnull-Anteile haben, also mit $\bar{s} = s{+}1$, in einem ähnlichen Grenzfall wie bei der niedrigdimensionalen Analyse erreichen. □

3.3.2 Konvergenzabschätzung für Ritzwerte

Mit weiteren geometrischen Argumenten können wir eine Konvergenzabschätzung für Ritzwerte formulieren. Die Schranke hat eine relativ komplizierte Darzustellung und lässt sich mehrstufig definieren.

Satz 3.17. *Unter Voraussetzung 3.13 gilt für $\mu(x) \in (\mu_2, \mu_1)$, dass*

$$0 < \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\Delta\big(\mu_1, \mu_2, \mu(x)\big)} \leq \frac{\Big(\sum_{j=2}^{s+1} \kappa_j^{-1}\Big)^2}{\Big(\sum_{j=2}^{s+1} |\kappa_j|^{-1}\Big)^2} \quad \textit{mit} \quad \kappa_j := \prod_{l=2,\ l\neq j}^{s+1} \frac{\mu_j - \mu_l}{\mu_1 - \mu_l}. \tag{3.42}$$

Die obere Schranke in (3.42) lässt sich in einem Grenzfall erreichen. (Damit ist (3.42) eine scharfe Abschätzung.)

Beweis. Basierend auf dem Beweis von Satz 3.15 sind die zwei Delta-Werte jeweils das Quadrat der größten Halbachse eines nicht-entarteten Ellipsoids, so dass deren Quotient positiv ist. Weiter stellen wir die Vektoren $w = y_1 + \sum_{j=2}^{s+1} \alpha_j y_j$ aus $\mathcal{E}(1)$ durch die entsprechenden s-dimensionalen Koeffizientenvektoren $Pw = (\alpha_2, \ldots, \alpha_{s+1})^T$ dar. Betrachten wir das Ellipoid $\mathcal{E}(1) \cap \mathcal{N}\big(\mu(x')\big)$ als die Einheitskugel der Vektornorm $\|\cdot\|_D$ mit

$$D = \operatorname{diag}(\delta_2, \ldots, \delta_{s+1}), \qquad \delta_j := 1/\Delta\big(\mu_1, \mu_j, \mu(x')\big), \quad j = 2, \ldots, s{+}1.$$

Da x' der Berührungspunkt der Tangentialhyperebene $\mathcal{E}(1) \cap \mathcal{K}_H^s(x)$ ist, liegt Px' D-orthogonal zum Richtungsanteil der $\mathcal{E}(1)$-Darstellung von $\mathcal{E}(1) \cap \mathcal{K}_H^s(x)$ (diese Darstellung ist ja auch ein affiner Raum). Um diese $\mathcal{E}(1)$-Darstellung zu konstruierten, bestimmen wir die Schnittpunkte von $\mathcal{E}(1) \cap \mathcal{K}_H^s(x)$ mit den affinen Räumen $y_1 + \operatorname{span}\{y_j\}$, $j = 2, \ldots, s{+}1$. Der Schnittpunkt für $j = 2$ wurde bereits im Beweis von Satz 3.15 ermittelt. In ähnlicher Weise werden die weiteren Schnittpunkte bestimmt. So sind sämtliche Schnittpunkte durch

$$y_1 + \kappa_j \beta_j y_j, \quad j = 2, \ldots, s+1 \quad \text{mit} \quad \kappa_j = \prod_{l=2,\ l\neq j}^{s+1} \frac{\mu_j - \mu_l}{\mu_1 - \mu_l}$$

gegeben. Die $\mathcal{E}(1)$-Darstellung von $\mathcal{E}(1) \cap \mathcal{K}_H^s(x)$ wird dann durch Verbindungsvektoren von $y_1 + \kappa_2 \beta_2 y_2$ mit anderen Schnittpunkten erhalten, nämlich

$$\begin{aligned} &(\kappa_2\beta_2, 0, \ldots, 0)^T + \bar{\mathcal{W}} \quad \text{mit} \quad \bar{\mathcal{W}} := \operatorname{span}\{w_3, \ldots, w_{s+1}\}, \\ &w_j := (\kappa_2\beta_2, 0, \ldots, 0, -\kappa_j\beta_j, 0, \ldots, 0)^T \in \mathbb{R}^s, \quad j = 3, \ldots, s{+}1. \end{aligned} \tag{3.43}$$

Damit liegt Px' D-orthogonal zum $(s-1)$-dimensionalen Unterraum $\bar{\mathcal{W}}$ von $\mathbb{R}^s$, so dass span$\{Px'\}$ das D-orthogonale Komplement von $\bar{\mathcal{W}}$ ist.

Wir konstruieren einen weiteren Nichtnullvektor, welcher zu diesem Komplement gehört. Mit

$$y := \left(\frac{1}{\delta_2\kappa_2\beta_2}, \ldots, \frac{1}{\delta_{s+1}\kappa_{s+1}\beta_{s+1}}\right)^T$$

werden die D-Orthogonalitäten $y^T D w_j = 0$ für $j = 3, \ldots, s+1$ erfüllt, daher ist das Komplement span$\{Px'\}$ auch durch y aufgespannt.

Da Px und Px' beide zum affinen Raum in (3.43) gehören, gilt $Px - Px' \in \bar{\mathcal{W}}$. Wegen der D-Orthogonalität $Px' \perp_D \bar{\mathcal{W}}$ gilt dann $Px - Px' \perp_D Px'$, so dass (siehe Definition 2.2) Px' die D-Projektion von Px auf span$\{Px'\}$ = span$\{y\}$ ist und

$$\begin{aligned}
\cos^2 \angle_D(Px, y) = \cos^2 \angle_D\big(Px, \text{span}\{y\}\big) = \frac{\|Px'\|_D^2}{\|Px\|_D^2} &= \frac{1}{\sum_{j=2}^{s+1} \delta_j \beta_j^2} = \frac{\delta_2^{-1}}{\beta_2^2 + \sum_{j=3}^{s+1} \frac{\delta_2^{-1}}{\delta_j^{-1}} \beta_j^2} \\
&= \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\beta_2^2 + \sum_{j=3}^{s+1} \frac{\mu(x') - \mu_j}{\mu(x') - \mu_2} \beta_j^2} \geq \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\beta_2^2 + \sum_{j=3}^{s+1} \frac{\mu(x) - \mu_j}{\mu(x) - \mu_2} \beta_j^2} \qquad (3.44) \\
&= \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\beta_2^2 + \sum_{j=3}^{s+1} \frac{\Delta\big(\mu_1, \mu_2, \mu(x)\big)}{\Delta\big(\mu_1, \mu_j, \mu(x)\big)} \beta_j^2} = \frac{\Delta\big(\mu_1, \mu_2, \mu(x')\big)}{\Delta\big(\mu_1, \mu_2, \mu(x)\big)}.
\end{aligned}$$

Dabei ist zu bemerken, dass x zu einem Ellipsoid mit Halbachsen $\Delta\big(\mu_1, \mu_j, \mu(x)\big)$, $j = 2, \ldots, s+1$ gehört. Mit den Komponenten von Px und y können wir $\cos^2 \angle_D(Px, y)$ nach oben abschätzen. Es gilt

$$\cos^2 \angle_D(Px, y) = \frac{(Px, y)_D^2}{\|Px\|_D^2 \|y\|_D^2} = \frac{\left(\sum_{j=2}^{s+1} \kappa_j^{-1}\right)^2}{\left(\sum_{j=2}^{s+1} \delta_j \beta_j^2\right) \left(\sum_{j=2}^{s+1} \left(\delta_j \kappa_j^2 \beta_j^2\right)^{-1}\right)} \leq \frac{\left(\sum_{j=2}^{s+1} \kappa_j^{-1}\right)^2}{\left(\sum_{j=2}^{s+1} |\kappa_j|^{-1}\right)^2},$$

denn nach der Cauchy-Schwarz-Ungleichung erhalten wir

$$\left(\sum_{j=2}^{s+1} |\kappa_j|^{-1}\right)^2 = \left(\sum_{j=2}^{s+1} \left(\sqrt{\delta_j}|\beta_j|\right)\left(\sqrt{\delta_j}|\kappa_j|\,|\beta_j|\right)^{-1}\right)^2 \leq \left(\sum_{j=2}^{s+1} \delta_j \beta_j^2\right) \left(\sum_{j=2}^{s+1} (\delta_j \kappa_j^2 \beta_j^2)^{-1}\right).$$

Einsetzen in (3.44) wird die Abschätzung (3.42) gezeigt. Die obere Schranke lässt sich im Grenzfall $\mu(x) \to \mu_1$ erreichen. Dabei gilt außerdem $\mu(x') \to \mu_1$. □

Ein Rückwärtseinsetzen der Abschätzung (3.42) ist für ein allgemeines $s \geq 3$ keine einfache Aufgabe. Die Monotonien der niedrigdimensionalen Schranke bezüglich der abstrakten Variablen μ_j sind für $s = 3$ bereits nicht alle eindeutig. Die symbolische Darstellung der allgemeinen Schranke wird deswegen komplizierter.

Satz 3.18. *Mit den Formulierungen in Lemma 3.2 Lemma 3.3 und Lemma 3.4 sei für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und das Verfahren (3.2) der nichttriviale Fall $s < \bar{s}$ betrachtet. Es gibt für $\rho(u) \in \big(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}\big)$ eine s-elementige Untermenge J der Indexmenge*

$\{\sigma(2),\dots,\sigma(\bar{s})\}$, *so dass*

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u)\big)} \le \frac{\Big(\sum_{j\in J}\hat{\kappa}_j^{-1}\Big)^2}{\Big(\sum_{j\in J}|\hat{\kappa}_j|^{-1}\Big)^2} \tag{3.45}$$

$$mit \quad \hat{\kappa}_j := \prod_{l\in J,\ l\neq j}\left(\frac{\lambda_{\sigma(1)}}{\lambda_j}\right)\left(\frac{\lambda_l-\lambda_j}{\lambda_l-\lambda_{\sigma(1)}}\right).$$

Beweis. Für $\rho(u)\in\big(\lambda_{\sigma(1)},\,\lambda_{\sigma(2)}\big)$, gilt zunächst die Bedingung $\mu(x)\in(\mu_2,\mu_1)$.
Dann folgt aus (3.10), (3.14), dass

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u)\big)} \le \frac{\Delta\big(\mu_1,\mu_2,\mu(x')\big)}{\Delta\big(\mu_1,\mu_2,\mu(x)\big)} \le \frac{\Big(\sum_{j=2}^{s+1}\kappa_j^{-1}\Big)^2}{\Big(\sum_{j=2}^{s+1}|\kappa_j|^{-1}\Big)^2}.$$

Diese Schranke lässt sich als eine Funktion von den Variablen $\mu_2,\dots,\mu_{s+1}$ betrachten. Für konkrete $\mu_2,\dots,\mu_{s+1}$ lassen sich die Monotonien bezüglich $\mu_2,\dots,\mu_{s+1}$ bestimmen. Die reslutierende Abschätzung hat dann die Form (3.45). □

Die obere Schranke in Satz 3.18 hat also nicht notwendig eine Abhängigkeit von bestimmten Eigenwerten . Dazu empfehlen wir, das Quadrat der Schranke $\hat{\kappa}$ aus Satz 3.16 als eine klarere (aber nicht scharfe) Schranke des Delta-Quotienten zu nutzen. Dafür ist eine weitere niedrigdimensionale Analyse nötig.

Satz 3.19. *Unter Voraussetzung 3.13 gilt für $\mu(x)\in(\mu_2,\,\mu_1)$, dass*

$$0 < \frac{\Delta\big(\mu_1,\mu_2,\mu(x')\big)}{\Delta\big(\mu_1,\mu_2,\mu(x)\big)} \le \kappa^2 \quad mit \quad \kappa := \prod_{j=3}^{s+1}\frac{\mu_2-\mu_j}{\mu_1-\mu_j}. \tag{3.46}$$

Beweis. Dass der Delta-Quotient positiv ist, wurde bereits in Satz 3.17 bewiesen. Basierend auf dem Beweis von Satz 3.15 ist $\kappa^2\beta_2^2$ eine obere Schranke von $\Delta\big(\mu_1,\mu_2,\mu(x')\big)$. Außerdem gilt

$$\beta_2^2 \le \beta_2^2 + \sum_{j=3}^{s+1}\frac{\Delta\big(\mu_1,\mu_2,\mu(x)\big)}{\Delta\big(\mu_1,\mu_j,\mu(x)\big)}\beta_j^2 = \Delta\big(\mu_1,\mu_2,\mu(x)\big).$$

Damit ist κ^2 eine obere Schranke des Delta-Quotienten. □

Als Folgerung erhalten wir die Abschätzung

$$0 < \frac{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u')\big)}{\Delta\big(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\rho(u)\big)} \le \hat{\kappa}^2 \tag{3.47}$$

für $\rho(u)\in\big(\lambda_{\sigma(1)},\,\lambda_{\sigma(2)}\big)$ unter Voraussetzungen von Satz 3.16.

Ausgehend von den Bedingungen der Form $\rho(u)\in(\lambda_i,\lambda_{i+1})$ können wir wie in Lemma 3.11 zeigen, dass eine Extremstelle in einem niedrigdimensionalen invarianten Unterraum liegen soll. Jedoch wird für s-dimensionale Krylovräume damit ein $(2s-1)$-dimensionaler invarianter Unterraum betrachtet. Wegen $2s-1 > s+1$ für $s\ge 3$ ist dieses Resultat nicht sehr nützlich zur Erweiterung der Abschätzungen (3.45) und (3.47) für allgemeinere Bedingungen.

Die schnellste Konvergenz des Krylovraum-Verfahrens (3.2) unter der Bedingung $\rho(u)\in(\lambda_i,\lambda_{i+1})$ bezieht sich wie INVIT(2) auch auf einen trivialen Fall. Man kann beispielsweise im Unterraum $\mathcal{V}_1\oplus\cdots\oplus\mathcal{V}_{s-1}\oplus\mathcal{V}_m$ einen Vektor finden, dessen Rayleigh-Quotient den gegebenen Wert hat. Die damit erzeugte neue Iterierte ist dann ein Eigenvektor zu λ_1.

Außerdem lassen sich Mehrschritt-Abschätzungen und Shift-Abschätzungen basierend auf den obigen Abschätzungen leicht konstruieren.

4 Vorkonditionierte Gradientenverfahren

In diesem Kapitel untersuchen wir die zwei Einschrittverfahren aus der PINVIT-Hierarchie, nämlich PINVIT(1) und PINVIT(2), wobei der Vorkonditionierer B^{-1} bezüglich der Voraussetzung 2.1 die Bedingung (1.4) erfüllen soll. PINVIT-Verfahren können als vorkonditionierte Gradientenverfahren beziehungsweise gestörte INVIT-Verfahren interpretiert werden. Die durch Vorkonditionierer verursachten Störungen können oft die Anzahl der äußeren Schritte erhöhen, jedoch auch den gesamten Aufwand reduzieren. Es gibt auch „Störungen“ für die mit der Konvergenzanalyse für INVIT-Verfahren methodisch ähnliche Konvergenzanalyse. Die Abschätzungen für Tangens-Quotienten sind für PINVIT-Verfahren nicht sinnvoll einsetzbar, denn die Tangenswerte zu den Vektoriterierten bilden nicht notwendig eine monoton fallende Folge. Trotzdem haben wir schon in Kapitel 2 gezeigt, dass die Rayleigh-Quotient-Verkleinerung durch die Bedingung (1.4) des Vorkonditionierers garantiert wird. Ausgehend von $\rho(u) \in (\lambda_i, \lambda_{i+1})$ kann man immer noch vernünftige Abschätzungen für Delta-Quotienten formulieren.

Eine klassische Konvergenzabschätzung für PINVIT(1) ist seit langem bekannt, siehe [60], [61]. Diese ist anders als eine Delta-Abschätzung und gibt eine längliche Schranke direkt für den Rayleigh-Quotienten der neuen Iterierten. In [49], [50], [4] sowie unserer weiteren Untersuchung wurden nacheinander einfachere Versionen formuliert, indem immer klarere geometrische Argumente verwendet wurden. Für PINVIT(2) wurde erst neulich die erste scharfe Delta-Abschätzung unter der Bedingung (1.4) entwickelt (eine scharfe Abschätzung unter einer relativ speziellen Bedingung wurde bereits durch Ovtchinnikov [74] vorgestellt), denn ein Schlüssel zu diesem Ergebnis wurde erst bei der Beweis-Vereinfachung der Konvergenzabschätzungen für INVIT(2), nämlich nach Anregung der Analyse in einem affinen Raum, erhalten.

Als der Ausgangspunkt unserer Konvergenznalyse für PINVIT-Verfahren wird in Abschnitt 4.1 eine geometrische Interpretation der Vorkonditionierer in Abhängigkeit von (1.4) vorgestellt. Da die anschließende Analyse in einer A-Geometrie durchgeführt wird, werden die originalen Problemstellungen der Einfachheit halber zunächst umformuliert. Für PINVIT(1) werden in Abschnitt 4.2 zwei Versionen der Konvergenzanlyse vergleichsweise formuliert, wobei einige neue Argumente eingesetzt werden und einige Anregungen zur Konvergenzanlyse für PINVIT(2) und blockweise PINVIT(1) erhalten werden. In Abschnitt 4.3 wird unsere neue Konvergenzabschätzung für PINVIT(2) präsentiert, wobei eine modifizierte Ellipsen-Analyse eine wichtige Rolle spielt.

4.1 Geometrische Interpretation der Verfahren

In Abschnitt 2.1 haben wir bereits einige Vorüberlegungen für PINVIT(1) und PINVIT(2) angestellt. Es wurde jeweils eine einfache Konvergenzabschätzung gezeigt, nämlich eine Rayleigh-Quotient-Verkleinerung unter der Bedingung (1.4), und durch eine Umschreibung der Vorschrift (1.5) von PINVIT(1) erhalten wir eine Abstand-Bedingung

$$\left\|\rho(u)A^{-1}Mu - u'\right\|_A \leq \gamma \left\|\rho(u)A^{-1}Mu - u\right\|_A$$

bezüglich der A-Vektornorm. Für genauere Konvergenzabschätzungen wird der Parameter γ zur Konstruktion der Schranken benötigt. Gemäß dieser Abstand-Bedingung sehen wir für ein festes $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$, dass die neue Iterierte u' als Punkt in einer A-Kugel (das heißt in einem Ellipsoid) mit dem Mittelpunkt $\rho(u)A^{-1}Mu$ und dem Radius $\gamma\left\|\rho(u)A^{-1}Mu - u\right\|_A$ eingeschlossen wird. Daher können wir ein Optimierungsproblem des Rayleigh-Quotienten auf dieser A-Kugel betrachten und Bedingungen für Extremstellen konstruieren, um ein $\rho(u')$ enthaltendes Intervall mittels

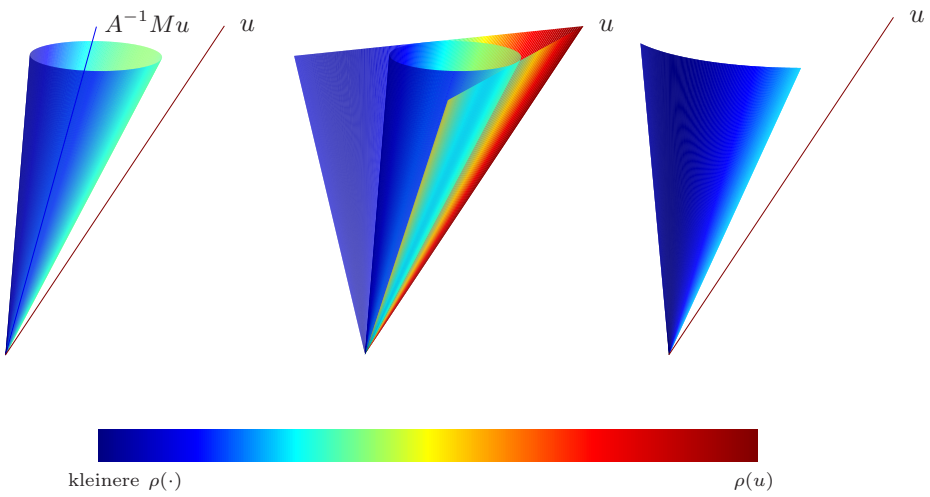


Abbildung 4.1: Geometrische Interpretation für PINVIT(1) und PINVIT(2)

Zwischengrößen zu bestimmen. Für $\gamma = 0$ wird diese A-Kugel entartet. Aber in diesem Fall liefert die Bedingung (1.4) des Vorkonditionierers die Gleichung $B^{-1} = A^{-1}$, so dass PINVIT-Verfahren INVIT-Verfahren werden. Für die zugehörige Konvergenzanalyse stehen also die Abschätzungen aus Kapitel 2 und Kapitel 3 zur Verfügung. Daher setzen wir im Folgenden $\gamma \in (0, 1)$. Um einen Zusammenhang zwischen PINVIT(1) und PINVIT(2) zu konstruieren, können wir einen der obigen A-Kugel entsprechenden A-Kegel betrachten. Um Verwechselung zu vermeiden, bezeichnen wir mit $\hat{u}$ die neue Iterierte von PINVIT(2). Mit den Iterierten u, u' von PINVIT(1) ist $\hat{u}$ dann ein Ritzvektor zum kleinsten Ritzwert von (A, M) bezüglich $\operatorname{span}\{u, u'\}$. Dieser Unterraum gehört zur Menge aller Unterräume, die jeweils durch u und einen im oben genannten A-Kegel eingeschlossen Vektor aufgespannt werden. So können wir für PINVIT(2) ebenfalls ein Optimierungsproblem aufstellen und Zwischengrößen konstruieren.

Um die Analyse klarer zu formulieren, nutzen wir im Folgenden eine A-Geometrie und die entsprechenden Darstellungen der betreffenden Matrizen und Vektoren, vergleiche Abschnitt 3.1.1. Mit beliebigen A-orthonormalen Basismatrizen $V_1, V_2, \ldots, V_m$ der Eigenräume von (A, M), siehe Voraussetzung 2.1, definieren wir $V := [V_1, V_2, \ldots, V_m]$, damit gilt

$$V^T A V = I \in \mathbb{R}^{n\times n}, \qquad V^T M V = \operatorname{diag}\big(\underbrace{\lambda_1^{-1}, \ldots, \lambda_1^{-1}}_{q_1-\text{mal}}, \underbrace{\lambda_2^{-1}, \ldots, \lambda_2^{-1}}_{q_2-\text{mal}}, \ldots, \underbrace{\lambda_m^{-1}, \ldots, \lambda_m^{-1}}_{q_m-\text{mal}}\big)$$

(alternativ können wir eine beliebige $V \in \mathbb{R}^{n\times n}$ mit $V^T A V = I$ verwenden, dann ist $V^T M V$ nicht notwendig eine Diagonalmatrix, aber immer noch symmetrisch und positiv definit, und besitzt λ_i^{-1} als Eigenwerte). Weiter bezeichnen wir die Diagonalmatrix $V^T M V$ mit H und dementsprechend die Diagonalkomponenten λ_i^{-1} von $V^T M V$ mit μ_i. Aus $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m < \infty$ folgt dann $\infty > \mu_1 > \mu_2 > \cdots > \mu_m > 0$. Damit ist H positiv definit. Außerdem definieren wir für jedes $\tilde{u} \in \mathbb{R}^n$ einen Koeffizientenvektor $\tilde{x} := V^T A \tilde{u}$. Mit den zu $V^T A V = I$ äquivalenten Gleichungen $A V V^T = I$ und $V V^T A = I$ erhalten wir eine Norm-Umrechnung

$$\|\tilde{x}\|_2 = \|V^T A \tilde{u}\|_2 = \sqrt{(V^T A \tilde{u})^T V^T A \tilde{u}} = \sqrt{\tilde{u}^T (A V V^T) A \tilde{u}} = \sqrt{\tilde{u}^T A \tilde{u}} = \|\tilde{u}\|_A$$

und für den bezüglich H definierten Rayleigh-Quotienten $\mu(\cdot)$ eine Rayleigh-Quotient-Umrechnung

$$\mu(\tilde{x}) = \frac{\tilde{x}^T H \tilde{x}}{\tilde{x}^T \tilde{x}} = \frac{(V^T A \tilde{u})^T (V^T M V)(V^T A \tilde{u})}{(V^T A \tilde{u})^T (V^T A \tilde{u})} = \frac{\tilde{u}^T (A V V^T) M (V V^T A) \tilde{u}}{\tilde{u}^T (A V V^T) A \tilde{u}} = \frac{\tilde{u}^T M \tilde{u}}{\tilde{u}^T A \tilde{u}} = \frac{1}{\rho(\tilde{u})} \,.$$

Mit den Koeffizientenvektoren x, x' von u, u' erhalten wir weiter

$$\frac{1}{\mu(x)}Hx = \rho(u)(V^TMV)(V^TAu) = \rho(u)V^TMu = \rho(u)V^T(AA^{-1})Mu = V^TA\big(\rho(u)A^{-1}Mu\big),$$

$$\big\|\rho(u)A^{-1}Mu - u'\big\|_A = \big\|V^TA(\rho(u)A^{-1}Mu - u')\big\|_2 = \left\|\frac{1}{\mu(x)}Hx - x'\right\|_2 = \frac{1}{\mu(x)}\big\|Hx - \mu(x)x'\big\|_2,$$

$$\text{und analog} \quad \big\|\rho(u)A^{-1}Mu - u\big\|_A = \frac{1}{\mu(x)}\big\|Hx - \mu(x)x\big\|_2,$$

so dass die Abstand-Bedingung von PINVIT(1) $\big($wegen $1/\mu(x) \neq 0\big)$ in

$$\big\|Hx - \mu(x)x'\big\|_2 \leq \gamma\big\|Hx - \mu(x)x\big\|_2 \tag{4.1}$$

umformuliert werden kann. Der umskalierte Koeffizientenvektor $\mu(x)x'$ von u' wird also als Punkt in einer von x abhängigen Euklidischen Kugel eingeschlossen. In anderen Worten wird x' im entsprechenden von x abhängigen Euklidischen Kegel eingeschlossen, denn aus (4.1) folgt eine Winkel-Bedingung

$$\angle_2(Hx, x') \leq \arcsin\big(\gamma \sin\angle_2(Hx, x)\big). \tag{4.2}$$

Dabei ist zunächst zu zeigen, dass $\angle_2(Hx, x)$ und $\angle_2(Hx, x')$ spitze Winkel sind. Der Winkel $\angle_2(Hx, x)$ ist spitz, denn mit der positiven Definitheit von H gilt $(Hx, x)_2 > 0$ beziehungsweise $\cos\angle_2(Hx, x) > 0$. Wegen $Hx - \mu(x)x \perp_2 x$ ist $\mu(x)x$ die Euklidische Projektion von Hx auf $\operatorname{span}\{x\}$, so dass

$$\big\|Hx - \mu(x)x\big\|_2 = \min_{\alpha\in\mathbb{R}} \|Hx - \alpha x\|_2 \leq \|Hx\|_2.$$

Weiter mit (4.1) und $\gamma < 1$ erhalten wir $\big\|Hx - \mu(x)x'\big\|_2 < \|Hx\|_2$, so dass $\angle_2\big(Hx, \mu(x)x'\big)$ nicht der größte Innenwinkel des von Hx und $\mu(x)x'$ gebildeten Dreiecks und daher spitz ist. Wegen $\mu(x) > 0$ ist $\angle_2(Hx, x')$ identisch zu $\angle_2\big(Hx, \mu(x)x'\big)$ und auch spitz. Anschließend lässt sich eine Sinusungleichung zeigen:

$$\sin\angle_2\big(Hx, x'\big) = \sin\angle_2\big(Hx, \operatorname{span}\{x'\}\big) = \frac{\min_{\alpha\in\mathbb{R}}\|Hx - \alpha x'\|_2}{\|Hx\|_2} \leq \frac{\|Hx - \mu(x)x'\|_2}{\|Hx\|_2}$$

$$\leq \gamma\frac{\|Hx - \mu(x)x\|_2}{\|Hx\|_2} = \gamma\frac{\min_{\alpha\in\mathbb{R}}\|Hx - \alpha x\|_2}{\|Hx\|_2} = \gamma\sin\angle_2\big(Hx, \operatorname{span}\{x\}\big) = \gamma\sin\angle_2\big(Hx, x\big).$$

Damit wird die Winkel-Bedingung (4.2) gezeigt, so dass x' in einem Euklidischen Kegel um Hx mit einem von x und γ abhängigen Öffnungswinkel eingeschlossen wird.

Im Folgenden werden wir für PINVIT(1) und PINVIT(2) Konvergenzabschätzungen bezüglich der Delta-Quotienten $\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big)/\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)$ und $\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\hat{u})\big)/\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)$ unter der Bedingung $\rho(u) \in (\lambda_i, \lambda_{i+1})$, $i \in \{1, \ldots, m-1\}$, präsentieren. Wir können diese Problemstellung zunächst gemäß obigen geometrischen Argumenten umformulieren. Mit den Eigenwerten μ_i, μ_{i+1} von H, dem Rayleigh-Quotienten $\mu(\cdot)$ bezüglich H sowie den Koeffizientenvektoren x, x', $\hat{x}$ von u, u', $\hat{u}$ gilt

$$\frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)} = \left(\frac{\rho(u') - \lambda_i}{\lambda_{i+1} - \rho(u')}\right)\left(\frac{\rho(u) - \lambda_i}{\lambda_{i+1} - \rho(u)}\right)^{-1} = \left(\frac{\frac{1}{\mu(x')} - \frac{1}{\mu_i}}{\frac{1}{\mu_{i+1}} - \frac{1}{\mu(x')}}\right)\left(\frac{\frac{1}{\mu(x)} - \frac{1}{\mu_i}}{\frac{1}{\mu_{i+1}} - \frac{1}{\mu(x)}}\right)^{-1}$$

$$= \left(\frac{\mu_{i+1}}{\mu_i}\right)\left(\frac{\mu_i - \mu(x')}{\mu(x') - \mu_{i+1}}\right)\left(\frac{\mu_{i+1}}{\mu_i}\right)^{-1}\left(\frac{\mu_i - \mu(x)}{\mu(x) - \mu_{i+1}}\right)^{-1}$$

$$= \left(\frac{\mu_i - \mu(x')}{\mu(x') - \mu_{i+1}}\right)\left(\frac{\mu_i - \mu(x)}{\mu(x) - \mu_{i+1}}\right)^{-1} = \frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(x')\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)},$$

$$\text{und analog} \quad \frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\hat{u})\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)} = \frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(\hat{x})\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)}.$$

Die Bedingung $\rho(u) \in (\lambda_i, \lambda_{i+1})$ ist äquivalent zu $\mu(u) \in (\mu_{i+1}, \mu_i)$. Außerdem ist das Maximum (Minimum) des Rayleigh-Quotienten $\rho(\cdot)$ auf einer A-Kugel identisch zum Kehrwert des Minimums (Maximums) des Rayleigh-Quotienten $\mu(\cdot)$ auf der entsprechenden Euklidischen Kugel, und der kleinste Ritzwert von (A, M) bezüglich eines Unterraums ist identisch zum Kehrwert des größten Ritzwert von H bezüglich des entsprechenden Unterraums der Koeffizientenvektoren.

Gemäß diesen Überlegungen formulieren wir eine Voraussetzung für die weitere Untersuchung.

Voraussetzung 4.1. *Eine positiv definite Diagonalmatrix $H \in \mathbb{R}^{n\times n}$ besitze m verschiedene Eigenwerte $\mu_1 > \mu_2 > \cdots > \mu_m$ mit zugehörigen Vielfachheiten $q_1, q_2, \ldots, q_m$ und zugehörigen Eigenräumen $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_m$. Mit dem bezüglich H definierten Rayleigh-Quotienten $\mu(\cdot)$ gelte $\mu(x) \in (\mu_{i+1}, \mu_i)$ für ein $x \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ und ein $i \in \{1, \ldots, m-1\}$. In Abhängigkeit von x und einem $\gamma \in (0, 1)$ werden eine Vollkugel und ein Vollkegel definiert:*

$$\mathscr{B}_\gamma(x) := \big\{w \in \mathbb{R}^n \ ; \ \|Hx - w\|_2 \le \gamma\big\|Hx - \mu(x)x\big\|_2\big\}, \tag{4.3}$$

$$\mathscr{C}_\gamma(x) := \big\{w \in \mathbb{R}^n \ ; \ \angle_2(Hx, w) \le \arcsin\big(\gamma \sin\angle_2(Hx, x)\big)\big\}. \tag{4.4}$$

Der Vektor $x' \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ erfülle die Bedingung $\mu(x)x' \in \mathscr{B}_\gamma(x)$ und es gelte damit $x' \in \mathscr{C}_\gamma(x)$. Weiter sei $\hat{x}$ ein Ritzvektor zum größten Ritzwert von H bezüglich $\operatorname{span}\{x, x'\}$. *Abzuschätzen sind die Delta-Quotienten $\Delta\big(\mu_i, \mu_{i+1}, \mu(x')\big)/\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)$ und $\Delta\big(\mu_i, \mu_{i+1}, \mu(\hat{x})\big)/\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)$.*

Wir können daher für PINVIT(1) ein Optimierungsproblem bezüglich des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ und für PINVIT(2) ein mit der Rayleigh-Ritz-Prozedur kombiniertes Optimierungsproblem auf $\mathscr{C}_\gamma(x)$ betrachten.

4.2 Konvergenzanalyse für PINVIT(1)

Zur Konvergenzanalyse für PINVIT-Verfahren können wir gleiche Grundidee für INVIT-Verfahren benutzen, nämlich die Kombination der Charakterisierung der extremen Konvergenz und der anschließenden niedrigdimensionalen Analyse. Wir formulieren zunächst einen gemeinsamen Anfangsteil der zwei zu präsentierenden Versionen der Konvergenzanalyse für PINVIT(1).

Lemma 4.2. *Gemäß Voraussetzung 4.1 gilt für jedes $w \in \mathscr{B}_\gamma(x)$, dass $\mu(w) > \mu(x)$. Jede Minimalstelle von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ gehört zum Rand von $\mathscr{B}_\gamma(x)$. Es gelten auch die gleichartigen Aussagen bezüglich des Kegels $\mathscr{C}_\gamma(x)$.*

Beweis. Für $w \in \mathscr{B}_\gamma(x)$ gilt $\|Hx - w\|_2 < \big\|Hx - \mu(x)x\big\|_2$. Damit wird $\mu(w) > \mu(x)$ analog zum Beweis von Lemma 2.10 gezeigt.

Gäbe es eine Minimalstelle von $\mu(\cdot)$ im echten Innern von $\mathscr{B}_\gamma(x)$, dann wäre diese eine lokale Minimalstelle von $\mu(\cdot)$ und würde zum Eigenraum des kleinsten Eigenwertes μ_m gehören, das heißt, das Minimum wäre μ_m. Nach dem ersten Teil des Beweises sowie Voraussetzung 4.1 gilt aber $\mu(w) > \mu(x) > \mu_i \ge \mu_m$ für jedes $w \in \mathscr{B}_\gamma(x)$, so dass ein Widerspruch entsteht.

Für $w \in \mathscr{C}_\gamma(x)$ gilt wegen $0 < \angle_2(Hx, w) \le \arcsin\big(\gamma \sin\angle_2(Hx, x)\big) < \angle_2(Hx, x) < \pi/2$, dass $\cos\angle_2(Hx, x) < \cos\angle_2(Hx, w)$. Damit erhalten wir

$$\frac{(Hx, x)_2}{\|Hx\|_2\|x\|_2} < \frac{(Hx, w)_2}{\|Hx\|_2\|w\|_2} = \frac{(x, w)_H}{\|Hx\|_2\|w\|_2} \le \frac{\|x\|_H\|w\|_H}{\|Hx\|_2\|w\|_2}.$$

Da die linke Seite identisch zu $\|x\|_H^2/(\|Hx\|_2\|x\|_2)$ ist, folgt aus dieser Ungleichungskette, dass $(\|x\|_H/\|x\|_2) < (\|w\|_H/\|w\|_2)$ und weiter $\mu(w) > \mu(x)$ gilt. Analog wird per Kontraposition gezeigt,

dass jede Minimalstelle von $\mu(\cdot)$ auf $\mathscr{C}_\gamma(x)$ zum Rand von $\mathscr{C}_\gamma(x)$ gehört und das Minimum größer als $\mu(x)$ ist.

□

Damit brauchen wir im Folgenden statt der ganzen Vollkugel beziehungsweise des ganzen Vollkegels nur deren Oberflächen zu betrachten.

4.2.1 Klassische Konvergenzanalyse

Mittels einer Kurven-Differentiation können wir implizite Darstellungen der Minimalstellen von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ konstruieren.

Lemma 4.3. *Gemäß Voraussetzung 4.1 gibt es für eine beliebige Minimalstelle z von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ ein $\alpha \in \mathbb{R}$ mit $\alpha \geq -\mu(z)$ und*

$$(H + \alpha I)z = \big(\mu(z) + \alpha\big)Hx. \tag{4.5}$$

z ist genau dann ein Eigenvektor, wenn $\alpha = -\mu(z)$. Falls z kein Eigenvektor ist, dann gilt $Hx - z \perp_2 z$, so dass z ein Randvektor des Kegels $\mathscr{C}_\gamma(x)$ ist.

Beweis. Nach Lemma 4.2 gehört z zum Rand $\partial\mathscr{B}_\gamma(x)$, außerdem ist z eine Minimalstelle von $\mu(\cdot)$ auf $\partial\mathscr{B}_\gamma(x)$. Weiter betrachten wir eine beliebige durch z laufende stetig differenzierbare Kurve $\{g(t)\ ;\ t \in T \subseteq \mathbb{R}\} \subset \partial\mathscr{B}_\gamma(x)$ mit $g(t_g) = z$, dann ist u auch eine Minimalstelle von $\mu(\cdot)$ auf der Kurve, so dass t_g eine Minimalstelle der Funktion $\mu \circ g$ ist und die notwendige Bedingung $\big(d\,(\mu \circ g)/dt\big)(t_g) = 0$ gilt. Daraus folgt mit $\big(d\,(\mu \circ g)/dt\big)(t) = \big(\nabla\mu\big(g(t)\big), \dot{g}(t)\big)_2$, wobei $\dot{g}(t)$ die Ableitung $(dg/dt)(t)$ bezeichnet, dass

$$0 = \big(\nabla\mu\big(g(t_g)\big), \dot{g}(t_g)\big)_2 = \big(\nabla\mu(z), \dot{g}(t_g)\big)_2.$$

Damit ist der Vektor $\nabla\mu(z)$ Euklidisch orthogonal zum Tangentialvektor $\dot{g}(t_g)$. Analog zum Beweis von Lemma 2.13 wird gezeigt, dass der Vektor $\nabla\mu(z)$ Euklidisch orthogonal zur durch solche Tangentialvektoren aufgespannten Tangentialhyperebene $\dot{z}$ liegt.

Auf $\partial\mathscr{B}_\gamma(x)$ ist $f(w) := \|Hx - w\|_2^2$ eine konstante Funktion. Mit einer ähnlichen Kurven-Differentiation wird dann gezeigt, dass der Vektor $Hx - z$ ebenfalls Euklidisch orthogonal $\dot{z}$ liegt. Wegen $\gamma > 0$ und $\big\|Hx - \mu(x)x\big\|_2 > 0$ gilt $\|Hx - z\| > 0$, so dass $Hx - z$ kein Nullvektor ist und ein $\tilde{\alpha} \in \mathbb{R}$ mit

$$\nabla\mu(z) = \tilde{\alpha}(Hx - z) \tag{4.6}$$

existiert. Mit den Substitutionen $\nabla\mu(z) = \dfrac{2}{z^T z}\big(Hz - \mu(z)z\big)$ und $\alpha = \dfrac{\tilde{\alpha}}{2}z^T z - \mu(z)$ lässt sich (4.6) in (4.5) umschreiben.

Zum Beweis von $\alpha \geq -\mu(z)$ genügt es zu zeigen, dass $\tilde{\alpha} \geq 0$. Für hinreichend kleines $\varepsilon \in \mathbb{R}^+$ ist $z + \varepsilon(Hx - z)$ ein innerer Punkt der Kugel $\mathscr{B}_\gamma(x)$, so dass

$$\mu\big(z + \varepsilon(Hx - z)\big) \geq \mu(z)$$

nach Lemma 4.2 gilt. Damit ist die Richtungsableitung von $\mu(\cdot)$ an z bezüglich der Richtung $Hx - z$ nicht negativ, das heißt

$$(Hx - z)^T \nabla\mu(z) \geq 0.$$

Multiplizieren wir (4.6) mit $(Hx - z)^T$, ergibt sich

$$\tilde{\alpha}(Hx - z)^T(Hx - z) = (Hx - z)^T \nabla\mu(z) \geq 0.$$

Das zeigt $\tilde{\alpha} \geq 0$, denn $(Hx-z)^T(Hx-z) = \|Hx-z\|_2^2 > 0$.
z ist genau dann ein Eigenvektor, wenn der Gradient $\nabla\mu(z)$ Nullvektor ist. Nach (4.6) gilt

$$\nabla\mu(z) = \mathbf{0} \quad \Leftrightarrow \quad \tilde{\alpha} = 0 \quad \Leftrightarrow \quad \alpha = -\mu(z).$$

Damit ist $\alpha = -\mu(z)$ äquivalent dazu, dass z ein Eigenvektor ist.
Weiter gilt nach (4.5), dass $\big(\mu(z)+\alpha\big)(Hx-z,z)_2 = \big(Hz-\mu(z)z,z\big)_2 = 0$. Falls z kein Eigenvektor ist, dann gilt $\big(\mu(z)+\alpha\big) \neq 0$, so dass $(Hx-z,z)_2 = 0$ gezeigt wird. Damit gilt die Orthogonalität $Hx - z \perp_2 z$ und z ist die Euklidische Projektion von Hx auf $\operatorname{span}\{z\}$, so dass

$$\sin\angle_2\big(Hx,z\big) = \frac{\|Hx-z\|_2}{\|Hx\|_2} \underset{\Longleftarrow}{^{z\in\partial\mathscr{B}_\gamma(x)}} \frac{\gamma\|Hx-\mu(x)x\|_2}{\|Hx\|_2} = \gamma\sin\angle_2\big(Hx,x\big),$$

das heißt, z gehört zum Rand des Kegels $\mathscr{C}_\gamma(x)$. □

Nach Lemma 4.3 entspricht $\alpha = -\mu(z)$ einem trivialen Fall, denn $\mu(z)$ ist dann ein Eigenwert und erfüllt nach Lemma 4.2 die Bedingung $\mu(z) > \mu(x) > \mu_{i+1}$, so dass $\mu(z) \in \{\mu_1,\ldots,\mu_i\}$ und anschließend $\mu(x') \geq \mu(z) \geq \mu_i$ gilt, das heißt, der Delta-Quotient $\Delta\big(\mu_i,\mu_{i+1},\mu(x')\big)/\Delta\big(\mu_i,\mu_{i+1},\mu(x)\big)$ wird nicht positiv. Für manches spezielle x wird ein solcher trivialer Fall aber ausgeschlossen.

