Numerik von Eigenwertaufgaben

Steffen Börm

Stand 10. Februar 2018

Inhaltsverzeichnis

1 Einleitung

Ein *Eigenwert* eines linearen Operators $L : V \to V$ auf einem K-Vektorraum V ist ein Element $\lambda \in K$ des zugehörigen Körpers, für das ein von null verschiedener Vektor $u \in V \setminus \{0\}$ mit

$$Lu = \lambda u \tag{1.1}$$

existiert. Diesen Vektor u nennt man einen *Eigenvektor* zu λ, das Paar (λ, u) nennt man ein *Eigenpaar* des Operators L.

Eigenwerte sind in naturwissenschaftlichen Anwendungen von Interesse, beispielsweise bei der Untersuchung des Resonanzverhaltens eines schwingenden Systems, aber auch bei der Klassifikation von Dokumenten, beispielsweise der Sortierung der Ergebnisse von Internet-Suchmaschinen, und bei der mathematischen Analyse von linearen Gleichungssystemen und linearen Anfangswertproblemen.

Obwohl die Gleichung (1.1) auf den ersten Blick einem gewöhnlichen linearen Gleichungssystem ähnelt, stellt sich bei genauerer Betrachtung heraus, dass durch den Zusammenhang zwischen λ und u ein nichtlineares System entsteht, das sich im Allgemeinen nicht mehr mit einem aus endlich vielen Rechenoperationen bestehenden Algorithmus lösen lässt.

Stattdessen kommen *iterative Verfahren* zum Einsatz, die beispielsweise eine Folge von Näherungen eines Eigenvektors oder Eigenwerts berechnen, die gegen exakte Lösungen konvergieren.

Dabei sind verschiedene Aufgabenstellung zu unterscheiden: In manchen Anwendungen ist man nur daran interessiert, einen bestimmten Eigenwert zu berechnen, beispielsweise um die niedrigste Resonanzfrequenz eines schwingungsfähigen Systems zu ermitteln. In anderen sind eine kleine Anzahl der größten oder kleinsten Eigenwerte von Interesse oder Eigenwerte, die in der Nähe eines gegebenen Punkts in der komplexen Ebene liegen. In wieder anderen ist eine vollständige Orthonormalbasis des gesamten Raums gesucht, mit deren Hilfe sich der Operator diagonalisieren lässt, um weitere Berechnungen zu vereinfachen.

Um die unterschiedlichen Anforderungen zu erfüllen werden eine Reihe von algorithmischen Ansätzen verwendet, beispielsweise lässt sich durch Potenzieren des Operators L eine Eigenvektor-Näherung bestimmen, Eigenwerte können als Minima oder Maxima geeigneter Funktionen beschrieben werden, alternativ im endlich-dimensionalen Fall auch als Nullstellen eines charakteristischen Polynoms.

Eine besondere Herausforderung stellt die Behandlung großer Matrizen dar, die beispielsweise bei der Analyse strukturmechanischer oder elektromagnetischer Schwingungen von großer Bedeutung sind. In diesem Fall ist L ein Differentialoperator, der im Rahmen einer Diskretisierung durch eine Matrix approximiert wird, deren kleinste Eigenwerte und zugehörige Eigenvektoren gesucht werden. Die Matrix ist in der Regel

schlecht konditioniert, so dass spezialisierte Iterationsverfahren zum Einsatz kommen müssen, die mit den bereits erwähnten verwandt sind, allerdings die besondere Form der Aufgabe berücksichtigen.

Ziel dieser Vorlesung ist es, einen Überblick über die grundlegende Theorie, die wichtigstens Verfahren und deren Analyse zu geben.

Danksagung

Ich bedanke mich bei Sabrina Reif und Robin-Thomas Léger für Hinweise auf Fehler in früheren Fassungen dieses Skripts und für Verbesserungsvorschläge.

2 Beispiele

In diesem Kapitel werden drei Beispiele für Eigenwertprobleme in der Praxis gegeben. Gleichzeitig werden drei Typen von Eigenwertproblemen charakterisiert: Die Berechnung eines einzelnen Eigenwert-Eigenvektor-Paares, die Berechnung von wenigen solchen Paaren und die Berechnung aller solcher Paare, also die Schur-Zerlegung einer Matrix.

2.1 Schwingende Saite

Wir untersuchen eine horizontal gespannte Saite der Länge ℓ. Ihre vertikale Auslenkung in Abhängigkeit von Ort und Zeit wird durch eine Funktion $u \in C^2([0,\ell] \times \mathbb{R}_{\geq 0})$ modelliert. Da die Saite links und rechts eingespannt ist, gelten die Randbedingungen

$$u(0,\cdot) = 0 \quad \text{und} \quad u(\ell,\cdot) = 0. \tag{2.1}$$

Das (vereinfachte) Verhalten der Saite wird durch die *Wellengleichung*

$$\frac{\partial^2 u}{\partial t^2}(x,t) = c\frac{\partial^2 u}{\partial x^2}(x,t) + f(x,t) \tag{2.2}$$

für alle $(x,t) \in]0,\ell[\times\mathbb{R}_{\geq 0}$ bestimmt. Hierbei bezeichnet $c \in \mathbb{R}$ einen Materialparameter, in den etwa Eigenschaften wie die Dicke der Saite, ihre Elastizität und das Maß der aufgewendeten Spannkraft eingehen. Die Funktion $f \in C([0,\ell] \times \mathbb{R}_{\geq 0})$ beschreibt die von Außen auf die Saite ausgeübte Kraft.

Wir interessieren uns für den Fall, in dem die äußere Kraft lediglich der Anregung dient und die Saite anschließend ohne weitere Beeinflussung schwingt. Wir sind also an Lösungen der Gleichung

$$\frac{\partial^2 u}{\partial t^2}(x,t) = c\frac{\partial^2 u}{\partial x^2}(x,t) \tag{2.3}$$

interessiert. Da die auftretenden Differentialoperatoren linear sind und keine vom Ort abhängigen Koeffizienten auftreten, entscheiden wir uns für den folgenden Separationsansatz:

$$u(x,t) = u_0(x)\cos(\omega t) \tag{2.4}$$

für eine Funktion $u_0 \in C^2([0,\ell])$. Dann gilt

$$\frac{\partial^2 u}{\partial t^2}(x,t) = -\omega^2 u(x,t)$$

für alle $(x,t) \in]0,\ell[\times\mathbb{R}_{\geq 0}$ und die Wellengleichung (2.3) nimmt die Form

$$c\frac{\partial^2 u}{\partial x^2}(x,t) = -\omega^2 u(x,t) \tag{2.5}$$

an. Mit den Abkürzungen

$$L := -c\frac{\partial^2}{\partial x^2} \quad \text{und} \quad \lambda := \omega^2$$

und durch Elimination des Cosinus aus u erhalten wir die Gleichung

$$Lu_0 = \lambda u_0. \tag{2.6}$$

Sie besitzt offenbar immer die triviale Lösung $u_0 = 0$, die einer ruhenden Saite entspricht. Für eine schwingende Saite fordern wir deshalb $u_0 \neq 0$ und erhalten das folgende Problem:

> Finde $u_0 \in C^2([0,\ell]) \setminus \{0\}$ und $\lambda \in \mathbb{R}$ derart, dass
>
> $$Lu_0 = \lambda u_0$$
>
> erfüllt ist.

In diesem Kontext bezeichnet man λ als *Eigenwert* und u_0 als *Eigenvektor* (beziehungsweise in diesem Fall als *Eigenfunktion*).

Da der Separationsansatz zur Elimination der Zeit aus der Gleichung (2.3) hilfreich war, liegt es nahe, ihn auch auf das Eigenwertproblem anzuwenden. Wir nehmen an, dass

$$u_0(x) = \sin(\alpha x)$$

für ein $\alpha \in \mathbb{R}$ gilt. Da diese Funktion nur dann gleich Null ist, falls $\alpha x \in \pi\mathbb{Z}$ erfüllt ist, folgt aus der Randbedingung (2.1) die Gleichung

$$\alpha\ell \in \pi\mathbb{Z}, \quad \text{also} \quad \alpha = k\pi/\ell$$

für ein $k \in \mathbb{Z}$. Der Operator L lässt sich für ein u_0 von dieser Gestalt als

$$Lu_0(x) = c\alpha^2 u_0(x)$$

schreiben, so dass aus der Gleichung (2.6) die Beziehung

$$c\alpha^2 = \lambda = \omega^2$$

folgt. Wir können also zu jedem α den entsprechenden Eigenwert λ und die korrespondierende *Eigenfrequenz* $\omega = \sqrt{\lambda}$ berechnen und erhalten das Ergebnis, dass für jedes $k \in \mathbb{N}$

$$\omega = \sqrt{c}\frac{\pi}{\ell}k$$

eine Eigenfrequenz ist, die zu einem nicht-trivalen Eigenvektor

$$u_0(x) = \sin(k\pi/\ell)$$

gehört.

Falls der Materialparameter c in der Wellengleichung nicht konstant, sondern vom Ort abhängig ist (falls sich beispielsweise die Dicke der Saite ändert), lässt sich der Separationsansatz $u_0(x) = \sin(\alpha x)$ nicht mehr anwenden.

In diesem Fall behilft man sich mit einer numerischen Approximation des Operators L, um aus der kontinuierlichen Gleichung (2.6) eine Matrixgleichung zu machen, die dann mit Standardverfahren behandelt werden kann.

Ein Ansatz für diese Approximation ist die Finite-Differenzen-Methode, die im Wesentlichen auf der Taylor-Entwicklung von u_0 basiert: Für $x \in]0, \ell[$ und $h \in \mathbb{R}$ mit $x + h, x - h \in]0, \ell[$ gibt es $\eta_+ \in]x, x + h[$ und $\eta_- \in]x - h, x[$ derart, dass

$$u_0(x + h) = u_0(x) + hu_0'(x) + h^2u_0''(x)/2 + h^3u_0^{(3)}(x)/6 + h^4u_0^{(4)}(\eta_+)/24,$$
$$u_0(x - h) = u_0(x) - hu_0'(x) + h^2u_0''(x)/2 - h^3u_0^{(3)}(x)/6 + h^4u_0^{(4)}(\eta_-)/24$$

gelten, wobei $u_0^{(3)}$ und $u_0^{(4)}$ die dritte und vierte Ableitung von u_0 bezeichnen. Durch Addition dieser Gleichungen erhalten wir

$$u_0(x + h) + u_0(x - h) = 2u_0(x) + h^2u_0''(x) + h^4(u_0^{(4)}(\eta_+) + u_0^{(4)}(\eta_-))/24$$

und können folgern, dass

$$Du_0(x, h) := (2u_0(x) - u_0(x + h) - u_0(x - h))/h^2 \tag{2.7}$$

eine Approximation für $-u_0''(x)$ ist, die die Fehlerabschätzung

$$|Du_0(x, h) + u_0''(x)| \leq h^2\|u_0^{(4)}\|_\infty/12 \tag{2.8}$$

erfüllt.

Um nun L zu approximieren, wählen wir eine äquidistante Partitionierung $0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = \ell$ des Intervalls $[0, \ell]$: Wir setzen $x_i = hi$ mit $h = \ell/(n+1)$. Für jedes $i \in \{1, \ldots, n\}$ ersetzen wir dann $-u_0''(x_i)$ durch

$$Du_0(x_i, h) = (2u_0(x_i) - u_0(x_{i+1}) - u_0(x_{i-1})/h^2.$$

Indem wir die Werte von u_0 in den Punkten $x_1, \ldots, x_n$ in dem Vektor $\mathbf{u}_0 = (u_0(x_i))_{i=1}^n$ zusammenfassen, nimmt das Eigenwertproblem (2.6) die Form

$$\mathbf{A}\mathbf{u}_0 = \lambda\mathbf{u}_0 \tag{2.9}$$

mit der Tridiagonalmatrix

$$\mathbf{A} = \frac{c}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \tag{2.10}$$

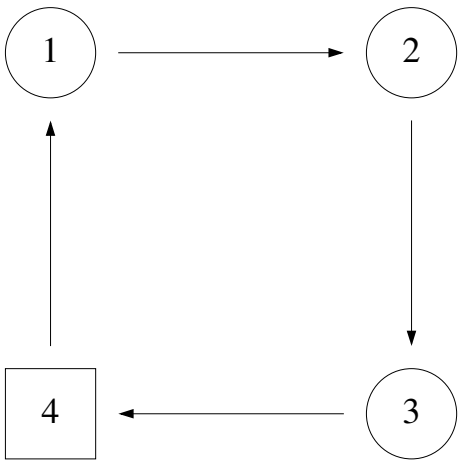
an. Damit ist das kontinuierliche Problem in ein diskretes Problem in $\mathbb{R}^n$ überführt. Da die Genauigkeit der Approximation wesentlich von dem Diskretisierungsparameter h abhängt, der wiederum proportional zum Kehrwert der Problemdimension n ist, ist klar, dass eine hohe Genauigkeit bei der Berechnung der Eigenwerte und Eigenvektoren auch eine hohe Dimension des Gesamtproblems erfordert.

Andererseits ist auch klar, dass beispielsweise für $u_0(x) = \sin(\alpha x)$ die für die Fehlerabschätzung (2.8) wichtige vierte Ableitung von u_0 sich wie α^4 verhalten wird. Damit der Fehler beschränkt bleibt, muss also $h^2 \approx \alpha^{-4}$ gelten, was zur Folge hat, dass Eigenvektoren zu höheren Frequenzen auch wesentlich feinere Gitterschrittweiten erfordern.

Aus diesem Grund wird man sich in der Regel nur für eine geringe Anzahl der zu den kleinsten Eigenfrequenzen gehörenden Eigenvektoren interessieren, man wird also nicht darauf abzielen, *alle* Eigenwerte und Eigenvektoren zu bestimmen, sondern nur eine sehr kleine und von n unabhängige Anzahl.

2.2 Minipoly

Als nächstes wenden wir uns einer stark reduzierten Version eines altbekannten Brettspiels zu. Das Spielfeld besteht aus vier Feldern und hat die folgende Form:



Die Spielregeln sind einfach:

- Zu Beginn steht eine Spielfigur auf einem beliebigen Feld.
- Falls die Figur auf einem der Felder 1, 2 oder 3 steht, wird ein üblicher sechsseitiger Würfel geworfen, dessen Augenzahl angibt, wieviele Schritte im Uhrzeigersinn durchzuführen sind.
- Falls die Figur auf dem „Gefängnisfeld“ 4 steht, wird wieder gewürfelt. Falls eine 6 herauskommt, kann die Figur auf Feld 1 fliehen, sonst bleibt sie im Gefängnis, also auf Feld 4.

Da die Position der Spielfigur nach einer Reihe von Schritten vom Zufall abhängt, können wir keine präzisen Vorhersagen treffen. Wir können allerdings nach der *Wahrscheinlichkeit* fragen, mit der eine Spielfigur auf einem bestimmten Feld stehen wird.

Eine ähnliche Technik wird beispielsweise bei der Bewertung von Internet-Seiten durch Suchmaschinen verwendet: Die ursprüngliche Idee der Suchmaschine Google beruhte auf dem Modell eines Anwenders, der ausgehend von einer Webseite zufällig einen der auf dieser Seite enthaltenen Verweise anklickt. Diejenigen Seiten, auf denen sich der simulierte Anwender mit hoher Wahrscheinlichkeit aufhielt, wurden als besonders attraktiv bewertet und in der Ergebnisliste der Suchmaschine als erste angegeben.

Die Folge der von einer Spielfigur eingenommenen Positionen bildet eine sogenannte *Markoff-Kette*, deren wahrscheinlichkeitstheoretische Eigenschaften durch die Übergangswahrscheinlichkeiten zwischen den einzelnen Feldern eindeutig bestimmt ist.

So beträgt die Wahrscheinlichkeit, von Feld 1 auf Feld 1 zu wechseln, gerade $1/6$, da dazu eine 4 gewürfelt werden muss. Um von Feld 1 auf Feld 2 zu wechseln, genügen dagegen eine 1 *oder* eine 5, so dass diese Wahrscheinlichkeit $1/3$ beträgt. Wir können die Übergangswahrscheinlichkeiten in einer Matrix $\mathbf{P} \in \mathbb{R}^{4\times 4}$ zusammenfassen, indem wir in der j-ten Spalte und i-ten Zeile eintragen, wie hoch die Wahrscheinlichkeit ist, von Feld j auf Feld i zu wechseln:

$$\mathbf{P} := \begin{pmatrix} 1/6 & 1/6 & 1/3 & 1/6 \\ 1/3 & 1/6 & 1/6 & 0 \\ 1/3 & 1/3 & 1/6 & 0 \\ 1/6 & 1/3 & 1/3 & 5/6 \end{pmatrix} \tag{2.11}$$

Für den Vektor $\mathbf{e}_2 = (0, 1, 0, 0)$ gibt $\mathbf{P}\mathbf{e}_2$ an, mit welcher Wahrscheinlichkeit die Spielfigur sich auf den einzelnen Feldern befindet, wenn von Feld 2 ausgehend ein Zug durchgeführt wurde. Der Vektor $\mathbf{P}^2\mathbf{e}_2$ beschreibt die Wahrscheinlichkeitsverteilung nach 2 Zügen, der Vektor $\mathbf{P}^n\mathbf{e}_2$ die nach n Zügen. Die Matrix $\mathbf{P}$ ordnet also der Wahrscheinlichkeitsverteilung in einem Zug die für den folgenden Zug zu.

Es stellt sich die Frage, ob ein Grenzwert existiert, ob sich also nach einer gewissen Anzahl von Spielzügen eine stabile Wahrscheinlichkeitsverteilung einstellt. Ein solches sogenanntes *invariantes Wahrscheinlichkeitsmaß* $\mathbf{w}$ ist durch die Fixpunktgleichung

$$\mathbf{P}\mathbf{w} = \mathbf{w} \tag{2.12}$$

charakterisiert. Offensichtlich kann $\mathbf{w}$ auch als Eigenvektor zum Eigenwert 1 aufgefasst werden, wir erhalten also wieder ein Eigenwertproblem, diesmal allerdings mit bekanntem Eigenwert und unbekanntem Eigenvektor.

Existenzsätze für Eigenwerte lassen sich auf die Matrix $\mathbf{P}$ anwenden und führen zu dem Ergebnis, dass sie einen Eigenwert 1 besitzt, zu dem es einen Eigenvektor mit positiven Koeffizienten gibt. Mit geeigneter Skalierung kann er als das gesuchte invariante Maß interpretiert werden.

Es lassen sich außerdem Standardverfahren anwenden, um diesen Eigenvektor zu be-

rechnen: Man erhält näherungsweise

$$\mathbf{w} \approx \begin{pmatrix} 0.185 \\ 0.097 \\ 0.113 \\ 0.605 \end{pmatrix},$$

die Spielfigur wird also im Durchschnitt über 60% der Spielzüge im Gefängnis verbringen und am seltensten auf dem Feld 2 anzutreffen sein.

Das wiederholte Anwenden von $\mathbf{P}$ auf Wahrscheinlichkeitsverteilungen lässt sich als ein Verfahren zur Bestimmung des größten Eigenwerts, in diesem Fall 1, interpretieren, so dass sich die im Rahmen der Analyse dieses Verfahrens erzielten Konvergenzaussagen direkt auf das Beispiel übertragen.

Das in diesem Kontext auftretende Eigenwertproblem kann als Spezialfall des im Falle der schwingenden Saite behandelten gesehen werden: Es wird lediglich nach *einem* Eigenvektor zu einem bestimmten Eigenwert gesucht.

2.3 Lineares Anfangswertproblem

Zum Abschluss des Beispiel-Kapitels untersuchen wir ein lineares Anfangswertproblem: Wir möchten zu einer Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ und einem Vektor $\mathbf{y}_0 \in \mathbb{R}^n$ eine vektorwertige Funktion $\mathbf{y} \in C^1(\mathbb{R}_{\geq 0}, \mathbb{R}^n)$ finden, die die Gleichungen

$$\mathbf{y}(0) = \mathbf{y}_0 \quad \text{und} \quad \mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) \tag{2.13}$$

für alle $t \in \mathbb{R}_{>0}$ erfüllt. Aus der Analysis ist bekannt, dass sich $\mathbf{y}$ durch

$$\mathbf{y}(t) = \exp(t\mathbf{A})\mathbf{y}_0$$

ausdrücken lässt. Um nun für ein konkretes $t \in \mathbb{R}_{>0}$ den Wert $\mathbf{y}(t)$ zu bestimmen, müssen wir die Exponentialfunktion der Matrix $t\mathbf{A}$ auswerten. Per Taylor-Entwicklung um 0 erhalten wir die Beziehung

$$\exp(t\mathbf{A}) = \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbf{A}^i.$$

Für numerische Zwecke ist diese Formel unbrauchbar, da die Berechnung der Matrizen $\mathbf{A}^i$ für i zwischen 0 und einer Obergrenze p immerhin $\mathcal{O}(pn^3)$ Operationen erfordert und p infolge der langsamen Konvergenz der Exponentialreihe für große Werte von t ebenfalls groß sein muss.

Das Problem lässt sich wesentlich vereinfachen, wenn man voraussetzt, dass $\mathbf{y}_0$ ein Eigenvektor von $\mathbf{A}$ zum Eigenwert λ ist, denn dann gilt

$$\mathbf{A}\mathbf{y}_0 = \lambda\mathbf{y}_0, \qquad \mathbf{A}^2\mathbf{y}_0 = \lambda^2\mathbf{y}_0, \qquad \mathbf{A}^i\mathbf{y}_0 = \lambda^i\mathbf{y}_0,$$

also lässt sich die Exponentialfunktion wesentlich einfacher darstellen:

$$\mathbf{y}(t) = \exp(t\mathbf{A})\mathbf{y}_0 = \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbf{A}^i\mathbf{y}_0 = \sum_{i=0}^{\infty} \frac{t^i}{i!}\lambda^i\mathbf{y}_0 = \exp(t\lambda)\mathbf{y}_0,$$

die Auswertung von $\mathbf{y}(t)$ erfordert also nur noch die Berechnung des Werts der Exponentialfunktion für eine reelle Zahl. Diese Aufgabe lässt sich mit einem Aufwand von höchstens $\mathcal{O}(p)$ bewerkstelligen, ist also *wesentlich* günstiger als der ursprüngliche Ansatz.

In der Praxis wird der Startvektor $\mathbf{y}_0$ eher selten ein Eigenvektor von $\mathbf{A}$ sein. Falls eine Basis $(\mathbf{v}_\ell)_{\ell=1}^n$ aus Eigenvektoren zu den Eigenwerten $(\lambda_\ell)_{\ell=1}^n$ von $\mathbf{A}$ existiert, lässt sich $\mathbf{y}_0$ durch

$$\mathbf{y}_0 = \sum_{\ell=1}^{n} \alpha_\ell \mathbf{v}_\ell$$

darstellen, so dass sich

$$\mathbf{y}(t) = \exp(t\mathbf{A})\mathbf{y}_0 = \sum_{\ell=1}^{n} \alpha_\ell \exp(t\mathbf{A})\mathbf{v}_\ell = \sum_{\ell=1}^{n} \alpha_\ell \exp(t\lambda_\ell)\mathbf{v}_\ell$$

ergibt und sich somit mit einem Aufwand von $\mathcal{O}(np + n^2)$ berechnen lässt, falls die Koeffizienten α_ℓ vorliegen.

In der Regel werden die Koeffizienten nicht vorliegen, so dass sie erst berechnet werden müssen. Für eine allgemeine Basis $(\mathbf{v}_\ell)_{\ell=1}^n$ erfordert das einen Aufwand von $\mathcal{O}(n^3)$, ist also unattraktiv.

Falls die Basis $(\mathbf{v}_\ell)_{\ell=1}^n$ allerdings *orthonormal* ist, kann auch diese Hürde überwunden werden, weil dann

$$\alpha_\ell = \langle \mathbf{v}_\ell, \mathbf{y}_0 \rangle$$

gilt und sich die Koeffizienten also mit einem Aufwand von $\mathcal{O}(n^2)$ bestimmen lassen.

Für die Berechnung von $\mathbf{y}(t)$ sind wir also daran interessiert, *alle* Eigenwerte und Eigenvektoren einer Matrix $\mathbf{A}$ zu berechnen und sicherzustellen, dass die Eigenvektoren eine orthonormale Basis bilden.

Schreibt man die Eigenvektoren $(\mathbf{v}_\ell)_{\ell=1}^n$ als Spalten einer Matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$, so folgt aus

$$\mathbf{A}\mathbf{Q}\mathbf{e}_i = \mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i = \lambda_i\mathbf{Q}\mathbf{e}_i$$

die Gleichung

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \underbrace{\begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \lambda_{n-1} & \\ & & & & \lambda_n \end{pmatrix}}_{=:\mathbf{D}},$$

wir haben also eine orthonormale Matrix $\mathbf{Q}$ gefunden, die $\mathbf{A}$ diagonalisiert. Daraus lässt sich wiederum

$$\mathbf{A} = \mathbf{QDQ}^T = \mathbf{Q}^{TT}\mathbf{D}^T\mathbf{Q}^T = (\mathbf{QDQ}^T)^T = \mathbf{A}^T$$

ableiten, also die negative Aussage, dass $\mathbf{A}$ symmetrisch sein muss, damit eine orthonormale Basis existiert, die diese Matrix diagonalisiert. Wir werden später nachweisen, dass diese Bedingung bereits hinreichend ist.

3 Theoretische Grundlagen

Dieses Kapitel hat zwei Ziele: Einerseits sollen die elementaren Aussagen über die Eigenwerte und Eigenvektoren quadratischer Matrizen zur Verfügung gestellt werden, andererseits werden einige grundlegende Begriffe aus der linearen Algebra rekapituliert, die für die Untersuchung der in den späteren Kapiteln eingeführten Verfahren nützlich sein werden.

3.1 Existenz von Eigenwerten

Wir arbeiten im Folgenden im Körper $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ der reellen oder komplexen Zahlen. An einigen Stellen greifen wir auf den Fundamentalsatz der Algebra zurück, so dass wir uns auf den Fall $\mathbb{K} = \mathbb{C}$ beschränken müssen.

Die Dimensionen der behandelten Matrizen werden wir mit $n, m \in \mathbb{N}$ bezeichnen.

Definition 3.1 (Eigenwerte und Eigenvektoren) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Eine Zahl $\lambda \in \mathbb{K}$ heißt* Eigenwert *von $\mathbf{A}$ genau dann, wenn es einen Vektor $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ gibt, der die Gleichung*

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{3.1}$$

erfüllt. Jeden derartigen Vektor bezeichnet man als Eigenvektor *der Matrix $\mathbf{A}$ zu dem Eigenwert λ.*

Falls $\mathbf{x}$ und $\mathbf{y}$ Eigenvektoren einer Matrix $\mathbf{A}$ zum gleichen Eigenwert λ sind, gilt dasselbe für alle von null verschiedenen Linearkombinationen der Vektoren:

$$\mathbf{A}(\mathbf{x} + \alpha\mathbf{y}) = \mathbf{A}\mathbf{x} + \alpha\mathbf{A}\mathbf{y} = \lambda\mathbf{x} + \alpha\lambda\mathbf{y} = \lambda(\mathbf{x} + \alpha\mathbf{y}).$$

Insbesondere sind Eigenvektoren zu einem Eigenwert niemals eindeutig bestimmt, stattdessen definieren sie einen Teilraum.

Definition 3.2 (Eigenräume) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$, sei $\lambda \in \mathbb{K}$ ein Eigenwert von $\mathbf{A}$. Dann bezeichnet $\mathcal{E}_A(\lambda)$ den Raum, der von den Eigenvektoren von $\mathbf{A}$ zu λ aufgespannt wird. Es gilt*

$$\mathcal{E}_A(\lambda) := \operatorname{Kern}(\lambda\mathbf{I} - \mathbf{A}).$$

Der Raum $\mathcal{E}_A(\lambda)$ heißt Eigenraum *von $\mathbf{A}$ zu dem Eigenwert λ.*

Lemma 3.3 (Charakterisierung der Eigenwerte) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und sei $\lambda \in \mathbb{K}$. Es sind äquivalent:*

1. *λ ist ein Eigenwert von $\mathbf{A}$,*
2. *$\lambda\mathbf{I} - \mathbf{A}$ ist singulär,*
3. $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$.

Beweis. „1 $\Rightarrow$ 2“: Sei zunächst λ ein Eigenwert. Dann gilt insbesondere

$$\mathrm{Kern}(\lambda\mathbf{I} - \mathbf{A}) = \mathcal{E}_A(\lambda) \neq \{\mathbf{0}\},$$

also ist $\lambda\mathbf{I} - \mathbf{A}$ nicht injektiv und damit insbesondere singulär.
„2 $\Rightarrow$ 1“: Sei nun $\lambda\mathbf{I}-\mathbf{A}$ singulär. Damit existiert insbesondere ein $\mathbf{x} \in \mathrm{Kern}(\lambda\mathbf{I}-\mathbf{A})\backslash\{\mathbf{0}\}$, und es folgt $\lambda\mathbf{x} = \mathbf{A}\mathbf{x}$, wir haben also einen Eigenvektor zu λ gefunden.
„2 $\Leftrightarrow$ 3“: Die Determinante einer Matrix verschwindet genau dann, wenn die Matrix singulär ist. ∎

Mit Hilfe der dritten Eigenschaft aus Lemma 3.3 können wir die Eigenwerte einer Matrix als Nullstellen eines Polynoms charakterisieren:

Definition 3.4 (Charakteristisches Polynom) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Dann ist*

$$p_A(\lambda) := \det(\lambda\mathbf{I} - \mathbf{A})$$

ein Polynom n-ten Grades. Es heißt das charakteristische Polynom *der Matrix $\mathbf{A}$.*

Lemma 3.5 (Nullstellen von p_A) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Dann ist $\lambda \in \mathbb{K}$ genau dann eine Nullstelle von p_A, wenn λ ein Eigenwert von $\mathbf{A}$ ist.*

Beweis. Folgt direkt aus Lemma 3.3. ∎

Übungsaufgabe 3.6 (Begleitmatrix) *(vgl. [2, Abschnitt 7.4.6]) Seien $n \in \mathbb{N}$ und $c_0, \ldots, c_{n-1} \in \mathbb{K}$ gegeben, und sei*

$$\mathbf{A} := \begin{pmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{pmatrix}.$$

Beweisen Sie

$$p_A(\lambda) = c_0 + c_1\lambda + \ldots + c_{n-1}\lambda^{n-1} + \lambda^n \qquad \text{für alle } \lambda \in \mathbb{K}.$$

Die Suche nach den Nullstellen eines beliebigen Polynoms lässt sich also immer auf ein Eigenwertproblem zurückführen.

Hinweis: *Man könnte den Laplace'schen Entwicklungssatz auf die erste Spalte der Matrix $\lambda\mathbf{I} - \mathbf{A}$ anwenden.*

Bevor wir aus dieser Charakterisierung der Eigenwerte einen ersten Existenzsatz gewinnen können, führen wir einige hilfreiche Begriffe ein:

Definition 3.7 (Spektrum, Vielfachheiten) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Die Menge*

$$\sigma(\mathbf{A}) = \{\lambda \in \mathbb{K} \ : \ \lambda\mathbf{I} - \mathbf{A} \text{ ist singulär}\}$$

heißt Spektrum *der Matrix $\mathbf{A}$. Für jedes $\lambda \in \sigma(\mathbf{A})$ bezeichnen wir mit $\mu_A^a(\lambda)$ die Vielfachheit der Nullstelle λ des Polynoms p_A und mit $\mu_A^g(\lambda)$ die Dimension des Kerns von $\lambda\mathbf{I} - \mathbf{A}$. $\mu_A^a(\lambda)$ und $\mu_A^g(\lambda)$ heißen* algebraische *und* geometrische Vielfachheit *des Eigenwerts λ.*

Nun lässt sich für den *komplexwertigen* Fall eine erste Existenzaussage für Eigenwerte mit Hilfe des Fundamentalsatzes der Algebra gewinnen:

Satz 3.8 (Existenz von Eigenwerten) *Sei $\mathbf{A} \in \mathbb{C}^{n\times n}$. Dann ist $\sigma(\mathbf{A}) \neq \emptyset$ und es gilt*

$$\sum_{\lambda\in\sigma(\mathbf{A})} \mu_A^a(\lambda) = n.$$

Beweis. Gemäß Fundamentalsatz der Algebra zerfällt p_A in Linearfaktoren, es gibt also $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ mit

$$p_A(\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i) \qquad \text{für alle } \lambda \in \mathbb{C}.$$

Da jedes λ_i eine Nullstelle von p_A ist, gilt nach Lemma 3.5

$$\sigma(\mathbf{A}) = \{\lambda_i \ : \ i \in [1:n]\}.$$

Wegen

$$\mu_A^a(\lambda_i) = \#\{j \ : \ \lambda_j = \lambda_i\}$$

folgt

$$\sum_{\lambda\in\sigma(\mathbf{A})} \mu_A^a(\lambda) = \sum_{\lambda\in\sigma(\mathbf{A})} \#\{j \ : \ \lambda_j = \lambda\} = n.$$

■

Dieser Satz gilt in dieser Form nur für den komplexwertigen Fall, wie das folgende Beispiel illustriert:

Beispiel 3.9 *Sei $\mathbf{A} \in \mathbb{R}^{2\times 2}$ durch*

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

gegeben. Es gilt

$$p_A(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^2 + 1,$$

also besitzt p_A keine reellen Nullstellen.

Fasst man $\mathbf{A}$ als Matrix über $\mathbb{C}$ auf, so gilt $\sigma(\mathbf{A}) = \{i, -i\}$, die Matrix besitzt also zwei rein imaginäre Eigenwerte, zu denen etwa

$$\mathbf{x}_1 := \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} \quad \textit{und} \quad \mathbf{x}_2 := \frac{1}{\sqrt{2}} \begin{pmatrix} i \\ 1 \end{pmatrix}$$

Eigenvektoren sind, die sogar eine Orthonormalbasis von $\mathbb{C}^2$ bilden.

Im dritten Beispiel aus Kapitel 2, dem linearen Anfangswertproblem, stellte sich die Frage nach der Existenz einer Basis aus Eigenvektoren einer gewissen Matrix. Da ein Eigenvektor nicht zu zwei Eigenwerten gehören kann, sind die Eigenräume zu unterschiedlichen Eigenwerten disjunkt (bis auf den in allen Eigenräumen enthaltenen Nullvektor), und es folgt, dass eine Basis aus Eigenvektoren genau dann existiert, wenn

$$\sum_{\lambda \in \sigma(A)} \mu_A^g(\lambda) = n$$

gilt. Das ist nicht immer der Fall, wie das folgende Beispiel demonstriert:

Beispiel 3.10 *Wir untersuchen die Matrix*

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Wegen

$$p_A(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - 1)^2$$

gilt $\sigma(\mathbf{A}) = \{1\}$. Um die geometrische Vielfachheit $\mu_A^g(1)$ von $\lambda = 1$ zu bestimmen, müssen wir die Dimension des Kerns von

$$\lambda \mathbf{I} - \mathbf{A} = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}$$

berechnen. Da das Bild dieser Matrix eindimensional ist, muss es ihr Kern ebenfalls sein. Demzufolge erhalten wir mit

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

zwar einen Eigenvektor, können aber keinen zweiten von ihm linear unabhängigen finden. Es gilt also $\mu_A^a(\lambda) = 2 > 1 = \mu_A^g(\lambda)$.

3.2 Ähnlichkeitstransformationen

Bei einer Diagonal- oder Dreiecksmatrix lassen sich die Eigenwerte und ihre algebraischen Vielfachheiten direkt an den Diagonalelementen ablesen, also wäre es nützlich, wenn wir allgemeine Matrizen auf diese spezielle Form bringen könnten.

Es stellt sich die Frage, wie eine Transformation aussehen muss, die mindestens die Eigenwerte unverändert lässt. Für die Klärung dieser Frage untersuchen wir eine Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ mit einem Eigenwert $\lambda \in \mathbb{K}$ und einem passenden Eigenvektor $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Nach Definition gilt

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

Für eine beliebige Matrix $\mathbf{B} \in \mathbb{K}^{n\times n}$ folgt daraus

$$\mathbf{B}\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}.$$

Das ist leider keine Eigenwert-Gleichung mehr. Falls $\mathbf{B}$ regulär ist, können wir allerdings einen neuen Vektor

$$\widehat{\mathbf{x}} := \mathbf{B}\mathbf{x}, \qquad \mathbf{x} = \mathbf{B}^{-1}\widehat{\mathbf{x}} \tag{3.2}$$

definieren und erhalten

$$\mathbf{B}\mathbf{A}\mathbf{B}^{-1}\widehat{\mathbf{x}} = \lambda\widehat{\mathbf{x}},$$

also wieder eine Eigenwert-Gleichung für die neue Matrix

$$\widehat{\mathbf{A}} := \mathbf{B}\mathbf{A}\mathbf{B}^{-1}.$$

Der Wechsel von $\mathbf{A}$ zu $\widehat{\mathbf{A}}$ lässt also Eigenwerte unverändert, während sich mit den Gleichungen (3.2) Eigenvektoren ineinander überführen lassen.

Definition 3.11 (Ähnliche Matrizen) *Seien $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{K}^{n\times n}$. Die Matrizen $\mathbf{A}$ und $\widehat{\mathbf{A}}$ heißen* ähnlich*, wenn eine reguläre Matrix $\mathbf{B} \in \mathbb{K}^{n\times n}$ mit*

$$\widehat{\mathbf{A}} = \mathbf{B}\mathbf{A}\mathbf{B}^{-1} \tag{3.3}$$

existiert. Den Übergang von $\mathbf{A}$ zu $\widehat{\mathbf{A}}$ bezeichnen wir als Ähnlichkeitstransformation*.*

Wir haben bereits gesehen, dass bei einer Ähnlichkeitstransformation Eigenwerte unverändert bleiben. Mit Hilfe der Determinanten-Multiplikationssatzes lässt sich sogar beweisen, dass das charakteristische Polynom ebenfalls unverändert bleibt, also insbesondere auch die algebraischen Vielfachheiten:

Lemma 3.12 (Eigenwerte ähnlicher Matrizen) *Seien $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{K}^{n\times n}$ ähnliche Matrizen. Dann gilt $p_A = p_{\widehat{A}}$, also folgt insbesondere $\sigma(\mathbf{A}) = \sigma(\widehat{\mathbf{A}})$ und für alle $\lambda \in \sigma(\mathbf{A})$ gilt $\mu_A^a(\lambda) = \mu_{\widehat{A}}^a(\lambda)$.*

Beweis. Sei $\mathbf{B} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix mit $\widehat{\mathbf{A}} = \mathbf{B}\mathbf{A}\mathbf{B}^{-1}$. Aus dem Determinanten-Multiplikationssatz folgt

$$\begin{aligned} p_A(\lambda) &= \det(\lambda\mathbf{I} - \mathbf{A}) = \det(\mathbf{I})\det(\lambda\mathbf{I} - \mathbf{A}) \\ &= \det(\mathbf{B}\mathbf{B}^{-1})\det(\lambda\mathbf{I} - \mathbf{A}) = \det(\mathbf{B})\det(\lambda\mathbf{I} - \mathbf{A})\det(\mathbf{B}^{-1}) \\ &= \det(\mathbf{B}(\lambda\mathbf{I} - \mathbf{A})\mathbf{B}^{-1}) = \det(\lambda\mathbf{B}\mathbf{B}^{-1} - \mathbf{B}\mathbf{A}\mathbf{B}^{-1}) \end{aligned}$$

$$= \det(\lambda \mathbf{I} - \widehat{\mathbf{A}}) = p_{\widehat{A}}(\lambda)$$

für alle $\lambda \in \mathbb{K}$. ∎

Eigenvektoren bleiben unter Ähnlichkeitstransformationen nicht unverändert, allerdings lässt sich explizit angeben, wie sie sich ändern. Insbesondere folgt, dass auch die geometrischen Vielfachheiten unverändert bleiben.

Lemma 3.13 (Eigenvektoren) *Seien $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{K}^{n\times n}$ ähnliche Matrizen und sei $\mathbf{B} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix mit $\widehat{\mathbf{A}} = \mathbf{B}\mathbf{A}\mathbf{B}^{-1}$. Für jeden Eigenwert $\lambda \in \sigma(\mathbf{A}) = \sigma(\widehat{\mathbf{A}})$ gilt die Gleichung*

$$\mathcal{E}_{\widehat{A}}(\lambda) = \mathbf{B}\mathcal{E}_A(\lambda).$$

Insbesondere folgt $\mu_A^g(\lambda) = \mu_{\widehat{A}}^g(\lambda)$.

Beweis. Sei $\lambda \in \sigma(\mathbf{A}) = \sigma(\widehat{\mathbf{A}})$.
„$\supseteq$“: Sei $x \in \mathcal{E}_A(\lambda)$. Mit $\widehat{\mathbf{x}} := \mathbf{B}\mathbf{x}$ gilt

$$\widehat{\mathbf{A}}\widehat{\mathbf{x}} = \mathbf{B}\mathbf{A}\mathbf{B}^{-1}\mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{x} = \mathbf{B}\lambda\mathbf{x} = \lambda\widehat{\mathbf{x}},$$

also ist $\widehat{\mathbf{x}} = \mathbf{B}\mathbf{x}$ ein Eigenvektor der Matrix $\widehat{\mathbf{A}}$ zu dem Eigenwert λ und somit ein Element des Eigenraums $\mathcal{E}_{\widehat{A}}(\lambda)$.
„$\subseteq$“: Sei $\widehat{\mathbf{x}} \in \mathcal{E}_{\widehat{A}}(\lambda)$. Mit $\mathbf{x} := \mathbf{B}^{-1}\widehat{\mathbf{x}}$ gilt

$$\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{B}^{-1}\widehat{\mathbf{x}} = \mathbf{B}^{-1}\mathbf{B}\mathbf{A}\mathbf{B}^{-1}\widehat{\mathbf{x}} = \mathbf{B}^{-1}\widehat{\mathbf{A}}\widehat{\mathbf{x}} = \mathbf{B}^{-1}\lambda\widehat{\mathbf{x}} = \lambda\mathbf{x},$$

also ist $\mathbf{x} = \mathbf{B}^{-1}\widehat{\mathbf{x}}$ ein Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert λ und somit ein Element des Eigenraums $\mathcal{E}_A(\lambda)$. ∎

Von besonderem Interesse sind Ähnlichkeitstransformationen, die eine Matrix auf Diagonalgestalt bringen, denn bei eine Diagonalmatrix lassen sich nicht nur die Eigenwerte unmittelbar ablesen, sondern wir können die durch

$$\delta_i^{(j)} := \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in [1:n]$$

definierten *kanonischen Einheitsvektoren* $\delta^{(1)}, \ldots, \delta^{(n)} \in \mathbb{K}^n$ als Eigenvektoren verwenden, die aufgrund ihrer einfachen Gestalt häufig große Vorzüge bieten.

Beispielsweise können wir kanonische Einheitsvektoren in Kombination mit einer geeignet gewählten Ähnlichkeitstransformation verwenden, um eine Beziehung zwischen der geometrischen und der algebraischen Vielfachheit eines Eigenwert zu gewinnen.

Lemma 3.14 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$, sei $\lambda \in \sigma(\mathbf{A})$. Dann gilt*

$$1 \leq \mu_A^g(\lambda) \leq \mu_A^a(\lambda).$$

Beweis. Die erste Ungleichung folgt direkt aus Lemma 3.3. Zum Nachweis der zweiten konstruieren wir eine geeignete Ähnlichkeitstransformation, indem wir $p := \mu_A^g(\lambda)$ linear unabhängige Eigenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_p \in \mathbb{K}^n$ von $\mathbf{A}$ zu dem Eigenwert λ wählen und sie zu einer Basis $(\mathbf{x}_i)_{i=1}^n$ ergänzen. Wir bezeichnen mit $\mathbf{X} \in \mathbb{K}^{n\times n}$ die reguläre Matrix, deren Spalten die Vektoren $(\mathbf{x}_i)_{i=1}^n$ sind, die also gerade

$$\mathbf{X}\delta^{(i)} = \mathbf{x}_i \qquad \text{für alle } i \in [1:n]$$

erfüllt. Daraus folgt

$$\mathbf{A}\mathbf{X}\delta^{(i)} = \mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i = \lambda_i \mathbf{X}\delta^{(i)} \qquad \text{für alle } i \in [1:p],$$

so dass wir insbesondere

$$\mathbf{X}^{-1}\mathbf{A}\mathbf{X}\delta^{(i)} = \lambda_i \delta^{(i)} \qquad \text{für alle } i \in [1:p]$$

erhalten. Also verschwinden in den ersten p Spalten der Matrix $\widehat{\mathbf{A}} := \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ jeweils die unteren $n-p$ Zeilen, so dass die Matrix die Form

$$\widehat{\mathbf{A}} = \begin{pmatrix} \lambda\mathbf{I} & \mathbf{R} \\ 0 & \mathbf{C} \end{pmatrix}$$

für Matrizen $\mathbf{R} \in \mathbb{K}^{p\times(n-p)}$ und $\mathbf{C} \in \mathbb{K}^{(n-p)\times(n-p)}$ aufweist. Da $\widehat{\mathbf{A}}$ und $\mathbf{A}$ ähnlich sind, gilt nach Lemma 3.12 $p_A = p_{\widehat{A}}$, also folgt für alle $\alpha \in \mathbb{K}$ die Gleichung

$$\begin{aligned} p_A(\alpha) = p_{\widehat{A}}(\alpha) &= \det(\alpha\mathbf{I} - \widehat{\mathbf{A}}) \\ &= \det(\alpha\mathbf{I} - \lambda\mathbf{I}) \det(\alpha\mathbf{I} - \mathbf{C}) = (\alpha - \lambda)^p p_C(\alpha), \end{aligned}$$

und damit ist λ eine mindestens p-fache Nullstelle von p_A. ∎

3.3 Hilberträume

Neben den bisher eingeführten algebraischen Techniken haben sich auch Konzepte der Analysis als sehr nützlich für die Untersuchung von Eigenwertproblemen erwiesen. Für uns sind dabei vor allem bestimmte Eigenschaften von Bedeutung, die in *Hilberträumen* gelten, also in Banach-Räumen, deren Norm von einem Skalarprodukt induziert ist.

Auf dem Raum $\mathbb{K}^n$ verwendet man typischerweise das *euklidische Skalarprodukt*, das durch

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^n \bar{x}_i y_i \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n$$

definiert ist. Es ist eine *Sesquilinearform*, erfüllt also die Gleichungen

$$\langle \mathbf{x} + \alpha\mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \bar{\alpha}\langle \mathbf{y}, \mathbf{z} \rangle, \tag{3.4a}$$

$$\langle \mathbf{x}, \mathbf{y} + \alpha\mathbf{z}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle + \alpha\langle \mathbf{x}, \mathbf{z}\rangle, \tag{3.4b}$$

$$\langle \mathbf{x}, \mathbf{y}\rangle = \overline{\langle \mathbf{y}, \mathbf{x}\rangle} \qquad \text{für alle } \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{K}^n,\ \alpha \in \mathbb{K}. \tag{3.4c}$$

Das euklidische Skalarprodukt induziert die *euklidische Norm*

$$\|\mathbf{x}\| := \langle \mathbf{x}, \mathbf{x}\rangle^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n,$$

und aus dieser Beziehung folgt die *Cauchy-Schwarz-Ungleichung*

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|\ \|\mathbf{y}\| \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n. \tag{3.5}$$

Beide Seiten dieser Ungleichung sind genau dann gleich, wenn die Vektoren $\mathbf{x}$ und $\mathbf{y}$ linear abhängig sind.

Eine nützliche Beziehung zwischen der Matrix-Vektor-Multiplikation und dem Skalarprodukt können wir mit Hilfe der *adjungierten Matrix* gewinnen:

Definition 3.15 (Adjungierte Matrix) *Sei $\mathbf{A} \in \mathbb{K}^{n\times m}$. Die durch*

$$b_{ij} = \bar{a}_{ji} \qquad \textit{für alle } i \in [1:m],\ j \in [1:n]$$

definierte Matrix $\mathbf{B} \in \mathbb{K}^{m\times n}$ heißt Adjungierte *von $\mathbf{A}$ und wird mit $\mathbf{A}^* = \mathbf{B}$ bezeichnet. Im Falle $\mathbb{K} = \mathbb{R}$ entspricht sie der transponierten Matrix $\mathbf{A}^T$, im Falle $\mathbb{K} = \mathbb{C}$ der hermiteschen Matrix $\mathbf{A}^H$.*

Bei unseren Untersuchungen spielt die Beziehung zwischen dem Skalarprodukt und der Adjungierten eine wichtige Rolle.