Lemma 4.4. *Gemäß Voraussetzung 4.1 betrachten wir den Fall, dass x genau auf den Eigenräumen $\mathcal{Y}_i$, $\mathcal{Y}_{i+1}$ Nichtnull-Anteile besitzt, also $x = y_i + y_{i+1}$ mit $y_i \in \mathcal{Y}_i\backslash\{\mathbf{0}\}$ und $y_{i+1} \in \mathcal{Y}_{i+1}\backslash\{\mathbf{0}\}$. Dann ist das Minimum von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ kleiner als μ_i und jede Minimalstelle z von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ gehört zum Unterraum $\operatorname{span}\{y_i,y_{i+1}\}$. Es gilt anschließend*

$$0 < \frac{\Delta\big(\mu_i,\mu_{i+1},\mu(z)\big)}{\Delta\big(\mu_i,\mu_{i+1},\mu(x)\big)} \leq \left((1-\gamma)\frac{\mu_{i+1}}{\mu_i}+\gamma\right)^2.$$

Die obere Schranke lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Der Vekotr $Hx = \mu_i y_i + \mu_{i+1}y_{i+1}$ hat wie x zwei Nichtnull-Anteile. Nach der Definition von $\mu(\cdot)$ liegen $\mu(x)$ und $\mu(Hx)$ im Intervall (μ_{i+1},μ_i). Wegen $Hx \in \mathscr{B}_\gamma(x)$ ist das Minimum von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ also kleiner als μ_i. Außerdem ist das Minimum nach Lemma 4.2 größer als $\mu(x)$ beziehungsweise μ_{i+1}. Damit ist das Minimum kein Eigenwert und jede Minimalstelle z kein Eigenvektor. Nach Lemma 4.3 lässt sich ein solches z durch (4.5) mit $\alpha > -\mu(z)$ darstellen. Weiter benutzen wir eine Darstellung $z = \sum_{j=1}^m Y_j b_j$ mit Euklidisch orthonormalen Basismatrizen Y_j der Eigenräume von H. Zusammen mit der gleichartigen Darstellung $x = Y_i c_i + Y_{i+1}c_{i+1}$ erhalten wir gemäß (4.5) die Gleichung

$$\sum_{j=1}^m (\mu_j+\alpha)Y_j b_j = \big(\mu(z)+\alpha\big)\big(\mu_i Y_i c_i + \mu_{i+1}Y_{i+1}c_{i+1}\big).$$

Wegen der Euklidischen Orthogonalitäten zwischen den Eigenräumen liefern Multiplikationen der obigen Gleichung mit $Y_k^T,\ k = 1,\ldots,m$ folgende Gleichungen

$$(\mu_k+\alpha)b_k \;=\; \begin{cases} \mathbf{0} & \text{für} \quad k \in \{1,\ldots,m\}\backslash\{i,\,i+1\}, \\ \big(\mu(z)+\alpha\big)\mu_k c_k & \text{für} \quad k \in \{i,\,i+1\}. \end{cases} \tag{4.7}$$

Da die Eigenwerte μ_k paarweise verschieden sind, gibt es höchstens einen Index k mit $\mu_k + \alpha = 0$. Damit gibt es höchstens einen Nichtnullvektor unter den Koeffizientenvektoren b_k mit $k \in \{1,\ldots,m\}\backslash\{i,\,i{+}1\}$. Dass z einen solchen Nichtnull-Koeffizientenvektor b_k hat, führt aber auf einen Widerspruch zu $\mu(z) > \mu(x)$: Wegen $b_k \neq \mathbf{0}$ müsste $\mu_k + \alpha = 0$ und damit $\mu_k = -\alpha < \mu(z) < \mu_i$

gelten. Wegen $k \neq i{+}1$ würde weiter $\mu_k < \mu_{i+1}$ gezeigt. Mit $\alpha = -\mu_k$ und (4.7) lassen sich die Koeffizientenvektoren b_i, b_{i+1} durch

$$b_i = \frac{\big(\mu(z) - \mu_k\big)\mu_i}{\mu_i - \mu_k} c_i, \quad b_{i+1} = \frac{\big(\mu(z) - \mu_k\big)\mu_{i+1}}{\mu_{i+1} - \mu_k} c_{i+1} \tag{4.8}$$

darstellen. Damit sind b_i, b_{i+1} wie c_i, c_{i+1} keine Nullvektoren. Außerdem gilt wegen

$$\left(\frac{\|b_{i+1}\|_2^2}{\|b_i\|_2^2}\right)\left(\frac{\|c_{i+1}\|_2^2}{\|c_i\|_2}\right)^{-1} = \left(\frac{\|c_i\|_2}{\|b_i\|_2}\,\frac{\|b_{i+1}\|_2}{\|c_{i+1}\|_2}\right)^2 = \left(\frac{(\mu_i - \mu_k)\mu_{i+1}}{(\mu_{i+1} - \mu_k)\mu_i}\right)^2 > 1,$$

dass $\|b_{i+1}\|_2^2/\|b_i\|_2^2 > \|c_{i+1}\|_2^2/\|c_i\|_2^2$. Daraus folgt $\mu(x) = \mu(Y_i c_i + Y_{i+1} c_{i+1}) > \mu(Y_i b_i, Y_{i+1} b_{i+1}) = \mu(z - Y_k b_k) > \mu(z)$ mittels der Definition des Rayleigh-Quotienten $\mu(\cdot)$. Dieser Widerspruch zu $\mu(z) > \mu(x)$ zeigt, dass z zu $\operatorname{span}\{Y_i b_i, Y_{i+1} b_{i+1}\} = \operatorname{span}\{Y_i c_i, Y_{i+1} c_{i+1}\} = \operatorname{span}\{y_i, y_{i+1}\}$ gehört. Damit gilt

$$\frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(z)\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)} = \left(\frac{\mu_i - \mu(z)}{\mu(z) - \mu_{i+1}}\right)\left(\frac{\mu_i - \mu(x)}{\mu(x) - \mu_{i+1}}\right)^{-1} = \left(\frac{\|b_{i+1}\|_2^2}{\|b_i\|_2^2}\right)\left(\frac{\|c_{i+1}\|_2^2}{\|c_i\|_2}\right)^{-1},$$

vergleiche Lemma 2.6. Anders als bei der widerspruchsführenden Annahme (4.8) werden b_i, b_{i+1} durch

$$b_i = \frac{\big(\mu(z) + \alpha\big)\mu_i}{\mu_i + \alpha} c_i, \quad b_{i+1} = \frac{\big(\mu(z) + \alpha\big)\mu_{i+1}}{\mu_{i+1} + \alpha} c_{i+1} \tag{4.9}$$

dargestellt, so dass

$$\frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(z)\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)} = \left(\frac{(\mu_i + \alpha)\mu_{i+1}}{(\mu_{i+1} + \alpha)\mu_i}\right)^2.$$

Daher können wir den Delta-Quotienten als das Quadrat der Funktion $f(\alpha) := \frac{(\mu_i+\alpha)\mu_{i+1}}{(\mu_{i+1}+\alpha)\mu_i}$ untersuchen. $f(\alpha)$ ist auf den Intervallen $(-\infty, -\mu_{i+1})$ und $(-\mu_{i+1}, \infty)$ stetig und monoton fallend. Weiter untersuchen wir den Wertbereich von α.

α lässt sich durch eine quadratische Gleichung implizit darstellen, wobei die Koeffizienten von μ_i, μ_{i+1}, $\mu(z)$ sowie γ abhängen. Zur Herleitung dieser Gleichung benutzen wir einige geometrische Eigenschaften von z. Da z kein Eigenvektor ist, gilt nach Lemma 4.4 die Orthogonalität $Hx - z \perp_2 z$. Zusammen mit der Definition (4.3) der Kugel $\mathscr{B}_\gamma(x)$ wird gezeigt, dass

$$\|z\|_2^2 = \|Hx\|_2^2 - \|Hx - z\|_2^2 = \|Hx\|_2^2 - \gamma^2 \|Hx - \mu(x)x\|_2^2. \tag{4.10}$$

Wenn wir (4.9) mit (4.10) direkt kombinieren, würde die resultierende Darstellung von c_i, c_{i+1} abhängen. Daher nutzen wir eine weitere Gleichung

$$(Hx, z)_2 = (Hx - z + z, z)_2 = (z, z)_2 = \|z\|_2^2$$

und erhalten

$$\frac{(Hx, z)_2^2}{\|z\|_2^2} = \|z\|_2^2 = \|Hx\|_2^2 - \gamma^2 \|Hx - \mu(x)x\|_2^2. \tag{4.11}$$

Einsetzen von $Hx = \mu_i Y_i c_i + \mu_{i+1} Y_{i+1} c_{i+1}$, $z = Y_i b_i + Y_{i+1} b_{i+1}$ sowie (4.9) in (4.11) liefert mittels der Substitution $\|c_{i+1}\|_2^2/\|c_i\|_2 = \big(\mu_i - \mu(x)\big)\big/\big(\mu(x) - \mu_{i+1}\big)$ eine quadratische Gleichung

$$a_2 \alpha^2 + a_1 \alpha + a_0 = 0 \qquad \text{mit}$$

$$a_2 = \big(\mu_i \mu(x) + \mu_{i+1}\mu(x) - \mu_i \mu_{i+1}\big)\gamma^2 > 0, \quad a_1 = 2\mu_i \mu_{i+1} \mu(x) \gamma^2 > 0, \quad a_0 = -\mu_i^2 \mu_{i+1}^2 (1 - \gamma^2) < 0.$$

Betrachten wir dann die Wurzeln

$$\alpha_1 = \frac{-a_1 + \sqrt{a_1^2 - 4a_2a_0}}{2a_2}, \quad \alpha_2 = \frac{-a_1 - \sqrt{a_1^2 - 4a_2a_0}}{2a_2}.$$

Wir zeigen zunächst, dass α_1 eine monoton fallende Funktion von $\mu(x)$ ist. Für feste μ_i, μ_{i+1} und γ sind a_2 und a_1 offensichtlich monoton steigende Funktionen von $\mu(x)$, und a_0 bleibt konstant. Damit kann man α_1 als eine Funktion mit den Variablen a_2 und a_1 untersuchen. Die partiellen Ableitungen haben stets negative Werte für $a_2 > 0$, $a_1 > 0$ und $a_0 < 0$:

$$\frac{\partial \alpha_1}{\partial a_2} = \frac{a_1\sqrt{a_1^2 - 4a_2a_0} - (a_1^2 - 2a_2a_0)}{2a_2^2\sqrt{a_1^2 - 4a_2a_0}} < 0, \qquad \frac{\partial \alpha_1}{\partial a_1} = \frac{a_1 - \sqrt{a_1^2 - 4a_2a_0}}{2a_2\sqrt{a_1^2 - 4a_2a_0}} < 0.$$

Nach Kettenregel ist dann die Ableitung von α_1 nach $\mu(x)$ stets negativ. Damit ist α_1 monoton fallend und $-\alpha_1$ monoton steigend bezüglich $\mu(x)$. Außerdem ist der Koeffizientenquotient $-a_1/a_2$ nach Betrachtung der Ableitung eine monoton steigende Funktion von $\mu(x)$, so dass $\alpha_2 = -a_1/a_2 - \alpha_1$ ebenfalls bezüglich $\mu(x)$ monoton steigend ist. Gemäß den Monotonien von α_1, α_2 und der Voraussetzung $\mu(x) \in (\mu_{i+1}, \mu_i)$ gilt

$$\frac{1-\gamma}{\gamma}\mu_{i+1} < \alpha_1 < \frac{1-\gamma}{\gamma}\mu_i, \quad -\frac{1+\gamma}{\gamma}\mu_i < \alpha_2 < -\frac{1+\gamma}{\gamma}\mu_{i+1},$$

Damit befinden sich α_1 und α_2 jeweils in den zwei Definitionsintervallen $(-\mu_{i+1}, \infty)$ und $(-\infty, -\mu_{i+1})$ von f und es gilt nach der Monotonie von f, dass

$$\beta_{11} := \frac{\mu_{i+1}}{(1-\gamma)\mu_i + \gamma\mu_{i+1}} < f(\alpha_1) < \frac{(1-\gamma)\mu_{i+1} + \gamma\mu_i}{\mu_i} =: \beta_{12},$$

$$\beta_{21} := \frac{(1+\gamma)\mu_{i+1} - \gamma\mu_i}{\mu_i} < f(\alpha_2) < \frac{\mu_{i+1}}{(1+\gamma)\mu_i - \gamma\mu_{i+1}} =: \beta_{22}.$$

Daher sind $\big(f(\alpha_1)\big)^2$ und $\big(f(\alpha_2)\big)^2$ beide durch

$$\max\{\beta_{11}^2, \beta_{12}^2, \beta_{21}^2, \beta_{22}^2\} = \beta_{12}^2 = \left((1-\gamma)\frac{\mu_{i+1}}{\mu_i} + \gamma\right)^2$$

beschränkt und eine obere Schranke des Delta-Quotienten wird damit erhalten. Der Delta-Quotient ist positiv, denn $\mu(x)$, $\mu(z)$ gehören beide zum Intervall (μ_{i+1}, μ_i). Es gilt außerdem

$$\lim_{\mu(x)\to\mu_i} f(\alpha_1) = f\left(\frac{1-\gamma}{\gamma}\mu_{i+1}\right) = \beta_{12},$$

so dass sich die obere Schranke β_{12}^2 im Grenzfall $\mu(x) \to \mu_i$ erreichen lässt. □

Zur Verallgemeinerung dieser niedrigdimensionalen Abschätzung vergleichen wir den in Voraussetzung 4.1 gegebenen Vektor x mit anderen möglichen Vektoren, welche den gleichen Rayleigh-Quotienten wie x haben. Zu nutzen ist der Durchschnitt einer Rayleigh-Quotient-Niveaumenge mit der Einheitskugel der Euklidischen Vektornorm.

Lemma 4.5. *Gelte für $\bar{x} \in \bar{\mathcal{N}}(\tilde{\mu}) := \{w \in \mathbb{R}^n \ ; \ \mu(w) = \tilde{\mu}, \ \|w\|_2 = 1\}$, $\tilde{\mu} \in (\mu_{i+1}, \mu_i)$ gemäß Voraussetzung 4.1, dass*

$$\|H\bar{x}\|_2 = \min_{w\in\bar{\mathcal{N}}(\tilde{\mu})} \|Hw\|_2.$$

Dann besitzt $\bar{x}$ genau auf den Eigenräumen $\mathcal{Y}_i$ und $\mathcal{Y}_{i+1}$ Nichtnull-Anteile.

Beweis. Betrachten wir hier ein Minimierungsproblem von $\|Hw\|_2^2$ $\left(= w^T H^2 w\right)$ für $w \in \bar{\mathcal{N}}(\tilde{\mu})$. Die Einschränkungsbedingungen für $\bar{\mathcal{N}}(\tilde{\mu})$ lassen sich als $w^T w = 1$, $w^T H w = \tilde{\mu}$ umschreiben. Mit der Methode der Lagrange-Multiplikatoren lässt sich eine notwendige Bedingung für Minimalstelle formulieren. Der Gradient der Lagrange-Funktion

$$L(w, \alpha, \beta) := w^T H^2 w + \alpha(w^T w - 1) + \beta(w^T H w - \tilde{\mu})$$

mit Lagrange-Multiplikatoren $\alpha, \beta \in \mathbb{R}$ lautet

$$\nabla L(w, \alpha, \beta) = \begin{bmatrix} 2H^2 w + 2\alpha w + 2\beta H w \\ 0 \\ 0 \end{bmatrix},$$

so dass für eine beliebige Minimalstelle $\bar{x}$

$$2H^2 \bar{x} + 2\alpha \bar{x} + 2\beta H \bar{x} = \mathbf{0}$$

gilt. Einsetzen einer Basis-Darstellung $\bar{x} = \sum_{j=1}^m Y_j c_j$ mit Euklidisch orthonormalen Basismatrizen Y_j der Eigenräume $\mathcal{Y}_j$ liefert $\sum_{j=1}^m 2\mu_j^2 Y_j c_j + 2\alpha Y_j c_j + 2\beta \mu_j Y_j c_j = \mathbf{0}$. Durch Multiplikationen mit Y_k^T erhalten wir weiter

$$(2\mu_k^2 + 2\alpha + 2\beta\mu_k) c_k = \mathbf{0}, \quad k = 1, \ldots, m.$$

Da μ_k paarweise verschieden sind, gibt es höchstens 2 Indizes mit $2\mu_k^2 + 2\alpha + 2\beta\mu_k = 0$, so dass $\bar{x}$ höchstens zwei Nichtnull-Koeffizientenvektoren besitzt. Wegen $\|\bar{x}\|_2 = 1$ und $\mu(\bar{x}) = \tilde{\mu} \in (\mu_{i+1}, \mu_i)$ ist $\bar{x}$ kein Nullvektor und kein Eigenvektor, so dass $\bar{x}$ genau zwei Nichtnull-Koeffizientenvektoren besitzt. Seien $\mu_{\sigma(1)} > \mu_{\sigma(2)}$ die entsprechenden Eigenwerte, dann gilt $\mu_{\sigma(1)} \geq \mu_i > \tilde{\mu} > \mu_{i+1} \geq \mu_{\sigma(2)}$. Mit den Einschränkungsbedingungen lassen sich $\|c_{\sigma(1)}\|_2^2$ und $\|c_{\sigma(2)}\|_2^2$ durch

$$\|c_{\sigma(1)}\|_2^2 = \frac{\tilde{\mu} - \mu_{\sigma(2)}}{\mu_{\sigma(1)} - \mu_{\sigma(2)}}, \qquad \|c_{\sigma(2)}\|_2^2 = \frac{\mu_{\sigma(1)} - \tilde{\mu}}{\mu_{\sigma(1)} - \mu_{\sigma(2)}}$$

darstellen. Weiter erhalten wir

$$\|H\bar{x}\|_2^2 = \mu_{\sigma(1)}^2 \|c_{\sigma(1)}\|_2^2 + \mu_{\sigma(2)}^2 \|c_{\sigma(2)}\|_2^2 = \tilde{\mu}\mu_{\sigma(1)} + \tilde{\mu}\mu_{\sigma(2)} - \mu_{\sigma(1)}\mu_{\sigma(2)} \geq \tilde{\mu}\mu_i + \tilde{\mu}\mu_{i+1} - \mu_i\mu_{i+1}.$$

Für $\sigma(1) = i$, $\sigma(2) = i{+}1$ wird diese Ungleichung eine Gleichung, so dass die Minimalstelle $\bar{x}$ nur die Nichtnull-Koeffizientenvektoren c_i, c_{i+1} besitzt. □

Zusammen mit $\|Hw\|_2$ können zwei weitere Maße durch $\bar{x}$ minimiert werden.

Lemma 4.6. *Unter Voraussetzung von Lemma 4.5 ist $\bar{x}$ auch eine Minimalstelle von $\angle_2(w, Hw)$ und $\|Hw - \tilde{\mu}w\|_2$ auf $w \in \bar{\mathcal{N}}(\tilde{\mu})$.*

Beweis. Für beliebiges $w \in \bar{\mathcal{N}}(\tilde{\mu})$ gilt

$$\cos \angle_2(w, Hw) = \frac{(w, Hw)_2}{\|w\|_2 \|Hw\|_2} = \frac{\mu(w)\|w\|_2^2}{\|w\|_2 \|Hw\|_2} = \frac{\tilde{\mu}}{\|Hw\|_2} \, (> 0),$$

$$Hw - \mu(w)w \perp w \quad \Rightarrow \quad \|Hw - \tilde{\mu}w\|_2^2 = \|Hw\|_2^2 - \|\tilde{\mu}w\|_2^2 = \|Hw\|_2^2 - \tilde{\mu}^2.$$

Damit sind

$$\angle_2(w, Hw) = \arccos\left(\frac{\tilde{\mu}}{\|Hw\|_2}\right) \quad \text{und} \quad \|Hw - \tilde{\mu}w\|_2 = \sqrt{\|Hw\|_2^2 - \tilde{\mu}^2}$$

monoton steigend bezüglich $\|Hw\|_2$, so dass $\bar{x}$ auch für diese Maße eine Minimalstelle ist. □

Zur Konstruktion eines Zusammenhangs zwischen einem niedrigdimensionalen Fall und einem allgemeinen Fall betrachten wir weiter eine Minimalstelle $\bar{x}$ von den Maßen aus Lemma 4.5 und Lemma 4.6 für $\tilde{\mu} = \mu(x)$. Wir können normierte Vektoren y_i, y_{i+1} der Nichtnull-Anteile $Y_i c_i$, $Y_{i+1} c_{i+1}$ konstruieren, so dass $\bar{x}$ sowie $H\bar{x}$ positive Koordinaten bezüglich y_i, y_{i+1} haben. Lemma 4.4 angewandt auf $\bar{x}$ zeigt dann, dass jede Minimalstelle $\bar{z}$ von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\bar{x})$ auch zu $\operatorname{span}\{y_i, y_{i+1}\}$ gehört. Lemma 4.3 zeigt weiter, dass $\bar{z}$ ein Randvektor des Kegels $\mathscr{C}_\gamma(\bar{x})$ beziehungsweise des Ebenensegments $\mathscr{C}_\gamma(\bar{x}) \cap \operatorname{span}\{y_i, y_{i+1}\}$ ist. Es gibt also insgesamt zwei mögliche Richtungen für $\bar{z}$, und der Euklidische Winkel zwischen $\bar{z}$ und y_i ist entweder $\varphi_1 := \angle_2(\bar{z}, H\bar{x}) + \angle_2(H\bar{x}, y_i)$ oder $\varphi_2 := |\angle_2(\bar{z}, H\bar{x}) - \angle_2(H\bar{x}, y_i)|$. Es gilt dann $0 < \varphi_2 < \varphi_1 < \angle_2(\bar{x}, y_i) < \pi/2$ gemäß der Definition von $\mathscr{C}_\gamma(\bar{x})$, so dass $\bar{z}$ mit φ_1 echt kleineren Rayleigh-Quotienten hat und die Minimalstelle ist. Dann hat $\bar{z}$ wie $\bar{x}$ und $H\bar{x}$ auch positive Koordinaten bezüglich y_i, y_{i+1}. Betrachten wir dann die Kurve $\{\bar{g}(t) \,;\, t \in (0, \pi/2)\}$ mit

$$\bar{g}(t) := \sin(t) y_i + \cos(t) y_{i+1},$$

die durch $\bar{x} = \bar{g}(\bar{t}_1)$ und $\bar{z}/\|\bar{z}\|_2 = \bar{g}(\bar{t}_2)$ und auf der Einheitskugel der Euklidischen Vektornorm läuft. Für jedes $t \in (0, \pi/2)$ lässt sich der Rayleigh-Quotient des entsprechenden Vektors durch Eigenwerte und Koordinaten darstellen:

$$\mu\big(\bar{g}(t)\big) = \frac{\mu_i \sin^2(t) + \mu_{i+1} \cos^2(t)}{\sin^2(t) + \cos^2(t)} = \mu_i \sin^2(t) + \mu_{i+1} \cos^2(t).$$

$\mu \circ \bar{g}$ ist monoton steigend auf $t \in (0, \pi/2)$, denn $\big(d(\mu \circ \bar{g})/dt\big)(t) = 2 \sin(t) \cos(t) (\mu_i - \mu_{i+1}) > 0$. Damit lässt sich $\bar{t}_2 > \bar{t}_1$ zeigen:

$$\mu(\bar{z}/\|\bar{z}\|_2) = \mu(\bar{z}) > \mu(\bar{x}) \quad \Rightarrow \quad \mu\big(\bar{g}(\bar{t}_2)\big) > \mu\big(\bar{g}(\bar{t}_1)\big) \quad \Rightarrow \quad \bar{t}_2 > \bar{t}_1.$$

Die Differenz $\bar{t}_2 - \bar{t}_1$ ist genau die Winkelgröße von $\angle_2(\bar{x}, \bar{z}/\|\bar{z}\|_2)$ beziehungsweise $\angle_2(\bar{x}, \bar{z})$, denn $\bar{t}_1$ und $\bar{t}_2$ sind jeweils die Winkelgrößen von $\angle_2(\bar{x}, y_{i+1})$ und $\angle_2(y_{i+1}, \bar{z})$. Alternativ kann man die Gleichheit $\angle_2(\bar{x}, \bar{z}) = \bar{t}_2 - \bar{t}_1$ mittels der Tangentialvektoren

$$\frac{d\bar{g}}{dt}(t) = \frac{d \sin}{dt}(t) y_i + \frac{d \cos}{dt}(t) y_{i+1} = \cos(t) y_i - \sin(t) y_{i+1}$$

an den Punkten des Kreisbogens $\{\bar{g}(t) \,;\, t \in [\,\bar{t}_1, \bar{t}_2\,]\}$ zeigen. Integration dieser bereits Euklidisch normierten Tangentialvektoren über $[\,\bar{t}_1, \bar{t}_2\,]$ liefert dann die Größe eines Zwischenwinkels:

$$\angle_2(\bar{x}, \bar{z}) = \angle_2(\bar{x}, \bar{z}/\|\bar{z}\|_2) = \angle_2\big(\bar{g}(\bar{t}_1), \bar{g}(\bar{t}_2)\big) = \int_{\bar{t}_1}^{\bar{t}_2} \left\| \frac{d\bar{g}}{dt}(t) \right\|_2 dt = \int_{\bar{t}_1}^{\bar{t}_2} 1 \; dt = \bar{t}_2 - \bar{t}_1. \tag{4.12}$$

Als eine Teilkurve verbindet $\{\bar{g}(t) \,;\, t \in [\,\bar{t}_1, \bar{t}_2\,]\}$ die Niveaukurven $\bar{\mathcal{N}}\big(\mu(x)\big)$ und $\bar{\mathcal{N}}\big(\mu(\bar{z})\big)$, und die Kurvenlänge ist nach Eigenschaft des Kreisbogens genau die Winkelgröße $\bar{t}_2 - \bar{t}_1$. Interessanterweise lassen sich die Tangentialvektoren $(d\bar{g}/dt)(t)$ als normierte Vektoren der Residuumkurve $\{\bar{r}(t) \,;\, t \in [\,\bar{t}_1, \bar{t}_2\,]\}$ mit

$$\bar{r}(t) := H\bar{g}(t) - \mu\big(\bar{g}(t)\big)\bar{g}(t)$$

betrachten, denn einfaches Einsetzen liefert $\bar{r}(t) = \sin(t) \cos(t) (\mu_i - \mu_{i+1}) (\cos(t) y_i - \sin(t) y_{i+1})$.

Betrachten wir ein weiteres $\tilde{x} \in \bar{\mathcal{N}}\big(\mu(x)\big)$ und eine Minimalstelle $\tilde{z}$ des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\tilde{x})$ in einem nichttrivialen Fall $\mu(\tilde{z}) < \mu_i$. Dafür können wir ähnliche Verbindungskurven auf der Einheitskugel der Euklidischen Vektornorm konstruieren. Mit $w_0 := \tilde{z}/\|\tilde{z}\|_2$ als Anfangsvektor kann man nach Anregung von $(d\bar{g}/dt)(t) = \bar{r}(t)/\|\bar{r}(t)\|_2$ ein Gradientenverfahren formulieren:

$$w_{l+1} = w_l - \varepsilon r_l/\|r_l\|_2 \quad \text{mit} \quad r_l = Hw_l - \mu(w_l) w_l, \quad \varepsilon > 0.$$

Im Grenzfall $\varepsilon \to 0$ bilden die umgekehrt geordneten Iterierten eine Kurve $\{\tilde{g}(t) \,;\, t \in T \subset \mathbb{R}\}$, die $\tilde{z}/\|\tilde{z}\|_2$ enthält und die Bedingung $\|\tilde{g}(t)\|_2 = 1$ erfüllt. Grenzwertbestimmung liefert dann

$$\frac{d\tilde{g}}{dt}(t) = \frac{\tilde{r}(t)}{\|\tilde{r}(t)\|_2} \quad \text{mit} \quad \tilde{r}(t) := H\tilde{g}(t) - \mu\big(\tilde{g}(t)\big)\tilde{g}(t).$$

Dann ist $\mu \circ \tilde{g}$ wegen

$$\frac{d(\mu \circ \tilde{g})}{dt}(t) = \Big(\nabla\mu\big(\tilde{g}(t)\big), \frac{d\tilde{g}}{dt}(t)\Big)_2 = \Big(\frac{2}{\|\tilde{g}(t)\|_2^2}\tilde{r}(t), \frac{\tilde{r}(t)}{\|\tilde{r}(t)\|_2}\Big)_2 = 2\|\tilde{r}(t)\|_2 \tag{4.13}$$

eine monoton steigende Funktion. Solange die Kurve nicht einen Eigenvektor erreicht, ist $\tilde{r}(t)$ kein Nullvektor, so dass $\mu \circ \tilde{g}$ sogar streng monoton steigend ist. Da $\mu(x)$ und $\mu(\tilde{z})$ beide im Intervall (μ_{i+1}, μ_i) liegen, besitzt das Intervall $\big(\mu(x), \mu(\tilde{z})\big)$ keinen Eigenwert, so dass $\mu\big(\tilde{g}(t)\big)$ den Wert $\mu(x)$ und die Kurve $\{\tilde{g}(t)\ ;\ t \in T \subset \mathbb{R}\}$ die Niveaukurve $\bar{\mathcal{N}}\big(\mu(x)\big)$ erreichen kann. Außerdem ist $\tilde{g}(t)$ im Fall $\mu(\tilde{g}(t)) \in (\mu_{i+1}, \mu_i)$ kein Eigenvektor, so dass $\mu \circ \tilde{g}$ lokal streng monoton steigend ist und eine Umkehrfunktion $\tilde{f}(\nu)$ von $(\mu \circ \tilde{g})(t)$ auf dem Intervall (μ_{i+1}, μ_i) definiert werden kann, welche nach Eigenschaft der Umkehrfunktionen ebenfalls streng monoton steigend ist. Mit $\tilde{t}_1 := \tilde{f}\big(\mu(x)\big)$ und $\tilde{t}_2 := \tilde{f}\big(\mu(\tilde{z})\big)$ gilt dann

$$\tilde{t}_1 < \tilde{t}_2, \qquad \tilde{g}(\tilde{t}_1) \in \bar{\mathcal{N}}\big(\mu(x)\big), \qquad \tilde{g}(\tilde{t}_2) = \tilde{z}/\|\tilde{z}\|_2.$$

Der Vektor $\tilde{g}(\tilde{t}_2)$ gehört dann nach Lemma 4.2 zum Rand des Kegels $\mathscr{C}_\gamma(\tilde{x})$ und ist ein echt innerer Vektor des Kegels $\mathscr{C}_1(\tilde{x})$. Analog zum Beweis von Lemma 4.2 wird gezeigt, dass das Minimum von $\mu(\cdot)$ auf $\mathscr{C}_1(\tilde{x})$ durch $\mu(\tilde{x}) = \mu(x)$ gegeben ist und jede Minimalstelle zum Rand dieses Kegels gehört. Wegen $\tilde{g}(\tilde{t}_1) \in \bar{\mathcal{N}}\big(\mu(x)\big)$ ist dann $\tilde{g}(\tilde{t}_1)$ kein echt innerer Vektor von $\mathscr{C}_1(\tilde{x})$, so dass ein $\tilde{t}_s \in [\tilde{t}_1, \tilde{t}_2)$ mit $\tilde{g}(\tilde{t}_s) \in \partial\mathscr{C}_1(\tilde{x})$ existiert. Weiter ist die Kurvenlänge zwischen $\tilde{g}(\tilde{t}_s)$ und $\tilde{g}(\tilde{t}_2)$ durch

$$L_{\tilde{g}}\big(\tilde{g}(\tilde{t}_s), \tilde{g}(\tilde{t}_2)\big) = \int_{\tilde{t}_s}^{\tilde{t}_2} \Big\|\frac{d\tilde{g}}{dt}(t)\Big\|_2 \, dt = \int_{\tilde{t}_s}^{\tilde{t}_2} 1 \, dt = \tilde{t}_2 - \tilde{t}_s$$

gegeben. Für alle Verbindungskurven zwischen $\tilde{g}(\tilde{t}_s)$ und $\tilde{g}(\tilde{t}_2)$ auf der Oberfläche der Einheitskugel ist das Minimum der Kurvenlängen identisch zur Winkelgröße $\angle_2\big(\tilde{g}(\tilde{t}_s), \tilde{g}(\tilde{t}_2)\big)$, so dass

$$\tilde{t}_2 - \tilde{t}_1 \geq \tilde{t}_2 - \tilde{t}_s = L_{\tilde{g}}\big(\tilde{g}(\tilde{t}_s), \tilde{g}(\tilde{t}_2)\big) \geq \angle_2\big(\tilde{g}(\tilde{t}_s), \tilde{g}(\tilde{t}_2)\big) \geq \min_{w_1 \in \partial\mathscr{C}_1(\tilde{x}), w_\gamma \in \partial\mathscr{C}_\gamma(\tilde{x})} \angle_2(w_1, w_\gamma).$$

Da $\mathscr{C}_1(\tilde{x})$ und $\mathscr{C}_\gamma(\tilde{x})$ dieselbe Achse $\{\delta H\tilde{x}\ ;\ \delta > 0\}$ haben, ist der obige minimale Winkel identisch zur Differenz der Öffnungswinkel, nämlich

$$\min_{w_1 \in \partial\mathscr{C}_1(\tilde{x}), w_\gamma \in \partial\mathscr{C}_\gamma(\tilde{x})} \angle_2(w_1, w_\gamma) = \tilde{\varphi} - \arcsin\big(\gamma \sin(\tilde{\varphi})\big) \quad \text{mit} \quad \tilde{\varphi} := \angle_2(\tilde{x}, H\tilde{x}).$$

Nach Lemma 4.5 und Lemma 4.6 gilt $\tilde{\varphi} \geq \bar{\varphi} := \angle_2(\bar{x}, H\bar{x})$. Da die Funktion

$$\vartheta(\varphi) := \varphi - \arcsin\big(\gamma \sin(\varphi)\big)$$

die Ableitung

$$\frac{d\vartheta}{d\varphi}(\varphi) = 1 - \frac{\gamma\cos(\varphi)}{\sqrt{(1-\gamma^2) + \big(\gamma\cos(\varphi)\big)^2}} > 0$$

hat, gilt dann $\vartheta(\tilde{\varphi}) \geq \vartheta(\bar{\varphi})$. Außerdem gehören der spezielle Vektor $\bar{x}$ und die entsprechende Minimalstelle $\bar{z}$ zu einem zweidimensionalen Unterraum $\{y_i, y_{i+1}\}$, so dass die Differenz der Kegel-Öffnungswinkel einfach durch $\angle_2(\bar{x}, \bar{z})$ gegeben ist. Nach obigen Betrachtungen sowie (4.12) erhalten wir

$$\tilde{t}_2 - \tilde{t}_1 \geq \vartheta(\tilde{\varphi}) \geq \vartheta(\bar{\varphi}) = \angle_2(\bar{x}, \bar{z}) = \bar{t}_2 - \bar{t}_1. \tag{4.14}$$

Betrachten wir weiter ein $\tilde{\mu} \in (\mu_{i+1}, \mu_i)$ und $\tilde{t}$, $\bar{t}$ mit $\mu\big(\tilde{g}(\tilde{t})\big) = \mu\big(\bar{g}(\bar{t})\big) = \tilde{\mu}$. Nach Lemma 4.5 und Lemma 4.6 gilt dann $\|\tilde{r}(\tilde{t})\|_2 \geq \|\bar{r}(\bar{t})\|_2$. Auf (μ_{i+1}, μ_i) lässt sich auch eine streng monoton steigende Umkehrfunktion $\bar{f}(\nu)$ von $(\mu \circ \bar{g})(t)$ definieren. Wegen (4.13) und der dazu analogen gleichartigen Gleichung für $\bar{g}$ und $\bar{r}$ gilt weiter

$$\Big(\frac{d\tilde{f}}{d\nu}(\tilde{\mu})\Big)^{-1} = \frac{d(\mu \circ \tilde{g})}{dt}(\tilde{t}) \overset{(4.13)}{\geq} \frac{d(\mu \circ \bar{g})}{dt}(\bar{t}) = \Big(\frac{d\bar{f}}{d\nu}(\tilde{\mu})\Big)^{-1} > 0 \quad \Rightarrow \quad \frac{d\tilde{f}}{d\nu}(\tilde{\mu}) \leq \frac{d\bar{f}}{d\nu}(\tilde{\mu}). \tag{4.15}$$

Unter der Annahme $\mu(\tilde{z}) < \mu(\bar{z})$ würde $\mu\big(\tilde{g}(\tilde{t}_2)\big) = \mu(\tilde{z}) < \mu(\bar{z}) = \mu\big(\bar{g}(\bar{t}_2)\big)$ gelten. Zusammen mit $\mu\big(\tilde{g}(\tilde{t}_1)\big) = \mu(\tilde{x}) = \mu(\bar{x}) = \mu\big(\bar{g}(\bar{t}_1)\big)$ und (4.15) wird ein Widerspruch

$$\tilde{t}_2 - \tilde{t}_1 = \int_{\mu(\tilde{g}(\tilde{t}_1))}^{\mu(\tilde{g}(\tilde{t}_2))} \frac{d\tilde{f}}{d\nu}(\nu)\, d\nu < \int_{\mu(\bar{g}(\bar{t}_1))}^{\mu(\bar{g}(\bar{t}_2))} \frac{d\bar{f}}{d\nu}(\nu)\, d\nu = t_2 - t_1$$

zu (4.14) erhalten. Das zeigt $\mu(\tilde{z}) \geq \mu(\bar{z})$. Formulieren wir diese Ergebnisse in einem Lemma.

Lemma 4.7. *Mit den Formulierungen von Lemma 4.5 und Lemma 4.6 sei ein $\tilde{x} \in \bar{\mathcal{N}}\big(\mu(x)\big)$ gegeben, wofür das Minimum $\mu(\tilde{z})$ des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\tilde{x})$ kleiner als μ_i ist. Dann gibt es ein $\bar{x} \in \bar{\mathcal{N}}\big(\mu(x)\big)$, welches genau auf den Eigenräumen $\mathcal{Y}_i$ und $\mathcal{Y}_{i+1}$ Nichtnull-Anteile besitzt, so dass $\mu(\tilde{z})$ nicht kleiner als das Minimum $\mu(\bar{z})$ des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\bar{x})$ ist.*

Kombination von Lemma 4.4 und Lemma 4.7 liefert eine Delta-Abschätzung für PINVIT(1).

Satz 4.8. *Gemäß Voraussetzung 4.1 gilt $\mu(x') > \mu(x)$. Im Fall $\mu(x') < \mu_i$ gilt*

$$0 < \frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(x')\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)} \leq \left((1-\gamma)\frac{\mu_{i+1}}{\mu_i} + \gamma\right)^2. \tag{4.16}$$

Die obere Schranke lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Wegen $x' \in \mathscr{C}_\gamma(x)$ gilt $\mu(x') > \mu(x)$ nach Lemma 4.2.
Für $\mu(x') < \mu_i$ ist der Delta-Quotient wegen $\mu_{i+1} < \mu(x) < \mu(x') < \mu_i$ positiv. Zur Bestimmung der oberen Schranke benutzen wir eine Minimalstelle z des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$. Wegen $\mu(x)x' \in \mathscr{B}_\gamma(x)$ gilt $\mu(z) \leq \mu\big(\mu(x)x'\big) = \mu(x') < \mu_i$. Lemma 4.7 angewandt auf x zeigt, dass ein $\bar{x}$ mit $\mu(\bar{x}) = \mu(x)$ und Nichtnull-Anteile auf $\mathcal{Y}_i$, $\mathcal{Y}_{i+1}$ existiert, so dass $\mu(z)$ nicht kleiner als das Minimum $\mu(\bar{z})$ des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\bar{x})$ ist. Lemma 4.4 angewandt auf $\bar{x}$ liefert dann eine obere Schranke $\big((1-\gamma)\mu_{i+1}/\mu_i + \gamma\big)^2$ für $\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{z})\big)/\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{x})\big)$, welche wegen $\mu(\bar{z}) \leq \mu(x')$ und $\mu(\bar{x}) = \mu(x)$ auch obere Schranke von $\Delta\big(\mu_i, \mu_{i+1}, \mu(x')\big)/\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)$ ist. Dass die obere Schranke erreichbar ist, wurde bereits bezüglich des niedrigdimensionalen Falls in Lemma 4.4 gezeigt. □

Für die originale Problemstellung lässt sich diese Abschätzung mittels der Umformulierung in Abschnitt 4.1 rückwärts umschreiben.

Satz 4.9. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 und einen Vektor $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ werde ein Schritt des PINVIT(1)-Verfahrens (1.5) durchgeführt, wobei der Vorkonditionierer B^{-1} die Bedingung (1.4) erfüllt.*
Falls u kein Eigenvektor ist, dann gilt $\rho(u') < \rho(u)$. Gelte zusätzlich $\rho(u) \in (\lambda_i, \lambda_{i+1})$ für ein $i \in \{1, \ldots, m-1\}$, dann gilt mit der Bezeichnung (2.4) entweder $\rho(u') \leq \lambda_i$ oder

$$0 < \frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u')\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u)\big)} \leq \left((1-\gamma)\frac{\lambda_i}{\lambda_{i+1}} + \gamma\right)^2.$$

Die obere Schranke der Delta-Abschätzung lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

4.2.2 Vereinfachungen

Wir formulieren die zweite Version der Konvergenzanalyse für PINVIT(1) als eine Zusammensetzung der Vereinfachungen der ersten Version. Ausgehend von Lemma 4.2 können wir zunächst gemäß Lemma 4.3 genauere Bedingungen zeigen.

Lemma 4.10. *Mit den Formuliereungen in Lemma 4.3 betrachten wir den nichttrivialen Fall, dass das Minimum von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ kleiner als μ_i ist. Dann gibt es eine eindeutige Minimalstelle z, welche durch (4.5) mit einem $\alpha > 0$ darstellbar ist.*

Beweis. Bezeichnen wir das Minimum von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ mit ϑ, dann gilt $\vartheta = \mu(z)$ für eine beliebige Minimalstelle z, und die Gleichung (4.5) lässt sich als

$$H(z - \vartheta x) = \alpha(Hx - z)$$

umschreiben. Die neue linke Seite ist der Gradient der Funktion

$$f(w) := \frac{1}{2}\|w - \vartheta x\|_H^2$$

an der Stelle z. Weiter zeigen wir, dass z eine Minimalstelle von f auf $\partial\mathscr{B}_\gamma(x)$ ist.

Nach den Voraussetzungen gilt $\mu_{i+1} < \mu(z) < \mu_i$, so dass z kein Eigenvektor ist. Nach Lemma 4.3 gilt weiter die Orthogonalität $Hx - z \perp_2 z$. Für beliebiges $w \in \partial\mathscr{B}_\gamma(x)$ folgt dann

$$\begin{aligned}(Hx, z)_2 &= \|z\|_2^2 = \|Hx\|_2^2 - \|Hx - z\|_2^2 = \|Hx\|_2^2 - \|Hx - w\|_2^2 \\ &= \|Hx\|_2^2 - \big(\|Hx\|_2^2 - 2(Hx, w)_2 + \|w\|_2^2\big) = 2(Hx, w)_2 - \|w\|_2^2.\end{aligned}$$

Mittels dieser Gleichheiten vergleichen wir die Funktionswerte $f(z)$ und $f(w)$:

$$\begin{aligned}2\big(f(z) - f(w)\big) &= \|z - \vartheta x\|_H^2 - \|w - \vartheta x\|_H^2 \\ &= \big(\|z\|_H^2 - 2(z, \vartheta x)_H + \|\vartheta x\|_H^2\big) - \big(\|w\|_H^2 - 2(w, \vartheta x)_H + \|\vartheta x\|_H^2\big) \\ &= \big(\|z\|_H^2 - 2\vartheta(Hx, z)_2\big) - \big(\|w\|_H^2 - 2\vartheta(Hx, w)_2\big) \\ &= \big(\vartheta\|z\|_2^2 - 2\vartheta\|z\|_2^2\big) - \big(\mu(w)\|w\|_2^2 - 2\vartheta(Hx, w)_2\big) \\ &= -\vartheta\big(2(Hx, w)_2 - \|w\|_2^2\big) - \big(\mu(w)\|w\|_2^2 - 2\vartheta(Hx, w)_2\big) \\ &= \big(\vartheta - \mu(w)\big)\|w\|_2^2 \le 0,\end{aligned}$$

damit gilt $f(z) \le f(w)$ beziehungsweise $\|z - \vartheta x\|_H \le \|w - \vartheta x\|_H$ für beliebiges $w \in \partial\mathscr{B}_\gamma(x)$. Der Punkt ϑx liegt außerhalb $\mathscr{B}_\gamma(x)$, denn nach Lemma 4.2 gilt für jedes $w \in \mathscr{B}_\gamma(x)$, dass $\mu(w) > \mu(x) = \mu(\vartheta x)$. Wegen der strikten Konvexität der Kugel $\mathscr{B}_\gamma(x)$ gibt es eine eindeutige Minimalstelle des Abstands von einem Punkt aus $\mathscr{B}_\gamma(x)$ zum Außenpunkt ϑx bezüglich der H-Vektornorm und diese Minimalstelle befindet sich auf dem Rand $\partial\mathscr{B}_\gamma(x)$. Daher ist z als eine Minimalstelle von $\|\cdot - \vartheta x\|_H$ beziehungsweise $f(\cdot)$ auf $\mathscr{B}_\gamma(x)$ eindeutig. Nach Betrachtung der Richtungsableitung ist der Gradient von f an z, nämlich $H(z - \vartheta x)$, ein nach Innern von $\mathscr{B}_\gamma(x)$ gerichteter Normalvektor an z und hat daher die gleiche Richtung wie $Hx - z$, damit gilt $\alpha > 0$. □

Zum Vergleich wurde in Lemma 4.3 tatsächlich eine andere lokale Extremstelle $\hat{z}$ von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ bei der Konstruktion der Kollinearitätsbedingung eingemischt. Wenn $\hat{z}$ eine lokale Minimalstelle oder Maximalstelle ist, befindet das entsprchende α im Intervall $\big(-\mu(\hat{z}), 0\big)$ oder $\big(-\infty, -\mu(\hat{z})\big)$. In Lemma 4.10 wurde diese Störung beseitigt, so dass die Bedingung zur echten Minimalstelle klarer wird. Mit den genaueren Bedingungen aus Lemma 4.10 können wir nicht nur den Beweis von Lemma 4.4 vereinfachen, sondern auch dieses Lemma für allgemeinere Indizes erweitern. Dabei sind auch einige Überlegungen vor Lemma 4.7 nützlich.

Lemma 4.11. *Gemäß Voraussetzung 4.1 und Lemma 4.10 betrachten wir den Fall, dass x auf genau zwei Eigenräumen von H Nichtnull-Anteile besitzt und damit durch $x = y_{\sigma(1)} + y_{\sigma(2)}$ mit $y_{\sigma(1)} \in \mathcal{Y}_{\sigma(1)}\backslash\{\mathbf{0}\}$, $y_{\sigma(2)} \in \mathcal{Y}_{\sigma(2)}\backslash\{\mathbf{0}\}$, $1 \le \sigma(1) < \sigma(2) \le m$ darstellbar ist. Das Minimum von*

$\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ sei kleiner als μ_i. Dann gehört die entsprechende eindeutige Minimalstelle z zum Unterraum $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}\}$. *Es gilt anschließend*

$$0 < \frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(z)\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(x)\big)} \leq \left((1-\gamma)\frac{\mu_{\sigma(2)}}{\mu_{\sigma(1)}} + \gamma\right)^2.$$

Die obere Schranke lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Nach Lemma 4.10 gibt es eine eindeutige Minimalstelle z, welche (4.5) mit einem $\alpha > 0$ erfüllt. Die Matrix $H + \alpha I$ ist damit regulär und z ist also kollinear zum Vektor $(H + \alpha I)^{-1}Hx$ und gehört zusammen mit diesem sowie x zum Unterraum $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}\}$.

Für die Delta-Abschätzung wird zunächst wie im Beweis von Lemma 4.10 gezeigt, dass z kein Eigenvektor ist und damit ein Randvektor des Kegels $\mathscr{C}_\gamma(x)$ beziehungsweise des Ebenensegments $\mathscr{C}_\gamma(x) \cap \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}\}$ ist. Es gibt insgesamt zwei mögliche Richtungen für z, und der Euklidische Winkel zwischen z und $y_{\sigma(1)}$ ist entweder $\varphi_1 := \angle_2(z, Hx) + \angle_2(Hx, y_{\sigma(1)})$ oder $\varphi_2 := |\angle_2(z, Hx) - \angle_2(Hx, y_{\sigma(1)})|$. Gemäß der Definition von $\mathscr{C}_\gamma(x)$ gilt $0 < \varphi_2 < \varphi_1 < \angle_2(x, y_{\sigma(1)}) < \pi/2$, so dass z mit φ_1 echt kleineren Rayleigh-Quotienten hat und die Minimalstelle ist. Dann hat z wie x und Hx auch positive Koordinaten bezüglich $y_{\sigma(1)}$, $y_{\sigma(2)}$. Weiter können wir den abzuschätzenden Delta-Quotienten durch einen Tangens-Quotienten ersetzen, nämlich

$$\frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(z)\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(x)\big)} = \frac{\tan^2 \angle_2\big(z, y_{\sigma(1)}\big)}{\tan^2 \angle_2\big(x, y_{\sigma(1)}\big)},$$

vergleiche Lemma 2.12. Dabei sind $\angle_2\big(z, y_{\sigma(1)}\big)$, $\angle_2\big(x, y_{\sigma(1)}\big)$ spitze Winkel, daher genügt es für die Delta-Abschätzung zu zeigen, dass

$$\tan \angle_2\big(z, y_{\sigma(1)}\big) \leq \left((1-\gamma)\frac{\mu_{\sigma(2)}}{\mu_{\sigma(1)}} + \gamma\right) \tan \angle_2\big(x, y_{\sigma(1)}\big). \tag{4.17}$$

Wegen $Hx = \mu_{\sigma(1)} y_{\sigma(1)} + \mu_{\sigma(2)} y_{\sigma(2)}$ gilt

$$\tan \angle_2\big(Hx, y_{\sigma(1)}\big) = \frac{\mu_{\sigma(2)}}{\mu_{\sigma(1)}} \tan \angle_2\big(x, y_{\sigma(1)}\big).$$

So ist (4.17) äquivalent zu

$$\tan \angle_2\big(z, y_{\sigma(1)}\big) \leq (1-\gamma) \tan \angle_2\big(Hx, y_{\sigma(1)}\big) + \gamma \tan \angle_2\big(x, y_{\sigma(1)}\big). \tag{4.18}$$

Mit den Bezeichungen

$$\varphi := \angle_2\big(Hx, x\big), \qquad \vartheta := \angle_2\big(Hx, y_{\sigma(1)}\big)$$

können wir (4.18) in

$$\tan\big(\vartheta + \arcsin(\gamma \sin\varphi)\big) \leq (1-\gamma)\tan\vartheta + \gamma \tan(\vartheta + \varphi). \tag{4.19}$$

umschreiben und die linke Seite als eine Funktion f mit der Variablen γ betrachten. Auf Intervall $[0, 1]$ ist diese eine konvexe Funktion: Wegen

$$\frac{df}{d\gamma}(\gamma) = \frac{\Big(1 + \big(f(\gamma)\big)^2\Big)\sin\varphi}{\sqrt{1 - (\gamma \sin\varphi)^2}} \geq 0$$

ist $f(\gamma)$ monoton steigend auf $[0, 1]$, und genauso der Zähler der Ableitung $(df/d\gamma)(\gamma)$. Im Gegensatz ist der Nenner der Ableitung monoton fallend. Daher ist $(df/d\gamma)(\gamma)$ ebenfalls monoton

steigend auf $[0,1]$, so dass $f(\gamma)$ konvex ist. Weiter gilt nach Eigenschaft einer allgemeinen konvexen Funktion, dass

$$\begin{aligned}\tan\big(\vartheta+\arcsin(\gamma\sin\varphi)\big) &= f(\gamma) = f\big((1-\gamma)\cdot 0+\gamma\cdot 1\big)\\ &\le (1-\gamma)f(0)+\gamma f(1) = (1-\gamma)\tan\vartheta+\gamma\tan(\vartheta+\varphi).\end{aligned}$$

Damit sind die Ungleichungen (4.19), (4.18), (4.17) sowie die Delta-Abschätzung bewiesen. □

Dementsprechend lässt sich die Begründung der Niedrigdimensionalen Analyse vereinfachen. Wir müssen nicht wie in der klassischen Analyse zeigen, dass der der langsamsten Konvergenz entsprechende H-invariante Unterraum durch Eigenvektoren zu μ_i, μ_{i+1} aufgespannt wird. Es genügt zu zeigen, dass ein solcher H-invarianter Unterraum zweidimensional ist.

Lemma 4.12. *Gemäß Voraussetzung 4.1, Lemma 4.3 und Lemma 4.10 betrachten wir eine Rayleigh-Quotient-Niveaumenge $\mathcal{N}(\tilde\mu) := \{\tilde x \in \mathbb{R}^n\backslash\{\mathbf{0}\}\ ;\ \mu(\tilde x)=\tilde\mu\}$ mit $\tilde\mu \in (\mu_{i+1},\mu_i)$. Für ein beliebiges $\tilde x \in \mathcal{N}(\tilde\mu)$ sei $\tilde z$ eine Minimalstelle des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\tilde x)$. Gelte für ein $\bar x \in \mathcal{N}(\tilde\mu)$, dass eine zugehörige Minimalstelle $\bar z$ eine (höherstufige) Minimalstelle des Rayleigh-Quotienten $\mu(\cdot)$ auf der durch alle $\tilde x \in \mathcal{N}(\tilde\mu)$ erzeugten Minimalstellenmenge $\mathcal{M}$ ist, dann besitzt $\bar x$ auf genau zwei Eigenräumen von H Nichtnull-Anteile.*

Beweis. Für eine solche $\bar x$ betrachten wir eine beliebige durch $\bar x$ laufende stetig differenzierbare Kurve $\{g_1(t)\ ;\ t\in T\subseteq\mathbb{R}\}\subset\mathcal{N}(\tilde\mu)$ mit $g_1(t_g)=\bar x$, dann gibt es eine entsprechende Minimalstellenkurve $\{g_2(t)\ ;\ t\in T\subseteq\mathbb{R}\}\subset\mathcal{M}$ mit $g_2(t_g)=\bar z$. Nach der Konstruktion und Lemma 4.2 ist $g_2(t)$ ein Randpunkt von $\mathscr{B}_\gamma\big(g_1(t)\big)$, daher gilt

$$\|Hg_1(t)-g_2(t)\|_2 = \gamma\|Hg_1(t)-\mu\big(g_1(t)\big)g_1(t)\|_2 = \gamma\|Hg_1(t)-\tilde\mu g_1(t)\|_2.$$

Dann ist $f(t) := \|Hg_1(t)-g_2(t)\|_2^2-\gamma^2\|Hg_1(t)-\tilde\mu g_1(t)\|_2^2$ eine konstante Funktion, so dass

$$0=\frac{df}{dt}(t) = 2\big(Hg_1(t)-g_2(t),\, H\dot g_1(t)-\dot g_2(t)\big)_2 - 2\gamma^2\big(Hg_1(t)-\tilde\mu g_1(t),\, H\dot g_1(t)-\tilde\mu\dot g_1(t)\big)_2,$$

wobei $\dot g_1(t)$, $\dot g_2(t)$ jeweils die Ableitungen $(dg_1/dt)(t)$, $(dg_2/dt)(t)$ bezeichnen. Das gilt insbesondere für $t=t_g$. Mit $g_1(t_g)=\bar x$ und $g_2(t_g)=\bar z$ entsteht dann

$$0 = 2\big(H\bar x-\bar z,\, H\dot g_1(t_g)-\dot g_2(t_g)\big)_2 - 2\gamma^2\big(H\bar x-\tilde\mu\bar x,\, H\dot g_1(t_g)-\tilde\mu\dot g_1(t_g)\big)_2. \tag{4.20}$$

Um diese Gleichung zu vereinfachen, zeigen wir weiter die Orthogonalitäten

$$\big(H\bar x-\bar z,\, \dot g_2(t_g)\big)_2 = 0, \qquad \big(H\bar x-\tilde\mu\bar x,\, \dot g_1(t_g)\big)_2 = 0. \tag{4.21}$$

Da $g_2(t_g)=\bar z$ eine Minimalstelle von $\mu(\cdot)$ auf $\{g_2(t)\ ;\ t\in T\subseteq\mathbb{R}\}$ ist, gilt t_g als eine Minimalstelle der Funktion $\mu\circ g_2$ und aus der notwendigen Bedingung $\big(d\,(\mu\circ g_2)/dt\big)(t_g)=0$ folgt

$$0 = \Big(\nabla\mu\big(g_2(t_g)\big)\Big)^T\dot g_2(t_g) = \big(\nabla\mu(\bar z),\dot g_2(t_g)\big)_2.$$

Nach Lemma 4.3 und Lemma 4.10 angewandt auf $\bar x$ erfüllen $\bar x$, $\bar z$ mit einem zugehörigen Parameter $\bar\alpha>0$ die Gleichung

$$(H+\bar\alpha I)\bar z = \big(\mu(\bar z)+\bar\alpha\big)H\bar x. \tag{4.22}$$

Damit gilt

$$H\bar z-\mu(\bar z)\bar z = (H+\bar\alpha I)\bar z-\big(\mu(\bar z)+\bar\alpha\big)\bar z = \big(\mu(\bar z)+\bar\alpha\big)\big(H\bar x-\bar z\big).$$

Das zeigt die Kollinearität von $\nabla\mu(\bar z)$ und $H\bar x-\bar z$ (vergleiche (4.6) im Beweis von Lemma 4.3), so dass $\big(H\bar x-\bar z,\, \dot g_2(t_g)\big)_2=0$ gilt.

Da $\mu(\cdot)$ eine konstante Funktion auf $\mathcal{N}(\tilde{\mu})$ und $\mu \circ g_1$ damit auch eine konstante Funktion ist, gilt $d\,(\mu \circ g_2)/dt = 0$ für beliebiges t und insbesondere für $t = t_g$. Daraus folgt

$$0 = \Big(\nabla\mu\big(g_1(t_g)\big)\Big)^T \dot{g}_1(t_g) = \big(\nabla\mu(\bar{x}), \dot{g}_1(t_g)\big)_2 = \frac{2}{\bar{x}^T\bar{x}}\big(H\bar{x} - \tilde{\mu}\bar{x}, \dot{g}_1(t_g)\big)_2$$

und anschließend $\big(H\bar{x} - \tilde{\mu}\bar{x},\, \dot{g}_1(t_g)\big)_2 = 0$.
Damit gelten die Orthogonalitäten in (4.21). Einsetzen in (4.20) liefert dann

$$\begin{aligned} 0 &= 2\big(H\bar{x} - \bar{z},\, H\dot{g}_1(t_g)\big)_2 - 2\gamma^2\big(H\bar{x} - \tilde{\mu}\bar{x},\, H\dot{g}_1(t_g)\big)_2 \\ &= 2\big((H\bar{x} - \bar{z}) - \gamma^2(H\bar{x} - \tilde{\mu}\bar{x}),\, H\dot{g}_1(t_g)\big)_2 = 2\big(w,\, \dot{g}_1(t_g)\big)_2 \end{aligned}$$

mit $w := H\big((1-\gamma^2)H\bar{x} + \gamma^2\tilde{\mu}\bar{x} - \bar{z}\big)$. Der Vektor w liegt also Euklidisch orthogonal zum Tangentialvektor $\dot{g}_1(t_g)$. Mit hinreichend vielen durch $\bar{x}$ und auf $\mathcal{N}(\tilde{\mu})$ laufenden stetig differenzierbaren Kurven bilden solche Tangentialvektoren eine Tangentialhyperebene $\dot{\bar{x}}$ an $\bar{x}$, so dass $w \perp_2 \dot{\bar{x}}$ gilt und w entweder ein Nullvektor oder ein Normalvektor an $\bar{x}$ ist. Wegen der Orthogonalität $\big(H\bar{x} - \tilde{\mu}\bar{x},\, \dot{g}_1(t_g)\big)_2 = 0$ ist $H\bar{x} - \tilde{\mu}\bar{x}$ ebenfalls Euklidisch orthogonal zu $\dot{\bar{x}}$. Wegen $\mu(\bar{x}) = \tilde{\mu} \in (\mu_{i+1}, \mu_i)$ ist $\bar{x}$ kein Eigenvektor, so dass das Residuum $H\bar{x} - \tilde{\mu}\bar{x}$ von Nullvektor verschieden und weiter ein Normalvektor an $\bar{x}$ ist. Damit gibt es ein $\beta \in \mathbb{R}$ mit $w = \beta(H\bar{x} - \tilde{\mu}\bar{x})$. Multiplizieren wir diese Gleichung mit $(H + \bar{\alpha}I)$, ergibt sich mit der Darstellung $w = H\big((1-\gamma^2)H\bar{x} + \gamma^2\tilde{\mu}\bar{x} - \bar{z}\big)$ von w, dass

$$(1-\gamma^2)H^2(H + \bar{\alpha}I)\bar{x} + \gamma^2\tilde{\mu}H(H + \bar{\alpha}I)\bar{x} - H(H + \bar{\alpha}I)\bar{z} = \beta(H + \bar{\alpha}I)(H - \tilde{\mu}I)\bar{x}.$$

Das lässt sich mittels (4.22) und der Darstellung $\bar{x} = \sum_{i=1}^m Y_i\bar{c}_i$ bezüglich Euklidisch orthonormaler Basismatrizen der Eigenräume in $\sum_{i=1}^m p(\mu_i)\, Y_i\bar{c}_i = \mathbf{0}$ umschreiben, wobei $p(\delta) = a_0 + a_1\delta + a_2\delta^2 + a_3\delta^3$ ein kubisches Polynom mit den Koeffizienten

$$a_0 = \bar{\alpha}\beta\tilde{\mu}, \quad a_1 = \bar{\alpha}\gamma^2\tilde{\mu} + \beta(\tilde{\mu} - \bar{\alpha}), \quad a_2 = \gamma^2(\tilde{\mu} - \bar{\alpha}) - \mu(\bar{z}) - \beta, \quad a_3 = 1 - \gamma^2$$

ist. Multiplizieren wir $\sum_{i=1}^m p(\mu_i)\, Y_i\bar{c}_i = \mathbf{0}$ mit den Matrizen Y_j^T, dann erhalten wir wegen der Euklidischen Orthonormalität der Eigenräume die Gleichungen

$$p(\mu_j)\,\bar{c}_j = \mathbf{0}, \qquad j = 1, \ldots, m.$$

Da $\bar{x}$ kein Nullvektor und kein Eigenvektor ist, besitzt $\bar{x}$ mindestens zwei Nichtnull-Koeffizietenvektoren der Form $\bar{c}_j$. Damit hat $p(\delta)$ wegen $\mu_j > 0\ \forall\ j$ mindestens zwei positive Nullstellen. Außerdem ist $p(\delta)$ ein kubisches Polynom mit reellen Koeffizienten und hat höchstens drei positive Nullstellen. Weiter untersuchen wir die Koeffizienten von $p(\delta)$, um den Fall, dass $p(\delta)$ drei positive Nullstellen hat, auszuschließen. Durch Koeffizientengleich der Darstellungen $p(\delta) = a_0 + a_1\delta + a_2\delta^2 + a_3\delta^3$ und $p(\delta) = a_3\prod_{i=1}^3(\delta - \delta_i)$, wobei $\delta_1, \delta_2, \delta_3$ die Nullstellen bezeichnen, gilt $\prod_{i=1}^3 \delta_i = -a_0/a_3$. Falls alle drei Nullstellen positiv sind, dann müsste $-a_0/a_3 > 0$ gelten. Wegen $\gamma \in (0,1)$ gilt $a_3 > 0$. Weiter zeigen wir $a_0 > 0$, so dass $-a_0/a_3 < 0$ gilt und $p(\delta)$ weniger als drei positive Nullstellen besitzt. Da $\bar{\alpha} > 0$ und $\tilde{\mu} > 0$ bereits bekannt sind, ist nur noch $\beta > 0$ zu zeigen. Multiplizieren wir die obige Kollinearitätsgleichung $\beta(H\bar{x} - \tilde{\mu}\bar{x}) = w$ mit $(H\bar{x} - \tilde{\mu}\bar{x})^T H^{-1}$, gilt

$$\begin{aligned} \beta(H\bar{x} - \tilde{\mu}\bar{x})^T H^{-1}(H\bar{x} - \tilde{\mu}\bar{x}) &= (H\bar{x} - \tilde{\mu}\bar{x})^T\big((1-\gamma^2)H\bar{x} + \gamma^2\tilde{\mu}\bar{x} - \bar{z}\big) \\ &= (H\bar{x} - \tilde{\mu}\bar{x})^T(H\bar{x} - \bar{z}) - \gamma^2\|H\bar{x} - \tilde{\mu}\bar{x}\|_2^2 \\ &= (H\bar{x} - \tilde{\mu}\bar{x})^T(H\bar{x} - \bar{z}) - \|H\bar{x} - \bar{z}\|_2^2 \\ &= \big((H\bar{x} - \tilde{\mu}\bar{x}) - (H\bar{x} - \bar{z})\big)^T(H\bar{x} - \bar{z}) \\ &= (\bar{z} - \tilde{\mu}\bar{x})^T(H\bar{x} - \bar{z}). \end{aligned}$$

Außerdem kann man in $\mathcal{N}(\tilde{\mu})$ ein $\tilde{x}$ finden, welches genau auf den Eigenräumen $\mathcal{Y}_i$, $\mathcal{Y}_{i+1}$ Nichtnull-Anteile besitzt. Damit liegt $\mu(H\tilde{x})$ im Intervall (μ_{i+1}, μ_i). Wegen $H\tilde{x} \in \mathscr{B}_\gamma(\tilde{x})$ ist das Minimum von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\tilde{x})$ also kleiner als μ_i. Da $\bar{z}$ eine höherstufige Minimalstelle von $\mu(\cdot)$ ist, gilt $\mu(\bar{z}) < \mu_i$. Nach Lemma 4.2 ist $\mu(\bar{z})$ größer als $\mu(x)$ beziehungsweise μ_{i+1}. Damit ist $\mu(\bar{z})$ kein Eigenwert und $\bar{z}$ kein Eigenvektor. Nach Lemma 4.3 angewandt auf $\bar{x}$ gilt dann die Orthogonalität $H\bar{x} - \bar{z} \perp_2 \bar{z}$, so dass

$$\begin{aligned}(\bar{z} - \tilde{\mu}\bar{x})^T(H\bar{x} - \bar{z}) &= \left(\frac{\tilde{\mu}}{\mu(\bar{z})}\bar{z} - \tilde{\mu}\bar{x}\right)^T (H\bar{x} - \bar{z}) \\ &= \frac{\tilde{\mu}}{\mu(\bar{z})}\big(\bar{z} - \mu(\bar{z})\bar{x}\big)^T(H\bar{x} - \bar{z}) \overset{(4.22)}{=} \frac{\bar{\alpha}\tilde{\mu}}{\mu(\bar{z})}(H\bar{x} - \bar{z})^T H^{-1}(H\bar{x} - \bar{z}).\end{aligned}$$

Damit gilt

$$\beta\|H\bar{x} - \tilde{\mu}\bar{x}\|^2_{H^{-1}} = \frac{\bar{\alpha}\tilde{\mu}}{\mu(\bar{z})}\|(H\bar{x} - \bar{z})\|^2_{H^{-1}}$$

und $\beta > 0$. So besitzt p genau zwei positive Nullstellen und $\bar{x}$ damit auf genau zwei Eigenräumen von H Nichtnull-Anteile. □

Mit Lemma 4.11 und Lemma 4.12 lässt sich die Delta-Abschätzung (4.16) in Satz 4.8 in vereinfachter Weise beweisen, vergleiche Satz 2.14 aus der Musteranalyse.

Für $\mu(x') < \mu_i$ gilt $\mu_{i+1} < \mu(x) < \mu(x') < \mu_i$, so dass der Delta-Quotient positiv ist. Zur Bestimmung der oberen Schranke des Delta-Quotienten setzen wir $\tilde{\mu} = \mu(x)$ gemäß Lemma 4.12 und betrachten x, $\bar{x} \in \mathcal{N}(\tilde{\mu})$ sowie die Minimalstellen z, $\bar{z}$ des Rayleigh-Quotienten $\mu(\cdot)$ auf den Kugeln $\mathscr{B}_\gamma(x)$, $\mathscr{B}_\gamma(\tilde{x})$, wobei $\bar{z}$ eine höherstufige Minimalstelle ist. Es gilt dann $\mu_{i+1} < \mu(x) = \mu(\bar{x}) < \mu(\bar{z}) \leq \mu(z) \leq \mu(x') < \mu_i$, so dass

$$\Delta\big(\mu_i, \mu_{i+1}, \mu(x')\big) \leq \Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{z})\big).$$

Außerdem gilt nach Lemma 4.12, dass $\bar{x}$ auf genau zwei Eigenräumen von H Nichtnull-Anteile besitzt. Seien $\sigma(1) < \sigma(2)$ die zugehörigen Indizes, dann liefert Lemma 4.11 angewandt auf $\bar{x}$ die Abschätzung

$$\frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{z})\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{x})\big)} \leq \left((1-\gamma)\frac{\mu_{\sigma(2)}}{\mu_{\sigma(1)}} + \gamma\right)^2.$$

Das Intervall (μ_{i+1}, μ_i) enthält keinen Eigenwert, damit gilt $\mu_{\sigma(2)} \leq \mu_{i+1} < \mu(\bar{x}) < \mu(\bar{z}) < \mu_i \leq \mu_{\sigma(1)}$ und weiter

$$\frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{z})\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{x})\big)} \leq \frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{z})\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{x})\big)}.$$

Aus diesen Ergebnissen folgt die Delta-Abschätzung (4.16):

$$0 < \frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(x')\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(x)\big)} \leq \frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{z})\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{x})\big)} \leq \left((1-\gamma)\frac{\mu_{\sigma(2)}}{\mu_{\sigma(1)}} + \gamma\right)^2 \leq \left((1-\gamma)\frac{\mu_{i+1}}{\mu_i} + \gamma\right)^2.$$

4.2.3 Ergänzungen

In der oben bewiesenen Delta-Abschätzung für PINVIT(1) ist die obere Schranke „nur“ im Grenzfall erreichbar. Daher ergänzen wir im Folgenden eine grenzfallfreie Abschätzung, welche eine komplexe Darstellung hat und der in [60], [61] vorgestellten scharfen (und auch komlexen) Schranke des Rayleigh-Quotienten der neuen Iterierten entspricht. Dafür brauchen wir die Konvergenzanalyse nur teilweise zu modifizieren. Im Beweis von Lemma 4.11 normieren wir die Nichtnull-Anteile $y_{\sigma(1)}$, $y_{\sigma(2)}$ bezüglich der Euklidischen Vektornorm und setzen $x = \xi_{\sigma(1)}y_{\sigma(1)} + \xi_{\sigma(2)}y_{\sigma(2)}$,

$z = \zeta_{\sigma(1)} y_{\sigma(1)} + \zeta_{\sigma(2)} y_{\sigma(2)}$ für z. Wegen $\mu_{\sigma(2)} < \mu(x) < \mu(z) \le \mu(Hx) < \mu_{\sigma(1)}$ ist z kein Eigenvektor und gehört damit nach Lemma 4.3 zum Rand des Kegels $\mathscr{C}_\gamma(x)$, so dass

$$\sin \angle_2(Hx, z) = \gamma \sin \angle_2(Hx, x)$$

gilt. Bezüglich der Koeffizientenvektoren lässt sich diese Bedingung als

$$\sin \angle_2 \left(\begin{pmatrix} \mu_{\sigma(1)} \xi_{\sigma(1)} \\ \mu_{\sigma(2)} \xi_{\sigma(2)} \end{pmatrix}, \begin{pmatrix} \zeta_{\sigma(1)} \\ \zeta_{\sigma(2)} \end{pmatrix} \right) = \gamma \sin \angle_2 \left(\begin{pmatrix} \mu_{\sigma(1)} \xi_{\sigma(1)} \\ \mu_{\sigma(2)} \xi_{\sigma(2)} \end{pmatrix}, \begin{pmatrix} \xi_{\sigma(1)} \\ \xi_{\sigma(2)} \end{pmatrix} \right)$$

umschreiben. Im Beweis von Lemma 4.11 wurde gezeigt, dass die Koeffizienten von x, z sämtlich positiv sind. Mit den Parametern

$$\xi := \xi_{\sigma(2)} / \xi_{\sigma(1)}, \quad \zeta := \zeta_{\sigma(2)} / \zeta_{\sigma(1)}, \quad \vartheta := \mu_{\sigma(2)} / \mu_{\sigma(1)}$$

gilt dann $\sin \angle_2\Big((1, \vartheta\xi)^T, (1, \zeta)^T \Big) = \gamma \sin \angle_2\Big((1, \vartheta\xi)^T, (1, \xi)^T \Big)$. Diese Sinusbedingung beleibt geltend, wenn diese zweidimensionalen Vektoren durch die Ergänzung einer gleichen dritten Komponente 0 dreidimensionale Vektoren werden, also

$$\sin \angle_2(a, b) = \gamma \sin \angle_2(a, c)$$

mit $a := (1, \vartheta\xi, 0)^T$, $b := (1, \zeta, 0)^T$, $c := (1, \xi, 0)^T$. Nach der Eigenschaft des äußeren Produktes $\langle \cdot, \cdot \rangle$ für dreidimensionale Vektoren gilt dann

$$\frac{\big\| \langle a, b \rangle \big\|_2}{\|a\|_2 \|b\|_2} = \gamma \frac{\big\| \langle a, c \rangle \big\|_2}{\|a\|_2 \|c\|_2}.$$

Daraus folgt eine Koeffizientenbedingung

$$\gamma^2 = \frac{\big\| \langle a, b \rangle \big\|_2^2 \|c\|_2^2}{\big\| \langle a, c \rangle \big\|_2^2 \|b\|_2^2} = \frac{(\zeta - \vartheta\xi)^2 (1 + \xi^2)}{(\xi - \vartheta\xi)^2 (1 + \zeta^2)} = \frac{\big((\zeta/\xi) - \vartheta\big)^2 (1 + \xi^2)}{(1 - \vartheta)^2 \big(1 + (\zeta/\xi)^2 \xi^2\big)}. \tag{4.23}$$

Das Quadrat des Quotienten ζ/ξ ist genau der abzuschätzenden Delta-Quotient:

$$\left(\frac{\zeta}{\xi} \right)^2 = \frac{\zeta_{\sigma(2)}^2 / \zeta_{\sigma(1)}^2}{\xi_{\sigma(2)}^2 / \xi_{\sigma(1)}^2} = \frac{\tan^2 \angle_2\big(z, y_{\sigma(1)}\big)}{\tan^2 \angle_2\big(x, y_{\sigma(1)}\big)} = \frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(z)\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(x)\big)}.$$

Weiter liefert (4.23) eine quadratische Gleichung für $\delta := \zeta/\xi$:

$$\big((1 + \xi^2) - \gamma^2 \xi^2 (1 - \vartheta)^2\big) \delta^2 - 2\vartheta(1 + \xi^2)\delta + \big(\vartheta^2 (1 + \xi^2) - \gamma^2 (1 - \vartheta)^2\big) = 0$$

mit den Wurzeln

$$\delta_\pm = \frac{\vartheta(1 + \xi^2) \pm \gamma(1 - \vartheta) \sqrt{(1 + \xi^2)(1 + \vartheta^2 \xi^2) - \gamma^2 \xi^2 (1 - \vartheta)^2}}{(1 + \xi^2) - \gamma^2 \xi^2 (1 - \vartheta)^2}. \tag{4.24}$$

Wegen $\vartheta = \mu_{\sigma(2)} / \mu_{\sigma(1)} \in (0, 1)$ und $\gamma \in (0, 1)$ gilt

$$(1 + \xi^2)(1 + \vartheta^2 \xi^2) - \gamma^2 \xi^2 (1 - \vartheta)^2 > 0, \qquad (1 + \xi^2) - \gamma^2 \xi^2 (1 - \vartheta)^2 > 0,$$

so dass $\delta_+ > 0$ gilt. Da z positive Koeffizienten besitzt, wird ζ/ξ durch δ_+ und damit der Delta-Quotient $\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(z)\big) / \Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(x)\big)$ durch δ_+^2 gegeben. Wegen $\vartheta = \mu_{\sigma(2)} / \mu_{\sigma(1)}$ und $\xi = \sqrt{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(x)}$ lässt sich der Delta-Quotient als eine Funktion von $\mu_{\sigma(1)}$, $\mu_{\sigma(2)}$, $\mu(x)$ und γ betrachten. Im Grenzfall $\mu(x) \to \mu_{\sigma(1)}$ gilt $\xi \to 0$, so dass

$$\lim_{\mu(x) \to \mu_{\sigma(1)}} \frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(z)\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(x)\big)} = \lim_{\xi \to 0} \delta_+^2 = \big(\vartheta + \gamma(1 - \vartheta)\big)^2 = \left((1 - \gamma) \frac{\mu_{\sigma(2)}}{\mu_{\sigma(1)}} + \gamma \right)^2.$$

Kombinieren wir dieses Ergebnis mit der klassischen Analyse, können wir weiter in Lemma 4.3, Satz 4.8 und Satz 4.9 scharfe Schranken wie δ_+^2 ergänzen.

Das Quadrat der Wurzel δ_- können wir bei der Abschätzung der schnellsten Konvergenz von PINVIT(1) verwenden. Die klassische Analyse in [62, 65] lässt sich mit ähnlicher Technik in Abschnitt 4.2.2 ebenfalls vereinfachen. Beim Ausgangspunkt wird ein spezieller Fall für die weitere Untersuchung ausgeschlossen, nämlich dass $\mathscr{B}_\gamma(x)$ einen Eigenvektor zum größten Eigenwert μ_1 enthält. Dann wird eine Kollinearität zur Beschreibung einer Maximalstelle erhalten.

Lemma 4.13. *Gemäß Voraussetzung 4.1 gilt für jedes $w \in \mathscr{B}_\gamma(x)$, dass $\mu(w) > \mu(x)$. Falls $\mathscr{B}_\gamma(x)$ keinen Eigenvektor zu μ_1 enthält, dann gehört jede Maximalstelle von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ zum Rand von $\mathscr{B}_\gamma(x)$. Es gelten auch die gleichartigen Aussagen bezüglich des Kegels $\mathscr{C}_\gamma(x)$. Weiter gibt es für eine beliebige Maximalstelle z von $\mu(\cdot)$ auf $\mathscr{B}_\gamma(x)$ ein $\alpha \in \mathbb{R}$ mit $\alpha \leq -\mu(z)$, so dass*

$$(H + \alpha I)z = \big(\mu(z) + \alpha\big)Hx.$$

z ist genau dann ein Eigenvektor, wenn $\alpha = -\mu(z)$. Falls z kein Eigenvektor ist, dann gilt $Hx - z \perp_2 z$, so dass z ein Randvektor des Kegels $\mathscr{C}_\gamma(x)$ ist.

Der Beweis ist analog zu den Beweisen von Lemma 4.2 und Lemma 4.3. Mit der Bedingung $\alpha \leq -\mu(z)$ kann man aber nicht ausschließen, dass $H + \alpha I$ singulär ist und z damit mehr Nichtnull-Koeffizientenvektoren als x besitzt. Daher ist eine niedrigdimensionale Analyse nicht für beliebiges x, z formulierbar. Glücklicherweise gibt es keine solche Schwierigkeit für $\bar{x}, \bar{z}$ zu einem höherstufigen Maximum.

Lemma 4.14. *Gemäß Voraussetzung 4.1 betrachten wir eine Rayleigh-Quotient-Niveaumenge $\mathcal{N}(\tilde{\mu}) := \{\tilde{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\} \ ; \ \mu(\tilde{x}) = \tilde{\mu}\}$ mit $\tilde{\mu} \in (\mu_{i+1}, \mu_i)$. Für ein beliebiges $\tilde{x} \in \mathcal{N}(\tilde{\mu})$ sei $\tilde{z}$ eine Maximalstelle des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{B}_\gamma(\tilde{x})$. Gemäß Lemma 4.13 werde angenommen, dass es $\tilde{x} \in \mathcal{N}(\tilde{\mu})$ gibt, dessen zugehörige Kugel $\mathscr{B}_\gamma(\tilde{x})$ keinen Eigenvektor zu μ_1 enthält. Die Maximalstellen $\tilde{z}$ zu solchen Vektoren $\tilde{x}$ bilden eine Maximalstellenmenge $\mathcal{M}$. Weiter gelte für ein $\bar{x}$ unter solchen $\tilde{x}$, dass eine zugehörige Maximalstelle $\bar{z}$ eine (höherstufige) Maximalstelle des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathcal{M}$ ist, dann gibt es genau zwei verschiedene Eigenwerte von H, zu deren Eigenräumen $\bar{x}$ nicht Eukldisch orthogonal liegt. Es gibt dann Eigenwerte $\mu_{\sigma(1)} > \mu_{\sigma(2)}$ mit zugehörigen Eigenvektoren $y_{\sigma(1)}, y_{\sigma(2)}$, so dass $\bar{x} \in \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}\}$ gilt. Weiter gehört $\bar{z}$ ebenfalls zu $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}\}$.*

Der Beweis ist analog zum Beweis von Lemma 4.12. Zunächst kann man in einer hinreichend kleinen Umgebung von $\bar{x}$ Kurven definieren, auf denen die Vektoren die Bedingung, dass die zugehörigen Kugeln keinen Eigenvektor zu μ_1 enthalten, erfüllen. Weiter ist ein Unterschied zu bemerken, dass die Parameter $\bar{\alpha}$ und β negativ sind. Das führt aber auch zu einem kubischen Polynom mit positiven Koeffizienten a_0, a_3, so dass $\bar{x}$ höchstens zwei Nichtnull-Koeffizientenvektoren hat. Weiter folgt aus der Darstellung $w = H\big((1-\gamma^2)H\bar{x} + \gamma^2\tilde{\mu}\bar{x} - \bar{z}\big)$ und der Kollinearitätsbedingung $w = \beta(H\bar{x} - \tilde{\mu}\bar{x})$, dass $\bar{z}$ wie $\bar{x}$ durch $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}\}$ enthalten wird. Mit der anschließenden niedrigdimensionalen Analyse und dem entsprechenden Rückwärtseinsetzen wird eine Delta-Abschätzung

$$\frac{\Delta\big(\mu_1, \mu_m, \mu(x')\big)}{\Delta\big(\mu_1, \mu_m, \mu(x)\big)} \geq \delta_-^2$$

erhalten, wobei δ_- mittels der Formel (4.24) mit $\vartheta = \mu_m/\mu_1$, $\xi = \sqrt{\Delta\big(\mu_1, \mu_m, \mu(x)\big)}$ gegeben wird. Diese Delta-Abschätzung lässt sich nach $\mu(x')$ umstellen und entspricht der Abschätzung aus [62]. Es gibt auch zwei nicht sehr bedeutende Abschätzungsvarianten, nämlich die „lokale“ langsamste/schnellste Konvergenz in $\operatorname{span}\{y_1, y_m\}$ / $\operatorname{span}\{y_i, y_{i+1}\}$. Dafür werden die Konvergenzfaktoren durch δ_+^2 beziehungsweise δ_-^2 mit passenden Indizes gegeben.

4.3 Konvergenzanalyse für PINVIT(2)

Als Ausgangspunkt der Konvergenzanalyse für PINVIT(2) definieren wir eine Menge, um einen zu $\hat{x}$ kollinearen Nichtnullvektor einzuschließen.

Lemma 4.15. *Gemäß Voraussetzung 4.1 gilt für ein beliebiges $w \in \mathscr{C}_\gamma(x)$, dass der durch x und w aufgespannte Unterraum* $\operatorname{span}\{x, w\}$ *zweidimensional ist. Es gibt ein eindeutiges $\beta(w) \in \mathbb{R}$, so dass $\beta(w)\, x + w$ ein Ritzvektor zum größten Ritzwert von H bezüglich* $\operatorname{span}\{x, w\}$ *ist. Dann enthält die Menge $\mathscr{D}_1(x) := \{\beta(w)\, x + w \ ; \ w \in \mathscr{C}_\gamma(x)\}$ einen zu $\hat{x}$ kollinearen Nichtnullvektor. Für jedes $z \in \mathscr{D}_1(x)$ gilt $\mu(z) > \mu(x)$.*

Beweis. Nach der Definition von $\mathscr{C}_\gamma(x)$ und $\gamma \in (0, 1)$ gehört x nicht zu $\mathscr{C}_\gamma(x)$, so dass x nicht kollinear zu w ist und $\operatorname{span}\{x, w\}$ zweidimensional ist. Nach Lemma 4.2 gilt $\mu(w) > \mu(x)$, so dass der größte Ritzwert von H bezüglich $\operatorname{span}\{x, w\}$ auch echt größer als $\mu(x)$ ist. Daher ist x kein zugehöriger Ritzvektor und $\operatorname{span}\{x, w\}$ besitzt zwei eindimensionale Ritzräume. Mit diesen Ritzräumen hat die Gerade $\tilde{\beta}\, x + w$ jeweils einen eindeutigen Schnittpunkt. Damit gibt es ein eindeutiges $\beta(w) \in \mathbb{R}$, so dass $\beta(w)\, x + w$ ein Ritzvektor zum größten Ritzwert von H bezüglich $\operatorname{span}\{x, w\}$ ist. Gemäß Voraussetzung 4.1 ist $\hat{x}$ ein Ritzvektor zum größten Ritzwert von H bezüglich $\operatorname{span}\{x, x'\}$ mit $x' \in \mathscr{C}_\gamma(x)$, so dass $\mathscr{D}_1(x)$ einen zu $\hat{x}$ kollinearen Nichtnullvektor enthält. Für jedes $z \in \mathscr{D}_1(x)$ gilt mittels eines z erzeugenden Vektors $w \in \mathscr{C}_\gamma(x)$, dass $\mu(z) \geq \mu(w) > \mu(x)$. □

Weiter zeigen wir wie in Lemma 4.2, dass die Betrachtung der Randvektoren bei der weiteren Untersuchung für PINVIT(2) genügt.

Lemma 4.16. *Mit den Formulierungen in Lemma 4.15 ist das Minimum des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{D}_1(x)$ identisch zum Minimum von $\mu(\cdot)$ auf $\mathscr{D}_2(x) := \{\beta(w)\, x + w \ ; \ w \in \partial\mathscr{C}_\gamma(x)\}$.*

Beweis. Mit den Unterraum-Mengen

$$\mathcal{M}_1 := \big\{\operatorname{span}\{x, w\} \ ; \ w \in \mathscr{C}_\gamma(x)\big\} \quad \text{und} \quad \mathcal{M}_2 := \big\{\operatorname{span}\{x, w\} \ ; \ w \in \partial\mathscr{C}_\gamma(x)\big\}$$

sind die Minima von $\mu(\cdot)$ auf $\mathscr{D}_1(x)$ und $\mathscr{D}_2(x)$ jeweils die Minima der größten Ritzwerte von H bezüglich der Unterräume aus $\mathcal{M}_1$ und $\mathcal{M}_2$.

Daher genügt es zu zeigen, dass $\mathcal{M}_1$ und $\mathcal{M}_2$ identisch sind. $\mathcal{M}_2 \subseteq \mathcal{M}_1$ ist trivial. Zu $\mathcal{M}_1 \subseteq \mathcal{M}_2$ brauchen wir nur zu zeigen, dass $\big(\mathcal{M}_1 \setminus \mathcal{M}_2\big) \subseteq \mathcal{M}_2$. Dafür betrachten wir einen beliebigen echt inneren Vektor w von $\mathscr{C}_\gamma(x)$. Der Durchschnitt von $\operatorname{span}\{x, w\}$ und $\mathscr{C}_\gamma(x)$ ist ein Teilgebiet $\tilde{\mathscr{C}}$ der Ebene $\operatorname{span}\{x, w\}$. Der Vektor x befindet sich außerhalb $\mathscr{C}_\gamma(x)$ und daher auch außerhalb $\tilde{\mathscr{C}}$. Damit ist x nicht kollinear zu den Randvektoren von $\tilde{\mathscr{C}}$, so dass $\operatorname{span}\{x, w\}$ durch x und einen beliebigen Randvektor von $\tilde{\mathscr{C}}$ aufgespannt werden kann. Da jeder Randvektor von $\tilde{\mathscr{C}}$ offenbar auch ein Randvektor von $\mathscr{C}_\gamma(x)$ ist, gilt $\big(\mathcal{M}_1 \setminus \mathcal{M}_2\big) \subseteq \mathcal{M}_2$. □

Damit ist das Minimum von $\mu(\cdot)$ auf $\mathscr{D}_2(x)$ eine obere Schranke von $\mu(\hat{x})$ und wird bei der folgenden Untersuchung betrachtet.