Lemma 3.16 (Adjungierte und Skalarprodukt) *Seien eine Matrix $\mathbf{A} \in \mathbb{K}^{n\times m}$ sowie Vektoren $\mathbf{x} \in \mathbb{K}^m$ und $\mathbf{y} \in \mathbb{K}^n$ gegeben. Dann gilt*

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle.$$

Beweis. Nach Definition des Skalarprodukts und der Matrix-Vektor-Multiplikation gilt

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{n} \overline{\left(\sum_{j=1}^{m} a_{ij}x_j\right)} y_i = \sum_{i=1}^{n}\sum_{j=1}^{m} \bar{a}_{ij}\bar{x}_j y_i = \sum_{j=1}^{m} \bar{x}_j \left(\sum_{i=1}^{n} \bar{a}_{ij} y_i\right) = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle.$$

Das ist die gesuchte Identität. ∎

Diese Gleichung lässt sich vielfältig einsetzen. Als Beispiel beweisen wir die folgende Aussage über die Adjungierte eines Produkts zweier Matrizen:

Lemma 3.17 (Adjungierte eines Produkts) *Seien $\mathbf{A} \in \mathbb{K}^{n\times m}$ und $\mathbf{B} \in \mathbb{K}^{m\times k}$ mit $n, m, k \in \mathbb{N}$ gegeben. Dann gilt*

$$(\mathbf{A}\mathbf{B})^* = \mathbf{B}^*\mathbf{A}^*.$$

Falls $\mathbf{A} \in \mathbb{K}^{n\times n}$ invertierbar ist, gilt $(\mathbf{A}^*)^{-1} = (\mathbf{A}^{-1})^*$.

Beweis. Nach Lemma 3.16 gelten die Gleichungen

$$\langle \mathbf{x}, (\mathbf{AB})^* \mathbf{y} \rangle = \langle \mathbf{ABx}, \mathbf{y} \rangle = \langle \mathbf{Bx}, \mathbf{A}^* \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{B}^* \mathbf{A}^* \mathbf{y} \rangle \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^k,\ \mathbf{y} \in \mathbb{K}^n.$$

Daraus folgt

$$\langle \mathbf{x}, ((\mathbf{AB})^* - \mathbf{B}^* \mathbf{A}^*) \mathbf{y} \rangle = 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^k,\ \mathbf{y} \in \mathbb{K}^n,$$

und indem wir $\mathbf{x} := ((\mathbf{AB})^* - \mathbf{B}^* \mathbf{A}^*) \mathbf{y}$ einsetzen erhalten wir

$$\| ((\mathbf{AB})^* - \mathbf{B}^* \mathbf{A}^*) \mathbf{y} \|^2 = 0 \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^n,$$

so dass sich unmittelbar die erste Gleichung ergibt.

Sei nun $\mathbf{A} \in \mathbb{K}^{n \times n}$ invertierbar. Dann gelten mit der ersten Gleichung

$$\mathbf{I} = \mathbf{I}^* = (\mathbf{A}^{-1} \mathbf{A})^* = \mathbf{A}^* (\mathbf{A}^{-1})^*,$$
$$\mathbf{I} = \mathbf{I}^* = (\mathbf{A} \mathbf{A}^{-1})^* = (\mathbf{A}^{-1})^* \mathbf{A}^*,$$

also ist $(\mathbf{A}^{-1})^*$ eine Rechts- und Linksinverse der Matrix $\mathbf{A}^*$, also deren Inverse. ∎

Bei der Analyse numerischer Näherungsverfahren stellt sich häufig die Frage danach, wie sich in den einzelnen Stufen des Algorithmus' eingeführte Approximationsfehler auf den Gesamtfehler auswirken. Deshalb müssen wir in der Lage sein, zu messen, wie sehr sich Matrizen unterscheiden. Zu diesem Zweck definieren wir die *Spektralnorm*:

Definition 3.18 (Spektralnorm) *Seien $n, m \in \mathbb{N}$. Die* Spektralnorm *auf $\mathbb{K}^{n \times m}$ ist durch*

$$\|\mathbf{A}\| := \max \left\{ \frac{\|\mathbf{Az}\|}{\|\mathbf{z}\|} \ : \ \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{n \times m}$$

definiert.

Einige wesentliche Eigenschaften der Spektralnorm fasst das folgende Lemma zusammen.

Lemma 3.19 (Spektralnorm) *Seien $n, m, k \in \mathbb{N}$. Die Spektralnorm ist* verträglich *mit der euklidischen Norm, es gilt nämlich*

$$\|\mathbf{Az}\| \leq \|\mathbf{A}\| \, \|\mathbf{z}\| \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{n \times m},\ \mathbf{z} \in \mathbb{K}^m. \tag{3.6a}$$

Die Spektralnorm ist auch submultiplikativ*, erfüllt also*

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\| \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{n \times m},\ \mathbf{B} \in \mathbb{K}^{m \times k}. \tag{3.6b}$$

Die Spektralnorm lässt sich alternativ durch

$$\|\mathbf{A}\| = \max \left\{ \frac{|\langle \mathbf{y}, \mathbf{Az} \rangle|}{\|\mathbf{y}\| \, \|\mathbf{z}\|} \ : \ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\},\ \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} \quad \text{für alle } \mathbf{A} \in \mathbb{K}^{n \times m} \tag{3.6c}$$

darstellen. Aus dieser Gleichung folgen die Beziehungen

$$\|\mathbf{A}\| = \|\mathbf{A}^*\| = \|\mathbf{A}^* \mathbf{A}\|^{1/2} = \|\mathbf{A} \mathbf{A}^*\|^{1/2} \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{n \times m}. \tag{3.6d}$$

Beweis. Seien $\mathbf{A} \in \mathbb{K}^{n\times m}$ und $\mathbf{z} \in \mathbb{K}^m$. Falls $\mathbf{z} = \mathbf{0}$ gilt, folgt $\|\mathbf{Az}\| = 0 = \|\mathbf{A}\|\|\mathbf{z}\|$ unmittelbar.

Ansonsten gilt nach Definition

$$\frac{\|\mathbf{Az}\|}{\|\mathbf{z}\|} \leq \|\mathbf{A}\|,$$

und Multiplikation mit $\|\mathbf{z}\|$ führt zu (3.6a).

Indem wir (3.6a) zweimal anwenden erhalten wir

$$\|\mathbf{ABz}\| \leq \|\mathbf{A}\|\,\|\mathbf{Bz}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|\,\|\mathbf{z}\| \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^k,$$

so dass sich aus Definition 3.18 unmittelbar die Ungleichung (3.6b) ergibt.

Sei $\mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\}$ gegeben. Mit der Cauchy-Schwarz-Ungleichung (3.5) erhalten wir

$$\frac{|\langle \mathbf{y}, \mathbf{Az}\rangle|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} \leq \frac{\|\mathbf{y}\|\,\|\mathbf{Az}\|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} = \frac{\|\mathbf{Az}\|}{\|\mathbf{z}\|} \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}. \tag{3.7}$$

Falls $\mathbf{y} := \mathbf{Az} \neq \mathbf{0}$ gilt, haben wir

$$\frac{|\langle \mathbf{y}, \mathbf{Az}\rangle|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} = \frac{|\langle \mathbf{Az}, \mathbf{Az}\rangle|}{\|\mathbf{Az}\|\,\|\mathbf{z}\|} = \frac{\|\mathbf{Az}\|^2}{\|\mathbf{Az}\|\,\|\mathbf{z}\|} = \frac{\|\mathbf{Az}\|}{\|\mathbf{z}\|}$$

und gemeinsam mit (3.7) ergibt sich

$$\frac{\|\mathbf{Az}\|}{\|\mathbf{z}\|} \leq \max\left\{ \frac{|\langle \mathbf{y}, \mathbf{Az}\rangle|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} \;:\; \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\} \leq \frac{\|\mathbf{Az}\|}{\|\mathbf{z}\|}.$$

Diese Abschätzung bleibt offenbar auch korrekt, falls $\mathbf{Az} = \mathbf{0}$ gilt. Indem wir zu dem Maximum über alle $\mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\}$ übergehen folgt (3.6c).

Aus dieser Gleichung folgt mit Lemma 3.16 und (3.4c) direkt

$$\begin{aligned}
\|\mathbf{A}\| &= \max\left\{ \frac{|\langle \mathbf{y}, \mathbf{Az}\rangle|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} \;:\; \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\},\ \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} \\
&= \max\left\{ \frac{|\langle \mathbf{A}^*\mathbf{y}, \mathbf{z}\rangle|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} \;:\; \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\},\ \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} \\
&= \max\left\{ \frac{|\langle \mathbf{z}, \mathbf{A}^*\mathbf{y}\rangle|}{\|\mathbf{z}\|\,\|\mathbf{y}\|} \;:\; \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\},\ \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} = \|\mathbf{A}^*\|.
\end{aligned}$$

Mit dieser Gleichung in Kombination mit Lemma 3.16, (3.6c) sowie (3.6b) erhalten wir

$$\begin{aligned}
\|\mathbf{A}\|^2 &= \max\left\{ \frac{\|\mathbf{Az}\|^2}{\|\mathbf{z}\|^2} \;:\; \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} \\
&= \max\left\{ \frac{|\langle \mathbf{Az}, \mathbf{Az}\rangle|}{\|\mathbf{z}\|^2} \;:\; \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\} \\
&= \max\left\{ \frac{|\langle \mathbf{z}, \mathbf{A}^*\mathbf{Az}\rangle|}{\|\mathbf{z}\|^2} \;:\; \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\}
\end{aligned}$$

$$\leq \max\left\{\frac{|\langle \mathbf{y}, \mathbf{A}^*\mathbf{A}\mathbf{z}\rangle|}{\|\mathbf{y}\|\,\|\mathbf{z}\|} \;:\; \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\},\ \mathbf{z} \in \mathbb{K}^m \setminus \{\mathbf{0}\}\right\}$$
$$= \|\mathbf{A}^*\mathbf{A}\| \leq \|\mathbf{A}^*\|\,\|\mathbf{A}\| = \|\mathbf{A}\|^2.$$

Indem wir dieses Resultat auf $\mathbf{A}^*$ anstelle von $\mathbf{A}$ anwenden folgt die letzte Gleichung. ∎

In einigen der folgenden Beweise benötigen wir eine Möglichkeit, rechteckige Matrizen zu „invertieren“. Diese Pseudo-Inverse können wir mit Hilfe der Adjungierten definieren. Ein erster Schritt in dieser Richtung ist das folgende Lemma, mit dem sich die Identität von Vektoren im Bild der Matrix überprüfen lässt.

Lemma 3.20 (Orthogonalität) *Sei $\mathbf{A} \in \mathbb{K}^{n\times m}$. Es gilt*

$$\langle \mathbf{x}, \mathbf{y}\rangle = 0 \qquad \text{für alle } \mathbf{x} \in \mathrm{Bild}(\mathbf{A}),\ \mathbf{y} \in \mathrm{Kern}(\mathbf{A}^*).$$

Insbesondere haben wir

$$\mathbf{A}\mathbf{z} = \mathbf{0} \iff \mathbf{A}^*\mathbf{A}\mathbf{z} = \mathbf{0} \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^m.$$

Beweis. Seien $\mathbf{x} \in \mathrm{Bild}(\mathbf{A})$ und $\mathbf{y} \in \mathrm{Kern}(\mathbf{A}^*)$ gegeben. Nach Definition finden wir $\mathbf{z} \in \mathbb{K}^m$ mit $\mathbf{x} = \mathbf{A}\mathbf{z}$. Mit Lemma 3.16 folgt

$$\langle \mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{A}\mathbf{z}, \mathbf{y}\rangle = \langle \mathbf{z}, \mathbf{A}^*\mathbf{y}\rangle = \langle \mathbf{z}, \mathbf{0}\rangle = 0.$$

Sei nun $\mathbf{z} \in \mathbb{K}^m$. Offenbar folgt aus $\mathbf{A}\mathbf{z} = \mathbf{0}$ unmittelbar auch $\mathbf{A}^*\mathbf{A}\mathbf{z} = \mathbf{A}^*\mathbf{0} = \mathbf{0}$.

Für den Nachweis der Umkehrung setzen wir voraus, dass $\mathbf{A}^*\mathbf{A}\mathbf{z} = \mathbf{0}$ gilt. Also liegt der Vektor $\mathbf{x} := \mathbf{A}\mathbf{z}$ sowohl im Bild der Matrix $\mathbf{A}$ als auch im Kern der Adjungierten $\mathbf{A}^*$. Mit dem ersten Teil unserer Aussage folgt

$$\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x}\rangle = 0,$$

also $\mathbf{A}\mathbf{z} = \mathbf{x} = \mathbf{0}$. ∎

Aus dieser Eigenschaft folgt bereits, dass zwei Vektoren aus dem Bild einer Matrix $\mathbf{A}$ genau dann identisch sind, falls ihre Produkte mit der Adjungierten $\mathbf{A}^*$ identisch sind. Um auch passende Urbilder rekonstruieren zu können, benötigen wir zusätzliche Eigenschaften der Matrix.

Definition 3.21 (Positiv definit) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Die Matrix $\mathbf{A}$ heißt* positiv definit*, falls*

$$\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle > 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \tag{3.8}$$

gilt. Sie heißt positiv semidefinit*, falls die Ungleichung*

$$\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle \geq 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \tag{3.9}$$

erfüllt ist.

Eine positiv definite Matrix ist immer invertierbar, denn aus (3.8) folgt unmittelbar, dass nur der Nullvektor im Kern der Matrix liegen kann, dass $\mathbf{A}$ also injektiv ist. Da $\mathbf{A}$ eine quadratische Matrix ist, muss sie nach dem Dimensionssatz für lineare Abbildungen auch invertierbar sein.

Wenn wir für einen Vektor $\mathbf{x}$ aus dem Bild einer Matrix $\mathbf{A}$ ein Urbild rekonstruieren wollen, also einen Vektor $\mathbf{y}$ mit $\mathbf{x} = \mathbf{A}\mathbf{y}$ suchen, können wir uns dem Problem nähern, indem wir beide Seiten der Gleichung mit der Adjungierten $\mathbf{A}^*$ multiplizieren und

$$\mathbf{A}^*\mathbf{x} = \mathbf{A}^*\mathbf{A}\mathbf{y}$$

erhalten. Falls die *Gramsche Matrix* $\mathbf{A}^*\mathbf{A}$ auf der rechten Seite invertierbar ist, können wir die Gleichung mit deren Inversen multiplizieren und eine explizite Formel erhalten, mit der sich $\mathbf{y}$ aus $\mathbf{x}$ berechnen lässt.

Lemma 3.22 (Gramsche Matrizen) *Sei $\mathbf{A} \in \mathbb{K}^{n\times m}$. Dann sind $\mathbf{A}^*\mathbf{A}$ und $\mathbf{A}\mathbf{A}^*$ positiv semidefinite Matrizen.*

$\mathbf{A}$ ist genau dann injektiv, wenn $\mathbf{A}^*\mathbf{A}$ positiv definit ist.

$\mathbf{A}$ ist genau dann surjektiv, wenn $\mathbf{A}\mathbf{A}^*$ positiv definit ist.

Beweis. Mit Lemma 3.16 erhalten wir

$$\langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}\rangle = \|\mathbf{A}\mathbf{x}\|^2 \geq 0, \tag{3.10a}$$

$$\langle \mathbf{A}\mathbf{A}^*\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}^*\mathbf{x}, \mathbf{A}^*\mathbf{x}\rangle = \|\mathbf{A}^*\mathbf{x}\|^2 \geq 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n, \tag{3.10b}$$

also sind $\mathbf{A}^*\mathbf{A}$ und $\mathbf{A}\mathbf{A}^*$ positiv semidefinit.

Falls $\mathbf{A}$ injektiv ist, gilt $\|\mathbf{A}\mathbf{x}\| > 0$ für alle $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$, also ist nach (3.10a) die Matrix $\mathbf{A}^*\mathbf{A}$ positiv definit.

Falls umgekehrt $\mathbf{A}^*\mathbf{A}$ positiv definit ist, gilt nach (3.10a) die Ungleichung $\|\mathbf{A}\mathbf{x}\| > 0$ für alle $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$, also kann der Kern der Matrix $\mathbf{A}$ nur den Nullvektor enthalten. Damit muss $\mathbf{A}$ injektiv sein.

Falls $\mathbf{A}$ surjektiv ist, falls wir also $\operatorname{Bild}(\mathbf{A}) = \mathbb{K}^n$ haben, folgt aus Lemma 3.20 bereits $\operatorname{Kern}(\mathbf{A}^*) = \{\mathbf{0}\}$, also gilt insbesondere $\|\mathbf{A}^*\mathbf{x}\| = 0$ genau dann, wenn $\mathbf{x} = \mathbf{0}$ gilt. Also muss nach (3.10b) die Matrix $\mathbf{A}\mathbf{A}^*$ positiv definit sein.

Falls umgekehrt $\mathbf{A}\mathbf{A}^*$ positiv definit, also insbesondere invertierbar ist, können wir zu jedem $\mathbf{x} \in \mathbb{K}^n$ den Vektor $\mathbf{y} := \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}\mathbf{x}$ definieren und erhalten

$$\mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}\mathbf{x} = \mathbf{x},$$

also folgt $\mathbf{x} \in \operatorname{Bild}(\mathbf{A})$ für beliebige $\mathbf{x} \in \mathbb{K}^n$, so dass $\mathbf{A}$ surjektiv sein muss. ∎

3.4 Invariante Unterräume

Wie bereits gesehen, lässt sich der Raum $\mathbb{K}^n$ im allgemeinen nicht in eine direkte Summe von Eigenräumen zerlegen. Es ist allerdings möglich, die Eigenräume durch allgemeinere Teilräume von $\mathbb{K}^n$ zu ersetzen, die dann eine direkte Summe bilden:

Definition 3.23 (Invariante Unterräume) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Ein Teilraum $\mathcal{V} \subseteq \mathbb{K}^n$ heißt* bezüglich $\mathbf{A}$ invarianter Unterraum*, falls für alle $\mathbf{x} \in \mathcal{V}$ die Gleichung*

$$\mathbf{A}\mathbf{x} \in \mathcal{V}$$

gilt, falls also $\mathbf{A}\mathcal{V} \subseteq \mathcal{V}$ erfüllt ist.

Beispiel 3.24 *Sei $p \in \mathbb{N}$ und $(\mathbf{x}_i)_{i=1}^p$ eine Familie von Eigenvektoren der Eigenwerte $(\lambda_i)_{i=1}^p$. Dann ist*

$$\mathcal{V} := \operatorname{span}\{\mathbf{x}_i \ : \ i \in [1:p]\}$$

ein bezüglich $\mathbf{A}$ invarianter Unterraum.

Beweis. Offensichtlich gilt $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \in \mathcal{V}$ für alle $i \in [1:p]$, also gilt dasselbe auch für den Aufspann dieser Vektoren. ∎

Beispiel 3.25 *Sei $\mathbf{R} \in \mathbb{K}^{n\times n}$ eine obere Dreiecksmatrix. Dann ist für jedes $p \in [1:n]$ der Raum*

$$\mathcal{V} := \operatorname{span}\{\delta^{(i)} \ : \ i \in [1:p]\} = \mathbb{K}^p \times \{\mathbf{0}\} \subseteq \mathbb{K}^n$$

invariant bezüglich $\mathbf{R}$.

Für Eigenräume gilt die Gleichung $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ mit dem Eigenwert λ. Für invariante Unterräume wird diese Gleichung etwas verallgemeinert, indem wir den Eigenwert λ durch eine kleine quadratische Matrix ersetzen und den Eigenvektor $\mathbf{x}$ durch eine aus mehreren Vektoren gebildete Matrix.

Lemma 3.26 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und sei $\mathcal{V} \subseteq \mathbb{K}^n$ ein bezüglich $\mathbf{A}$ invarianter Unterraum. Sei $\mathbf{X} \in \mathbb{K}^{n\times p}$ eine Matrix, deren Spalten den Raum $\mathcal{V}$ aufspannen, die also die Gleichung*

$$\operatorname{Bild}(\mathbf{X}) = \mathcal{V}$$

erfüllt. Dann gibt es eine Matrix $\mathbf{\Lambda} \in \mathbb{K}^{p\times p}$ mit

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}. \tag{3.11}$$

Falls $\mathbf{X}$ injektiv ist, ist $\mathbf{\Lambda}$ durch die obige Gleichung eindeutig festgelegt.

Beweis. Sei $j \in [1:p]$. Offenbar gilt $\mathbf{X}\delta^{(j)} \in \mathcal{V}$, und aus der Invarianz folgt $\mathbf{A}\mathbf{X}\delta^{(j)} \in \mathcal{V} = \operatorname{Bild}(\mathbf{X})$. Also existiert ein Vektor $\mathbf{z}^{(j)} \in \mathbb{K}^p$ mit

$$\mathbf{A}\mathbf{X}\delta^{(j)} = \mathbf{X}\mathbf{z}^{(j)}.$$

Wir definieren die Matrix

$$\mathbf{\Lambda} := \begin{pmatrix} \mathbf{z}^{(1)} & \cdots & \mathbf{z}^{(p)} \end{pmatrix},$$

so dass gerade

$$\mathbf{\Lambda}\delta^{(j)} = \mathbf{z}^{(j)} \qquad \text{für alle } j \in [1:p]$$

gilt. Nun folgt

$$\mathbf{A}\mathbf{X}\delta^{(j)} = \mathbf{X}\mathbf{z}^{(j)} = \mathbf{X}\mathbf{\Lambda}\delta^{(j)} \qquad \text{für alle } j \in [1:p],$$

so dass unmittelbar (3.11) folgt.

Sei nun $\mathbf{X}$ injektiv, und sei $\mathbf{\Lambda} \in \mathbb{K}^{p\times p}$ eine Matrix, die (3.11) erfüllt. Nach Lemma 3.22 ist $\mathbf{X}^*\mathbf{X}$ invertierbar, so dass aus

$$\begin{aligned}\mathbf{A}\mathbf{X} &= \mathbf{X}\mathbf{\Lambda},\\ \mathbf{X}^*\mathbf{A}\mathbf{X} &= \mathbf{X}^*\mathbf{X}\mathbf{\Lambda}\end{aligned}$$

bereits

$$\mathbf{\Lambda} = (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*\mathbf{A}\mathbf{X}$$

folgt. Durch diese Gleichung ist $\mathbf{\Lambda}$ eindeutig festgelegt. ∎

Bemerkung 3.27 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Falls Matrizen $\mathbf{X} \in \mathbb{K}^{n\times p}$ und $\mathbf{\Lambda} \in \mathbb{K}^{p\times p}$ so gegeben sind, dass (3.11) gilt, muss $\mathcal{V} := \mathrm{Bild}(\mathbf{X})$ bereits ein invarianter Teilraum sein: Für jedes $\mathbf{x} \in \mathrm{Bild}(\mathbf{X})$ existiert ein Urbild $\mathbf{y} \in \mathbb{K}^p$ mit $\mathbf{x} = \mathbf{X}\mathbf{y}$, so dass*

$$\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{X}\mathbf{y} = \mathbf{X}\mathbf{\Lambda}\mathbf{y} \in \mathrm{Bild}(\mathbf{X})$$

unmittelbar folgt.

Bemerkung 3.28 *Seien $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{K}^{n\times n}$ ähnliche Matrizen mit $\mathbf{B}\mathbf{A}\mathbf{B}^{-1} = \widehat{\mathbf{A}}$ für eine reguläre Matrix $\mathbf{B}$. Dann ist für jeden bezüglich $\mathbf{A}$ invarianten Unterraum $\mathcal{V} \subseteq \mathbb{K}^n$ die Menge*

$$\widehat{\mathcal{V}} := \mathbf{B}^{-1}\mathcal{V}$$

ein bezüglich $\mathbf{B}$ invarianter Unterraum.

Bemerkung 3.29 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$, und seien Matrizen $\mathbf{X} \in \mathbb{K}^{n\times p}$ und $\mathbf{\Lambda} \in \mathbb{K}^{p\times p}$ so gegeben sind, dass (3.11) gilt. Falls $\mathbf{X}$ injektiv ist, ist jeder Eigenwert der Matrix $\mathbf{\Lambda}$ auch ein Eigenwert der Matrix $\mathbf{A}$: Für jedes $\lambda \in \sigma(\mathbf{\Lambda})$ existiert ein Eigenvektor $\mathbf{y} \in \mathbb{K}^p \setminus \{\mathbf{0}\}$. Da $\mathbf{X}$ injektiv ist, ist $\mathbf{x} := \mathbf{X}\mathbf{y}$ nicht der Nullvektor. Es folgt*

$$\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{X}\mathbf{y} = \mathbf{X}\mathbf{\Lambda}\mathbf{y} = \mathbf{X}\lambda\mathbf{y} = \lambda\mathbf{X}\mathbf{y} = \lambda\mathbf{x}.$$

3.5 Selbstadjungierte und unitäre Matrizen

Einen invarianten Unterraum können wir durch eine beliebige Basis darstellen. Aus der Perspektive der Numerik ist es ratsam, eine *Orthonormalbasis* zu verwenden, denn derartige Basen zeichnen sich durch eine besondere Unempfindlichkeit gegenüber Rundungsfehlern aus.

Definition 3.30 (Isometrisch und unitär) *Sei* $\mathbf{Q} \in \mathbb{K}^{n\times m}$*. Falls die Gleichung*

$$\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$$

gilt, nennen wir $\mathbf{Q}$ isometrisch*.*

Falls $\mathbf{Q}$ *isometrisch und quadratisch ist, falls also auch noch* $n = m$ *gilt, nennen wir* $\mathbf{Q}$ unitär*.*

Um die Bezeichnung „isometrische Matrix“ zu rechtfertigen müssen wir auf eine auch noch für andere Zwecke nützliche Klasse von Matrizen zurückgreifen, nämlich auf die *selbstadjungierten Matrizen.*

Definition 3.31 (Selbstadjungierte Matrix) *Eine Matrix* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *heißt* selbstadjungiert*, falls* $\mathbf{A} = \mathbf{A}^*$ *gilt.*

Neben vielen anderen vorteilhaften Eigenschaften bieten selbstadjungierte Matrizen den Vorteil, dass sich viele wichtige Eigenschaften bereits an dem Skalarprodukt $\langle \mathbf{Ax}, \mathbf{x}\rangle$ ablesen lassen. Für unseren Zweck genügt zunächst die folgende Identitätsaussage:

Lemma 3.32 (Identität selbstadjungierter Matrizen) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine selbstadjungierte Matrix. Falls*

$$\langle \mathbf{Ax}, \mathbf{x}\rangle = 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \tag{3.12}$$

gilt, folgt bereits $\mathbf{A} = \mathbf{0}$*.*

Beweis. Wir nehmen an, dass (3.12) gilt. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$. Nach Voraussetzung folgt mit Lemma 3.16 die Gleichung

$$\begin{aligned} 0 &= \langle \mathbf{A}(\mathbf{x}+\mathbf{y}), \mathbf{x}+\mathbf{y}\rangle = \langle \mathbf{Ax}, \mathbf{x}\rangle + \langle \mathbf{Ay}, \mathbf{x}\rangle + \langle \mathbf{Ax}, \mathbf{y}\rangle + \langle \mathbf{Ay}, \mathbf{y}\rangle \\ &= \langle \mathbf{Ay}, \mathbf{x}\rangle + \langle \mathbf{Ax}, \mathbf{y}\rangle = \langle \mathbf{y}, \mathbf{A}^*\mathbf{x}\rangle + \langle \mathbf{Ax}, \mathbf{y}\rangle = \langle \mathbf{y}, \mathbf{Ax}\rangle + \langle \mathbf{Ax}, \mathbf{y}\rangle. \end{aligned}$$

Nun können wir $\mathbf{y} = \mathbf{Ax}$ setzen und erhalten $0 = 2\|\mathbf{Ax}\|^2$, also $\mathbf{Ax} = \mathbf{0}$ für jeden beliebigen Vektor $\mathbf{x} \in \mathbb{K}^n$, und damit insbesondere $\mathbf{A} = \mathbf{0}$. ∎

Mit diesem Hilfsmittel können wir nun eine alternative Charakterisierung isometrischer Matrizen angeben:

Lemma 3.33 (Isometrische Matrix) *Sei* $\mathbf{Q} \in \mathbb{K}^{n\times m}$*.* $\mathbf{Q}$ *ist genau dann isometrisch, wenn*

$$\|\mathbf{Qx}\| = \|\mathbf{x}\| \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \tag{3.13}$$

gilt, wenn die Multiplikation mit $\mathbf{Q}$ *also die Norm unverändert lässt.*

Beweis. „⇒“: Sei zunächst $\mathbf{Q}$ isometrisch. Nach Lemma 3.16 folgt

$$\|\mathbf{Qx}\|^2 = \langle \mathbf{Qx}, \mathbf{Qx}\rangle = \langle \mathbf{Q}^*\mathbf{Qx}, \mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle = \|\mathbf{x}\|^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n,$$

also gerade (3.13).

„⇐“: Gelte nun umgekehrt (3.13). Es folgt

$$\begin{aligned} 0 &= \|\mathbf{Qx}\|^2 - \|\mathbf{x}\|^2 = \langle \mathbf{Qx}, \mathbf{Qx}\rangle - \langle \mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{Q}^*\mathbf{Qx}, \mathbf{x}\rangle - \langle \mathbf{x}, \mathbf{x}\rangle \\ &= \langle (\mathbf{Q}^*\mathbf{Q} - \mathbf{I})\mathbf{x}, \mathbf{x}\rangle \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n. \end{aligned}$$

Da die Matrix $\mathbf{Q}^*\mathbf{Q} - \mathbf{I}$ selbstadjungiert ist, erhalten wir mit Lemma 3.32 bereits die Gleichung $\mathbf{Q}^*\mathbf{Q} - \mathbf{I} = \mathbf{0}$, also muss $\mathbf{Q}$ isometrisch sein. ∎

In Hinblick auf die für die Behandlung von Eigenwertproblemen sehr wichtigen Ähnlichkeitstransformationen bieten unitäre Matrizen den Vorteil, dass sich ihre Inversen besonders einfach berechnen lassen.

Lemma 3.34 (Unitäre Matrix) *Sei $\mathbf{Q} \in \mathbb{K}^{n\times n}$ unitär.*
Dann gilt $\mathbf{QQ}^* = \mathbf{I}$, also ist die Adjungierte $\mathbf{Q}^*$ die Inverse der Matrix $\mathbf{Q}$.

Beweis. Da die Matrix $\mathbf{Q}$ nach (3.13) insbesondere injektiv ist, folgt mit der Dimensionsformel, dass sie als quadratische Matrix auch surjektiv sein muss, es gilt also $\text{Bild}(\mathbf{Q}) = \mathbb{K}^n$.

Sei $\mathbf{x} \in \mathbb{K}^n$. Dank der Surjektivität gilt $\mathbf{x} \in \text{Bild}(\mathbf{Q})$, also finden wir ein Urbild $\mathbf{y} \in \mathbb{K}^n$ mit $\mathbf{x} = \mathbf{Qy}$. Mit Definition 3.30 erhalten wir

$$\mathbf{x} = \mathbf{Qy} = \mathbf{Q}(\mathbf{Q}^*\mathbf{Q})\mathbf{y} = \mathbf{QQ}^*\mathbf{Qy} = \mathbf{QQ}^*\mathbf{x}.$$

Da $\mathbf{x}$ beliebig gewählt wurde, folgt daraus bereits $\mathbf{I} = \mathbf{QQ}^*$. Da wegen Definition 3.30 auch $\mathbf{I} = \mathbf{Q}^*\mathbf{Q}$ gilt, muss $\mathbf{Q}^*$ die Inverse der Matrix $\mathbf{Q}$ sein. ∎

3.6 Schur-Zerlegung

Nun können wir daran gehen, nach einer Möglichkeit zu suchen, um eine Matrix durch *unitäre* Ähnlichkeitstransformationen auf obere Dreiecksgestalt zu bringen.

Wir werden dabei induktiv vorgehen und zunächst versuchen, die erste Spalte der Matrix auf ein Vielfaches des ersten kanonischen Einheitsvektors abzubilden. Für derartige Aufgaben bietet die numerische lineare Algebra ein nützliches Hilfsmittel: *Householder-Spiegelungen* sind unitäre Abbildungen, die einen beliebigen Vektor in den Aufspann eines beliebigen anderen (von null verschiedenen) Vektors, beispielsweise eines kanonischen Einheitsvektors, überführen.

Erinnerung 3.35 (Householder-Spiegelung) *Zu jedem Vektor $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ist die* Householder-Spiegelung

$$\mathbf{Q}_v := \mathbf{I} - 2\frac{\mathbf{vv}^*}{\mathbf{v}^*\mathbf{v}}$$

eine unitäre Matrix.

Für beliebige Vektoren $\mathbf{x} \in \mathbb{K}^n$ *und* $\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ *können wir einen* Householder-Vektor $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ *so finden, dass*

$$\mathbf{Q}_v \mathbf{x} = \alpha \mathbf{y}$$

mit einem $\alpha \in \mathbb{K}$ *gilt, dass also* $\mathbf{x}$ *in den von* $\mathbf{y}$ *aufgespannten Raum abgebildet wird.*

Erinnerung 3.36 (QR-Zerlegung) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times m}$*. Dann existieren eine unitäre Matrix* $\mathbf{Q} \in \mathbb{K}^{n\times n}$ *und eine obere Dreiecksmatrix* $\mathbf{R} \in \mathbb{K}^{n\times m}$ *mit* $\mathbf{A} = \mathbf{QR}$*.*

Mit Hilfe der QR-Zerlegung können wir eine isometrische Matrix zu einer unitären Matrix ergänzen.

Lemma 3.37 (Basisergänzung) *Sei* $\widehat{\mathbf{Q}} \in \mathbb{K}^{n\times p}$ *eine isometrische Matrix. Dann existiert eine unitäre Matrix* $\mathbf{Q} \in \mathbb{K}^{n\times n}$ *mit* $\mathbf{Q}|_{n\times p} = \widehat{\mathbf{Q}}$*.*

Beweis. Nach Erinnerung 3.36 existiert eine QR-Zerlegung $\mathbf{Q}_0\mathbf{R}_0 = \widehat{\mathbf{Q}}$ der Matrix $\widehat{\mathbf{Q}}$.

Nach Lemma 3.33 ist $\widehat{\mathbf{Q}}$ injektiv, also muss $p \leq n$ gelten. Da $\mathbf{R}_0$ eine obere Dreiecksmatrix mit mehr Zeilen als Spalten ist, existiert eine Matrix $\mathbf{R} \in \mathbb{K}^{p\times p}$ mit

$$\mathbf{R}_0 = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}.$$

Es gilt

$$\mathbf{R}^*\mathbf{R} = \mathbf{R}_0^*\mathbf{R}_0 = \mathbf{R}_0^*\mathbf{Q}_0^*\mathbf{Q}_0\mathbf{R}_0 = \widehat{\mathbf{Q}}^*\widehat{\mathbf{Q}} = \mathbf{I},$$

also ist $\mathbf{R}$ eine unitäre Dreiecksmatrix. Wir definieren

$$\mathbf{Q} := \mathbf{Q}_0 \begin{pmatrix} \mathbf{R} & \\ & \mathbf{I} \end{pmatrix}$$

und stellen fest, dass

$$\mathbf{Q}|_{n\times p} = \mathbf{Q}_0 \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}_0\mathbf{R}_0 = \widehat{\mathbf{Q}}$$

gilt. Also haben wir die gewünschte Fortsetzung gefunden. ∎

Tatsächlich können wir sogar beweisen, dass sich die ersten p Spalten der Matrix $\mathbf{Q}_0$ lediglich durch ihre Vorzeichen von den ersten Spalten der Matrix $\widehat{\mathbf{Q}}$ unterscheiden: Man kann sich überlegen, dass jede unitäre Dreiecksmatrix bereits eine Diagonalmatrix sein muss, und dieses Resultat lässt sich auf die Matrix $\mathbf{R}$ anwenden.

Lemma 3.38 (Deflation) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$*, sei* $p \in \mathbb{N}$ *und sei* $\mathbf{X} \in \mathbb{K}^{n\times p}$ *eine isometrische Matrix, die (3.11) mit einer geeigneten Matrix* $\mathbf{\Lambda} \in \mathbb{K}^{p\times p}$ *erfüllt. Dann existieren eine unitäre Matrix* $\mathbf{Q} \in \mathbb{K}^{n\times n}$ *und weitere Matrizen* $\mathbf{D} \in \mathbb{K}^{(n-p)\times(n-p)}, \mathbf{R} \in \mathbb{K}^{p\times(n-p)}$ *derart, dass die Gleichung*

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{\Lambda} & \mathbf{R} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \tag{3.14}$$

erfüllt ist und $\mathbf{Q}|_{n\times p} = \mathbf{X}$ *gilt.*

Beweis. Gemäß Lemma 3.37 können wir $\mathbf{X}$ zu einer unitären Matrix $\mathbf{Q}$ mit $\mathbf{Q}|_{n\times p} = \mathbf{X}$ ergänzen.

Wir zerlegen $\mathbf{Q}$ dementsprechend in

$$\mathbf{Q} = \begin{pmatrix}\mathbf{X} & \mathbf{Y}\end{pmatrix}$$

und stellen fest, dass auch $\mathbf{Y} \in \mathbb{K}^{n\times(n-p)}$ isometrisch ist und aus

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} = \mathbf{I} = \mathbf{Q}^*\mathbf{Q} = \begin{pmatrix}\mathbf{X}^*\\ \mathbf{Y}^*\end{pmatrix}\begin{pmatrix}\mathbf{X} & \mathbf{Y}\end{pmatrix} = \begin{pmatrix}\mathbf{X}^*\mathbf{X} & \mathbf{X}^*\mathbf{Y}\\ \mathbf{Y}^*\mathbf{X} & \mathbf{Y}^*\mathbf{Y}\end{pmatrix}$$

die Identität $\mathbf{Y}^*\mathbf{X} = \mathbf{0}$ folgt.

Aus der Gleichung (3.11) erhalten wir

$$\begin{aligned}\mathbf{Q}^*\mathbf{A}\mathbf{Q} &= \begin{pmatrix}\mathbf{X}^*\\ \mathbf{Y}^*\end{pmatrix}\mathbf{A}\begin{pmatrix}\mathbf{X} & \mathbf{Y}\end{pmatrix} = \begin{pmatrix}\mathbf{X}^*\mathbf{A}\mathbf{X} & \mathbf{X}^*\mathbf{A}\mathbf{Y}\\ \mathbf{Y}^*\mathbf{A}\mathbf{X} & \mathbf{Y}^*\mathbf{A}\mathbf{Y}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{X}^*\mathbf{X}\mathbf{\Lambda} & \mathbf{X}^*\mathbf{A}\mathbf{Y}\\ \mathbf{Y}^*\mathbf{X}\mathbf{\Lambda} & \mathbf{Y}^*\mathbf{A}\mathbf{Y}\end{pmatrix} = \begin{pmatrix}\mathbf{\Lambda} & \mathbf{X}^*\mathbf{A}\mathbf{Y}\\ \mathbf{0} & \mathbf{Y}^*\mathbf{A}\mathbf{Y}\end{pmatrix},\end{aligned}$$

also können wir den Beweis abschließen, indem wir

$$\mathbf{R} := \mathbf{X}^*\mathbf{A}\mathbf{Y}, \qquad\qquad \mathbf{D} := \mathbf{Y}^*\mathbf{A}\mathbf{Y}$$

setzen. ∎

Da nach Satz 3.8 eine komplexe Matrix mindestens einen Eigenwert besitzt, ist der zugehörige Eigenraum nach Beispiel 3.24 ein invarianter Unterraum, der mittels Lemma 3.38 Anlass zu einer vereinfachenden Ähnlichkeitstransformation gibt. Bei wiederholter Anwendung erhält man eine sogenannte *Schur-Zerlegung* der Matrix:

Satz 3.39 (Schur-Zerlegung) *Sei $\mathbf{A} \in \mathbb{C}^{n\times n}$. Dann existieren eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine obere Dreiecksmatrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ so, dass die Gleichung*

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R} \tag{3.15}$$

erfüllt ist.

Beweis. Per Induktion über die Dimension n. Der Induktionsanfang $n = 1$ ist trivial.

Wir nehmen an, dass $n > 1$ gilt und die Aussage für $n-1$ bewiesen ist. Nach Satz 3.8 besitzt $\mathbf{A}$ mindestens einen Eigenwert $\lambda \in \mathbb{C}$. Sei $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ ein zugehöriger Eigenvektor, den wir auf $\|\mathbf{x}\| = 1$ normieren. Indem wir $\mathbf{x}$ als Matrix $\mathbf{X} \in \mathbb{C}^{n\times 1}$ und λ als Matrix $\mathbf{\Lambda} \in \mathbb{C}^{1\times 1}$ interpretieren, können wir Lemma 3.38 anwenden, um eine unitäre Matrix $\mathbf{Q}_1 \in \mathbb{C}^{n\times n}$ und Matrizen $\mathbf{R}_1 \in \mathbb{C}^{1\times n}$ und $\mathbf{A}_1 \in \mathbb{C}^{(n-1)\times(n-1)}$ zu erhalten, für die

$$\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1 = \begin{pmatrix}\lambda & \mathbf{R}_1\\ \mathbf{0} & \mathbf{A}_1\end{pmatrix}$$

gilt. Wir wenden die Induktionsvoraussetzung auf $\mathbf{A}_1$ an und erhalten eine orthonormale Matrix $\mathbf{Q}_2 \in \mathbb{C}^{(n-1)\times(n-1)}$ und eine obere Dreiecksmatrix $\mathbf{R}_2 \in \mathbb{C}^{(n-1)\times(n-1)}$ mit

$$\mathbf{Q}_2^* \mathbf{A}_1 \mathbf{Q}_2 = \mathbf{R}_2.$$

Nun können wir

$$\mathbf{Q} := \mathbf{Q}_1 \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}, \qquad \mathbf{R} := \begin{pmatrix} \lambda & \mathbf{R}_1 \mathbf{Q}_2 \\ \mathbf{0} & \mathbf{R}_2 \end{pmatrix}$$

definieren und erhalten

$$\begin{aligned} \mathbf{Q}^* \mathbf{A} \mathbf{Q} &= \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^* \end{pmatrix} \mathbf{Q}_1^* \mathbf{A} \mathbf{Q}_1 \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^* \end{pmatrix} \begin{pmatrix} \lambda & \mathbf{R}_1 \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix} \\ &= \begin{pmatrix} \lambda & \mathbf{R}_1 \mathbf{Q}_2 \\ \mathbf{0} & \mathbf{Q}_2^* \mathbf{A}_1 \mathbf{Q}_2 \end{pmatrix} = \begin{pmatrix} \lambda & \mathbf{R}_1 \mathbf{Q}_2 \\ \mathbf{0} & \mathbf{R}_2 \end{pmatrix} = \mathbf{R}. \end{aligned}$$

Da $\mathbf{R}_2$ eine obere Dreiecksmatrix ist, muss auch $\mathbf{R}$ eine obere Dreiecksmatrix sein, also ist der Beweis vollständig. ∎

Bemerkung 3.40 *Im Beweis von Satz 3.39 können wir offenbar die Reihenfolge der Eigenwerte beliebig wählen und so beispielsweise auf der Diagonalen von* $\mathbf{R}$ *jede Anordnung erreichen.*

Die Gleichung $\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R}$ *ist äquivalent zu* $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{R}$*, und indem wir diese Gleichung auf die ersten p Spalten einschränken und die Dreiecksstruktur von* $\mathbf{R}$ *ausnutzen, folgt sofort, dass der Aufspann der ersten p Spalten von* $\mathbf{Q}$ *jeweils ein invarianter Teilraum von* $\mathbf{A}$ *sein muss.*

3.7 Diagonalisierbarkeit durch unitäre Transformationen

Die Matrix auf obere Dreiecksgestalt bringen zu können hilft uns zwar bei der Bestimmung der Eigenwerte, allerdings nur sehr bedingt bei der Untersuchung der Eigenvektoren. Insbesondere wird für viele Beweise eine *Orthonormalbasis* von Eigenvektoren benötigt, die wir Satz 3.39 nicht unmittelbar entnehmen können.

Für reelle Eigenwerte lässt sich relativ leicht klären, welche Eigenschaften eine Matrix aufweisen muss, um sich unitär diagonalisieren zu lassen: Wenn eine *reelle* Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ und einer unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ mit

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D}$$

existieren, folgt

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$$

und damit bereits

$$\mathbf{A}^* = \mathbf{Q}^{**}\mathbf{D}^*\mathbf{Q}^* = \mathbf{Q}\mathbf{D}\mathbf{Q}^* = \mathbf{A},$$

also muss $\mathbf{A}$ mindestens selbstadjungiert sein. Dank Satz 3.39 können wir feststellen, dass diese Eigenschaft auch schon ausreicht:

Folgerung 3.41 (Selbstadjungierte Matrizen) *Sei $\mathbf{A} \in \mathbb{C}^{n\times n}$. Es existieren genau dann eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ und eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ mit*

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D}, \tag{3.16}$$

wenn $\mathbf{A}$ selbstadjungiert ist.

Beweis. Die Richtung „⇒“ haben wir oben bereits bewiesen.

Sei nun also $\mathbf{A}$ selbstadjungiert. Nach Satz 3.39 existieren eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine rechte obere Dreiecksmatrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ derart, dass

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R}$$

gilt. Indem wir zu der Adjungierten übergehen erhalten wir

$$\mathbf{R}^* = \mathbf{Q}^*\mathbf{A}^*\mathbf{Q}^{**} = \mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R}.$$

Also ist die Dreiecksmatrix $\mathbf{R}$ selbstadjungiert. Daraus folgt $r_{ij} = \bar{r}_{ji}$ für alle $i, j \in [1 : n]$, also muss $\mathbf{R}$ eine Diagonalmatrix mit reellen Einträgen sein. ∎

Für eine reelle Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ haben wir bisher nur bewiesen, dass sie sich mit einer *komplexen* unitären Matrix diagonalisieren lässt. Indem wir den Beweis des Satzes 3.39 in leicht modifizierter Form nachvollziehen, können wir zeigen, dass sich auch eine *reelle* unitäre Matrix finden lässt.

Satz 3.42 (Courant-Fischer) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert. Die Rayleigh-Quotientenabbildung*

$$\Lambda_A \colon \mathbb{K}^n \setminus \{\mathbf{0}\} \to \mathbb{R}, \qquad \mathbf{x} \mapsto \frac{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle},$$

besitzt ein Maximum, dieses Maximum ist gerade der größte Eigenwert der Matrix $\mathbf{A}$, und jedes seiner Urbilder ist ein Eigenvektor zu diesem Eigenwert.

Beweis. Zunächst müssen wir nachprüfen, dass Λ_A wohldefiniert ist. Sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Dann gilt mit Lemma 3.16 die Gleichung

$$\Lambda_A(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\langle \mathbf{A}^*\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\overline{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle}}{\langle \mathbf{x}, \mathbf{x}\rangle} = \overline{\Lambda_A(\mathbf{x})},$$

also muss $\Lambda_A(\mathbf{x})$ reell sein.

Die Rayleigh-Quotientenabbildung ist invariant unter Skalierung, denn für alle $\alpha \in \mathbb{K}$ gilt

$$\Lambda_A(\alpha\mathbf{x}) = \frac{\langle \alpha\mathbf{x}, \alpha\mathbf{A}\mathbf{x}\rangle}{\langle \alpha\mathbf{x}, \alpha\mathbf{x}\rangle} = \frac{|\alpha|^2\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle}{|\alpha|^2\langle \mathbf{x}, \mathbf{x}\rangle} = \Lambda_A(\mathbf{x}), \tag{3.17}$$

also dürfen wir uns auf der Suche nach dem Maximum auf die Einheitssphäre

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{K}^n \ : \ \|\mathbf{x}\| = 1\}$$

beschränken. Diese Menge ist beschränkt und abgeschlossen, also nach dem Satz von Heine-Borel kompakt. Als stetige Abbildung nimmt Λ_A auf der kompakten Menge $\mathbf{S}$ ein Maximum an, das wir mit $\lambda \in \mathbb{R}$ bezeichnen. Da es sich um ein Maximum handelt, finden wir einen Vektor $\mathbf{e} \in \mathcal{S}$ mit $\lambda = \Lambda_A(\mathbf{e})$.