4.3.1 Begründung der niedrigdimensionalen Analyse

Mittels einer Kurven-Differentiation lassen sich implizite Darstellungen der Minimalstellen konstruieren.

Lemma 4.17. *Mit den Formulierungen in Lemma 4.16 betrachten wir eine beliebige Minimalstelle z von $\mu(\cdot)$ auf $\mathscr{D}_2(x)$ und einen erzeugenden Vektor $w \in \partial\mathscr{C}_\gamma(x)$ von z. Es gibt dann ein $\alpha \in \mathbb{R}$, so dass die Gleichung*

$$\big(H + (\alpha\eta - \mu(z))I\big)w = \big((\alpha - \beta(w))H + \beta(w)\mu(z)\big)x \tag{4.25}$$

durch z, w, x mit $\eta := (Hx, w)_2/\|w\|_2^2$ erfüllen.

Beweis. Für beliebiges $\tilde{w} \in \partial\mathscr{C}_\gamma(x)$ ist nach Lemma 4.15 eine Funktion definierbar:

$$f(\tilde{w}) := \mu\big(\beta(\tilde{w})x + \tilde{w}\big).$$

Sei $w \in \partial\mathscr{C}_\gamma(x)$ ein z erzeugender Vektor, dann gilt $\mu(z) = f(w)$. Weiter betrachten wir eine beliebige durch w laufende stetig differenzierbare Kurve $\{g(t)\ ;\ t \in T \subseteq \mathbb{R}\} \subset \partial\mathscr{C}_\gamma(x)$ mit $g(t_g) = w$, dann ist w eine Minimalstelle von f auf der Kurve, so dass t_g eine Minimalstelle der Funktion $f \circ g$ ist und die notwendige Bedingung $\big(d\,(f \circ g)/dt\big)(t_g) = 0$ gilt. Daraus folgt mit

$$(f \circ g)(t) = \mu\big(\beta\big(g(t)\big)x + g(t)\big),$$

$$\frac{d(f \circ g)}{dt}(t) = \big(\nabla\mu\big(\beta\big(g(t)\big)x + g(t)\big)\big)^T \left(\frac{d(\beta \circ g)}{dt}(t)x + \dot{g}(t)\right),$$

wobei $\dot{g}(t)$ die Ableitung $(dg/dt)(t)$ bezeichnet, dass

$$0 = \big(\nabla\mu\big(\beta\big(g(t_g)\big)x + g(t_g)\big)\big)^T \left(\frac{d(\beta \circ g)}{dt}(t_g)x + \dot{g}(t_g)\right) = \big(\nabla\mu(z)\big)^T \left(\frac{d(\beta \circ g)}{dt}(t_g)x + \dot{g}(t_g)\right).$$

Da z ein Ritzvektor zum Unterraum $\operatorname{span}\{x, w\}$ ist, gilt die Orthogonalität $\nabla\mu(z) \perp_2 x$, so dass aus der obigen notwendigen Bedingung weiter die Orthogonalität $\nabla\mu(z) \perp_2 \dot{g}(t_g)$ folgt. Analog zum Beweis von Lemma 2.13 wird gezeigt, dass der Vektor $\nabla\mu(z)$ Euklidisch orthogonal zur durch solche Tangentialvektoren aufgespannten Tangentialhyperebene $\dot{w}$ liegt. Auf $\partial\mathscr{C}_\gamma(x)$ ist

$$h(w) := \cos^2 \angle_2(Hx, w) = \frac{(Hx, w)_2^2}{\|Hx\|_2^2\|w\|_2^2}$$

eine konstante Funktion. Mit einer ähnlichen Kurven-Differentiation wird dann gezeigt, dass der Vektor $Hx - \eta w$ mit $\eta := (Hx, w)_2/\|w\|_2^2$ ebenfalls Euklidisch orthogonal $\dot{w}$ liegt. Wegen $w \in \partial\mathscr{C}_\gamma(x)$ und $\gamma \in (0, 1)$ ist w nicht kollinear zu Hx, so dass $Hx - \eta w$ kein Nullvektor ist und ein $\tilde{\alpha} \in \mathbb{R}$ mit

$$\nabla\mu(z) = \tilde{\alpha}(Hx - \eta w) \tag{4.26}$$

existiert. Mit den Substitutionen $\nabla\mu(z) = \dfrac{2}{z^T z}\big(Hz - \mu(z)z\big)$, $z = \beta(w)x + w$ und $\alpha = \dfrac{\tilde{\alpha}}{2}z^T z$ lässt sich (4.26) in (4.25) umschreiben. □

Wenn w eine Minimalstelle z erzeugt, dann erzeugt $\tilde{\eta}w$ mit einem beliebigen $\tilde{\eta} \in \mathbb{R}\backslash\{0\}$ einen zu z kollinearen Vektor, welcher auch eine Minimalstelle ist. Deswegen können wir ein solches w umskalieren, so dass $1 = \eta = (Hx, w)_2/\|w\|_2^2$. Damit ist w auch ein Randpunkt der Kugel $\mathscr{B}_\gamma(x)$, denn wegen

$$\cos \angle_2(Hx, w) = \frac{(Hx, w)_2}{\|Hx\|_2\|w\|_2} = \eta\frac{\|w\|_2}{\|Hx\|_2} = \frac{\|w\|_2}{\|Hx\|_2}$$

bilden Hx und w ein recheckiges Dreieck, so dass

$$\|Hx - w\|_2 = \|Hx\|_2 \sin \angle_2(Hx, w) = \gamma\|Hx\|_2 \sin \angle_2(Hx, x) = \gamma\big\|Hx - \mu(x)x\big\|_2.$$

Weiter untersuchen wir solche umskalierte Minimalstellen-Erzeuger gemäß der entsprchenden leicht modifizierten Bedingung

$$\big(H + (\alpha - \mu(z))I\big)w = \big((\alpha - \beta(w))H + \beta(w)\mu(z)\big)x. \tag{4.27}$$

Dabei kann man nicht ausschließen, dass $H + (\alpha - \mu(z))I$ singulär ist, wie ein einfaches Beispiel zeigt: Gegeben sei eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{3\times 3}$ mit Eigenwerten $\mu_1 > \mu_2 > \mu_3$ und zugehörigen Eigenvektoren y_1, y_2, y_3. Ein $x \in \operatorname{span}\{y_1, y_2\}$ sei kein Nullvektor und kein Eigenvektor von H. Für eine Minimalstelle $z = \beta(w)x + w$ von $\mu(\cdot)$ auf $\mathscr{D}_2(x)$ soll dann $w \notin \operatorname{span}\{y_1, y_2\}$ gelten, denn sonst wäre $\operatorname{span}\{x, w\} = \operatorname{span}\{y_1, y_2\}$, so dass $\mu(z) = \mu_1$ gelten würde und $\mu(\cdot)$ eine konstante Funktion auf $\mathscr{D}_2(x)$ wäre; aber in $\mathscr{D}_2(x)$ gibt es offenbar auch Ritzvektoren zu den von $\operatorname{span}\{y_1, y_2\}$ verschiedenen Unterräumen, so dass $\mu(z) < \mu_1$ gelten soll. Deswegen besitzt w auf y_3 einen Nichtnull-Koeffizienten, und nach (4.27) gilt $\mu_3 + (\alpha - \mu(z)) = 0$, so dass $H + (\alpha - \mu(z))I$ singulär ist. Interessanterweise gibt es keine solche Singularität für eine höherstufige Minimalstelle bezüglich einer Rayleigh-Quotienten-Niveaumenge.

Lemma 4.18. *Gemäß Voraussetzung 4.1, Lemma 4.17 und der modifizierten Bedingung (4.27) betrachten wir eine Rayleigh-Quotient-Niveaumenge $\mathcal{N}(\tilde{\mu}) := \{\tilde{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\} \ ; \ \mu(\tilde{x}) = \tilde{\mu}\}$ mit $\tilde{\mu} \in (\mu_{i+1}, \mu_i)$. Für ein beliebiges $\tilde{x} \in \mathcal{N}(\tilde{\mu})$ sei $\tilde{z} = \beta(\tilde{w})\tilde{x} + \tilde{w}$ eine Minimalstelle des Rayleigh-Quotienten $\mu(\cdot)$ auf $\mathscr{D}_2(\tilde{x})$ mit $\tilde{w} \in \partial\mathscr{B}_\gamma(\tilde{x}) \cap \partial\mathscr{C}_\gamma(\tilde{x})$. Gelte für ein $\bar{x} \in \mathcal{N}(\tilde{\mu})$, dass eine zugehörige Minimalstelle $\bar{z}$ eine (höherstufige) Minimalstelle des Rayleigh-Quotienten $\mu(\cdot)$ auf der durch alle $\tilde{x} \in \mathcal{N}(\tilde{\mu})$ erzeugten Minimalstellenmenge $\mathcal{M}$ ist, dann besitzt $\bar{x}$ auf höchstens drei Eigenräumen von H Nichtnull-Anteile. Es gibt also Eigenwerte $\mu_{\sigma(1)} > \mu_{\sigma(2)} > \mu_{\sigma(3)}$ mit zugehörigen Eigenvektoren $y_{\sigma(1)}$, $y_{\sigma(2)}$, $y_{\sigma(3)}$, so dass $\bar{x} \in \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$ gilt. Weiter gehören $\bar{z}$ und der $\bar{z}$ erzeugende Vektor $\bar{w}$ ebenfalls zu* $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$.

Beweis. Der erste Anteil des Beweises ist ganz analog zum Beweis von Satz 4.12. Für eine solche $\bar{x}$ betrachten wir eine beliebige durch $\bar{x}$ laufende stetig differenzierbare Kurve $\{g_1(t) \ ; \ t \in T \subseteq \mathbb{R}\} \subset \mathcal{N}(\tilde{\mu})$ mit $g_1(t_g) = \bar{x}$, dann gibt es dementsprechend eine Minimalstellenkurve $\{g_2(t) \ ; \ t \in T \subseteq \mathbb{R}\}$ mit $g_2(t_g) = \bar{z}$ und eine Erzeugerkurve $\{g_3(t) \ ; \ t \in T \subseteq \mathbb{R}\}$ mit $g_3(t_g) = \bar{w}$. Gemäß der Konstruktion ist $g_3(t)$ ein Randpunkt von $\mathscr{B}_\gamma\big(g_1(t)\big)$, daher gilt

$$\|Hg_1(t) - g_3(t)\|_2 = \gamma\|Hg_1(t) - \mu\big(g_1(t)\big)g_1(t)\|_2 = \gamma\|Hg_1(t) - \tilde{\mu}g_1(t)\|_2.$$

Dann ist $f(t) := \|Hg_1(t) - g_3(t)\|_2^2 - \gamma^2\|Hg_1(t) - \tilde{\mu}g_1(t)\|_2^2$ eine konstante Funktion, so dass

$$0 = \frac{df}{dt}(t) = 2\big(Hg_1(t) - g_3(t),\, H\dot{g}_1(t) - \dot{g}_3(t)\big)_2 - 2\gamma^2\big(Hg_1(t) - \tilde{\mu}g_1(t),\, H\dot{g}_1(t) - \tilde{\mu}\dot{g}_1(t)\big)_2,$$

wobei $\dot{g}_1(t)$, $\dot{g}_3(t)$ jeweils die Ableitungen $(dg_1/dt)(t)$, $(dg_3/dt)(t)$ bezeichnen. Das gilt insbesondere für $t = t_g$. Mit $g_1(t_g) = \bar{x}$ und $g_3(t_g) = \bar{w}$ entsteht dann

$$0 = 2\big(H\bar{x} - \bar{w},\, H\dot{g}_1(t_g) - \dot{g}_3(t_g)\big)_2 - 2\gamma^2\big(H\bar{x} - \tilde{\mu}\bar{x},\, H\dot{g}_1(t_g) - \tilde{\mu}\dot{g}_1(t_g)\big)_2. \tag{4.28}$$

Um diese Gleichung zu vereinfachen, zeigen wir weiter die Orthogonalitäten

$$\big(H\bar{x} - \bar{w},\, \dot{g}_3(t_g)\big)_2 = \big(H\bar{x} - \bar{w},\, -\beta(\bar{w})\dot{g}_1(t_g)\big)_2, \qquad \big(H\bar{x} - \tilde{\mu}\bar{x},\, \dot{g}_1(t_g)\big)_2 = 0. \tag{4.29}$$

Da $g_2(t_g) = \bar{z}$ eine Minimalstelle von $\mu(\cdot)$ auf $\{g_2(t) \ ; \ t \in T \subseteq \mathbb{R}\}$ ist, gilt t_g als eine Minimalstelle der Funktion $\mu \circ g_2$ und aus der notwendigen Bedingung $\big(d\,(\mu \circ g_2)/dt\big)(t_g) = 0$ folgt

$$0 = \Big(\nabla\mu\big(g_2(t_g)\big)\Big)^T \dot{g}_2(t_g) = \big(\nabla\mu(\bar{z}), \dot{g}_2(t_g)\big)_2,$$

wobei $\dot{g}_2(t)$ die Ableitung $(dg_2/dt)(t)$ bezeichnet. Diese Orthogonalität lässt sich wegen

$$g_2(t) = \beta\big(g_3(t)\big)g_1(t) + g_3(t), \qquad \dot{g}_2(t) := \frac{dg_2}{dt}(t) = \frac{d(\beta \circ g_3)}{dt}(t)g_1(t) + \beta\big(g_3(t)\big)\dot{g}_1(t) + \dot{g}_3(t)$$

sowie $g_1(t_g) = \bar{x}$, $g_3(t_g) = \bar{w}$ als

$$\Big(\nabla\mu(\bar{z}),\, \frac{d(\beta \circ g_3)}{dt}(t_g)\bar{x} + \beta(\bar{w})\dot{g}_1(t_g) + \dot{g}_3(t_g)\Big)_2 = 0$$

umschreiben. Da $\bar{z}$ ein Ritzvektor zum Unterraum $\operatorname{span}\{\bar{x}, \bar{w}\}$ ist, gilt $\nabla\mu(\bar{z}) \perp_2 \bar{x}$, so dass die obige Bedingung in

$$\big(\nabla\mu(\bar{z}),\, \beta(\bar{w})\dot{g}_1(t_g) + \dot{g}_3(t_g)\big)_2 = 0$$

vereinfacht wird. Außerdem folgt aus $\bar{z} = \beta(\bar{w})\bar{x} + \bar{w}$ und der Bedingung (4.27) angewandt auf $\bar{x}$, nämlich

$$\big(H + (\bar{\alpha} - \mu(\bar{z}))I\big)\bar{w} = \big((\bar{\alpha} - \beta(\bar{w}))H + \beta(\bar{w})\mu(\bar{z})\big)\bar{x}, \tag{4.30}$$

eine Kollinearität von $\nabla\mu(\bar{z})$ und $H\bar{x} - \bar{w}$, so dass $\big(H\bar{x} - \bar{w},\, \beta(\bar{w})\dot{g}_1(t_g) + \dot{g}_3(t_g)\big)_2 = 0$ und weiter die erste Gleichung in (4.29) gilt. Da $\mu(\cdot)$ eine konstante Funktion auf $\mathcal{N}(\tilde{\mu})$ und $\mu \circ g_1$ damit auch eine konstante Funktion ist, gilt $d\,(\mu \circ g_2)/dt = 0$ für beliebiges t und insbesondere für $t = t_g$. Daraus folgt $\big(H\bar{x} - \tilde{\mu}\bar{x},\, \dot{g}_1(t_g)\big)_2 = 0$.

Damit gelten die Orthogonalitäten in (4.29). Einsetzen in (4.28) liefert dann

$$\begin{aligned} 0 &= 2\big(H\bar{x} - \bar{w},\, H\dot{g}_1(t_g) + \beta(\bar{w})\dot{g}_1(t_g)\big)_2 - 2\gamma^2\big(H\bar{x} - \tilde{\mu}\bar{x},\, H\dot{g}_1(t_g)\big)_2 \\ &= 2\Big(\big(H + \beta(\bar{w})I\big)(H\bar{x} - \bar{w}) - \gamma^2 H\big(H\bar{x} - \tilde{\mu}\bar{x}\big),\, \dot{g}_1(t_g)\Big)_2 = 2\big(v, \dot{g}_1(t_g)\big)_2 \end{aligned}$$

mit $v := (1-\gamma^2)H^2\bar{x} + \big(\beta(\bar{w}) + \gamma^2\tilde{\mu}\big)H\bar{x} - \big(H + \beta(\bar{w})I\big)\bar{w}$. Der Vektor v liegt also Euklidisch orthogonal zum Tangentialvektor $\dot{g}_1(t_g)$. Analog zum Beweis von Satz 4.12 wird gezeigt, dass es ein $\tau \in \mathbb{R}$ mit $v = \tau(H\bar{x} - \tilde{\mu}\bar{x})$ gibt. Multiplizieren wir diese Gleichung mit $H + (\bar{\alpha} - \mu(\bar{z}))I$ und nutzen (4.30) als Substitution, erhalten wir die Darstellung $p(H)\bar{x} = \mathbf{0}$ mit einem kubischen Polynom p mit reellen Koeffizienten (Details der Koeffizienten von p sind nicht relevant für diesen Beweis). Mittels der Darstellung $\bar{x} = \sum_{i=1}^{m} Y_i\bar{c}_i$ bezüglich Euklidisch orthonormaler Basismatrizen der Eigenräume entstehen die Gleichungen

$$p(\mu_j)\,\bar{c}_j = \mathbf{0}, \qquad j = 1, \ldots, m.$$

Da p höchstens drei positive Nullstellen hat, gibt es höchstens drei Nichtnull-Koeffizientenvektoren $\bar{c}_j$. In anderen Worten besitzt $\bar{x}$ auf höchstens drei Eigenräumen Nichtnull-Anteile. Wegen $\mu(\bar{x}) = \tilde{\mu} \in (\mu_{i+1}, \mu_i)$ ist $\bar{x}$ kein Eigenvektor und hat daher zwei oder drei Nichtnull-Koeffizientenvektoren. Damit können wir die Nichtnull-Anteile von $\bar{x}$ beziehungsweise eventuell einen zusätzlichen Eigenvektor aus einem anderen Eigenraum als Basisvektoren wählen, um einen $\bar{x}$ enthaltenden Unterraum $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$ mit angeordneten Eigenwerten $\mu_{\sigma(1)} > \mu_{\sigma(2)} > \mu_{\sigma(3)}$ zu definieren.

Wäre $\bar{w} \notin \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$, dann würde $\bar{w}$ wegen (4.30) höchstens vier Nichtnull-Koeffizientenvektoren bezüglich der Eigenbasen besitzen und wäre durch

$$\bar{w} = \omega_{\sigma(1)}y_{\sigma(1)} + \omega_{\sigma(2)}y_{\sigma(2)} + \omega_{\sigma(3)}y_{\sigma(3)} + \omega_s y_s$$

mit einem $\omega_s \neq 0$ und ein Eigenvektor y_s zu einem von $\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu_{\sigma(3)}$ verschiedenen Eigenwert μ_s darstellbar. Dafür müsste noch $\bar{\alpha} - \mu(\bar{z}) = -\mu_s$ gelten.

Weiter folgt aus der Kollinearitätsbedingung $v = \tau(H\bar{x} - \tilde{\mu}\bar{x})$, dass

$$\big(H + \beta(\bar{w})I\big)\bar{w} = (1-\gamma^2)H^2\bar{x} + \big(\beta(\bar{w}) + \gamma^2\tilde{\mu}\big)H\bar{x} - \tau(H\bar{x} - \tilde{\mu}\bar{x}) \in \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\},$$

so dass $\beta(\bar{w}) = -\mu_s = \bar{\alpha} - \mu(\bar{z})$ gelten müsste. Einsetzten dieser Gleichung in (4.30) liefert aber

$$(H - \mu_s I)\bar{w} = \big(\mu(\bar{z})H - \mu_s\mu(\bar{z})\big)\bar{x} = \mu(\bar{z})(H - \mu_s I)\bar{x}.$$

Wegen $\mu_s \notin \{\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu_{\sigma(3)}\}$ ist die Matrix $H - \mu_s I$ in $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$ „lokal invertierbar“, so dass $\bar{w}$ durch eine orthogonale Aufspaltung

$$\bar{w} = \mu(\bar{z})\bar{x} + \omega_s y_s$$

darstellbar wäre. Damit würde $(H\bar{x}, \omega_s y_s)_2 = 0$, $\|\bar{w}\|_2 \geq \|\mu(\bar{z})\bar{x}\|_2$, und weiter

$$\cos\angle_2(H\bar{x}, \bar{w}) = \frac{(H\bar{x}, \bar{w})_2}{\|H\bar{x}\|_2\|\bar{w}\|_2} = \frac{\big(H\bar{x}, \mu(\bar{z})\bar{x}\big)_2 + (H\bar{x}, \omega_s y_s)_2}{\|H\bar{x}\|_2\|\bar{w}\|_2} = \frac{\big(H\bar{x}, \mu(\bar{z})\bar{x}\big)_2}{\|H\bar{x}\|_2\|\bar{w}\|_2}$$

$$\leq \frac{\big(H\bar{x}, \mu(\bar{z})\bar{x}\big)_2}{\|H\bar{x}\|_2\|\mu(\bar{z})\bar{x}\|_2} = \frac{(H\bar{x}, \bar{x})_2}{\|H\bar{x}\|_2\|\bar{x}\|_2} = \cos\angle_2(H\bar{x}, \bar{x})$$

gelten, das heißt $\angle_2(H\bar{x}, \bar{w}) \geq \angle_2(H\bar{x}, \bar{x})$, so dass $\bar{w}$ außerhalb des Kegels $\mathscr{C}_\gamma(\tilde{x})$ liegen müsste. Dieser Widerspruch zur Voraussetzung $\bar{w} \in \partial\mathscr{C}_\gamma(\tilde{x})$ zeigt $\bar{w} \in \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$ und anschließend $\bar{z} = \beta(\bar{w})\bar{x} + \bar{w} \in \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$. □

Eine alternative Version der Begründung der niedrigdimensionalen Analyse wird durch Neymeyr [67] präsentiert. Dabei zu betrachten ist ein Kegel, welcher auch die Kugel $\mathscr{B}_\gamma(x)$ umfasst, dessen Achse jedoch durch $\mu(x)x$ und den Mittelpunkt Hx von $\mathscr{B}_\gamma(x)$ bestimmt wird.

4.3.2 Ellipsen-Analyse

Gemäß Lemma 4.18 können wir eine Delta-Abschätzung mittels einer ähnlichen niedrigdimensionalen Analyse wie bei INVIT(2) entwickeln. Hauptsächliche Beweistechnik dafür ist eine modifizierte Ellipsen-Analyse.

Lemma 4.19. *Mit den Formulierungen von Lemma 4.18 gilt entweder $\mu(\bar{z}) \geq \mu_i$ oder*

$$0 < \frac{\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{z})\big)}{\Delta\big(\mu_i, \mu_{i+1}, \mu(\bar{x})\big)} \leq \left(\frac{\hat{\kappa} + \gamma(2 - \hat{\kappa})}{(2 - \hat{\kappa}) + \gamma\hat{\kappa}}\right)^2 \quad \textit{mit} \quad \hat{\kappa} := \frac{\mu_{i+1} - \mu_m}{\mu_i - \mu_m}.$$

Die obere Schranke lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Zu betrachten ist also der Fall $\mu(\bar{z}) < \mu_i$. Dafür ist 0 offenbar eine untere Schranke des Delta-Quotienten. Weiter ist eine obere Schranke zu konstruieren.

Nach Lemma 4.18 befinden sich die Vektoren $\bar{x}$, $\bar{z}$ sowie $\bar{w}$ in einem H-invarianten Raum $\mathcal{Y} := \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}\}$. Da $\bar{z}$ ein Ritzvektor zum größten Ritzwert eines zweidimensionalen Unterraum von $\mathcal{Y}$ ist, gilt nach dem Courant-Fischer-Prinzip, dass $\mu(\bar{z}) \geq \mu_{\sigma(2)}$.

Aus $\mu(\bar{x}) = \tilde{\mu} \in (\mu_{i+1}, \mu_i)$ folgt entweder $\mu_{\sigma(1)} > \mu_{\sigma(2)} \geq \mu_i > \mu(\bar{x}) > \mu_{i+1} \geq \mu_{\sigma(3)}$ oder $\mu_{\sigma(1)} \geq \mu_i > \mu(\bar{x}) > \mu_{i+1} \geq \mu_{\sigma(2)} > \mu_{\sigma(3)}$. Im ersten Fall gilt $\mu(\bar{z}) \geq \mu_{\sigma(2)} \geq \mu_i$. Im zweiten Fall erhalten wir eine Voraussetzung $\mu_{\sigma(1)} > \mu(\bar{x}) > \mu_{\sigma(2)}$ für die niedrigdimensionale Analyse. Damit hat $\bar{x}$ bezüglich $y_{\sigma(1)}$ einen Nichtnull-Koeffizienten und wir können die Basisvektoren $y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}$ normieren beziehungsweise deren Richtungen umkehren, so dass $\bar{x}$ bezüglich $y_{\sigma(1)}$ einen positiven Koeffizienten ξ und bezüglich $y_{\sigma(2)}, y_{\sigma(3)}$ nicht-negative Koeffizienten hat. Es gibt dann $\alpha_2 \geq 0$, $\alpha_3 \geq 0$ mit

$$\bar{x}/\xi = y_{\sigma(1)} + \alpha_2 y_{\sigma(2)} + \alpha_3 y_{\sigma(3)}.$$

Der Vektor $\bar{x}/\xi$ gehört zum Durchschnitt des affinen Raums $\mathcal{E}(1) := y_{\sigma(1)} + \operatorname{span}\{y_{\sigma(2)}, y_{\sigma(3)}\}$ mit der Rayleigh-Quotienten-Niveaumenge $\mathcal{N}(\tilde{\mu})$. Weiter ist

$$\mathscr{E} := \mathcal{E}(1) \cap \mathcal{N}(\tilde{\mu})$$

eine Ellipse, denn ein $v = y_{\sigma(1)} + \vartheta_2 y_{\sigma(2)} + \vartheta_3 y_{\sigma(3)} \in \mathcal{E}(1)$ hat den Rayleigh-Quotienten $\mu(v) = \tilde{\mu}$ genau dann, wenn die Koeffizienten ϑ_2, ϑ_3 die Ellipsengleichung

$$\frac{\vartheta_2^2}{a^2} + \frac{\vartheta_3^2}{b^2} = 1 \quad \text{mit} \quad a := \sqrt{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \tilde{\mu}\big)}, \quad b := \sqrt{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(3)}, \tilde{\mu}\big)}$$

erfüllen. Daher können wir $\bar{x}/\xi$ mit anderen Punkten auf dieser Ellipse vergleichen, um den Delta-Quotienten nach oben abzuschätzen. Wegen $\alpha_2 \geq 0$, $\alpha_3 \geq 0$ können wir

$$\alpha_2 = a\cos(\varphi), \quad \alpha_3 = b\sin(\varphi) \quad \text{mit} \quad \varphi \in [0,\, \pi/2] \tag{4.31}$$

setzen und nur ein Viertel der Ellipse betrachten. Weiter sind „Vertreter“ von $\bar{w}$, $\bar{z}$ in $\mathcal{E}(1)$ zu bestimmen. Nach Lemma 4.2 und Lemma 4.15 angewandt auf $\bar{x}$ gilt $\mu(\bar{z}) \geq \mu(\bar{w}) > \mu(\bar{x}) > \mu_{\sigma(2)}$, so dass $\bar{w}$, $\bar{z}$ bezüglich $y_{\sigma(1)}$ Nichtnull-Koeffizienten ω, ζ haben und die umskalierten Vektoren $\bar{w}/\omega$, $\bar{z}/\zeta$ zu $\mathcal{E}(1)$ gehören. Da $\bar{z}$ ein Ritzvektor zum größten Ritzwert von H bezüglich des Unterraums $\operatorname{span}\{\bar{x}, \bar{w}\}$ ist, gilt $\operatorname{span}\{\bar{x}, \bar{w}\}$ als eine Tangentialebene der Rayleigh-Quotienten-Niveaumenge $\mathcal{N}(\mu(\bar{z}))$ (diese ist ein elliptischer Doppelkegel ohne Nullpunkt) an $\operatorname{span}\{\bar{z}\}$, so dass

$$\mathscr{G} := \mathcal{E}(1) \cap \operatorname{span}\{\bar{x}, \bar{w}\}$$

eine Tangente der Ellipse

$$\bar{\mathscr{E}} := \mathcal{E}(1) \cap \mathcal{N}(\mu(\bar{z}))$$

an $\bar{z}/\zeta$ ist. $\bar{\mathscr{E}}$ hat die Halbachsen

$$\bar{a} := \sqrt{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{z})\big)}, \quad \bar{b} := \sqrt{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(3)}, \mu(\bar{z})\big)}.$$

Im Vergleich mit den Halbachsen a, b der Ellipse $\mathscr{E}$ gilt $1 < (\bar{a}/\bar{b}) < (a/b)$. Als eine Hilfsgröße nutzen wir eine weitere Ellipse $\tilde{\mathscr{E}}$ mit den Halbachsen $\tilde{a}$, $\tilde{b}$, wofür $(\tilde{a}/\tilde{b}) = (a/b)$ gilt und $\mathscr{G}$ auch eine Tagente von $\tilde{\mathscr{E}}$ ist. Dann lässt sich $\tilde{a} \geq \bar{a}$ durch Kontraposition zeigen: Unter der Annahme $\tilde{a} < \bar{a}$ sei $y_{\sigma(1)} + \vartheta_2 y_{\sigma(2)} + \vartheta_3 y_{\sigma(3)}$ ein beliebiger Punkt auf $\tilde{\mathscr{E}}$. Dann würde

$$\vartheta_2^2 + \frac{\bar{a}^2}{\bar{b}^2}\vartheta_3^2 < \vartheta_2^2 + \frac{a^2}{b^2}\vartheta_3^2 = \vartheta_2^2 + \frac{\tilde{a}^2}{\tilde{b}^2}\vartheta_3^2 = \tilde{a}^2 < \bar{a}^2 \quad \Rightarrow \quad \frac{\vartheta_2^2}{\bar{a}^2} + \frac{\vartheta_3^2}{\bar{b}^2} < 1$$

gelten, so dass $\tilde{\mathscr{E}}$ echt innerhalb der Ellipse $\bar{\mathscr{E}}$ liegen würde und keine Schnittpunkte mit $\mathscr{G}$ hätte. Als Folgerung gilt

$$\frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{z})\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{x})\big)} = \frac{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \mu(\bar{z})\big)}{\Delta\big(\mu_{\sigma(1)}, \mu_{\sigma(2)}, \tilde{\mu}\big)} = \frac{\bar{a}^2}{a^2} \leq \frac{\tilde{a}^2}{a^2}. \tag{4.32}$$

Im Folgenden wird $\tilde{a}^2/a^2$ als eine obere Schranke des Delta-Quotienten weiter nach oben abgeschätzt. $\tilde{a}^2/a^2$ lässt sich mittels der Schnittpunkte der Tangente $\mathscr{G}$ mit $y_{\sigma(1)} + \operatorname{span}\{y_{\sigma(2)}\}$ und $y_{\sigma(1)} + \operatorname{span}\{y_{\sigma(3)}\}$ umschreiben. Seien diese Schnittpunkte durch

$$y_{\sigma(1)} + \eta_2 y_{\sigma(2)} \quad \text{und} \quad y_{\sigma(1)} + \eta_3 y_{\sigma(3)}$$

gegeben, dann liegt der Berührungspunkt $\tilde{y}$ der Ellipse $\tilde{\mathscr{E}}$ mit der Tangente $\mathscr{G}$ auf der geraden Strecke zwischen $y_{\sigma(1)} + \eta_2 y_{\sigma(2)}$ und $y_{\sigma(1)} + \eta_3 y_{\sigma(3)}$ und lässt sich durch

$$\tilde{y} = y_{\sigma(1)} + (1-\tau)\eta_2 y_{\sigma(2)} + \tau\eta_3 y_{\sigma(3)}$$

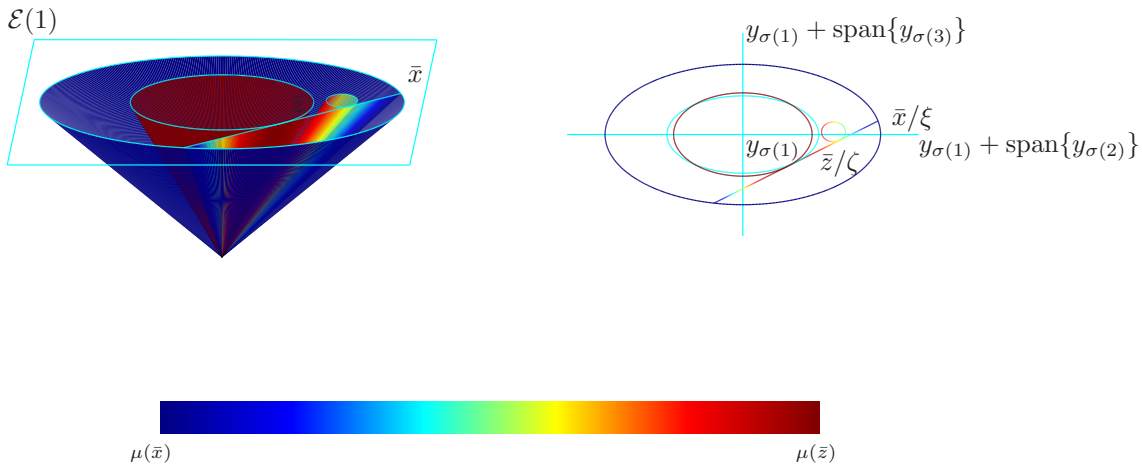


Abbildung 4.2: Darstellung für PINVIT(2) in einem affinen Raum

mit einem $\tau \in [0,1]$ darstellen. Die Ellipsengleichung $\vartheta_2^2/\tilde{a}^2 + \vartheta_3^2/\tilde{b}^2 = 1$ von $\tilde{\mathscr{E}}$ definiert den Einheitskreis der D-Vektornorm mit $D := \mathrm{diag}(1/\tilde{a}^2,\, 1/\tilde{b}^2)$. Bezüglich des $\mathcal{E}(1)$-Koordinatensystems entspricht der Vektor $(\eta_2, -\eta_3)^T$ der Richtung der Tangente $\mathscr{G}$ und der Vektor $\big((1-\tau)\eta_2,\, \tau\eta_3\big)^T$ der Richtung der Verbindungsstrecke von dem Mittelpunkt $y_{\sigma(1)}$ zum Berührungspunkt $\tilde{y}$. Damit gilt die Orthogonalität

$$(\eta_2, -\eta_3)^T \perp_D \big((1-\tau)\eta_2,\, \tau\eta_3\big)^T,$$

welche $\tau = \eta_2^2/\big(\eta_2^2 + (\tilde{a}^2/\tilde{b}^2)\eta_3^2\big)$ liefert. Außerdem erfüllt die Koordinaten $(1-\tau)\eta_2$ und $\tau\eta_3$ von $\tilde{y}$ die Ellipsengleichung von $\tilde{\mathscr{E}}$, so dass

$$\tilde{a}^2 = \big((1-\tau)\eta_2\big)^2 + \frac{\tilde{a}^2}{\tilde{b}^2}\big(\tau\eta_3\big)^2 = \frac{\eta_2^2\eta_3^2}{(\tilde{b}^2/\tilde{a}^2)\eta_2^2 + \eta_3^2} = \frac{\eta_2^2\eta_3^2}{(b^2/a^2)\eta_2^2 + \eta_3^2}$$

und weiter

$$\frac{\tilde{a}^2}{a^2} = \frac{\eta_2^2\eta_3^2}{b^2\eta_2^2 + a^2\eta_3^2} \tag{4.33}$$

gilt. Dabei lassen sich η_2, η_3 noch bezüglich der Koeffizienten von $\bar{x}/\xi$ sowie der Eigenwerte umschreiben. Die Punkte $y_{\sigma(1)} + \eta_2 y_{\sigma(2)}$ und $y_{\sigma(1)} + \eta_3 y_{\sigma(3)}$ liegen auf der Tangente $\mathscr{G}$, welche auch die Punkte $\bar{x}/\xi$ und $\bar{w}/\omega$ enthält. Ein beliebiger Punkt auf $\mathscr{G}$ lässt sich dann mit einem zugehörigen $\tau \in \mathbb{R}$ durch $(1-\tau)(\bar{x}/\xi) + \tau(\bar{w}/\omega)$ beziehungsweise mit den $y_{\sigma(2)}$, $y_{\sigma(3)}$-Koeffizieten α_2, α_3 von $\bar{x}/\xi$ und β_2, β_3 von $\bar{w}/\omega$ durch

$$y_{\sigma(1)} + \big((1-\tau)\alpha_2 + \tau\beta_2\big)y_{\sigma(2)} + \big((1-\tau)\alpha_3 + \tau\beta_3\big)y_{\sigma(3)}$$

darstellen. Zur Bestimmung der Parameter η_2, η_3 für $y_{\sigma(1)} + \eta_2 y_{\sigma(2)}$, $y_{\sigma(1)} + \eta_3 y_{\sigma(3)}$ werden lineare Gleichungen aufgestellt und jeweils ein zugehöriges τ ermittelt. Rückwärtseinsetzen liefert

$$\eta_2 = \frac{\alpha_3\beta_2 - \alpha_2\beta_3}{\alpha_3 - \beta_3}, \qquad \eta_3 = \frac{\alpha_2\beta_3 - \alpha_3\beta_2}{\alpha_2 - \beta_2}.$$

Um η_2, η_3 unabhängig von den Koeffizienten β_2, β_3 darzustellen, benötigen wir weitere geometrische Argumente in $\mathcal{Y}$. Mit der orthonormalen Basismatrix $Y = [y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)}]$ von $\mathcal{Y}$ sind die

dreidimensionalen Vektoren $Y^T(\bar{x}/\xi) = (1, \alpha_2, \alpha_3)^T$ und $Y^T(\bar{w}/\omega) = (1, \beta_2, \beta_3)^T$ die Koeffizientenvektoren von $\bar{x}/\xi$ und $\bar{w}/\omega$ und haben das äußere Produkt

$$\big(\alpha_2\beta_3 - \alpha_3\beta_2,\ \alpha_3 - \beta_3,\ \beta_2 - \alpha_2\big)^T.$$

Im Beweis von Lemma 4.18 wurde mit (4.30) gezeigt, dass $H\bar{x}-\bar{w}$ zu $\nabla\mu(\bar{z})$ kollinear ist. Damit liegt $H\bar{x}-\bar{w}$ Euklidisch orthogonal zum Unterraum $\operatorname{span}\{\bar{x}, \bar{w}\}$, und der Koeffizientenvektor $Y^T(H\bar{x}-\bar{w})$ liegt dementsprechend Euklidisch orthogonal zu

$$Y^T\operatorname{span}\{\bar{x}, \bar{w}\} = \operatorname{span}\big\{Y^T(\bar{x}/\xi),\ Y^T(\bar{w}/\omega)\big\}.$$

Daher ist $Y^T(H\bar{x} - \bar{w})$ kollinear zum äußeren Produkt von $Y^T(\bar{x}/\xi)$ und $Y^T(\bar{w}/\omega)$. Wenn eine Darstellung $Y^T(H\bar{x} - \bar{w}) = (\bar{\nu}_1, \bar{\nu}_2, \bar{\nu}_3)^T$ bekannt ist, dann soll $\eta_2 = \bar{\nu}_1/\bar{\nu}_2$, $\eta_3 = -\bar{\nu}_1/\bar{\nu}_3$ gelten.

Ein weiteres geometrisches Argument basiert auf dem Dreieck mit den Eckpunkten $H\bar{x}$, $\tilde{\mu}\bar{x}$, $\bar{w}$. Mit der Bezeichnung $\bar{r} := H\bar{x} - \tilde{\mu}\bar{x}$ sind zwei Seitenlängen durch

$$\|H\bar{x} - \tilde{\mu}\bar{x}\|_2 = \|\bar{r}\|_2 \quad \text{und} \quad \|H\bar{x} - \bar{w}\|_2 = \gamma\|H\bar{x} - \tilde{\mu}\bar{x}\|_2 = \gamma\|\bar{r}\|_2$$

$\big($denn $\bar{w} \in \partial\mathscr{B}_\gamma(\bar{x})\big)$ gegeben. Wegen $H\bar{x} - \bar{w} \perp_2 \operatorname{span}\{\bar{x}, \bar{w}\}$ liegen die Seitenvektoren $H\bar{x} - \bar{w}$ und $\tilde{\mu}\bar{x} - \bar{w}$ Euklidisch orthogonal zueinander, so dass dieses Dreieck rechteckig ist. Auf der Seite zwischen $H\bar{x}$ und $\tilde{\mu}\bar{x}$ hat der Punkt

$$\bar{y} := (1 - \gamma^2)H\bar{x} + \gamma^2\tilde{\mu}\bar{x}$$

den minimalen Abstand zum Eckpunkt $\bar{w}$ bezüglich der Euklidischen Vektornorm. Dementsprechend gilt die Orthogonalität

$$\bar{w} - \bar{y} \perp_2 H\bar{x} - \tilde{\mu}\bar{x}.$$

Außerdem gilt wegen $H\bar{x} - \tilde{\mu}\bar{x} \perp_2 \bar{x}$ und $H\bar{x} - \bar{w} \perp_2 \bar{x}$, dass

$$\bar{w} - \bar{y} = \gamma^2(H\bar{x} - \tilde{\mu}\bar{x}) - (H\bar{x} - \bar{w}) \perp_2 \bar{x}.$$

Damit erhalten wir die Orthogonalität

$$\bar{w} - \bar{y} \perp_2 \operatorname{span}\{H\bar{x} - \tilde{\mu}\bar{x},\ \bar{x}\} = \operatorname{span}\{\bar{x}/\xi,\ H\bar{x}/\xi\}.$$

Dementsprechend liegt der Koeffizientenvektor $Y^T(\bar{w} - \bar{y})$ Euklidisch orthogonal zum Unterraum $\operatorname{span}\{Y^T(\bar{x}/\xi),\ Y^T(H\bar{x}/\xi)\}$ und ist kollinear zum äußeren Produkt von

$$Y^T(\bar{x}/\xi) = (1, \alpha_2, \alpha_3)^T \quad \text{und} \quad Y^T(H\bar{x}/\xi) = \big(\mu_{\sigma(1)}, \mu_{\sigma(2)}\alpha_2, \mu_{\sigma(3)}\alpha_3\big)^T,$$

nämlich $\big((\mu_{\sigma(3)}-\mu_{\sigma(2)})\alpha_2\alpha_3,\ (\mu_{\sigma(1)}-\mu_{\sigma(3)})\alpha_3,\ (\mu_{\sigma(2)}-\mu_{\sigma(1)})\alpha_2\big)^T$, welches im Folgenden mit e bezeichnet wird. Dann hat $Y^T(\bar{w}-\bar{y})$ zwei mögliche Richtungen $\pm e$ (diese entsprechen nicht unbedingt beide dem echten Minimum), so dass

$$Y^T(H\bar{x} - \bar{w}) = Y^T(H\bar{x} - \bar{y}) - Y^T(\bar{w} - \bar{y}) = Y^T\gamma^2\bar{r} \pm \big\|Y^T(\bar{w} - \bar{y})\big\|_2 e/\|e\|_2.$$

Der Faktor $\big\|Y^T(\bar{w} - \bar{y})\big\|_2$ lässt sich mit γ und $\|\bar{r}\|_2$ darstellen:

$$\big\|Y^T(\bar{w} - \bar{y})\big\|_2 = \|\bar{w} - \bar{y}\|_2 = \cos\angle_2(H\bar{x} - \bar{w},\ \bar{y} - \bar{w})\|H\bar{x} - \bar{w}\|_2$$

$$= \cos\angle_2(H\bar{x} - \tilde{\mu}\bar{x},\ \bar{w} - \tilde{\mu}\bar{x})\|H\bar{x} - \bar{w}\|_2 = \sqrt{1 - \gamma^2}\,\gamma\|\bar{r}\|_2.$$

Damit gilt $Y^T(H\bar{x}-\bar{w}) = Y^T\gamma^2\bar{r} \pm \sqrt{1-\gamma^2}\,\gamma\, e\big(\|\bar{r}\|_2/\|e\|_2\big)$. Der Quotient $\|\bar{r}\|_2/\|e\|_2$ lässt sich noch als $\xi^2/\|\bar{x}\|_2$ umschreiben: $\|e\|_2$ ist der Flächeninhalt des von $Y^T(\bar{x}/\xi)$ und $Y^T(H\bar{x}/\xi)$ aufgespannten Parallelogramms. Durch $\|\bar{x}/\xi\|_2 = \|Y^T(\bar{x}/\xi)\|_2$ und $\|\bar{r}/\xi\|_2 = \|Y^T(\bar{r}/\xi)\|_2$ werden eine Seitenlänge und eine zugehörige Höhenlänge dieses Parallelogramms gegeben, so dass $\|\bar{x}/\xi\|_2\|\bar{r}/\xi\|_2 = \|e\|_2$ beziehungsweise $\|\bar{r}\|_2/\|e\|_2 = \xi^2/\|\bar{x}\|_2$ gilt. Zusammengefasst ergibt sich

$$Y^T(H\bar{x}-\bar{w}) = Y^T\gamma^2\bar{r} \pm \sqrt{1-\gamma^2}\,\gamma\,\xi^2 e/\|\bar{x}\|_2.$$

Mit den Substitutionen

$$\begin{aligned} Y^T\gamma^2\bar{r} &= \gamma^2\xi\big(Y^T(H\bar{x}/\xi) - \tilde{\mu}Y^T(\bar{x}/\xi)\big) = \gamma^2\xi\big(\mu_{\sigma(1)}-\tilde{\mu},\, (\mu_{\sigma(2)}-\tilde{\mu})\alpha_2,\, (\mu_{\sigma(3)}-\tilde{\mu})\alpha_3\big)^T,\\ e &= \big((\mu_{\sigma(3)}-\mu_{\sigma(2)})\alpha_2\alpha_3,\, (\mu_{\sigma(1)}-\mu_{\sigma(3)})\alpha_3,\, (\mu_{\sigma(2)}-\mu_{\sigma(1)})\alpha_2\big)^T,\\ \|\bar{x}\|_2 &= \xi\,\|\bar{x}/\xi\|_2 = \xi\sqrt{1+\alpha_2^2+\alpha_3^2} \end{aligned}$$

hat $Y^T(H\bar{x}-\bar{w})$ die Komponenten

$$\begin{aligned} \bar{\nu}_1 &= \gamma^2\xi(\mu_{\sigma(1)}-\tilde{\mu}) \pm \sqrt{1-\gamma^2}\,\gamma\,\xi\,\alpha_2\alpha_3(\mu_{\sigma(3)}-\mu_{\sigma(2)})\big/\sqrt{1+\alpha_2^2+\alpha_3^2},\\ \bar{\nu}_2 &= \gamma^2\xi\alpha_2(\mu_{\sigma(2)}-\tilde{\mu}) \pm \sqrt{1-\gamma^2}\,\gamma\,\xi\,\alpha_3(\mu_{\sigma(1)}-\mu_{\sigma(3)})\big/\sqrt{1+\alpha_2^2+\alpha_3^2},\\ \bar{\nu}_3 &= \gamma^2\xi\alpha_3(\mu_{\sigma(3)}-\tilde{\mu}) \pm \sqrt{1-\gamma^2}\,\gamma\,\xi\,\alpha_2(\mu_{\sigma(2)}-\mu_{\sigma(1)})\big/\sqrt{1+\alpha_2^2+\alpha_3^2}, \end{aligned}$$

so dass die Parameter η_2, η_3 mit Eigenwerten, γ und den Koeffizienten von $\bar{x}/\xi$ darstellbar sind:

$$\begin{aligned} \eta_2 &= \bar{\nu}_1/\bar{\nu}_2 = \frac{\gamma(\mu_{\sigma(1)}-\tilde{\mu})\sqrt{1+\alpha_2^2+\alpha_3^2} \pm \sqrt{1-\gamma^2}\,\alpha_2\alpha_3(\mu_{\sigma(3)}-\mu_{\sigma(2)})}{\gamma\alpha_2(\mu_{\sigma(2)}-\tilde{\mu})\sqrt{1+\alpha_2^2+\alpha_3^2} \pm \sqrt{1-\gamma^2}\,\alpha_3(\mu_{\sigma(1)}-\mu_{\sigma(3)})},\\ \eta_3 &= -\bar{\nu}_1/\bar{\nu}_3 = \frac{\gamma(\tilde{\mu}-\mu_{\sigma(1)})\sqrt{1+\alpha_2^2+\alpha_3^2} \pm \sqrt{1-\gamma^2}\,\alpha_2\alpha_3(\mu_{\sigma(2)}-\mu_{\sigma(3)})}{\gamma\alpha_3(\mu_{\sigma(3)}-\tilde{\mu})\sqrt{1+\alpha_2^2+\alpha_3^2} \pm \sqrt{1-\gamma^2}\,\alpha_2(\mu_{\sigma(2)}-\mu_{\sigma(1)})}. \end{aligned}$$

Im nächsten Schritt können wir diese Darstellungen in (4.33) einsetzen, um $\tilde{a}^2/a^2$ nach oben abzuschätzen. Gemäß unserem Versuch ist dessen Kehrwert $a^2/\tilde{a}^2$ einfacher zu analysieren. Dafür formulieren wir $a^2/\tilde{a}^2$ bezüglich (4.31) und (4.33) als eine Funktion mit den Variablen

$$\delta := a^2, \qquad t := \tan(\varphi).$$

Das wird mit einigen technischen Substitutionen verwirklicht. Zu benutzen sind die Parameter

$$\psi := \frac{1}{\gamma}\sqrt{1-\gamma^2}, \qquad \kappa := \frac{\mu_{\sigma(2)}-\mu_{\sigma(3)}}{\mu_{\sigma(1)}-\mu_{\sigma(3)}}$$

und die Umschreibungen

$$\frac{\mu_{\sigma(1)}-\tilde{\mu}}{\tilde{\mu}-\mu_{\sigma(2)}} = a^2 = \delta, \qquad \frac{\mu_{\sigma(1)}-\tilde{\mu}}{\tilde{\mu}-\mu_{\sigma(3)}} = b^2 = \frac{\delta(1-\kappa)}{1+\kappa\delta},$$

$$\frac{\mu_{\sigma(2)}-\mu_{\sigma(3)}}{\mu_{\sigma(1)}-\tilde{\mu}} = \frac{\kappa(1+\delta)}{\delta(1-\kappa)}, \qquad \frac{\mu_{\sigma(1)}-\mu_{\sigma(3)}}{\tilde{\mu}-\mu_{\sigma(2)}} = \frac{1+\delta}{1-\kappa}, \qquad \frac{\mu_{\sigma(1)}-\mu_{\sigma(2)}}{\tilde{\mu}-\mu_{\sigma(3)}} = \frac{(1-\kappa)(1+\delta)}{1+\kappa\delta},$$

$$\alpha_2 = a\cos(\varphi) = \sqrt{\frac{\delta}{1+t^2}}, \qquad \alpha_3 = b\sin(\varphi) = t\sqrt{\frac{\delta(1-\kappa)}{(1+t^2)(1+\kappa\delta)}}.$$

Dann wird $a^2/\tilde{a}^2$ entweder durch die Funktion

$$\begin{aligned} f_1(\delta,t) :=& \Big((1+\delta)\big(\psi^2(1-\kappa)^2+\kappa(1-\kappa)+\psi^2t^2\big)+(1-\kappa)^2+t^2(1-\kappa) \\ &+2\kappa\psi t\sqrt{(1+\delta)(1-\kappa)(1+t^2+\delta\kappa)/(1+t^2)}\Big) \\ &\Big(\sqrt{(1-\kappa)(1+t^2+\delta\kappa)}+\kappa\psi t\sqrt{(1+\delta)/(1+t^2)}\Big)^{-2} \end{aligned}$$

oder durch eine sehr ähnliche Funktion $f_2(\delta,t)$ mit $f_2(\delta,t) = f_1(\delta,-t)$ dargestellt. Nach den Voraussetzungen und den Konstruktionen der Parameter werden die Bedingungen

$$\delta > 0, \qquad t \geq 0, \qquad \psi > 0, \qquad \kappa \in (0,1)$$

erfüllt. Damit gilt $f_1(\delta,t) \leq f_2(\delta,t)$, denn

$$\begin{aligned} f_1(\delta,t) - f_2(\delta,t) =& -4\kappa\psi^3t(1+\delta)^{3/2}\big(1+t^2-\kappa\big)^2\sqrt{(1-\kappa)(1+t^2+\delta\kappa)}\big(1+t^2\big)^{-3/2} \\ &\Big(\sqrt{(1-\kappa)(1+t^2+\delta\kappa)}+\kappa\psi t\sqrt{(1+\delta)/(1+t^2)}\Big)^{-2} \\ &\Big(\sqrt{(1-\kappa)(1+t^2+\delta\kappa)}-\kappa\psi t\sqrt{(1+\delta)/(1+t^2)}\Big)^{-2} \leq 0. \end{aligned}$$

Daher ist $f_1(\delta,t)$ eine untere Schranke von $a^2/\tilde{a}^2$ und wird weiter betrachetet. Wegen

$$\frac{\partial f_1}{\partial\delta}(\delta,t) = \frac{\psi^2\sqrt{1-\kappa}\big((1-\kappa)^3+3(1-\kappa)^2t^2+3(1-\kappa)t^4+t^6\big)}{(1+t^2)\sqrt{1+t^2+\delta\kappa}\big(\sqrt{(1-\kappa)(1+t^2+\delta\kappa)}+\kappa\psi t\sqrt{(1+\delta)/(1+t^2)}\big)^3} > 0$$

ist f_1 bezüglich δ monoton steigend, so dass

$$f_1(\delta,t) \geq f_1(0,t) = \frac{(1+t^2)\big(\psi^2(1-\kappa)^2+(1+t^2)(1-\kappa)+\psi^2t^2+2\kappa\psi t\sqrt{1-\kappa}\big)}{\big(\sqrt{1-\kappa}(1+t^2)+\kappa\psi t\big)^2} =: g(t),$$

wobei die Gleichheit im Grenzfall $\delta \to 0$ beziehungsweise $\mu(\bar{x}) \to \mu_{\sigma(1)}$ gilt. Weiter betrachten wir die Ableitung

$$\frac{dg}{dt}(t) = \frac{2\kappa\psi^2(1-\kappa+t^2)\big(\psi t^2+2t\sqrt{1-\kappa}-\psi(1-\kappa)\big)}{\big(\sqrt{1-\kappa}(1+t^2)+\kappa\psi t\big)^3}$$

und bestimmen die Extremstellen. Es gilt

$$\begin{aligned} \frac{dg}{dt}(t) = 0 \quad &\Rightarrow \quad \psi t^2+2t\sqrt{1-\kappa}-\psi(1-\kappa) = 0 \\ &\Rightarrow \quad t_{1,2} = \frac{-2\sqrt{1-\kappa}\pm\sqrt{4(1-\kappa)+4\psi^2(1-\kappa)}}{2\psi} = \frac{\sqrt{1-\kappa}(-1\pm\sqrt{1+\psi^2})}{\psi}, \end{aligned}$$

und $t_2 < 0 < t_1$. Da $t \geq 0$ gelten soll, betrachten wir weiter die Ableitung $(dg/dt)(t)$ für $0 \leq t < t_1$ und $t > t_1$. Dann ist $(dg/dt)(t)$ im ersten Fall negativ und im zweiten Fall positiv, so dass t_1 die Minimalstelle von $g(t)$ über $t \geq 0$ ist. Mit

$$t_1 = \frac{\sqrt{1-\kappa}(-1+\sqrt{1+\psi^2})}{\psi} = \frac{\sqrt{1-\kappa}(1-\gamma)}{\sqrt{1-\gamma^2}}$$

gilt daher

$$a^2/\tilde{a}^2 \geq f_1(\delta,t) \geq g(t) \geq g(t_1) = \left(\frac{(2-\kappa)+\gamma\,\kappa}{\kappa+\gamma\,(2-\kappa)}\right)^2.$$

Zusammen mit (4.32) wird eine Zwischen-Abschätzung bewiesen:

$$\frac{\Delta\big(\mu_{\sigma(1)},\mu_{\sigma(2)},\mu(\bar z)\big)}{\Delta\big(\mu_{\sigma(1)},\mu_{\sigma(2)},\mu(\bar x)\big)} \le (\tilde a^2/a^2) = \big(a^2/\tilde a^2\big)^{-1} \le \left(\frac{\kappa+\gamma\,(2-\kappa)}{(2-\kappa)+\gamma\,\kappa}\right)^2 =: h(\kappa).$$

Das zeigt die Delta-Abschätzung mit der oberen Schranke $h(\hat\kappa)$, denn es gilt

$$\kappa = \frac{\mu_{\sigma(2)}-\mu_{\sigma(3)}}{\mu_{\sigma(1)}-\mu_{\sigma(3)}} \le \frac{\mu_{i+1}-\mu_m}{\mu_i-\mu_m} = \hat\kappa$$

und $h(\kappa)$ ist bezüglich κ monoton steigend. Die obere Schranke lässt sich mit $\mu_{\sigma(1)}=\mu_i,\ \mu_{\sigma(2)}=\mu_{i+1},\ \mu_{\sigma(3)}=\mu_m$, im Grenzfall $\mu(\bar x)\to\mu_i$ erreichen. □

Kombination von Lemma 4.18 und Lemma 4.19 liefert eine Delta-Abschätzung für PINVIT(2).

Satz 4.20. *Gemäß Voraussetzung 4.1 gilt $\mu(\hat x)>\mu(x)$. Im Fall $\mu(\hat x)<\mu_i$ gilt*

$$0<\frac{\Delta\big(\mu_i,\mu_{i+1},\mu(\hat x)\big)}{\Delta\big(\mu_i,\mu_{i+1},\mu(x)\big)} \le \left(\frac{\hat\kappa+\gamma(2-\hat\kappa)}{(2-\hat\kappa)+\gamma\hat\kappa}\right)^2 \quad \textit{mit} \quad \hat\kappa := \frac{\mu_{i+1}-\mu_m}{\mu_i-\mu_m}. \tag{4.34}$$

Die obere Schranke lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Da $\hat x$ ein Ritzvektor zum größten Ritzwert von H bezüglich $\operatorname{span}\{x,x'\}$ mit $x'\in\mathscr{C}_\gamma(x)$ ist, gilt $\mu(\hat x)\ge\mu(x')>\mu(x)$ nach Lemma 4.2.

Für $\mu(\hat x)<\mu_i$ ist der Delta-Quotient wegen $\mu_{i+1}<\mu(x)<\mu(\hat x)<\mu_i$ positiv. Zur Bestimmung der oberen Schranke des Delta-Quotienten setzen wir $\tilde\mu=\mu(x)$ gemäß Lemma 4.18 und betrachten $x,\ \bar x\in\mathcal{N}(\tilde\mu)$ sowie die gemäß (4.27) umskalierte Minimalstellen z, $\bar z$ des Rayleigh-Quotienten $\mu(\cdot)$ auf den Mengen $\mathscr{D}_2(x)$, $\mathscr{D}_2(\bar x)$, wobei $\bar z$ eine höherstufige Minimalstelle ist. Es gilt dann $\mu_{i+1}<\mu(x)=\mu(\bar x)<\mu(\bar z)\le\mu(z)\le\mu(\hat x)<\mu_i$, so dass

$$\Delta\big(\mu_i,\mu_{i+1},\mu(x)\big)=\Delta\big(\mu_i,\mu_{i+1},\mu(\bar x)\big),\qquad \Delta\big(\mu_i,\mu_{i+1},\mu(\hat x)\big)\le\Delta\big(\mu_i,\mu_{i+1},\mu(\bar z)\big).$$

Einsetzen in die Abschätzung von Lemma 4.19 werden die restlichen Aussagen bewiesen. □

Für die originale Problemstellung lässt sich diese Abschätzung mittels der Umformulierung in Abschnitt 4.1 rückwärts umschreiben.

Satz 4.21. *Für das Matrixbüschel (A,M) gemäß Voraussetzung 2.1 und einen Vektor $u\in\mathbb{R}^n\backslash\{\mathbf{0}\}$ werde ein Schritt des PINVIT(2)-Verfahrens, vergleiche die Vorschrift (1.3) zu $H=B$, durchgeführt, wobei der Vorkonditionierer B^{-1} die Bedingung (1.4) erfüllt.*

Falls u kein Eigenvektor ist, dann gilt $\rho(u')<\rho(u)$. Gelte zusätzlich $\rho(u)\in(\lambda_i,\lambda_{i+1})$ für ein $i\in\{1,\dots,m-1\}$, dann gilt mit der Bezeichnung (2.4) entweder $\rho(u')\le\lambda_i$ oder

$$0<\frac{\Delta\big(\lambda_i,\lambda_{i+1},\rho(u')\big)}{\Delta\big(\lambda_i,\lambda_{i+1},\rho(u)\big)} \le \left(\frac{\hat\kappa+\gamma(2-\hat\kappa)}{(2-\hat\kappa)+\gamma\hat\kappa}\right)^2 \quad \textit{mit} \quad \hat\kappa := \left(\frac{\lambda_i}{\lambda_{i+1}}\right)\left(\frac{\lambda_m-\lambda_{i+1}}{\lambda_m-\lambda_i}\right).$$

Die obere Schranke der Delta-Abschätzung lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Die schnellste Konvergenz von PINVIT(2) unter der Bedingung $\rho(u)\in(\lambda_i,\lambda_{i+1})$ ist trivial zu untersuchen. Im Unterraum $\mathcal{V}_1\oplus\mathcal{V}_m$ kann man einen Vektor $u=\alpha_1v_1+\alpha_mv_m$ mit Eigenvektoren v_1, v_m finden, welche diese Bedingung erfüllt. Dann lässt sich durch den Durchschnitt des zweidimensionalen Unterraums $\operatorname{span}\{u,\rho(u)A^{-1}Mu\}$ mit dem PINVIT-Kegel zu u einen Vorkonditionier konstruieren, so dass die neue PINVIT(2)-Iterierte durch $\operatorname{span}\{u,\rho(u)A^{-1}Mu\}$ enthalten wird und damit ein Eigenvektor zu λ_1 ist.

Außerdem lassen sich Mehrschritt-Abschätzungen und Shift-Abschätzungen für PINVIT(1) und PINVIT(2) basierend auf den obigen Abschätzungen leicht konstruieren.

5 Vorkonditionierte blockweise Iterationen

Mit den oben untersuchten vorkonditionierten Gradientenverfahren werden einzelne Vektoren als Iterierte berechnet. Zur Berechnung von mehreren Eigenvektor-Näherungen können wir Blockversionen verwenden. Dabei werden Matrizen, deren Spaltenanzahlen gleich oder ein wenig größer als die Anzahl der gewünschten Eigenvektor-Näherungen sind, als Iterierte berechnet. Solche Matrizen bilden dann Unterräume, welche einen invarianten Unterraum approximieren.

Durch Neymeyr [63] wurde eine scharfe Konvergenzabschätzung für das blockweise PINVIT(1)-Verfahren vorgestellt. Diese beruht auf der Analyse für das (verktorweise) PINVIT(1)-Verfahren und ist im Wesentlichen eine spaltenweise Konvergenzabschätzung. Ein Nachteil ist, dass die sogenannte „Clusterrobustheit“ der blockweisen Iterationen (das heißt, die Konvergenzgeschwindigkeit wird nicht wie bei manchen Vektoriterationen wegen der dicht liegenden relevanten Eigenwerte verschlechtert) durch diese Konvergenzabschätzung nicht reflektiert wird. Einige bekannte Cluster-Robust-Abschätzungen für vorkonditionierte blockweise Iterationen, siehe [15, 73], haben aber auch Nachteile, wie zum Beispiel, dass eine nicht sehr natürliche Annahme benötigt wird oder dass der Konvergenzfaktor eine zu komplexe Darstellung hat.

Mit Anregungen der neuen Ergebnisse für vorkonditionierte Gradientenverfahren erhalten wir einige Zwischenergebnisse für vorkonditionierte blockweise Iterationen, welche in diesem Kapitel diskutiert werden. In Abschnitt 5.1 werden die zu untersuchenden vorkonditionierten blockweisen Iterationen definiert und einige geometrische Argumente für PINVIT-Verfahren aus Abschnitt 4.1 verallgemeinert. Die resultierenden Aussagen beziehen sich auf die Begriffe Unterraumwinkel und kanonische Winkel, welche bereits in [91, 10, 45, 46, 47] vorgestellt sowie für verschiedene theoretische Anwendungen untersucht wurden. In Abschnitt 5.2 wird die spaltenweise Konvergenzabschätzung aus [63] vorgestellt und deren Beweis vereinfacht. In Abschnitt 5.3 werden anhand der Analyse aus [79] für die Clusterrobustheit der inversen Unterraumiteration einige neue Argumente zur gewünschten scharfen Clusterrobustheit-Abschätzung für das blockweise PINVIT(1)-Verfahren präsentiert.

5.1 Blockweise PINVIT-Verfahren und Winkelbedingungen

Ein blockweises PINVIT-Verfahren heißt nicht, gemäß dem dafür grundlegenden PINVIT-Verfahren mehrere Vektoriterationen unabhängig voneinander durchzuführen. Ritzwerte und Ritzvektoren sollen dabei wichtige Rollen spielen. Betrachten wir zunächst das blockweise PINVIT(1)-Verfahren gemäß Voraussetzung 2.1. Bei der Initialisierung wird durch die gegebenen Anfangsvektoren eine Basismatrix $\check{U} \in \mathbb{R}^{n\times s}$ konstruiert und damit eine Rayleigh-Ritz-Prozedur durchgeführt. Damit wird eine Diagonalisierung des Matrixbüschels $(\check{U}^T A\check{U}, \check{U}^T M\check{U})$ erhalten, nämlich

$$\check{C}^T(\check{U}^T A\check{U})\check{C} = \Theta = \operatorname{diag}(\theta_1, \ldots, \theta_s), \qquad \check{C}^T(\check{U}^T M\check{U})\check{C} = I,$$

und eine M-orthonormale Ritzbasis $U = \check{U}\check{C}$ von $\operatorname{span}\{U\}$ sowie die entsprechende Ritzwert-Matrix Θ werden konstruiert. Das blockweise PINVIT(1)-Verfahren lässt sich dann durch PINVIT(1) angewandt auf die Spaltenvektoren $u_1, \ldots, u_s$ der Matrix U aufbauen. Mit der Vorschrift (1.5) von PINVIT(1) und den Spaltenvektoren $e_1, \ldots, e_s$ der Einheitsmatrix I erhalten wir

$$u_k - B^{-1}r(u_k) = u_k - B^{-1}\big(Au_k - \rho(u_k)Mu_k\big) = u_k - B^{-1}(Au_k - MU\Theta e_k), \qquad k = 1, \ldots, s$$

als neue Vektor-Iterierte, welche die Matrix

$$\tilde{U} := U - B^{-1}\big(AU - MU\Theta\big) \tag{5.1}$$

bilden. Als Vorbereitung für den nächsten Iterationsschritt wird bezüglich $\tilde{U}$ wieder eine Rayleigh-Ritz-Prozedur durchgeführt, welche eine Ritzbasis U' liefert. Zur Formulierung der Verfahrensvorschrift stellt es sich die Frage, ob die Matrix $\tilde{U}$ wie U den Rang s besitzt. Unter der Bedingung (1.4) des Vorkonditionierers B^{-1} gibt es für diese Frage eine eindeutige Antwort.

Lemma 5.1. *Gemäß Voraussetzung 2.1 sei $U \in \mathbb{R}^{n\times s}$ eine M-orthonormale Ritzbasis und Θ die entsprechende Ritzwert-Matrix. Dann hat die Matrix $\tilde{U}$ aus (5.1) unter der Bedingung (1.4) den gleichen Rang s wie U.*

Beweis. Offenbar hat $\tilde{U}$ höchstens den Rang s. Wäre der Rang von $\tilde{U}$ kleiner als s, dann müsste es einen Nichtnullvektor $c \in \mathbb{R}^s$ mit $\tilde{U}c = \mathbf{0}$ geben. Mit der Umformulierung

$$A^{-1}MU\Theta - (I - B^{-1}A)(A^{-1}MU\Theta - U)$$

von $\tilde{U}$ müsste dann $A^{-1}MU\Theta c = (I - B^{-1}A)(A^{-1}MU\Theta c - Uc)$ gelten. Weiter hat $A^{-1}MU\Theta$ wegen der Regularitäten von A^{-1}, M und Θ (denn A, M sind symmetrisch und positiv definit und Ritzwerte sind damit positiv) den gleichen Rang s wie U, so dass $A^{-1}MU\Theta c \neq \mathbf{0}$ gelten müsste. Gemäß Definition 2.2 ist Uc die A-Projektion von $A^{-1}MU\Theta c$ auf $\operatorname{span}\{U\}$, denn

$$\big(U(U^TAU)^{-1}U^TA\big)(A^{-1}MU\Theta c) = (U\Theta^{-1}U^TA)(A^{-1}MU\Theta c) = U\Theta^{-1}(U^TMU)\Theta c = Uc.$$

Weiter würde unter der Bedingung (1.4) und wegen der Projektionseigenschaft eine Ungleichung

$$\|A^{-1}MU\Theta c\|_A \leq \underbrace{\|I - B^{-1}A\|_A}_{=\mathscr{R}(I-B^{-1}A)\leq\gamma<1} \underbrace{\|A^{-1}MU\Theta c - Uc\|_A}_{\leq\|A^{-1}MU\Theta c\|_A} < \|A^{-1}MU\Theta c\|_A$$

folgen, welche offenbar nicht gilt. Daher hat $\tilde{U}$ den Rang s. □

Nach Lemma 5.1 erhält das blockweise PINVIT(1)-Verfahren die Vorschrift

$$U' = \mathrm{RR}_{\rho,\min,s}\Big(U - B^{-1}\big(AU - MU\Theta\big)\Big), \tag{5.2}$$

wobei die Rayleigh-Ritz-Prozedur $\mathrm{RR}_{\rho,\min,s}$ zum Rayleigh-Quotienten ρ die kleinsten s Ritzwerte und eine zugehörige M-orthonormale Ritzbasis liefern soll. $AU - MU\Theta$ besteht aus den Residuen der Ritzvektoren und lässt sich als ein blockweises Rayleigh-Quotient-Residuum betrachten. Weitere blockweise PINVIT-Verfahren lassen sich gemäß der PINVIT-Hierarchie aus Anschnitt 1.1 aufbauen. Das blockweise PINVIT(2)-Verfahren hat dann die Vorschrift

$$U' = \mathrm{RR}_{\rho,\min,s}\Big(\operatorname{span}\big\{U,\, U - B^{-1}\big(AU - MU\Theta\big)\big\}\Big),$$

und das blockweise PINVIT(k)-Verfahren mit $k > 2$ hat die Vorschrift

$$U' = \mathrm{RR}_{\rho,\min,s}\Big(\operatorname{span}\big\{U_{2-k},\ldots,U_{-1},\, U,\, U - B^{-1}\big(AU - MU\Theta\big)\big\}\Big)$$

mit $k-2$ weiteren vorherigen Iterierten. Für blockweise PINVIT-Verfahren lassen sich einige geometrische Argumente für PINVIT-Verfahren aus Abschnitt 4.1 verallgemeinern. Zunächst liefert die Zwischenvorschrift (5.1) unter der Bedingung (1.4) eine blockweise „Abstand-Bedingung“

$$\big\|\big(A^{-1}MU\Theta - \tilde{U}\big)C\big\|_A \leq \gamma\big\|\big(A^{-1}MU\Theta - U\big)C\big\|_A \quad \text{für beliebiges } C \in \mathbb{R}^{s\times l},\ l \in \{1,\ldots,s\}, \tag{5.3}$$

welche eine Bedingung bezüglich der Unterraumwinkel impliziert.

Definition 5.2. *Gegeben seien eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ und zwei Unterräume $\mathcal{W}, \mathcal{Z}$ von $\mathbb{R}^n$ mit $1 \le \dim \mathcal{W} \le \dim \mathcal{Z}$. Gemäß Definition 2.2 heißt dann*

$$\angle_H(\mathcal{W},\mathcal{Z}) := \max_{w\in\mathcal{W}\setminus\{\mathbf{0}\}} \angle_H(w,\mathcal{Z}) = \max_{w\in\mathcal{W}\setminus\{\mathbf{0}\}} \min_{z\in\mathcal{Z}\setminus\{\mathbf{0}\}} \angle_H(w,z)$$

der H-Winkel zwischen $\mathcal{W}$ und $\mathcal{Z}$. Offenbar haben alle solchen Winkel eine Größe zwischen 0 und $\pi/2$. Im Fall $\dim\mathcal{W} = \dim\mathcal{Z}$ gilt $\angle_H(\mathcal{W},\mathcal{Z}) = \angle_H(\mathcal{Z},\mathcal{W})$. Im Fall $H = I$ werden diese Winkel Euklidische Winkel und werden mit $\angle_2(\cdot,\cdot)$ bezeichnet.

Im Folgenden werden eventuell der Einfachheit halber statt der Bezeichnungen wie $\operatorname{span}\{W\}$ *die Bezeichnungen wie $\langle W\rangle$ verwendet.*

In dieser Definition ist die Voraussetzung $\dim\mathcal{W} \le \dim\mathcal{Z}$ sehr relevant, denn für $\dim\mathcal{W} > \dim\mathcal{Z}$ ist der Durchschnitt von $\mathcal{W}$ mit dem H-orthogonalen Komplement von $\mathcal{Z}$ nach dem Dimensionsvergleich $\dim\mathcal{W} + (n - \dim\mathcal{Z}) > n$ mindestens eindimensional. Mit einem Nichtnullvektor w aus diesem Durchschnitt gilt dann $\angle_H(w,\mathcal{Z}) = \pi/2$, so dass $\angle_H(\mathcal{W},\mathcal{Z}) = \pi/2$ gelten müsste und $\angle_H(\mathcal{W},\mathcal{Z})$ kein sinnvolles Maß für eine Approximationsrelation zwischen $\mathcal{W}$ und $\mathcal{Z}$ wäre.
Gemäß Definition 5.2 liefert (5.3) eine Winkelbedingung

$$\sin\angle_A\big(\langle A^{-1}MU\rangle,\, \langle\tilde{U}\rangle\big) \le \gamma \sin\angle_A\big(\langle A^{-1}MU\rangle,\, \langle U\rangle\big). \tag{5.4}$$

Zum Beweis können wir einen Zusammenhang zwischen dem Sinuswert eines Unterraumwinkels und den entsprechenden Projektionsmatrizen benutzen.

Lemma 5.3. *Gemäß Definition 5.2 gilt mit den H-Projektionsmatrizen P, Q von $\mathcal{W}$, $\mathcal{Z}$, dass*

$$\|(I-Q)P\|_H = \sin\angle_H\big(\mathcal{W},\mathcal{Z}\big).$$

Beweis. Zunächst gilt wegen

$$\max_{y\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|(I-Q)Py\|_H}{\|y\|_H} \le \max_{y\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|(I-Q)Py\|_H}{\|Py\|_H} = \max_{w\in\mathcal{W}\setminus\{\mathbf{0}\}} \frac{\|(I-Q)w\|_H}{\|w\|_H},$$

$$\max_{y\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|(I-Q)Py\|_H}{\|y\|_H} \ge \max_{w\in\mathcal{W}\setminus\{\mathbf{0}\}} \frac{\|(I-Q)Pw\|_H}{\|w\|_H} = \max_{w\in\mathcal{W}\setminus\{\mathbf{0}\}} \frac{\|(I-Q)w\|_H}{\|w\|_H},$$

dass

$$\|(I-Q)P\|_H = \max_{y\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|(I-Q)Py\|_H}{\|y\|_H} = \max_{w\in\mathcal{W}\setminus\{\mathbf{0}\}} \frac{\|(I-Q)w\|_H}{\|w\|_H}.$$

Gemäß Definition 2.2 gilt für beliebiges $w \in \mathcal{W}\setminus\{\mathbf{0}\}$, dass

$$\frac{\|(I-Q)w\|_H}{\|w\|_H} = \sin\angle_H\big(w,\mathcal{Z}\big).$$

Daraus folgt die Behauptung nach Definition 5.2 und der Monotonie der Sinusfunktion auf dem Intervall $[0,\, \pi/2]$. □

Zum Beweis der Implikation „(5.3) ⇒ (5.4)“ haben die relevanten Unterräume eine gleiche Dimensionszahl. Dafür können wir Lemma 5.3 noch spezialisieren.

Lemma 5.4. *Gemäß Definition 5.2 gilt mit beliebigen H-orthonormalen Basismatrizen W, Z von $\mathcal{W}$, $\mathcal{Z}$ und den H-Projektionsmatrizen P, Q von $\mathcal{W}$, $\mathcal{Z}$ im Fall* $\dim\mathcal{W} = \dim\mathcal{Z}$, *dass*

$$\begin{aligned}\sin\angle_H\big(\mathcal{W},\mathcal{Z}\big) &= \|(I-Q)W\|_H = \|(I-Q)P\|_H = \|P-Q\|_H \\ &= \|(I-P)Q\|_H = \|(I-P)Z\|_H = \sin\angle_H\big(\mathcal{Z},\mathcal{W}\big).\end{aligned}$$

Beweis. Analog zum Beweis von Lemma 5.3. □

Nach Lemma 5.4 lässt sich die Implikation „(5.3) ⇒ (5.4)“ mittels der A-Projektionsmatrix des Unterraums $\langle U\rangle$, nämlich

$$P := U(U^TAU)^{-1}U^TA = U\Theta^{-1}U^TA,$$

und einer beliebigen A-orthonormalen Basismatrix Y vom Unterraum $\langle A^{-1}MU\rangle$ beweisen. Da $A^{-1}MU\Theta$ vollrangig und daher eine Basismatrix von $\langle A^{-1}MU\rangle$ ist, gibt es eine reguläre Matrix $C \in \mathbb{R}^{s\times s}$ mit $Y = A^{-1}MU\Theta C$, so dass

$$PY = (U\Theta^{-1}U^TA)(A^{-1}MU\Theta C) = UC.$$

Daraus folgt nach Lemma 5.4 der Zusammenhang

$$\sin\angle_A\big(\langle A^{-1}MU\rangle, \langle U\rangle\big) = \|(I-P)Y\|_A = \big\|A^{-1}MU\Theta C - UC\big\|_A.$$

Mit der Bedingung (5.3) gilt weiter

$$\gamma\sin\angle_A\big(\langle A^{-1}MU\rangle, \langle U\rangle\big) = \gamma\big\|\big(A^{-1}MU\Theta - U\big)C\big\|_A \geq \big\|\big(A^{-1}MU\Theta - \tilde U\big)C\big\|_A = \|Y - \tilde UC\|_A.$$

Außerdem gilt gemäß Definition 5.2, dass

$$\sin\angle_A\big(\langle A^{-1}MU\rangle,\ \langle\tilde U\rangle\big) = \sin\max_{y\in\langle A^{-1}MU\rangle\setminus\{\mathbf{0}\}}\angle_A\big(y, \langle\tilde U\rangle\big).$$

Damit gibt es einen Koeffizientenvektor $\tilde c \in \mathbb{R}^s\setminus\{\mathbf{0}\}$, so dass das obige Maximum an $y = YC^{-1}\tilde c$ angenommen wird. Dann ergibt sich

$$\begin{aligned}\sin\angle_A\big(\langle A^{-1}MU\rangle,\ \langle\tilde U\rangle\big) = \sin\angle_A\big(YC^{-1}\tilde c,\ \langle\tilde U\rangle\big) &= \min_{\tilde u\in\langle\tilde U\rangle}\frac{\|YC^{-1}\tilde c - \tilde u\|_A}{\|YC^{-1}\tilde c\|_A}\\ &\leq \frac{\|YC^{-1}\tilde c - \tilde UC\tilde c\|_A}{\|YC^{-1}\tilde c\|_A} = \frac{\big\|\big(Y-\tilde UC\big)C^{-1}\tilde c\big\|_A}{\|C^{-1}\tilde c\|_A} \leq \|Y-\tilde UC\|_A.\end{aligned}$$

Zusammensetzen der obigen Ungleichungen liefert die Winkelbedingung (5.4). Mit (5.4) lässt sich der Beweis einer nützlichen Zwischenabschätzung bei der in [15] präsentierten Cluster-Robust-Abschätzung teilweise klarer formulieren. Außerdem kann diese Zwischenabschätzung ein wenig verallgemeinert beziehungsweise verbessert werden, vergleiche Lemma 4.2 in [15].

Lemma 5.5. *Gemäß Voraussetzung 2.1 sei $U \in \mathbb{R}^{n\times s}$ eine M-orthonormale Ritzbasis und Θ die entsprechende Ritzwert-Matrix mit den Ritzwerten $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_s$. Weiter sei $\mathcal{V}$ ein durch Eigenvektoren von (A, M) aufgespannter Unterraum mit der Dimension s. Dann gilt für die Matrix $\tilde U$ in (5.1) unter der Bedingung (1.4), dass*

$$\sin\angle_A\big(\mathcal{V}, \langle\tilde U\rangle\big) \leq \Big(1 + (1+\gamma)\theta_s\big(\lambda_1^{-1} - \lambda_m^{-1}\big)\Big)\sin\angle_A\big(\mathcal{V}, \langle U\rangle\big).$$

Beweis. Nach Lemma 5.1 hat $\langle\tilde U\rangle$ die gleiche Dimension wie $\langle U\rangle$, $\langle A^{-1}MU\rangle$ beziehungsweise $\mathcal{V}$. Mit den zugehörigen A-Projektionsmatrizen $P_{\langle\tilde U\rangle}$, $P_{\langle U\rangle}$, $P_{\langle A^{-1}MU\rangle}$, $P_{\mathcal{V}}$ gilt nach Lemma 5.4, dass

$$\begin{aligned}\sin\angle_A\big(\mathcal{V}, \langle\tilde U\rangle\big) &= \|P_{\mathcal{V}} - P_{\langle\tilde U\rangle}\|_A\\ &\leq \|P_{\mathcal{V}} - P_{\langle U\rangle}\|_A + \|P_{\langle U\rangle} - P_{\langle A^{-1}MU\rangle}\|_A + \|P_{\langle A^{-1}MU\rangle} - P_{\langle\tilde U\rangle}\|_A\\ &= \sin\angle_A\big(\mathcal{V}, \langle U\rangle\big) + \sin\angle_A\big(\langle A^{-1}MU\rangle, \langle U\rangle\big) + \sin\angle_A\big(\langle A^{-1}MU\rangle, \langle\tilde U\rangle\big).\end{aligned}$$

Mit der Winkelbedingung (5.4) gilt also

$$\sin\angle_A\big(\mathcal{V},\langle\tilde{U}\rangle\big) \le \sin\angle_A\big(\mathcal{V},\langle U\rangle\big) + (1+\gamma)\sin\angle_A\big(\langle A^{-1}MU\rangle,\langle U\rangle\big).$$

Mittels einer Maximalstelle $\tilde{c}$ der Funktion $\varphi(c) := \sin\angle_A\big(A^{-1}MUc,\langle U\rangle\big)$ auf $c \in \mathbb{R}^s\backslash\{\mathbf{0}\}$ lässt sich $\sin\angle_A\big(\langle A^{-1}MU\rangle,\langle U\rangle\big)$ nach oben abschätzen. Mit der Bezeichnung $y := U\tilde{c}$ gilt

$$\begin{aligned}\sin\angle_A\big(\langle A^{-1}MU\rangle,\langle U\rangle\big) &= \sin\angle_A\big(A^{-1}MU\tilde{c},\langle U\rangle\big) = \sin\angle_A\big(A^{-1}My,\langle U\rangle\big)\\ &= \frac{\big\|\big(I-P_{\langle U\rangle}\big)A^{-1}My\big\|_A}{\|A^{-1}My\|_A} = \frac{\big\|\big(I-P_{\langle U\rangle}\big)(A^{-1}M)P_{\langle U\rangle}y\big\|_A}{\|A^{-1}My\|_A}\\ &\le \big\|\big(I-P_{\langle U\rangle}\big)(A^{-1}M)P_{\langle U\rangle}\big\|_A\left(\frac{\|y\|_A}{\|A^{-1}My\|_A}\right).\end{aligned}$$

Für den Quotienten $\|y\|_A/\|A^{-1}My\|_A$ wird wegen $y = U\tilde{c} \in \langle U\rangle$ eine obere Schranke durch den größten Ritzwert θ_s zu $\langle U\rangle$ gegeben:

$$\frac{\|y\|_A}{\|A^{-1}My\|_A} = \frac{\|y\|_A^2}{\|y\|_A\|A^{-1}My\|_A} \le \frac{\|y\|_A^2}{(y,A^{-1}My)_A} = \frac{y^TAy}{y^TMy} = \rho(y) \le \theta_s,$$

wobei eine Cauchy-Schwarz-Ungleichung verwendet wird. Anschließend ist die Operatornorm $\big\|\big(I-P_{\langle U\rangle}\big)(A^{-1}M)P_{\langle U\rangle}\big\|_A$ nach oben abzuschätzen. Dafür nutzen wir die A-Projektionsmatrix $P_{\mathcal{V}}$ und zeigen zunächst, dass $P_{\mathcal{V}}$ und $A^{-1}M$ bezüglich der Multiplikation kommutativ sind. Durch die $\mathcal{V}$ erzeugenden Eigenvektoren lässt sich eine A-orthonormale Basismatrix $\tilde{V} \in \mathbb{R}^{n\times s}$ konstruieren und es gilt $A\tilde{V} = M\tilde{V}\tilde{\Lambda}$ mit einer Diagonalmatrix $\tilde{\Lambda} \in \mathbb{R}^{s\times s}$, deren Diagonalkomponenten Eigenwerte zu den Spalten von $\tilde{V}$ sind. Da alle Eigenwerte von (A,M) positiv sind, ist $\tilde{\Lambda}$ invertierbar. Weiter gilt wegen $P_{\mathcal{V}} = \tilde{V}\tilde{V}^TA$ und $A\tilde{V} = M\tilde{V}\tilde{\Lambda}$, dass

$$\begin{aligned}(A^{-1}M)P_{\mathcal{V}} &= A^{-1}M\tilde{V}\tilde{V}^TA = \tilde{V}\tilde{\Lambda}^{-1}\tilde{V}^TA = \tilde{V}(\tilde{V}\tilde{\Lambda}^{-1})^TA\\ &= \tilde{V}(A^{-1}M\tilde{V})^TA = \tilde{V}\tilde{V}^TM = \tilde{V}\tilde{V}^T(AA^{-1})M = P_{\mathcal{V}}(A^{-1}M).\end{aligned}$$

Für ein beliebiges $\sigma \in \mathbb{R}$ gilt dann (wegen $P_{\langle U\rangle}^2 = P_{\langle U\rangle}$ und der obigen Kommutativität)

$$\begin{aligned}&\big(I-P_{\langle U\rangle}\big)(A^{-1}M)P_{\langle U\rangle} = \big(I-P_{\langle U\rangle}\big)\big(A^{-1}M-\sigma I\big)P_{\langle U\rangle}\\ &= \big(I-P_{\langle U\rangle}\big)\big(A^{-1}M-\sigma I\big)\big(P_{\mathcal{V}}+(I-P_{\mathcal{V}})\big)P_{\langle U\rangle}\\ &= \big(I-P_{\langle U\rangle}\big)P_{\mathcal{V}}\big(A^{-1}M-\sigma I\big)P_{\langle U\rangle} + \big(I-P_{\langle U\rangle}\big)\big(A^{-1}M-\sigma I\big)(I-P_{\mathcal{V}})P_{\langle U\rangle},\end{aligned}$$

so dass $\big\|\big(I-P_{\langle U\rangle}\big)(A^{-1}M)P_{\langle U\rangle}\big\|_A$ durch

$$\big\|\big(I-P_{\langle U\rangle}\big)P_{\mathcal{V}}\big\|_A\big\|A^{-1}M-\sigma I\big\|_A\big\|P_{\langle U\rangle}\big\|_A + \big\|I-P_{\langle U\rangle}\big\|_A\big\|A^{-1}M-\sigma I\big\|_A\big\|(I-P_{\mathcal{V}})P_{\langle U\rangle}\big\|_A \quad (5.5)$$

von oben beschränkt wird. Dazu gilt nach Lemma 5.4, dass

$$\big\|\big(I-P_{\langle U\rangle}\big)P_{\mathcal{V}}\big\|_A = \big\|(I-P_{\mathcal{V}})P_{\langle U\rangle}\big\|_A = \sin\angle_A\big(\mathcal{V},\langle U\rangle\big).$$

Außerdem gilt für die Projektionsmatrix $P_{\langle U\rangle}$ wegen $\|P_{\langle U\rangle}w\|_A \le \|w\|_A$ für beliebiges $w \in \mathbb{R}^n$ und $\|P_{\langle U\rangle}u\|_A = \|u\|_A$ für beliebiges $u \in \langle U\rangle$, dass $\|P_{\langle U\rangle}\|_A = 1$, und analog $\|I-P_{\langle U\rangle}\|_A = 1$. Zusammen mit der Darstellung (5.5) wird gezeigt, dass

$$\big\|\big(I-P_{\langle U\rangle}\big)(A^{-1}M)P_{\langle U\rangle}\big\|_A \le 2\big\|A^{-1}M-\sigma I\big\|_A\sin\angle_A\big(\mathcal{V},\langle U\rangle\big).$$

Zum Schluss ist $\|A^{-1}M - \sigma I\|_A$ abzuschätzen. Dafür betrachten wir ein beliebiges $w \in \mathbb{R}^n$ und dessen Darstellung $w = \sum_{i=1}^m V_i c_i$ mit A-orthonormalen Basismatrizen der Eigenräume. Es gilt

$$\|(A^{-1}M - \sigma I)w\|_A = \left(\sum_{i=1}^m (\lambda_i^{-1} - \sigma)^2 \|c_i\|_2^2\right)^{1/2} \leq \left(\max_{i=1,\dots,m} |\lambda_i^{-1} - \sigma|\right) \|w\|_A.$$

Dabei gilt die Gleichheit, falls w ein Eigenvektor zu einem Eigenwert mit maximalem Wert von $|\lambda_i^{-1} - \sigma|$ ist. Daher gilt $\|A^{-1}M - \sigma I\|_A = \max_{i=1,\dots,m} |\lambda_i^{-1} - \sigma|$. Weiter lässt sich ein optimales σ zur Minimierung dieser Operatornorm (mittels einfacher Fallunterscheidung) bestimmen, nämlich der Mittelpunkt des Intervalls $[\lambda_m^{-1}, \lambda_1^{-1}]$. Die optimale Schranke wird dann durch die Hälfte der Intervalllänge gegeben, so dass die Behauptung bewiesen wird. □

Bei Lemma 4.2 in [15] wurde der Term λ_m^{-1} vernachlässigt. Ein vermutlicher Grund ist etwa, dass λ_m bei den Modellproblemen oft sehr groß ist. Es ist zu bemerken, dass diese Zwischenabschätzung keine sinnvolle Schranke von $\sin \angle_A(\mathcal{V}, \langle \tilde{U} \rangle)$ liefert, denn der Faktor $1 + (1+\gamma)\theta_s(\lambda_1^{-1} - \lambda_m^{-1})$ ist offenbar größer als 1. Trotzdem lassen sich sinnvolle Abschätzungen der Sinuswerte von einigen Vektor-Untteraum-Winkeln mittels dieser Zwischenabschätzung unter zusätzlichen technischen Annahmen entwickeln, siehe [15] für Details.