Seien nun $\mathbf{y} \in \mathbb{K}^n$ und $\alpha \in (0, 1/\|\mathbf{y}\|)$ (mit der Konvention $1/0 = \infty$) gegeben. Dann gilt

$$\|\mathbf{e} + \alpha\mathbf{y}\| \geq \|\mathbf{e}\| - |\alpha|\,\|\mathbf{y}\| > 1 - 1 = 0,$$

also $\mathbf{e} + \alpha\mathbf{y} \neq \mathbf{0}$. Da λ das Maximum der Rayleigh-Quotientenabbildung ist, gilt

$$\frac{\langle \mathbf{e} + \alpha\mathbf{y}, \mathbf{A}(\mathbf{e} + \alpha\mathbf{y})\rangle}{\langle \mathbf{e} + \alpha\mathbf{y}, \mathbf{e} + \alpha\mathbf{y}\rangle} = \Lambda_A(\mathbf{e} + \alpha\mathbf{y}) \leq \lambda.$$

Es folgt

$$\begin{aligned}
\lambda\|\mathbf{e}\|^2 + 2\lambda\alpha\,\mathrm{Re}\langle \mathbf{y}, \mathbf{e}\rangle + \lambda\alpha^2\|\mathbf{y}\|^2 &= \lambda(\langle \mathbf{e}, \mathbf{e}\rangle + \langle \alpha\mathbf{y}, \mathbf{e}\rangle + \langle \mathbf{e}, \alpha\mathbf{y}\rangle + \langle \alpha\mathbf{y}, \alpha\mathbf{y}\rangle) \\
&= \lambda\langle \mathbf{e} + \alpha\mathbf{y}, \mathbf{e} + \alpha\mathbf{y}\rangle \geq \langle \mathbf{e} + \alpha\mathbf{y}, \mathbf{A}(\mathbf{e} + \alpha\mathbf{y})\rangle \\
&= \langle \mathbf{e}, \mathbf{A}\mathbf{e}\rangle + \langle \alpha\mathbf{y}, \mathbf{A}\mathbf{e}\rangle + \langle \mathbf{e}, \alpha\mathbf{A}\mathbf{y}\rangle + \langle \alpha\mathbf{y}, \alpha\mathbf{A}\mathbf{y}\rangle \\
&= \langle \mathbf{e}, \mathbf{A}\mathbf{e}\rangle + 2\alpha\,\mathrm{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e}\rangle + \alpha^2\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle \\
&= \lambda\|\mathbf{e}\|^2 + 2\alpha\,\mathrm{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e}\rangle + \alpha^2\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle.
\end{aligned}$$

Wir dürfen $\lambda\|\mathbf{e}\|^2$ auf beiden Seiten streichen, die Terme umsortieren und durch α dividieren, um

$$\begin{aligned}
\alpha^2(\langle \mathbf{y}, \lambda\mathbf{y}\rangle - \langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle) &\geq 2\alpha(\mathrm{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e}\rangle - \mathrm{Re}\langle \mathbf{y}, \lambda\mathbf{e}\rangle), \\
\alpha\langle \mathbf{y}, \lambda\mathbf{y} - \mathbf{A}\mathbf{y}\rangle &\geq 2\,\mathrm{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e} - \lambda\mathbf{e}\rangle
\end{aligned}$$

zu erhalten. Wir setzen $\mathbf{y} := \mathbf{A}\mathbf{e} - \lambda\mathbf{e}$ und erhalten

$$\alpha\langle \mathbf{y}, \lambda\mathbf{y} - \mathbf{A}\mathbf{y}\rangle \geq 2\|\mathbf{A}\mathbf{e} - \lambda\mathbf{e}\|.$$

Da wir α beliebig klein wählen dürfen, folgt $0 \geq \|\mathbf{A}\mathbf{e} - \lambda\mathbf{e}\|^2$, also muss $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$ gelten. Damit ist $\mathbf{e}$ ein Eigenvektor zu dem Eigenwert λ.

Sei nun $\tilde{\lambda} \in \mathbb{K}$ ein beliebiger weiterer Eigenwert der Matrix $\mathbf{A}$, und sei $\widetilde{\mathbf{e}} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein zugehöriger Eigenvektor, der Einfachheit halber so skaliert, dass $\|\widetilde{\mathbf{e}}\| = 1$ gilt. Dann gilt wegen $\widetilde{\mathbf{e}} \in \mathcal{S}$ und der Wahl des Maximums λ die Ungleichung

$$\lambda \geq \Lambda_A(\widetilde{\mathbf{e}}) = \frac{\langle \widetilde{\mathbf{e}}, \mathbf{A}\widetilde{\mathbf{e}}\rangle}{\langle \widetilde{\mathbf{e}}, \widetilde{\mathbf{e}}\rangle} = \frac{\langle \widetilde{\mathbf{e}}, \tilde{\lambda}\widetilde{\mathbf{e}}\rangle}{\langle \widetilde{\mathbf{e}}, \widetilde{\mathbf{e}}\rangle} = \tilde{\lambda}\frac{\langle \widetilde{\mathbf{e}}, \widetilde{\mathbf{e}}\rangle}{\langle \widetilde{\mathbf{e}}, \widetilde{\mathbf{e}}\rangle} = \tilde{\lambda},$$

also ist λ nicht nur das Maximum der Rayleigh-Quotientenabbildung, sondern tatsächlich der größte Eigenwert. ∎

Satz 3.43 (Reelle Diagonalisierbarkeit) *Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$. Eine* reelle *unitäre Matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ mit (3.16) existieren genau dann, wenn $\mathbf{A}$ selbstadjungiert ist.*

Beweis. „⇒“: Es gelte (3.16). Dann folgt

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^* = \mathbf{Q}\mathbf{D}^*\mathbf{Q}^* = \mathbf{A}^*,$$

also ist $\mathbf{A}$ selbstadjungiert.

„⇐“: Sei $\mathbf{A}$ selbstadjungiert. Wie schon bei der Schur-Zerlegung gehen wir induktiv vor. Der Fall $n = 1$ ist trivial.

Gelte die Behauptung für $n-1 \in \mathbb{N}$. Wir fassen $\mathbf{A}$ zunächst als komplexe Matrix auf, die wegen Satz 3.8 mindestens einen Eigenwert $\lambda \in \mathbb{C}$ besitzen muss. Sei $\mathbf{x} \in \mathbb{C}^n$ ein zugehöriger Eigenvektor mit $\|\mathbf{x}\| = 1$. Da das Skalarprodukt eine Sesquilinearform ist, gilt

$$\begin{aligned}\lambda = \lambda\|\mathbf{x}\|^2 = \lambda\langle\mathbf{x},\mathbf{x}\rangle = \langle\mathbf{x},\lambda\mathbf{x}\rangle = \langle\mathbf{x},\mathbf{A}\mathbf{x}\rangle = \langle\mathbf{A}^*\mathbf{x},\mathbf{x}\rangle \\ = \langle\mathbf{A}\mathbf{x},\mathbf{x}\rangle = \langle\lambda\mathbf{x},\mathbf{x}\rangle = \bar{\lambda}\langle\mathbf{x},\mathbf{x}\rangle = \bar{\lambda}\|\mathbf{x}\|^2 = \bar{\lambda},\end{aligned}$$

also erhalten wir $\lambda \in \mathbb{R}$. Damit ist $\lambda\mathbf{I} - \mathbf{A}$ eine reelle singuläre Matrix, also muss es auch einen reellen Eigenvektor $\mathbf{x} \in \mathbb{R}^n$ mit $\|\mathbf{x}\| = 1$ geben.

Sei $\mathbf{Q}_1 \in \mathbb{R}^{n\times n}$ eine Householder-Matrix, die $\mathbf{x}$ unitär auf den ersten kanonischen Einheitsvektor $\delta^{(1)}$ abbildet. Dann gilt

$$\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1\delta^{(1)} = \mathbf{Q}_1^*\mathbf{A}\mathbf{x} = \lambda\mathbf{Q}_1^*\mathbf{x} = \lambda\delta^{(1)}.$$

Da $\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1$ wieder eine selbstadjungierte Matrix ist, muss es die Form

$$\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1 = \begin{pmatrix}\lambda & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{A}}\end{pmatrix}$$

mit einer selbstadjungierten Matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{(n-1)\times(n-1)}$ aufweisen.

Also können wir die Induktionsvoraussetzung auf $\widehat{\mathbf{A}}$ anwenden und wie im Beweis von Satz 3.39 fortfahren. ∎

Definition 3.44 (Normale Matrizen) *Eine Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ heißt* normal, *wenn*

$$\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$$

gilt.

Lemma 3.45 (Metrische Äquivalenz) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine Matrix. $\mathbf{A}$ ist genau dann normal, wenn*

$$\|\mathbf{A}\mathbf{x}\| = \|\mathbf{A}^*\mathbf{x}\|$$

für alle $\mathbf{x} \in \mathbb{K}^n$ gilt.

Beweis. Zunächst stellen wir fest, dass für alle $\mathbf{x} \in \mathbb{K}^n$ die Gleichungen

$$\|\mathbf{A}\mathbf{x}\|^2 = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}\rangle = \langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle, \qquad \|\mathbf{A}^*\mathbf{x}\|^2 = \langle \mathbf{A}^*\mathbf{x}, \mathbf{A}^*\mathbf{x}\rangle = \langle \mathbf{A}\mathbf{A}^*\mathbf{x}, \mathbf{x}\rangle$$

gelten.
„⇒“: Sei $\mathbf{A}$ normal. Dann folgt aus diesen Gleichungen bereits

$$\|\mathbf{A}\mathbf{x}\|^2 = \langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}\mathbf{A}^*\mathbf{x}, \mathbf{x}\rangle = \|\mathbf{A}^*\mathbf{x}\|^2$$

für alle $\mathbf{x} \in \mathbb{K}^n$.
„⇐“: Gelte nun die Gleichheit der Normen. Dann folgt

$$\langle (\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*)\mathbf{x}, \mathbf{x}\rangle = \|\mathbf{A}\mathbf{x}\|^2 - \|\mathbf{A}^*\mathbf{x}\|^2 = 0$$

für alle $\mathbf{x} \in \mathbb{K}^n$, und da $\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*$ eine selbstadjungierte Matrix ist, können wir Lemma 3.32 anwenden, um zu folgern, dass sie gleich null sein muss. ∎

Lemma 3.46 (Normale Dreiecksmatrizen) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine obere Dreiecksmatrix. Dann ist $\mathbf{A}$ genau dann normal, wenn die Matrix diagonal ist.*

Beweis. Übungsaufgabe. ∎

Folgerung 3.47 (Komplexe Diagonalisierbarkeit) *Sei $\mathbf{A} \in \mathbb{C}^{n\times n}$. Eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ mit*

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D} \tag{3.18}$$

existieren genau dann, wenn $\mathbf{A}$ normal ist.

Beweis. „⇒“: Setze zunächst voraus, dass $\mathbf{Q}, \mathbf{D}$ wie in (3.18) existieren. Wegen $\mathbf{D}\mathbf{D}^* = \mathbf{D}^*\mathbf{D}$ folgt

$$\begin{aligned}\mathbf{A}^*\mathbf{A} &= \mathbf{Q}^*\mathbf{D}^*\mathbf{Q}\mathbf{Q}^*\mathbf{D}\mathbf{Q} = \mathbf{Q}^*\mathbf{D}^*\mathbf{D}\mathbf{Q}\\ &= \mathbf{Q}^*\mathbf{D}\mathbf{D}^*\mathbf{Q} = \mathbf{Q}^*\mathbf{D}\mathbf{Q}\mathbf{Q}^*\mathbf{D}^*\mathbf{Q} = \mathbf{A}\mathbf{A}^*,\end{aligned}$$

also ist $\mathbf{A}$ normal.
„⇐“: Für die umgekehrte Implikation seien $\mathbf{Q}, \mathbf{R} \in \mathbb{C}^{n\times n}$ die Matrizen der Schur-Zerlegung (3.15). Da $\mathbf{A}$ normal ist, gilt

$$\begin{aligned}\mathbf{R}^*\mathbf{R} &= \mathbf{Q}^*\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{Q}^*\mathbf{A}^*\mathbf{A}\mathbf{Q}\\ &= \mathbf{Q}^*\mathbf{A}\mathbf{A}^*\mathbf{Q} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}\mathbf{Q}^*\mathbf{A}^*\mathbf{Q} = \mathbf{R}\mathbf{R}^*,\end{aligned}$$

also ist $\mathbf{R}$ normal. Nach Lemma 3.46 muss $\mathbf{R}$ damit bereits eine Diagonalmatrix sein. ∎

Der Name *Spektralnorm* legt nahe, dass diese Norm in enger Beziehung zu dem Spektrum stehen dürfte. Das folgende Lemma bestätigt diese Vermutung.

Lemma 3.48 (Spektralradius) *Sei $\mathbf{A} \in \mathbb{C}^{n\times n}$ eine Matrix. Den Betrag ihres betragsgrößten Eigenwerts*

$$\varrho(\mathbf{A}) := \max\{|\lambda| \ : \ \lambda \in \sigma(\mathbf{A})\}$$

nennen wir ihren Spektralradius. *Falls $\mathbf{A}$ normal ist, gilt $\varrho(\mathbf{A}) = \|\mathbf{A}\|$.*

Beweis. Nach Folgerung 3.47 finden wir eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ mit $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$.

Mit Definition 3.18 und (3.13) sowie der Substitution $\widehat{\mathbf{z}} = \mathbf{Q}^*\mathbf{z}$ erhalten wir

$$\begin{aligned}\|\mathbf{A}\| &= \max\left\{\frac{\|\mathbf{A}\mathbf{z}\|}{\|\mathbf{z}\|} \ : \ \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} = \max\left\{\frac{\|\mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{z}\|}{\|\mathbf{z}\|} \ : \ \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} \\ &= \max\left\{\frac{\|\mathbf{D}\mathbf{Q}^*\mathbf{z}\|}{\|\mathbf{z}\|} \ : \ \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} = \max\left\{\frac{\|\mathbf{D}\widehat{\mathbf{z}}\|}{\|\mathbf{Q}\widehat{\mathbf{z}}\|} \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} \\ &= \max\left\{\frac{\|\mathbf{D}\widehat{\mathbf{z}}\|}{\|\widehat{\mathbf{z}}\|} \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} = \|\mathbf{D}\|.\end{aligned}$$

Da $\mathbf{A}$ und $\mathbf{D}$ ähnliche Matrizen sind, muss nach Lemma 3.12 ein $j \in [1:n]$ mit $d_{jj} = \varrho(\mathbf{A})$ existieren. Für jeden Vektor $\widehat{\mathbf{z}} \in \mathbb{K}^n$ gilt

$$\|\mathbf{D}\widehat{\mathbf{z}}\|^2 = \sum_{i=1}^{n} |d_{ii}|^2|\widehat{z}_i|^2 \leq \sum_{i=1}^{n} |d_{jj}|^2|\widehat{z}_i|^2 = \varrho(\mathbf{A})^2 \sum_{i=1}^{n} |\widehat{z}_i|^2 = \varrho(\mathbf{A})^2\|\widehat{\mathbf{z}}\|^2.$$

Andererseits gilt auch

$$\|\mathbf{D}\delta^{(j)}\| = \|d_{jj}\delta^{(j)}\| = \varrho(\mathbf{A})\|\delta^{(j)}\|,$$

also folgt $\varrho(\mathbf{A}) = \|\mathbf{D}\|$ und der Beweis ist abgeschlossen. ∎

3.8 Was kommt nach der Schur-Zerlegung?

Es stellt sich die Frage, ob sich eine Matrix über die Schur-Zerlegung hinaus durch Ähnlichkeitstransformationen weiter vereinfachen lässt. Um die „Einfachheit“ messen zu können, benötigen wir eine geeignete Norm:

Definition 3.49 (Frobenius-Norm) *Die Abbildung*

$$\|\cdot\|_F : \mathbb{K}^{n\times m} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{A} \mapsto \left(\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2\right)^{1/2},$$

ist eine Norm auf $\mathbb{K}^{n\times m}$ und wird als Frobenius-Norm *bezeichnet.*

Die für unsere Untersuchungen wesentliche Eigenschaft der Frobenius-Norm ist ihre Invarianz unter orthonormalen Transformationen:

Lemma 3.50 (Invarianz) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times m}$. *Seien* $\mathbf{P} \in \mathbb{K}^{n\times n}$ *und* $\mathbf{Q} \in \mathbb{K}^{m\times m}$ *unitäre Matrizen. Dann gilt*

$$\|\mathbf{PA}\|_F = \|\mathbf{A}\|_F = \|\mathbf{AQ}\|_F.$$

Beweis. Für $i \in [1:m]$ definieren wir den Vektor $\mathbf{a}_i$ durch

$$(\mathbf{a}_i)_j = A_{ji}, \qquad \text{für alle } j \in [1:n],$$

er entspricht also gerade der i-ten Spalte von $\mathbf{A}$. Nach Definition 3.49 und weil die euklidische Norm invariant unter unitären Transformationen ist gilt

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{m} \|\mathbf{a}_i\|_2^2 = \sum_{i=1}^{m} \|\mathbf{P}\mathbf{a}_i\|_2^2 = \|\mathbf{PA}\|_F^2.$$

Wegen

$$\|\mathbf{A}\|_F^2 = \|\mathbf{A}^*\|_F^2$$

folgt aus der Orthogonalität von $\mathbf{Q}^*$ auch

$$\|\mathbf{A}\|_F^2 = \|\mathbf{A}^*\|_F^2 = \|\mathbf{Q}^*\mathbf{A}^*\|_F^2 = \|\mathbf{AQ}\|_F^2.$$

Offenbar lässt sich der Beweis auch auf rechteckige orthogonale Matrizen $\mathbf{P}$ und $\mathbf{Q}^*$ erweitern. ∎

Als Maß für die „Diagonalität“ einer Matrix verwenden wir die Norm ihrer Außerdiagonaleinträge:

Definition 3.51 *Für* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *definieren wir*

$$\operatorname{off}(\mathbf{A}) := \left(\sum_{\substack{i,j=1 \\ i\neq j}}^{n} |A_{ij}|^2 \right)^{1/2} = \|\mathbf{A} - \operatorname{diag}(A_{11}, \ldots, A_{nn})\|_F.$$

Jetzt lässt sich zeigen, dass diese Größe für jede Schur-Zerlegung einer festen Matrix dieselbe ist, sich also eine Matrix mit orthonormalen Transformationen nicht beliebig weit einer Diagonalmatrix annähern lässt:

Satz 3.52 (Invariante) *Sei* $\mathbf{A} \in \mathbb{C}^{n\times n}$. *Wir bezeichnen die Nullstellen des charakteristischen Polynoms* p_A *mit* $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, *so dass*

$$p_A(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i)$$

gilt. Auf dieser Grundlage definieren wir

$$\Delta^2(\mathbf{A}) := \|\mathbf{A}\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2.$$

Für alle unitären Matrizen $\mathbf{Q} \in \mathbb{C}^{n\times n}$ *und für alle oberen Dreiecksmatrizen* $\mathbf{R} \in \mathbb{C}^{n\times n}$ *mit*

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R}$$

gilt dann die Gleichung

$$\operatorname{off}^2(\mathbf{R}) = \Delta^2(\mathbf{A}).$$

Jede unitäre Ähnlichkeitstransformation auf obere Dreiecksgestalt wird also denselben Außerdiagonalanteil haben, insofern ist eine beliebige Schur-Zerlegung in dieser Hinsicht schon optimal.

Beweis. Sei $\mathbf{D} = \operatorname{diag}(R_{11}, \ldots, R_{nn})$. Dann gilt

$$p_D(\lambda) = \det(\lambda\mathbf{I} - \mathbf{D}) = \det(\lambda\mathbf{I} - \mathbf{R}) = p_R(\lambda) = p_A(\lambda),$$

also gibt es zu jedem μ-fachen Eigenwert von $\mathbf{A}$ einen μ-fachen Eigenwert von $\mathbf{D}$, und da die Eigenwerte von $\mathbf{D}$ durch die Diagonalelemente gegeben sind, muss

$$\|\mathbf{D}\|_F^2 = \sum_{i=1}^{n} |\lambda_i|^2$$

gelten. Wir erhalten

$$\begin{aligned}\operatorname{off}^2(\mathbf{R}) &= \|\mathbf{R} - \operatorname{diag}(R_{11}, \ldots, R_{nn})\|_F^2 = \|\mathbf{R} - \mathbf{D}\|_F^2 = \|\mathbf{R}\|_F^2 - \|\mathbf{D}\|_F^2 \\ &= \|\mathbf{Q}^*\mathbf{A}\mathbf{Q}\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2 = \|\mathbf{A}\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2 = \Delta^2(\mathbf{A}),\end{aligned}$$

also die gesuchte Identität. ∎

Da wir bereits wissen, dass nur bei normalen Matrizen eine unitäre Transformation auf Diagonalgestalt möglich ist, können wir die Größe $\Delta^2(\mathbf{A})$ als Maß dafür ansehen, wie „unnormal“ eine Matrix $\mathbf{A}$ ist: Für normale Matrizen verschwindet sie, für alle anderen muss sie größer als null sein.

3.9 Nicht-unitäre Transformationen

Lässt man statt der orthonormalen Transformationen allgemeinere Ähnlichkeitstransformationen zu, kann man die in der Schur-Zerlegung auftretende Dreiecksmatrix durch eine Block-Diagonalmatrix ersetzen.

Dazu betrachten wir eine Blockmatrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ & \mathbf{A}_{22} \end{pmatrix}$$

mit $\mathbf{A} \in \mathbb{K}^{n\times n}$, $\mathbf{A}_{11} \in \mathbb{K}^{m\times m}$, $\mathbf{A}_2 \in \mathbb{K}^{(n-m)\times(n-m)}$ und $\mathbf{A}_{12} \in \mathbb{K}^{m\times(n-m)}$ und suchen nach einer Ähnlichkeitstransformation der Form

$$\mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{X} \\ & \mathbf{I} \end{pmatrix}, \qquad \mathbf{B}^{-1} = \begin{pmatrix} \mathbf{I} & -\mathbf{X} \\ & \mathbf{I} \end{pmatrix}$$

mit $\mathbf{X} \in \mathbb{K}^{n\times(n-m)}$, die $\mathbf{A}$ in Block-Diagonalform überführt. Aufgrund der Gleichung

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{B} = \begin{pmatrix} \mathbf{I} & -\mathbf{X} \\ & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{11}\mathbf{X} + \mathbf{A}_{12} \\ & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{11}\mathbf{X} - \mathbf{X}\mathbf{A}_{22} + \mathbf{A}_{12} \\ & \mathbf{A}_{22} \end{pmatrix}$$

ist das äquivalent dazu, ein $\mathbf{X} \in \mathbb{K}^{m\times(n-m)}$ so zu finden, dass

$$\mathbf{A}_{11}\mathbf{X} - \mathbf{X}\mathbf{A}_{22} = -\mathbf{A}_{12}$$

gilt. Gleichungen dieser Form nennt man *Sylvester-Gleichungen.*

Satz 3.53 (Sylvester-Gleichung) *Seien $\mathbf{A} \in \mathbb{C}^{n\times n}$ und $\mathbf{B} \in \mathbb{C}^{m\times m}$ gegeben. Die* Sylvester-Gleichung

$$\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B} = \mathbf{C} \tag{3.19}$$

besitzt genau dann für alle $\mathbf{C} \in \mathbb{C}^{n\times m}$ eine Lösung, wenn $\sigma(\mathbf{A}) \cap \sigma(\mathbf{B}) = \emptyset$ gilt.

Beweis. Wir werden zeigen, dass die lineare Abbildung

$$\mathcal{L}\colon \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}, \qquad \mathbf{X} \mapsto \mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B},$$

genau dann injektiv ist, wenn $\sigma(\mathbf{A}) \cap \sigma(\mathbf{B}) = \emptyset$ gilt. Da $\mathcal{L}$ ein Automorphismus zwischen endlich-dimensionalen Räumen ist, sind Injektivität und Surjektivität dank des Dimensionssatzes äquivalent.

Gelte zunächst $\sigma(\mathbf{A}) \cap \sigma(\mathbf{B}) \neq \emptyset$. Dann existiert ein Eigenwert $\lambda \in \sigma(\mathbf{A}) \cap \sigma(\mathbf{B})$. Da das charakteristische Polynom der adjungierten Matrix $\mathbf{B}^*$ die Gleichung

$$p_{B^*}(z) = \det(z\mathbf{I} - \mathbf{B}^*) = \overline{\det(\bar{z}\mathbf{I} - \mathbf{B})} = \overline{p_B(\bar{z})} \qquad \text{für alle } z \in \mathbb{C}$$

erfüllt, ist $\bar{\lambda}$ ein Eigenwert der Matrix $\mathbf{B}^*$.

Also finden wir $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ und $\mathbf{y} \in \mathbb{C}^m \setminus \{\mathbf{0}\}$ mit

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \qquad \mathbf{B}^*\mathbf{y} = \bar{\lambda}\mathbf{y}.$$

Die Matrix

$$\mathbf{X} := \mathbf{x}\mathbf{y}^*$$

ist dann ungleich null und erfüllt

$$\mathcal{L}[\mathbf{X}] = \mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B} = \mathbf{A}\mathbf{x}\mathbf{y}^* - \mathbf{x}(\mathbf{B}^*\mathbf{y})^* = \lambda\mathbf{x}\mathbf{y}^* - \mathbf{x}(\bar{\lambda}\mathbf{y})^* = \mathbf{0}.$$

Wir haben also ein $\mathbf{X} \in \mathbb{C}^{n\times n} \setminus \{\mathbf{0}\}$ mit $\mathcal{L}[\mathbf{X}] = \mathbf{0}$ gefunden, somit ist $\mathcal{L}$ nicht injektiv.

Sei nun umgekehrt ein $\mathbf{X} \in \mathbb{C}^{n\times m} \setminus \{\mathbf{0}\}$ mit $\mathcal{L}[\mathbf{X}] = \mathbf{0}$ gegeben. Dann gilt

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B}.$$

Wir nehmen zunächst an, dass $\mathbf{B}$ eine obere Dreiecksmatrix ist. Sei $j \in [1 : m]$ der kleinste Index mit $\mathbf{X}\delta^{(j)} \neq \mathbf{0}$, also der Index der ersten Spalte der Matrix $\mathbf{X}$, die nicht

gleich null ist. Da wir $\mathbf{X} \neq \mathbf{0}$ vorausgesetzt haben, existiert ein derartiger Index. Wir setzen $\mathbf{e} := \mathbf{X}\delta^{(j)}$ und halten $\mathbf{e} \neq \mathbf{0}$ fest. Da $\mathbf{B}$ eine obere Dreiecksmatrix ist, gilt

$$\mathbf{Ae} = \mathbf{AX}\delta^{(j)} = \mathbf{XB}\delta^{(j)} = \mathbf{X}\sum_{i=1}^{j} b_{ij}\delta^{(i)} = \sum_{i=1}^{j} b_{ij}\mathbf{X}\delta^{(i)}.$$

Aufgrund unserer Wahl des Index j gilt $\mathbf{X}\delta^{(i)} = \mathbf{0}$ für alle $i \in [1 : j-1]$, so dass wir

$$\mathbf{Ae} = b_{jj}\mathbf{X}\delta^{(j)} = b_{jj}\mathbf{e}$$

erhalten. Also ist $\mathbf{e}$ ein Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert b_{jj}. Da $\mathbf{B}$ eine obere Dreiecksmatrix ist, ist b_{jj} auch ein Eigenwert der Matrix $\mathbf{B}$, es gilt also $b_{jj} \in \sigma(\mathbf{A}) \cap \sigma(\mathbf{B})$.

Den allgemeinen Fall können wir mit dem Satz 3.39 über die Schur-Zerlegung auf den bereits behandelten Sonderfall zurückführen: Wir finden eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{m\times m}$ und eine obere Dreiecksmatrix $\mathbf{R} \in \mathbb{K}^{m\times m}$ mit

$$\mathbf{B} = \mathbf{QRQ}^*,$$

und es gilt

$$\mathbf{AX} = \mathbf{XB} \iff \mathbf{AX} = \mathbf{XQRQ}^* \iff \mathbf{AXQ} = \mathbf{XQR}.$$

Wir setzen $\widehat{\mathbf{X}} := \mathbf{XQ}$ und erhalten

$$\mathbf{A}\widehat{\mathbf{X}} = \widehat{\mathbf{X}}\mathbf{R},$$

und da $\mathbf{R}$ nun eine obere Dreiecksmatrix ist, können wir mit dem bereits Gezeigten folgern, dass $\mathbf{A}$ und $\mathbf{R}$ einen gemeinsamen Eigenwert besitzen müssen. Da $\mathbf{R}$ und $\mathbf{B}$ ähnlich sind, folgt auch $\sigma(\mathbf{A}) \cap \sigma(\mathbf{B}) \neq \emptyset$. ∎

Satz 3.54 (Jordan-Normalform) *Sei $A \in \mathbb{C}^{n\times n}$. Dann existieren eine reguläre Matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ sowie $q \in \mathbb{N}$, $(m_i)_{i=1}^q$ in $\mathbb{N}$ und $(\lambda_i)_{i=1}^q$ in $\mathbb{C}$ mit*

$$\mathbf{X}^{-1}\mathbf{AX} = \operatorname{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_q),$$

wobei der Jordan-Block $\mathbf{J}_i \in \mathbb{C}^{m_i\times m_i}$ *die Gestalt*

$$\mathbf{J}_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}$$

besitzt, m_i die algebraische Vielfachheit von λ_i ist und $n = m_1 + m_2 + \ldots + m_q$ gilt.

Beweis. Die entscheidende Vorstufe findet sich als Theorem 7.1.6 in Golub/VanLoan, „Matrix Computations“, Johns Hopkins University Press 1996. Dort gibt es auch eine Referenz auf den vollständigen Beweis. ∎

Bei der Jordan-Normalform ist allerdings zu beachten, dass nicht-unitäre Ähnlichkeitstransformationen numerisch nicht ungefährlich sind, wie das folgende Beispiel zeigt:

Beispiel 3.55 *Wir betrachten die Matrix*

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

aus Beispiel 3.10. Wenn wir das untere Diagonalelement mit $\epsilon \in \mathbb{R} \setminus \{0\}$ stören, erhalten wir

$$\mathbf{A}_\epsilon = \begin{pmatrix} 1 & 1 \\ 0 & 1+\epsilon \end{pmatrix},$$

und diese Matrix besitzt das charakteristische Polynom

$$p_{A_\epsilon}(\lambda) = (1-\lambda)(1+\epsilon-\lambda),$$

also zwei verschiedene Eigenwerte. Da zu jedem Eigenwert mindestens ein Eigenvektor gehören muss und wir in einem zweidimensionalen Raum arbeiten, muss eine Basis aus Eigenvektoren exitieren. Beispielsweise können wir

$$\mathbf{X}_\epsilon := \begin{pmatrix} 1 & 1 \\ 0 & \epsilon \end{pmatrix}, \qquad \mathbf{X}_\epsilon^{-1} = \begin{pmatrix} 1 & -1/\epsilon \\ 0 & 1/\epsilon \end{pmatrix}$$

verwenden und erhalten die Gleichung

$$\mathbf{X}_\epsilon^{-1}\mathbf{A}_\epsilon\mathbf{X}_\epsilon = \begin{pmatrix} 1 & 0 \\ 0 & 1+\epsilon \end{pmatrix}.$$

Also ist $\mathbf{A}_\epsilon$ für jedes $\epsilon \neq 0$ diagonalisierbar.

Vom numerischen Standpunkt aus gesehen ist diese Diagonalisierbarkeit allerdings fragwürdig, da die Kondition beispielsweise in der Zeilensummennorm durch

$$\kappa_\infty(\mathbf{X}_\epsilon) = \|X_\epsilon\|_\infty \|X_\epsilon^{-1}\|_\infty = 2\left(1+\frac{1}{\epsilon}\right) = 2+\frac{2}{\epsilon}$$

gegeben ist und deshalb für kleiner werdendes ϵ schnell gegen unendlich strebt.

3.10 Eigenwerte nicht-negativer Matrizen *

Die Matrix (2.11) des „Minipoly“-Beispiels weist die Besonderheit auf, ausschließlich nicht-negative Einträge zu besitzen. Da wir den Eigenvektor zu dem maximalen Eigenwert 1 als invariantes Wahrscheinlichkeitsmaß interpretieren wollen, sind wir daran interessiert, einen Vektor zu erhalten, der keine negativen Einträge enthält.

Derartige Aufgabenstellungen lassen sich mit einer von Oskar Perron [3] und Georg Frobenius [1] entwickelten Methode untersuchen, die naheliegenderweise in der Literatur als *Perron-Frobenius-Theorie* bezeichnet wird.

Definition 3.56 (Nicht-negative Matrizen und Vektoren) *Seien* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *und* $\mathbf{x} \in \mathbb{R}^n$*. Wir definieren*

$$\begin{aligned}
\mathbf{A} \geq 0 &:\Longleftrightarrow \forall i,j \in [1:n] \ : \ a_{ij} \geq 0,\\
\mathbf{A} > 0 &:\Longleftrightarrow \forall i,j \in [1:n] \ : \ a_{ij} > 0,\\
\mathbf{x} \geq 0 &:\Longleftrightarrow \forall i \in [1:n] \ : \ x_i \geq 0,\\
\mathbf{x} > 0 &:\Longleftrightarrow \forall i \in [1:n] \ : \ x_i > 0.
\end{aligned}$$

Mit $K := \{\mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{x} \geq 0\}$ *bezeichnen wir den Kegel der nicht-negativen Vektoren.*

Wir gehen im Folgenden davon aus, dass eine Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ mit $\mathbf{A} \geq 0$ gegeben ist.

Für unsere Untersuchung verwenden wir einen auf Helmut Wieland [4] zurückgehenden Ansatz, der auf der Abbildung

$$r : K \setminus \{\mathbf{0}\} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{x} \mapsto \min\left\{ \frac{(\mathbf{A}\mathbf{x})_i}{x_i} \ : \ i \in [1:n],\ x_i \neq 0 \right\} \tag{3.20}$$

beruht. Für einen gegebenen Vektor $\mathbf{x} \in K$ gilt dann

$$r(\mathbf{x})x_i \leq (\mathbf{A}\mathbf{x})_i \qquad \text{für alle } i \in [1:n],$$

wobei für den Fall $x_i = 0$ die Nicht-Negativität und für den Fall $x_i \neq 0$ die Definition der Abbildung r herangezogen wird. Diese Eigenschaft können wir kompakt als

$$\mathbf{A}\mathbf{x} - r(\mathbf{x})\mathbf{x} \geq 0 \tag{3.21}$$

schreiben. Wenn $\mathbf{x}$ ein Eigenvektor wäre, würde in dieser Formel Gleichheit gelten, also bietet es sich an, nach Vektoren zu suchen, für die die linke Seite möglichst klein, also $r(\mathbf{x})$ möglichst groß wird.

Lemma 3.57 (Extremalvektoren) *Es gibt mindestens einen Vektor* $\mathbf{x}^* \in K \setminus \{\mathbf{0}\}$*, für den*

$$r(\mathbf{x}) \leq r(\mathbf{x}^*) \qquad \textit{für alle } \mathbf{x} \in K \setminus \{\mathbf{0}\} \tag{3.22}$$

gilt, für den also r *sein Maximum annimmt.*

Beweis. Sei $\mathbf{x} \in K \setminus \{\mathbf{0}\}$, und sei $\mathbf{e} \in K$ derjenige Vektor, dessen Einträge alle gleich eins sind. Aus (3.21) folgt

$$\langle \mathbf{e}, \mathbf{A}\mathbf{x}\rangle_2 - r(\mathbf{x})\langle \mathbf{e}, \mathbf{x}\rangle_2 = \sum_{i=1}^{n} (\mathbf{A}\mathbf{x})_i - r(\mathbf{x})x_i \geq 0,$$

also insbesondere auch

$$r(\mathbf{x}) \leq \frac{\langle \mathbf{e}, \mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{e}, \mathbf{x}\rangle_2}.$$

Wenn wir mir α den größten Eintrag des Vektors $\mathbf{A}^*\mathbf{e}$ bezeichnen, gilt

$$\langle \mathbf{e}, \mathbf{A}\mathbf{x}\rangle_2 = \langle \mathbf{A}^*\mathbf{e}, \mathbf{x}\rangle_2 = \sum_{i=1}^{n} (\mathbf{A}^*\mathbf{e})_i x_i \leq \sum_{i=1}^{n} \alpha x_i = \alpha \langle \mathbf{e}, \mathbf{x}\rangle_2,$$

also erhalten wir

$$r(\mathbf{x}) \leq \frac{\langle \mathbf{e}, \mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{e}, \mathbf{x}\rangle_2} \leq \frac{\alpha\langle \mathbf{e}, \mathbf{x}\rangle_2}{\langle \mathbf{e}, \mathbf{x}\rangle_2} = \alpha,$$

so dass der Wertebereich der Abbildung r in $[0, \alpha]$ enthalten sein muss. Demzufolge ist das Supremum

$$\varrho := \sup\{r(\mathbf{x}) \ : \ \mathbf{x} \in K_{\geq} \setminus \{\mathbf{0}\}\}$$

wohldefiniert und nicht-negativ.

Nach Definition des Supremums finden wir eine Folge von Vektoren $(\mathbf{x}^{(m)})_{m=1}^{\infty}$ in $K \setminus \{\mathbf{0}\}$ derart, dass

$$r(\mathbf{x}^{(m)}) \geq \varrho - 1/m \qquad \text{für alle } m \in \mathbb{N}$$

gilt. Ein Blick auf die Definition (3.20) zeigt

$$r(\alpha\mathbf{x}) = r(\mathbf{x}) \qquad \text{für alle } \alpha \in \mathbb{R}_{>0},\ \mathbf{x} \in K \setminus \{\mathbf{0}\},$$

die Skalierung der Vektoren spielt also keine Rolle, so dass wir ohne Beschränkung der Allgemeinheit davon ausgehen können, dass

$$\langle \mathbf{e}, \mathbf{x}^{(m)}\rangle_2 = 1 \qquad \text{für alle } m \in \mathbb{N}$$

gilt. Die Menge

$$T := \{\mathbf{x} \in K \ : \ \langle \mathbf{e}, \mathbf{x}\rangle_2 = 1\}$$

ist abgeschlossen und beschränkt, also nach dem Satz von Heine-Borel auch kompakt. Die Elemente der Folge $(\mathbf{x}^{(m)})_{m=1}^{\infty}$ liegen in dieser Menge, also muss die Folge mindestens einen Häufungspunkt $\mathbf{x}^* \in T$ besitzen.

Wenn r stetig wäre, könnten wir unmittelbar schließen, dass $r(\mathbf{x}^*) = \varrho$ gelten muss. Leider ist r „nicht ganz“ stetig, da sich die Menge, über die in (3.20) das Minimum gebildet wird, abhängig von den Nulleinträgen des Vektors $\mathbf{x}$ ändert.

Deshalb müssen wir die gewünschte Gleichung etwas ausführlicher nachprüfen. Dazu wählen wir ein beliebiges $\epsilon \in \mathbb{R}_{>0}$ und setzen

$$\delta_1 := \min\{x_i^* \ : \ x_i^* \neq 0\}.$$

Wegen $\mathbf{x}^* \in K \setminus \{\mathbf{0}\}$ dürfen wir $\delta_1 > 0$ festhalten.

Aufgrund der Stetigkeit der Funktionen

$$\mathbf{x} \mapsto \frac{(\mathbf{A}\mathbf{x})_i}{x_i} \qquad \text{für alle } i \in [1:n],\ x_i^* \neq 0$$

in der Nähe des Vektors $\mathbf{x}^*$ können wir ein $\delta_2 > 0$ so finden, dass

$$\frac{(\mathbf{A}\mathbf{x})_i}{x_i} \geq \frac{(\mathbf{A}\mathbf{x}^*)_i}{x_i} - \epsilon/2 \qquad \begin{array}{c}\text{für alle } \mathbf{x} \in K \text{ mit } \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \delta_2 \\ \text{und alle } i \in [1:n] \text{ mit } x_i^* \neq 0\end{array}$$

erfüllt ist. Wir setzen $\delta := \min\{\delta_1, \delta_2\}$.

Da $\mathbf{x}^*$ ein Häufungspunkt ist, finden wir ein $m \in \mathbb{N}$ derart, dass

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_2 < \delta, \qquad r(\mathbf{x}^{(m)}) \geq \varrho - \epsilon/2$$

gelten. Daraus folgt insbesondere

$$\begin{aligned} x_i^{(m)} &= x_i^* + x_i^{(m)} - x_i^* \geq x_i^* - |x_i^{(m)} - x_i^*| \\ &> x_i^* - \delta \geq 0 \qquad \text{für alle } i \in [1:n] \text{ mit } x_i^* \neq 0, \end{aligned}$$

so dass wir

$$\{i \in [1:n] \ : \ x_i^* \neq 0\} \subseteq \{i \in [1:n] \ : \ x_i^{(m)} \neq 0\}$$

erhalten. Daraus folgt insbesondere

$$\begin{aligned} r(\mathbf{x}^*) &= \min\left\{\frac{(\mathbf{A}\mathbf{x}^*)_i}{x_i^*} \ : \ i \in [1:n],\ x_i^* \neq 0\right\} \\ &\geq \min\left\{\frac{(\mathbf{A}\mathbf{x}^*)_i}{x_i^*} \ : \ i \in [1:n],\ x_i^{(m)} \neq 0\right\} \\ &\geq \min\left\{\frac{(\mathbf{A}\mathbf{x}^{(m)})_i}{x_i^{(m)}} - \epsilon/2 \ : \ i \in [1:n],\ x_i^{(m)} \neq 0\right\} \\ &= r(\mathbf{x}^{(m)}) - \epsilon/2 \geq \varrho - \epsilon. \end{aligned}$$

Da ϵ beliebig gewählt wurde, haben wir $r(\mathbf{x}^*) = \varrho$ bewiesen. ∎

Wir sind daran interessiert, Eigenvektoren zu finden, bei denen alle Komponenten positiv sind. Leider ist das nicht bei allen nicht-negativen Matrizen möglich, beispielsweise verschwindet bei allen Eigenvektoren der Matrix

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{3.23}$$

die zweite Komponente. Um diesen Fall auszuschließen, müssen wir eine zusätzliche Bedingung an die Matrix $\mathbf{A}$ stellen.

Definition 3.58 (Reduzible und irreduzible Matrizen) *Eine Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ nennen wir* reduzibel, *falls ein $k \in [1 : n-1]$, eine Permutationsmatrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ und Matrizen $\mathbf{A}_{11} \in \mathbb{K}^{k\times k}$, $\mathbf{A}_{12} \in \mathbb{K}^{k\times(n-k)}$ und $\mathbf{A}_{22} \in \mathbb{K}^{(n-k)\times(n-k)}$ so existieren, dass*

$$\mathbf{P}\mathbf{A}\mathbf{P}^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ & \mathbf{A}_{22} \end{pmatrix} \tag{3.24}$$

gilt. Falls eine Matrix nicht reduzibel ist, nennen wir sie irreduzibel.

Irreduzible Matrizen sind also gerade diejenigen, die sich nicht durch das Umsortieren von Indizes auf Block-Dreiecksform bringen lassen.

Lemma 3.59 (Irreduzible Matrix) *Falls* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *irreduzibel ist und* $\mathbf{A} \geq 0$ *gilt, existiert für jeden Vektor* $\mathbf{x} \in K \setminus \{\mathbf{0}\}$ *ein* $m \in \mathbb{N}_0$ *mit*

$$(\mathbf{A} + \mathbf{I})^m \mathbf{x} > 0.$$

Beweis. Sei $\mathbf{x} \in K \setminus \{\mathbf{0}\}$ gegeben. Wir definieren

$$\mathbf{x}^{(m)} := (\mathbf{A} + \mathbf{I})^m \mathbf{x} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Mit $\mathbf{A} \geq 0$ folgt daraus unmittelbar

$$\mathbf{x}^{(m+1)} = \mathbf{A}\mathbf{x}^{(m)} + \mathbf{x}^{(m)} \geq \mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0.$$

In dem Vektor $\mathbf{x}^{(m+1)}$ können also höchstens diejenigen Koeffizienten gleich null sein, die auch in $\mathbf{x}^{(m)}$ bereits gleich null waren. Da $\mathbf{x}^{(0)} = \mathbf{x} \neq \mathbf{0}$ gilt, kann insbesondere keiner der Vektoren $\mathbf{x}^{(m)}$ der Nullvektor sein.

Wären für ein $m \in \mathbb{N}_0$ dieselben Koeffizienten in $\mathbf{x}^{(m)}$ und $\mathbf{x}^{(m+1)}$ gleich null, könnten wir diese Koeffizienten mit einer Permutationsmatrix $\mathbf{P}$ in die letzten $n-k$ Komponenten umsortieren und so

$$\mathbf{P}\mathbf{x}^{(m)} = \begin{pmatrix} \widehat{\mathbf{x}}^{(m)} \\ \mathbf{0} \end{pmatrix}, \qquad \mathbf{P}\mathbf{x}^{(m+1)} = \begin{pmatrix} \widehat{\mathbf{x}}^{(m+1)} \\ \mathbf{0} \end{pmatrix}$$

mit $\widehat{\mathbf{x}}^{(m)}, \widehat{\mathbf{x}}^{(m+1)} \in \mathbb{R}^k$, $k \in [1 : n-1]$ sowie $\widehat{\mathbf{x}}^{(m)} > 0$ und $\widehat{\mathbf{x}}^{(m+1)} > 0$ zu erhalten.

Wir zerlegen die permutierte Matrix in der Form

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} := \mathbf{P}\mathbf{A}\mathbf{P}^{-1}, \qquad \mathbf{A}_{11} \in \mathbb{R}^{k\times k}, \qquad \mathbf{A}_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$$

und erhalten

$$\begin{aligned} \begin{pmatrix} \widehat{\mathbf{x}}^{(m+1)} \\ \mathbf{0} \end{pmatrix} = \mathbf{P}\mathbf{x}^{(m+1)} = \mathbf{P}(\mathbf{x}^{(m)} + \mathbf{A}\mathbf{x}^{(m)}) = \mathbf{P}\mathbf{x}^{(m)} + \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \mathbf{P}\mathbf{x}^{(m)} \\ = \begin{pmatrix} \widehat{\mathbf{x}}^{(m)} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{x}}^{(m)} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{x}}^{(m)} + \mathbf{A}_{11}\widehat{\mathbf{x}}^{(m)} \\ \mathbf{A}_{21}\widehat{\mathbf{x}}^{(m)} \end{pmatrix} \end{aligned}$$

folgen. Dank $\widehat{\mathbf{x}}^{(m)} > 0$ und $\mathbf{A}_{21} \geq 0$ müsste dann bereits $\mathbf{A}_{21} = \mathbf{0}$ gelten, also wäre (3.24) erfüllt, im Widerspruch zur Voraussetzung.

Damit kann für jedes $m \in \mathbb{N}$ der Vektor $\mathbf{x}^{(m+1)}$ nicht dieselben Nulleinträge wie $\mathbf{x}^{(m)}$ enthalten. Wir haben bereits gesehen, dass er auch nicht weitere Nulleinträge enthalten kann, also muss die Zahl der Nulleinträge sinken. Da $\mathbf{x}^{(0)} = \mathbf{x} \neq \mathbf{0}$ höchstens $n-1$ Nulleinträge aufweisen kann, muss deshalb $\mathbf{x}^{(m)} > 0$ spätestens für $m = n-1$ gelten. ∎

Bemerkung 3.60 (Irreduzible Matrix) *Ein Blick auf den vorangehenden Beweis zeigt, dass* $(\mathbf{I}+\mathbf{A})^{n-1}\mathbf{x} > 0$ *für* alle *Vektoren* $\mathbf{x} \in K \setminus \{\mathbf{0}\}$ *gilt. Indem wir für* $\mathbf{x}$ *die kanonischen Einheitsvektoren einsetzen folgt, dass jede Spalte der Matrix* $(\mathbf{I}+\mathbf{A})^{n-1}$ *in jeder Komponente echt größer als Null ist, also haben wir sogar* $(\mathbf{I}+\mathbf{A})^{n-1} > 0$ *erhalten.*

Falls $\mathbf{A}$ irreduzibel ist, können wir wie bereits angedeutet folgern, dass ein Vektor $\mathbf{x}^* \in K \setminus \{\mathbf{0}\}$ mit der in (3.22) gegebenen Extremaleigenschaft ein Eigenvektor sein muss.