Im Folgenden wird eine Verallgemeinerung der Winkelbedingung (5.4) vorgestellt, welche sich auf die entsprechenden kanonischen Winkel bezieht und möglicherweise bei der Clusterrobustheit-Analyse für blockweise PINVIT-Verfahren nützlich ist.

Definition 5.6. *Gegeben seien eine symmetrische und positiv definite Matrix $H \in \mathbb{R}^{n\times n}$ und H-orthonormale Basismatrizen W, $Z \in \mathbb{R}^{n\times s}$ $(1 \leq s \leq n)$ der Unterräume $\langle W \rangle$, $\langle Z \rangle$, sowie die Singulärwerte $\sigma_1 \leq \dots \leq \sigma_s$ der Matrix $Z^T H W$, dann heißen $\varphi_k := \arccos(\sigma_{s+1-k})$, $k = 1, \dots, s$ die H-kanonischen Winkel zwischen $\langle W \rangle$ und $\langle Z \rangle$.*

Es gilt $0 \leq \varphi_1 \leq \dots \leq \varphi_s \leq \pi/2$, denn die Singulärwerte sind sämtlich nicht-negativ. Kanonische Winkel können auch wie bei Definition 5.2 mittels Vektor-Unteraum-Winkel definiert werden.

Lemma 5.7. *Gemäß Definition 5.6 gilt*

$$\varphi_k = \min_{\substack{D\in\mathbb{R}^{s\times k},\\ Rang(D)=k}} \angle_H\big(\langle WD\rangle, \langle Z\rangle\big) = \min_{\substack{D\in\mathbb{R}^{s\times k},\\ Rang(D)=k}} \angle_H\big(\langle W\rangle, \langle ZD\rangle\big).$$

Diese Aussage gilt auch, wenn W, Z beliebige Basismatrizen der entsprechenden Unterräume sind.

Beweis. Wir zeigen diese Aussage zunächst für H-orthonormale Basismatrizen. Für ein beliebiges $d \in \mathbb{R}^s \backslash \{\mathbf{0}\}$ gilt mit der H-Projektionsmatrix ZZ^TH von $\langle Z \rangle$ und der die H-kanonischen Winkel erzeugenden Matrix $C := Z^T H W$, dass

$$\cos^2 \angle_H\big(Wd, \langle Z\rangle\big) = \cos^2 \angle_H\big(Wd, ZZ^THWd\big) = \frac{\big((Wd)^T H(ZZ^THWd)\big)^2}{\|Wd\|_H^2 \|ZZ^THWd\|_H^2} = \frac{d^T(C^TC)d}{d^Td},$$

das heißt, $\cos^2 \angle_H\big(Wd, \langle Z\rangle\big)$ ist gleich dem Rayleigh-Quotienten von d bezüglich der Matrix C^TC. Da C^TC die Eiegenwerte $\sigma_1^2 \leq \dots \leq \sigma_s^2$ hat, gilt nach dem Courant-Fischer-Prinzip, dass

$$\sigma_{s+1-k}^2 = \max_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \min_{d\in\langle D\rangle\backslash\{\mathbf{0}\}} \frac{d^T(C^TC)d}{d^Td} = \max_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \min_{d\in\langle D\rangle\backslash\{\mathbf{0}\}} \cos^2 \angle_H\big(Wd, \langle Z\rangle\big)$$

$$\Rightarrow \quad \sigma_{s+1-k} = \max_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \min_{e\in\mathbb{R}^k\backslash\{\mathbf{0}\}} \cos \angle_H\big(WDe, \langle Z\rangle\big)$$

$$\Rightarrow \quad \varphi_k = \arccos(\sigma_{s+1-k}) = \min_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \max_{e\in\mathbb{R}^k\backslash\{\mathbf{0}\}} \angle_H\big(WDe, \langle Z\rangle\big) = \min_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \angle_H\big(\langle WD\rangle, \langle Z\rangle\big).$$

Analog wird die zweite Darstellung mittels der Betrachtung von $\angle_H\big(\langle W\rangle, Zd\big)$ hergeleitet.

Wenn W, Z beliebige Basismatrizen sind, dann gibt es reguläre Matrizen $E_1, E_2 \in \mathbb{R}^{s\times s}$, so dass WE_1, ZE_2 H-orthonormale Basismatrizen sind. Weiter lässt sich die Aussage zu W und Z durch einfache Umformulierungen aus der Aussage zu WE_1, ZE_2 herleiten. □

Mit Lemma 5.7 können wir die Winkelbedingung (5.4) auf kanonische Winkel übertragen.

Satz 5.8. *Gemäß Voraussetzung 2.1 sei $U \in \mathbb{R}^{n\times s}$ eine M-orthonormale Ritzbasis und Θ die entsprechende Ritzwert-Matrix. Dann erfüllt die Matrix $\tilde{U}$ aus (5.1) unter der Bedingung (1.4) die Ungleichungen*

$$\sin\varphi_k \le \gamma \sin\psi_k, \qquad k = 1,\dots,s, \tag{5.6}$$

wobei $\varphi_1 \le \cdots \le \varphi_s$ die A-kanonischen Winkel zwischen $\langle A^{-1}MU\rangle$, $\langle\tilde{U}\rangle$ sind und $\psi_1 \le \cdots \le \psi_s$ die A-kanonischen Winkel zwischen $\langle A^{-1}MU\rangle$, $\langle U\rangle$ sind.

Beweis. Als Vorbereitung zeigen wir zunächst die Sinus-Ungleichung

$$\sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle\tilde{U}\rangle\big) \le \gamma \sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle U\rangle\big) \tag{5.7}$$

für beliebiges vollrangiges $D \in \mathbb{R}^{s\times k}$ mit $k \in \{1,\dots,s\}$. Wegen $\dim\langle A^{-1}MU\Theta D\rangle = k \le s = \dim\langle\tilde{U}\rangle$ gilt nach Definition 5.2 und der Monotonie der Sinusfunktion auf Intervall $[0,\pi/2]$, dass

$$\sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle\tilde{U}\rangle\big) = \max_{c\in\mathbb{R}^k\setminus\{\mathbf{0}\}} \sin\angle_A\big(A^{-1}MU\Theta Dc, \langle\tilde{U}\rangle\big).$$

Es gibt also einen Koeffizientenvektor $\tilde{c} \in \mathbb{R}^k\setminus\{\mathbf{0}\}$, mit dem das obige Maximum angenommen wird und anschließend

$$\begin{aligned}&\sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle\tilde{U}\rangle\big) = \sin\angle_A\big(A^{-1}MU\Theta D\tilde{c}, \langle\tilde{U}\rangle\big)\\ =\ &\min_{\tilde{u}\in\langle\tilde{U}\rangle\setminus\{\mathbf{0}\}} \sin\angle_A\big(A^{-1}MU\Theta D\tilde{c}, \langle\tilde{u}\rangle\big) \le \sin\angle_A\big(A^{-1}MU\Theta D\tilde{c}, \langle\tilde{U}D\tilde{c}\rangle\big)\end{aligned} \tag{5.8}$$

(gemäß Definition 2.2) gilt. Unter der Bedingung (1.4) liefert die Multiplikation der Darstellung (5.1) mit $D\tilde{c}$, dass

$$\big\|A^{-1}MU\Theta D\tilde{c} - \tilde{U}D\tilde{c}\big\|_A \le \gamma\big\|A^{-1}MU\Theta D\tilde{c} - UD\tilde{c}\big\|_A.$$

Mit $U^TAU = \Theta$ und $U^TMU = I$ lässt sich zeigen, dass $UD\tilde{c}$ die A-Projektion von $A^{-1}MU\Theta D\tilde{c}$ auf $\langle U\rangle$ ist. Damit gilt

$$\sin\angle_A\big(A^{-1}MU\Theta D\tilde{c}, \langle U\rangle\big) = \frac{\big\|A^{-1}MU\Theta D\tilde{c} - UD\tilde{c}\big\|_A}{\big\|A^{-1}MU\Theta D\tilde{c}\big\|_A}.$$

Mit der obigen Abstand-Bedingung wird anschließend die Relation

$$\sin\angle_A\big(A^{-1}MU\Theta D\tilde{c}, \langle\tilde{U}D\tilde{c}\rangle\big) \le \gamma\sin\angle_A\big(A^{-1}MU\Theta D\tilde{c}, \langle U\rangle\big) \le \gamma\sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle U\rangle\big)$$

(gemäß Definition 5.2) gezeigt. Kombiniert mit (5.8) liefert das die Ungleichung (5.7).

Da $A^{-1}MU\Theta$ eine Basismatrix von $\langle A^{-1}MU\rangle$ ist, gilt nach (5.7) und Lemma 5.7 schließlich

$$\sin\varphi_k = \sin\min_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle\tilde{U}\rangle\big) = \min_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle\tilde{U}\rangle\big)$$

$$\le \min_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \gamma\sin\angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle U\rangle\big) = \gamma\sin\min_{\substack{D\in\mathbb{R}^{s\times k},\\ \mathrm{Rang}(D)=k}} \angle_A\big(\langle A^{-1}MU\Theta D\rangle, \langle U\rangle\big) = \gamma\sin\psi_k$$

für jedes $k \in \{1,\dots,s\}$. □

5.2 Spaltenweise Konvergenzabschätzungen

Basierend auf der Zwischenvorschrift (5.1) lassen sich spaltenweise Konvergenzabschätzungen für das blockweise PINVIT(1)-Verfahren entwickeln. Die Spaltenvektoren von $\tilde{U}$ sind wesentlich durch PINVIT(1) mit den Ritzvektoren aus dem Unterraum $\operatorname{span}\{U\}$ konstruiert, so dass die Delta-Abschätzungen bezüglich der Rayleigh-Quotienten der Spaltenvektoren von $\tilde{U}$ sofort formulierbar sind. Jedoch sind diese Rayleigh-Quotienten normalerweise keine Ritzwerte zu $\operatorname{span}\{\tilde{U}\}$. Um Delta-Abschätzungen der Ritzwerte zu beweisen, benötigen wir gewisse Zwischenvektoren. Im Folgenden formulieren wir eine spaltenweise Konvergenzabschätzung aus [63] mit einem vereinfachten Beweis.

Satz 5.9. *Gemäß Voraussetzung 2.1 sei $U = [u_1, \dots, u_s] \in \mathbb{R}^{n\times s}$ eine M-orthonormale Ritzbasis und Θ die entsprechende Ritzwert-Matrix mit den Ritzwerten $\theta_1 \le \dots \le \theta_s$. Weiter sei $\tilde{U}$ durch (5.1) und unter der Bedingung (1.4) konstruiert. Durch $\theta'_1 \le \dots \le \theta'_s$ werden die Ritzwerte von (A, M) bezüglich $\operatorname{span}\{\tilde{U}\}$ gegeben.*

Gelte $\theta_k \in (\lambda_i, \lambda_{i+1})$ für ein $k \in \{1, \dots, s\}$ und ein $i \in \{1, \dots, m-1\}$, dann gilt mit der Bezeichnung (2.4) entweder $\theta'_k \le \lambda_i$ oder

$$0 < \frac{\Delta\big(\lambda_i, \lambda_{i+1}, \theta'_k\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \theta_k\big)} \le \left((1-\gamma)\frac{\lambda_i}{\lambda_{i+1}} + \gamma\right)^2.$$

Die obere Schranke der Delta-Abschätzung lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Zeigen wir diese Aussage zunächst für den Index $k = s$. Konstruieren wir dafür einen Ritzvektor u'_s zum Ritzwert θ'_s gemäß (5.1). Wegen $u'_s \in \operatorname{span}\{\tilde{U}\}\backslash\{\mathbf{0}\}$ lässt sich u'_s mit einem $c \in \mathbb{R}^s\backslash\{\mathbf{0}\}$ durch $u'_s = \tilde{U}c$ darstellen. Mit (5.1) gilt dann

$$u'_s = Uc - B^{-1}\big(AU - MU\Theta\big)c,$$

welches mit der Bezeichnung $\breve{u} := U\Theta c$ in

$$A^{-1}M\breve{u} - u'_s = (I - B^{-1}A)(A^{-1}M\breve{u} - Uc)$$

umformuliert werden kann. Da θ_s der größte Ritzwert und damit das Maximum des Rayleigh-Quotienten auf $\operatorname{span}\{U\}$ ist, gilt $\theta_s \ge \rho(\breve{u})$. Aus der Bedingung $\theta_s \in (\lambda_i, \lambda_{i+1})$ folgt dann entweder $\rho(\breve{u}) \le \lambda_i$ oder $\rho(\breve{u}) \in (\lambda_i, \lambda_{i+1})$. Außerdem lässt sich mit $U^TAU = \Theta$ und $U^TMU = I$ die A-Orthogonalität $A^{-1}M\breve{u} - Uc \perp_A U$ zeigen, so dass

$$\|A^{-1}M\breve{u} - Uc\|_A = \min_{u\in\operatorname{span}\{U\}} \|A^{-1}M\breve{u} - u\|_A \le \|A^{-1}M\breve{u} - \breve{u}/\rho(\breve{u})\|_A$$

gilt. Mit der Bedingung (1.4) wird anschließend

$$\|A^{-1}M\breve{u} - u'_s\|_A \le \gamma\|A^{-1}M\breve{u} - Uc\|_A \le \gamma\|A^{-1}M\breve{u} - \breve{u}/\rho(\breve{u})\|_A$$

gezeigt. Damit gehört u'_s zu einer A-Kugel mit dem Mittelpunkt $A^{-1}M\breve{u}$. Seien x'_s und $\breve{x}$ die Koeffizientenvektoren von u'_s und $\breve{u}$ bezüglich einer aus A-orthonormalen Basismatrizen der Eigenräume von (A, M) bestehenden Matrix $V \in \mathbb{R}^{n\times n}$, dann gilt $x'_s \in \mathscr{B}_\gamma(\breve{x})$ gemäß der geometrischen Interpretation in Voraussetzung 4.1. Da die Delta-Abschätzungen bei Umskalierungen der Vektoren ungeändert bleiben, können wir weiter $\breve{x}'_s := x'_s/\mu(\breve{x})$ betrachten, welches wegen $\mu(\breve{x})\breve{x}'_s \in \mathscr{B}_\gamma(\breve{x})$ die Voraussetzung 4.1 bezüglich $\breve{x}$ erfüllt. So sind die in Abschnitt 4.2 bewiesenen Aussagen anwendbar. Nach Satz 4.8 und einem Rückwärtseinsetzen erhalten wir folgende Aussagen: a) Im Fall $\rho(\breve{u}) \le \lambda_i$ gilt $\rho(u'_s) \le \rho(\breve{u}) \le \lambda_i$. b) Im Fall $\rho(\breve{u}) \in (\lambda_i, \lambda_{i+1})$ gilt entweder $\rho(u'_s) \le \lambda_i$ oder

$$0 < \frac{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(u'_s)\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \rho(\breve{u})\big)} \le \left((1-\gamma)\frac{\lambda_i}{\lambda_{i+1}} + \gamma\right)^2$$

mit einer scharfen oberen Schranke. Das zeigt mit $\theta'_s = \rho(u'_s)$ und $\theta_s \geq \rho(\check{u})$ die behauptete Aussage für $k = s$. Die obere Schranke lässt sich erreichen, wenn zum Beispiel die ersten $s-1$ Spalten von U bereits Eigenvektoren sind und u_s durch einen zweidimensionalen invarianten Unterraum zu den Eigenwerten λ_i, λ_{i+1} enthalten wird.

Weiter zeigen wir die Aussage für einen beliebigen Index $k \in \{1, \ldots, s-1\}$. Dazu betrachten wir die Untermatrix $U_k := [u_1, \ldots, u_k]$ von U und konstruieren mit $\Theta_k := U_k^T A U_k = \operatorname{diag}(\theta_1, \ldots, \theta_k)$ gemäß (5.1) die Matrix

$$\tilde{U}_k := U_k - B^{-1}\big(AU_k - MU_k\Theta_k\big).$$

So können wir den ersten Teil des Beweises auf U_k und $\tilde{U}_k$ übertragen und erhalten für den größten Ritzwert $\tilde{\theta}_k$ von (A, M) bezüglich $\operatorname{span}\{\tilde{U}_k\}$ die folgende Aussage: Im Fall $\theta_k \in (\lambda_i, \lambda_{i+1})$ gilt entweder $\tilde{\theta}_k \leq \lambda_i$ oder

$$0 < \frac{\Delta\big(\lambda_i, \lambda_{i+1}, \tilde{\theta}_k\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \theta_k\big)} \leq \left((1-\gamma)\frac{\lambda_i}{\lambda_{i+1}} + \gamma\right)^2$$

mit einer scharfen oberen Schranke.

Außerdem ist $\tilde{U}_k$ eine Untermatrix von $\tilde{U}$ und $\operatorname{span}\{\tilde{U}_k\}$ ein Unterraum von $\operatorname{span}\{\tilde{U}\}$. Nach dem Courant-Fischer-Prinzip (siehe Satz 2.4) gilt dann $\theta'_k \leq \tilde{\theta}_k$, so dass die Aussage auch bezüglich θ'_k gezeigt wird. Die obere Schranke lässt sich erreichen, wenn zum Beispiel u_k durch einen zweidimensionalen invarianten Unterraum zu den Eigenwerten λ_i, λ_{i+1} enthalten wird und die anderen $s-1$ Spalten von U bereits Eigenvektoren sind. □

Für den Index $k = 1$ gibt es eine alternative Beweisidee, welche auch auf das blockweise PINVIT(2)-Verfahren anwendbar ist.

Lemma 5.10. *Gemäß Voraussetzung 2.1 sei $U = [u_1, \ldots, u_s] \in \mathbb{R}^{n\times s}$ eine M-orthonormale Ritzbasis und Θ die entsprechende Ritzwert-Matrix mit den Ritzwerten $\theta_1 \leq \cdots \leq \theta_s$. Durch das blockweise PINVIT(2)-Verfahren werde dann ein Unterraum $\mathcal{U}' := \operatorname{span}\big\{U, U - B^{-1}\big(AU - MU\Theta\big)\big\}$ konstruiert. Sei θ'_1 der kleinste Ritzwert von (A, M) bezüglich $\mathcal{U}'$, dann gilt im Fall $\theta_1 \in (\lambda_i, \lambda_{i+1})$, $i \in \{1, \ldots, m-1\}$ mit der Bezeichnung (2.4) entweder $\theta'_1 \leq \lambda_i$ oder*

$$0 < \frac{\Delta\big(\lambda_i, \lambda_{i+1}, \theta'_1\big)}{\Delta\big(\lambda_i, \lambda_{i+1}, \theta_1\big)} \leq \left(\frac{\hat{\kappa} + \gamma(2-\hat{\kappa})}{(2-\hat{\kappa}) + \gamma\hat{\kappa}}\right)^2 \quad \textit{mit} \quad \hat{\kappa} := \left(\frac{\lambda_i}{\lambda_{i+1}}\right)\left(\frac{\lambda_m - \lambda_{i+1}}{\lambda_m - \lambda_i}\right).$$

Die obere Schranke der Delta-Abschätzung lässt sich in einem Grenzfall erreichen. (Damit ist diese Abschätzung scharf.)

Beweis. Betrachten wir die neue PINVIT(2)-Iterierte zu u_1, nämlich

$$w = \mathrm{RR}_{\rho,\min}(\mathcal{W}) \quad \text{mit} \quad \mathcal{W} := \operatorname{span}\{u_1,\, u_1 - B^{-1}(Au_1 - \rho(u_1)Mu_1)\}.$$

Wegen $\rho(u_1) = \theta_1$ gilt dann nach Satz 4.21 eine gleichartige Aussage wie die Behauptung, aber mit $\rho(w)$ statt θ'_1. Wegen $\rho(u_1) = \theta_1$ gilt außerdem, dass $u_1 - B^{-1}(Au_1 - \rho(u_1)Mu_1)$ eine Spalte von $U - B^{-1}\big(AU - MU\Theta\big)$ und damit $\mathcal{W}$ ein Unterraum von $\mathcal{U}'$ ist, daher gilt $w \in \mathcal{U}'$. Da θ'_1 der kleinste Ritzwert und damit das Minimum des Rayleigh-Quotienten auf $\mathcal{U}'$ ist, gilt $\theta'_1 \leq \rho(w)$, so dass die Aussage auch bezüglich θ'_1 gezeigt wird. Die obere Schranke lässt sich erreichen, wenn zum Beispiel u_1 durch einen zweidimensionalen invarianten Unterraum zu den Eigenwerten λ_i, λ_{i+1} enthalten wird und die anderen $s-1$ Spalten von U bereits Eigenvektoren sind. □

Für die Indizes $k = 2, \ldots, s$ lässt sich zur Zeit nur grobe Abschätzungen formulieren, nämlich mittels der Übertragung der Abschätzungen für blockweise PINVIT(1) nach dem Courant-Fischer-Prinzip. Natürlich werden durch die Abschätzungen für blockweise PINVIT(1) ebenfalls grobe Abschätzungen für weitere blockweise PINVIT-Verfahren ermöglicht.

5.3 Über Cluster-Robust-Konvergenzabschätzungen

Für manche klassische blockweise Verfahren sind Cluster-Robust-Konvergenzabschätzungen schon seit langem bekannt. Als ein Beispiel betrachten wir eine Umformulierung der Abschätzung in Theorem 14.4.1 von Parlett [79] für die inverse Unterraumiteration.

Satz 5.11. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische und positiv definite Matrix mit Eigenwerten $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s < \alpha_{s+1} \leq \cdots \leq \alpha_n$ und $Z := [z_1, \ldots, z_s] \in \mathbb{R}^{n\times s}$ eine Euklidisch orthonormale Matrix, deren Spaltenvektoren zugehörige Eigenvektoren von $\alpha_1, \ldots, \alpha_s$ sind.*

Dann gilt für den Unterraum $\mathcal{Z} := \operatorname{span}\{Z\}$ und einen s-dimensionalen Unterraum $\mathcal{U}$ von $\mathbb{R}^n$ mit $\tan\angle_2(\mathcal{Z},\mathcal{U}) < \infty$, dass

$$\tan\angle_2(z_i, A^{-k}\mathcal{U}) \leq \left(\frac{\alpha_i}{\alpha_{s+1}}\right)^k \tan\angle_2(\mathcal{Z},\mathcal{U}) \qquad \forall\, i \in \{1,\ldots,s\}.$$

Mit den Formulierung in [79] wurde statt $\mathbb{R}^n$ ein abstrakter n-dimensionaler Innenproduktraum betrachtet und A sollte ein selbstadjungierter Operator sein. Der Unterschied ist nur „oberflächlich“, denn mittels einer gewissen Basis lässt sich ein äquivalentes Problem zu einer hermiteschen Matrix stellen, und Satz 5.11 lässt sich auch für eine hermitesche Matrix leicht umformulieren und in gleicher Weise beweisen.

Im originalen Beweis in [79] wurde eine blockweise orthogonale Aufspaltung $U = ZC + JS$ mit zwei Diagonalmatrizen C und S benutzt, deren Komponenten jeweils Kosinuswerte und Sinuswerte der kanonischen Winkel zwischen $\mathcal{Z}$ und $\mathcal{U}$ sind. Es ist zu bemerken, dass bei der zu zeigenden Abschätzung nur ein kanonischer Winkel, nämlich der Unterraumwinkel, von Interesse ist. Daher ist die Betrachtung von C und S möglicherweise nicht sehr relevant. Außerdem wurde ein Zwischenvektor, welcher die wesentlich wichtige Rolle des Beweises spielte, relativ technisch konstruiert. Gemäß diesen Bemerkungen formulieren wir einen vereinfachten Beweis.

Beweis. (von Satz 5.11) Benutzen wir eine Euklidisch orthonormale Matrix $W := [z_{s+1}, \ldots, z_n] \in \mathbb{R}^{n\times(n-s)}$, deren Spaltenvektoren Eigenvektoren zu den Eigenwerten $\alpha_{s+1}, \ldots, \alpha_n$ sind. Da die Eigenwert-Mengen $\{\alpha_1, \ldots, \alpha_s\}$ und $\{\alpha_{s+1}, \ldots, \alpha_n\}$ disjunkt sind, ist $\mathcal{W} := \operatorname{span}\{W\}$ das Euklidisch orthogonale Komplement von $\mathcal{Z}$. Dann hat der Unterraum $\tilde{\mathcal{W}}_i := \operatorname{span}\{z_i, W\}$ für beliebiges $i \in \{1,\ldots,s\}$ die Dimension $n-s+1$, so dass der Unterraum $\tilde{\mathcal{W}}_i \cap \mathcal{U}$ wegen

$$\dim\tilde{\mathcal{W}}_i + \dim\mathcal{U} = (n-s+1) + s = n+1 > n$$

mindestens eindimensional ist und es ein $u \in \mathcal{U}\backslash\{\mathbf{0}\}$ mit $u \in \tilde{\mathcal{W}}$ gibt. Weiter lässt sich u durch $u = \beta_i z_i + \sum_{j=s+1}^n \beta_j z_j$ darstellen. Es gilt $\beta_i \neq 0$, denn sonst würde $\angle_2(u,\mathcal{Z}) = \pi/2$ und damit $\angle_2(\mathcal{Z},\mathcal{U}) = \pi/2$ gelten, was der Voraussetzung $\tan\angle_2(\mathcal{Z},\mathcal{U}) < \infty$ widerspricht. Weiter ist $\beta_i z_i$ die Euklidische Projektion von u sowohl auf $\mathcal{Z}$, als auch auf $\operatorname{span}\{z_i\}$. Gemäß Definition 2.2 gilt dann

$$\tan\angle_2(u,\operatorname{span}\{z_i\}) = \tan\angle_2(u,\mathcal{Z}) = \frac{\|u-\beta_i z_i\|_2}{\|\beta_i z_i\|_2} = \frac{\left\|\sum_{j=s+1}^n \beta_j z_j\right\|_2}{\|\beta_i z_i\|_2} = \frac{\sqrt{\sum_{j=s+1}^n \beta_j^2}}{|\beta_i|}.$$

Analog gilt für $A^{-k}u$ mit der Darstellung $A^{-k}u = \alpha_i^{-k}\beta_i z_i + \sum_{j=s+1}^n \alpha_j^{-k}\beta_j z_j$, dass

$$\tan\angle_2(A^{-k}u,\operatorname{span}\{z_i\}) = \tan\angle_2(A^{-k}u,\mathcal{Z}) = \frac{\left\|\sum_{j=s+1}^n \alpha_j^{-k}\beta_j z_j\right\|_2}{\|\alpha_i^{-k}\beta_i z_i\|_2} = \frac{\sqrt{\sum_{j=s+1}^n \alpha_j^{-2k}\beta_j^2}}{\alpha_i^{-k}|\beta_i|}$$

$$\leq \frac{\sqrt{\sum_{j=s+1}^n \alpha_{s+1}^{-2k}\beta_j^2}}{\alpha_i^{-k}|\beta_i|} = \frac{\alpha_{s+1}^{-k}\sqrt{\sum_{j=s+1}^n \beta_j^2}}{\alpha_i^{-k}|\beta_i|} = \left(\frac{\alpha_i}{\alpha_{s+1}}\right)^k \tan\angle_2(u,\mathcal{Z}).$$

Gemäß Definition 2.2 und Definition 5.2 gilt außerdem

$$0 \leq \angle_2(z_i, A^{-k}\mathcal{U}) \leq \angle_2(z_i, \operatorname{span}\{A^{-k}u\}) = \angle_2(A^{-k}u, \operatorname{span}\{z_i\}) < \pi/2,$$

$$0 \leq \angle_2(u, \mathcal{Z}) \leq \angle_2(\mathcal{U}, \mathcal{Z}) = \angle_2(\mathcal{Z}, \mathcal{U}) < \pi/2,$$

so dass der Beweis vervollständigt wird. □

Satz 5.11 lässt sich für die inverse Unterraumiteration zu Matrixbüscheln verallgemeinern.

Korollar 5.12. *Für das Matrixbüschel (A, M) gemäß Voraussetzung 2.1 sei $\mathcal{V}$ die direkte Summe der Eigenräume zu den s kleinsten Eigenwerten, nämlich $\mathcal{V} = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_s$. Falls ein Unterraum $\mathcal{U}$ die gleiche Dimension wie $\mathcal{V}$ hat und die Bedingung $\tan\angle_M(\mathcal{V},\mathcal{U}) < \infty$ erfüllt, dann gilt*

$$\tan\angle_M\big(\mathcal{V}_i, \big(A^{-1}M\big)^k\mathcal{U}\big) \leq \left(\frac{\lambda_i}{\lambda_{s+1}}\right)^k \tan\angle_M(\mathcal{V},\mathcal{U}) \qquad \forall\ i \in \{1,\ldots,s\}.$$

Es gilt auch die gleichartige Abschätzung bezüglich der A-Winkel.

Im Beweis ist wesentlich zu zeigen, dass

$$\tan\angle_M\big(v_i, \big(A^{-1}M\big)^k\mathcal{U}\big) \leq \left(\frac{\lambda_i}{\lambda_{s+1}}\right)^k \tan\angle_M(\mathcal{V},\mathcal{U})$$

für beliebiges $v_i \in \mathcal{V}_i$ gilt. Der Kernanteil des Beweises bezieht sich wieder auf Vektor-Winkel.

In Kapitel 4 wurde erwähnt, dass die Tangens-Abschätzungen wegen der Störung des Vorkonditionierers im Allgemeinen nicht auf PINVIT-Verfahren übertragen werden können. Nur in einem niedrigdimensionalen Fall kann man noch Tangens-Abschätzungen formulieren, welche im Beweis der Delta-Abschätzungen einsetzbar sind, siehe Lemma 4.11. Nach dieser Überlegung wird für das blockweise PINVIT(1)-Verfahren eine niedrigdimensionale Tangens-Abschätzung als ein Zwischenergebnis entwickelt. Um mit Lemma 4.11 zu vergleichen, wird die Problemstellung wie in Abschnitt 4.1 bezüglich einer aus A-orthonormalen Basismatrizen der Eigenräume von (A, M) bestehenden Matrix V umformuliert. Mit den Substitutionen

$$V^TAV = I, \qquad V^TMV =: H, \qquad V^TAU =: X, \qquad V^TA\tilde{U} =: \tilde{X}, \qquad \Theta^{-1} =: \Psi \tag{5.9}$$

werden wegen $AVV^T = VV^TA = I$ und

$$V^TA(A^{-1}MU\Theta) = V^TMU\Theta = V^TM(VV^TA)U\Theta = HX\Psi^{-1}$$

eine Umfomulierung

$$\big\|(HX - \tilde{X}\Psi)\tilde{C}\big\|_2 \leq \gamma\big\|(HX - X\Psi)\tilde{C}\big\|_2 \quad \text{für beliebiges } \tilde{C} \in \mathbb{R}^{s\times l},\ l \in \{1,\ldots,s\} \tag{5.10}$$

von (5.3) erhalten. Damit entsteht eine Voraussetzung für die weitere Untersuchung.

Vorausetzung 5.13. *Eine positiv definite Diagonalmatrix $H \in \mathbb{R}^{n\times n}$ besitze m verschiedene Eigenwerte $\mu_1 > \mu_2 > \cdots > \mu_m$ mit zugehörigen Vielfachheiten $q_1, q_2, \ldots, q_m$ und zugehörigen Eigenräumen $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_m$. Der bezüglich H definierte Rayleigh-Quotient sei mit μ bezeichnet. Weiter erfülle eine Matrix $X \in \mathbb{R}^{n\times s}$ die Bedingungen*

$$X^THX = I, \qquad X^TX = \Psi^{-1},$$

wobei Ψ eine Diagonalmatrix mit angeordneten Diagonalkomponenten $\psi_1 \geq \psi_2 \geq \cdots \geq \psi_s$ ist. In anderen Worten sind H-orthonormale Ritzvektoren zu den Ritzwerten $\psi_1, \ldots, \psi_s$ von H bezüglich $\operatorname{span}\{X\}$ *durch die Spaltenvektoren von X gegeben. Mit X und einer weiteren Matrix $\tilde{X} \in \mathbb{R}^{n\times s}$ gelte die Bedingung (5.10) mit $\gamma \in (0,1)$. Zu untersuchen sind die Ritzwerte $\psi'_1, \ldots, \psi'_s$ zu* $\operatorname{span}\{\tilde{X}\}$ *in Abhängigkeit von den Ritzwerten $\psi_1, \ldots, \psi_s$ zu* $\operatorname{span}\{X\}$.

Für einen niedrigdimensionalen Fall lässt sich zunächst eine Tangens-Abschätzung beweisen.

Satz 5.14. *Gemäß Voraussetzung 5.13 sei* $\langle X\rangle := \operatorname{span}\{X\}$ *ein Unterraum des durch den Eigenvektoren* $y_{\sigma(1)}, y_{\sigma(2)}, \dots, y_{\sigma(s+1)}$ *zu den Eigenwerten* $\mu_{\sigma(1)} > \mu_{\sigma(2)} > \cdots > \mu_{\sigma(s+1)}$ *aufgespannten invarianten Unterraums. Falls* $\langle X\rangle$ *keinen Eigenvektor besitzt, dann gilt mit der Bezeichnung* $\mathcal{Y} := \operatorname{span}\{y_{\sigma(1)}, \dots, y_{\sigma(s)}\}$ *die Bedingung* $\tan\angle_2\big(\mathcal{Y}, \langle X\rangle\big) < \infty$. *Weiter gilt für* $\langle \tilde{X}\rangle := \operatorname{span}\{\tilde{X}\}$ *und beliebiges* $i \in \{1, \dots, s\}$, *dass*

$$\tan\angle_2\big(y_{\sigma(i)}, \langle \tilde{X}\rangle\big) \le \left((1-\gamma)\frac{\mu_{\sigma(s+1)}}{\mu_{\sigma(i)}} + \gamma\right) \tan\angle_2\big(\mathcal{Y}, \langle X\rangle\big). \tag{5.11}$$

Beweis. Würde $\tan\angle_2\big(\mathcal{Y}, \langle X\rangle\big) < \infty$ nicht gelten, dann gäbe es ein $x \in \langle X\rangle$ mit $\angle_2(x, \mathcal{Y}) = \pi/2$. Damit würde x zum Euklidisch orthogonalen Komplement von $\mathcal{Y}$ bezüglich des Unterraums $\mathcal{Y} \oplus \operatorname{span}\{y_{\sigma(s+1)}\}$ gehören, das aber genau durch $\operatorname{span}\{y_{\sigma(s+1)}\}$ gegeben wird. So würde $\langle X\rangle$ einen Eigenvektor besitzen.

Weiter gibt es für beliebiges $i \in \{1, \dots, s\}$ nach dem Dimensionsvergleich

$$\dim \operatorname{span}\{y_{\sigma(i)}, y_{\sigma(s+1)}\} + \dim\langle X\rangle = 2 + s > s + 1$$

Nichtnullvektoren im Durchschnitt $\operatorname{span}\{y_{\sigma(i)}, y_{\sigma(s+1)}\} \cap \langle X\rangle$. Sei ein solcher Vektor durch Xc mit einem $c \in \mathbb{R}^s \backslash \{\mathbf{0}\}$ gegeben, dann lässt sich Xc auch durch $\alpha_1 y_{\sigma(i)} + \alpha_2 y_{\sigma(s+1)}$ mit $\alpha_1, \alpha_2 \in \mathbb{R}\backslash\{0\}$ darstellen (denn Xc ist kein Nullvektor und wegen $Xc \in \langle X\rangle$ kein Eigenvektor). Da die zu zeigende Abschätzung bei Nichtnull-Umskalierungen der Eigenvektoren ungeändert bleibt, wird im Folgenden o.B.d.A. angenommen, dass $y_{\sigma(i)}$, $y_{\sigma(s+1)}$ Euklidisch normiert sind und die Koeffizienten α_1, α_2 positiv sind. Damit hat der Vektor HXc die positiven Koeffizienten $\mu_{\sigma(i)}\alpha_1$, $\mu_{\sigma(s+1)}\alpha_2$. und es gilt

$$\tan\angle_2(y_{\sigma(i)}, HXc) = \frac{\mu_{\sigma(s+1)}\alpha_2}{\mu_{\sigma(i)}\alpha_1} = \frac{\mu_{\sigma(s+1)}}{\mu_{\sigma(i)}} \tan\angle_2(y_{\sigma(i)}, Xc) < \tan\angle_2(y_{\sigma(i)}, Xc),$$

$$\angle_2(y_{\sigma(i)}, HXc) + \angle_2(HXc, Xc) = \angle_2(y_{\sigma(i)}, Xc) < \pi/2. \tag{5.12}$$

Weiter betrachten wir den Vektor $\tilde{X}\Psi c$. Nach der Bedingung (5.10) gilt

$$\|HXc - \tilde{X}\Psi c\|_2 \le \gamma \|HXc - X\Psi c\|_2.$$

Dabei ist $X\Psi c$ wegen

$$\big(X(X^TX)^{-1}X^T\big)(HXc) = X\Psi(X^THX)c = X\Psi c$$

die Euklidische Projektion von HXc auf $\langle X\rangle$, so dass

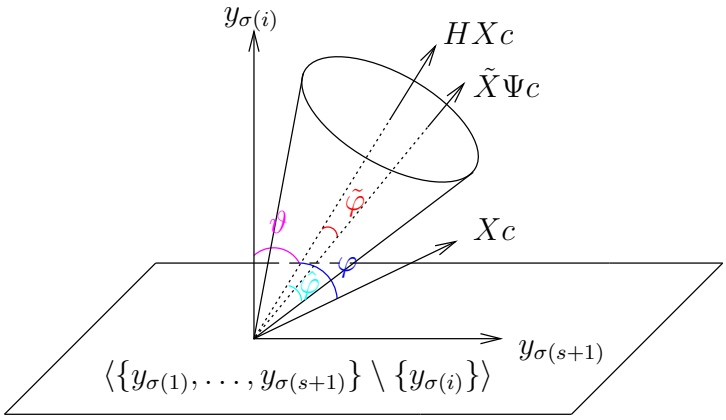
$$\|HXc - X\Psi c\|_2 \le \|HXc - \mu(Xc)Xc\|_2.$$

Außerdem gilt $\|HXc - \mu(Xc)Xc\|_2 \le \|HXc\|_2$ wegen $HXc - \mu(Xc)Xc \perp_2 Xc$. So entsteht

$$\|HXc - \tilde{X}\Psi c\|_2 \le \gamma \|HXc - \mu(Xc)Xc\|_2 < \|HXc\|_2.$$

Damit ist $\angle_2(HXc, \tilde{X}\Psi c)$ nicht der größte Innenwinkel des von HXc und $\tilde{X}\Psi c$ gebildeten Dreiecks und daher ein spitzer Winkel, und es gilt genauer

$$\angle_2(HXc, \tilde{X}\Psi c) \le \arcsin\big(\gamma \sin\angle_2(HXc, Xc)\big) < \angle_2(HXc, Xc). \tag{5.13}$$



$$\angle_2(y_{\sigma(i)},\tilde{X}\Psi c)\le\vartheta+\tilde{\varphi}\le\vartheta+\hat{\varphi},$$

$$\sin\hat{\varphi}=\gamma\sin\varphi\quad\Rightarrow\quad\tan\angle_2(y_{\sigma(i)},\tilde{X}\Psi c)\le(1-\gamma)\tan\vartheta+\gamma\tan(\vartheta+\varphi).$$

Abbildung 5.1: Eine Winkel-Ungleichung

Da der Euklidische Winkel zwischen zwei Nichtnullvektoren x, y identisch zum kürzesten Weg zwischen $x/\|x\|_2$, $y/\|y\|_2$ auf der Einheitskugel der Euklidischen Vektornorm ist, gilt eine Dreiecksungleichung der Form $\angle_2(x,y)\le\angle_2(x,z)+\angle_2(z,y)$. Damit ist $\angle_2(y_{\sigma(i)},\tilde{X}\Psi c)$ wegen

$$\begin{aligned}\angle_2(y_{\sigma(i)},\tilde{X}\Psi c)&\le\angle_2(y_{\sigma(i)},HXc)+\angle_2(HXc,\tilde{X}\Psi c)\\&<\angle_2(y_{\sigma(i)},HXc)+\angle_2(HXc,Xc)=\angle_2(y_{\sigma(i)},Xc)\end{aligned}$$

auch ein spitzer Winkel, so dass

$$\tan\angle_2\big(y_{\sigma(i)},\langle\tilde{X}\rangle\big)\le\tan\angle_2(y_{\sigma(i)},\tilde{X}\Psi c)\le\tan\big(\angle_2(y_{\sigma(i)},HXc)+\angle_2(HXc,\tilde{X}\Psi c)\big).$$

Nach (5.12) und (5.13) hat $\tan\angle_2\big(y_{\sigma(i)},\langle\tilde{X}\rangle\big)$ dann eine obere Schranke

$$\tan\big(\angle_2(y_{\sigma(i)},HXc)+\arcsin\big(\gamma\sin\angle_2(HXc,Xc)\big)\big),$$

welche weiter durch

$$\left((1-\gamma)\frac{\mu_{\sigma(s+1)}}{\mu_{\sigma(i)}}+\gamma\right)\tan\angle_2(y_{\sigma(i)},Xc)$$

nach oben abgeschätzt werden kann. Dabei wird analog zum Beweis von Lemma 4.11 eine konvexe Funktion für γ betrachtet. Weiter ist $\alpha_1 y_{\sigma(i)}$ die Euklidische Projektion von Xc sowohl auf $\operatorname{span}\{y_{\sigma(i)}\}$, als auch auf $\mathcal{Y}$, so dass

$$\tan\angle_2(y_{\sigma(i)},Xc)=\tan\angle_2\big(Xc,\mathcal{Y}\big)\le\tan\angle_2\big(\mathcal{Y},\langle X\rangle\big)$$

gilt (siehe Abbildung 5.1) und der Beweis damit vervollständigt wird. □

Für den Fall, dass $\langle X\rangle$ Eigenvektoren besitzt, lässt sich die Analyse auf den größten keinen Eigenvektor besitzenden Unterraum von $\langle X\rangle$ reduzieren. Satz 5.14 ist möglicherweise zur Entwicklung einer verallgemeinerten Delta-Abschätzung nützlich, denn $\tan\angle_2(y_{\sigma(i)},Xc)$ für $i=1$ lässt sich wie folgt mit Eigenwerten und Ritzwerten darstellen.

Lemma 5.15. *Mit den Formulierungen in Satz 5.14 gilt die strikte Einschachtelung*

$$\mu_{\sigma(1)} > \psi_1 > \mu_{\sigma(2)} > \psi_2 > \cdots > \mu_{\sigma(s)} > \psi_s > \mu_{\sigma(s+1)},$$

falls $\langle X\rangle$ keinen Eigenvektor besitzt. Weiter gilt für $\mathcal{W} := \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(s+1)}\} \cap \langle X\rangle$, dass

$$\tan^2 \measuredangle_2(y_{\sigma(1)}, \mathcal{W}) = \left(\prod_{i=1}^{s} \frac{\mu_{\sigma(1)} - \psi_i}{\psi_i - \mu_{\sigma(s+1)}}\right)\left(\prod_{i=2}^{s} \frac{\mu_{\sigma(i)} - \mu_{\sigma(s+1)}}{\mu_{\sigma(1)} - \mu_{\sigma(i)}}\right). \tag{5.14}$$

Beweis. Falls $\langle X\rangle$ keinen Eigenvektor besitzt, dann sind die Ritzvektoren zu $\langle X\rangle$ keine Eigenvektoren, und deren Residuen sind Nichtnullvektoren und gehören zum Euklidisch orthogonalen Komplement $\langle X\rangle^{\perp_2}$ von $\langle X\rangle$ bezüglich des Unterraums $\operatorname{span}\{y_{\sigma(1)}, \ldots, y_{\sigma(s+1)}\}$. Wegen $\dim\langle X\rangle = s$ ist $\langle X\rangle^{\perp_2}$ eindimensional und lässt sich durch das Residuum eines beliebigen Ritzvektors erzeugen. Betrachten wir dann einen beliebigen Ritzwert ψ_i. Würde $\psi_i = \mu_{\sigma(i)}$ gelten, dann müsste das Residuum eines zugehörigen Ritzvektors x_i Euklidisch orthogonal zum Eigenvektor $y_{\sigma(i)}$ liegen, denn

$$y_{\sigma(i)}^T\big((H - \psi_i)x_i\big) = \big((H - \mu_{\sigma(i)})y_{\sigma(i)}\big)^T x_i = \mathbf{0}^T x_i = 0.$$

Damit würde $y_{\sigma(i)}$ Euklidisch orthogonal zu $\langle X\rangle^{\perp_2}$ liegen, das heißt aber $y_{\sigma(i)} \in \langle X\rangle$, so dass $\langle X\rangle$ einen Eigenvektor besitzen müsste. Daher gilt $\psi_i \neq \mu_{\sigma(i)}$ und analog $\psi_i \neq \mu_{\sigma(i+1)}$ für beliebiges $i = 1, \ldots, s$, so dass die strikte Einschachtelung bewiesen wird.

Damit sind die Matrizen $H - \psi_i I$ invertierbar. Betrachten wir weiter einen beliebigen Nichtnullvektor r aus $\langle X\rangle^{\perp_2}$ und beliebige Ritzvektoren x_i zu den Ritzwerten ψ_i. Dann gibt es für jedes $i \in \{1, \ldots, s\}$ ein $\omega_i \in \mathbb{R}\backslash\{0\}$ mit

$$(H - \psi_i I)x_i = \omega_i r, \quad \text{also} \quad x_i = \omega_i (H - \psi_i I)^{-1} r.$$

Nach der strikten Einschachtelung sind die Ritzwerte ψ_i paarweise verschieden, so dass die Ritzvektoren x_i paarweise Euklidisch orthogonal sind und damit eine Basis von $\langle X\rangle$ bilden. Dann ist ein beliebiger Nichtnullvektor $w \in \mathcal{W}$ wegen $\mathcal{W} \subset \langle X\rangle$ durch

$$w = \sum_{i=1}^{s} \beta_i x_i = \sum_{i=1}^{s} \beta_i \omega_i (H - \psi_i I)^{-1} r$$

mit Koeffizienten $\beta_i \in \mathbb{R}$ darstellbar. Einsetzen der Eigenbasisdarstellung $r = \sum_{j=1}^{s+1} \alpha_j z_j$ mit $z_j := y_{\sigma(j)}/\|y_{\sigma(j)}\|_2$ liefert dann

$$\begin{aligned} w &= \sum_{i=1}^{s}\sum_{j=1}^{s+1} \beta_i \omega_i \alpha_j (H - \psi_i I)^{-1} z_j = \sum_{j=1}^{s+1} \alpha_j \sum_{i=1}^{s} \beta_i \omega_i (\mu_{\sigma(j)} - \psi_i)^{-1} z_j \\ &= \sum_{j=1}^{s+1} \alpha_j \left(\prod_{k=1}^{s} (\mu_{\sigma(j)} - \psi_k)^{-1}\right) p_{s-1}\big(\mu_{\sigma(j)}\big) z_j \end{aligned}$$

mit einem Polynom

$$p_{s-1}(\delta) := \sum_{i=1}^{s} \beta_i \omega_i \prod_{k=1, k\neq i}^{s} (\delta - \psi_k).$$

Wegen $w \in \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(s+1)}\}$ gilt $\alpha_j \left(\prod_{k=1}^{s} (\mu_{\sigma(j)} - \psi_k)^{-1}\right) p_{s-1}\big(\mu_{\sigma(j)}\big) = 0$ für $j = 2, \ldots, s$. Außerdem sind die Koeffizienten α_j sämtlich von Null verschieden, denn sonst gäbe es Eigenvektoren, die zu r beziehungsweise zu $\langle X\rangle^{\perp_2}$ Euklidisch orthogonal liegen und damit zu $\langle X\rangle$ gehören. Zusammen mit der strikten Einschachtelung der Ritzwerte wird dann gezeigt, dass die Eigenwerte

$\mu_{\sigma(2)}, \ldots, \mu_{\sigma(s)}$ Nullstellen von p_{s-1} sind. Da diese $s-1$ Eigenwerte paarweise verschieden sind und p_{s-1} offenbar höchstens den Grad $s-1$ hat, gibt es ein $\tilde{\alpha} \in \mathbb{R}$, so dass

$$p_{s-1}(\delta) = \tilde{\alpha} \prod_{j=2}^{s} \big(\delta - \mu_{\sigma(j)}\big).$$

Außerdem gilt $\tilde{\alpha} \neq 0$, denn sonst wäre w Nullvektor. Weiter ist $\mathcal{W}$ eindimensional, denn sonst müsste $\mathcal{W}$ die Dimensionszahl 2 haben und identisch mit $\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(s+1)}\}$ wäre, so dass

$$\operatorname{span}\{y_{\sigma(1)}, y_{\sigma(s+1)}\} = \mathcal{W} \subset \langle X \rangle$$

gelten würde, das heißt, $\langle X \rangle$ würde Eigenvektoren besitzen. Damit gilt $\mathcal{W} = \operatorname{span}\{w\}$, und weiter

$$\tan^2 \angle_2(y_{\sigma(1)}, \mathcal{W}) = \tan^2 \angle_2(w, z_1) = \frac{\alpha_{s+1}^2 \left(\prod_{k=1}^{s} (\mu_{\sigma(s+1)} - \psi_k)^{-1}\right)^2 \left(p_{s-1}\big(\mu_{\sigma(s+1)}\big)\right)^2}{\alpha_1^2 \left(\prod_{k=1}^{s} (\mu_{\sigma(1)} - \psi_k)^{-1}\right)^2 \left(p_{s-1}\big(\mu_{\sigma(1)}\big)\right)^2}$$

$$= \left(\frac{\alpha_{s+1}^2}{\alpha_1^2}\right) \left(\prod_{k=1}^{s} \frac{\mu_{\sigma(1)} - \psi_k}{\psi_k - \mu_{\sigma(s+1)}}\right)^2 \left(\prod_{j=2}^{s} \frac{\mu_{\sigma(j)} - \mu_{\sigma(s+1)}}{\mu_{\sigma(1)} - \mu_{\sigma(j)}}\right)^2. \tag{5.15}$$

Um $\alpha_{s+1}^2/\alpha_1^2$ umzuschreiben, nutzen wir die Orthogonalität $r \perp_2 x_i$ und erhalten

$$\omega_i r^T (H - \psi_i I)^{-1} r = 0 \quad \text{sowie} \quad r^T (H - \psi_i I)^{-1} r = 0.$$

Einsetzen der Eigenbasisdarstellung $r = \sum_{j=1}^{s+1} \alpha_j z_j$ liefert dann

$$0 = \sum_{j=1}^{s+1} \alpha_j^2 (\mu_{\sigma(j)} - \psi_i)^{-1} = \left(\prod_{k=1}^{s+1} (\mu_{\sigma(k)} - \psi_i)^{-1}\right) p_s(\psi_i)$$

mit einem Polynom

$$p_s(\delta) := \sum_{j=1}^{s+1} \alpha_j^2 \prod_{k=1, k\neq j}^{s+1} (\mu_{\sigma(k)} - \delta). \tag{5.16}$$

Da die Koeffizienten α_j von Null verschieden sind, ist p_s keine Nullfunktion. Mit der strikten Einschachtelung der Ritzwerte wird dann gezeigt, dass die Ritzwerte $\psi_1, \ldots, \psi_s$ Nullstellen von p_s sind. Da diese s Ritzwerte paarweise verschieden sind und p_s höchstens den Grad s hat, gibt es ein $\tilde{\beta} \in \mathbb{R}\backslash\{0\}$ mit

$$p_s(\delta) = \tilde{\beta} \prod_{i=1}^{s} \big(\delta - \psi_i\big).$$

Zusammen mit der Darstellung (5.16) liefert das für $\delta = \mu_{\sigma(1)}$ die Gleichung

$$\tilde{\beta} \prod_{i=1}^{s} \big(\mu_{\sigma(1)} - \psi_i\big) = p_s(\mu_{\sigma(1)}) = \alpha_1^2 \prod_{k=2}^{s+1} \big(\mu_{\sigma(k)} - \mu_{\sigma(1)}\big)$$

und analog für $\delta = \mu_{\sigma(s+1)}$ die Gleichung

$$\tilde{\beta} \prod_{i=1}^{s} \big(\mu_{\sigma(s+1)} - \psi_i\big) = p_s(\mu_{\sigma(s+1)}) = \alpha_{s+1}^2 \prod_{k=1}^{s} \big(\mu_{\sigma(k)} - \mu_{\sigma(s+1)}\big),$$

so dass

$$\frac{\alpha_{s+1}^2}{\alpha_1^2} = \left(\frac{\prod_{i=1}^{s} \big(\mu_{\sigma(s+1)} - \psi_i\big)}{\prod_{i=1}^{s} \big(\mu_{\sigma(1)} - \psi_i\big)}\right) \left(\frac{\prod_{k=1}^{s} \big(\mu_{\sigma(k)} - \mu_{\sigma(s+1)}\big)}{\prod_{k=2}^{s+1} \big(\mu_{\sigma(k)} - \mu_{\sigma(1)}\big)}\right)^{-1}$$

$$= \left(\prod_{i=1}^{s} \frac{\mu_{\sigma(1)} - \psi_i}{\psi_i - \mu_{\sigma(s+1)}} \right)^{-1} \left(\prod_{k=2}^{s} \frac{\mu_{\sigma(k)} - \mu_{\sigma(s+1)}}{\mu_{\sigma(1)} - \mu_{\sigma(k)}} \right)^{-1}.$$

Zusammen mit (5.15) und Index-Umbezeichnungen wird (5.14) gezeigt. □

Natürlich kann $\tan^2 \angle_2(y_{\sigma(i)}, \mathcal{W})$ auch für $i = 2, \ldots, s$ analog umformuliert werden, aber die resultierende neue Form ist wenig geeignet für Ritzwert-Abschätzungen.

Lemma 5.15 lässt sich leicht auf $\tilde{\mathcal{W}} := \operatorname{span}\{y_{\sigma(1)}, y_{\sigma(s+1)}\} \cap \langle \tilde{X} \rangle$ und die Ritzwerte $\psi'_1, \ldots, \psi'_s$ zu $\langle \tilde{X} \rangle$ übertragen. Als Folgerung erhalten wir eine Gleichung zwischen einem Tangens-Quotienten und einem verallgemeinerten Delta-Quotienten, nämlich

$$\frac{\tan^2 \angle_2(y_{\sigma(1)}, \tilde{\mathcal{W}})}{\tan^2 \angle_2(y_{\sigma(1)}, \mathcal{W})} = \left(\prod_{i=1}^{s} \frac{\mu_{\sigma(1)} - \psi'_i}{\psi'_i - \mu_{\sigma(s+1)}} \right) \left(\prod_{i=1}^{s} \frac{\mu_{\sigma(1)} - \psi_i}{\psi_i - \mu_{\sigma(s+1)}} \right)^{-1}.$$

Dazu wird eine Tangens-Abschätzung

$$\tan \angle_2(y_{\sigma(1)}, \tilde{\mathcal{W}}) \leq \left((1 - \gamma) \frac{\mu_{\sigma(s+1)}}{\mu_{\sigma(1)}} + \gamma \right) \tan \angle_2(y_{\sigma(1)}, \mathcal{W})$$

vermutet, welche insbesondere für $\gamma = 0$ gilt (das folgt aus Korollar 5.12). Außerdem ist der verallgemeinerte Delta-Quotient möglicherweise bei der Begründung der niedrigdimensionalen Analyse nützlich.

6 Schlussfolgerung und Ausblick

In dieser Arbeit haben wir vorkonditionierte Gradientenverfahren zur numerischen Behandlung der diskretisierten Eigenwertprobleme elliptischer Differentialoperatoren betrachtet und neue Ergebnisse zur Konvergenzanalyse präsentiert. Diese Verfahren lassen sich sowohl auf Basen von Gradienten des Rayleigh-Quotienten in verschiedenen Geometrien konstruieren, als auch als gestörte Versionen der inversen Vektoriteration beziehungsweise mancher krylovraumartigen Verfahren interpretieren und hierarchisch auflisten, zum Beispiel als verwandte Verfahren der inversen Vektoriteration in einer von Neymeyr [62] vorgestellten (P)INVIT-Hierarchie. Trotz der langen Anwendungsgeschichte solcher Gradientenverfahren sind bisher nur für einige relativ einfache Verfahren scharfe Konvergenzabschätzungen entwickelt worden.

Durch die (P)INVIT-Hierarchie wird eine Aufgabenserie der „scharfen Konvergenzanalyse“ erstellt. Mit den bereits gelösten Aufgaben für die Verfahren der ersten Stufe, nämlich eine skalierte inverse Vektoriteration und eine durch Vorkonditionierung gestörte inverse Vektoriteration, wird versucht, die weiteren Aufgaben in ähnlicher Weise zu lösen. Die Entwicklung der Konvergenzabschätzung soll etwa aus einer Charakterisierung der extremalen Konvergenz und einer folgenden Analyse in einem niedrigdimensionalen invarianten Unterraum bestehen.

Für die (P)INVIT-Verfahren der zweiten Stufe ist diese Idee erst neulich völlig realisiert worden. In der Zusammenarbeit [68] wurde die gewünschte Konvergenzanalyse für die INVIT-Variante vervollständigt. Mit einer Kurven-Differentiation auf einer Rayleigh-Quotienten-Niveaumenge wurde die extremale Konvergenz charakterisiert, und mit einer Ellipsen-Analyse in einem affinen Raum wurde die vorher bereits bekannte niedrigdimensionale Analyse deutlich vereinfacht. Als Nebenprodukte gelten eine Verschärfung der klassischen winkelmässigen Konvergenzabschätzung und eine Übertragung der Konvergenzanalyse auf eine inverse Krylovraum-Iteration. Bei dieser Übertragung wurde die oben genannte Ellipsen-Analyse in eine Ellipsoide-Analyse verallgemeinert. Außerdem wurden durch Grundeigenschaften der Krylovräume einige interessante Argumente gezeigt, zum Beispiel die Koordinaten-Darstellung der relevanten Schnittpunkte durch Eigenwerte. Aus diesen Resultaten haben wir auch Anregungen zur gewünschten Konvergenzanalyse für die PINVIT-Variante erhalten. Zur Charakterisierung der extremalen Konvergenz wurde eine kompliziertere Kurven-Differentiation mit einer Kegel-Interpretation der möglichen Vorkonditionierer formuliert. Zur niedrigdimensionalen Analyse wurde eine Übergangsellipse konstruiert und der Qualitätsparameter der Vorkonditionierung wurde in technischer Weise in das abzuschätzende Verhältnis eingesetzt. Die erhaltene Schranke lässt sich überraschend in einer einfachen Form darstellen. Eine Vereinfachung des aktuellen Beweises ist also noch wünschenswert.

Ab der dritten Stufe haben die (P)INVIT-Verfahren die Form der Mehrschrittverfahren. Durch zahlreiche numerische Experimente wird gemerkt, dass diese Verfahren höherer Stufe im Gegensatz zu denen aus den ersten zwei Stufen bereits die Clusterrobustheit besitzen sollen. Diese Eigenschaft wurde vorher hauptsächlich in Bezug auf blockweise Verfahren untersucht. Einige bekannte Abschätzungen lassen sich noch verbessern, so dass deren unnatürlichen Voraussetzungen durch bessere Interpretation der Verfahren weggelassen werden können. Zu diesem Zweck haben wir einige Sinusungleichungen bezüglich Unterraumwinkel und kanonischer Winkel, eine niedrigdimensionale Abschätzung sowie eine Verallgemeinerung der klassischen Konvergenzmaße gezeigt.

Zur Konvergenzanalyse für die Shift-Versionen der (P)INVIT-Verfahren, wobei die Inverse beziehungsweise eine Näherungsinverse der geshifteten Matrix $\bar{A} := A - \bar{\lambda}M$ mit einem Shift $\bar{\lambda} < \lambda_1$ eingesetzt wird, lassen sich die oben genannten Abschätzungen wegen der positiven Definitheit von $\bar{A}$ auf das Matrixbüschel $(\bar{A}, M)$ übertragen. Nach einfachen Umrechnungen werden ähnliche

Abschätzungen mit leicht geänderten Konvergenzfaktoren erhalten.

Eines der zukünftigen Forschungsthemen ist die Entwicklung scharfer Konvergenzabschätzungen für die Mehrschrittverfahren der (P)INVIT-Hierarchie. Der Untersuchungsablauf ist vorstellbar. Ausgehend von einem grundlegenden Argument, nämlich der Positionierung der neuen Iterierten durch Kugeln beziehungsweise Kegel, lässt sich ein mehrstufiges Optimierungsproblem des Rayleigh-Quotienten aufstellen. Die damit begründete niedrigdimensionale Analyse bezieht sich dann auf einen invarianten Unterraum, dessen Dimension mindestens vier ist. Mit solchen „größeren“ Dimensionszahlen wird die niedrigdimensionale Analyse wenig intuitiv. Dass es mehr relevante vorherige Iterierte gibt, führt zu weiteren zu lösenden Teilaufgaben. Möglicherweise können wir aus dem beobachteten treppenförmigen Konvergenzverhalten bei unseren numerischen Experimenten (siehe Anhang) sowie aus der Konvergenzanalyse für ähnliche Verfahren wie Jacobi-Davidson-Verfahren, siehe etwa [89, 90, 71, 72, 70, 75, 76], manche Anregungen erhalten.

Ein weiteres Thema ist die Vervollständigung der Konvergenzanalyse für blockweise (P)INVIT-Verfahren. Es wird versucht, die entsprechenden Argumente zu den vektorweisen Verfahren zu verallgemeinern. Zum Beispiel können gewisse von Ritzwerten abhängigen Mengen der Unterräume als Verallgemeinerungen der Rayleigh-Quotienten-Niveaumengen der Vektoren untersucht werden.

Eine andere Problemstellung aus der Praxis ist die Berechnung der inneren Eigenwerte und zugehöriger Eigenvektoren. Dafür kann man die Shift-Versionen der (P)INVIT-Verfahren mit einem aus der Mitte des Spektrums gewählten Shift $\bar{\lambda}$ verwenden. Der Vorkonditionierer B^{-1} soll dann eine Näherungsinverse der indefiniten geshifteten Matrix $\bar{A} := A - \bar{\lambda}M$ sein (solche Vorkonditionierer lassen sich zum Beispiel mit ILUPACK [11] konstruieren, siehe auch [103, 92, 69] für einige Varianten), so dass unsere bezüglich der positiven definiten Matrizen formulierten Konvergenzabschätzungen schwierig darauf zu übertragen sind. Dazu wird versucht, die Vorkonditionierer der indefiniten Matrizen auch durch geometrische Argumente zu interpretieren. Es gibt auch einen einfacher zu analysierenden Verfahrenstyp, welcher nach einer Umformulierung des Eigenwertproblems konstruiert werden kann. Normalerweise ist der Shift $\bar{\lambda}$ noch kein Eigenwert, und die geshiftete Matrix $\bar{A}$ ist damit invertierbar. Wegen $M^{-1}\bar{A} = \bar{A}^{-1}(\bar{A}M^{-1}\bar{A})$ besitzen die Matrixbüschel $(\bar{A}, M)$ und $(\bar{A}M^{-1}\bar{A},\ \bar{A})$ die gleichen Eigenwerte und Eigenvektoren, damit lassen sich die gewünschten betragskleinsten Eigenwerte von $(\bar{A}, M)$ durch die Minimierung beziehungsweise die Maximierung des Rayleigh-Quotienten $\tilde{\rho}(u) := \big(u^T\bar{A}u\big)/\big(u^T\bar{A}M^{-1}\bar{A}u\big)$ bezüglich des umgekehrten Matrixbüschels $(\bar{A},\ \bar{A}M^{-1}\bar{A})$ erhalten. Wegen der positiven Definitheit von $\bar{A}M^{-1}\bar{A}$ lassen sich entsprechende Gradientenverfahren konstruieren und gemäß der klassischen Form analysieren. Der Kehrwert von $\tilde{\rho}(u)$ ist wesentlich ein harmonisches Mittel der Eigenwerte von $(\bar{A}, M)$ und heißt daher auch ein harmonischer Rayleigh-Quotient. In diesem Sinne lässt sich unsere Untersuchung auf den harmonischen Rayleigh-Quotienten und zugehörige Verfahren (siehe [58, 78, 88, 36, 59]) erweitern.

Anhang: Numerische Experimente

Bisher haben wir klassische und neue Konvergenzanalyse für Gradientenverfahren der (P)INVIT-Hierarchie sowie einen Typ von Krylovraum-Verfahren vorgestellt. In diesem Anhang werden diese Verfahren bei konkreten Modellproblemen eingesetzt und deren Konvergenzverhalten illustriert, um mögliche Anregungen für zukünftige Untersuchungen zu erhalten. In Abschnitt A.1 werden die ausgewählten Modellprobleme vorgestellt. Dabei wird der negative Laplace-Operator auf verschiedenen zweidimensionalen Gebieten betrachtet. Zur Diskretisierung der entsprechenden kontinuierlichen Eigenwertprobleme wird die Finite-Elemente-Methode mit der Delaunay-Triangulierung (siehe [21, 7]) und stückweisen linearen Polynomen verwendet. Die Vorkonditionierer werden durch nichtvollständige Cholesky-Faktorisierungen (siehe [85, 104]) konstruiert. Anschließend werden durch oben erwähnte Gradientenverfahren und Krylovraum-Verfahren einige Eigenwertnäherungen und Eigenfunktionsnäherungen berechnet, um die Durchführbarkeit und die Konvergenz dieser Verfahren zu überprüfen. In weiteren Tests wird die Clusterrobustheit untersucht. Als Beispiele werden in Abschnitt A.2 die Tests zu einem Schriftzeichen-Gebiet berichtet.

A.1 Modellprobleme für den negativen Laplace-Operator

Ein typischer ellptischer Operator ist der negative Laplace-Operator, der im Folgenden auf einigen zweidimensionalen Gebieten betrachtet wird. Die entsprechenden kontinuierlichen Eigenwertprobleme haben dann die Form

$$-\frac{\partial^2 \mathsf{u}}{\partial x^2}(x,y) - \frac{\partial^2 \mathsf{u}}{\partial y^2}(x,y) = \lambda \mathsf{u}(x,y), \qquad (x,y)^T \in \Omega \subset \mathbb{R}^2$$

mit gewissen homogenen Randbedingungen, und es sind einige der kleinsten Eigenwerte sowie zugehörige Eigenfunktionen von Interesse.

Zu diesem Problemtyp bringen die Quadrat-Gebiete mit homogener Dirichlet-Randbedingung die wohl einfachsten Aufgabenstellungen. Wählen wir insbesondere die Gebiete

$$\Omega_1 := \big\{(x,y)^T \in \mathbb{R}^2 \;;\; |x| + |y| \leq 1\big\}, \qquad \Omega_\infty := \big\{(x,y)^T \in \mathbb{R}^2 \;;\; \max\{|x|,|y|\} \leq 1\big\},$$

die jeweils durch die Einheitskreise der Vektornormen $\|\cdot\|_1$ und $\|\cdot\|_\infty$ eingeschlossen werden. Auf Ω_1 werden durch

$$\cos\left(\frac{k\pi}{2}(x+y)\right)\cos\left(\frac{l\pi}{2}(x-y)\right), \qquad k,\, l \in \mathbb{N}\backslash\{0\}$$

(sowie deren Nichtnull-Umskalierungen) Eigenfunktionen zu den Eigenwerten $(k^2+l^2)\pi^2/2$ gegeben. Auf Ω_∞ erhält man in ähnlicher Form

$$\cos\left(\frac{k\pi}{2}x\right)\cos\left(\frac{l\pi}{2}y\right), \qquad k,\, l \in \mathbb{N}\backslash\{0\}$$

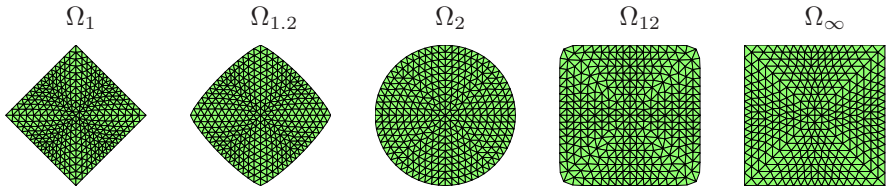
und $(k^2+l^2)\pi^2/4$ als Eigenfunktionen und Eigenwerte. Eine kompliziertere, aber auch analytisch behandelbare Aufgabenstellung bezieht sich auf das durch den Einheitskreis der Euklidischen Vektornorm $\|\cdot\|_2$ eingeschlossene Gebiet

$$\Omega_2 := \big\{(x,y)^T \in \mathbb{R}^2 \;;\; x^2+y^2 \leq 1\big\} = \big\{\big(r\cos(\vartheta), r\sin(\vartheta)\big)^T \;;\; r \in [0,1],\; \vartheta \in [0,2\pi)\big\}$$

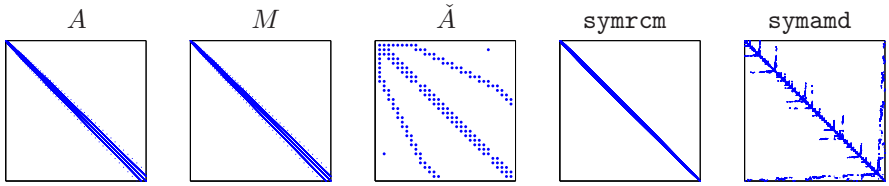
und homogene Dirichlet-Randbedingung. Dafür sind Eigenfunktionen bezüglich Polarkoordinaten und Bessel-Funktionen analytisch darstellbar. Eine Eigenfunktionsbasis besteht aus den Funktionen

$$\cos(k\vartheta)J_k(\xi_{kl}\,r), \quad k \in \mathbb{N},\ l \in \mathbb{N}\backslash\{0\} \quad \text{und} \quad \sin(k\vartheta)J_k(\xi_{kl}\,r), \quad k,\, l \in \mathbb{N}\backslash\{0\},$$

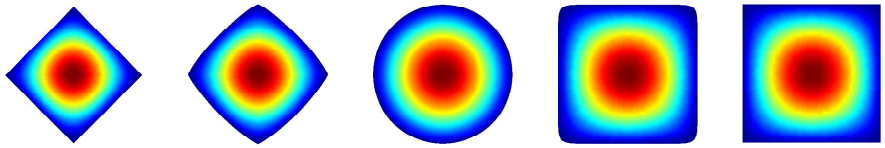
wobei J_k eine Bessel-Funktion der ersten Art ist und der Faktor ξ_{kl} in der Variablen die l-te positive Nullstelle von J_k ist. Durch ξ_{kl}^2 werden die entsprechenden Eigenwerte gegeben.
Wählen wir zusätzlich zwei Übergangsgebiete $\Omega_{1.2}$, Ω_{12}, welche in ähnlicher Weise durch Einheitskreise der p-Normen $\|\cdot\|_{1.2}$, $\|\cdot\|_{12}$ definiert werden. Für diese fünf Einheitskreis-Gebieten können wir mit Polarkoordinaten und Normierungen quasi gleichmäßig verteilte Gitterpunkte konstruieren und anschließend mit dem MATLAB-Befehl `delaunay` Triangulierungen erstellen.



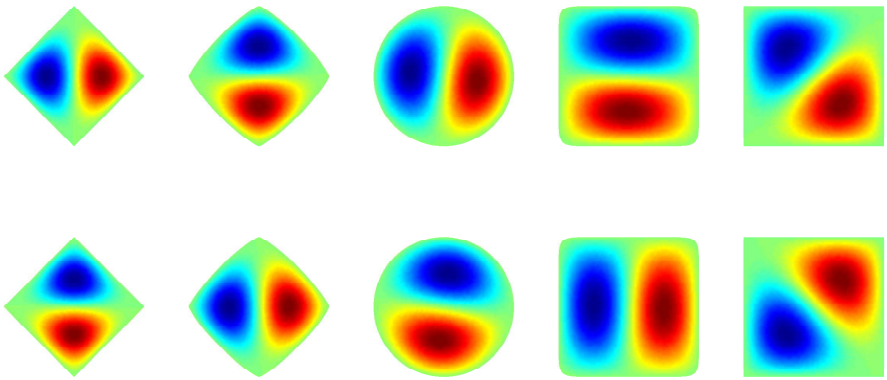
Bezüglich der erhaltenen Dreiecke werden stückweise lineare Polynome definiert, welche Basisfunktionen bilden und Komponenten der Steifigkeitsmatrix A und der Massenmatrix M erzeugen. A und M haben gleiche oder fast gleiche Besetzungsstrukturen (für manche innerhalb der Träger der Ansatzfunktionen liegende Gitterkanten können Nullkomponenten in die Steifigkeitsmatrix A eingetragen werden). Als Beispiel zeigen wir bezüglich des Gebiets Ω_2 die Besetzungsstrukturen von A, M und einer führenden Untermatrix $\check{A} := A(1:30,\, 1:30)$ von A. Zur Berechnung des kleinsten Eigenwerts und zugehöriger Eigenvektoren von (A, M) wird das PINVIT(1)-Verfahren eingesetzt, wobei der Vorkonditionierer durch eine nichtvollständige Cholesky-Faktorisierung konstruiert wird. Eventuell kann man die MATLAB-Befehle `symrcm` (siehe [20, 26, 27]) und `symamd` (siehe [1, 2]) benutzen, um die Matrizen zu permutieren und dann die Anzahl der Nichtnullkomponenten der näherungsweisen Cholesky-Faktoren zu reduzieren.



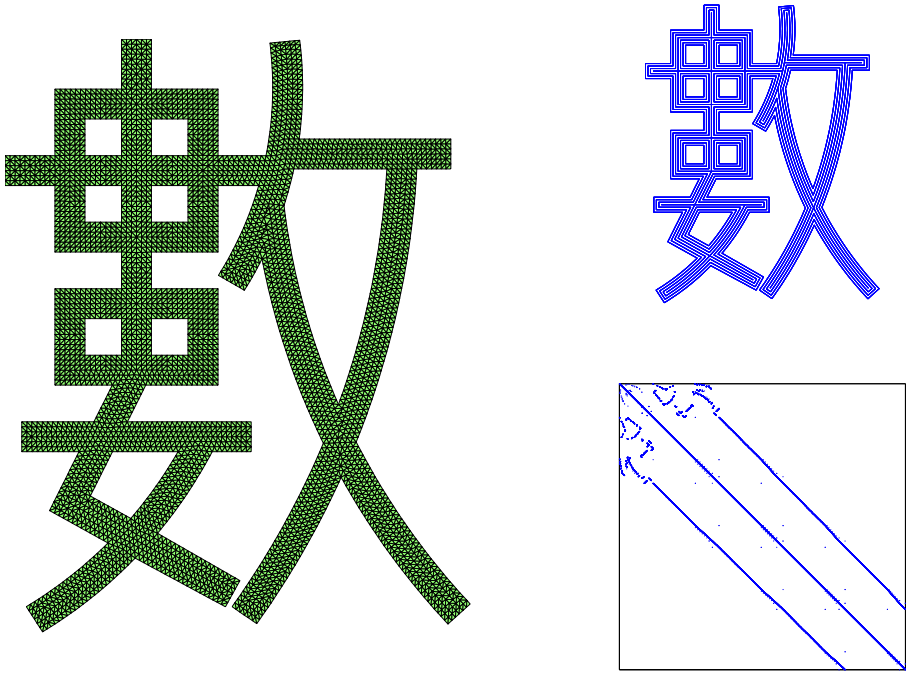
Mit den berechneten Eigenvektoren lassen sich die entsprechenden Eigenfunktionen näherungsweise darstellen. Für jedes der obigen Einheitskreis-Gebiete sind die Näherungen der Eigenfunktionen zum kleinsten Eigenwert symmetrisch bezüglich der Koordinaten-Achsen.



Weitere Eigenwerte und Eigenvektoren von (A, M) können mit dem blockweisen PINVIT(1)-Verfahren ermittelt werden. Für Ω_1, Ω_2 und Ω_∞ ist der zweite analytische Eigenwert jeweils ein verdoppelter Eigenwert. Trotzdem werden dessen Näherungen oft durch zwei fast gleiche einfache Eigenwerte von (A, M) gegeben. Gemäß dem Test ist der zweite analytische Eigenwert für $\Omega_{1.2}$ beziehungsweise Ω_{12} vermutlich auch verdoppelt.



Weiter betrachten wir Aufgabenstellungen mit Schriftzeichen-Gebieten und homogener Dirichlet-Randbedingung. Dabei können wir zur Delaunay-Triangulierung gemäß der Randfigur einige nacheinander einschachtelnde Kurven konstruieren und darauf Gitterpunkte legen. Die restlichen Berechnungen werden wie beim obigen Test für Einheitskreis-Gebiete durchgeführt. Als Beispiel wählen wir ein Gebiet zum chinesischen Schriftzeichen Shu (das kann Zahl bedeuten und als Abkürzung des Wortes Shuxue, das heißt Mathematik, verwendet werden). Da die Matrizen A und M fast gleiche Besetzungsstrukturen haben, zeigen wir nur die Besetzungsstruktur von A.



Mit den berechneten Eigenvektoren zu den kleinsten fünf diskreten Eigenwerten erhalten wir fol-

gende Eigenfunktionsnäherungen.

In [55] wurde zum Test eines adaptiven Eigenlösers eine interessante Aufgabenstellung konstruiert. Dabei ist das Eigenwertproblem auf einer Kreisscheibe mit einem Schlitz betrachtet, nämlich auf dem Gebiet

$$\big\{\big(r\cos(\vartheta), r\sin(\vartheta)\big)^T \;;\; r\in[0,1],\ \vartheta\in(0,2\pi)\big\}.$$

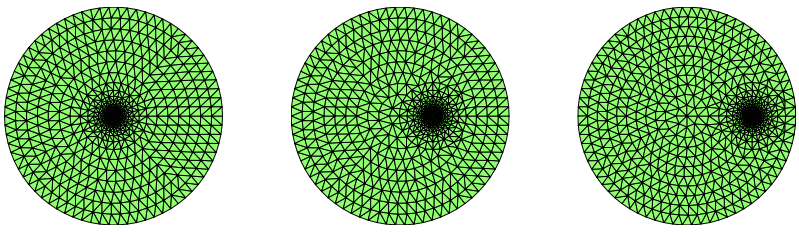
Auf dem Randteil mit $r\in[0,1],\ \vartheta=0$ und dem Randteil mit $r=1,\ \vartheta\in(0,2\pi]$ wird homogene Dirichlet-Randbedingung gestellt, während auf dem Randteil mit $r\in(0,1),\ \vartheta=2\pi$ homogene Neumann-Randbedingung gestellt wird. Dafür sind die Eigenfunktionen des negativen Laplace-Operators wieder bezüglich Polarkoordinaten und Bessel-Funktionen analytisch darstellbar. Eine Eigenfunktionsbasis besteht aus den Funktionen

$$\sin\big((k/2+1/4)\vartheta\big)J_{k/2+1/4}(\tilde{\xi}_{kl}\,r), \qquad k\in\mathbb{N},\ l\in\mathbb{N}\backslash\{0\}$$

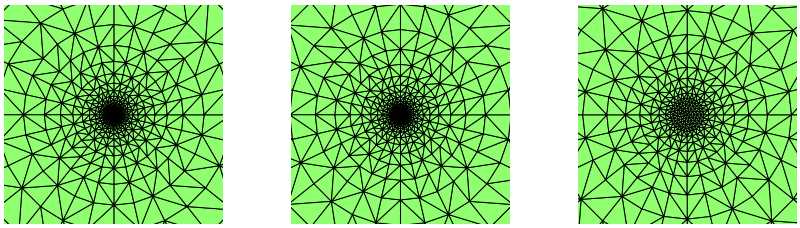
mit Bessel-Funktionen $J_{k/2+1/4}$ und deren positiven Nullstellen $\tilde{\xi}_{kl}$. Die entsprechenden Eigenwerte werden durch $\tilde{\xi}_{kl}^2$ gegeben. (Für eine ähnliche Aufgabenstellung auf Kreisringen mit Schlitz lassen sich die Eigenfunktionen durch zwei Arten von Bessel-Funktionen beschreiben.) Eine allgemeinere Aufgabenstellung bezieht sich auf Gebiete der Form

$$\Omega_{2,a} := \big\{\big(r\cos(\vartheta), r\sin(\vartheta)\big)^T \;;\; r\in[0,a],\ \vartheta\in[0,2\pi) \quad\text{oder}\quad r\in(a,1],\ \vartheta\in(0,2\pi)\big\}$$

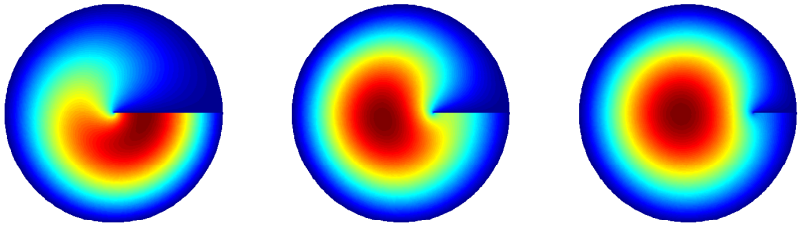
und homogene Dirichlet-Randbedingung auf dem Randteil mit $r\in[a,1],\ \vartheta=0$ und dem Randteil mit $r=1,\ \vartheta\in(0,2\pi]$ sowie homogene Neumann-Randbedingung auf dem Randteil mit $r\in(a,1),\ \vartheta=2\pi$. Zum Test wählen wir die $\Omega_{2,a}$-Gebiete zu $a\in\{0,\,0.3,\,0.6\}$. Mit dem adaptiven Eigenlöser aus [55] kann man für ausgewählte Eigenfunktionen die Teilgebiete, wo die Residuen (gemäß vorherigen Berechnungen) betragsmäßig große Komponenten haben, feiner als andere Teilgebiete triangulieren. Alternativ können wir mit Polarkoordinaten eine ähnliche Verteilung der Gitterpunkte konstruieren und anschließend eine Delaunay-Triangulierung durchführen. Zum Beispiel verfeinern wir gemäß den Residuen der Eigenfunktionen zum kleinsten Eigenwert jeweils eine Umgebung des Punktes $(a,0)^T$.



Für Details der Verfeinerung können wir das Teilgebiet $[a-\varepsilon,a+\varepsilon]\times[-\varepsilon,\varepsilon]$ mit $\varepsilon\in\{0.1,\,0.01,\,0.001\}$ beobachten.



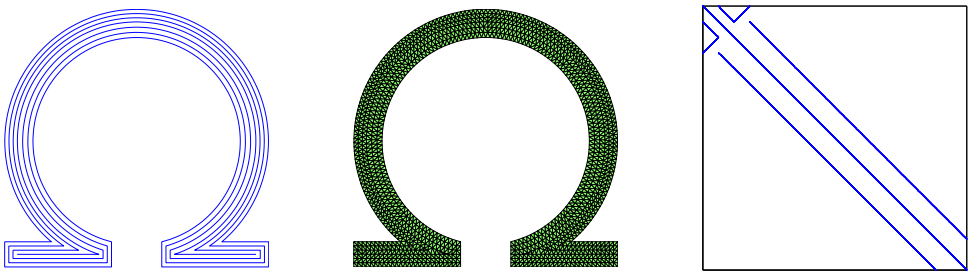
Mit ähnlichen Anzahlen der Gitterpunkte liefert eine solche Triangulierung bessere Eigenwert-Näherungen als eine quasi-gleichmäßige Triangulierung. Außerdem haben die Eigenfunktionsnäherungen bessere Darstellungen nahe dem Punkt $(a, 0)^T$.