Lemma 3.61 (Eigenvektor) *Sei* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *irreduzibel mit* $\mathbf{A} \geq 0$. *Sei* $\mathbf{x}^* \in K \setminus \{\mathbf{0}\}$ *ein Vektor mit der Eigenschaft (3.22) und sei* $\varrho := r(\mathbf{x}^*)$. *Dann gelten* $\mathbf{x}^* > 0$ *und*

$$\mathbf{A}\mathbf{x}^* = \varrho\mathbf{x}^*,$$

$\mathbf{x}^*$ *ist also ein Eigenvektor zu dem Eigenwert* ϱ.

Beweis. Aufgrund der Ungleichung (3.21) gilt

$$\mathbf{y} := \mathbf{A}\mathbf{x}^* - \varrho\mathbf{x}^* \geq 0.$$

Wir müssen nachweisen, dass $\mathbf{y} = \mathbf{0}$ gilt.

Dazu verwenden wir einen Widerspruchsbeweis: Wir nehmen $\mathbf{y} \neq \mathbf{0}$ an. Nach Lemma 3.59 existiert dann ein $m \in \mathbb{N}_0$ so, dass

$$(\mathbf{I}+\mathbf{A})^m\mathbf{y} > 0$$

erfüllt ist. Für den Vektor

$$\mathbf{z} := (\mathbf{I}+\mathbf{A})^m\mathbf{x}^*$$

folgt daraus

$$\begin{aligned}\mathbf{A}\mathbf{z} - \varrho\mathbf{z} &= \mathbf{A}(\mathbf{I}+\mathbf{A})^m\mathbf{x}^* - \varrho(\mathbf{I}+\mathbf{A})^m\mathbf{x}^* \\ &= (\mathbf{I}+\mathbf{A})(\mathbf{I}+\mathbf{A})^m\mathbf{x}^* - (\varrho+1)(\mathbf{I}+\mathbf{A})^m\mathbf{x}^* \\ &= (\mathbf{I}+\mathbf{A})^m(\mathbf{I}+\mathbf{A})\mathbf{x}^* - (\varrho+1)(\mathbf{I}+\mathbf{A})^m\mathbf{x}^* \\ &= (\mathbf{I}+\mathbf{A})^m(\mathbf{A}\mathbf{x}^* - \varrho\mathbf{x}^*) = (\mathbf{I}+\mathbf{A})^m\mathbf{y} > 0.\end{aligned}$$

Da $\varrho \geq r(\mathbf{z})$ nach (3.22) gilt, erhalten wir insbesondere

$$\mathbf{A}\mathbf{z} - r(\mathbf{z})\mathbf{z} \geq \mathbf{A}\mathbf{z} - \varrho\mathbf{z} > 0.$$

Nach Definition des Minimums (3.20) muss aber ein $i \in [1:n]$ mit $z_i \neq 0$ und

$$r(\mathbf{z}) = \frac{(\mathbf{A}\mathbf{z})_i}{z_i}, \qquad r(\mathbf{z})z_i = (\mathbf{A}\mathbf{z})_i, \qquad (\mathbf{A}\mathbf{z} - r(\mathbf{z})\mathbf{z})_i = 0$$

existieren, im Widerspruch zu der obigen Ungleichung.

Also muss $\mathbf{y} = \mathbf{0}$ gelten, somit ist $\mathbf{x}^*$ ein Eigenvektor zu dem Eigenwert ϱ. ∎

In Kombination mit Lemma 3.57 folgt also, dass eine nicht-negative irreduzible Matrix mindestens einen Eigenvektor zu dem Eigenwert ϱ besitzt und dass die Koeffizienten dieses Eigenvektors alle echt positiv sind.

Der Eigenwert ϱ lässt sich sogar als im Betrag maximaler Eigenwert der gesamten Matrix $\mathbf{A}$ identifizieren:

Lemma 3.62 (Maximaler Eigenwert) *Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine irreduzible Matrix mit $\mathbf{A} \geq 0$. Sei ϱ wie in Lemma 3.61 definiert, und sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein beliebiger Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert $\lambda \in \mathbb{K}$. Dann gilt $|\lambda| \leq \varrho$.*

Beweis. Wir definieren den Vektor $\mathbf{y} \in K \setminus \{\mathbf{0}\}$ durch

$$y_i := |x_i| \qquad \text{für alle } i \in [1:n]$$

und folgern mit der Dreiecksungleichung

$$|\lambda| y_i = |\lambda x_i| = |(\mathbf{A}\mathbf{x})_i| = \left|\sum_{j=1}^{n} a_{ij} x_j\right| \leq \sum_{j=1}^{n} a_{ij}|x_j| = (\mathbf{A}\mathbf{y})_i \qquad \text{für alle } i \in [1:n].$$

Für jedes $i \in [1:n]$ mit $y_i \neq 0$ folgt daraus

$$|\lambda| \leq \frac{(\mathbf{A}\mathbf{y})_i}{y_i},$$

also nach (3.20) insbesondere

$$|\lambda| \leq r(\mathbf{y}) \leq \varrho,$$

und damit ist die gewünschte Aussage bewiesen. ∎

Wir können die Ergebnisse dieses Abschnitts in dem folgenden Satz zusammenfassen:

Satz 3.63 (Perron-Frobenius) *Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ irreduzibel mit $\mathbf{A} \geq 0$. Der* Spektralradius *der Matrix ist durch*

$$\varrho(\mathbf{A}) := \max\{|\lambda| \ : \ \lambda \in \sigma(\mathbf{A})\}$$

gegeben. Er ist ein Eigenwert der Matrix $\mathbf{A}$, zu dem ein Eigenvektor $\mathbf{x} \in \mathbb{R}^n$ mit $\mathbf{x} > 0$ existiert.

Beweis. Sei $\mathbf{x}^* \in K \setminus \{\mathbf{0}\}$ der nach Lemma 3.57 existierende Vektor mit der Eigenschaft (3.22). Nach Lemma 3.61 ist er ein Eigenvektor zu dem Eigenwert $\varrho = r(\mathbf{x}^*)$ und erfüllt $\mathbf{x}^* > 0$. Nach Lemma 3.62 muss auch $\varrho = \varrho(\mathbf{A})$ gelten. ∎

Bemerkung 3.64 (Einfacher Eigenwert) *Es lässt sich beweisen, dass $\varrho(\mathbf{A})$ unter den im vorangehenden Satz gegebenen Bedingungen sogar ein* einfacher *Eigenwert der Matrix $\mathbf{A}$ ist. Der betreffende Beweis ist in [4, Beweis von I.f-g] zu finden.*

4 Die Jacobi-Iteration

In diesem Kapitel soll eines der einfachsten Iterationsverfahren zur Bestimmung der Schur-Zerlegung einer selbstadjungierte Matrix eingeführt werden: Die Jacobi-Iteration.

4.1 Iterierte Ähnlichkeitstransformationen

Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix. Unser Ziel ist die Bestimmung der Schur-Zerlegung

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D}$$

für unitäre Matrizen $\mathbf{Q}$ und eine Diagonalmatrix $\mathbf{D}$.

Wenn es möglich wäre, eine derartige Zerlegung in endlich vielen Rechenschritten für eine beliebige Dimension n zu berechnen, könnte man auch sämtliche Nullstellen von Polynomen beliebig hoher Ordnung mit endlich vielen Schritten berechnen. Da bekannt ist, dass diese Aufgabe nicht lösbar ist, dürfen wir auch nicht darauf hoffen, die Schur-Zerlegung mit endlich vielen Rechenoperationen bestimmen zu können.

Stattdessen müssen wir auf Näherungsverfahren zurückgreifen. Unser Ziel ist es, die Matrix $\mathbf{A}$ mit Hilfe einer unitären Ähnlichkeitstransformation auf Diagonalgestalt zu bringen. Die Diagonalgestalt ist durch $\operatorname{off}(\mathbf{D}) = 0$ charakterisiert, also ist die Idee naheliegend, nach Verfahren zu suchen, die diese Größe reduzieren: Wir beginnen mit $\mathbf{A}^{(0)} := \mathbf{A}$ und bestimmen zu jedem $m \in \mathbb{N}_0$ ein unitäres $\mathbf{Q}^{(m)}$, das $\operatorname{off}((\mathbf{Q}^{(m)})^*\mathbf{A}^{(m)}\mathbf{Q}^{(m)})$ verkleinert und setzen dann

$$\mathbf{A}^{(m+1)} := (\mathbf{Q}^{(m)})^*\mathbf{A}^{(m)}\mathbf{Q}^{(m)}.$$

Wir brechen ab, sobald $\operatorname{off}(\mathbf{A}^{(m+1)})$ klein genug ist und verwenden

$$\widetilde{\mathbf{Q}} := \mathbf{Q}^{(0)}\mathbf{Q}^{(1)}\dots\mathbf{Q}^{(m)}, \qquad \widetilde{\mathbf{D}} := \mathbf{A}^{(m+1)} = \widetilde{\mathbf{Q}}^*\mathbf{A}\widetilde{\mathbf{Q}}$$

als Approximationen von $\mathbf{Q}$ und $\mathbf{D}$.

Da jeder einzelne Schritt des Verfahrens eine unitäre Ähnlichkeitstransformation ist, muss $\widetilde{\mathbf{Q}}$ ebenfalls unitär sein, und an der Größe $\operatorname{off}(\widetilde{\mathbf{D}})$ lässt sich direkt ablesen, wie nahe wir einer durch $\operatorname{off}(\mathbf{D}) = 0$ charakterisierten exakten Schur-Zerlegung gekommen sind.

Ein Schritt des Verfahrens, also die Berechnung von $\mathbf{A}^{(m+1)} = (\mathbf{Q}^{(m)})^*\mathbf{A}^{(m)}\mathbf{Q}^{(m)}$, kann in zwei Teilschritte zerlegt werden:

$$\mathbf{B}^{(m)} := (\mathbf{Q}^{(m)})^*\mathbf{A}^{(m)}, \qquad \mathbf{A}^{(m+1)} := \mathbf{B}^{(m)}\mathbf{Q}^{(m)} = ((\mathbf{Q}^{(m)})^*(\mathbf{B}^{(m)})^*)^*.$$

Der erste Schritt entspricht der Anwendung von $(\mathbf{Q}^{(m)})^*$ auf jede Spalte von $\mathbf{A}^{(m)}$, der zweite der Anwendung derselben Matrix auf jede Zeile von $\mathbf{B}^{(m)}$. Sofern sich also die Multiplikation mit $(\mathbf{Q}^{(m)})^*$ effizient gestalten lässt, kann auch die Iteration effizient durchgeführt werden.

4.2 Zweidimensionaler Fall

Wie bei vielen anderen numerischen Verfahren empfiehlt es sich, die Strategie zur Lösung eines großen und komplizierten Problems auf eine Folge kleinerer und einfacherer Probleme zurückzuführen.

Wir untersuchen deshalb zunächst den Fall einer zweidimensionalen selbstadjungierten Matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix}$$

mit $a, d \in \mathbb{R}$ und $b \in \mathbb{K}$. Die Eigenwerte dieser Matrix können wir mit Hilfe des charakteristischen Polynoms einfach bestimmen: Für jeden Eigenwert $\lambda \in \sigma(\mathbf{A})$ muss

$$\begin{aligned} 0 = \det(\lambda\mathbf{I} - \mathbf{A}) &= \det \begin{pmatrix} \lambda - a & -b \\ -\bar{b} & \lambda - d \end{pmatrix} = (\lambda - a)(\lambda - d) - |b|^2 \\ &= \lambda^2 - (a+d)\lambda + ad - |b|^2 = \lambda^2 - (a+d)\lambda + \frac{(a+d)^2}{4} - \frac{(a+d)^2}{4} - \frac{4ad}{4} - |b|^2 \\ &= \left(\lambda - \frac{a+d}{2}\right)^2 - \frac{(a-d)^2 + 4|b|^2}{4} \end{aligned}$$

gelten, also folgt

$$\lambda = \frac{a+d}{2} + \sigma \frac{\sqrt{(a-d)^2 + 4|b|^2}}{2} \qquad \text{für ein } \sigma \in \{-1, 1\}.$$

Der Fall $b = 0$, also $|b| = 0$, interessiert uns nicht, denn in diesem Fall ist $\mathbf{A}$ bereits diagonal. Im Falle $b \neq 0$ können wir

$$\tau := \frac{d-a}{2b}$$

definieren und erhalten wegen $4|\tau|^2|b|^2 = (a-d)^2$ die Gleichung

$$\lambda = \frac{a+d}{2} + \sigma|b|\sqrt{|\tau|^2 + 1}.$$

Nun kennen wir die Eigenwerte, also müssen wir als nächstes die Eigenvektoren bestimmen. Da wir wissen, dass die Eigenvektoren nur bis auf ein skalares Vielfaches bestimmt sind, verwenden wir den Ansatz

$$\mathbf{x} = \begin{pmatrix} c \\ tc \end{pmatrix}$$

für $t, c \in \mathbb{K}$ mit $c \neq 0$. Der Eigenvektor $\mathbf{x}$ zu dem Eigenwert

$$\lambda = \frac{a+d}{2} + \sigma|b|\sqrt{|\tau|^2 + 1}$$

ergibt sich als Lösung der definierenden Gleichung

$$\mathbf{0} = (\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \begin{pmatrix} \frac{d-a}{2} + \sigma|b|\sqrt{|\tau|^2+1} & -b \\ -\bar{b} & \frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1} \end{pmatrix} \begin{pmatrix} c \\ tc \end{pmatrix}$$

Aus der ersten Zeile dieser Gleichung folgt

$$btc = \left(\frac{d-a}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right)c, \qquad t = \frac{d-a}{2b} + \sigma\frac{|b|}{b}\sqrt{|\tau|^2+1},$$

und diese Gleichung lässt sich kompakt als

$$t = \tau + \sigma\,\overline{\operatorname{sgn}(b)}\sqrt{|\tau|^2+1} \tag{4.1}$$

schreiben. Dank

$$\begin{aligned}\left(\frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right)t &= \left(\frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right)\left(\frac{d-a}{2b} + \sigma\frac{|b|}{b}\sqrt{|\tau|^2+1}\right)\\ &= b\left(-\frac{d-a}{2b} + \sigma\frac{|b|}{b}\sqrt{|\tau|^2+1}\right)\left(\frac{d-a}{2b} + \sigma\frac{|b|}{b}\sqrt{|\tau|^2+1}\right)\\ &= b\left(\frac{|b|^2}{b^2}(|\tau|^2+1) - \frac{(d-a)^2}{4b^2}\right)\\ &= \frac{1}{b}\left(|b|^2\frac{(d-a)^2}{4|b|^2} + |b|^2 - \frac{(d-a)^2}{4}\right) = \frac{|b|^2}{b} = \bar{b}\end{aligned} \tag{4.2}$$

erhalten wir in der zweiten Zeile

$$-\bar{b}c + \left(\frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right)tc = -\bar{b}c + \bar{b}c = 0,$$

also ist $\mathbf{x}$ für das in (4.1) definierte t tatsächlich ein Eigenvektor. Da wir an einem normierten Eigenvektor interessiert sind, also $\|\mathbf{x}\|_2 = 1$ gelten soll, müssen wir c so bestimmen, dass

$$1 = |c|^2 + |t|^2|c|^2 = |c|^2(1+|t|^2)$$

erfüllt ist, somit können wir c als

$$c = \frac{1}{\sqrt{1+|t|^2}}$$

wählen. Zur Abkürzung der Notation verwenden wir

$$s := tc, \qquad \mathbf{x} = \begin{pmatrix} c \\ s \end{pmatrix}.$$

Die Bestimmung des zweiten Eigenvektors gestaltet sich wesentlich einfacher: Da $\mathbf{A}$ selbstadjungiert ist, müssen Eigenvektoren zu verschiedenen Eigenwerten senkrecht aufeinander stehen, und im zweidimensionalen Raum ist es leicht, einen auf $\mathbf{x}$ senkrecht stehenden Vektor zu berechnen. Eine naheliegende Wahl ist

$$\mathbf{x}^\perp = \begin{pmatrix} -\bar{s} \\ \bar{c} \end{pmatrix},$$

und da dieser Vektor ebenfalls normiert ist, muss die gesuchte orthogonale Transformation $\mathbf{Q}$ durch

$$\mathbf{Q} = \begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix}$$

gegeben sein. Da $|c|^2 + |s|^2 = 1$ gilt, können wir diese Matrix als Rotation um die Null interpretieren, wobei der Rotationswinkel φ durch $c = \cos(\varphi)$ und, zumindest im Fall $\mathbb{K} = \mathbb{R}$, durch $s = \sin(\varphi)$ gegeben ist. Dann folgt $t = s/c = \tan(\varphi)$.

Lemma 4.1 (Zweidimensionale Schur-Zerlegung) *Für eine beliebige selbstadjungierte Matrix*

$$\mathbf{A} = \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix}$$

mit $a, d \in \mathbb{R}$ und $b \in \mathbb{K} \setminus \{0\}$ setzen wir

$$\tau := \frac{d-a}{2b},$$

wählen ein Vorzeichen $\sigma \in \{-1, 1\}$ und verwenden

$$t := \tau + \sigma\,\overline{\operatorname{sgn}(b)}\sqrt{|\tau|^2+1}, \qquad c := \frac{1}{\sqrt{1+|t|^2}}, \qquad s := tc \tag{4.3}$$

zur Definition von

$$\mathbf{Q} := \begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix}. \tag{4.4}$$

Dann gilt

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \begin{pmatrix} \frac{a+d}{2} + \sigma|b|\sqrt{|\tau|^2+1} & 0 \\ 0 & \frac{a+d}{2} - \sigma|b|\sqrt{|\tau|^2+1} \end{pmatrix},$$

wir haben also eine explizite Formel für die Schur-Zerlegung gefunden.

Beweis. Die Eigenwerte kürzen wir mit

$$\lambda_1 := \frac{a+d}{2} + \sigma|b|\sqrt{|\tau|^2+1}, \qquad \lambda_2 := \frac{a+d}{2} - \sigma|b|\sqrt{|\tau|^2+1} \tag{4.5}$$

ab und berechnen zunächst

$$\mathbf{A}\mathbf{Q} = \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix}\begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix} = \begin{pmatrix} ac+bs & -a\bar{s}+b\bar{c} \\ \bar{b}c+ds & -\bar{b}\bar{s}+d\bar{c} \end{pmatrix} = \begin{pmatrix} (a+bt)c & (b-a\bar{t})\bar{c} \\ (\bar{b}+dt)c & (d-\bar{b}\bar{t})\bar{c} \end{pmatrix}.$$

Um diesen Ausdruck zu vereinfachen verwenden wir (4.2) und die Definitionen, um

$$\begin{aligned} a + bt &= a + \frac{d-a}{2} + \sigma|b|\sqrt{|\tau|^2+1} = \lambda_1, \\ \bar{b} + dt &= \left(\frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right)t + dt = \lambda_1 t, \end{aligned}$$

$$b - a\bar{t} = \overline{\left(\frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right)}\bar{t} - a\bar{t} = -\lambda_2\bar{t},$$
$$d - \bar{b}\bar{t} = d - \frac{d-a}{2} - \sigma|b|\sqrt{|\tau|^2+1} = \lambda_2$$

zu erhalten und so zu

$$\mathbf{AQ} = \begin{pmatrix} (a+bt)c & (b-a\bar{t})\bar{c} \\ (\bar{b}+dt)c & (d-\bar{b}\bar{t})\bar{c} \end{pmatrix} = \begin{pmatrix} \lambda_1 c & -\lambda_2\bar{s} \\ \lambda_1 s & \lambda_2\bar{c} \end{pmatrix}$$

zu erhalten. Multiplikation mit $\mathbf{Q}^*$ ergibt

$$\mathbf{Q}^*\mathbf{AQ} = \begin{pmatrix} \bar{c} & \bar{s} \\ -s & c \end{pmatrix}\begin{pmatrix} \lambda_1 c & -\lambda_2\bar{s} \\ \lambda_1 s & \lambda_2\bar{c} \end{pmatrix} = \begin{pmatrix} \lambda_1(|c|^2+|s|^2) & \lambda_2(-\bar{c}\bar{s}+\bar{s}\bar{c}) \\ \lambda_1(-sc+cs) & \lambda_2(|s|^2+|c|^2) \end{pmatrix} = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix},$$

so dass die gewünschte Aussage bewiesen ist. ∎

Falls $\mathbf{A}$ schon „fast“ diagonal ist, könnte es passieren, dass durch die Transformation $\mathbf{Q}$ die Diagonalelemente den Platz tauschen. Das würde zu unerwünschten Problemen bei der Untersuchung der Konvergenz des Verfahrens führen und sollte deshalb vermieden werden. Glücklicherweise lässt sich dieser Effekt vermeiden, indem das Vorzeichen σ geschickt gewählt wird.

Lemma 4.2 *Seien $\mathbf{A}, \mathbf{Q}, \tau, t, c$ und s wie in Lemma 4.1 gegeben. Dann gilt*

$$\|\mathbf{A} - \mathbf{Q}^*\mathbf{AQ}\|_F^2 = \frac{2}{c^2}|b|^2.$$

Beweis. Wir kürzen die Eigenwerte wie in (4.5) ab und erhalten

$$\begin{aligned}
|a-\lambda_1|^2 &= \left|a - \frac{a+d}{2} - \sigma|b|\sqrt{|\tau|^2+1}\right|^2 = \left|\frac{a-d}{2} - \sigma|b|\sqrt{|\tau|^2+1}\right|^2 \\
&= \frac{(a-d)^2}{4} + |b|^2(|\tau|^2+1) - 2\sigma|b|\frac{a-d}{2}\sqrt{|\tau|^2+1} \\
&= |b|^2\left(\frac{(a-d)^2}{4|b|^2} + (|\tau|^2+1) - 2\sigma\frac{a-d}{2|b|}\sqrt{|\tau|^2+1}\right) \\
&= |b|^2\left(\frac{(d-a)^2}{4|b|^2} + (|\tau|^2+1) + 2\sigma\frac{d-a}{2|b|}\sqrt{|\tau|^2+1}\right) \\
&= |b|^2\left(\frac{(d-a)^2}{4b\bar{b}} + \left(\sigma^2\frac{|b|^2}{b\bar{b}}\right)(|\tau|^2+1) + 2\frac{d-a}{2\bar{b}}\sigma\frac{|b|}{b}\sqrt{|\tau|^2+1}\right) \\
&= |b|^2\left(\frac{d-a}{2b} + \sigma\frac{|b|}{b}\sqrt{|\tau|^2+1}\right)\overline{\left(\frac{d-a}{2b} + \sigma\frac{|b|}{b}\sqrt{|\tau|^2+1}\right)} = |b|^2|t|^2.
\end{aligned}$$

Für den zweiten Diagonaleintrag ergibt sich

$$|d-\lambda_2|^2 = \left|d - \frac{a+d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right|^2 = \left|-\frac{a-d}{2} + \sigma|b|\sqrt{|\tau|^2+1}\right|^2 = |a-\lambda_1|^2$$

und da die beiden Außerdiagonaleinträge von $\mathbf{Q}^*\mathbf{AQ}$ verschwinden, erhalten wir

$$\|\mathbf{A} - \mathbf{Q}^*\mathbf{AQ}\|_F^2 = 2|b|^2|t|^2 + 2|b|^2 = 2|b|^2(|t|^2+1) = \frac{2|b|^2}{c^2},$$

also die gewünschte Gleichung. ∎

Wenn wir sicher stellen wollen, dass die Matrix so wenig wie möglich verändert wird, müssen wir also darauf achten, dass c möglichst groß ist, also $|t|$ möglichst klein, also sollte das Vorzeichen von $\sigma\overline{\operatorname{sgn}(b)}$ gerade dem Vorzeichen von τ entgegengesetzt sein, so dass wir

$$\begin{aligned}\sigma\overline{\operatorname{sgn}(b)} = \frac{\sigma}{\operatorname{sgn}(b)} &\stackrel{!}{=} -\operatorname{sgn}(\tau),\\ \sigma = -\operatorname{sgn}(b)\operatorname{sgn}(\tau) &= -\operatorname{sgn}(b)\operatorname{sgn}\left(\frac{d-a}{2}\right) = -\operatorname{sgn}(d-a) = \operatorname{sgn}(a-d)\end{aligned}$$

erhalten. Dann entsteht allerdings ein Problem: Bei der Berechnung von

$$t = \tau - \operatorname{sgn}(\tau)\sqrt{|\tau|^2+1}$$

kann der Algorithmus instabil werden, falls $|\tau|$ groß ist, denn dann gilt $\sqrt{|\tau|^2+1} \approx |\tau|$, also

$$\operatorname{sgn}(\tau)\sqrt{|\tau|^2+1} \approx \operatorname{sgn}(\tau)|\tau| \approx \tau,$$

so dass sich die beiden Summanden näherungsweise auslöschen.

Das Problem lässt sich lösen, indem wir t indirekt berechnen: Wir wählen das entgegengesetzte Vorzeichen $-\sigma$ und berechnen die zweite Nullstelle

$$\hat{t} := \tau + \operatorname{sgn}(\tau)\sqrt{|\tau|^2+1}.$$

Offenbar sind t und $\hat{t}$ Nullstellen des Polynoms

$$\begin{aligned}z \mapsto (z-t)(z-\hat{t}) = z^2 - (t+\hat{t})z + t\hat{t} &= z^2 - 2\tau z + \tau^2 - \operatorname{sgn}(\tau)^2(|\tau|^2+1)\\ &= z^2 - 2\tau z + \tau^2 - \operatorname{sgn}(\tau)^2|\tau|^2 - \operatorname{sgn}(\tau)^2\\ &= z^2 - 2\tau z - \operatorname{sgn}(\tau)^2,\end{aligned}$$

und mittels eines Koeffizientenvergleichs folgt

$$t\hat{t} = -\operatorname{sgn}(\tau)^2, \qquad t = -\frac{\operatorname{sgn}(\tau)^2}{\hat{t}} = -\frac{\operatorname{sgn}(\tau)}{|\tau| + \sqrt{|\tau|^2+1}}.$$

Diese Gleichung erlaubt es uns, die benötigte Größe t auch für große Werte von τ noch stabil zu berechnen.

4.3 Höherdimensionaler Fall

Sei $A \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix. Zwei Indizes $i, j \in \{1, \ldots, n\}$ mit $i < j$ legen einen Außerdiagonaleintrag a_{ij} in $\mathbf{A}$ fest. Gesucht ist eine orthonormale Ähnlichkeitstransformation $\mathbf{Q} \in \mathbb{K}^{n\times n}$, die a_{ij} zu 0 macht, die also für $\mathbf{B} := \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ die Gleichung $b_{ij} = 0$ erfüllt.

Wenn $n = 2$ gelten würde, ließe sich $\mathbf{Q}$ mit Lemma 4.1 bestimmen. Es stellt sich also die Frage, ob sich das n-dimensionale Problem auf das zweidimensionale reduzieren läßt. Dazu setzen wir

$$\widehat{\mathbf{A}} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix} \tag{4.6}$$

und wählen

$$\widehat{\mathbf{Q}} = \begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix}$$

gemäß Lemma 4.1. Wir definieren $\mathbf{Q} \in \mathbb{K}^{n\times n}$, indem wir $\widehat{\mathbf{Q}}$ auf den Unterraum $\mathcal{E}_{ij} := \operatorname{span}\{\mathbf{e}^{(i)}, \mathbf{e}^{(j)}\}$ anwenden und seinen Senkrechtraum $\mathcal{E}_{ij}^{\perp} := \{\mathbf{x} \in \mathbb{K}^n \ : \ x_i = x_j = 0\}$ unverändert lassen.

Definition 4.3 *Seien $i, j \in \{1, \ldots, n\}$ mit $i < j$, sei $\widehat{\mathbf{A}}$ gemäß (4.6) definiert. Aus Lemma 4.1 erhalten wir eine Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{2\times 2}$. Sei $\mathbf{P}_{i,j} := (\mathbf{e}_i \ \mathbf{e}_j) \in \mathbb{K}^{n\times 2}$. Die durch*

$$\mathbf{Q} := (\mathbf{I} - \mathbf{P}_{i,j}\mathbf{P}_{i,j}^*) + \mathbf{P}_{i,j}\widehat{\mathbf{Q}}\mathbf{P}_{i,j}^* \tag{4.7}$$

beziehungsweise in „Pünktchennotation“ durch

$$\mathbf{Q} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -\bar{s} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & \bar{c} & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}. \tag{4.8}$$

definierte Matrix $Q \in \mathbb{K}^{n\times n}$ ist eine orthonormale Matrix, nämlich eine Rotation, und wird als Givens-Rotation *zu der Matrix A und dem Indexpaar (i, j) bezeichnet.*

Seien $i, j \in \{1, \ldots, n\}$ mit $i < j$. Sei $\mathbf{Q}$ die Givens-Rotation zu der Matrix $\mathbf{A}$ und dem Indexpaar (i, j). Sei $\mathbf{B} := \mathbf{Q}^*\mathbf{A}\mathbf{Q}$. Dann gilt

$$\begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} = \mathbf{P}_{i,j}^*\mathbf{B}\mathbf{P}_{i,j} = \mathbf{P}_{i,j}^*\mathbf{Q}^*\mathbf{A}\mathbf{Q}\mathbf{P}_{i,j} = \widehat{\mathbf{Q}}^*\mathbf{P}_{i,j}^*\mathbf{A}\mathbf{P}_{i,j}\widehat{\mathbf{Q}} = \widehat{\mathbf{Q}}^*\widehat{\mathbf{A}}\widehat{\mathbf{Q}} = \widehat{\mathbf{B}}.$$

Da $\widehat{\mathbf{B}}$ diagonal ist, folgt $b_{ij} = b_{ji} = 0$, also lassen sich mit Hilfe von Givens-Rotationen gezielt einzelne Außerdiagonaleinträge durch orthonormale Ähnlichkeitstransformationen eliminieren.

Es bleibt die Frage, ob die Ähnlichkeitstransformation mit $\mathbf{Q}$ tatsächlich den Außerdiagonalanteil von $\mathbf{B}$ gegenüber $\mathbf{A}$ reduziert hat. Diese Frage beantwortet das folgende Lemma:

Lemma 4.4 *Sei $\mathbf{Q}$ wie in (4.7) für $i, j \in \{1, \dots, n\}$ mit $i < j$ definiert. Dann gilt für die Matrix $\mathbf{B} := \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ die Gleichung*

$$\operatorname{off}^2(\mathbf{B}) = \operatorname{off}^2(\mathbf{A}) - 2|a_{ij}|^2.$$

Beweis. Wir wenden Lemma 3.50 auf die Matrizen $\widehat{\mathbf{A}}$ und $\widehat{\mathbf{B}}$ an und erhalten

$$|a_{ii}|^2 + 2|a_{ij}|^2 + |a_{jj}|^2 = \|\widehat{\mathbf{A}}\|_F^2 = \|\widehat{\mathbf{Q}}^*\widehat{\mathbf{A}}\widehat{\mathbf{Q}}\|_F^2 = \|\widehat{\mathbf{B}}\|_F^2 = |b_{ii}|^2 + |b_{jj}|^2.$$

Die einzigen Diagonalelemente, in denen sich $\mathbf{B}$ von $\mathbf{A}$ unterscheidet, sind b_{ii} und b_{jj}, so dass sich für die Norm die Außerdiagonalelemente die Gleichung

$$\begin{aligned}
\operatorname{off}^2(\mathbf{B}) &= \|\mathbf{B}\|_F^2 - \sum_{k=1}^{n} |b_{kk}|^2 = \|\mathbf{Q}^*\mathbf{A}\mathbf{Q}\|_F^2 - \sum_{\substack{k=1 \\ k\neq i, k\neq j}}^{n} |b_{kk}|^2 - |b_{ii}|^2 - |b_{jj}|^2 \\
&= \|\mathbf{A}\|_F^2 - \sum_{\substack{k=1 \\ k\neq i, k\neq j}}^{n} |a_{kk}|^2 - |a_{ii}|^2 - |a_{jj}|^2 - 2|a_{ij}|^2 \\
&= \|\mathbf{A}\|_F^2 - \sum_{k=1}^{n} |a_{kk}|^2 - 2|a_{ij}|^2 = \operatorname{off}^2(\mathbf{A}) - 2|a_{ij}|^2
\end{aligned}$$

ergibt. Also bewirkt die Ähnlichkeitstransformation von $\mathbf{A}$ mit $\mathbf{Q}$ dieselbe Reduktion der Außerdiagonalelemente wie die von $\widehat{\mathbf{A}}$ mit $\widehat{\mathbf{Q}}$. ∎

4.4 Algorithmus

Seien $i, j \in \{1, \dots, n\}$ mit $i < j$ gegeben. Wir berechnen $c, s \in \mathbb{K}$ entsprechend Lemma 4.1 und sind daran interessiert, die in Definition 4.3 gegebene Transformation $\mathbf{Q}$ möglichst effizient anzuwenden. Zunächst betrachten wir dazu die Matrix $\mathbf{M} = \mathbf{A}\mathbf{Q}$, deren Einträge durch

$$\mathbf{M} = \begin{pmatrix}
a_{11} & \cdots & ca_{1i} + sa_{1j} & \cdots & -\bar{s}a_{1i} + \bar{c}a_{1j} & \cdots & a_{1n} \\
\vdots & & \vdots & & \vdots & & \vdots \\
a_{i1} & \cdots & ca_{ii} + sa_{ij} & \cdots & -\bar{s}a_{ii} + \bar{c}a_{ij} & \cdots & a_{in} \\
\vdots & & \vdots & & \vdots & & \vdots \\
a_{j1} & \cdots & ca_{ji} + sa_{jj} & \cdots & -\bar{s}a_{ji} + \bar{c}a_{jj} & \cdots & a_{jn} \\
\vdots & & \vdots & & \vdots & & \vdots \\
a_{n1} & \cdots & ca_{ni} + sa_{nj} & \cdots & -\bar{s}a_{ni} + \bar{c}a_{nj} & \cdots & a_{nn}
\end{pmatrix}$$

gegeben sind: Nur die i-te und j-te Spalte wurden verändert, ein Algorithmus könnte diesen Schritt also in $6n$ Operationen durchführen.

Die Matrix $\mathbf{B} = \mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{Q}^*\mathbf{M}$ ist entsprechend durch

$$\mathbf{B} = \begin{pmatrix} m_{11} & \cdots & m_{1i} & \cdots & m_{1j} & \cdots & m_{1n} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \bar{c}m_{i1} + \bar{s}m_{j1} & \cdots & \bar{c}m_{ii} + \bar{s}m_{ji} & \cdots & \bar{c}m_{ij} + \bar{s}m_{jj} & \cdots & \bar{c}m_{in} + \bar{s}m_{jn} \\ \vdots & & \vdots & & \vdots & & \vdots \\ -sm_{i1} + cm_{j1} & \cdots & -sm_{ii} + cm_{ji} & \cdots & -sm_{ij} + cm_{jj} & \cdots & -sm_{in} + cm_{jn} \\ \vdots & & \vdots & & \vdots & & \vdots \\ m_{n1} & \cdots & m_{ni} & \cdots & m_{nj} & \cdots & m_{nn} \end{pmatrix}$$

beschrieben: Hier wurden nur die i-te und die j-te Zeile verändert, also kann auch dieser Schritt in $6n$ Operationen durchgeführt werden. Wir erhalten den folgenden Algorithmus:

Algorithmus 4.5 (Jacobi-Iteration) *Der folgende Algorithmus überschreibt* $\mathbf{A}$ *mit einer zu* $\mathbf{A}$ *ähnlichen „bis auf* ϵ *diagonalen“ Matrix und speichert die korrespondierende Ähnlichkeitstransformation in der Matrix* $\widetilde{\mathbf{Q}}$*:*

$\widetilde{\mathbf{Q}} \leftarrow \mathbf{I}$;
$\alpha \leftarrow \operatorname{off}(\mathbf{A})^2$;
while $\alpha > \epsilon^2$ **do begin**
  Wähle $i, j \in \{1, \ldots, n\}$ *mit* $i < j$;
  $\tau \leftarrow \frac{a_{jj} - a_{ii}}{2a_{ij}}$; $t \leftarrow \frac{\operatorname{sgn}(\tau)}{|\tau| + \sqrt{|\tau|^2 + 1}}$;
  $c \leftarrow 1/\sqrt{1 + |t|^2}$; $s \leftarrow tc$;
  for $k = 1$ **to** n **do begin** { *Berechne* $\mathbf{A} \leftarrow \mathbf{A}\mathbf{Q}$ *und* $\widetilde{\mathbf{Q}} \leftarrow \widetilde{\mathbf{Q}}\mathbf{Q}$ }
    $h \leftarrow a_{ki}$; $a_{ki} \leftarrow hc + a_{kj}s$; $a_{kj} \leftarrow -h\bar{s} + a_{kj}\bar{c}$;
    $h \leftarrow \tilde{q}_{ki}$; $q_{ki} \leftarrow hc + \tilde{q}_{kj}s$; $q_{kj} \leftarrow -h\bar{s} + \tilde{q}_{kj}\bar{c}$
  end;
  for $k = 1$ **to** n **do begin** { *Berechne* $\mathbf{A} \leftarrow \mathbf{Q}^*\mathbf{A}$ }
    $h \leftarrow a_{ik}$; $a_{ik} \leftarrow \bar{c}h + \bar{s}a_{jk}$; $a_{jk} \leftarrow -sh + ca_{jk}$
  end;
  $\alpha \leftarrow \alpha - 2|a_{ij}|^2$
end

Falls wir in diesem Algorithmus das Paar $1 \le i < j \le n$ so wählen, dass $|a_{ij}|$ den maximalen Wert annimmt, können wir eine einfache Konvergenzaussage herleiten:

Satz 4.6 (Konvergenz) *Falls wir* $i, j \in \{1, \ldots, n\}$ *mit* $i < j$ *so wählen, dass*

$$|a_{k\ell}| \le |a_{ij}| \qquad \text{für alle } k, \ell \in \{1, \ldots, n\} \text{ mit } k < \ell \tag{4.9}$$

erfüllt ist, erfüllt die Matrix $\mathbf{B}$ *aus Lemma 4.4 die Abschätzung*

$$\operatorname{off}^2(\mathbf{B}) \le \left(1 - \frac{2}{n(n-1)}\right) \operatorname{off}^2(\mathbf{A}),$$

also konvergiert der Algorithmus mit mindestens linearer Geschwindigkeit.

Beweis. Seien $i, j \in \{1, \dots, n\}$ mit $i < j$ so gewählt, dass (4.9) gilt. Dann folgt

$$\operatorname{off}^2(\mathbf{A}) = \sum_{\substack{k,\ell=1 \\ k\neq\ell}}^{n} |a_{k\ell}|^2 = 2 \sum_{\substack{k,\ell=1 \\ k<\ell}}^{n} |a_{k\ell}|^2 \leq 2 \sum_{\substack{k,\ell=1 \\ k<\ell}}^{n} |a_{ij}|^2 = 2\frac{n(n-1)}{2}|a_{ij}|^2 = n(n-1)|a_{ij}|^2.$$

und damit nach Lemma 4.4

$$\operatorname{off}^2(\mathbf{B}) = \operatorname{off}^2(\mathbf{A}) - 2|a_{ij}|^2 \leq \operatorname{off}^2(\mathbf{A}) - \frac{2}{n(n-1)} \operatorname{off}^2(\mathbf{A}) = \left(1 - \frac{2}{n(n-1)}\right) \operatorname{off}^2(\mathbf{A}).$$

Das ist die gesuchte Abschätzung. ∎

Lemma 4.1 legt die Rotation $\mathbf{Q}$ nur bis auf das Vorzeichen $\sigma \in \{1, -1\}$ fest. Es stellt sich die Frage, ob eine geschickte Wahl des Vorzeichens Vorteile für das Verfahren bietet. Das folgende Lemma deutet ein mögliches Auswahlkriterium an:

Lemma 4.7 *Sei $\mathbf{Q} \in \mathbb{K}^{n\times n}$ wie in Gleichung (4.8) definiert, sei $\mathbf{B} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$. Dann gilt*

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = 4(1-c) \sum_{k\neq i,j} (|a_{ki}|^2 + |a_{kj}|^2) + \frac{2}{c^2}|a_{ij}|^2.$$

Beweis. Sei $\mathbf{M} := \mathbf{A}\mathbf{Q}$. Sei $k \in \{1, \dots, n\}$. Dann gilt

$$\begin{aligned}
|m_{ki} - a_{ki}|^2 &= (m_{ki} - a_{ki})(\bar{m}_{ki} - \bar{a}_{ki}) = |m_{ki}|^2 + |a_{ki}|^2 - m_{ki}\bar{a}_{ki} - a_{ki}\bar{m}_{ki} \\
&= |m_{ki}|^2 + |a_{ki}|^2 - 2\operatorname{Re}(m_{ki}\bar{a}_{ki}) \\
&= |m_{ki}|^2 + |a_{ki}|^2 - 2\operatorname{Re}((a_{ki}c + a_{kj}s)\bar{a}_{ki}), \\
|m_{kj} - a_{kj}|^2 &= |m_{kj}|^2 + |a_{kj}|^2 - 2\operatorname{Re}(a_{kj}\bar{m}_{kj}) \\
&= |m_{kj}|^2 + |a_{kj}|^2 - 2\operatorname{Re}(a_{kj}(-\bar{a}_{ki}s + \bar{a}_{kj}c))
\end{aligned}$$

und durch Addition beider Gleichungen erhalten wir

$$|m_{ki} - a_{ki}|^2 + |m_{kj} - a_{kj}|^2 = |m_{ki}|^2 + |m_{kj}|^2 + |a_{ki}|^2 + |a_{kj}|^2 - 2(|a_{ki}|^2 + |a_{kj}|^2)\operatorname{Re} c.$$

Da die Transformation mit $\widehat{\mathbf{Q}}$ unitär ist, muss $|m_{ki}|^2 + |m_{kj}|^2 = |a_{ki}|^2 + |a_{kj}|^2$ gelten und wir erhalten

$$|m_{ki} - a_{ki}|^2 + |m_{kj} - a_{kj}|^2 = 2(1 - \operatorname{Re} c)(|a_{ki}|^2 + |a_{kj}|^2).$$

Indem wir dieselbe Argumentation auf $\mathbf{B} = \mathbf{Q}^*\mathbf{M}$ anwenden, erhalten wir

$$|b_{ik} - m_{ik}|^2 + |b_{jk} - m_{jk}|^2 = 2(1 - \operatorname{Re} c)(|m_{ik}|^2 + |m_{jk}|^2).$$

Da die Spaltentransformation nur die Spalten i, j betrifft und die Zeilentransformation nur die Zeilen i, j verändert, folgt

$$|b_{ki} - a_{ki}|^2 + |b_{kj} - a_{kj}|^2 = 2(1 - \operatorname{Re} c)(|a_{ki}|^2 + |a_{kj}|^2) \qquad \text{für alle } k \notin \{i, j\},$$
$$|b_{ik} - a_{ik}|^2 + |b_{jk} - a_{jk}|^2 = 2(1 - \operatorname{Re} c)(|a_{ki}|^2 + |a_{kj}|^2) \qquad \text{für alle } k \notin \{i, j\},$$

wobei im letzten Schritt ausgenutzt wurde, dass $\mathbf{A}$ selbstadjungiert ist. Da c in unserem Fall immer reell und positiv ist, erhalten wir den ersten Summanden unserer Gleichung.

Es fehlt nur noch die Betrachtung der Elemente a_{ii}, a_{ij}, a_{ji} und a_{jj}. Dazu greifen wir auf Lemma 4.2 zurück und erhalten

$$|a_{ii} - b_{ii}|^2 + |a_{ij} - b_{ij}|^2 + |a_{ji} - b_{ji}|^2 + |a_{jj} - b_{jj}|^2 = \frac{2}{c^2}|a_{ij}|^2,$$

und der Beweis ist vollständig. ∎

Wir können Sprünge in der Folge der Iterierten vermeiden, indem wir dafür sorgen, dass c möglichst große Werte annimmt. Nach Konstruktion ist das gerade dann der Fall, wenn $|t|$ möglichst klein ist, wenn also das Vorzeichen von $a_{ii} - a_{jj}$ dem von $\sigma|a_{ij}|\sqrt{\tau^2 + 1}$ entspricht. Oder kürzer: Wir müssen σ negativ wählen, falls $\tau < 0$ gilt, und ansonsten positiv. Beispielsweise wird auf diese Weise verhindert, dass bei bereits kleinen Außerdiagonaleinträgen die Diagonaleinträge die Plätze tauschen und so zwar $\operatorname{off}(\mathbf{B})^2$ konvergiert, nicht aber die Folge der Matrizen selbst.

Bemerkung 4.8 (Quadratische Konvergenz) *Wählt man die Vorzeichen wie oben erläutert, so kann man unter einigen zusätzlichen Voraussetzungen nachweisen, dass das Jacobi-Verfahren lokal quadratisch konvergiert.*

Bemerkung 4.9 (Parallelisierung) *Falls zwei Indexpaare (i, j) und $(i', j)'$ die Bedingung $\{i, j\} \cap \{i', j'\} = \emptyset$ erfüllen, können die Berechnung der Zeilen zu beiden Rotationen parallel erfolgen: Die Rotation zu (i, j) betrifft nur die i-te und die j-te Zeile, während die Rotation zu (i', j') nur die i'-te und die j'-te Zeile betrifft. Entsprechend können wir auch mit den Spaltenrotationen verfahren.*

5 Die Vektoriteration

Die im letzten Kapitel vorgestellte Jacobi-Iteration berechnet im Falle einer symmetrischen Matrix sämtliche Eigenwerte und Eigenvektoren. Es gibt viele Fälle, in denen man lediglich an einen Teil des Spektrums und den zugehörigen Eigenvektoren interessiert ist, beispielsweise bei der Bestimmung von Eigenschwingungen oder eines invarianten Wahrscheinlichkeitsmaßes.

Zur Lösung derartiger Aufgaben werden sehr oft Verfahren auf der Basis der sogenannten *Vektoriteration* (engl. „power iteration“, weil Potenzen der Matrix eine entscheidende Rolle spielen) verwendet. Einigen dieser Methoden ist dieses Kapitel gewidmet.

5.1 Grundidee

Wir erinnern uns an das zweite Beispiel aus Kapitel 2, in dem die Aufgabe darin bestand, eine Wahrscheinlichkeitsverteilung $\mathbf{y} \in \mathbb{R}^4_{\geq 0} \setminus \{0\}$ zu finden, die stabil bleibt, sich also bei der „Durchführung eines Spielzugs“ nicht ändert. Ein derartiger Vektor ist durch die Gleichung

$$\mathbf{M}\mathbf{y} = \mathbf{y} \tag{5.1}$$

beschrieben. Ein Ansatz zur Bestimmung eines derartigen Vektors besteht darin, zu untersuchen, unter welchen Bedingungen die Folge $(\mathbf{M}^m\mathbf{z})_{m\in\mathbb{N}}$ für eine Startverteilung $\mathbf{z} \in \mathbb{R}^4_{\geq 0}$ konvergiert.

Falls wir annehmen, dass die Folge gegen einen Vektor $\mathbf{z}^* \in \mathbb{R}^4$ konvergiert, finden wir zu jedem $\epsilon \in \mathbb{R}_{>0}$ ein $n_0 \in \mathbb{N}$ mit

$$\|\mathbf{M}^m\mathbf{z} - \mathbf{z}^*\| < \epsilon \qquad \text{für alle } m \in \mathbb{N}_{\geq n_0},$$

so dass wir

$$\begin{aligned}\|\mathbf{M}\mathbf{z}^* - \mathbf{z}^*\| &= \|\mathbf{M}\mathbf{z}^* - \mathbf{M}^{m+1}\mathbf{z} + \mathbf{M}^{m+1}\mathbf{z} - \mathbf{z}^*\| \leq \|\mathbf{M}(\mathbf{z}^* - \mathbf{M}^m\mathbf{z})\| + \|\mathbf{M}^{m+1}\mathbf{z} - \mathbf{z}^*\| \\ &\leq \|\mathbf{M}\|\ \|\mathbf{z}^* - \mathbf{M}^m\mathbf{z}\| + \|\mathbf{M}^{m+1}\mathbf{z} - \mathbf{z}^*\| < (\|\mathbf{M}\| + 1)\epsilon\end{aligned}$$

für alle $m \in \mathbb{N}_{\geq n_0}$ erhalten. Da diese Abschätzung für beliebige $\epsilon \in \mathbb{R}_{>0}$ gilt, folgt

$$\mathbf{M}\mathbf{z}^* = \mathbf{z}^*,$$

also ist der Grenzwert der Folge, sofern er existiert, auch Lösung der Gleichung (5.1).

Unsere Aufgabe besteht also darin, für eine Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ und einen Startvektor $\mathbf{z}^{(0)}$ die Folge $(\mathbf{A}^m\mathbf{z}^{(0)})_{m\in\mathbb{N}_0}$ zu untersuchen.

Dazu untersuchen wir zunächst eine Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},$$

mit Eigenwerten $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$. Ohne Beschränkung der Allgemeinheit können wir annehmen, dass die Eigenwerte nach absteigendem Betrag sortiert sind, dass also

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n| \tag{5.2}$$

gilt. Wenn wir mit einem Vektor $\widehat{\mathbf{x}}^{(0)} \in \mathbb{K}^n$ anfangen, erhalten wir

$$\widehat{\mathbf{x}}^{(m)} = \mathbf{D}^m \widehat{\mathbf{x}}^{(0)} = \begin{pmatrix} \lambda_1^m \hat{x}_1^{(0)} \\ \vdots \\ \lambda_n^m \hat{x}_n^{(0)} \end{pmatrix} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Wir können sehen, dass die erste Komponente dieses Vektors schneller als alle anderen wachsen wird, falls $|\lambda_1| > |\lambda_2|$ und $\hat{x}_1^{(0)} \neq 0$ gelten.

Falls allerdings $\lambda_1 \neq 1$ gilt, dürfen wir nicht erwarten, dass die Folge $(\widehat{\mathbf{x}}^{(m)})_{m=0}^{\infty}$ im konventionellen Sinn konvergiert, da asymptotisch $\widehat{\mathbf{x}}^{(m+1)} \approx \lambda_1 \widehat{\mathbf{x}}^{(m)}$ gelten wird. Es ist deshalb hilfreich, lediglich die von den Vektoren aufgespannten Räume miteinander zu vergleichen statt die Vektoren selbst. Dafür ist der *Winkel* zwischen Vektoren nützlich.

Definition 5.1 (Winkel) *Seien* $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. *Der* Winkel *zwischen den Vektoren ist definiert durch*

$$\angle(\mathbf{x}, \mathbf{y}) = \arccos \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{\|\mathbf{x}\| \, \|\mathbf{y}\|}.$$

Diese Definition hat den Vorteil, dass eine Skalierung der Vektoren $\mathbf{x}$ und $\mathbf{y}$ mit beliebigen von null verschiedenen Faktoren den Winkel nicht ändert, so dass wir auf Konvergenz hoffen dürfen.

In der Praxis ist es häufig handlicher, mit von dem Winkel abgeleiteten trigonometrischen Funktionen zu arbeiten, die durch

$$\begin{aligned} \cos \angle(\mathbf{x}, \mathbf{y}) &= \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \\ \sin \angle(\mathbf{x}, \mathbf{y}) &:= \sqrt{1 - \cos^2 \angle(\mathbf{x}, \mathbf{y})}, \\ \tan \angle(\mathbf{x}, \mathbf{y}) &:= \frac{\sin \angle(\mathbf{x}, \mathbf{y})}{\cos \angle(\mathbf{x}, \mathbf{y})} \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \end{aligned}$$

gegeben sind. Ein Blick auf die Taylor-Entwicklung zeigt, dass für kleine Winkel sowohl der Sinus als auch der Tangens ungefähr gleich dem Winkel sind, so dass Konvergenzaussagen über diese Funktionen auch Aussagen über den Winkel zulassen.

Lemma 5.2 (Konvergenz für Diagonalmatrizen) *Sei $\widehat{\mathbf{x}}^{(0)} \in \mathbb{K}^n$ mit $\hat{x}_1^{(0)} \neq 0$ gegeben. Dann gilt*

$$\tan \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) \leq \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^m \tan \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(0)}) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweis. Da $\delta^{(1)} \in \mathbb{K}^n$ der erste kanonische Einheitsvektor ist, gelten

$$\begin{aligned}
\cos^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) &= \frac{|\hat{x}_1^{(m)}|^2}{\|\widehat{\mathbf{x}}^{(m)}\|^2} = \frac{|\hat{x}_1^{(m)}|^2}{\sum_{i=1}^n |\hat{x}_i^{(m)}|^2}, \\
\sin^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) &= 1 - \cos^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) = \frac{\sum_{i=1}^n |\hat{x}_i^{(m)}|^2 - |\hat{x}_1^{(m)}|^2}{\sum_{i=1}^n |\hat{x}_i^{(m)}|^2} = \frac{\sum_{i=2}^n |\hat{x}_i^{(m)}|^2}{\sum_{i=1}^n |\hat{x}_i^{(m)}|^2}, \\
\tan^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) &= \frac{\sin^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)})}{\cos^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)})} = \frac{\sum_{i=2}^n |\hat{x}_i^{(m)}|^2}{|\hat{x}_1^{(m)}|^2} \qquad \text{für alle } m \in \mathbb{N}_0.
\end{aligned}$$

Sei $m \in \mathbb{N}_0$. Mit (5.2) folgt

$$\begin{aligned}
\tan^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) &= \frac{\sum_{i=2}^n |\hat{x}_i^{(m)}|^2}{|\hat{x}_1^{(m)}|^2} = \frac{\sum_{i=2}^n |\lambda_i^{2m}|\,|\hat{x}_i^{(0)}|^2}{|\lambda_1^{2m}|\,|\hat{x}_1^{(0)}|^2} \leq \frac{\sum_{i=2}^n |\lambda_2^{2m}|\,|\hat{x}_i^{(0)}|^2}{|\lambda_1^{2m}|\,|\hat{x}_1^{(0)}|^2} \\
&= \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{2m} \frac{\sum_{i=2}^n |\hat{x}_i^{(0)}|^2}{|\hat{x}_1^{(0)}|^2} = \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{2m} \tan^2 \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}).
\end{aligned}$$

Die Behauptung folgt, indem wir auf beiden Seiten die Wurzel ziehen. ∎

Auf Konvergenz des Tangens, und damit des Winkels, gegen null dürfen wir demnach hoffen, falls $|\lambda_1| > |\lambda_2|$ gilt. Gemäß unserer Definition bedeutet das gerade, dass ein Eigenwert im Betrag echt größer ist als alle anderen.

Definition 5.3 (Dominanter Eigenwert) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Ein Eigenwert $\lambda_1 \in \sigma(\mathbf{A})$ heißt* dominant*, falls*

$$|\lambda_1| > |\lambda| \qquad \text{für alle } \lambda \in \sigma(\mathbf{A}) \setminus \{\lambda_1\}$$

gilt, falls der Betrag von λ_1 also echt größer als die Beträge aller anderen Eigenwerte der Matrix $\mathbf{A}$ ist (Es sei daran erinnert, dass das Spektrum grundsätzlich über dem Körper $\mathbb{C}$ der komplexen Zahlen definiert wird, so dass $\sigma(\mathbf{A})$ nicht leer sein kann).

Für die Praxis ist ein Verfahren, das sich nur auf Diagonalmatrizen anwenden lässt, natürlich uninteressant. Als erste Verallgemeinerung untersuchen wir eine normale Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ (vgl. Definition 3.44). Nach Folgerung 3.47 existieren eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ mit

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D}.$$

Nach Bemerkung 3.40 können wir dafür sorgen, dass

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| \tag{5.3}$$

gelten. Im Folgenden gehen wir davon aus, dass die Matrix $\mathbf{D}$ diese Eigenschaft besitzt.

Den durch die Gleichung

$$\mathbf{x}^{(m)} = \mathbf{A}^m \mathbf{x}^{(0)} \qquad \text{für alle } m \in \mathbb{N}_0$$

definierten Iterationsvektoren können wir nun transformierte Vektoren

$$\widehat{\mathbf{x}}^{(m)} := \mathbf{Q}^* \mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

zuordnen. Dann gilt

$$\begin{aligned} \widehat{\mathbf{x}}^{(m)} &= \mathbf{Q}^* \mathbf{x}^{(m)} = \mathbf{Q}^* \mathbf{A}^m \mathbf{x}^{(0)} = \mathbf{Q}^* \mathbf{A}^m \mathbf{Q} \widehat{\mathbf{x}}^{(0)} \\ &= (\mathbf{Q}^* \mathbf{A} \mathbf{Q})^* \widehat{\mathbf{x}}^{(0)} = \mathbf{D}^m \widehat{\mathbf{x}}^{(0)} \qquad \text{für alle } m \in \mathbb{N}_0. \end{aligned}$$

Damit können wir Lemma 5.2 auf die Vektoren $(\widehat{\mathbf{x}}^{(m)})_{m=0}^{\infty}$ anwenden und müssen lediglich untersuchen, wie sich die Winkel unter unitären Transformationen verändern.

Satz 5.4 (Konvergenz für normale Matrizen) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine normale Matrix. Sie besitzt eine Schur-Zerlegung $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ mit einer unitären Matrix $\mathbf{Q} \in \mathbb{C}^{n \times n}$ und einer Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n \times n}$ der Form (5.3).*

Dann ist $\mathbf{e} := \mathbf{Q}\delta^{(1)}$ ein Eigenvektor zu dem betragsgrößten Eigenwert λ_1.

Sei $\mathbf{x}^{(0)} \in \mathbb{K}^n$ mit $\langle \mathbf{e}, \mathbf{x}^{(0)} \rangle_2 \neq 0$ gegeben. Dann gilt

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(m)}) \leq \left(\frac{|\lambda_2|}{|\lambda_1|} \right)^m \tan \angle(\mathbf{e}, \mathbf{x}^{(0)}) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Mit Lemma 3.33 gilt

$$\begin{aligned} \cos \angle(\mathbf{e}, \mathbf{x}^{(m)}) &= \frac{|\langle \mathbf{e}, \mathbf{x}^{(m)} \rangle|}{\|\mathbf{e}\| \, \|\mathbf{x}^{(m)}\|} = \frac{|\langle \mathbf{Q}\delta^{(1)}, \mathbf{Q}\widehat{\mathbf{x}}^{(m)} \rangle|}{\|\mathbf{Q}\delta^{(1)}\| \, \|\mathbf{Q}\widehat{\mathbf{x}}^{(m)}\|} = \frac{|\langle \delta^{(1)}, \mathbf{Q}^*\mathbf{Q}\widehat{\mathbf{x}}^{(m)} \rangle|}{\|\delta^{(1)}\| \, \|\widehat{\mathbf{x}}^{(m)}\|} \\ &= \frac{|\langle \delta^{(1)}, \widehat{\mathbf{x}}^{(m)} \rangle|}{\|\delta^{(1)}\| \, \|\widehat{\mathbf{x}}^{(m)}\|} = \cos \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0. \end{aligned}$$

Nach Definition folgt daraus unmittelbar

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(m)}) = \tan \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Nun können wir Lemma 5.2 anwenden und erhalten

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(m)}) = \tan \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}) \leq \left(\frac{|\lambda_2|}{|\lambda_1|} \right)^m \tan \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(0)})$$

$$= \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^m \tan\angle(\mathbf{e}, \mathbf{x}^{(0)}) \qquad \text{für alle } m \in \mathbb{N}_0.$$

∎

Falls wir Aussagen für Matrizen erhalten wollen, die nicht normal sind, können wir nicht länger auf unitäre Ähnlichkeitstransformationen zurückgreifen. In dieser Situation kann es nützlich sein, eine alternative Charakterisierung des Winkels zu verwenden.

Lemma 5.5 (Sinus als Minimum) *Es gilt*

$$\sin\angle(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{\|\mathbf{x} - \alpha\mathbf{y}\|}{\|\mathbf{x}\|} \;:\; \alpha \in \mathbb{K}\right\} \qquad \textit{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}.$$

Beweis. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ gegeben. Wir setzen $\beta := \langle \mathbf{y}, \mathbf{x}\rangle / \|\mathbf{y}\|^2$. Sei $\alpha \in \mathbb{K}$. Es gilt

$$\begin{aligned}\|\mathbf{x} - \alpha\mathbf{y}\|^2 &= \|\mathbf{x} - \beta\mathbf{y} + (\beta - \alpha)\mathbf{y}\|^2 \\ &= \langle \mathbf{x} - \beta\mathbf{y} + (\beta - \alpha)\mathbf{y}, \mathbf{x} - \beta\mathbf{y} + (\beta - \alpha)\mathbf{y}\rangle \\ &= \|\mathbf{x} - \beta\mathbf{y}\|^2 - \overline{(\beta - \alpha)}\langle \mathbf{y}, \mathbf{x} - \beta\mathbf{y}\rangle - (\beta - \alpha)\langle \mathbf{x} - \beta\mathbf{y}, \mathbf{y}\rangle + |\beta - \alpha|^2\|\mathbf{y}\|^2.\end{aligned}$$

Wir haben

$$\begin{aligned}\langle \mathbf{x} - \beta\mathbf{y}, \mathbf{y}\rangle &= \langle \mathbf{x}, \mathbf{y}\rangle - \bar{\beta}\|\mathbf{y}\|^2 = \langle \mathbf{x}, \mathbf{y}\rangle - \overline{\langle \mathbf{y}, \mathbf{x}\rangle} = 0, \\ \langle \mathbf{y}, \mathbf{x} - \beta\mathbf{y}\rangle &= \overline{\langle \mathbf{x} - \beta\mathbf{y}, \mathbf{y}\rangle} = 0,\end{aligned}$$

so dass wir

$$\|\mathbf{x} - \alpha\mathbf{y}\|^2 = \|\mathbf{x} - \beta\mathbf{y}\|^2 + |\beta - \alpha|^2\|\mathbf{y}\|^2$$

erhalten. Offenbar nimmt dieser Ausdruck sein Minimum für $\alpha = \beta$ an, und dieses Minimum ist gerade

$$\begin{aligned}\|\mathbf{x} - \beta\mathbf{y}\|^2 &= \|\mathbf{x}\|^2 - \bar{\beta}\langle \mathbf{y}, \mathbf{x}\rangle - \beta\langle \mathbf{x}, \mathbf{y}\rangle + |\beta|^2\|\mathbf{y}\|^2 \\ &= \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y}\rangle|^2}{\|\mathbf{y}\|^2} - \frac{|\langle \mathbf{x}, \mathbf{y}\rangle|^2}{\|\mathbf{y}\|^2} + \frac{|\langle \mathbf{x}, \mathbf{y}\rangle|^2}{\|\mathbf{y}\|^4}\|\mathbf{y}\|^2 = \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y}\rangle|^2}{\|\mathbf{y}\|^2}.\end{aligned}$$

Es folgt

$$\sin^2\angle(\mathbf{x}, \mathbf{y}) = 1 - \cos^2\angle(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x}\|^2\|\mathbf{y}\|^2 - |\langle \mathbf{x}, \mathbf{y}\rangle|^2}{\|\mathbf{x}\|^2\|\mathbf{y}\|^2} = \frac{\|\mathbf{x} - \beta\mathbf{y}\|^2}{\|\mathbf{x}\|^2},$$

und damit die Behauptung. ∎

Diese alternative Charakterisierung des Winkels ist nützlich, weil sie es uns erlaubt, auch allgemeine Transformationen in Betracht zu ziehen.

Lemma 5.6 (Transformierter Sinus) *Sei $\mathbf{B} \in \mathbb{K}^{n\times n}$ eine invertierbare Matrix. Es gilt*

$$\sin\angle(\mathbf{Bx}, \mathbf{By}) \leq \|\mathbf{B}\|\,\|\mathbf{B}^{-1}\|\sin\angle(\mathbf{x}, \mathbf{y}) \qquad \textit{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}.$$

Beweis. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Mit Lemma 5.5 finden wir ein $\alpha \in \mathbb{K}$ so, dass

$$\sin \angle(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x} - \alpha\mathbf{y}\|}{\|\mathbf{x}\|}$$

gilt. Mit (3.6a) folgen

$$\begin{aligned} \|\mathbf{B}\mathbf{x} - \alpha\mathbf{B}\mathbf{y}\| &= \|\mathbf{B}(\mathbf{x} - \alpha\mathbf{y})\| \leq \|\mathbf{B}\| \, \|\mathbf{x} - \alpha\mathbf{y}\|, \\ \|\mathbf{B}^{-1}\| \, \|\mathbf{B}\mathbf{x}\| &\geq \|\mathbf{B}^{-1}\mathbf{B}\mathbf{x}\| = \|\mathbf{x}\|, \end{aligned}$$

und wir erhalten schließlich

$$\sin \angle(\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{y}) \leq \frac{\|\mathbf{B}\mathbf{x} - \alpha\mathbf{B}\mathbf{y}\|}{\|\mathbf{B}\mathbf{x}\|} \leq \|\mathbf{B}\| \, \|\mathbf{B}^{-1}\| \frac{\|\mathbf{x} - \alpha\mathbf{y}\|}{\|\mathbf{x}\|} = \|\mathbf{B}\| \, \|\mathbf{B}^{-1}\| \sin \angle(\mathbf{x}, \mathbf{y}).$$

■

Wenden wir uns also dem allgemeinen Fall zu: Für einen Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^n$ und die allgemeine Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ definieren wir die Folge der Iterierten durch

$$\mathbf{x}^{(m)} := \mathbf{A}^m \mathbf{x}^{(0)} = \mathbf{A}\mathbf{x}^{(m-1)} \qquad \text{für alle } m \in \mathbb{N}. \tag{5.4}$$

Wir setzen voraus, dass $\mathbf{A}$ diagonalisierbar und die Eigenwerte nach ihrem Betrag absteigend sortiert sind, dass also eine reguläre Matrix $\mathbf{T} \in \mathbb{K}^{n \times n}$ mit

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

und der bereits aus (5.2) bekannten Anordnung

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$$

existiert. Mit Hilfe des Lemmas 5.6 können wir den Satz 5.4 wie folgt verallgemeinern:

Folgerung 5.7 (Konvergenz) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ diagonalisierbar mit $\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1}$ und $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ sowie (5.2).*

Dann ist $\mathbf{e} := \mathbf{T}\delta^{(1)}$ ein Eigenvektor zu dem betragsgrößten Eigenwert λ_1 und es gilt

$$\sin \angle(\mathbf{e}, \mathbf{x}^{(m)}) \leq \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^m \|\mathbf{T}\| \, \|\mathbf{T}^{-1}\| \tan \angle(\delta^{(1)}, \mathbf{T}^{-1}\mathbf{x}^{(0)}) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir definieren

$$\widehat{\mathbf{x}}^{(m)} := \mathbf{T}^{-1}\mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

und stellen fest, dass

$$\widehat{\mathbf{x}}^{(m)} = \mathbf{T}^{-1}\mathbf{x}^{(m)} = \mathbf{T}^{-1}\mathbf{A}^m\mathbf{x}^{(0)} = \mathbf{T}^{-1}\mathbf{A}^m\mathbf{T}\widehat{\mathbf{x}}^{(0)} = (\mathbf{T}^{-1}\mathbf{A}\mathbf{T})^m\widehat{\mathbf{x}}^{(0)} = \mathbf{D}^m\widehat{\mathbf{x}}^{(0)}$$

für alle $m \in \mathbb{N}_0$ gilt. Mit Lemma 5.6 erhalten wir

$$\sin \angle(\mathbf{e}, \mathbf{x}^{(m)}) = \sin \angle(\mathbf{T}\delta^{(1)}, \mathbf{T}\widehat{\mathbf{x}}^{(m)}) \leq \|\mathbf{T}\| \, \|\mathbf{T}^{-1}\| \sin \angle(\delta^{(1)}, \widehat{\mathbf{x}}^{(m)}).$$

Den Sinus können wir wie bisher mit Lemma 5.2 abschätzen, um die gewünschte Aussage zu erhalten. ∎

In der Praxis kann die Bestimmung der Iterierten $\mathbf{x}^{(m)}$ zu Schwierigkeiten führen, weil die Komponenten der Vektoren durch Maschinenzahlen dargestellt werden, die nicht beliebig groß oder klein werden können: Der Vektor wird asymptotisch in jedem Schritt ungefähr mit dem Faktor λ_1 multipliziert werden. Falls $|\lambda_1| > 1$ gilt, wird er also exponentiell wachsen, bis die Menge der Maschinenzahlen ausgeschöpft ist. Im IEEE-754-Standard würden dann die „übergelaufenen“ Koeffizienten gleich unendlich gesetzt werden und wären für unsere Zwecke unbrauchbar. Falls $|\lambda_1| < 1$ gilt, werden die Vektoren exponentiell schrumpfen, bis sie so nahe an der Null sind, dass sie zu null abgerundet werden und damit ebenfalls unbrauchbar werden.

Um zu verhindern, dass die Koeffizienten der Iterationsvektoren die Menge der Maschinenzahlen verlassen, empfiehlt es sich, eine Normierung einzuführen, also die Vektoren im Zuge des Verfahrens so zu skalieren, dass Über- und Unterläufe ausgeschlossen werden.

Da gemäß Lemma 5.5 auch

$$\begin{aligned}\sin \angle(\beta\mathbf{x}, \mathbf{y}) &= \min\left\{ \frac{\|\beta\mathbf{x} - \gamma\mathbf{y}\|}{\|\beta\mathbf{x}\|} \;:\; \gamma \in \mathbb{K} \right\} = \min\left\{ \frac{\|\beta\mathbf{x} - \beta\gamma'\mathbf{y}\|}{\|\beta\mathbf{x}\|} \;:\; \gamma' \in \mathbb{K} \right\} \\ &= \min\left\{ \frac{\|\mathbf{x} - \gamma'\mathbf{y}\|}{\|\mathbf{x}\|} \;:\; \gamma' \in \mathbb{K} \right\} = \sin \angle(\mathbf{x}, \mathbf{y})\end{aligned}$$

für alle $\beta \in \mathbb{K}\backslash\{0\}$ gilt, beeinflusst eine beliebige Skalierung die Konvergenz der Vektoren nicht im Geringsten. Häufig wählt man die Skalierung so, dass die Iterierten Einheitsvektoren bezüglich einer geeigneten Norm sind. Der korrespondierende Algorithmus nimmt dann die folgende Form an:

Algorithmus 5.8 (Vektoriteration) *Seien $\mathbf{A} \in \mathbb{K}^{n\times n}$ und $\mathbf{x}^{(0)} \in \mathbb{K}^n$ wie in Folgerung 5.7 gegeben und gelte $\gamma \neq 0$. Dann berechnet der folgende Algorithmus die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}}$, die gegen ein Vielfaches des ersten Eigenvektors $\mathbf{e}^{(1)} = \mathbf{T}\delta^{(1)}$ konvergiert:*

$m \leftarrow 0$
$\mathbf{x}^{(m)} \leftarrow \mathbf{x}^{(m)}/\|\mathbf{x}^{(m)}\|$
while „Fehler zu groß“ do begin
  $\mathbf{w}^{(m+1)} \leftarrow \mathbf{A}\mathbf{x}^{(m)}$
  $\mathbf{x}^{(m+1)} \leftarrow \mathbf{w}^{(m+1)}/\|\mathbf{w}^{(m+1)}\|$
  $m \leftarrow m + 1$
end

Selbstverständlich können wir den Algorithmus nicht unendlich lange arbeiten lassen, sondern wir müssen den Fehler der Approximation des Eigenvektors einschätzen und die

Iteration abbrechen, sobald der Fehler klein genug geworden ist. Da uns im Allgemeinen weder der Eigenvektor $\mathbf{e}$ noch der Eigenwert λ_1 zur Verfügung stehen, müssen wir beides auf Grundlage verfügbarer Daten approximieren.

Definition 5.9 (Rayleigh-Quotient) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine Matrix. Die Funktion*

$$\Lambda_A : \mathbb{K}^n \setminus \{\mathbf{0}\} \to \mathbb{K}, \qquad \mathbf{y} \mapsto \frac{\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2}{\langle \mathbf{y}, \mathbf{y}\rangle_2},$$

nennen wir die Rayleigh-Quotienten-Funktion *zu $\mathbf{A}$, für einen Vektor $\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ nennen wir den Funktionswert $\Lambda_A(\mathbf{y})$ den* Rayleigh-Quotienten *zu $\mathbf{A}$ und $\mathbf{y}$.*

Der Rayleigh-Quotient leistet gute Dienste dabei, einem Eigenvektor einen Eigenwert zuzuordnen:

Lemma 5.10 (Eigenwert) *Sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein Eigenvektor der Matrix $\mathbf{A}$ zum Eigenwert $\lambda \in \mathbb{K}$. Dann gilt $\Lambda_A(\mathbf{x}) = \lambda$.*

Beweis. Da das Skalarprodukt sesquilinear ist, ist es insbesondere linear im zweiten Argument, und wir erhalten

$$\Lambda_A(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \frac{\langle \mathbf{x}, \lambda\mathbf{x}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \lambda\frac{\langle \mathbf{x}, \mathbf{x}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \lambda.$$

Das ist die gesuchte Gleichung. ∎

Eine einfache Strategie zur Schätzung des Fehlers der Vektoriteration besteht nun darin, in jedem Schritt des Verfahrens den Rayleigh-Quotienten $\lambda_m := \Lambda_A(\mathbf{z}^{(m)})$ der aktuellen Iterierten $\mathbf{z}^{(m)}$ zu bestimmen und den Fehler gegenüber der exakten Eigenwertgleichung $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ zu bestimmen, also

$$\epsilon_m := \|\mathbf{A}\mathbf{z}^{(m)} - \lambda_m\mathbf{z}^{(m)}\| = \|\mathbf{w}^{(m+1)} - \langle \mathbf{z}^{(m)}, \mathbf{w}^{(m+1)}\rangle\mathbf{z}^{(m)}\| \tag{5.5}$$

als Maß des Fehlers zu verwenden. Der für die Auswertung der Norm erforderliche Vektor $\mathbf{A}\mathbf{z}^{(m)}$ stimmt gerade mit dem in Algorithmus 5.8 berechneten Vektor $\mathbf{w}^{(m+1)}$ überein, erfordert also keine zusätzliche Matrix-Vektor-Multiplikation.

Offenbar gilt für einen exakten Eigenvektor $\epsilon_m = 0$, und auch bei genäherten Eigenvektoren erhalten wir noch eine gute Schätzung für den Eigenwert:

Lemma 5.11 (Eigenwert-Approximation) *Sei $\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ eine Näherung eines Eigenvektors $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ der Matrix $\mathbf{A}$ zum Eigenwert λ. Dann gilt*

$$|\Lambda_A(\mathbf{y}) - \lambda| \leq \|\mathbf{A} - \lambda\mathbf{I}\|_2 \sin\angle(\mathbf{y}, \mathbf{x}) \leq \|\mathbf{A} - \lambda\mathbf{I}\|_2 \frac{\|\mathbf{y} - \alpha\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \qquad \text{für alle } \alpha \in \mathbb{K}.$$

Falls $\mathbf{A}$ eine normale Matrix ist, gilt sogar

$$|\Lambda_A(\mathbf{y}) - \lambda| \leq \|\mathbf{A} - \lambda\mathbf{I}\|_2 \sin^2\angle(\mathbf{y}, \mathbf{x}) \leq \|\mathbf{A} - \lambda\mathbf{I}\|_2 \left(\frac{\|\mathbf{y} - \alpha\mathbf{x}\|_2}{\|\mathbf{y}\|_2}\right)^2 \qquad \text{für alle } \alpha \in \mathbb{K}.$$

Beweis. Sei $\alpha \in \mathbb{K}$. Nach Definition gilt $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, also folgt die Abschätzung

$$\begin{aligned}|\Lambda_A(\mathbf{y}) - \lambda| &= \left|\frac{\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2}{\langle \mathbf{y}, \mathbf{y}\rangle_2} - \frac{\langle \mathbf{y}, \lambda\mathbf{y}\rangle_2}{\langle \mathbf{y}, \mathbf{y}\rangle_2}\right| = \left|\frac{\langle \mathbf{y}, (\mathbf{A} - \lambda\mathbf{I})\mathbf{y}\rangle_2}{\langle \mathbf{y}, \mathbf{y}\rangle_2}\right| = \left|\frac{\langle \mathbf{y}, (\mathbf{A} - \lambda\mathbf{I})(\mathbf{y} - \alpha\mathbf{x})\rangle_2}{\langle \mathbf{y}, \mathbf{y}\rangle_2}\right| \\ &\leq \frac{\|\mathbf{y}\|_2 \|(\mathbf{A} - \lambda\mathbf{I})(\mathbf{y} - \alpha\mathbf{x})\|_2}{\|\mathbf{y}\|_2^2} \leq \frac{\|\mathbf{A} - \lambda\mathbf{I}\|_2 \|\mathbf{y} - \alpha\mathbf{x}\|_2}{\|\mathbf{y}\|_2}\end{aligned}$$

Da wir diese Ungleichung für jedes $\alpha \in \mathbb{K}$ bewiesen haben, folgt aus Definition 5.1 die erste Aussage. Sei nun $\mathbf{A}$ eine normale Matrix. Nach Lemma 3.45 gilt

$$\|(\mathbf{A}^* - \bar{\lambda}\mathbf{I})\mathbf{x}\|_2 = \|(\mathbf{A} - \lambda\mathbf{I})^*\mathbf{x}\|_2 = \|(\mathbf{A} - \lambda\mathbf{I})\mathbf{x}\|_2 = 0,$$

also insbesondere auch $(\mathbf{A}^* - \bar{\lambda}\mathbf{I})\mathbf{x} = \mathbf{0}$, $\mathbf{x}$ ist also ein Eigenvektor von $\mathbf{A}^*$ zum Eigenwert $\bar{\lambda}$. Mit Hilfe dieser Gleichung können wir unsere Abschätzung verbessern:

$$\begin{aligned}|\Lambda_A(\mathbf{y}) - \lambda| &= \left|\frac{\langle \mathbf{y}, (\mathbf{A} - \lambda\mathbf{I})(\mathbf{y} - \alpha\mathbf{x})\rangle_2}{\|\mathbf{y}\|_2^2}\right| = \left|\frac{\langle (\mathbf{A} - \lambda\mathbf{I})^*\mathbf{y}, \mathbf{y} - \alpha\mathbf{x}\rangle_2}{\|\mathbf{y}\|_2^2}\right| \\ &= \left|\frac{\langle (\mathbf{A} - \lambda\mathbf{I})^*(\mathbf{y} - \alpha\mathbf{x}), \mathbf{y} - \alpha\mathbf{x}\rangle_2}{\|\mathbf{y}\|_2^2}\right| = \left|\frac{\langle \mathbf{y} - \alpha\mathbf{x}, (\mathbf{A} - \lambda\mathbf{I})(\mathbf{y} - \alpha\mathbf{x})\rangle_2}{\|\mathbf{y}\|_2^2}\right| \\ &\leq \frac{\|\mathbf{y} - \alpha\mathbf{x}\|_2 \|\mathbf{A} - \lambda\mathbf{I}\|_2 \|\mathbf{y} - \alpha\mathbf{x}\|_2}{\|\mathbf{y}\|_2^2} = \|\mathbf{A} - \lambda\mathbf{I}\|_2 \left(\frac{\|\mathbf{y} - \alpha\mathbf{x}\|_2}{\|\mathbf{y}\|_2}\right)^2,\end{aligned}$$

und da auch diese Abschätzung für beliebiges $\alpha \in \mathbb{K}$ gezeigt wurde, folgt die zweite Fehlerabschätzung. ∎

Es ist bemerkenswert, dass bei normalen Matrizen der mit Hilfe des Rayleigh-Quotienten berechnete Eigenwert wesentlich schneller als der Eigenvektor konvergieren kann. Falls wir also nur an einer Approximation des Eigenwerts interessiert sind, kann eine wesentlich geringere Anzahl von Iterationen bereits eine gute Genauigkeit erreichen.

Wir können die in (5.5) gegebene Größe für die Konstruktion eines Abbruchkriteriums der Vektoriteration verwenden. Um sicherzustellen, dass das Kriterium unabhängig von der Skalierung der Matrix $\mathbf{A}$ arbeitet, vergleichen wir ϵ_m mit $\epsilon|\lambda|$ für eine gegebene Fehlerschranke $\epsilon > 0$.

Algorithmus 5.12 (Vektoriteration mit Abbruchkriterium) *Seien $\mathbf{A} \in \mathbb{K}^{n\times n}$ und $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ wie in Satz 5.4 gegeben und gelte $\gamma \neq 0$. Sei $\epsilon \in \mathbb{R}_{>0}$. Dann berechnet der folgende Algorithmus eine Näherung λ des Eigenwerts λ_1 und eine Näherung $\mathbf{z}$ eines zugehörigen Eigenvektors:*

$\mathbf{x} \leftarrow \mathbf{x}/\|\mathbf{z}\|$
$\mathbf{a} \leftarrow \mathbf{A}\mathbf{x}$
$\lambda \leftarrow \langle \mathbf{x}, \mathbf{a}\rangle$
while $\|\mathbf{a} - \lambda\mathbf{x}\| > \epsilon|\lambda|$ *do begin*
  $\mathbf{x} \leftarrow \mathbf{a}/\|\mathbf{a}\|$
  $\mathbf{a} \leftarrow \mathbf{A}\mathbf{x}$
  $\lambda \leftarrow \langle \mathbf{x}, \mathbf{a}\rangle$
end

Die Näherungen erfüllen die Abschätzung $\|\mathbf{Ax} - \lambda\mathbf{x}\| \leq \epsilon|\lambda|$.

Bemerkung 5.13 (Implementierung) *Für den Algorithmus 5.8 ist es lediglich erforderlich, dass sich die Matrix* $\mathbf{A}$ *mit einem Vektor multiplizieren lässt. Diese Eigenschaft ist* sehr *wichtig, da es sehr viele Fälle gibt, in denen das Auswerten einer Matrix relativ kostengünstig durchzuführen ist, etwa bei Matrizen mit vielen Nulleinträgen (z.B. Bandmatrizen wie die Tridiagonalmatrix aus Beispiel 2.1) oder bei Matrizen, die implizit als Lösungsoperator eines linearen Gleichungssystems gegeben sind.*

Bemerkung 5.14 (Iteration im Teilraum) *Falls* $\mathbf{A}$ *eine normale Matrix ist, stehen ihre Eigenvektoren senkrecht aufeinander. Nehmen wir an, dass ein Eigenvektor* $\mathbf{e}^{(1)}$ *zu dem Eigenwert* λ_1 *berechnet wurde. Dann können wir einen Anfangsvektor* $\mathbf{x}^{(0)}$ *wählen, der senkrecht auf dem bereits berechneten Eigenvektor steht. Es lässt sich einfach nachrechnen, dass dann auch die gesamte Vektoriteration in dem orthogonalen Komplement dieses Eigenvektors ablaufen wird, dass also alle* $\mathbf{x}^{(m)}$ *senkrecht auf* $\mathbf{e}^{(1)}$ *stehen werden.*

Wenn wir $\mathbf{A}$ *als Abbildung dieses invarianten Teilraums in sich interpretieren, ist ihr betragsgrößter Eigenwert nun* λ_2*, und falls* $|\lambda_2| > |\lambda_3|$ *gilt, wird die Vektoriteration gegen einen Eigenvektor zu dem Eigenwert* λ_2 *konvergieren.*

Falls $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$ *gilt, können wir in dieser Weise nach und nach Eigenvektoren zu allen Eigenwerten berechnen und die Matrix diagonalisieren.*

5.2 Fehleranalyse

Der Algorithmus 5.12 endet, sobald die Norm des *Residuums*

$$\mathbf{r}_A(\mathbf{z}) := \Lambda_A(\mathbf{z})\mathbf{z} - \mathbf{Az}$$

den Wert $\epsilon|\lambda|$ unterschreitet. Es stellt sich die Frage, ob diese Eigenschaft bereits bedeutet, dass wir gute Näherungen des Eigenwerts und des Eigenvektors berechnet haben.

Wir beschränken uns bei der Untersuchung auf den Fall einer selbstadjungierten Matrix $\mathbf{A} = \mathbf{A}^* \in \mathbb{K}^{n\times n}$.

Bei der Analyse des Fehlers verwenden wir ein allgemeines Prinzip, das auch in anderen Gebieten der numerischen Mathematik von Bedeutung ist: Die Idee der *Rückwärtsanalyse* beruht darauf, zu einer Näherungslösung eines Problems ein zweites Problem zu konstruieren, das durch die Näherungslösung exakt gelöst wird. Falls Aussagen darüber zur Verfügung stehen, wie sich die Lösungen eines Problems unter Störungen der Problemstellung verändern, kann man dann Rückschlüsse auf die Genauigkeit der Näherungslösung ziehen.

In unserem Fall suchen wir einen Eigenvektor $\mathbf{x}$, also eine Lösung der Gleichung

$$\mathbf{Ax} = \lambda\mathbf{x}.$$

Uns steht ein genäherter Eigenvektor $\widetilde{\mathbf{x}}$ zur Verfügung, zu dem $\tilde{\lambda} = \Lambda_A(\widetilde{\mathbf{x}})$ eine Näherung des Eigenwerts darstellt. Wir suchen eine Matrix $\widetilde{\mathbf{A}}$ derart, dass

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \tilde{\lambda}\widetilde{\mathbf{x}}$$

gilt. Wenn wir das Residuum mit

$$\mathbf{r} := \Lambda_A(\widetilde{\mathbf{x}})\widetilde{\mathbf{x}} - \mathbf{A}\widetilde{\mathbf{x}} = \tilde{\lambda}\widetilde{\mathbf{x}} - \mathbf{A}\widetilde{\mathbf{x}}$$

bezeichnen, erhalten wir wegen

$$\langle \mathbf{r}, \widetilde{\mathbf{x}}\rangle = \overline{\Lambda_A(\widetilde{\mathbf{x}})}\langle \widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle - \langle \mathbf{A}\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle = \frac{\langle \mathbf{A}\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle}{\langle \widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle}\langle \widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle - \langle \mathbf{A}\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle = 0$$

für die Matrix

$$\widetilde{\mathbf{A}} := \mathbf{A} + \frac{\widetilde{\mathbf{x}}\mathbf{r}^*}{\|\widetilde{\mathbf{x}}\|^2} + \frac{\mathbf{r}\widetilde{\mathbf{x}}^*}{\|\widetilde{\mathbf{x}}\|^2}$$

die Gleichung

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \mathbf{A}\widetilde{\mathbf{x}} + \widetilde{\mathbf{x}}\frac{\langle \mathbf{r}, \widetilde{\mathbf{x}}\rangle}{\|\widetilde{\mathbf{x}}\|^2} + \mathbf{r}\frac{\langle \widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle}{\|\widetilde{\mathbf{x}}\|^2} = \mathbf{A}\widetilde{\mathbf{x}} + \mathbf{r} = \mathbf{A}\widetilde{\mathbf{x}} + \tilde{\lambda}\widetilde{\mathbf{x}} - \mathbf{A}\widetilde{\mathbf{x}} = \tilde{\lambda}\widetilde{\mathbf{x}},$$

also erfüllt $\widetilde{\mathbf{A}}$ unsere Anforderungen.

Die Spektralnorm der Störung lässt sich besonders elegant darstellen:

Lemma 5.15 (Rang-2-Matrix) *Seien* $\mathbf{a}, \mathbf{b} \in \mathbb{K}^n$ *mit* $\langle \mathbf{a}, \mathbf{b}\rangle = 0$ *gegeben. Sei* $\mathbf{E} := \mathbf{a}\mathbf{b}^* + \mathbf{b}\mathbf{a}^*$. *Dann gilt* $\|\mathbf{E}\| = \|\mathbf{a}\|\,\|\mathbf{b}\|$.

Beweis. Wir untersuchen zunächst den Sonderfall $\|\mathbf{a}\| = 1 = \|\mathbf{b}\|$.

Sei $\mathbf{x} \in \mathbb{K}^n$. Wir zerlegen den Vektor in Anteile aus dem Aufspann der Vektoren $\mathbf{a}$ und $\mathbf{b}$ sowie einen Rest, der senkrecht auf beiden steht:

$$\alpha := \langle \mathbf{a}, \mathbf{x}\rangle, \qquad \beta := \langle \mathbf{b}, \mathbf{x}\rangle, \qquad \mathbf{x}_0 := \mathbf{x} - \alpha\mathbf{a} - \beta\mathbf{b}.$$

Dann erhalten wir

$$\begin{aligned}
\langle \mathbf{a}, \mathbf{x}_0\rangle &= \langle \mathbf{a}, \mathbf{x}\rangle - \alpha\langle \mathbf{a}, \mathbf{a}\rangle - \beta\langle \mathbf{a}, \mathbf{b}\rangle = \alpha - \alpha = 0,\\
\langle \mathbf{b}, \mathbf{x}_0\rangle &= \langle \mathbf{b}, \mathbf{x}\rangle - \alpha\langle \mathbf{b}, \mathbf{a}\rangle - \beta\langle \mathbf{b}, \mathbf{b}\rangle = \beta - \beta = 0,\\
\mathbf{E}\mathbf{x}_0 &= \mathbf{a}\langle \mathbf{b}, \mathbf{x}_0\rangle + \mathbf{b}\langle \mathbf{a}, \mathbf{x}_0\rangle = 0,
\end{aligned}$$

also liegt $\mathbf{x}_0$ im Kern der Matrix $\mathbf{E}$. Es folgt

$$\mathbf{E}\mathbf{x} = \mathbf{E}(\mathbf{x}_0 + \alpha\mathbf{a} + \beta\mathbf{b}) = \alpha\mathbf{a}\langle \mathbf{b}, \mathbf{a}\rangle + \alpha\mathbf{b}\langle \mathbf{a}, \mathbf{a}\rangle + \beta\mathbf{a}\langle \mathbf{b}, \mathbf{b}\rangle + \beta\mathbf{b}\langle \mathbf{a}, \mathbf{b}\rangle = \alpha\mathbf{b} + \beta\mathbf{a},$$

und infolge der Orthogonalität der Vektoren $\mathbf{x}_0$, $\mathbf{a}$ und $\mathbf{b}$ ergibt sich

$$\begin{aligned}
\|\mathbf{E}\mathbf{x}\|^2 &= \|\alpha\mathbf{b} + \beta\mathbf{a}\|^2 = |\alpha|^2\|\mathbf{b}\|^2 + |\beta|^2\|\mathbf{a}\|^2 = |\alpha|^2 + |\beta|^2 \le \|\mathbf{x}_0\|^2 + |\alpha|^2 + |\beta|^2\\
&= \|\mathbf{x}_0\|^2 + \|\alpha\mathbf{a}\|^2 + \|\beta\mathbf{b}\|^2 = \|\mathbf{x}_0 + \alpha\mathbf{a} + \beta\mathbf{b}\|^2 = \|\mathbf{x}\|^2,
\end{aligned}$$

also $\|\mathbf{E}\| \le 1 = \|\mathbf{a}\|\,\|\mathbf{b}\|$ nach Definition der Spektralnorm. Wegen

$$\|\mathbf{E}\mathbf{a}\| = \|\mathbf{b}\langle \mathbf{a}, \mathbf{a}\rangle + \mathbf{a}\langle \mathbf{b}, \mathbf{a}\rangle\| = \|\mathbf{b}\| = 1$$

muss auch $\|\mathbf{E}\| \geq 1$ gelten, so dass wir $\|\mathbf{E}\| = 1$ bewiesen haben.

Widmen wir uns nun dem allgemeinen Fall. Sollte $\mathbf{a} = \mathbf{0}$ oder $\mathbf{b} = \mathbf{0}$ gelten, so folgen $\mathbf{E} = \mathbf{0}$ und damit die Behauptung.

Anderenfalls setzen wir $\widehat{\mathbf{a}} := \mathbf{a}/\|\mathbf{a}\|$ und $\widehat{\mathbf{b}} := \mathbf{b}/\|\mathbf{b}\|$ und wenden den bereits bewiesenen Sonderfall auf $\widehat{\mathbf{E}} := \widehat{\mathbf{a}}\widehat{\mathbf{b}}^* + \widehat{\mathbf{b}}\widehat{\mathbf{a}}^*$ an, um $\|\widehat{\mathbf{E}}\| = 1$ zu erhalten. Mit $\|\mathbf{E}\| = \|\mathbf{a}\|\,\|\mathbf{b}\|\,\|\widehat{\mathbf{E}}\|$ folgt die Behauptung. ∎

Die Anwendung dieses Lemmas auf unseren Fall führt zu

$$\|\mathbf{A} - \widetilde{\mathbf{A}}\| = \frac{\|\mathbf{r}\|}{\|\widetilde{\mathbf{x}}\|}, \tag{5.6}$$

die Störung der Matrix ist also unmittelbar proportional zu der relativen Größe des Residuums, die wir in unserem Algorithmus explizit berechnen können.

Unser Ziel ist es, aus dieser Abschätzung Rückschlüsse auf die Genauigkeit der Approximation des Eigenwerts und des Eigenvektors zu gewinnen. Als Hilfsmittel verwenden wir die folgende sehr vereinfachte Fassung des *Courant-Fischer-Weyl-Minimax-Prinzips*:

Satz 5.16 (Courant-Minimierung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert. Dann sind nach Folgerung 3.41 alle Eigenwerte der Matrix reell. Wir bezeichnen mit $\lambda_{min} \in \sigma(\mathbf{A})$ den minimalen Eigenwert. Dann gilt*

$$\lambda_{min} = \min\{\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle \ : \ \mathbf{z} \in \mathbb{K}^n, \ \|\mathbf{z}\| = 1\},$$

der minimale Eigenwert ist also gerade das Minimum des Rayleigh-Quotienten.

Jeder Vektor $\mathbf{z} \in \mathbb{K}^n$ mit $\|\mathbf{z}\| = 1$ und $\lambda_{min} = \langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle$ ist ein Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert λ_{min}.