A.2 Über Clusterrobustheit

Wohlbekannt ist die Clusterrobustheit der blockweisen vorkonditionierten Gradientenverfahren, siehe [49, 62] für einige experimentielle Beispiele. Eine bisher wenig betrachtete Frage ist, welche Vektoriterationen der (P)INVIT-Hierarchie clusterrobust sind.

Betrachten wir dafür die Aufgabenstellung auf einem Gebiet zum Schriftzeichen Omega mit homogener Dirichlet-Randbedingung und illustrieren bezüglich eines zugehörigen diskreten Eigenwertproblems unsere Konvergenzabschätzungen für Gradientenverfahren. Das Gebiet besteht aus einem Teil des Kreisrings mit dem Mittelpunkt $(0.5,\ 0.5)^T$ und dem Radiusintervall $[\sqrt{0.17},\ \sqrt{0.225}+0.05]$ und einem Teil des Rechtecks $[-0.025,\ 1.025]\times[0,\ 0.1]$. Mittels einiger einschachtelnder Kurven wird eine quasi gleichmäßige Gitterpunktverteilung konstruiert. Gemäß einer Delaunay-Triangulierung und stückweise definierten linearen Ansatzfunktionen wird ein Matrixbüschel mit dünn besetzten Matrizen A und M aufgestellt.



Mit den durch blockweise PINVIT(1) berechneten Eigenvektoren zu den kleinsten fünf diskreten Eigenwerten erhalten wir folgende Eigenfunktionsnäherungen. Anscheinend können die ersten zwei Eigenfunktionen einem verdoppelten analytischen Eigenwert entsprechen.

Wählen wir in weiteren Tests eine feinere Diskretisierung mit 10044 inneren Gitterpunkten. Die Gitterpunkte werden weiter durch den MATLAB-Befehl `symamd` umnumeriert, so dass die entsprechenden permutierten Matrizen A, M bessere Besetzungsstruktur haben. Zur Konstruktion der Vorkonditionierer werden nichtvollständige Cholesky-Faktorisierungen mit dem MATLAB-Befehl `cholinc` durchgeführt. Mit den erhaltenen oberen Dreiecksmatrizen $\tilde{C}$ werden die Vorkonditionierer B^{-1} durch $\tilde{C}^{-1}\tilde{C}^{-T}$ gegeben. Wählen wir beispielsweise die folgenden drei Vorkonditionierer.

	Anzahl der Nichtnullen in $\tilde{C}$	$\mathscr{R}(I - B^{-1}A)$
`cholinc`$(A, 0.012)$	75933	$\leq \gamma \approx 0.85099$
`cholinc`$(A, 0.003)$	112752	$\leq \gamma \approx 0.50183$
`cholinc`$(A, 0.001)$	135646	$\leq \gamma \approx 0.17783$

Weiter lassen sich die kleinsten 6 Eigenwerte $\lambda_1, \ldots, \lambda_6$ von (A, M) mit blockweisen PINVIT-Verfahren bis zu 14 signifikanten Stellen bestimmen.

$$\begin{aligned} \lambda_1 &\approx 553.84293037914, & \lambda_2 &\approx 553.86938364353, \\ \lambda_3 &\approx 789.17739621791, & \lambda_4 &\approx 794.53805958656, \\ \lambda_5 &\approx 803.43416066468, & \lambda_6 &\approx 815.81147549699. \end{aligned}$$

Diese Eigenwerte sind alle einfach. λ_1 und λ_2 bilden einen sogenannten Eigenwert-Cluster. Wegen $\lambda_1 \approx \lambda_2$ liegen die Schlussphase-Konvergenzfaktoren

$$\left((1-\gamma)\frac{\lambda_1}{\lambda_2} + \gamma\right)^2, \quad \left(\frac{\kappa + \gamma(2-\kappa)}{(2-\kappa) + \gamma\kappa}\right)^2 \quad \text{mit} \quad \kappa = \frac{\lambda_1}{\lambda_2}\left(\frac{\lambda_m - \lambda_2}{\lambda_m - \lambda_1}\right)$$

der Einschittverfahren der PINVIT-Hierarchie nahe 1. Dementsprechend beobachtet man für diese Verfahren im folgenden Test sehr langsame Konvergenz.

Test I. Mit den oben gewählten Vorkonditionierern und 100 zufälligen Anfangsiterierten zum gleichen Rayleigh-Quotienten 10^5 werden jeweils die ersten zehn PINVIT-Verfahren durchgeführt und die resultierenden Rayleigh-Quotient-Folgen betrachtet. Mit nur wenigen Schritten erreichen alle Folgen das Intervall (λ_1, λ_2). Weiter können die Folgen von PINVIT(1) und PINVIT(2) ein Hindernis nicht mit einem angemessenen Aufwand überschreiten. Die Folgen von anderen Verfahren haben bezüglich des Abstands zu λ_1 treppenförmige Konvergenzverhalten. In Abbildung A.1 wird jeweils die langsamste Konvergenz zu den gegebenen 100 Anfangsiterierten dargestellt. Als Ergänzung wird der Test auch für den exakten Vorkonditionierer $B^{-1}{=}A^{-1}$ (dieser lässt sich durch iteratives Lösen der Gleichungssysteme realisieren), das heißt für $\gamma = 0$ durchgeführt. Die entsprechenden Verfahren sind also im Wesentlichen INVIT-Verfahren.

Anscheinend wird die benötigte Treppenanzahl eines Mehrschrittverfahrens der PINVIT-Hierarchie durch die Qualität des Vorkonditionierers sowie den Index der PINVIT-Stufe beeinflusst. Die relativ schnelle Konvergenz dieser Mehrschittverfahren lässt sich vermutlich durch einen von gewissen Eigenwerten aus $\{\lambda_3, \ldots, \lambda_m\}$ abhängigen Konvergenzfaktor erklären.

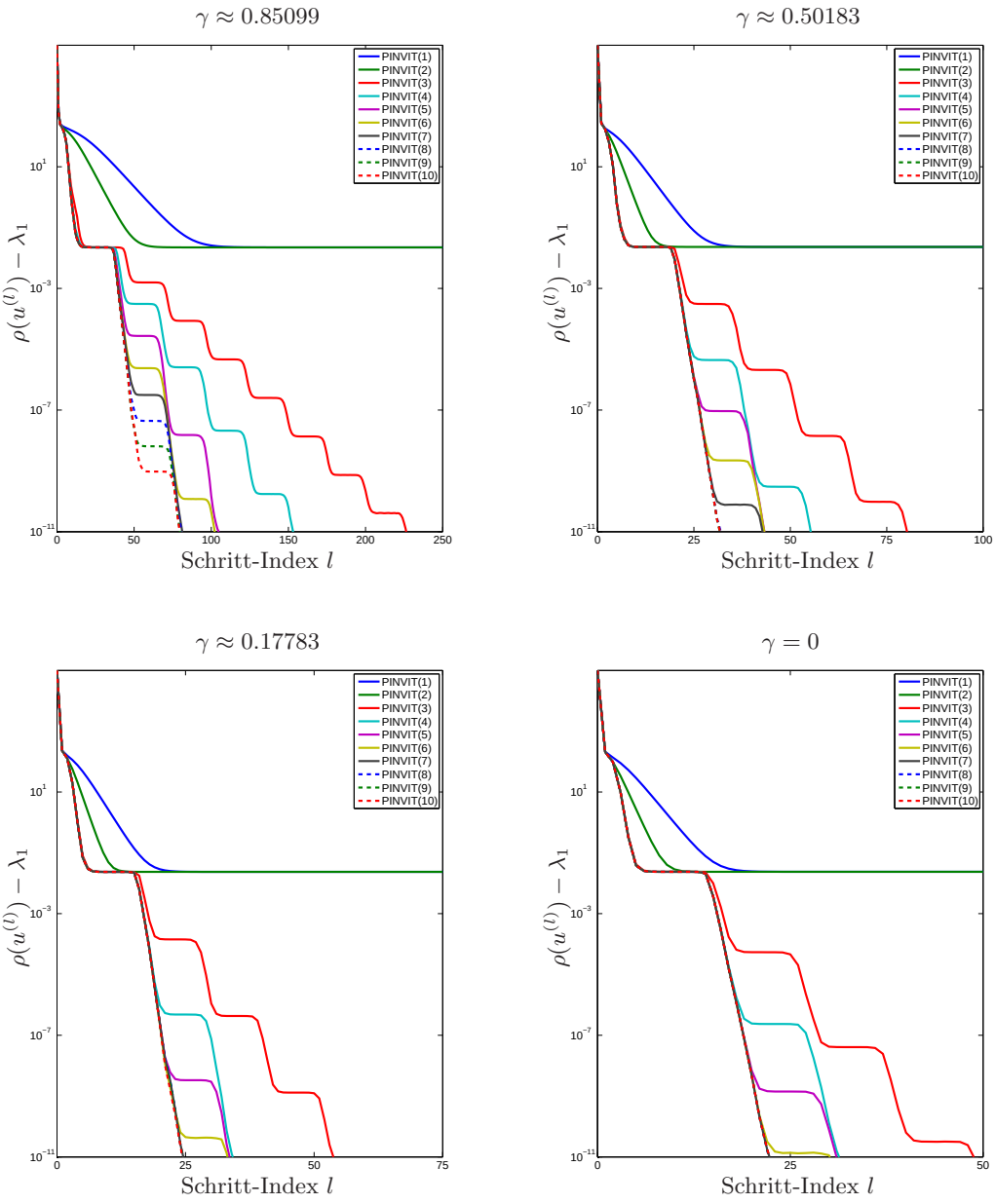


Abbildung A.1: PINVIT-Verfahren bei einem zweielementigen Eigenwert-Cluster

Es stellt sich die Frage, ob PINVIT(3) bei einem dreielementigen Eigenwert-Cluster sehr langsam konvergieren könnte. Dafür können wir mittels einer Rang-1-Modifikation von A ein Matrixbüschel $(\tilde{A}, M)$ konstruieren, welcher einen solchen Eigenwert-Cluster besitzt. Für beliebige M-orthonormale Basismatrizen V_i der Eigenräume von (A, M) gilt mit $V = [V_1, \ldots, V_m]$, dass

$$AV = MV\Lambda \quad \Leftrightarrow \quad A = MV\Lambda V^T M = \sum_{i=1}^{m} \lambda_i (MV_i)(MV_i)^T.$$

Da λ_3 einfach ist, können wir einen zugehörigen M-normierten Eigenvektor v_3 als eine M-orthonormale Basismatrix benutzen. Mit $\tilde{A} := A + (\lambda_2 - \lambda_3 + 0.01)(Mv_3)(Mv_3)^T$ werden die kleinsten drei Eigenwerte von $(\tilde{A}, M)$ durch λ_1, λ_2, λ_2+0.01 gegeben, welche einen Eigenwert-Cluster bilden. Da $\tilde{A}$ normalerweise nicht dünn besetzt ist, benutzen wir zur Vorkonditionierung von $\tilde{A}$ die obigen Vorkonditionierer B^{-1} von A und eine näherungsweise Sherman-Morrison-Woodbury-Formel:

$$\tilde{B}^{-1} := B^{-1} - \frac{\alpha(B^{-1}w)(B^{-1}w)^T}{1 + \alpha(w^T B^{-1} w)} \quad \text{mit} \quad \alpha := \lambda_2 - \lambda_3 + 0.01, \quad w := Mv_3.$$

Damit ist Qualitätsschranke $\mathscr{R}(I - \tilde{B}^{-1}\tilde{A})$ nur leicht von $\mathscr{R}(I - B^{-1}A)$ verschieden.

	$\mathscr{R}(I - B^{-1}A)$	$\mathscr{R}(I - \tilde{B}^{-1}\tilde{A})$
$\texttt{cholinc}(A, 0.012)$	$\leq \gamma \approx 0.85099$	$\leq \tilde{\gamma} \approx 0.85095$
$\texttt{cholinc}(A, 0.003)$	$\leq \gamma \approx 0.50183$	$\leq \tilde{\gamma} \approx 0.50186$
$\texttt{cholinc}(A, 0.001)$	$\leq \gamma \approx 0.17783$	$\leq \tilde{\gamma} \approx 0.17783$

Zu diesem modifizierten Eigenwertproblem konvergieren die Rayleigh-Quotient-Folgen der Mehrschittverfahren der PINVIT-Hierarchie etwas langsamer und besitzen mehr Treppen in der Darstellung des Abstands zu λ_1. Außerdem muss ein PINVIT-Verfahren höherer Stufe nicht eine kleinere Anzahl der äußeren Schritte haben. In Abbildung A.2 werden diese Konvergenzverhalten für 100 zufällige Anfangsiterierten zum gleichen Rayleigh-Quotienten 10^5 dargestellt.

Mit einer anderen Rang-1-Modifikation $\hat{A} := A + (\lambda_1 - \lambda_2)(Mv_2)(Mv_2)^T$ sind die kleinsten Eigenwerte des Matrixbüschels $(\hat{A}, M)$ gut getrennt, darunter ist λ_1 ein zweifacher Eigenwert. Zur Vorkonditionierung von $\hat{A}$ werden wieder die Vorkonditionierer B^{-1} von A und eine näherungsweise Sherman-Morrison-Woodbury-Formel benutzt.

	$\mathscr{R}(I - B^{-1}A)$	$\mathscr{R}(I - \hat{B}^{-1}\hat{A})$
$\texttt{cholinc}(A, 0.012)$	$\leq \gamma \approx 0.85099$	$\leq \hat{\gamma} \approx 0.85096$
$\texttt{cholinc}(A, 0.003)$	$\leq \gamma \approx 0.50183$	$\leq \hat{\gamma} \approx 0.50183$
$\texttt{cholinc}(A, 0.001)$	$\leq \gamma \approx 0.17783$	$\leq \hat{\gamma} \approx 0.17783$

Zu diesem modifizierten Eigenwertproblem sind die Schlussphase-Konvergenzfaktoren der Einschittverfahren der PINVIT-Hierarchie von λ_1 und λ_3 abhängig und liegen viel weiter entfernt von 1. Dementsprechend können die resultierenden Rayleigh-Quotient-Folgen relativ schnell konvergieren. Die Rayleigh-Quotient-Folgen der Mehrschittverfahren der PINVIT-Hierarchie haben jeweils nur eine Treppe und benötigen fast gleiche Anzahl der äußeren Schritte. In Abbildung A.3 werden diese Konvergenzverhalten für 100 zufällige Anfangsiterierten zum gleichen Rayleigh-Quotienten 10^5 dargestellt.

Mit der Rang-1-Modifikation $\hat{A} := A + 100(Mv_2)(Mv_2)^T$ lässt sich ein weiteres Beispiel für gut getrennte Eigenwerte konstruieren, wobei alle relevanten Eigenwerte einfach sind. Die Konvergenzverhalten sehen ähnlich aus wie diejenigen in Abbildung A.3.

Gemäß den Ergebnissen von Test I sind die Mehrschittverfahren der PINVIT-Hierarchie anscheinend alle clusterrobust. PINVIT(3) muss nicht viel mehr äußere Schritte als die PINVIT-Verfahren höherer Stufe benötigen und ist ein quasi optimales PINVIT-Verfahren im Sinne des gesamten Aufwands. Für unsere zukünftige Konvergenzanalyse für Mehrschittverfahren der PINVIT-Hierarchie

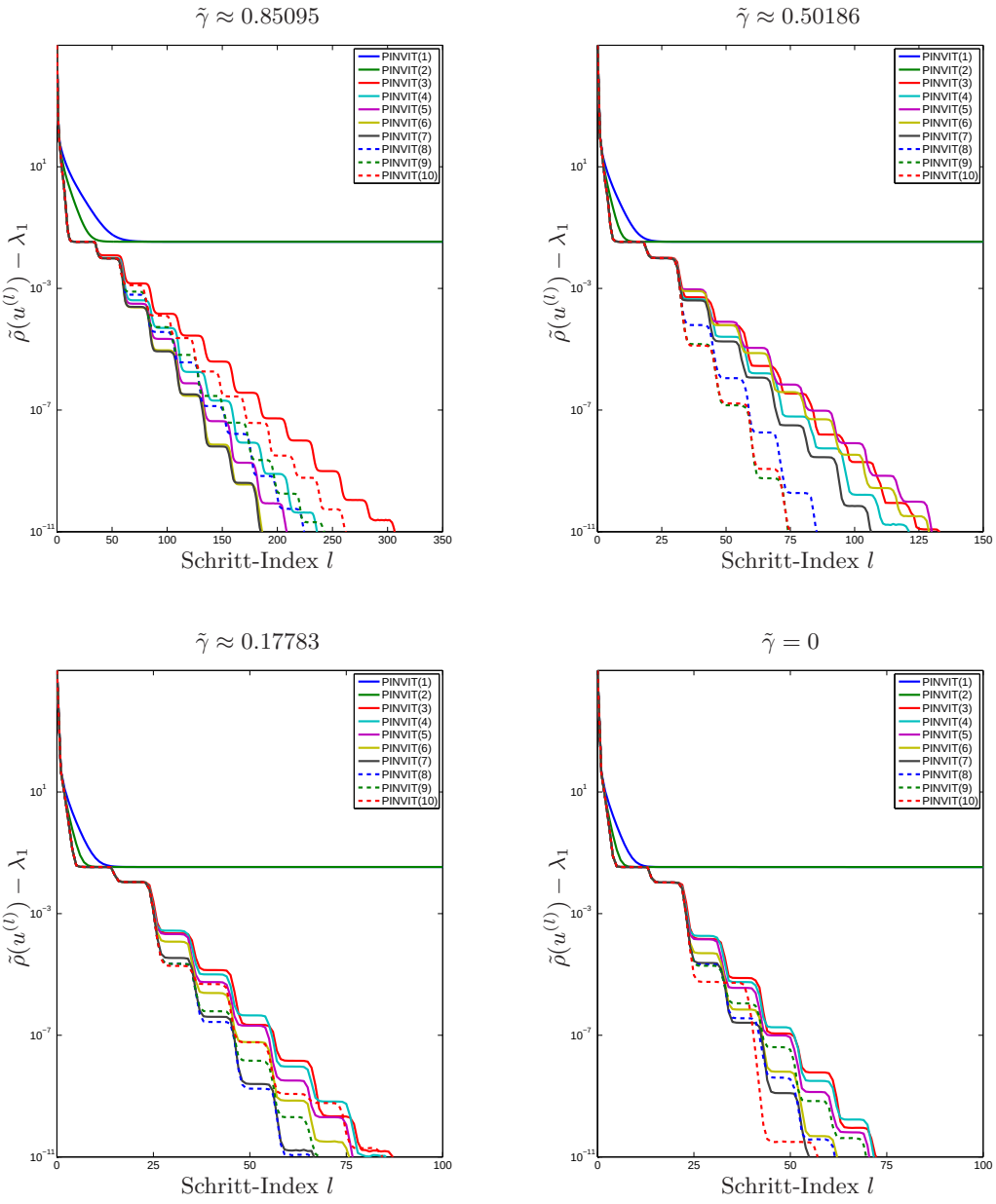


Abbildung A.2: PINVIT-Verfahren bei einem dreielementigen Eigenwert-Cluster

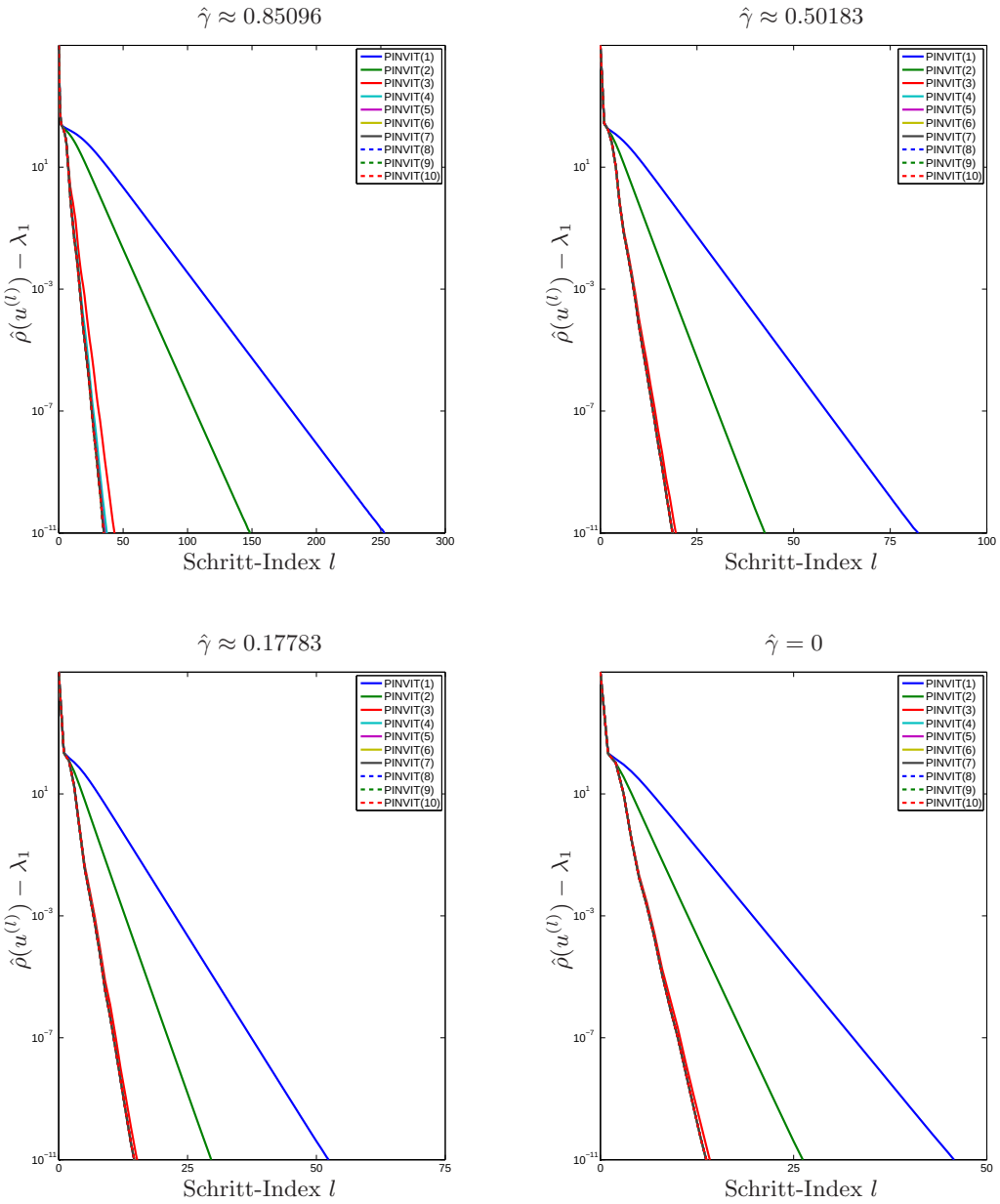


Abbildung A.3: PINVIT-Verfahren bei gut getrennten Eigenwerten

sind die treppenförmigen Konvergenzverhalten bei den Eigenwert-Clustern möglicherweise interessante Argumente.

Test II. Für die Matrixbüschel (A, M) und $(\hat{A}, M)$ aus Test I werden Krylovraum-Verfahren der Form (3.2) durchgeführt, wobei die Krylovräume durch Lanczos-Verfahren mit vollständigen Orthogonalisierungen konstruiert werden, siehe (3.3). In Abbildung A.4 werden die Konvergenzverhalten einiger solchen Verfahren (genannt KRIT) für 100 zufällige Anfangsiterierten zum gleichen Rayleigh-Quotienten 10^5 dargestellt und mit den Konvergenzverhalten von INVIT(1) und INVIT(2) verglichen. Dabei sind diese Krylovraum-Verfahren ab Stufe 6 clusterrobust. Außerdem muss ein Verfahren höherer Stufe nicht eine kleinere Anzahl der äußeren Schritte haben. Das stimmt mit unseren Argumenten, dass die Indizes der relevanten Eigenwerte in der scharfen Schranke der Delta-Abschätzung für diese Krylovraum-Verfahren problemabhängig sind, überein (siehe Satz 3.18).

Krylovraum-Verfahren der Form (3.2) lassen sich auch durch Vorkonditionierer modifizieren. Dafür kann man einfach beim zugehörigen Lanczos-Verfahren (3.3) einen Vorkonditionierer an der Stelle von A^{-1} einsetzen. In Abbildung A.4 werden weiter die Ergebnisse zu den durch `cholinc`$(A, 0.012)$ konstruierten relativ groben Vorkonditionierern dargestellt. Dabei sind diese vorkonditionierten Verfahren (genannt PKRIT) erst ab Stufe 10 clusterrobust.

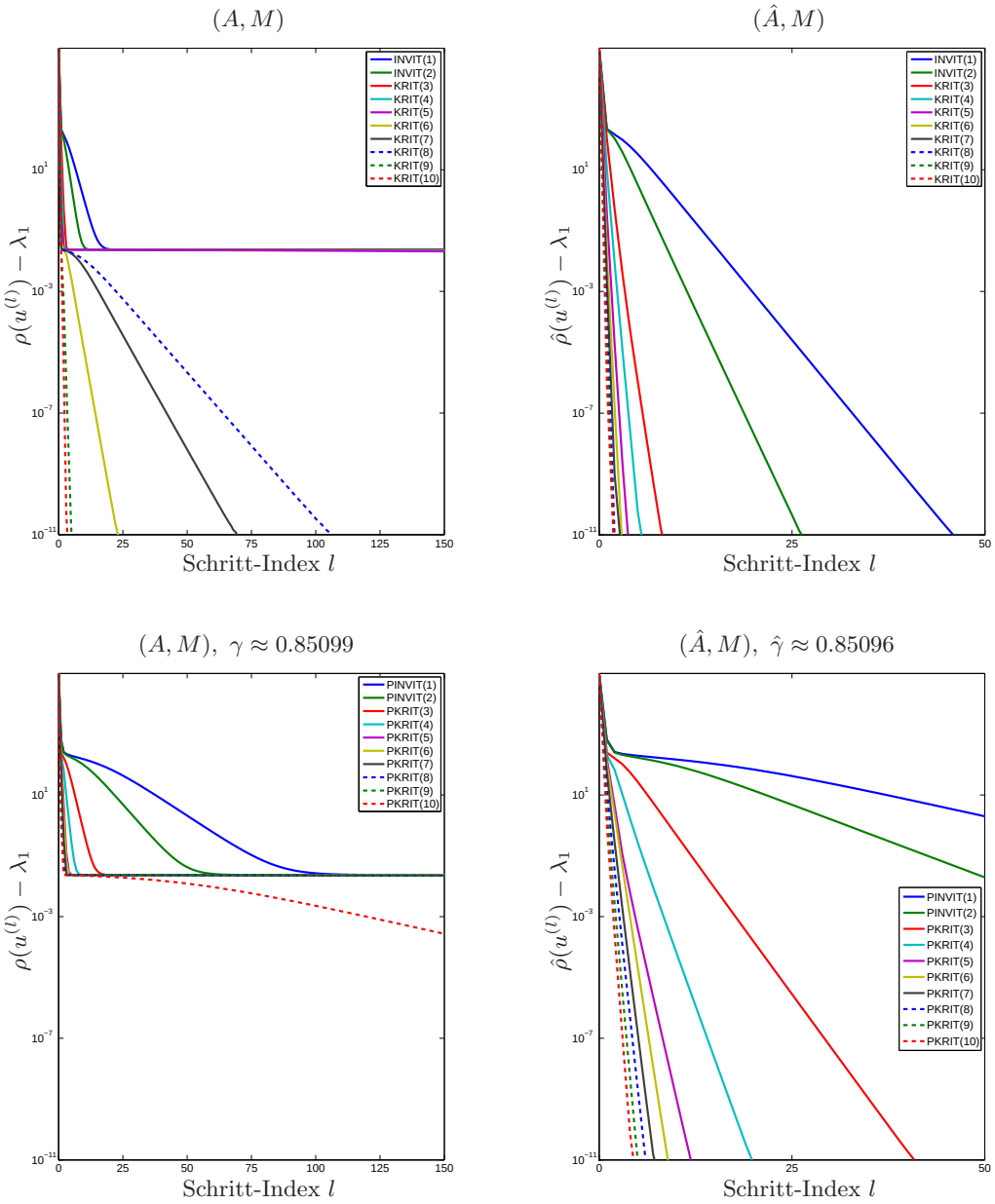


Abbildung A.4: KRIT-Verfahren und PKRIT-Verfahren

Literaturverzeichnis

[1] P.R. Amestoy, T.A. Davis, and I.S. Duff. An approximate minimum degree ordering algorithm. *SIAM J. Matrix Anal. Appl.*, 17(4):886–905, 1996.

[2] P.R. Amestoy, T.A. Davis, and I.S. Duff. Algorithm 837: AMD, an approximate minimum degree ordering algorithm. *ACM Transactions on Mathematical Software*, 30(3):381–388, 2004.

[3] 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. Meth. Engng.*, 64(2):204–236, 2005.

[4] R. Argentati, A. Knyazev, K. Neymeyr, and E. Ovtchinnikov. Preconditioned eigensolver convergence theory in a nutshell. Technical report, in preparation, 2010.

[5] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[6] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the solution of algebraic eigenvalue problems: A practical guide*. SIAM, Philadelphia, 2000.

[7] C.B. Barber, D.P. Dobkin, and H. Huhdanpaa. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[8] M. Beeler, R.W. Gosper, and R. Schroeppel. *HAKMEM, Memo 239*. MIT AI Lab, Cambridge, Massachusetts, 1972.

[9] M. Benzi. Preconditioning techniques for large linear systems: A survey. *J. Comput. Phys.*, 182:418–477, 2002.

[10] R. Bhatia. *Matrix Analysis*. Springer, 1997.

[11] M. Bollhöfer. *ILUPACK Guide*. `ilupack.tu-bs.de`, 2004.

[12] M. Bollhöfer and Y. Saad. Multilevel preconditioners constructed from inverse-based ILUs. *SIAM J. Sci. Comput.*, 27(5):1627–1650, 2006.

[13] A. Borzì and G. Borzì. Algebraic multigrid methods for solving generalized eigenvalue problems. *Int. J. Numer. Meth. Engng.*, 65(8):1186–1196, 2006.

[14] D. Braess. *Finite Elemente*. Springer, 2007.

[15] J.H. Bramble, J.E. Pasciak, and A.V. Knyazev. A subspace preconditioning algorithm for eigenvector/eigenvalue computation. *Adv. Comput. Math.*, 6:159–189, 1996.

[16] J.H. Bramble, J.E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990.

[17] D. Calvetti, L. Reichel, and D.C. Sorensen. An implicit restarted Lanczos method for large symmetric eigenvalue problems. *ETNA, Electron. Trans. Numer. Anal.*, 2:1–21, 1994.

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

[19] J. K. Cullum and R. A. Willoughby. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations. Vol. II: Programs*. Birkhäuser, Boston, 1985.

[20] E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. *Proceedings of the 24th national conference ACM*, pages 157–172, 1969.

[21] B.N. Delaunay. Sur la sphère vide. *Bulletin of Academy of Sciences of the USSR*, 6:793–800, 1934.

[22] E.G. D'yakonov. *Optimization in solving elliptic problems.* CRC Press, Boca Raton, Florida, 1996.

[23] T. Ericsson and A. Ruhe. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Math. Comput.*, 35:1251–1268, 1980.

[24] J.G.F. Francis. The QR Transformation–Part 1. *The Computer Journal*, 4:265–271, 1961.

[25] J.G.F. Francis. The QR Transformation–Part 2. *The Computer Journal*, 4:332–345, 1962.

[26] A. George and J. Liu. *Computer Solution of Large Sparse Positive Definite Systems.* Prentice Hall, Englewood Cliffs, New Jersey, 1981.

[27] J.R. Gilbert, C. Moler, and R. Schreiber. Sparse matrices in MATLAB: Design and implementation. *SIAM J. Matrix Anal. Appl.*, 13(1):333–356, 1992.

[28] G.H. Golub and C.F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1996.

[29] G.H. Golub and F. Uhlig. The QR algorithm: 50 years later its genesis by John Francis and Vera Kublanovskaya and subsequent developments. *IMA Journal of Numerical Analysis*, 29:467–485, 2009.

[30] G.H. Golub and H.A. van der Vorst. Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.*, 123(1–2):35–282, 2000 (Special issue on numerical analysis 2000, Vol. III: Linear algebra).

[31] G.H. Golub and Q. Ye. An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems. *SIAM J. Sci. Comput.*, 24:312–334, 2002.

[32] W. Hackbusch. Ein iteratives Verfahren zur schnellen Auflösung elliptischer Randwertprobleme. Technical report, 76-12, Mathematisches Institut der Universität zu Köln, 1976.

[33] 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(2):201–215, 1979.

[34] W. Hackbusch. *Multi-grid methods and applications.* Springer, 1985.

[35] M.R. Hestenes and W. Karush. A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix. *J. Res. Nat. Bureau Standards*, 47:45–61, 1951.

[36] M.E. Hochstenbach. Generalizations of harmonic and refined Rayleigh-Ritz. *ETNA, Electron. Trans. Numer. Anal.*, 20:235–252, 2005.

[37] C.G.J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *Journal für die reine und angewandte Mathematik*, 30:51–94, 1846.

[38] S. Kaniel. Estimates for some computational techniques in linear algebra. *Math. Comput.*, 20:369–378, 1966.

[39] L.V. Kantorovich. *Functional analysis and applied mathematics.* U. S. Department of Commerce National Bureau of Standards, Los Angeles, Calif., 1952. Translated by C. D. Benster.

[40] L.V. Kantorovich and G.P. Akilov. *Functional analysis in normed spaces.* Pergamon, London, 1964.

[41] A.V. Knyazev. Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem. *Russian J. Numer. Anal. Math. Modelling*, 2:371–396, 1987.

[42] A.V. Knyazev. A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace. Proc. Eigenwertaufgaben in Natur- und Ingenieurwissenschaften und ihre numerische Behandlung, Oberwolfach, 1990. *International Ser. Numerical Mathematics, Birkhauser, Basel*, 96:143–154, 1991.

[43] A.V. Knyazev. Preconditioned eigensolvers - an oxymoron? *ETNA, Electron. Trans. Numer. Anal.*, 7:104–123, 1998.

[44] A.V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comput.*, 23:517–541, 2001.

[45] A.V. Knyazev and M.E. Argentati. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM J. Sci. Comput.*, 23(6):2008–2040, 2002.

[46] A.V. Knyazev and M.E. Argentati. Majorization for changes in angles between subspaces, Ritz values, and graph Laplacian spectra. *SIAM J. Matrix Anal. Appl.*, 29(1):15–32, 2006.

[47] A.V. Knyazev and M.E. Argentati. RayleighRitz majorization error bounds with applications to FEM. *SIAM J. Matrix Anal. Appl.*, 31(3):1521–1537, 2010.

[48] A.V. Knyazev and K. Neymeyr. Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. *ETNA, Electron. Trans. Numer. Anal.*, 15:38–55, 2003.

[49] 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.*, 358:95–114, 2003.

[50] A.V. Knyazev and K. Neymeyr. Gradient flow approach to geometric convergence analysis of preconditioned eigensolvers. *SIAM J. Matrix Anal. Appl.*, 31:621–628, 2009.

[51] A.V. Knyazev and A.L. Skorokhodov. On exact estimates of the convergence rate of the steepest ascent method in the symmetric eigenvalue problem. *Linear Algebra Appl.*, 154–156:245–257, 1991.

[52] V.N. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *Z. Vycisl. Mat. i Mat. Fiz.*, 1:555–570, 1961. English version (ScienceDirect): *USSR Computational Mathematics and Mathematical Physics*, 1(3):637–657, 1962.

[53] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, 45:255–282, 1950.

[54] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, Philadelphia, 1998.

[55] P. Leinen, W. Lembach, and K. Neymeyr. An adaptive subspace method for elliptic eigenproblems with hierarchical basis preconditioning. Technical report, Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, 1997.

[56] D. E. Longsine and S. F. McCormick. Simultaneous Rayleigh-quotient minimization methods for $Ax = \lambda Bx$. *Linear Algebra Appl.*, 34:195–234, 1980.

[57] A. Meyer. *Modern algorithms for large sparse eigenvalue problems*. Akademie-Verlag, Berlin, 1987.

[58] R.B. Morgan. Computing interior eigenvalues of large matrices. *Linear Algebra Appl.*, 154–156:289–309, 1991.

[59] R.B. Morgan and M. Zeng. A harmonic restarted Arnoldi algorithm for calculating eigenvalues and determining multiplicity. *Linear Algebra Appl.*, 415:96–113, 2006.

[60] K. Neymeyr. A geometric theory for preconditioned inverse iteration, I: Extrema of the Rayleigh quotient. *Linear Algebra Appl.*, 332:61–85, 2001.

[61] K. Neymeyr. A geometric theory for preconditioned inverse iteration, II: Convergence estimates. *Linear Algebra Appl.*, 332:87–104, 2001.

[62] K. Neymeyr. *A hierarchy of preconditioned eigensolvers for elliptic differential operators*. Habilitation thesis, Universität Tübingen, Germany, 2001.

[63] K. Neymeyr. A geometric theory for preconditioned inverse iteration applied to a subspace. *Math. Comp.*, 71:197–216, 2002.

[64] K. Neymeyr. A note on inverse iteration. *Numer. Linear Algebra Appl.*, 12:1–8, 2005.

[65] K. Neymeyr. A geometric theory for preconditioned inverse iteration, IV: On the fastest convergence cases. *Linear Algebra Appl.*, 415:114–139, 2006.

[66] K. Neymeyr. On preconditioned eigensolvers and Invert-Lanczos processes. *Linear Algebra Appl.*, 430:1039–1056, 2009.

[67] K. Neymeyr. A geometric convergence theory for the preconditioned steepest descent iteration. Technical report, Universität Rostock, 2010.

[68] K. Neymeyr, E. Ovtchinnikov, and M. Zhou. Convergence analysis of gradient iterations for the symmetric eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 32(2):443–456, 2011.

[69] M.G. Neytcheva and P.S. Vassilevski. Preconditioning of indefinite and almost singular finite element elliptic equations. *SIAM J. Sci. Comput.*, 19(5):1471–1485, 1998.

[70] Y. Notay. Is Jacobi-Davidson faster than Davidson? *SIAM J. Matrix Anal. Appl.*, 26(2):522–543, 2004.

[71] E.E. Ovtchinnikov. Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems, I: The preconditioning aspect. *SIAM J. Numer. Anal.*, 41(1):258–271, 2003.

[72] E.E. Ovtchinnikov. Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems II: The subspace acceleration. *SIAM J. Numer. Anal.*, 41(1):272–286, 2003.

[73] E.E. Ovtchinnikov. Cluster robustness of preconditioned gradient subspace iteration eigensolvers. *Linear Algebra Appl.*, 415:140–166, 2006.

[74] E.E. Ovtchinnikov. Sharp convergence estimates for the preconditioned steepest descent method for Hermitian eigenvalue problems. *SIAM J. Numer. Anal.*, 43(6):2668–2689, 2006.

[75] E.E. Ovtchinnikov. Jacobi correction equation, line search, and conjugate gradients in Hermitian eigenvalue computation, I: Computing an extreme eigenvalue. *SIAM J. Numer. Anal.*, 46(5):2567–2592, 2008.

[76] E.E. Ovtchinnikov. Jacobi correction equation, line search, and conjugate gradients in Hermitian eigenvalue computation, II: Computing several extreme eigenvalues. *SIAM J. Numer. Anal.*, 46(5):2593–2619, 2008.

[77] C.C. Paige. *The computation of eigenvalues and eigenvectors of very large sparse matrices.* PhD thesis, London University, Institute of computer science, 1971.

[78] C.C. Paige, B.N. Parlett, and H.A. van der Vorst. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numer. Linear Algebra Appl.*, 2:115–133, 1995.

[79] B.N. Parlett. *The symmetric eigenvalue problem.* Prentice Hall, Englewood Cliffs, New Jersey, 1980.

[80] A. Perdon and G. Gambolati. Extreme eigenvalues of large sparse matrices by Rayleigh quotient and modified conjugate gradients. *Comput. Methods Appl. Mech. Eng.*, 56:251–264, 1986.

[81] V.G. Prikazchikov. Strict estimates of the rate of convergence of an iterative method of computing eigenvalues. *USSR J. Comput. Math. and Math. Physics*, 15:235–239, 1975.

[82] A. Ruhe. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra Appl.*, 58:391–405, 1984.

[83] Y. Saad. On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM J. Numer. Anal.*, 17(5):687–706, 1980.

[84] Y. Saad. *Numerical methods for large eigenvalue problems.* Manchester University Press, Manchester, 1992.

[85] Y. Saad. *Iterative methods for sparse linear systems.* PWS Publishing Company, Boston, 1996.

[86] B.A. Samokish. The steepest descent method for an eigenvalue problem with semi-bounded operators. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 5:105–114, 1958.

[87] O. Schenk, M. Bollhöfer, and R.A. Römer. On large-scale diagonalization techniques for the anderson model of localization. *SIAM J. Sci. Comput.*, 28(3):963–983, 2006.

[88] G.L.G. Sleijpen and J. van den Eshof. Accurate approximations to eigenpairs using the harmonic Rayleigh-Ritz method. *Linear Algebra Appl.*, 358:115–137, 2003.

[89] G.L.G. Sleijpen and A. van der Sluis. Further results on the convergence behavior of conjugate-gradients and Ritz values. *Linear Algebra Appl.*, 246:233–278, 1996.

[90] G.L.G. Sleijpen and H. A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17(2):401–425, 1996.

[91] G. W. Stewart and J. Sun. *Matrix Perturbation Theory.* Academic Press, New York, 1990.

[92] P.S. Vassilevski. Preconditioning nonsymmetric and indefinite finite element matrices. *Numer. Linear Algebra Appl.*, 1:59–76, 1992.

[93] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil I: Abzählung der Eigenwerte komplexer Matrizen. Technical report, B43/J/9, Aerodynamische Versuchsanstalt Göttingen, 1943.

[94] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil II: Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben. Technical report, B43/J/21, Aerodynamische Versuchsanstalt Göttingen, 1943. *Mathematische Zeitschrift*, 50:93–143, 1944.

[95] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil III: Das Iterationsverfahren in der Flatterrechnung. Technical report, B44/J/21, Aerodynamische Versuchsanstalt Göttingen, 1944.

[96] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil IV: Konvergenzbeweis für das Iterationsverfahren. Technical report, B44/J/38, Aerodynamische Versuchsanstalt Göttingen, 1944.

[97] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil V: Bestimmung höherer Eigenwerte durch gebrochene Iteration. Technical report, B44/J/38, Aerodynamische Versuchsanstalt Göttingen, 1944.

[98] H. Wielandt. *Mathematische Werke, Mathematical Works: Linear algebra and analysis.* Walter de Gruyter, 1996.

[99] J.H. Wilkinson. *The algebraic eigenvalue problem.* Oxford University Press, USA, 1988.

[100] K. Wu and H. Simon. Thick-restart Lanczos method for large symmetric eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 22(2):602–616, 2000.

[101] X. Yang. A survey of various conjugate gradient algorithms for iterative solution of the largest/smallest eigenvalue and eigenvector of a symmetric matrix, Collection: Application of conjugate gradient method to electromagnetic and signal analysis. *Progress Electromagnetic Res.*, 5:567–588, 1991.

[102] H. Yserentant. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49:379–412, 1986.

[103] H. Yserentant. Preconditioning indefinite discretization matrices. *Numer. Math.*, 54(6):719–734, 1989.

[104] Y. Zhang. Solving large-scale linear programs by Interior-Point methods under the MATLAB environment. Technical report, TR96-01, Department of Mathematics and Statistics, University of Maryland Baltimore County, 1996.

[105] P.F. Zhuk. Asymptotic behavior of the s-step method of steepest descent for eigenvalue problems in a Hilbert space. *Russ. Acad. Sci., Sb., Math.*, 80(2):467–495, 1993.

[106] P.F. Zhuk and L.N. Bondarenko. Sharp estimates for the rate of convergence of the s-step method of steepest descent in eigenvalue problems. *Ukraïn. Mat. Zh.*, 49(12):1694–1699, 1997.