Beweis. Nach Folgerung 3.41 finden wir eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ mit

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*.$$

Für einen beliebigen Vektor $\mathbf{z} \in \mathbb{K}^n$ und $\widehat{\mathbf{z}} := \mathbf{Q}^*\mathbf{z}$ gilt

$$\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle = \langle \mathbf{z}, \mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{z}\rangle = \langle \mathbf{Q}^*\mathbf{z}, \mathbf{D}\mathbf{Q}^*\mathbf{z}\rangle = \langle \widehat{\mathbf{z}}, \mathbf{D}\widehat{\mathbf{z}}\rangle = \sum_{i=1}^{n} d_{ii}|\widehat{z}_i|^2.$$

Mit Lemma 3.33 folgt

$$\begin{aligned}
\min\{\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle \ : \ \mathbf{z} \in \mathbb{K}^n, \ \|\mathbf{z}\| = 1\} &= \min\{\langle \widehat{\mathbf{z}}, \mathbf{D}\widehat{\mathbf{z}}\rangle \ : \ \mathbf{z} \in \mathbb{K}^n, \ \|\mathbf{z}\| = 1, \ \widehat{\mathbf{z}} = \mathbf{Q}^*\mathbf{z}\} \\
&= \min\{\langle \widehat{\mathbf{z}}, \mathbf{D}\widehat{\mathbf{z}}\rangle \ : \ \mathbf{z} \in \mathbb{K}^n, \ \widehat{\mathbf{z}} = \mathbf{Q}^*\mathbf{z}, \ \|\widehat{\mathbf{z}}\| = 1\} \\
&= \min\{\langle \widehat{\mathbf{z}}, \mathbf{D}\widehat{\mathbf{z}}\rangle \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n, \ \|\widehat{\mathbf{z}}\| = 1\} \\
&= \min\left\{\sum_{i=1}^{n} d_{ii}|\widehat{z}_i|^2 \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n, \ \|\widehat{\mathbf{z}}\| = 1\right\}.
\end{aligned}$$

Da $\mathbf{A}$ und $\mathbf{D}$ ähnlich sind, ist $\lambda_{\min}$ auch der minimale Eigenwert der Matrix $\mathbf{D}$. Aufgrund derer Diagonalgestalt finden wir $j \in \{1, \ldots, n\}$ mit

$$\lambda_{\min} = d_{jj} \leq d_{ii} \qquad \text{für alle } i \in \{1, \ldots, n\}.$$

Damit folgt einerseits

$$\begin{aligned}
\min\{\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle \ : \ \mathbf{z} \in \mathbb{K}^n, \ \|\mathbf{z}\| = 1\} &= \min\left\{\sum_{i=1}^n d_{ii}|\widehat{z}_i|^2 \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n, \ \|\widehat{\mathbf{z}}\| = 1\right\} \\
&\geq \min\left\{\sum_{j=1}^n \lambda_{\min}|\widehat{z}_i|^2 \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n, \ \|\widehat{\mathbf{z}}\| = 1\right\} \\
&= \lambda_{\min} \min\left\{\sum_{j=1}^n |\widehat{z}_i|^2 \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n, \ \|\widehat{\mathbf{z}}\| = 1\right\} = \lambda_{\min}
\end{aligned}$$

und andererseits für den j-ten kanonischen Einheitsvektor

$$\begin{aligned}
\min\{\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle \ : \ \mathbf{z} \in \mathbb{K}^n, \ \|\mathbf{z}\| = 1\} &= \min\left\{\sum_{i=1}^n d_{ii}|\widehat{z}_i|^2 \ : \ \widehat{\mathbf{z}} \in \mathbb{K}^n, \ \|\widehat{\mathbf{z}}\| = 1\right\} \\
&\leq \sum_{i=1}^n d_{ii}|\delta_i^{(j)}|^2 = d_{jj} = \lambda_{\min}.
\end{aligned}$$

Damit ist die gewünschte Gleichung bewiesen.

Sei nun $\mathbf{z} \in \mathbb{K}^n$ mit $\|\mathbf{z}\| = 1$ und $\lambda_{\min} = \langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle$ gegeben, und sei $\widehat{\mathbf{z}} := \mathbf{Q}^*\mathbf{z}$. Nach Voraussetzung gilt

$$\begin{aligned}
0 &= \langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle - \lambda_{\min}\|\mathbf{z}\|^2 = \langle \widehat{\mathbf{z}}, \mathbf{D}\widehat{\mathbf{z}}\rangle - \lambda_{\min}\|\widehat{\mathbf{z}}\|^2 \\
&= \sum_{i=1}^n d_{ii}|\widehat{z}_i|^2 - \lambda_{\min}\sum_{i=1}^n |\widehat{z}_i|^2 = \sum_{i=1}^n (d_{ii} - \lambda_{\min})|\widehat{z}_i|^2.
\end{aligned}$$

Da keiner der Summanden negativ sein kann, müssen sie alle gleich null sein. Insbesondere muss aus $d_{ii} \neq \lambda_{\min}$ bereits $\widehat{z}_i = 0$ für alle $i \in \{1, \ldots, n\}$ folgen. Damit ist $\widehat{\mathbf{z}}$ ein Eigenvektor der Matrix $\mathbf{D}$ zu dem Eigenwert $\lambda_{\min}$, also ist auch $\mathbf{z}$ ein Eigenvektor der Matrix $\mathbf{A}$ zu demselben Eigenwert. ∎

Neben seiner Bedeutung für den folgenden Störungssatz kann das Courant-Minimierungsprinzip auch als Ausgangspunkt bei der Konstruktion eines numerischen Näherungsverfahrens dienen: Statt unmittelbar nach einem Eigenvektor zu suchen, können wir uns auch um ein Minimum des Rayleigh-Quotienten bemühen. Mit diesem Ansatz werden wir uns später eingehender befassen.

Zunächst formulieren wir allerdings den *Bauer-Fike-Störungssatz* für selbstadjungierte Matrizen, mit dessen Hilfe wir eine Aussage über die Konvergenz der Eigenwerte gewinnen können.

Satz 5.17 (Bauer-Fike) *Seien $\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{K}^{n\times n}$ selbstadjungierte Matrizen und sei $\tilde{\lambda} \in \sigma(\widetilde{\mathbf{A}})$. Dann existiert ein Eigenwert $\lambda \in \sigma(\mathbf{A})$ mit*

$$|\lambda - \tilde{\lambda}| \leq \|\mathbf{A} - \widetilde{\mathbf{A}}\|.$$

Beweis. Wir wählen ein $\lambda \in \sigma(\mathbf{A})$ mit minimalem Abstand zu $\tilde{\lambda}$, es soll also

$$|\lambda - \tilde{\lambda}| \leq |\mu - \tilde{\lambda}| \qquad \text{für alle } \mu \in \sigma(\mathbf{A})$$

gelten. Mit Folgerung 3.41 finden wir eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine reelle Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ mit $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$. Es folgt

$$(\mathbf{A} - \tilde{\lambda}\mathbf{I})^2 = (\mathbf{Q}\mathbf{D}\mathbf{Q}^* - \tilde{\lambda}\mathbf{Q}\mathbf{Q}^*)^2 = \mathbf{Q}(\mathbf{D} - \tilde{\lambda}\mathbf{I})^2\mathbf{Q}^*,$$

und da die Eigenwerte der Matrix $\mathbf{A}$ die Diagonalelemente der Matrix $\mathbf{D}$ sind, muss $(\lambda - \tilde{\lambda})^2$ der kleinste Eigenwert der Matrix $(\mathbf{A} - \tilde{\lambda}\mathbf{I})^2$ sein.

Sei $\widetilde{\mathbf{x}} \in \mathbb{K}^n$ ein Eigenvektor der Matrix $\widetilde{\mathbf{A}}$ zu dem Eigenwert $\tilde{\lambda}$ mit $\|\widetilde{\mathbf{x}}\| = 1$. Dann gilt nach Satz 5.16 und (3.6a) die Abschätzung

$$\begin{aligned}(\lambda - \tilde{\lambda})^2 &= \min\{\langle(\mathbf{A} - \tilde{\lambda}\mathbf{I})^2\mathbf{z}, \mathbf{z}\rangle \;:\; \mathbf{z} \in \mathbb{K}^n,\ \|\mathbf{z}\| = 1\} \\ &\leq \langle(\mathbf{A} - \tilde{\lambda}\mathbf{I})^2\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}\rangle = \langle(\mathbf{A} - \tilde{\lambda}\mathbf{I})\widetilde{\mathbf{x}}, (\mathbf{A} - \tilde{\lambda}\mathbf{I})\widetilde{\mathbf{x}}\rangle \\ &= \langle(\mathbf{A} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}}, (\mathbf{A} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}}\rangle = \|(\mathbf{A} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}}\|^2 \leq \|\mathbf{A} - \widetilde{\mathbf{A}}\|^2\|\widetilde{\mathbf{x}}\|^2 = \|\mathbf{A} - \widetilde{\mathbf{A}}\|^2,\end{aligned}$$

und das ist unsere Behauptung. ∎

Indem wir diese Abschätzung mit (5.6) kombinieren, erhalten wir für ein $\lambda \in \sigma(\mathbf{A})$ schon die Ungleichung

$$|\lambda - \tilde{\lambda}| \leq \|\mathbf{A} - \widetilde{\mathbf{A}}\| = \frac{\|\mathbf{r}\|}{\|\widetilde{\mathbf{x}}\|},$$

können also die Genauigkeit des genäherten Eigenwerts durch die praktisch berechenbaren Größen $\|\mathbf{r}\|$ und $\|\mathbf{x}\|$ beschreiben.

Natürlich sind wir auch daran interessiert, Aussagen über die Qualität der Approximation des Eigenvektors zu gewinnen. Diese Aufgabe erweist sich als schwieriger:

Beispiel 5.18 (Keine Konvergenz) *Wir untersuchen die Matrizen*

$$\mathbf{A} := \begin{pmatrix} 1+2\epsilon & \\ & 1 \end{pmatrix}, \qquad \widetilde{\mathbf{A}} := \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1+\epsilon & \\ & 1-\epsilon \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Für $\epsilon \to 0$ konvergieren beide gegen die Einheitsmatrix, also insbesondere auch gegen einander, wir können die Differenz mit Lemma 3.48 sogar explizit berechnen, indem wir die Nullstellen des charakteristischen Polynoms bestimmen:

$$\begin{aligned}\|\mathbf{A} - \widetilde{\mathbf{A}}\| &= \left\|\begin{pmatrix} 2\epsilon & -\epsilon \\ -\epsilon & 0 \end{pmatrix}\right\| = \max\{|\lambda| \;:\; (\lambda - 2\epsilon)\lambda - \epsilon^2 = 0\} \\ &= \max\{(\sqrt{2}-1)\epsilon, (\sqrt{2}+1)\epsilon\} = (\sqrt{2}+1)\epsilon.\end{aligned}$$

Der Winkel zwischen den kanonischen Einheitsvektoren, die die Eigenvektoren der Matrix $\mathbf{A}$ *sind, und den Eigenvektoren der Matrix* $\widehat{\mathbf{A}}$ *dagegen ist durch*

$$\cos \angle(\mathbf{x}, \widetilde{\mathbf{x}}) = \frac{\left| \left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\rangle \right|}{\sqrt{2}} = 1/\sqrt{2}$$

gegeben, beträgt also unabhängig von ϵ *immer* $\pi/4$*. Die Eigenvektoren der Matrizen konvergieren demzufolge nicht gegeneinander.*

Für $\epsilon = 0$ sind in unserem Beispiel *alle* von null verschiedenen Vektoren Eigenvektoren, so dass die Eigenschaft, Eigenvektor zu sein, keine Aussagen über einen Vektor mehr zulässt. Diesen Sonderfall können wir ausschließen, indem wir messen, wie nahe die Eigenwerte beieinander liegen, um so eine „Durchmischung der Eigenräume“ zu vermeiden.

Definition 5.19 (Spektrallücke) *Sei* $\mathbf{A} \in \mathbb{K}^{n \times n}$ *eine normale Matrix. Für alle* $\lambda \in \mathbb{K}$ *definieren wir die* Spektrallücke *zu* λ *in Bezug auf* $\mathbf{A}$ *durch*

$$\gamma_A(\lambda) := \inf\{|\mu - \lambda| \ : \ \mu \in \sigma(\mathbf{A}) \setminus \{\lambda\}\},$$

also gerade als den Abstand von λ *zu dem nächstgelegenen Eigenwert. Wie üblich setzen wir* $\inf \emptyset = \infty$ *und verwenden in den folgenden Argumenten die Konvention* $1/\infty = 0$*.*

Ausgehend von einer Abschätzung für $\|\mathbf{A} - \widetilde{\mathbf{A}}\|$ ist es relativ leicht, Aussagen im Bildbereich der Matrizen zu formulieren. Da wir daran interessiert sind, eine Aussage im Definitionsbereich, nämlich über die Störung der Eigenvektoren, zu erhalten, brauchen wir eine Möglichkeit, aus ersterem letztere zu gewinnen. Falls λ ein Eigenwert der Matrix $\mathbf{A}$ ist, kann $\lambda\mathbf{I} - \mathbf{A}$ nicht injektiv sein, also müssen wir Vektoren aus dem Kern gesondert berücksichtigen.

Lemma 5.20 (Urbild-Abschätzung) *Sei* $\mathbf{A} \in \mathbb{K}^{n \times n}$ *eine selbstadjungierte Matrix. Dann existiert eine Matrix* $\mathbf{\Pi} \in \mathbb{K}^{n \times n}$ *derart, dass*

$$\|\mathbf{z} - \mathbf{\Pi}\mathbf{z}\| \leq \frac{1}{\gamma_A(0)} \|\mathbf{A}\mathbf{z}\| \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^n \tag{5.7}$$

und $\mathbf{A}\mathbf{\Pi} = \mathbf{0}$ *gelten. Letztere Gleichung bedeutet, dass das Bild der Matrix* $\mathbf{\Pi}$ *im Kern der Matrix* $\mathbf{A}$ *enthalten ist.*

Beweis. Da $\mathbf{A}$ selbstadjungiert ist, finden wir nach Folgerung 3.41 beziehungsweise Satz 3.43 eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n \times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{K}^{n \times n}$ mit

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*.$$

Falls $\gamma_A(0) = \infty$ gilt, folgt $\gamma_D(0) = \infty$, also $\mathbf{D} = \mathbf{0}$ und damit $\mathbf{A} = \mathbf{0}$. In diesem Fall setzen wir $\mathbf{\Pi} = \mathbf{I}$ und sind fertig.

Anderenfalls definieren wir eine Diagonalmatrix $\widehat{\mathbf{\Pi}} \in \mathbb{R}^{n\times n}$ durch

$$\widehat{\pi}_{ij} := \begin{cases} 1 & \text{falls } i = j \text{ und } d_{ii} = 0, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i,j \in \{1,\ldots,n\}$$

und setzen $\mathbf{\Pi} := \mathbf{Q}\widehat{\mathbf{\Pi}}\mathbf{Q}^*$. Aus der Definition folgt $\mathbf{D}\widehat{\mathbf{\Pi}} = \mathbf{0}$, so dass sich

$$\mathbf{A\Pi} = \mathbf{QDQ}^*\mathbf{Q}\widehat{\mathbf{\Pi}}\mathbf{Q}^* = \mathbf{QD}\widehat{\mathbf{\Pi}}\mathbf{Q}^* = \mathbf{Q0Q}^* = \mathbf{0}$$

ergibt. Also ist das Bild der Matrix $\mathbf{\Pi}$ im Kern der Matrix $\mathbf{A}$ enthalten.

Sei $\mathbf{z} \in \mathbb{K}^n$. Wir definieren $\widehat{\mathbf{z}} := \mathbf{Q}^*\mathbf{z}$ und halten fest, dass nach Lemma 3.33 die Gleichung

$$\|\mathbf{Az}\| = \|\mathbf{QDQ}^*\mathbf{z}\| = \|\mathbf{D}\widehat{\mathbf{z}}\|$$

gilt. Die Norm können wir einfach berechnen und erhalten

$$\begin{aligned} \|\mathbf{D}\widehat{\mathbf{z}}\|^2 &= \sum_{i=1}^{n} |d_{ii}|^2|\widehat{z}_i|^2 = \sum_{\substack{i=1\\ d_{ii}\neq 0}}^{n} |d_{ii}|^2|\widehat{z}_i|^2 \geq \sum_{\substack{i=1\\ d_{ii}\neq 0}}^{n} \gamma_A(0)^2|\widehat{z}_i|^2 \\ &= \gamma_A(0)^2 \sum_{i=1}^{n} (1-\widehat{\pi}_{ii})^2|\widehat{z}_i|^2 = \gamma_A(0)^2\|(\mathbf{I}-\widehat{\mathbf{\Pi}})\widehat{\mathbf{z}}\|^2. \end{aligned}$$

Insgesamt ergibt sich

$$\gamma_A(0)\|\mathbf{z}-\mathbf{\Pi z}\| = \gamma_A(0)\|\mathbf{Q}^*(\mathbf{z}-\mathbf{\Pi z})\| = \gamma_A(0)\|\widehat{\mathbf{z}}-\widehat{\mathbf{\Pi}}\widehat{\mathbf{z}}\| \leq \|\mathbf{D}\widehat{\mathbf{z}}\| = \|\mathbf{Az}\|,$$

und die Division durch $\gamma_A(0) > 0$ führt zu der gewünschten Abschätzung. ∎

Bemerkung 5.21 (Projektion) *Die in Lemma 5.20 definierte Matrix* $\mathbf{\Pi}$ *ist eine* orthogonale Projektion*, erfüllt also* $\mathbf{\Pi}^2 = \mathbf{\Pi} = \mathbf{\Pi}^*$.

Aus (5.7) folgt, dass für jedes $\mathbf{z} \in \operatorname{Kern}(\mathbf{A})$ *die rechte Seite gleich null ist, also auch die linke, so dass* $\mathbf{\Pi z} = \mathbf{z}$ *folgt. Damit ist das Bild der Matrix* $\mathbf{\Pi}$ *nicht nur im Kern der Matrix enthalten, sondern es ist bereits der gesamte Kern.*

Damit ist $\mathbf{\Pi}$ *eine orthogonale Projektion auf den Kern der Matrix* $\mathbf{A}$*. Für jede Matrix* $\mathbf{A}$ *gibt es nur eine solche Projektion, und sie ordnet jedem Vektor* $\mathbf{z}$ *denjenigen Vektor* $\mathbf{\Pi z}$ *aus dem Kern zu, der ihm in der euklidischen Norm am nächsten kommt.*

Satz 5.22 (Gestörtes Eigenwertproblem) *Seien* $\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{K}^{n\times n}$ *selbstadjungierte Matrizen, sei* $\tilde{\lambda} \in \sigma(\widetilde{\mathbf{A}})$ *ein Eigenwert der Matrix* $\widetilde{\mathbf{A}}$ *und* $\widetilde{\mathbf{x}} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ *ein zugehöriger Eigenvektor.*

Dann existieren ein Eigenwert $\lambda \in \sigma(\mathbf{A})$ *und ein Vektor* $\mathbf{x} \in \mathcal{E}_A(\lambda)$ *aus dem zugehörigen Eigenraum der Matrix* $\mathbf{A}$ *mit*

$$|\lambda - \tilde{\lambda}| \leq \|\mathbf{A}-\widetilde{\mathbf{A}}\|, \qquad \|\mathbf{x}-\widetilde{\mathbf{x}}\| \leq \frac{2}{\gamma_A(\lambda)}\|\mathbf{A}-\widetilde{\mathbf{A}}\|\,\|\widetilde{\mathbf{x}}\|.$$

Beweis. Mit dem Satz 5.17 finden wir einen Eigenwert $\lambda \in \sigma(\mathbf{A})$ mit

$$|\lambda - \tilde{\lambda}| \leq \|\mathbf{A} - \widetilde{\mathbf{A}}\|.$$

Mit der Dreiecksungleichung, $\tilde{\lambda}\widetilde{\mathbf{x}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{x}}$ und (3.6a) erhalten wir

$$\begin{aligned}\|(\lambda\mathbf{I} - \mathbf{A})\widetilde{\mathbf{x}}\| &= \|(\tilde{\lambda}\mathbf{I} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}} - (\mathbf{A} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}} + (\lambda - \tilde{\lambda})\widetilde{\mathbf{x}}\| \\ &\leq \|(\tilde{\lambda}\mathbf{I} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}}\| + \|(\mathbf{A} - \widetilde{\mathbf{A}})\widetilde{\mathbf{x}}\| + |\lambda - \tilde{\lambda}|\,\|\widetilde{\mathbf{x}}\| \\ &\leq \|\mathbf{A} - \widetilde{\mathbf{A}}\|\,\|\widetilde{\mathbf{x}}\| + \|\mathbf{A} - \widetilde{\mathbf{A}}\|\,\|\widetilde{\mathbf{x}}\| = 2\|\mathbf{A} - \widetilde{\mathbf{A}}\|\,\|\widetilde{\mathbf{x}}\|.\end{aligned}$$

Wir wenden Lemma 5.20 auf $\lambda\mathbf{I} - \mathbf{A}$ an und erhalten mit $\gamma_{\lambda I - A}(0) = \gamma_A(\lambda)$ die Abschätzung

$$\|\widetilde{\mathbf{x}} - \mathbf{\Pi}\widetilde{\mathbf{x}}\| \leq \frac{1}{\gamma_A(\lambda)}\|(\lambda\mathbf{I} - \mathbf{A})\widetilde{\mathbf{x}}\| \leq \frac{2}{\gamma_A(\lambda)}\|\mathbf{A} - \widetilde{\mathbf{A}}\|\,\|\widetilde{\mathbf{x}}\|.$$

Das Bild der in Lemma 5.20 definierten Matrix $\mathbf{\Pi}$ liegt im Kern der Matrix $\lambda\mathbf{I} - \mathbf{A}$, also gerade im Eigenraum $\mathcal{E}_A(\lambda)$. Demnach erhalten wir mit $\mathbf{x} := \mathbf{\Pi}\widetilde{\mathbf{x}}$ die gewünschte Abschätzung. ∎

Indem wir diesen Satz auf die im Rahmen der Rückwärtsanalyse konstruierte Matrix $\widetilde{\mathbf{A}}$ anwenden und das Abbruchkriterium der in Algorithmus 5.12 gegebenen Vektoriteration einsetzen, erhalten wir die folgende Aussage über das von ihr berechnete Ergebnis:

Folgerung 5.23 (Ergebnis der Vektoriteration) *Sei $\widetilde{\mathbf{x}} \in \mathbb{K}^n$ ein Vektor, der*

$$\|\widetilde{\mathbf{x}}\| = 1, \qquad \|\mathbf{A}\widetilde{\mathbf{x}} - \tilde{\lambda}\widetilde{\mathbf{x}}\| \leq \epsilon|\tilde{\lambda}|$$

mit $\tilde{\lambda} := \Lambda_A(\widetilde{\mathbf{x}}) = \langle\widetilde{\mathbf{x}}, \mathbf{A}\widetilde{\mathbf{x}}\rangle$ erfüllt.

Dann existieren ein Eigenwert $\lambda \in \sigma(\mathbf{A})$ der Matrix $\mathbf{A}$ und ein Vektor $\mathbf{x} \in \mathcal{E}_A(\lambda)$ aus dem zugehörigen Eigenraum mit

$$\frac{|\lambda - \tilde{\lambda}|}{|\tilde{\lambda}|} \leq \epsilon, \qquad |\lambda - \tilde{\lambda}| \leq \frac{\|\lambda\mathbf{I} - \mathbf{A}\|\,|\tilde{\lambda}|^2}{\gamma_A(\lambda)^2}\epsilon^2, \qquad \|\mathbf{x} - \widetilde{\mathbf{x}}\| \leq \frac{2|\tilde{\lambda}|}{\gamma_A(\lambda)}\epsilon.$$

Beweis. Wir bezeichnen das Residuum wieder mit

$$\mathbf{r} := \tilde{\lambda}\widetilde{\mathbf{x}} - \mathbf{A}\widetilde{\mathbf{x}}$$

und setzen

$$\widetilde{\mathbf{A}} := \mathbf{A} + \frac{\widetilde{\mathbf{x}}\mathbf{r}^*}{\|\widetilde{\mathbf{x}}\|^2} + \frac{\mathbf{r}\widetilde{\mathbf{x}}^*}{\|\widetilde{\mathbf{x}}\|^2} = \mathbf{A} + \widetilde{\mathbf{x}}\mathbf{r}^* + \mathbf{r}\widetilde{\mathbf{x}}^*.$$

Nach Lemma 5.15 und Voraussetzung gilt

$$\|\mathbf{A} - \widetilde{\mathbf{A}}\| = \|\widetilde{\mathbf{x}}\mathbf{r}^* + \mathbf{r}\widetilde{\mathbf{x}}^*\| = \|\widetilde{\mathbf{x}}\|\,\|\mathbf{r}\| = \|\mathbf{r}\| \leq \epsilon|\tilde{\lambda}|.$$

Mit Satz 5.22 finden wir $\lambda \in \sigma(\mathbf{A})$ und $\mathbf{x} \in \mathcal{E}_A(\lambda)$ mit

$$|\lambda - \tilde{\lambda}| \leq \epsilon|\tilde{\lambda}|, \qquad \|\mathbf{x} - \widetilde{\mathbf{x}}\| \leq \frac{2}{\gamma_A(\lambda)}\epsilon|\tilde{\lambda}|.$$

Mit Lemma 5.11 folgt wegen $\|\widetilde{\mathbf{x}}\| = 1$ auch

$$|\lambda - \tilde{\lambda}| = |\Lambda_A(\widetilde{\mathbf{x}}) - \lambda| \leq \|\lambda\mathbf{I} - \mathbf{A}\|\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 \leq \frac{\|\lambda\mathbf{I} - \mathbf{A}\|\,|\tilde{\lambda}|^2}{\gamma_A(\lambda)^2}\epsilon^2,$$

also die noch fehlende Abschätzung. ∎

Bemerkung 5.24 (Relativer Fehler) *Wir können aus den Ergebnissen der vorangehenden Folgerung auch Abschätzungen des relativen Fehlers gewinnen, falls wir $\epsilon < 1$ voraussetzen: Für den Eigenwert gilt*

$$\frac{|\lambda - \tilde{\lambda}|}{|\lambda|} = \frac{|\lambda - \tilde{\lambda}|}{|\tilde{\lambda} + \lambda - \tilde{\lambda}|} \leq \frac{|\lambda - \tilde{\lambda}|}{|\tilde{\lambda}| - |\lambda - \tilde{\lambda}|} = \frac{|\lambda - \tilde{\lambda}|/|\tilde{\lambda}|}{1 - |\lambda - \tilde{\lambda}|/|\tilde{\lambda}|} \leq \frac{\epsilon}{1-\epsilon},$$

für den Eigenvektor erhalten wir

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\| \leq \frac{2|\tilde{\lambda}|}{\gamma_A(\lambda)}\epsilon \leq \frac{2(|\lambda| + |\tilde{\lambda} - \lambda|)}{\gamma_A(\lambda)}\epsilon = \frac{2|\lambda|}{\gamma_A(\lambda)}\left(1 + \frac{|\tilde{\lambda} - \lambda|}{|\lambda|}\right)\epsilon \leq \frac{2|\lambda|}{\gamma_A(\lambda)}\left(1 + \frac{\epsilon}{1-\epsilon}\right)\epsilon.$$

Bei dieser Abschätzung ist von besonderem Interesse, dass die relative *Spektrallücke $\gamma_A(\lambda)/|\lambda|$ ausreicht.*

Die Aussage über den Fehler des Eigenvektors können wir auch mit Hilfe des Winkels formulieren: Falls ϵ klein genug ist, folgt aus $\|\widetilde{\mathbf{x}}\| = 1$ bereits $\|\mathbf{x}\| > 0$, so dass wir mit Definition 5.1 unmittelbar abschätzen, um

$$\sin\angle(\widetilde{\mathbf{x}}, \mathbf{x}) \leq \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|}{\|\widetilde{\mathbf{x}}\|} = \|\widetilde{\mathbf{x}} - \mathbf{x}\| \leq \frac{2|\lambda|}{\gamma_A(\lambda)}\left(1 + \frac{\epsilon}{1-\epsilon}\right)\epsilon$$

zu erhalten. Dieses Resultat können wir natürlich auch wieder in Lemma 5.11 einsetzen, um eine verbesserte Aussage über den Fehler des Eigenwerts zu erhalten.

5.3 Inverse Iteration mit und ohne Shift

Im Beispiel 2.1 ist die Anwendung der Vektoriteration in der Form des Algorithmus 5.8 nicht sinnvoll, da die Eigenvektoren zu den größten Eigenwerten gerade diejenigen sind, die wegen des Diskretisierungsfehlers besonders wenig mit den Eigenvektoren des kontinuierlichen Problems zu tun haben.

Interessanter sind in diesem Fall die Eigenvektoren zu den *kleinsten* Eigenwerten, da diese Vektoren in der Regel relativ gut approximiert werden und auch für ingenieurtechnische Anwendungen von weitaus größerem Interesse sind, schließlich will man häufig die *niedrigste* Frequenz kennen, bei der es zu Oszillationen kommen kann.

Falls die Matrix $\mathbf{A}$ regulär und λ einer ihrer Eigenwerte mit Eigenvektor $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ist, gilt

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \qquad \mathbf{x} = \lambda\mathbf{A}^{-1}\mathbf{x}, \qquad \frac{1}{\lambda}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x},$$

also ist der Kehrwert jedes Eigenwerts von $\mathbf{A}$ auch ein Eigenwert von $\mathbf{A}^{-1}$. Da die Inverse von $\mathbf{A}^{-1}$ wieder $\mathbf{A}$ ist, folgt, dass das Spektrum von $\mathbf{A}^{-1}$ nur aus den Kehrwerten der Eigenwerte von $\mathbf{A}$ besteht.

Insbesondere ist der Kehrwert des betragskleinsten Eigenwerts von $\mathbf{A}$ gerade der betragsgrößte Eigenwert von $\mathbf{A}^{-1}$. Falls wir an dem betragskleinsten Eigenwert interessiert sind, liegt es also nahe, den Algorithmus 5.8 auf die inverse Matrix $\mathbf{A}^{-1}$ anzuwenden.

Wir definieren also die m-te Iterierte unseres neuen Verfahrens durch

$$\mathbf{x}^{(m)} := \mathbf{A}^{-m}\mathbf{x}^{(0)} = \mathbf{A}^{-1}\mathbf{x}^{(m-1)} \qquad \text{für alle } m \in \mathbb{N}. \tag{5.8}$$

Da es aus der Anwendung der Vektoriteration auf die Inverse entsteht, trägt dieses Iterationsverfahren den Namen *inverse Iteration*.

Für die Konvergenzuntersuchung müssen wir die Voraussetzungen so wählen, dass sich Satz 5.4 oder Folgerung 5.7 auf $\mathbf{A}^{-1}$ statt $\mathbf{A}$ anwenden lassen.

Wir beschränken uns auf den einfacheren der beiden Fälle: Sei $\mathbf{A}$ im Folgenden eine normale Matrix. Dann existieren nach Folgerungz 3.47 eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{C}^n$ mit

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D}.$$

Nach Bemerkung 3.40 können wir die Reihenfolge der Eigenwerte frei wählen, so dass wir

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad |\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_n| \tag{5.9}$$

erhalten. Daraus folgt

$$\left|\frac{1}{\lambda_1}\right| \ge \left|\frac{1}{\lambda_2}\right| \ge \cdots \ge \left|\frac{1}{\lambda_n}\right|.$$

Mit $\mathbf{e} := \mathbf{Q}\delta^{(1)}$ bezeichnen wir wieder einen Eigenvektor, der zu dem Eigenwert λ_1 der Matrix $\mathbf{A}$ gehört.

Satz 5.25 (Konvergenz) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine invertierbare normale Matrix. Sie besitzt eine Schur-zerlegung $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ mit einer unitären Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und einer Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ der Form (5.9).*

Dann ist $\mathbf{e} := \mathbf{Q}\delta^{(1)}$ ein Eigenvektor zu dem betragskleinsten Eigenwert λ_1.

Sei $\mathbf{x}^{(0)} \in \mathbb{K}^n$ mit $\langle \mathbf{e}, \mathbf{x}^{(0)}\rangle \neq 0$ gegeben. Dann gilt

$$\tan\angle(\mathbf{e}, \mathbf{x}^{(m)}) \le \left(\frac{|\lambda_1|}{|\lambda_2|}\right)^m \tan\angle(\mathbf{e}, \mathbf{x}^{(0)}) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir setzen $\widehat{\mathbf{A}} := \mathbf{A}^{-1}$. Es gilt

$$\mathbf{Q}^* \widehat{\mathbf{A}} \mathbf{Q} = \mathbf{Q}^* \mathbf{A}^{-1} \mathbf{Q} = \mathbf{Q}^* (\mathbf{Q}\mathbf{D}\mathbf{Q}^*)^* \mathbf{Q} = \mathbf{Q}^* \mathbf{Q} \mathbf{D}^{-1} \mathbf{Q}^* \mathbf{Q} = \mathbf{D}^{-1},$$

also ist insbesondere $\widehat{\mathbf{A}}$ diagonalisierbar mit derselben Transformation $\mathbf{Q}$. Die Eigenwerte von $\widehat{\mathbf{A}}$ sind offenbar gerade die Diagonalelemente von

$$\mathbf{D}^{-1} = \begin{pmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_n \end{pmatrix},$$

und sie sind nach Voraussetzung dem Betrag nach absteigend sortiert. Also dürfen wir Satz 5.4 anwenden und erhalten

$$\begin{aligned} \tan \angle(\mathbf{e}, \mathbf{x}^{(m)}) &\leq \left(\frac{1/|\lambda_2|}{1/|\lambda_1|} \right)^m \tan \angle(\mathbf{e}, \mathbf{x}^{(0)}) \\ &= \left(\frac{|\lambda_1|}{|\lambda_2|} \right)^m \tan \angle(\mathbf{e}, \mathbf{x}^{(0)}) \qquad \text{für alle } m \in \mathbb{N}_0. \end{aligned}$$

■

Bemerkung 5.26 *Mit der inversen Iteration lässt sich der Eigenvektor zu dem kleinsten Eigenwert der Matrix aus Beispiel 2.1 berechnen. Da der kontinuierliche Operator die Eigenwerte*

$$\lambda_k = c \frac{\pi^2}{\ell^2} k^2$$

besitzt und die Eigenwerte der diskreten Matrix gegen diese Werte konvergieren, wird die Konvergenzgeschwindigkeit der inversen Iteration gegen

$$\frac{|\lambda_1|}{|\lambda_2|} = 1/4$$

streben, also unabhängig von der Auflösung der Diskretisierung sein.

Während bei der Jacobi-Iteration die Konvergenzgeschwindigkeit potentiell von der Dimension der Matrix abhängt, ist sie also bei der inversen Iteration davon unabhängig. Das ist insbesondere bei hohen Auflösungen ein sehr großer Vorteil.

An diesem Beispiel zeigt sich auch die Wichtigkeit einer dem Problem angemessenen Implementierung des Verfahrens: Falls man bei der Lösung des tridiagonalen Gleichungssystems die Bandstruktur (etwa mit Hilfe einer LU-Zerlegung) ausnutzt, benötigt der gesamte Schleifenrumpf lediglich $\mathcal{O}(n)$ Operationen, so dass sich mit fast linearem Aufwand der gesuchte Eigenvektor bestimmen lässt.

Würde man stattdessen $\mathbf{A}^{-1}$ explizit berechnen, wäre dafür im Allgemeinen ein Aufwand von $\mathcal{O}(n^3)$ Operationen erforderlich, während die Multiplikation mit $\mathbf{A}^{-1}$ in jedem Schritt der inversen Iteration einen Aufwand von $\mathcal{O}(n^2)$ nach sich ziehen würde.

Wie wir gesehen haben steht uns $\mathbf{A}^{-1}$ in der Praxis oft nicht zur Verfügung, oder der Aufwand für die Berechnung der Inversen ist inakzeptabel, deshalb empfiehlt es sich, den Schritt

$$\mathbf{w}^{(m+1)} := \mathbf{A}^{-1}\mathbf{x}^{(m)}$$

des Originalverfahrens durch das Lösen des Gleichungssystems

$$\mathbf{A}\mathbf{w}^{(m+1)} = \mathbf{x}^{(m)}$$

zu ersetzen, das sich in vielen praktischen Anwendungen mit Hilfe einer Faktorisierung oder eines iterativen Lösungsverfahrens effizient durchführen lässt.

Algorithmus 5.27 (Inverse Iteration) *Seien* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *und* $\mathbf{x}^{(0)} \in \mathbb{K}^n$ *wie in Satz 5.25 gegeben. Dann berechnet der folgende Algorithmus die normierte Folge* $(\mathbf{x}^{(m)})_{m\in\mathbb{N}}$*, die gegen ein Vielfaches des ersten Eigenvektors konvergiert:*

$m \leftarrow 0$
$\mathbf{x}^{(m)} \leftarrow \mathbf{x}^{(m)}/\|\mathbf{x}^{(m)}\|$
while „Fehler zu groß“ do begin
  Löse $\mathbf{A}\mathbf{w}^{(m+1)} = \mathbf{x}^{(m)}$
  $\mathbf{x}^{(m+1)} \leftarrow \mathbf{w}^{(m+1)}/\|\mathbf{w}^{(m+1)}\|$
  $m \leftarrow m + 1$
end

Auch dieser Algorithmus lässt sich wieder mit dem Rayleigh-Quotienten kombinieren, um eine Approximation des betragskleinsten Eigenwerts und damit auch eine Schätzung des Iterationsfehlers zu gewinnen. Wir können den im Zuge der Iteration ohnehin berechneten Vektor $\mathbf{w}^{(m+1)}$ verwenden, um $\lambda_m := \Lambda_{A^{-1}}(\mathbf{x}^{(m)})$ effizient zu berechnen und

$$\epsilon_m := \|\mathbf{A}^{-1}\mathbf{x}^{(m)} - \lambda_m\mathbf{x}^{(m)}\| = \|\mathbf{w}^{(m+1)} - \langle\mathbf{x}^{(m)}, \mathbf{w}^{(m+1)}\rangle\mathbf{x}^{(m)}\|$$

als Maß des Fehlers benutzen. Dabei sollte man natürlich nicht vergessen, dass λ_m nun gegen den *Kehrwert* des kleinsten Eigenwerts konvergieren wird, nicht mehr gegen den Eigenwert selbst.

Bei der Betrachtung der inversen Iteration stellen sich zwei Fragen:

1. Die Forderung nach der Regularität von $\mathbf{A}$ ist im Kontext von Eigenwertverfahren unerwartet, schließlich kam sie bei keiner der bisherigen theoretischen Aussagen vor und schränkt die Anwendbarkeit der inversen Iteration ein. Lässt sie sich vermeiden?

2. Mit der Vektoriteration und der inversen Iteration stehen Verfahren zur Bestimmung der größten und kleinsten Eigenwerte zur Verfügung. Lassen sich auch „mittlere“ Eigenwerte berechnen?

Beide Fragen lassen sich positiv beantworten, wenn man von der Inversen $\mathbf{A}^{-1}$ von $\mathbf{A}$ zu der einer um einen gewissen Betrag $\mu \in \mathbb{K}$ „verschobenen“ Matrix übergeht, also $\mathbf{A}^{-1}$ durch $(\mathbf{A} - \mu\mathbf{I})^{-1}$ ersetzt.

Falls $\mu \notin \sigma(\mathbf{A})$ gilt, folgt für $\lambda \in \mathbb{K}, x \in \mathbb{K}^n \setminus \{0\}$ die Äquivalenz

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \iff (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{x} = \frac{1}{\lambda - \mu}\mathbf{x},$$

so dass alle Eigenvektoren von $\mathbf{A}$ auch Eigenvektoren von $(\mathbf{A} - \mu\mathbf{I})^{-1}$ sind. Man kann einfach nachweisen, dass jeder Eigenwert der letzteren Matrix sich als $1/(\lambda - \mu)$ mit einem $\lambda \in \sigma(\mathbf{A})$ darstellen lässt, wir gewinnen oder verlieren also keine Eigenwerte, ähnlich wie bei der inversen Iteration.

Die Konvergenz der inversen Iteration hängt von dem Verhältnis des vom Betrag her kleinsten zum vom Betrag her zweitkleinsten Eigenwerts ab. Verwendet man die Matrix $(\mathbf{A}-\mu\mathbf{I})^{-1}$ anstelle der Matrix $\mathbf{A}^{-1}$, so ist der vom Betrag her größte Eigenwert derjenige, der dem Wert μ am nächsten liegt. Durch die Wahl des sogenannten *Shift-Parameters* μ lässt sich also festlegen, welche Eigenwerte und Eigenvektoren berechnet werden sollen.

Für die *inverse Iteration mit Shift* μ definieren wir die Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$ der Iterierten durch

$$\mathbf{x}^{(m)} := (\mathbf{A} - \mu\mathbf{I})^{-m}\mathbf{x}^{(0)} = (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{x}^{(m-1)} \qquad \text{für alle } m \in \mathbb{N}. \tag{5.10}$$

Auch in diesem Fall lässt sich Satz 5.4 anwenden, um eine Konvergenaussage zu erhalten. Sei $\mu \in \mathbb{K} \setminus \sigma(\mathbf{A})$. Wie gehabt soll $\mathbf{A}$ normal sein, so dass wir mit Folgerung 3.47 eine unitäre Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und eine Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ mit

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D}$$

finden. Dank Bemerkung 3.40 können wir diesmal die Eigenwerte so anordnen, dass

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad |\lambda_1 - \mu| \le |\lambda_2 - \mu| \le \cdots \le |\lambda_n - \mu| \tag{5.11}$$

gilt. Daraus folgt nun

$$\left|\frac{1}{\lambda_1 - \mu}\right| \ge \left|\frac{1}{\lambda_2 - \mu}\right| \ge \cdots \ge \left|\frac{1}{\lambda_n - \mu}\right|.$$

Mit $\mathbf{e} := \mathbf{Q}\delta^{(1)}$ bezeichnen wir wieder einen Eigenvektor, der zu dem Eigenwert λ_1 der Matrix $\mathbf{A}$ gehört. Dann erhalten wir die folgende Konvergenzaussage:

Satz 5.28 (Konvergenz) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine normale Matrix, sei $\mu \in \mathbb{K} \setminus \sigma(\mathbf{A})$. Die Matrix $\mathbf{A}$ besitzt eine Schurzerlegung $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ mit einer unitären Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und einer Diagonalmatrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ der Form (5.11).*

Dann ist $\mathbf{e} := \mathbf{Q}\delta^{(1)}$ ein Eigenvektor zu dem Eigenwert λ_1, der μ am nächsten liegt.

Sei $\mathbf{x}^{(0)} \in \mathbb{K}^n$ mit $\langle \mathbf{e}, \mathbf{x}^{(0)}\rangle \neq 0$ gegeben. Dann gilt

$$\tan\angle(\mathbf{e}, \mathbf{x}^{(m)}) \le \left(\frac{|\lambda_1 - \mu|}{|\lambda_2 - \mu|}\right)^m \tan\angle(\mathbf{e}, \mathbf{x}^{(0)}) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir setzen $\widehat{\mathbf{A}} := (\mathbf{A} - \mu\mathbf{I})^{-1}$. Es gilt

$$\mathbf{Q}^*\widehat{\mathbf{A}}\mathbf{Q} = \mathbf{Q}^*(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{Q} = (\mathbf{D} - \mu\mathbf{I})^{-1} = \begin{pmatrix} 1/(\lambda_1 - \mu) & & \\ & \ddots & \\ & & 1/(\lambda_n - \mu) \end{pmatrix},$$

also ist $\widehat{\mathbf{A}}$ diagonalisierbar und die Eigenwerte von $(\mathbf{D}-\mu\mathbf{I})^{-1}$ sind dem Betrag nach absteigend sortiert. Also können wir den Satz 5.4 anwenden, um die gesuchte Abschätzung zu erhalten. ∎

Die Folge der normierten Iterierten lässt sich einfach mit Hilfe des folgenden Algorithmus berechnen:

Algorithmus 5.29 (Inverse Iteration mit Shift) *Seien $\mathbf{A} \in \mathbb{K}^{n\times n}$ und $\mathbf{x}^{(0)} \in \mathbb{K}^n$ wie in Satz 5.28 gegeben. Dann berechnet der folgende Algorithmus die normierte Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}}$, die gegen ein Vielfaches des ersten Eigenvektors konvergiert:*

$m \leftarrow 0$
$\mathbf{x}^{(0)} \leftarrow \mathbf{x}^{(0)}/\|\mathbf{x}^{(0)}\|$
while „Fehler zu groß“ **do begin**
  Löse $(\mathbf{A} - \mu\mathbf{I})\mathbf{w}^{(m+1)} = \mathbf{x}^{(m)}$
$\mathbf{x}^{(m+1)} \leftarrow \mathbf{w}^{(m+1)}/\|\mathbf{w}^{(m+1)}\|$
$m \leftarrow m + 1$
end

Der Shift-Parameter ermöglicht es uns nicht nur, beliebige Eigenwerte zu approximieren, er kann bei geschickter Wahl auch zu einer erheblichen Beschleunigung der Konvergenz führen, wie die folgenden Beispiele zeigen:

Beispiel 5.30 (Trennung von Eigenwerten) *Wir untersuchen die Matrix*

$$\mathbf{A}_\epsilon := \begin{pmatrix} 1 & 0 \\ 0 & 1+\epsilon \end{pmatrix}$$

für $\epsilon \in \mathbb{R}_{>0}$. Offenbar gilt $\sigma(\mathbf{A}_\epsilon) = \{1, 1+\epsilon\}$. Wendet man die inverse Iteration auf diese Matrix und einen Startvektor aus $(\mathbb{R}\setminus\{0\})\times(\mathbb{R}\setminus\{0\})$ an, so konvergiert sie gemäß Satz 5.25 mit einer Geschwindigkeit von $1/(1+\epsilon)$. Falls ϵ klein ist, erhalten wir sehr langsame Konvergenz.

Wendet man die inverse Iteration mit einem Shift von $\mu = 1+\epsilon/3$ auf die Matrix und den Startvektor an, so konvergiert sie mit einer Geschwindigkeit von

$$\left|\frac{1-\mu}{1+\epsilon-\mu}\right| = \left|\frac{-\epsilon/3}{2\epsilon/3}\right| = 1/2.$$

Durch geschickte Wahl des Shift-Parameters lässt sich also auch bei nahe beieinander liegenden (auch als „schlecht separiert“ bezeichneten) Eigenwerten eine gute Konvergenzrate erzielen.

Betrachten wir als nächstes die Matrix

$$\mathbf{B} := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Ihr charakteristisches Polynom ist $p_B(\lambda) = \lambda^2 + 1$*, also folgt* $\sigma(\mathbf{B}) = \{i, -i\}$*. Wegen* $|i| = |-i| = 1$ *ist nicht zu erwarten, dass die Vektoriteration oder die inverse Iteration bei dieser Matrix konvergieren. Selbst in diesem Fall lässt sich durch Verwendung eines Shift-Wertes von* $\mu = i/3$ *noch die gute Konvergenzrate von* $1/2$ *erzielen:*

$$\left|\frac{i-\mu}{-i-\mu}\right| = \left|\frac{2i/3}{-4i/3}\right| = 1/2.$$

Zum Abschluß dieses Abschnittes sollen noch einmal die Vor- und Nachteile der einzelnen vorgestellten Verfahren zusammengefasst werden:

1. Die *Vektoriteration* erfordert keine Inversion der Matrix $\mathbf{A}$ und konvergiert unter den in Satz 5.4 angegebenen Bedingungen gegen einen Eigenvektor zu dem betragsgrößten Eigenwert. Die Konvergenz ist linear und hängt vom Verhältnis zwischen den Eigenwerten mit größtem und zweitgrößtem Betrag ab.

2. Die *inverse Iteration* erfordert das wiederholte Lösen eines Gleichungssystems (potentiell zeitaufwendig) und konvergiert unter den in Satz 5.25 angegebenen Bedingungen gegen einen Eigenvektor zu dem betragskleinsten Eigenwert. Die Konvergenz ist linear und hängt vom Verhältnis zwischen den Eigenwerten mit kleinstem und zweitkleinstem Betrag ab.

3. Die *inverse Iteration mit Shift* erfordert das wiederholte Lösen eines Gleichungssystems und konvergiert gegen den Eigenvektor, dessen Eigenwert dem Shift am nächsten liegt. Die Konvergenz ist linear und hängt von dem Verhältnis der Abstände der Eigenwerte zu μ ab.

5.4 Inverse Iteration mit Rayleigh-Shift

Die geschickte Wahl des Shift-Parameters μ ist offensichtlich von großer Bedeutung für die Geschwindigkeit des Verfahrens. Da die Konvergenz desto besser wird, je näher μ an dem gesuchten Eigenwert liegt (beziehungsweise desto weiter es von allen anderen entfernt ist), sind wir daran interessiert, μ als Approximation des uns interessierenden Eigenwerts zu wählen. Lemma 5.11 legt die Idee nahe, für die Berechnung von μ den Rayleigh-Quotienten zu verwenden.

Wenn wir davon ausgehen, dass $\mathbf{x}^{(0)}$ bereits eine relativ gute Approximation eines Eigenvektors $\mathbf{e}$ zu dem Eigenwert λ_1 ist, wird nach Lemma 5.11 die Zahl $\mu_0 := \Lambda_A(\mathbf{x}^{(0)})$ eine gute Approximation von λ_1 sein. Wenn wir nun μ_0 als Shift verwenden und

$$\mathbf{x}^{(1)} := (\mathbf{A} - \mu_0 \mathbf{I})^{-1} \mathbf{x}^{(0)}$$

berechnen, erhalten wir nach Satz 5.28 eine Abschätzung der Form

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(1)}) \leq \frac{|\lambda_1 - \mu_0|}{|\lambda_2 - \mu_0|} \tan \angle(\mathbf{e}, \mathbf{x}^{(0)}),$$

wobei die Eigenwerte wieder in der Form

$$|\lambda_1 - \mu_0| \leq |\lambda_2 - \mu_0| \leq \cdots |\lambda_n - \mu_0|$$

angeordnet sind und

$$\mathbf{Q}^* \mathbf{A} \mathbf{Q} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{5.12}$$

gelten soll. Aus Lemma 5.11 erhalten wir eine Abschätzung für $|\lambda_1 - \mu_0|$, so dass wir

$$\begin{aligned} \tan \angle(\mathbf{e}, \mathbf{x}^{(1)}) &\leq \frac{\|\mathbf{A} - \lambda_1 \mathbf{I}\|}{|\lambda_2 - \mu_0|} \sin \angle(\mathbf{e}, \mathbf{x}^{(0)}) \tan \angle(\mathbf{e}, \mathbf{x}^{(0)}) \\ &\leq \frac{\|\mathbf{A} - \lambda_1 \mathbf{I}\|}{|\lambda_2 - \mu_0|} \tan^2 \angle(\mathbf{e}, \mathbf{x}^{(0)}) \end{aligned} \tag{5.13}$$

bewiesen haben. Diese Formel suggeriert, dass der Fehler der neuen Iterierten $\mathbf{x}^{(1)}$ sich wie das Quadrat des Fehlers der alten Iterierten $\mathbf{x}^{(0)}$ verhält, dass wir also auf *quadratische Konvergenz* hoffen dürfen. Das würde bedeuten, dass das Verfahren sehr schnell konvergiert, sobald $\mathbf{x}^{(0)}$ dem gesuchten Eigenraum hinreichend nahe ist.

Ausgehend von $\mathbf{x}^{(1)}$ können wir dann einen neuen Shift-Parameter $\mu_1 := \Lambda_A(\mathbf{x}^{(1)})$ bestimmen und den Vorgang wiederholen, um eine weitere Näherung $\mathbf{x}^{(2)}$ zu gewinnen. Da die Berechnung des Rayleigh-Quotienten nichtlinear ist, wird die so definierte *inverse Iteration mit Rayleigh-Shift*, anders als die Vektoriteration oder die konventionelle inverse Iteration, ein nichtlineares Verfahren für die Approximation des Eigenvektors sein.

Algorithmus 5.31 (Inverse Iteration mit Rayleigh-Shift) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine normale Matrix, und sei $\mathbf{x}^{(0)}$ eine hinreichend gute Approximation des Eigenvektors $\mathbf{e} := \mathbf{Q}\delta^{(1)}$. Dann berechnet der folgende Algorithmus die Folge $(\mathbf{x}^{(m)})_{m \in \mathbb{N}}$, die gegen ein Vielfaches dieses Vektors konvergiert:*

$m \leftarrow 0$
$\mathbf{x}^{(m)} \leftarrow \mathbf{x}^{(m)} / \|\mathbf{x}^{(m)}\|_2$
while „*Fehler zu groß*“ *do begin*
  $\mu_m \leftarrow \langle \mathbf{x}^{(m)}, \mathbf{A}\mathbf{x}^{(m)} \rangle_2$
  Löse $(\mathbf{A} - \mu_m \mathbf{I})\mathbf{w}^{(m+1)} = \mathbf{x}^{(m)}$
  $\mathbf{x}^{(m+1)} \leftarrow \mathbf{w}^{(m+1)} / \|\mathbf{w}^{(m+1)}\|_2$
  $m \leftarrow m + 1$
end

Auf den ersten Blick ist die inverse Iteration mit Rayleigh-Shift nicht viel aufwendiger als die inverse Iteration mit konstantem Shift. In praktischen Implementierungen ist das leider häufig nicht der Fall:

Bemerkung 5.32 (Rechenaufwand) *In der Praxis werden die Gleichungssysteme*

$$(\mathbf{A} - \mu\mathbf{I})\mathbf{w}^{(m+1)} = \mathbf{x}^{(m)}, \qquad (\mathbf{A} - \mu_m\mathbf{I})\mathbf{w}^{(m+1)} = \mathbf{x}^{(m)},$$

die bei der inversen Iteration mit konstantem bzw. mit Rayleigh-Shift auftreten, häufig mit Hilfe einer Faktorisierung gelöst, etwa mit einer LR-Faktorisierung mit Pivotsuche.

Im Falle eines konstanten Shift-Parameters kann die Berechnung der Faktorisierung vor dem Eintritt in die zentrale Schleife der inversen Iteration stattfinden, so dass ein Iterationsschritt lediglich das Vorwärts- und Rückwärtseinsetzen in die bereits berechneten Faktoren erfordert.

Für den Rayleigh-Shift dagegen ändert sich die Matrix potentiell in jedem Schritt, so dass für jede Iterierte die Faktorisierung erneut berechnet werden muss. Deshalb kann die inverse Iteration mit Rayleigh-Shift unter Umständen wesentlich *aufwendiger als die Variante mit konstantem Shift werden.*

Ein einfacher Ausweg besteht darin, den Shift-Parameter nicht in jedem Schritt zu aktualisieren, sondern in größeren Abständen, um die Anzahl der Faktorisierungen zu reduzieren. Dann verliert man zwar potentiell die quadratische Konvergenz, gewinnt aber ein schnelleres Verfahren.

Um aus der Ungleichung (5.13) eine quadratische Konvergenzaussage zu gewinnen, müssen wir den Nenner $|\lambda_2 - \mu_0|$ durch eine von m unabhängige Konstante abschätzen. Das gelingt uns, falls wir voraussetzen, dass μ_0 hinreichend nahe an λ_1 liegt, und das folgt, falls der Winkel zwischen $\mathbf{x}^{(0)}$ und dem gesuchten Eigenvektor $\mathbf{e}$ hinreichend klein ist.

Satz 5.33 (Quadratische Konvergenz) *Sei $\delta \in [0, 1)$. Sei $\mathbf{x}^{(0)} \in \mathbb{K}^n$ mit*

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(0)}) \leq \delta \frac{|\lambda_2 - \lambda_1|}{\|\mathbf{A} - \lambda_1\mathbf{I}\|}$$

gegeben, und sei $\mu_0 := \Lambda_A(\mathbf{x}^{(0)})$. Für den Vektor $\mathbf{x}^{(1)} := (\mathbf{A} - \mu_0\mathbf{I})^{-1}\mathbf{x}^{(0)}$ gilt dann

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(1)}) \leq \frac{\|\mathbf{A} - \lambda_1\mathbf{I}\|}{(1 - \delta)|\lambda_2 - \lambda_1|} \tan^2 \angle(\mathbf{e}, \mathbf{x}^{(0)}).$$

Dann gilt auch $\delta < 1$ und wir erhalten insbesondere

$$\tan \angle(\mathbf{e}, \mathbf{x}^{(1)}) \leq \delta \tan \angle(\mathbf{e}, \mathbf{x}^{(0)}),$$

so dass auch $\mathbf{x}^{(1)}$ die Voraussetzungen des Satzes erfüllt und die Rayleigh-Iteration konvergiert.

Beweis. Den größten Teil der Arbeit haben wir bereits in (5.13) geleistet, wir müssen lediglich noch den Nenner abschätzen. Mit der umgekehrten Dreiecksungleichung erhalten wir

$$|\lambda_2 - \mu_0| \geq |\lambda_2 - \lambda_1| - |\lambda_1 - \mu_0|,$$

und mit Lemma 5.11 und der Voraussetzung folgt

$$|\lambda_1 - \mu_0| \leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \sin\angle(\mathbf{e}, \mathbf{x}^{(0)}) \leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \tan\angle(\mathbf{e}, \mathbf{x}^{(0)}) \leq \delta|\lambda_2 - \lambda_1|,$$

so dass wir insgesamt zu

$$|\lambda_2 - \mu_0| \geq |\lambda_2 - \lambda_1| - \delta|\lambda_2 - \lambda_1| = (1-\delta)|\lambda_2 - \lambda_1|$$

gelangen. Es bleibt noch $\delta < 1$ zu zeigen. Dazu setzen wir $\mathbf{y} := \mathbf{Q}\delta^{(2)}$ und halten

$$\begin{aligned}\|\mathbf{A} - \lambda_1\mathbf{I}\| &\geq \frac{\|(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{y}\|}{\|\mathbf{y}\|} = \frac{\|\mathbf{Q}^*(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{Q}\delta^{(2)}\|}{\|\mathbf{Q}\delta^{(2)}\|} = \frac{\|(\mathbf{D} - \lambda_1\mathbf{I})\delta^{(2)}\|}{\|\delta^{(2)}\|} \\ &= \|(\lambda_2 - \lambda_1)\delta^{(2)}\| = |\lambda_2 - \lambda_1|\end{aligned}$$

fest. Es folgt

$$\frac{|\lambda_2 - \lambda_1|}{\|\mathbf{A} - \lambda_1\mathbf{I}\|} \leq \frac{|\lambda_2 - \lambda_1|}{|\lambda_2 - \lambda_1|} = 1,$$

so dass sich aus unsere Voraussetzung auch $\delta < 1$ ergibt. ∎

Diese Aussage gilt auch noch für allgemeine diagonalisierbare Matrizen, wenn man die Konditionszahl der diagonalisierenden Ähnlichkeitstransformation an geeigneter Stelle berücksichtigt.

Für normale Matrizen lässt sich sogar *kubische* Konvergenz beweisen:

Satz 5.34 (Kubische Konvergenz) *Sei* $\mathbf{A}$ *normal. Sei* $\delta \in [0, 1)$. *Sei* $\mathbf{x}^{(0)} \in \mathbb{K}^n$ *mit*

$$\tan^2\angle(\mathbf{z}^{(0)}, \mathbf{x}^{(1)}) \leq \delta\frac{|\lambda_2 - \lambda_1|}{\|\mathbf{A} - \lambda_1\mathbf{I}\|}$$

gegeben, und sei $\mu_0 := \Lambda_A(\mathbf{x}^{(0)})$. *Für den Vektor* $\mathbf{x}^{(1)} := (\mathbf{A} - \mu_0\mathbf{I})^{-1}\mathbf{x}^{(0)}$ *gilt dann*

$$\tan\angle(\mathbf{e}, \mathbf{x}^{(1)}) \leq \frac{\|\mathbf{A} - \lambda_1\mathbf{I}\|_2}{(1-\delta)|\lambda_2 - \lambda_1|}\tan^3\angle(\mathbf{e}, \mathbf{x}^{(0)}).$$

Insbesondere erfüllt $\mathbf{x}^{(1)}$ *wieder die Voraussetzung des Satzes, so dass die Rayleigh-Iteration konvergiert.*

Beweis. Nach Lemma 5.11 gilt

$$|\lambda_1 - \mu_0| \leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \sin^2\angle(\mathbf{e}, \mathbf{x}^{(0)}) \leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \tan^2\angle(\mathbf{e}, \mathbf{x}^{(0)}),$$

und Einsetzen in die Abschätzung des Satzes 5.28 ergibt

$$\tan\angle(\mathbf{e}, \mathbf{x}^{(1)}) \leq \frac{|\lambda_1 - \mu_0|}{|\lambda_2 - \mu_0|}\tan\angle(\mathbf{e}, \mathbf{x}^{(0)}) \leq \frac{\|\mathbf{A} - \lambda_1\mathbf{I}\|}{|\lambda_2 - \mu_0|}\tan^3\angle(\mathbf{e}, \mathbf{x}^{(0)}).$$

Mit der umgekehrten Dreiecksungleichung folgt

$$|\lambda_2 - \mu_0| \geq |\lambda_2 - \lambda_1| - |\lambda_1 - \mu_0| \geq |\lambda_2 - \lambda_1| - \|\mathbf{A} - \lambda_1\mathbf{I}\| \tan^2\angle(\mathbf{e}, \mathbf{x}^{(0)})$$

$$\geq |\lambda_2 - \lambda_1| - \delta|\lambda_1 - \lambda_1| = (1-\delta)|\lambda_2 - \lambda_1|.$$

Das gewünschte Resultat folgt durch Einsetzen in den Nenner der vorangehenden Ungleichung. ∎

In der Praxis dürfte es sinnvoll sein, die inverse Iteration mit konstantem Shift zu verwenden, um eine erste Näherung des Eigenvektors zu gewinnen, die genau genug ist, um dann unter Einsatz des Rayleigh-Shifts in wenigen Schritten eine hohe Genauigkeit zu erreichen. Das so zusammengesetzte Verfahren würde von der Iteration mit konstantem Shift die globale Konvergenz und den geringen Rechenaufwand pro Schritt erben und von der Iteration mit Rayleigh-Shift die schnelle lokale Konvergenz und deshalb hohe Genauigkeit.

5.5 Orthogonale Iteration

Die Vektoriteration und die von ihr abgeleiteten Verfahren dienen der Berechnung eines Eigenvektors zu einem bestimmten Eigenwert, der in geeigneter Weise dominant sein muss: Bei der einfachen Vektoriteration mussten wir voraussetzen, dass $|\lambda_1|$ echt größer als die Beträge aller anderen Eigenwerte ist. Insbesondere mussten wir dabei mehrfache Eigenwerte ausschließen, denn im Falle $|\lambda_1| = |\lambda_2|$ würde unser Konvergenzsatz 5.4 keine brauchbare Fehlerabschätzung mehr zur Verfügung stellen. Falls immerhin noch $\lambda_1 \neq \lambda_2$ gilt, können wir dieses Problem mit einem geeigneten Shift-Parameter beheben, im Falle $\lambda_1 = \lambda_2$ dagegen, also bei einem doppelten Eigenwert, erhalten wir mit der bisherigen Theorie keine Konvergenzaussage.

Wenn wir eine Diagonalmatrix $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit

$$|\lambda_1| = |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$$

näher untersuchen, stellen wir fest, dass in den Vektoren

$$\mathbf{x}^{(m)} := \frac{1}{|\lambda_1|^m}\mathbf{D}^m\mathbf{x}^{(0)} \qquad \text{für } m \in \mathbb{N}_0$$

immer noch alle Komponenten außer den *ersten beiden* gegen null konvergieren. Die Vektoren werden also zwar nicht gegen einen Vektor des Eigenwerts λ_1 konvergieren, aber immer noch gegen einen Vektor aus dem von Eigenvektoren zu λ_1 und λ_2 aufgespannten *invarianten Teilraum*.

Konvergenzaussagen wie Satz 5.4 basieren darauf, dass die Iterierte $\mathbf{x}^{(m)}$ mit einem geeigneten Vielfachen des ersten Eigenvektors $\mathbf{e}$ verglichen wird. Wenn wir die Konvergenz gegen einen Teilraum untersuchen wollen, liegt es also nahe $\mathbf{x}^{(m)}$ mit einem Vektor aus dem Teilraum zu vergleichen, der ihm möglichst nahe liegt.

Ein einfacher Zugang besteht darin, die Approximation einer Iterierten durch eine Matrix $\mathbf{P}$ zu beschreiben, deren Bild der gewünschte Teilraum ist. Dann vergleichen wir die Iterierte $\mathbf{x}^{(m)}$ mit dem Element $\mathbf{P}\mathbf{x}^{(m)}$ des Teilraums, und falls die Differenz klein ist, liegt $\mathbf{x}^{(m)}$ „fast“ im Teilraum. Besonders günstig ist es, wenn $\mathbf{P}$ eine Projektion auf den Teilraum ist, also $\mathbf{P}^2 = \mathbf{P}$ gilt. Für unsere Zwecke ideal unter derartigen Matrizen sind die *orthogonalen Projektionen*.

Definition 5.35 (Orthogonale Projektion) *Sei $\mathbf{P} \in \mathbb{K}^{n\times n}$. Falls $\mathbf{P}^2 = \mathbf{P} = \mathbf{P}^*$ gilt, nennen wir $\mathbf{P}$ eine* orthogonale Projektion.

Lemma 5.36 (Orthogonale Projektion) *Sei $\mathbf{P} \in \mathbb{K}^{n\times n}$ eine orthogonale Projektion. Dann gilt*

$$\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n,\ \mathbf{y} \in \operatorname{Bild}(\mathbf{P}). \tag{5.14a}$$

Daraus folgen

$$\|\mathbf{x}-\mathbf{P}\mathbf{x}\| \le \|\mathbf{x}-\mathbf{y}\| \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n,\ \mathbf{y} \in \operatorname{Bild}(\mathbf{P}), \tag{5.14b}$$

$$\|\mathbf{P}\mathbf{x}\| \le \|\mathbf{x}\| \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n, \tag{5.14c}$$

die Projektion bildet also jeden Vektor auf seine beste Approximation in ihrem Bild ab und vergrößert dabei nicht seine Norm.

Beweis. Seien $\mathbf{x} \in \mathbb{K}^n$ und $\mathbf{y} \in \operatorname{Bild}(\mathbf{P})$ gegeben. Dann existiert ein $\mathbf{z} \in \mathbb{K}^n$ mit $\mathbf{y} = \mathbf{P}\mathbf{z}$. Wir haben

$$\begin{aligned}
\|\mathbf{x}-\mathbf{y}\|^2 &= \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle \\
&= \langle (\mathbf{x}-\mathbf{P}\mathbf{x}) + (\mathbf{P}\mathbf{x}-\mathbf{y}), (\mathbf{x}-\mathbf{P}\mathbf{x}) + (\mathbf{P}\mathbf{x}-\mathbf{y})\rangle \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \langle \mathbf{x}-\mathbf{P}\mathbf{x}, \mathbf{P}\mathbf{x}-\mathbf{y}\rangle + \langle \mathbf{P}\mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{P}\mathbf{x}\rangle + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \langle \mathbf{x}-\mathbf{P}\mathbf{x}, \mathbf{P}(\mathbf{x}-\mathbf{z})\rangle + \langle \mathbf{P}(\mathbf{x}-\mathbf{z}), \mathbf{x}-\mathbf{P}\mathbf{x}\rangle + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \langle \mathbf{P}^*(\mathbf{x}-\mathbf{P}\mathbf{x}), \mathbf{x}-\mathbf{z}\rangle + \langle \mathbf{x}-\mathbf{z}, \mathbf{P}^*(\mathbf{x}-\mathbf{P}\mathbf{x})\rangle + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \langle \mathbf{P}(\mathbf{x}-\mathbf{P}\mathbf{x}), \mathbf{x}-\mathbf{z}\rangle + \langle \mathbf{x}-\mathbf{z}, \mathbf{P}(\mathbf{x}-\mathbf{P}\mathbf{x})\rangle + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \langle \mathbf{P}\mathbf{x}-\mathbf{P}^2\mathbf{x}, \mathbf{x}-\mathbf{z}\rangle + \langle \mathbf{x}-\mathbf{z}, \mathbf{P}\mathbf{x}-\mathbf{P}^2\mathbf{x}\rangle + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \langle \mathbf{0}, \mathbf{x}-\mathbf{z}\rangle + \langle \mathbf{x}-\mathbf{z}, \mathbf{0}\rangle + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2 \\
&= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \|\mathbf{P}\mathbf{x}-\mathbf{y}\|^2.
\end{aligned}$$

Daraus folgt unmittelbar (5.14b). Durch Einsetzen von $\mathbf{y} = \mathbf{0}$ folgt (5.14c). ∎

Lemma 5.37 (Existenz und Eindeutigkeit) *Sei $\mathcal{V} \subseteq \mathbb{K}^n$ ein Teilraum. Dann existiert genau eine orthogonale Projektion $\mathbf{P} \in \mathbb{K}^{n\times n}$ mit $\operatorname{Bild}(\mathbf{P}) = \mathcal{V}$. Wir nennen sie* die orthogonale Projektion auf $\mathcal{V}$.

Beweis. Zunächst zeigen wir, dass eine orthogonale Projektion existiert. Falls $\mathcal{V} = \{\mathbf{0}\}$ gilt, setzen wir $\mathbf{P} = \mathbf{0}$ und sind fertig.

Anderenfalls sei $k \in [1:n]$ die Dimension des Raums $\mathcal{V}$. Indem wir eine Basis $(a_i)_{i=1}^k$ des Raums wählen und ihre Elemente als Spalten einer Matrix $\mathbf{A} \in \mathbb{K}^{n\times k}$ verwenden, erhalten wir eine injektive Matrix mit $\operatorname{Bild}(\mathbf{A}) = \mathcal{V}$.

Nach Lemma 3.22 ist $\mathbf{A}^*\mathbf{A}$ positiv definit, also insbesondere invertierbar. Wir definieren

$$\mathbf{P} := \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$$

und stellen mit Lemma 3.17 fest, dass $\mathbf{P} = \mathbf{P}^*$ gilt. Es gilt auch

$$\mathbf{P}^2 = \mathbf{A}\underbrace{(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}}_{=\mathbf{I}}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{P},$$

also haben wir eine orthogonale Projektion gefunden.

Sei nun $\widehat{\mathbf{P}} \in \mathbb{K}^{n\times n}$ eine weitere orthogonale Projektion auf $\mathcal{V}$. Sei $\mathbf{x} \in \mathbb{K}^n$. Wir setzen

$$\mathbf{y} := \mathbf{P}\mathbf{x}, \qquad \widehat{\mathbf{y}} := \widehat{\mathbf{P}}\mathbf{x}$$

und erhalten mit Lemma 5.36

$$\|\mathbf{x} - \widehat{\mathbf{y}}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|^2,$$
$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \mathbf{y}\|^2.$$

Indem wir die erste Gleichung in die zweite einsetzen, folgt

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 + 2\|\mathbf{y} - \widehat{\mathbf{y}}\|^2,$$

also muss $\mathbf{P}\mathbf{x} = \mathbf{y} = \widehat{\mathbf{y}} = \widehat{\mathbf{P}}\mathbf{x}$ gelten. Da wir diese Gleichung für beliebige $\mathbf{x} \in \mathbb{K}^n$ bewiesen haben, folgt $\mathbf{P} = \widehat{\mathbf{P}}$. ∎

Lemma 5.38 (Konvergenz für diagonale Matrizen) *Sei $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ gegeben und gelte*

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_k| > |\lambda_{k+1}| \geq \cdots \geq |\lambda_n| \tag{5.15}$$

für ein $k \in [1 : n-1]$. Sei $\mathbf{P} \in \mathbb{K}^{n\times n}$ die orthogonale Projektion auf den Teilraum $\mathbb{K}^k \times \{\mathbf{0}\}$, der von den ersten k Eigenvektoren der Matrix $\mathbf{D}$ aufgespannt wird.

Sei $\mathbf{x}^{(0)} \in \mathbb{K}^n$ mit $\mathbf{P}\mathbf{x} \neq \mathbf{0}$ gegeben. Dann gilt für die Iterierten

$$\mathbf{x}^{(m)} := \mathbf{D}^m\mathbf{x}^{(0)} \qquad \text{für alle } m \in \mathbb{N}.$$

die Abschätzung

$$\frac{\|\mathbf{x}^{(m)} - \mathbf{P}\mathbf{x}^{(m)}\|_2}{\|\mathbf{P}\mathbf{x}^{(m)}\|_2} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|\mathbf{x}^{(0)} - \mathbf{P}\mathbf{x}^{(0)}\|_2}{\|\mathbf{P}\mathbf{x}^{(0)}\|_2} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweis. Sei $\mathbf{I}_k \in \mathbb{K}^{k\times k}$ die Identität auf $\mathbb{K}^k$, und sei

$$\mathbf{P} = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{K}^{n\times n}$$

ihre Fortsetzung durch null. Dann gilt

$$\mathbf{P}\mathbf{z}^{(m)} = \begin{pmatrix} \lambda_1^m z_1^{(0)} \\ \cdots \\ \lambda_k^m z_k^{(0)} \\ 0 \\ \cdots \\ 0 \end{pmatrix}, \qquad \mathbf{z}^{(m)} - \mathbf{P}\mathbf{z}^{(m)} = \begin{pmatrix} 0 \\ \cdots \\ 0 \\ \lambda_{k+1}^m z_{k+1}^{(0)} \\ \cdots \\ \lambda_n^m z_n^{(0)} \end{pmatrix} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Für die Normen erhalten wir

$$\|\mathbf{z}^{(m)} - \mathbf{P}\mathbf{z}^{(m)}\|_2^2 = \sum_{i=k+1}^{n} |\lambda_i^m z_i^{(0)}|^2 \leq |\lambda_{k+1}|^{2m} \sum_{i=k+1}^{n} |z_i^{(0)}|^2 \leq |\lambda_{k+1}|^{2m} \|\mathbf{z}^{(0)} - \mathbf{P}\mathbf{z}^{(0)}\|_2^2,$$
$$\|\mathbf{P}\mathbf{z}^{(m)}\|_2^2 = \sum_{i=1}^{k} |\lambda_i^m z_i^{(0)}|^2 \geq |\lambda_k|^{2m} \sum_{i=1}^{k} |z_i^{(0)}|^2 = |\lambda_k|^{2m} \|\mathbf{P}\mathbf{z}^{(0)}\|_2^2,$$

also folgt

$$\frac{\|\mathbf{z}^{(m)} - \mathbf{P}\mathbf{z}^{(m)}\|_2}{\|\mathbf{P}\mathbf{z}^{(m)}\|_2} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|\mathbf{z}^{(0)} - \mathbf{P}\mathbf{z}^{(0)}\|_2}{\|\mathbf{P}\mathbf{z}^{(0)}\|_2} \qquad \text{für alle } m \in \mathbb{N}_0,$$

und das ist die zu zeigende Abschätzung. ∎

Auch dieses Ergebnis können wir durch einen Winkelbegriff ausdrücken: Analog zu Lemma 5.5 definieren wir für einen Vektor $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ und einen Teilraum $\mathcal{V} \subseteq \mathbb{K}^n$ den Winkel durch

$$\sin \angle(\mathbf{x}, \mathcal{V}) := \min \left\{ \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{x}\|} \;:\; \mathbf{y} \in \mathcal{V} \right\}.$$

Falls $\mathbf{P} \in \mathbb{K}^{n \times n}$ die orthogonale Projektion auf $\mathcal{V}$ ist, können wir mit Lemma 5.36 das Minimum explizit darstellen und erhalten

$$\sin \angle(\mathbf{x}, \mathcal{V}) = \frac{\|\mathbf{x} - \mathbf{P}\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Indem wir (5.14a) auf $\mathbf{y} = \mathbf{0}$ anwenden, erhalten wir $\|\mathbf{x}\|^2 = \|\mathbf{x} - \mathbf{P}\mathbf{x}\|^2 + \|\mathbf{P}\mathbf{x}\|^2$ und können den Cosinus des Winkels durch

$$\cos^2 \angle(\mathbf{x}, \mathcal{V}) = 1 - \sin^2 \angle(\mathbf{x}, \mathcal{V}) = \frac{\|\mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{P}\mathbf{x}\|^2}{\|\mathbf{x}\|^2} = \frac{\|\mathbf{P}\mathbf{x}\|^2}{\|\mathbf{x}\|^2}$$

einführen. Insgesamt haben wir

$$\sin \angle(\mathbf{x}, \mathcal{V}) = \frac{\|\mathbf{x} - \mathbf{P}\mathbf{x}\|}{\|\mathbf{x}\|}, \quad \cos \angle(\mathbf{x}, \mathcal{V}) = \frac{\|\mathbf{P}\mathbf{x}\|}{\|\mathbf{x}\|}, \quad \tan \angle(\mathbf{x}, \mathcal{V}) = \frac{\|\mathbf{x} - \mathbf{P}\mathbf{x}\|}{\|\mathbf{P}\mathbf{x}\|}. \tag{5.16a}$$

Die Aussage des Lemma 5.38 nimmt damit die uns vertraute Form

$$\tan \angle(\mathbf{x}^{(m)}, \mathcal{V}) \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \tan \angle(\mathbf{x}^{(0)}, \mathcal{V}) \qquad \text{für alle } m \in \mathbb{N}_0$$

an, wobei $\mathcal{V} = \mathbb{K}^k \times \{\mathbf{0}\} \subseteq \mathbb{K}^n$ der von den Eigenvektoren der ersten k Eigenwerte aufgespannte Teilraum ist.

Falls nicht zufällig $\lambda_1 = \ldots = \lambda_k$ gilt, erhalten wir in diesem Fall *keine* Konvergenz gegen einen Eigenraum, sondern nur gegen den invarianten Teilraum $\mathcal{V}$.

Unser Ziel ist es nun, wenigstens diesen Teilraum vollständig zu beschreiben, indem wir eine Basis konstruieren. Die Idee ist einfach: Wenn die Vektoriteration zu *einem*

Startvektor gegen *einen* Vektor aus dem Unterraum konvergiert, dann wird die Vektoriteration zu k Startvektoren gegen k Vektoren aus dem Unterraum konvergieren, und falls wir sicherstellen können, dass diese Vektoren linear unabhängig sind, erhalten wir eine Basis.

Zur Abkürzung der Notation fassen wir die Vektoren in einer Matrix zusammen: Die k Spalten einer Matrix $\mathbf{X}^{(m)} \in \mathbb{K}^{n\times k}$ interpretieren wir als die Iterierten von k simultan ausgeführten Vektoriterationen mit den k Spalten der Matrix $\mathbf{X}^{(0)} \in \mathbb{K}^{n\times k}$ als Startvektoren. Die Iterierten sind dann durch

$$\mathbf{X}^{(m)} := \mathbf{D}^m \mathbf{X}^{(0)} \qquad \text{für alle } m \in \mathbb{N} \tag{5.17}$$

definiert. Es stellt sich die Frage, unter welchen Bedingungen wir darauf hoffen dürfen, dass die Spalten von $\mathbf{X}^{(m)}$ linear unabhängig sind. Offenbar müssen dafür zumindest die Spalten von $\mathbf{X}^{(0)}$ linear unabhängig sein, die Matrix muss also vollen Rang besitzen. Das reicht allerdings noch nicht: Falls eine Linearkombination der Spalten in den Kern der Matrix $\mathbf{D}^m$ fallen sollte, würde $\mathbf{X}^{(m)}$ trotzdem keinen vollen Rang besitzen.

Dieser Fall lässt sich einfach ausschließen, indem wir die in Lemma 5.38 eingeführte Projektion $\mathbf{P}$ auf den Unterraum verwenden: Statt zu fordern, dass $\mathbf{X}^{(0)}$ vollen Rang hat, fordern wir diese Eigenschaft von $\mathbf{PX}^{(0)}$. Diese Voraussetzung ist ausreichend:

Lemma 5.39 (Basis) *Seien $\mathbf{D}, \mathbf{P}, \lambda_1, \ldots, \lambda_n$ wie in Lemma 5.38 gegeben. Sei $\mathbf{X}^{(0)} \in \mathbb{K}^{n\times k}$ so gegeben, dass $\mathbf{PX}^{(0)}$ vollen Rang hat.*

Dann haben für jedes $m \in \mathbb{N}_0$ auch die Matrizen $\mathbf{PX}^{(m)}$ und $\mathbf{X}^{(m)}$ vollen Rang.

Beweis. Sei $m \in \mathbb{N}_0$. Dazu führen wir

$$\widehat{\mathbf{D}} := \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix}, \qquad \widehat{\mathbf{X}}^{(m)} := \begin{pmatrix} x_{11}^{(m)} & \ldots & x_{1k}^{(m)} \\ \vdots & \ddots & \vdots \\ x_{k1}^{(m)} & \ldots & x_{kk}^{(m)} \end{pmatrix}$$

ein und können $\mathbf{P}^2 = \mathbf{P}$ sowie $\mathbf{PD} = \mathbf{DP}$ ausnutzen, um

$$\mathbf{PX}^{(m)} = \mathbf{PD}^m\mathbf{X}^{(0)} = \mathbf{P}^2\mathbf{D}^m\mathbf{X}^{(0)} = \mathbf{PD}^m\mathbf{PX}^{(0)} = \begin{pmatrix} \widehat{\mathbf{D}}^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{X}}^{(0)} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{D}}^m\widehat{\mathbf{X}}^{(0)} \\ \mathbf{0} \end{pmatrix}$$

zu erhalten. Nach Voraussetzung hat

$$\mathbf{PX}^{(0)} = \begin{pmatrix} \widehat{\mathbf{X}}^{(0)} \\ \mathbf{0} \end{pmatrix}$$

vollen Rang, also muss $\widehat{\mathbf{X}}^{(0)}$ regulär sein. Aus (5.15) folgt, dass $\widehat{\mathbf{D}}$ regulär ist, also muss auch $\widehat{\mathbf{D}}^m$ regulär sein, und somit auch $\widehat{\mathbf{D}}\widehat{\mathbf{X}}^{(0)}$. Damit ist bewiesen, dass $\mathbf{PX}^{(m)}$ vollen Rang hat. Die Dimension des Kerns beträgt also $n - k$, und da $\mathbf{P}$ eine quadratische Matrix ist, kann der Kern der Matrix $\mathbf{X}^{(m)}$ höchstens kleiner sein. Also muss auch $\mathbf{X}^{(m)}$ vollen Rang haben. ∎

Indem wir Lemma 5.38 und Lemma 5.39 kombinieren, können wir folgern, dass die Matrizen $\mathbf{X}^{(m)}$ gegen Basen des für uns interessanten invarianten Unterraums konvergieren werden.

Dieses Konvergenzverhalten können wir auch quantifizieren: Nach Lemma 3.26 bilden die Spalten einer Matrix $\mathbf{V} \in \mathbb{K}^{n\times k}$ genau dann eine Basis eines invarianten Unterraums, wenn es eine Matrix $\mathbf{\Lambda} \in \mathbb{K}^{k\times k}$ gibt, mit der $\mathbf{DV} = \mathbf{V\Lambda}$ gilt. Wenn $\mathbf{X}^{(m)}$ „fast“ eine solche Basis ist, sollte eine Eigenschaft der Form

$$\|\mathbf{DX}^{(m)} - \mathbf{X}^{(m)}\mathbf{\Lambda}\|_2 \leq \epsilon$$

für ein geeignetes $\epsilon \in \mathbb{R}_{>0}$ gelten. Wenn wir die Skalierung der Matrix $\mathbf{X}^{(m)}$ geeignet berücksichtigen, erhalten wir das folgende Resultat:

Lemma 5.40 (Konvergenz gegen Teilraum) *Seien $\mathbf{D}, \mathbf{P}, \lambda_1, \ldots, \lambda_n$ wie in Lemma 5.38 gegeben. Sei $\mathbf{X}^{(0)} \in \mathbb{K}^{n\times k}$ so gegeben, dass $\mathbf{PX}^{(0)}$ vollen Rang hat.*

Dann gibt es eine Matrix $\mathbf{\Lambda} \in \mathbb{K}^{k\times k}$ so, dass für alle $\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ und alle $m \in \mathbb{N}$ die Abschätzungen

$$\frac{\|(\mathbf{DX}^{(m)} - \mathbf{X}^{(m)}\mathbf{\Lambda})\mathbf{y}\|_2}{\|\mathbf{PX}^{(m)}\mathbf{y}\|_2} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{DX}^{(0)} - \mathbf{X}^{(0)}\mathbf{\Lambda})\mathbf{y}\|_2}{\|\mathbf{PX}^{(0)}\mathbf{y}\|_2},$$
$$\frac{\|(\mathbf{X}^{(m)} - \mathbf{PX}^{(m)})\mathbf{y}\|_2}{\|\mathbf{PX}^{(m)}\mathbf{y}\|_2} \leq \left(\frac{|\lambda_{m+1}|}{|\lambda_m|}\right)^m \frac{\|(\mathbf{X}^{(0)} - \mathbf{PX}^{(0)})\mathbf{y}\|_2}{\|\mathbf{PX}^{(0)}\mathbf{y}\|_2}$$

erfüllt sind.

Beweis. Wir definieren Hilfsmatrizen

$$\widehat{\mathbf{X}} := \begin{pmatrix} x_{11}^{(0)} & \ldots & x_{1k}^{(0)} \\ \vdots & \ddots & \vdots \\ x_{k1}^{(0)} & \ldots & x_{kk}^{(0)} \end{pmatrix}, \qquad \mathbf{X}_0 := \begin{pmatrix} x_{k+1,1}^{(0)} & \ldots & x_{k+1,k}^{(0)} \\ \vdots & \ddots & \vdots \\ x_{n,1}^{(0)} & \ldots & x_{n,k}^{(0)} \end{pmatrix},$$
$$\widehat{\mathbf{D}} := \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix}, \qquad \mathbf{D}_0 := \begin{pmatrix} \lambda_{k+1} & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

und erhalten

$$\mathbf{X}^{(0)} = \begin{pmatrix} \widehat{\mathbf{X}} \\ \mathbf{X}_0 \end{pmatrix}, \qquad \mathbf{PX}^{(0)} = \begin{pmatrix} \widehat{\mathbf{X}} \\ \mathbf{0} \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} \widehat{\mathbf{D}} & \\ & \mathbf{D}_0 \end{pmatrix}.$$

Nach Voraussetzung hat $\mathbf{PX}^{(0)}$ vollen Rang, also muss $\widehat{\mathbf{X}} \in \mathbb{K}^{k\times k}$ regulär sein. Wir definieren

$$\mathbf{\Lambda} := \widehat{\mathbf{X}}^{-1}\widehat{\mathbf{D}}\widehat{\mathbf{X}}$$

und erhalten

$$\widehat{\mathbf{X}}\mathbf{\Lambda} = \widehat{\mathbf{X}}\widehat{\mathbf{X}}^{-1}\widehat{\mathbf{D}}\widehat{\mathbf{X}} = \widehat{\mathbf{D}}\widehat{\mathbf{X}}.$$

Nun können wir die Terme unserer Abschätzung untersuchen. Sei $\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ und $m \in \mathbb{N}$. Es gilt

$$\begin{aligned}
\|(\mathbf{D}\mathbf{X}^{(m)} - \mathbf{X}^{(m)}\mathbf{\Lambda})\mathbf{y}\|_2 &= \|\mathbf{D}^{m+1}\mathbf{X}^{(0)}\mathbf{y} - \mathbf{D}^m\mathbf{X}^{(0)}\mathbf{\Lambda}\mathbf{y}\|_2 \\
&= \left\| \begin{pmatrix} (\widehat{\mathbf{D}}^{m+1}\widehat{\mathbf{X}} - \widehat{\mathbf{D}}^m\widehat{\mathbf{X}}\mathbf{\Lambda})\mathbf{y} \\ (\mathbf{D}_0^{m+1}\mathbf{X}_0 - \mathbf{D}_0^m\mathbf{X}_0\mathbf{\Lambda})\mathbf{y} \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} (\widehat{\mathbf{D}}^{m+1}\widehat{\mathbf{X}} - \widehat{\mathbf{D}}^{m+1}\widehat{\mathbf{X}})\mathbf{y} \\ (\mathbf{D}_0^{m+1}\mathbf{X}_0 - \mathbf{D}_0^m\mathbf{X}_0\mathbf{\Lambda})\mathbf{y} \end{pmatrix} \right\|_2 \\
&= \left\| \begin{pmatrix} \mathbf{0} \\ \mathbf{D}_0^m(\mathbf{D}_0\mathbf{X}_0 - \mathbf{X}_0\mathbf{\Lambda})\mathbf{y} \end{pmatrix} \right\|_2 \leq |\lambda_{k+1}|^m \|(\mathbf{D}_0\mathbf{X}_0 - \mathbf{X}_0\mathbf{\Lambda})\mathbf{y}\|_2 \\
&\leq |\lambda_{k+1}|^m \left\| \begin{pmatrix} (\widehat{\mathbf{D}}\widehat{\mathbf{X}} - \widehat{\mathbf{X}}\mathbf{\Lambda})\mathbf{y} \\ (\mathbf{D}_0\mathbf{X}_0 - \mathbf{X}_0\mathbf{\Lambda})\mathbf{y} \end{pmatrix} \right\|_2 = |\lambda_{k+1}|^m \|(\mathbf{D}\mathbf{X}^{(0)} - \mathbf{X}^{(0)}\mathbf{\Lambda})\mathbf{y}\|_2,
\end{aligned}$$

und wir erhalten außerdem

$$\|\mathbf{P}\mathbf{X}^{(m)}\mathbf{y}\|_2 = \left\| \begin{pmatrix} \widehat{\mathbf{D}}^m\widehat{\mathbf{X}}\mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\|_2 \geq |\lambda_k|^m \left\| \begin{pmatrix} \widehat{\mathbf{X}}\mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\|_2 = |\lambda_k|^m \|\mathbf{P}\mathbf{X}^{(0)}\mathbf{y}\|_2. \tag{5.18}$$

Einsetzen dieser Ungleichungen in den zu untersuchenden Bruch führt zu der ersten Abschätzung.

Für die zweite Abschätzung verwenden wir einfach

$$\begin{aligned}
\|(\mathbf{X}^{(m)} - \mathbf{P}\mathbf{X}^{(m)})\mathbf{y}\|_2 &= \left\| \begin{pmatrix} \widehat{\mathbf{D}}^m\widehat{\mathbf{X}} - \widehat{\mathbf{D}}^m\widehat{\mathbf{X}} \\ \mathbf{D}_0^m\mathbf{X}_0 \end{pmatrix} \mathbf{y} \right\| = \|\mathbf{D}_0^m\mathbf{X}_0\mathbf{y}\|_2 \leq |\lambda_{k+1}|^m \|\mathbf{X}_0\mathbf{y}\|_2 \\
&= |\lambda_{k+1}|^m \left\| \begin{pmatrix} \widehat{\mathbf{X}} - \widehat{\mathbf{X}} \\ \mathbf{X}_0 \end{pmatrix} \mathbf{y} \right\| = |\lambda_{k+1}|^m \|(\mathbf{X}^{(0)} - \mathbf{P}\mathbf{X}^{(0)})\mathbf{y}\|
\end{aligned}$$

in Kombination mit (5.18). ∎

Auch die Konvergenz von Teilräumen können wir durch Winkel ausdrücken: Für zwei nicht leere Teilräume $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{K}^n$ definieren wir

$$\begin{aligned}
\sin\angle(\mathcal{X}, \mathcal{Y}) &:= \max\{\sin\angle(\mathbf{x}, \mathcal{Y}) \ : \ \mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}\}, \\
\cos\angle(\mathcal{X}, \mathcal{Y}) &:= \min\{\cos\angle(\mathbf{x}, \mathcal{Y}) \ : \ \mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}\}, \\
\tan\angle(\mathcal{X}, \mathcal{Y}) &:= \max\{\tan\angle(\mathbf{x}, \mathcal{Y}) \ : \ \mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}\}.
\end{aligned}$$

Dann können wir die dritte Aussage des Lemmas 5.40 auch kurz in der Form

$$\tan\angle(\operatorname{Bild}\mathbf{X}^{(m)}, \mathcal{V}) \leq \left(\frac{|\lambda_{m+1}|}{|\lambda_m|}\right)^m \tan\angle(\operatorname{Bild}\mathbf{X}^{(0)}, \mathcal{V}) \qquad \text{für alle } m \in \mathbb{N}_0$$

schreiben.

Wie üblich können wir diese Konvergenzaussage auf den Fall normaler Matrizen $\mathbf{A} \in \mathbb{K}^{n\times n}$ übertragen, indem wir sie mit Hilfe der Folgerung 3.47 diagonalisieren.

Satz 5.41 (Konvergenz) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine normale Matrix. Sie besitzt eine Schurzerlegung $\mathbf{A} = \mathbf{QDQ}^*$ mit einer unitären Matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ und einer Diagonalmatrix*

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \qquad |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Sei $k \in [1 : n-1]$ so gegeben, dass $|\lambda_k| > |\lambda_{k+1}|$ gilt, sei $\mathcal{V} := \mathbf{Q}(\mathbb{K}^k \times \{\mathbf{0}\})$ der von den ersten k Eigenvektoren aufgespannte invariante Teilraum von $\mathbb{K}^n$, und sei $\mathbf{P} \in \mathbb{K}^{n\times n}$ die Projektion auf diesen Teilraum.

Sei $\mathbf{X}^{(0)} \in \mathbb{K}^{n\times k}$ so gegeben, dass die Matrix $\mathbf{PX}^{(0)}$ vollen Rang hat.

Dann existiert eine Matrix $\mathbf{\Lambda} \in \mathbb{K}^{k\times k}$ so, dass

$$\frac{\|(\mathbf{AX}^{(m)} - \mathbf{X}^{(m)}\mathbf{\Lambda})\mathbf{y}\|_2}{\|\mathbf{PX}^{(m)}\mathbf{y}\|_2} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{AX}^{(0)} - \mathbf{X}^{(0)}\mathbf{\Lambda})\mathbf{y}\|_2}{\|\mathbf{PX}^{(0)}\mathbf{y}\|_2}, \tag{5.19a}$$

$$\frac{\|(\mathbf{X}^{(m)} - \mathbf{PX}^{(m)})\mathbf{y}\|_2}{\|\mathbf{PX}^{(m)}\mathbf{y}\|_2} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{X}^{(0)} - \mathbf{PX}^{(0)})\mathbf{y}\|_2}{\|\mathbf{PX}^{(0)}\mathbf{y}\|_2} \tag{5.19b}$$

für alle $m \in \mathbb{N}_0$ und $\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ gelten. Insbesondere haben die Matrizen $\mathbf{PX}^{(m)}$ und $\mathbf{X}^{(m)}$ vollen Rang.

Beweis. Sei $\widehat{\mathbf{P}} \in \mathbb{K}^{n\times n}$ die Projektion aus Lemma 5.38. Durch Rücktransformation erhalten wir die orthogonale Projektion

$$\mathbf{P} := \mathbf{Q}\widehat{\mathbf{P}}\mathbf{Q}^*$$

auf $\mathcal{V}$. Wie schon im Beweis von Satz 5.4 definieren wir durch

$$\widehat{\mathbf{X}}^{(m)} := \mathbf{Q}^*\mathbf{X}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

die transformierte Folge der Iterierten, für die wegen

$$\widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(0)} = \mathbf{Q}^*\mathbf{PQ}\widehat{\mathbf{X}}^{(0)} = \mathbf{Q}^*\mathbf{PX}^{(0)}$$

und der Gleichung

$$\widehat{\mathbf{X}}^{(m)} = \mathbf{Q}^*\mathbf{X}^{(m)} = \mathbf{Q}^*\mathbf{A}^m\mathbf{X}^{(0)} = \mathbf{Q}^*\mathbf{A}^m\mathbf{Q}\widehat{\mathbf{X}}^{(0)} = \mathbf{D}^m\widehat{\mathbf{X}}^{(0)} \qquad \text{für alle } m \in \mathbb{N}_0$$

die Voraussetzungen von Lemma 5.40 erfüllt sind. Also existiert ein $\mathbf{\Lambda} \in \mathbb{K}^{k\times k}$ derart, dass die Abschätzungen

$$\frac{\|(\mathbf{D}\widehat{\mathbf{X}}^{(m)} - \widehat{\mathbf{X}}^{(m)}\mathbf{\Lambda})\mathbf{y}\|_2}{\|\widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(m)}\mathbf{y}\|_2} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{D}\widehat{\mathbf{X}}^{(0)} - \widehat{\mathbf{X}}^{(0)}\mathbf{\Lambda})\mathbf{y}\|_2}{\|\widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(0)}\mathbf{y}\|_2},$$

$$\frac{\|(\widehat{\mathbf{X}}^{(m)} - \widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(m)})\mathbf{y}\|}{\|\widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(m)}\mathbf{y}\|} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\widehat{\mathbf{X}}^{(0)} - \widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(0)})\mathbf{y}\|}{\|\widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(0)}\mathbf{y}\|}$$

für alle $\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ und alle $m \in \mathbb{N}$ gelten. Mit (3.13) erhalten wir

$$\begin{aligned}\|(\mathbf{D}\widehat{\mathbf{X}}^{(m)} - \widehat{\mathbf{X}}^{(m)}\mathbf{\Lambda})\mathbf{y}\| &= \|(\mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{X}^{(m)} - \mathbf{Q}\widehat{\mathbf{X}}^{(m)}\mathbf{\Lambda})\mathbf{y}\| = \|(\mathbf{A}\mathbf{X}^{(m)} - \mathbf{X}^{(m)}\mathbf{\Lambda})\mathbf{y}\|,\\ \|(\widehat{\mathbf{X}}^{(m)} - \widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(m)})\mathbf{y}\| &= \|(\mathbf{Q}\widehat{\mathbf{X}}^{(m)} - \mathbf{Q}\widehat{\mathbf{P}}\mathbf{Q}^*\mathbf{X}^{(m)})\mathbf{y}\| = \|(\mathbf{X}^{(m)} - \mathbf{P}\mathbf{X}^{(m)})\mathbf{y}\|,\\ \|\widehat{\mathbf{P}}\widehat{\mathbf{X}}^{(m)}\mathbf{y}\| &= \|\mathbf{Q}\widehat{\mathbf{P}}\mathbf{Q}^*\mathbf{X}^{(m)}\mathbf{y}\| = \|\mathbf{P}\mathbf{X}^{(m)}\mathbf{y}\|\end{aligned}$$

für alle $m \in \mathbb{N}_0$ und alle $\mathbf{y} \in \mathbb{K}^k$, und damit folgen die gewünschten Aussagen. ∎

Wie schon im Falle der konventionellen Vektoriteration kann auch bei dieser Iteration das Problem auftreten, dass es durch die wiederholte Multiplikation mit $\mathbf{A}$ zu Über- oder Unterläufen kommt. Sehr viel schlimmer ist allerdings, dass in der Regel alle Spalten der Matrizen $\mathbf{X}^{(m)}$ gegen Vielfache des Eigenvektors eines dominanten Eigenwerts konvergieren werden, so dass sie zwar theoretisch linear unabhängig bleiben, in der numerischen Praxis aber fast linear abhängig werden können.

Beide Probleme lassen sich wieder durch eine geschickte Normierung beheben: Da wir nur an dem invarianten Unterraum, also dem Bild von $\mathbf{Z}^{(m)}$, interessiert sind, können wir die Matrix fast beliebig skalieren und Linearkombinationen ihrer Spalten bilden, ohne diesen Unterraum zu verändern. Insbesondere können wir mit Hilfe der Gram-Schmidt-Orthonormalisierung dafür sorgen, dass die von den Spalten beschriebene Basis orthonormal ist, und damit insbesondere aus linear unabhängigen Einheitsvektoren besteht.

Da die Gram-Schmidt-Orthonormalisierung numerisch instabil sein kann, verwenden wir stattdessen allerdings die übliche QR-Zerlegung: Zu jeder Matrix $\mathbf{X}^{(m)} \in \mathbb{K}^{n\times k}$ existieren eine unitäre Matrix $\widehat{\mathbf{Q}}^{(m)} \in \mathbb{K}^{n\times n}$ und eine obere Dreiecksmatrix $\widehat{\mathbf{R}}^{(m)} \in \mathbb{K}^{n\times k}$ so, dass

$$\mathbf{X}^{(m)} = \widehat{\mathbf{Q}}^{(m)}\widehat{\mathbf{R}}^{(m)}$$

gilt. Wenn wir die ersten k Zeilen von $\widehat{\mathbf{R}}^{(m)}$ mit $\mathbf{R}^{(m)}$ sowie die ersten k Spalten von $\widehat{\mathbf{Q}}^{(m)}$ mit $\mathbf{Q}^{(m)}$ sowie die restlichen Spalten mit $\widetilde{\mathbf{Q}}^{(m)}$ bezeichnen, erhalten wir

$$\mathbf{X}^{(m)} = \begin{pmatrix}\mathbf{Q}^{(m)} & \widetilde{\mathbf{Q}}^{(m)}\end{pmatrix}\begin{pmatrix}\mathbf{R}^{(m)}\\ \mathbf{0}\end{pmatrix} = \mathbf{Q}^{(m)}\mathbf{R}^{(m)},$$

da $\widehat{\mathbf{R}}^{(m)}$ eine obere Dreiecksmatrix ist. Wir können also zu jedem $\mathbf{X}^{(m)} \in \mathbb{K}^{n\times k}$ eine orthogonale Matrix $\mathbf{Q}^{(m)} \in \mathbb{K}^{n\times k}$ und eine obere Dreiecksmatrix $\mathbf{R}^{(m)} \in \mathbb{K}^{k\times k}$ mit

$$\mathbf{X}^{(m)} = \mathbf{Q}^{(m)}\mathbf{R}^{(m)} \tag{5.20}$$

konstruieren. Die Spalten von $\mathbf{Q}^{(m)}$ beschreiben dann die gesuchte Orthonormalbasis.

Für derartige Matrizen vereinfacht sich die Aussage von Satz 5.41 wesentlich, wenn wir berücksichtigen, dass durch die Orthogonalisierung in jedem Schritt die Basis verändert wird.

Folgerung 5.42 (Orthogonale Iteration) *Unter den Voraussetzungen von Satz 5.41 und mit (5.20) finden wir eine Familie $(\widetilde{\mathbf{\Lambda}}^{(m)})_{m=0}^{\infty}$ in $\mathbb{K}^{k\times k}$ mit der folgenden Eigenschaft: Für jedes $m \in \mathbb{N}_0$ und jeden Vektor $\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ existiert ein Vektor $\mathbf{z} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ mit*

$$\frac{\|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)})\mathbf{y}\|}{\|\mathbf{P}\mathbf{Q}^{(m)}\mathbf{y}\|} \leq \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{A}\mathbf{Q}^{(0)} - \mathbf{Q}^{(0)}\widetilde{\mathbf{\Lambda}}^{(0)})\mathbf{z}\|}{\|\mathbf{P}\mathbf{Q}^{(0)}\mathbf{z}\|}.$$

Beweis. Sei $\mathbf{X}^{(0)} \in \mathbb{K}^{n\times k}$ mit $\operatorname{rank}(\mathbf{P}\mathbf{X}^{(0)}) = k$ gegeben. Da $\mathbf{X}^{(m)}$ nach Satz 5.41 für alle $m \in \mathbb{N}_0$ vollen Rang hat, müssen nach (5.20) auch die Matrizen $\mathbf{R}^{(m)}$ vollen Rang haben, also regulär sein. Damit ist

$$\widetilde{\mathbf{\Lambda}}^{(m)} := \mathbf{R}^{(m)}\mathbf{\Lambda}(\mathbf{R}^{(m)})^{-1} \qquad \text{für alle } m \in \mathbb{N}_0$$

mit der Matrix $\mathbf{\Lambda}$ aus demselben Satz wohldefiniert.

Seien nun $m \in \mathbb{N}$ und ein Vektor $\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}$ fixiert. Wir definieren

$$\widetilde{\mathbf{y}} := (\mathbf{R}^{(m)})^{-1}\mathbf{y}, \qquad \mathbf{z} := \mathbf{R}^{(0)}\widetilde{\mathbf{y}},$$

und erhalten mit (5.19a) und (5.20) die Abschätzung

$$\begin{aligned}
\frac{\|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)})\mathbf{y}\|_2}{\|\mathbf{P}\mathbf{Q}^{(m)}\mathbf{y}\|_2} &= \frac{\|(\mathbf{A}\mathbf{Q}^{(m)}\mathbf{R}^{(m)} - \mathbf{Q}^{(m)}\mathbf{R}^{(m)}\mathbf{\Lambda})(\mathbf{R}^{(m)})^{-1}\mathbf{y}\|_2}{\|\mathbf{P}\mathbf{Q}^{(m)}\mathbf{R}^{(m)}(\mathbf{R}^{(m)})^{-1}\mathbf{y}\|_2} \\
&= \frac{\|(\mathbf{A}\mathbf{X}^{(m)} - \mathbf{X}^{(m)}\mathbf{\Lambda})\widetilde{\mathbf{y}}\|_2}{\|\mathbf{P}\mathbf{X}^{(m)}\widetilde{\mathbf{y}}\|_2} \\
&\le \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{A}\mathbf{X}^{(0)} - \mathbf{X}^{(0)}\mathbf{\Lambda})\widetilde{\mathbf{y}}\|_2}{\|\|\mathbf{P}\mathbf{X}^{(0)}\widetilde{\mathbf{y}}\|_2} \\
&= \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{A}\mathbf{X}^{(0)}(\mathbf{R}^{(0)})^{-1} - \mathbf{X}^{(0)}(\mathbf{R}^{(0)})^{-1}\widetilde{\mathbf{\Lambda}}^{(0)})\mathbf{z}\|_2}{\|\|\mathbf{P}\mathbf{X}^{(0)}(\mathbf{R}^{(0)})^{-1}\mathbf{z}\|_2} \\
&= \left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \frac{\|(\mathbf{A}\mathbf{Q}^{(0)} - \mathbf{Q}^{(0)}\widetilde{\mathbf{\Lambda}}^{(0)})\mathbf{z}\|_2}{\|\|\mathbf{P}\mathbf{Q}^{(0)}\mathbf{z}\|_2}.
\end{aligned}$$

■

Bemerkung 5.43 (Konvergenz in der Spektralnorm) *Da $\mathbf{P} \in \mathbb{K}^{n\times n}$ eine orthogonale Projektion und $\mathbf{Q}^{(m)}$ isometrisch ist, erhalten wir*

$$\|\mathbf{P}\mathbf{Q}^{(m)}\mathbf{y}\| \le \|\mathbf{Q}^{(m)}\mathbf{y}\| = \|\mathbf{y}\|,$$

also

$$\|\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)}\| \le \sup\left\{\frac{\|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)})\mathbf{y}\|}{\|\mathbf{P}\mathbf{Q}^{(m)}\mathbf{y}\|} \;:\; \mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}\right\}.$$

Mit der Konstanten

$$C := \sup\left\{\frac{\|(\mathbf{A}\mathbf{Q}^{(0)} - \mathbf{Q}^{(0)}\widetilde{\mathbf{\Lambda}}^{(0)})\mathbf{y}\|}{\|\mathbf{P}\mathbf{Q}^{(0)}\mathbf{y}\|} \;:\; \mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}\right\}.$$

nimmt dann Folgerung 5.42 die kompakte Form

$$\|\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)}\| \le C\left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \qquad \textit{für alle } m \in \mathbb{N}_0$$

an, wir erhalten also Konvergenz gegen eine Basis eines invarianten Teilraums in der Spektralnorm.

Für die Praxis ist es nicht sinnvoll, die Matrizen $\mathbf{X}^{(m)}$ zu berechnen, da sie aus den erwähnten Gründen zunehmend schlecht konditioniert sein werden. Stattdessen wollen wir möglichst direkt mit den orthogonalen Matrizen $\mathbf{Q}^{(m)}$ arbeiten, bei denen diese Gefahr nicht besteht.

Wir verwenden (5.20), um

$$\mathbf{X}^{(m+1)} = \mathbf{A}\mathbf{X}^{(m)} = \mathbf{A}\mathbf{Q}^{(m)}\mathbf{R}^{(m)} \tag{5.21}$$

zu erhalten. Nun berechnen wir die infolge der Orthogonalität von $\mathbf{Q}^{(m)}$ hoffentlich gut konditionierte Matrix

$$\mathbf{W}^{(m+1)} := \mathbf{A}\mathbf{Q}^{(m)} \in \mathbb{K}^{n\times k}$$

und bestimmen ihre QR-Faktorisierung

$$\mathbf{W}^{(m+1)} = \mathbf{Q}^{(m+1)}\widetilde{\mathbf{R}}^{(m+1)}$$

mit einer geeigneten orthogonalen Matrix $\mathbf{Q}^{(m+1)} \in \mathbb{K}^{n\times k}$ sowie einer oberen Dreiecksmatrix $\widetilde{\mathbf{R}}^{(m+1)} \in \mathbb{K}^{k\times k}$. Wir setzen diese Zerlegung in die Gleichung (5.21) ein, um

$$\mathbf{X}^{(m+1)} = \mathbf{W}^{(m+1)}\mathbf{R}^{(m)} = \mathbf{Q}^{(m+1)}\widetilde{\mathbf{R}}^{(m+1)}\mathbf{R}^{(m)}$$

zu erhalten, und da $\widetilde{\mathbf{R}}^{(m+1)}$ und $\mathbf{R}^{(m)}$ obere Dreiecksmatrizen sind, muss auch

$$\mathbf{R}^{(m+1)} := \widetilde{\mathbf{R}}^{(m+1)}\mathbf{R}^{(m)}$$

eine obere Dreiecksmatrix sein. Wir können also eine Zerlegung der gewünschten Form (5.20) berechnen, indem wir die Matrix $\mathbf{A}$ mit der *unitären* — und deshalb gut konditionierten — Matrix $\mathbf{Q}^{(m)}$ multiplizieren und eine QR-Zerlegung des Ergebnisses berechnen.

Die resultierende Verallgemeinerung der Vektoriteration bezeichnet man als *orthogonale Iteration*:

Algorithmus 5.44 (Orthogonale Iteration) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und $\mathbf{X}^{(0)} \in \mathbb{K}^{n\times k}$ so gegeben, dass die Voraussetzungen von Satz 5.41 erfüllt sind. Dann berechnet der folgende Algorithmus die Folge $(\mathbf{Q}^{(m)})_{m\in\mathbb{N}}$ aus (5.20). Die Spalten dieser Matrizen konvergieren gegen eine orthonormale Basis des invarianten Unterraums, der von den ersten k Eigenvektoren aufgespannt wird.*

$m \leftarrow 0$
Berechne die Zerlegung $\mathbf{Q}^{(0)}\mathbf{R}^{(0)} = \mathbf{X}^{(0)}$
while „*Fehler zu groß*“ **do begin**
  Berechne $\mathbf{W}^{(m+1)} \leftarrow \mathbf{A}\mathbf{Q}^{(m)}$
  Berechne die QR-Zerlegung $\mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{W}^{(m+1)}$
  $m \leftarrow m + 1$
end

Selbstverständlich müssen wir uns auch bei der orthogonalen Iteration Gedanken über ein geeignetes Abbruchkriterium machen. Wenn wir die Matrix $\widetilde{\mathbf{\Lambda}}^{(m)}$ aus Folgerung 5.42 zur Verfügung hätten, könnten wir relativ einfach nachprüfen, ob

$$\|\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)}\|_2$$

hinreichend klein ist, um eine gute Approximation des invarianten Unterraums sicherzustellen. Leider bietet uns die Folgerung nur eine Existenzaussage, so dass wir auf eine Approximation von $\mathbf{\Lambda}^{(m)}$ zurückgreifen müssen: Wenn $\mathbf{Q}^{(m)}$ eine exakte Basis eines invarianten Unterraums wäre, würde

$$\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\widetilde{\mathbf{\Lambda}}^{(m)} = \mathbf{0}$$

gelten. Indem wir diese Gleichung mit $(\mathbf{Q}^{(m)})^*$ multiplizieren und ausnutzen, dass $\mathbf{Q}^{(m)}$ orthogonal ist, würden wir dann

$$(\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} - \widetilde{\mathbf{\Lambda}}^{(m)} = \mathbf{0}$$

erhalten. Also dürfen wir hoffen, dass

$$\mathbf{\Lambda}^{(m)} := (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} \tag{5.22}$$

eine gute Approximation von $\widetilde{\mathbf{\Lambda}}^{(m)}$ sein wird. In der Tat ist $\mathbf{\Lambda}^{(m)}$ sogar eine bessere Wahl als $\widetilde{\mathbf{\Lambda}}^{(m)}$:

Lemma 5.45 *Es gilt*

$$\|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda})\mathbf{y}\|_2^2 = \|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}\|_2^2 + \|(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\|_2^2$$

für alle $m \in \mathbb{N}$ *und* $\mathbf{y} \in \mathbb{K}^k$ *und* beliebige *Matrizen* $\mathbf{\Lambda} \in \mathbb{K}^{k\times k}$.

Insbesondere minimiert die Wahl $\mathbf{\Lambda} = \mathbf{\Lambda}^{(m)}$ *die rechte Seite der Gleichung.*

Beweis. Seien $m \in \mathbb{N}$, $\mathbf{y} \in \mathbb{K}^k$ und $\mathbf{\Lambda} \in \mathbb{K}^{k\times k}$ gegeben. Dann folgt

$$\begin{aligned}
\|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda})\mathbf{y}\|_2^2 &= \|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y} + \mathbf{Q}^{(m)}(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\|_2^2 \\
&= \|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}\|_2^2 + \|\mathbf{Q}^{(m)}(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\|_2^2 \\
&\quad + 2\operatorname{Re}\langle(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}, \mathbf{Q}^{(m)}(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\rangle_2^2 \\
&= \|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}\|_2^2 + \|\mathbf{Q}^{(m)}(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\|_2^2 \\
&\quad + 2\operatorname{Re}\langle(\mathbf{Q}^{(m)})^*(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}, (\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\rangle_2^2 \\
&= \|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}\|_2^2 + \|\mathbf{Q}^{(m)}(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\|_2^2 \\
&\quad + 2\operatorname{Re}\langle\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda}^{(m)}\mathbf{y}, (\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\rangle_2^2 \\
&= \|(\mathbf{A}\mathbf{Q}^{(m)} - \mathbf{Q}^{(m)}\mathbf{\Lambda}^{(m)})\mathbf{y}\|_2^2 + \|(\mathbf{\Lambda}^{(m)} - \mathbf{\Lambda})\mathbf{y}\|_2^2,
\end{aligned}$$

und das ist die gewünschte Darstellung des Fehlers. ∎

Die Fehlerabschätzung aus Folgerung 5.42 gilt also auch, wenn wir auf der linken Seite $\widetilde{\mathbf{\Lambda}}^{(m)}$ durch $\mathbf{\Lambda}^{(m)}$ ersetzen, und letztere Matrix können wir mit (5.22) beziehungsweise etwas effizienter mit

$$\mathbf{\Lambda}^{(m)} = (\mathbf{Q}^{(m)})^* \mathbf{W}^{(m+1)}$$

konkret berechnen, um den Approximationsfehler zu bestimmen.

Bemerkung 5.46 (Orthogonale inverse Iteration) *Selbstverständlich können wir auch die inverse Iteration mit und ohne Shift in ähnlicher Weise modifizieren, um Näherungen invarianter Unterräume zu berechnen. Es können sogar unterschiedliche Shift-Parameter für die einzelnen Spalten von $\mathbf{Q}^{(m)}$ verwendet werden, um gezielt die Konvergenz zu beschleunigen.*

6 Die QR-Iteration

Wir wenden uns zunächst wieder der Frage zu, wie sich *alle* Eigenvektoren einer gegebenen Matrix bestimmen lassen. Für selbstadjungierte Matrizen haben wir mit der Jacobi-Iteration bereits ein erstes Verfahren zur Lösung dieser Aufgabe kennengelernt, allerdings zeigt sich, dass diese Iteration den Nachteil hat, dass bei bereits „fast“ diagonalen Matrizen, etwa Tridiagonalmatrizen, diese Struktur wieder zunichte gemacht wird.

In diesem Kapitel entwickeln wir einen alternativen Algorithmus, der diesen Nachteil nicht aufweist. Obwohl wir ihn nur für den Fall reeller symmetrischer Matrizen diskutieren, lässt er sich auch zur Bestimmung der Schur-Zerlegung komplexwertiger Matrizen einsetzen.

6.1 Grundidee

Am Beispiel der orthogonalen Iteration haben wir gesehen, dass sich mehrere Eigenvektoren simultan berechnen lassen. Also liegt es nahe, die Iteration so zu erweitern, dass *alle* Eigenvektoren berechnet werden, indem man sie mit einer vollständigen Basis startet, also beispielsweise mit $\mathbf{Q}^{(0)} = \mathbf{I}$ und $k = n$.

Wir erhalten damit das folgende Verfahren:

$\mathbf{Q}^{(0)} \leftarrow \mathbf{I}$
$m \leftarrow 0$
`repeat`
  $\mathbf{W}^{(m+1)} \leftarrow \mathbf{A}\mathbf{Q}^{(m)}$
  Berechne die QR-Zerlegung $\mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{W}^{(m+1)}$
  $m \leftarrow m + 1$
`until` „genau genug“

Bisher haben wir bei der Untersuchung der orthogonalen Iteration nicht ausgenutzt, dass die Matrizen $\mathbf{R}^{(m)}$ obere Dreiecksmatrizen sind. Diese Eigenschaft hat nützliche Konsequenzen, denen wir uns jetzt widmen werden.

Wir wählen ein $k \in [1 : n-1]$ und zerlegen die im Rahmen der orthogonalen Iteration auftretenden Matrizen in der Form

$$\mathbf{Q}^{(m)} = \begin{pmatrix} \mathbf{Q}_k^{(m)} & \mathbf{Q}_*^{(m)} \end{pmatrix}, \qquad \mathbf{Q}_k^{(m)} \in \mathbb{K}^{n\times k},$$

$$\mathbf{R}^{(m)} = \begin{pmatrix} \mathbf{R}_{kk}^{(m)} & \mathbf{R}_{k*}^{(m)} \\ & \mathbf{R}_{**}^{(m)} \end{pmatrix}, \qquad \mathbf{R}_{kk}^{(m)} \in \mathbb{K}^{k\times k} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Aus den definierenden Gleichungen der Iteration folgt dann

$$\begin{aligned}
&\left(\mathbf{Q}_k^{(m+1)}\mathbf{R}_{kk}^{(m+1)} \quad \mathbf{Q}_k^{(m+1)}\mathbf{R}_{k*}^{(m+1)} + \mathbf{Q}_*^{(m+1)}\mathbf{R}_{**}^{(m+1)}\right)\\
&= \left(\mathbf{Q}_k^{(m+1)} \quad \mathbf{Q}_*^{(m+1)}\right)\begin{pmatrix}\mathbf{R}_{kk}^{(m+1)} & \mathbf{R}_{k*}^{(m+1)}\\ & \mathbf{R}_{**}^{(m+1)}\end{pmatrix} = \mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)}\\
&= \mathbf{W}^{(m+1)} = \mathbf{A}\mathbf{Q}^{(m)} = \mathbf{A}\left(\mathbf{Q}_k^{(m)} \quad \mathbf{Q}_*^{(m)}\right)\\
&= \left(\mathbf{A}\mathbf{Q}_k^{(m)} \quad \mathbf{A}\mathbf{Q}_*^{(m)}\right) \qquad \text{für alle } m \in \mathbb{N}_0,
\end{aligned}$$

also insbesondere

$$\mathbf{Q}_k^{(m+1)}\mathbf{R}_{kk}^{(m+1)} = \mathbf{A}\mathbf{Q}_k^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Die ersten k Spalten $\mathbf{Q}_k^{(m)}$ der Matrizen $\mathbf{Q}^{(m)}$ können also auch als Ergebnis einer orthogonalen Iteration mit der Anfangsmatrix $\mathbf{Q}_k^{(0)}$ interpretiert werden. Also lassen sich unsere Konvergenzresultate auch auf diese Teilmatrizen anwenden.

Falls insbesondere

$$|\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_k| < |\lambda_{k+1}| \leq \ldots \leq |\lambda_n|$$

gilt, können wir Bemerkung 5.43 verwenden, um eine Abschätzung der Form

$$\|\mathbf{A}\mathbf{Q}_k^{(m)} - \mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)}\|_2 \leq C\left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \qquad \text{für alle } m \in \mathbb{N} \tag{6.1}$$

zu erhalten, indem wir entsprechend Lemma 5.45 die Matrix

$$\mathbf{\Lambda}_{kk}^{(m)} := (\mathbf{Q}_k^{(m)})^*\mathbf{A}\mathbf{Q}_k^{(m)}$$

einsetzen. Um diese Abschätzung auszunutzen, untersuchen wir, wie sich die Matrix $\mathbf{A}$ in der durch $\mathbf{Q}^{(m)}$ gegebenen Basis darstellt, wir sind also an den Matrizen

$$\mathbf{A}^{(m)} := (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0 \tag{6.2}$$

interessiert. Aus der Blockdarstellung der Matrizen $\mathbf{Q}^{(m)}$ folgt

$$\begin{aligned}
\mathbf{A}^{(m)} = (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} &= \begin{pmatrix}(\mathbf{Q}_k^{(m)})^*\\ (\mathbf{Q}_*^{(m)})^*\end{pmatrix}\mathbf{A}\left(\mathbf{Q}_k^{(m)} \quad \mathbf{Q}_*^{(m)}\right)\\
&= \begin{pmatrix}(\mathbf{Q}_k^{(m)})^*\mathbf{A}\mathbf{Q}_k^{(m)} & (\mathbf{Q}_k^{(m)})^*\mathbf{A}\mathbf{Q}_*^{(m)}\\ (\mathbf{Q}_*^{(m)})^*\mathbf{A}\mathbf{Q}_k^{(m)} & (\mathbf{Q}_*^{(m)})^*\mathbf{A}\mathbf{Q}_*^{(m)}\end{pmatrix}.
\end{aligned}$$

Wir interessieren uns für den linken unteren Block dieser Matrix und möchten beweisen, dass er gegen null konvergiert. Aus (6.1) folgt $\mathbf{A}\mathbf{Q}_k^{(m)} \approx \mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)}$, also

$$(\mathbf{Q}_*^{(m)})^*\mathbf{A}\mathbf{Q}_k^{(m)} \approx (\mathbf{Q}_*^{(m)})^*\mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)}.$$

Da die Spalten der Matrix $\mathbf{Q}^{(m)}$ eine Orthonormalbasis sind, müssen sie senkrecht aufeinander stehen. Konkret können wir nachrechnen, dass

$$\begin{aligned}\begin{pmatrix}(\mathbf{Q}_k^{(m)})^*\mathbf{Q}_k^{(m)} & (\mathbf{Q}_k^{(m)})^*\mathbf{Q}_*^{(m)}\\ (\mathbf{Q}_*^{(m)})^*\mathbf{Q}_k^{(m)} & (\mathbf{Q}_*^{(m)})^*\mathbf{Q}_*^{(m)}\end{pmatrix} &= \begin{pmatrix}(\mathbf{Q}_k^{(m)})^*\\ (\mathbf{Q}_*^{(m)})^*\end{pmatrix}\begin{pmatrix}\mathbf{Q}_k^{(m)} & \mathbf{Q}_*^{(m)}\end{pmatrix}\\ &= (\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m)} = \mathbf{I} = \begin{pmatrix}\mathbf{I} & \\ & \mathbf{I}\end{pmatrix}\end{aligned}$$

gilt, also insbesondere $(\mathbf{Q}_*^{(m)})^*\mathbf{Q}_k^{(m)} = \mathbf{0}$. Wegen $(\mathbf{Q}_k^{(m)})^*\mathbf{Q}_k^{(m)} = \mathbf{I}$ und $(\mathbf{Q}_*^{(m)})^*\mathbf{Q}_*^{(m)} = \mathbf{I}$ sind die Matrizen $\mathbf{Q}_k^{(m)}$ und $\mathbf{Q}_*^{(m)}$ außerdem isometrisch. Für den rechten unteren Block der Matrix $\mathbf{A}^{(m)}$ erhalten wir mit (6.1) und Lemma 3.19 die Abschätzung

$$\begin{aligned}\|(\mathbf{Q}_*^{(m)})^*\mathbf{A}\mathbf{Q}_k^{(m)}\| &= \|(\mathbf{Q}_*^{(m)})^*\mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)} + (\mathbf{Q}_*^{(m)})^*(\mathbf{A}\mathbf{Q}_k^{(m)} - \mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)})\|\\ &= \|(\mathbf{Q}_*^{(m)})^*(\mathbf{A}\mathbf{Q}_k^{(m)} - \mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)})\|\\ &\leq \|(\mathbf{Q}_*^{(m)})^*\|\,\|\mathbf{A}\mathbf{Q}_k^{(m)} - \mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)}\|\\ &= \|\mathbf{Q}_*^{(m)}\|\,\|\mathbf{A}\mathbf{Q}_k^{(m)} - \mathbf{Q}_k^{(m)}\mathbf{\Lambda}_{kk}^{(m)}\|\\ &\leq C\left(\frac{|\lambda_{k+1}|}{|\lambda_k|}\right)^m \qquad \text{für alle } m \in \mathbb{N}_0.\end{aligned}$$

Die Matrizen $\mathbf{A}^{(m)}$ werden sich also einer oberen Block-Dreiecksform annähern.

Falls wir dieses Argument auf alle $k \in [1 : n-1]$ anwenden können, falls also

$$|\lambda_1| < |\lambda_2| < \ldots < |\lambda_{n-1}| < |\lambda_n|$$

gilt, erhalten wir sogar, dass die Matrix gegen obere Dreiecksgestalt konvergiert, dass wir uns also iterativ der Schur-Zerlegung (vgl. Satz 3.39) nähern.

In diesem Fall bietet uns die orthogonale Iteration, angewendet auf eine vollständige Basis, eine Möglichkeit, die Schur-Zerlegung zu approximieren. Da $\mathbf{Q}^{(m)}$ in diesem Fall immer eine vollständige Basis ist, ist $\mathbf{A}^{(m)}$ das Ergebnis einer unitären Ähnlichkeitstransformation der Ausgangsmatrix $\mathbf{A}$, wir berechnen also eine Folge unitär ähnlicher Matrizen, die gegen eine obere Dreiecksmatrix konvergieren.

Wir würden diese Matrizen gerne berechnen, ohne die vollen orthogonalen Basen $\mathbf{Q}^{(m)}$ mitführen zu müssen, denn es gibt Anwendungen, bei denen wir nur an den Eigenwerten, aber nicht an den Eigenvektoren interessiert sind. Nach Konstruktion haben wir

$$\begin{aligned}\mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)} &= \mathbf{W}^{(m+1)} = \mathbf{A}\mathbf{Q}^{(m)},\\ \mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)}(\mathbf{Q}^{(m)})^* &= \mathbf{A}, \qquad (6.3)\\ (\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)} &= (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} = \mathbf{A}^{(m)}. \qquad (6.4)\end{aligned}$$

Aus (6.3) erhalten wir die Darstellung

$$\mathbf{A}^{(m+1)} = (\mathbf{Q}^{(m+1)})^*\mathbf{A}\mathbf{Q}^{(m+1)} = (\mathbf{Q}^{(m+1)})^*\mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)}(\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m+1)}$$

$$= \mathbf{R}^{(m+1)}(\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m+1)}.$$

Das Produkt

$$\widehat{\mathbf{Q}}^{(m+1)} := (\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m+1)} \tag{6.5}$$

tritt auch in (6.4) auf, wir erhalten also die Gleichungen

$$\mathbf{A}^{(m)} = \widehat{\mathbf{Q}}^{(m+1)}\mathbf{R}^{(m+1)}, \qquad \mathbf{A}^{(m+1)} = \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)}. \tag{6.6}$$

Die erste dieser Gleichungen entspricht einer QR-Zerlegung der Matrix $\mathbf{A}^{(m)}$, die wir auch ohne die Ausgangsmatrix $\mathbf{A}$ und die Matrix $\mathbf{Q}^{(m)}$ konstruieren können. Um unser Ziel zu erreichen, müssen wir also nur nachweisen, dass eine derartige QR-Zerlegung mit der für die orthogonale Iteration benötigten Zerlegung korrespondiert. Seien also nun $\widehat{\mathbf{Q}}^{(m+1)}$ und $\mathbf{R}^{(m+1)}$ durch (6.6) definiert. In Anlehnung an (6.5) definieren wir dann die Matrix $\mathbf{Q}^{(m+1)}$ durch die Gleichung

$$\mathbf{Q}^{(m+1)} := \mathbf{Q}^{(m)}\widehat{\mathbf{Q}}^{(m+1)} \tag{6.7}$$

und erhalten

$$\mathbf{Q}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{Q}^{(m)}\widehat{\mathbf{Q}}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{Q}^{(m)}\mathbf{A}^{(m)} = \mathbf{Q}^{(m)}(\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} = \mathbf{A}\mathbf{Q}^{(m)},$$

also die definierende Gleichung der orthogonalen Iteration. Damit haben wir unser Ziel erreicht: Wir können die orthogonale Iteration durchführen, ohne auf die Matrizen $\mathbf{A}$ und $\mathbf{Q}^{(m)}$ zurückzugreifen.

Dieser Zugang ähnelt wegen

$$\mathbf{A}^{(m+1)} = \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)} = (\widehat{\mathbf{Q}}^{(m+1)})^*\mathbf{A}^{(m)}\widehat{\mathbf{Q}}^{(m+1)}$$

dem bereits bei der Jacobi-Iteration verwendeten Ansatz: Alle Iterierten sind unitär ähnlich, und wir versuchen, die unitären Transformationen so zu wählen, dass sie gegen die obere Dreiecksgestalt konvergieren.

Algorithmus 6.1 (QR-Iteration) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Der folgende Algorithmus berechnet die Folge $(\mathbf{A}^{(m)})_{m\in\mathbb{N}}$ aus (6.2).*

$\mathbf{A}^{(0)} \leftarrow \mathbf{A}$
$m \leftarrow 0$
repeat
  Bestimme die QR-Zerlegung $\widehat{\mathbf{Q}}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{A}^{(m)}$
  $\mathbf{A}^{(m+1)} \leftarrow \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)}$
  $m \leftarrow m + 1$
until „genau genug“

Bemerkung 6.2 *Nach der durch Gleichung (6.2) gegebenen Definition sind alle Matrizen $\mathbf{A}^{(m)}$ unitär ähnlich zu $\mathbf{A}$, besitzen also dieselben Eigenwerte.*

Die in (6.2) verwendete Transformation $\mathbf{Q}^{(m)}$ kann mit Hilfe der Gleichung (6.7) während der Durchführung der QR-Iteration berechnet werden, falls wir nicht nur an den Eigenwerten, sondern auch an den Eigenvektoren von $\mathbf{A}$ interessiert sind.

6.2 Shift-Strategien und Deflation

Zur Verbesserung der Konvergenzgeschwindigkeit der Vektoriteration haben wir in Kapitel 5 die inverse Iteration mit Shift verwendet: Statt mit der Matrix $\mathbf{A}$ haben wir mit der Matrix $(\mathbf{A} - \mu\mathbf{I})^{-1}$ gearbeitet, die dieselben Eigenvektoren wie $\mathbf{A}$ besitzt, deren Eigenwerte aber durch

$$\lambda \in \sigma(\mathbf{A}) \qquad \Longleftrightarrow \qquad \frac{1}{\lambda - \mu} \in \sigma((\mathbf{A} - \mu\mathbf{I})^{-1})$$

gegeben sind, so dass wir durch Wahl des Shift-Parameters μ in der Nähe eines Eigenwerts für schnelle Konvergenz sorgen können. Dieser Ansatz lässt sich verfeinern, indem μ mit Hilfe des Rayleigh-Quotienten automatisch gewählt wird, und in diesem Fall konvergiert die entsprechende Iteration quadratisch, für normale Matrizen sogar kubisch.

Der Einsatz eines Shift-Parameters ist also von großem Vorteil zur Beschleunigung des Verfahrens. Leider macht er es bei der inversen Iteration auch erforderlich, mit der Inversen von $\mathbf{A} - \mu\mathbf{I}$ zu arbeiten, die uns bei der QR-Iteration nicht ohne weiteres zur Verfügung steht.

Ein genauerer Blick auf die die Konvergenz des Verfahrens beschreibende Abschätzung (6.1) legt nahe, dass wir auch eine andere Shift-Strategie verwenden können: Wenn μ hinreichend nahe an einem α-fachen Eigenwert von $\mathbf{A}$ liegt, können wir die Eigenwerte der Matrix $\mathbf{A} - \mu\mathbf{I}$ in die Reihenfolge

$$|\lambda_1 - \mu| \geq \ldots \geq |\lambda_{n-\alpha} - \mu| > |\lambda_{n-\alpha+1} - \mu| = \ldots = |\lambda_n - \mu|$$

mit $\lambda_{n-\alpha+1} = \ldots = \lambda_n$ bringen. Angewendet auf diese Matrix sollte also der linke untere Matrixblock mit α Zeilen und $n - \alpha$ Spalten gegen null konvergieren, und der Iterationsfehler sollte sich durch

$$\left(\frac{|\lambda_n - \mu|}{|\lambda_{n-\alpha} - \mu|}\right)^m$$

abschätzen lassen. Insbesondere sollte die Konvergenz um so schneller werden, je geringer der Abstand zwischen μ und dem Eigenwert $\lambda_{n-\alpha+1} = \lambda_n$ ist.

Eine Verschiebung des Spektrums von $\mathbf{A}$ kann also auch dann von Vorteil sein, wenn wir nicht die inverse Iteration verwenden. Da bei diesem Ansatz keine Berechnung der Inversen erforderlich ist, können wir den Shift-Parameter praktisch ohne zusätzlichen Rechenaufwand in jedem Schritt anpassen und so beispielsweise die Rayleigh-Quotienten-Strategie verwenden.

Für unsere Variante der QR-Iteration wollen wir weiterhin möglichst mit den Matrizen $\mathbf{A}^{(m)}$ statt $\mathbf{A}^{(m)} - \mu\mathbf{I}$ arbeiten. Dieses Ziel können wir erreichen, indem wir (6.6) durch

$$\mathbf{A}^{(m)} - \mu\mathbf{I} = \widehat{\mathbf{Q}}^{(m+1)}\mathbf{R}^{(m+1)}, \qquad \mathbf{A}^{(m+1)} = \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)} + \mu\mathbf{I} \tag{6.8}$$

ersetzen, denn dann gilt weiterhin

$$\mathbf{A}^{(m+1)} = \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)} + \mu\mathbf{I}$$

$$= (\widehat{\mathbf{Q}}^{(m+1)})^*(\mathbf{A}^{(m)} - \mu\mathbf{I})\widehat{\mathbf{Q}}^{(m+1)} + \mu(\widehat{\mathbf{Q}}^{(m+1)})^*\widehat{\mathbf{Q}}^{(m+1)}$$
$$= (\widehat{\mathbf{Q}}^{(m+1)})^*(\mathbf{A} - \mu\mathbf{I} + \mu\mathbf{I})\widehat{\mathbf{Q}}^{(m+1)} = (\widehat{\mathbf{Q}}^{(m+1)})^*\mathbf{A}^{(m)}\widehat{\mathbf{Q}}^{(m+1)},$$

wir berechnen also wie gehabt eine Ähnlichkeitstransformation von $\mathbf{A}^{(m)}$, geändert hat sich nur die Wahl der unitären Transformation $\widehat{\mathbf{Q}}^{(m+1)}$.

Es stellt sich die Frage nach der geeigneten Wahl des Shift-Parameters μ. Da wir hoffen, dass alle Außerdiagonalelemente der n-ten Zeile der Matrix $\mathbf{A}^{(m)}$ gegen null konvergieren, können wir annehmen, dass das letzte Diagonalelement $a_{nn}^{(m)}$ gegen einen Eigenwert konvergieren wird. Unter diesen Annahmen ist also $\mu = a_{nn}^{(m)}$ eine gute Wahl für den Shift-Parameter.

Diese Wahl lässt sich auch begründen, indem wir auf den Rayleigh-Quotienten zurückgreifen, der sich bereits bei der inversen Iteration als nützlich erwiesen hat: Wir bezeichnen den n-ten kanonischen Einheitsvektor mit $\mathbf{e}^{(n)} \in \mathbb{K}^n$ und die letzte Spalte von $\mathbf{Q}^{(m)}$ mit $\mathbf{q}^{(m)} = \mathbf{Q}^{(m)}\mathbf{e}^{(n)} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ und stellen fest, dass

$$\mu = a_{nn}^{(m)} = \frac{\langle \mathbf{e}^{(n)}, \mathbf{A}^{(m)}\mathbf{e}^{(n)}\rangle_2}{\langle \mathbf{e}^{(n)}, \mathbf{e}^{(n)}\rangle_2} = \frac{\langle \mathbf{e}^{(n)}, (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)}\mathbf{e}^{(n)}\rangle_2}{\langle \mathbf{e}^{(n)}, (\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m)}\mathbf{e}^{(n)}\rangle_2}$$
$$= \frac{\langle \mathbf{Q}^{(m)}\mathbf{e}^{(n)}, \mathbf{A}\mathbf{Q}^{(m)}\mathbf{e}^{(n)}\rangle_2}{\langle \mathbf{Q}^{(m)}\mathbf{e}^{(n)}, \mathbf{Q}^{(m)}\mathbf{e}^{(n)}\rangle_2} = \frac{\langle \mathbf{q}^{(m)}, \mathbf{A}\mathbf{q}^{(m)}\rangle_2}{\langle \mathbf{q}^{(m)}, \mathbf{q}^{(m)}\rangle_2} = \Lambda_A(\mathbf{q}^{(m)})$$

gilt. Damit entspricht unsere Wahl des Shift-Parameters gerade der Verwendung des Rayleigh-Quotienten, und analog zu Satz 5.33 lässt sich zeigen, dass $\mathbf{q}^{(m)}$ quadratisch gegen einen Eigenraum der adjungierten Matrix $\mathbf{A}^*$ konvergieren wird. Für eine normale Matrix können wir entsprechend Satz 5.34 sogar kubische Konvergenz nachweisen.

Mit Hilfe dieser Modifikation können wir also darauf hoffen, dass die Matrizen $\mathbf{A}^{(m)}$ schnell gegen die Form

$$\begin{pmatrix} \widetilde{\mathbf{A}} & \mathbf{C} \\ & \lambda \end{pmatrix}$$

konvergieren werden, dass also

$$\mathbf{A} \approx \mathbf{Q}^{(m)} \begin{pmatrix} \widetilde{\mathbf{A}} & \mathbf{C} \\ & \lambda \end{pmatrix} (\mathbf{Q}^{(m)})^*$$

für ein relativ kleines m gelten wird. Sobald die Einträge im linken unteren Block klein genug sind, können wir uns darauf konzentrieren, nur noch den linken oberen Block $\widetilde{\mathbf{A}}$ auf obere Dreiecksgestalt zu bringen. Falls wir nämlich eine näherungsweise Schur-Zerlegung

$$\widetilde{\mathbf{A}} \approx \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}\widetilde{\mathbf{Q}}^*$$

gefunden haben, ist durch

$$\mathbf{Q}^{(m)} \begin{pmatrix} \widetilde{\mathbf{Q}} & \\ & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{R}} & \widetilde{\mathbf{Q}}^*\mathbf{C} \\ & \lambda \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Q}}^* & \\ & 1 \end{pmatrix} (\mathbf{Q}^{(m)})^* \approx \mathbf{Q}^{(m)} \begin{pmatrix} \widetilde{\mathbf{A}} & \mathbf{C} \\ & \lambda \end{pmatrix} (\mathbf{Q}^{(m)})^* \approx \mathbf{A}$$

eine näherungsweise Schur-Zerlegung der Gesamtmatrix $\mathbf{A}$ gegeben. Indem wir das Konvergenzverhalten der Blöcke unterhalb der Diagonalen kontrollieren, können wir also

bereits konvergierte Teile der Matrix abspalten und so die Problemdimension nach und nach reduzieren.

Dieser Ansatz, also das Entfernen bereits konvergierter Diagonalblöcke aus dem Verfahren, trägt den Namen *Deflation* und sorgt einerseits für eine Reduktion des Rechenaufwands, während andererseits auch dafür gesorgt wird, dass bereits konvergierte Eigenwerte nicht mehr weiter als Shift-Parameter verwendet werden.

In der Praxis zeigt sich, dass mit Rayleigh-Shift in der Regel bereits nach sehr wenigen Schritten ein Eigenwert hinreichend gut approximiert ist, um die Deflation anwenden zu können. Insgesamt werden bei diesem Ansatz dann nur $\sim n$ Iterationen benötigt, um die Matrix auf obere Dreiecksgestalt zu bringen.

6.3 Hessenberg-Form

Bisher haben wir uns auf die Konvergenzrate der QR-Iteration konzentriert und den Aufwand für die Durchführung eines QR-Schrittes vernachlässigt. Dieser Aufwand ist sehr hoch: Die Berechnung einer QR-Zerlegung erfordert in der Regel $\sim n^3$ Operationen, und die Berechnung des Produkts $\mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)}$ hat einen ähnlich hohen Rechenaufwand. Wenn wir davon ausgehen, dass wir $\sim n$ Iterationen benötigen, um die Matrix auf obere Dreiecksgestalt zu bringen, würde unser Algorithmus $\sim n^4$ Operationen benötigen und wäre damit sehr aufwendig.

Unser Ziel ist es nun, die QR-Iteration zu beschleunigen. Dazu soll die Matrix so transformiert werden, daß sich ein QR-Schritt mit $\sim n^2$ oder im symmetrischen Fall sogar $\sim n$ Operationen durchführen lässt. Der Schlüssel zu dieser erheblichen Effizienzsteigerung ist eine Verallgemeinerung der oberen Dreiecksgestalt:

Definition 6.3 (Hessenberg-Form) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$. *Falls*

$$a_{ij} = 0 \qquad \text{für alle } i,j \in \{1,\dots,n\} \text{ mit } i > j+1$$

gilt, bezeichnet man $\mathbf{A}$ *als* Matrix in (oberer) Hessenberg-Form.

Jede obere Dreiecksmatrix ist auch eine Hessenberg-Matrix, aber bei einer Hessenberg-Matrix sind auch noch Einträge in der unteren Nebendiagonalen der Matrix erlaubt: Typisch ist eine Struktur der Form

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ & a_{32} & a_{33} & \dots & a_{3n} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}. \tag{6.9}$$

Diese Struktur bietet den Vorteil, dass sich die QR-Zerlegung einer Hessenberg-Matrix besonders einfach berechnen lässt: Wir wählen eine Givens-Rotation $\mathbf{G}_1 \in \mathbb{K}^{n\times n}$, die den

Eintrag a_{21} eliminiert, indem die erste und zweite Zeile miteinander kombiniert werden, beispielsweise als

$$\mathbf{G}_1 = \begin{pmatrix} \bar{c} & \bar{s} & & & \\ -s & c & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}, \qquad c = \frac{a_{11}}{r}, \qquad s = \frac{a_{21}}{r}, \qquad r = \sqrt{|a_{11}|^2 + |a_{21}|^2}.$$

Entsprechend wählen wir weitere Rotationen $\mathbf{G}_2, \ldots, \mathbf{G}_{n-1} \in \mathbb{K}^{n \times n}$, die der Reihe nach die Einträge $a_{32}, \ldots, a_{n,n-1}$ eliminieren und so die Matrix $\mathbf{A}$ in obere Dreiecksform überführen: Es gilt dann

$$\mathbf{G}_{n-1} \ldots \mathbf{G}_1 \mathbf{A} = \mathbf{R}, \qquad \mathbf{A} = \mathbf{G}_1^* \ldots \mathbf{G}_{n-1}^* \mathbf{R},$$

also haben wir die QR-Zerlegung in $\sim n^2$ Operationen berechnet, und der Faktor $\mathbf{Q}$ ist durch

$$\mathbf{Q} = \mathbf{G}_1^* \ldots \mathbf{G}_{n-1}^* \tag{6.10}$$

gegeben. Für die zweite Hälfte des QR-Schritts müssen wir nun

$$\mathbf{RQ} = (\mathbf{Q}^* \mathbf{R}^*)^* = (\mathbf{G}_{n-1} \ldots \mathbf{G}_1 \mathbf{R}^*)^*$$

berechnen. Das entspricht gerade der Anwendung der Givens-Rotationen auf die Spalten der Matrix $\mathbf{R}$, die Reihenfolge der Rotationen ist dieselbe wie bei der Berechnung der Zerlegung. Damit benötigt auch diese Berechnung $\sim n^2$ Operationen, so dass sich ein Schritt der QR-Iteration für eine Hessenberg-Matrix in $\sim n^2$ Operationen durchführen lässt.

Man kann sich leicht überlegen, dass $\mathbf{RQ}$ bei der beschriebenen Vorgehensweise wieder eine Matrix in Hessenberg-Form sein wird. Diese Eigenschaft ist von entscheidender Bedeutung: Sofern die Ausgangsmatrix $\mathbf{A}^{(0)}$ in Hessenberg-Form vorliegt, werden alle gemäß der oben beschriebenen Vorgehensweise berechneten Matrizen $\mathbf{A}^{(m)}$ der QR-Iteration ebenfalls in Hessenberg-Form sein, so dass jeder Iterationsschritt nur $\sim n^2$ Operationen erfordert. Das ist wesentlich effizienter als die $\sim n^3$ Operationen, die ohne Ausnutzung der Hessenberg-Form erforderlich wären.

Unser Ziel sollte es nun also sein, die erste Iterierte $\mathbf{A}^{(0)}$ möglichst in Hessenberg-Form zu überführen, indem wir die Basis $\mathbf{Q}^{(0)}$ geschickter als bisher wählen. Auch dieses Ziel lässt sich mit Hilfe geeigneter Givens-Rotationen erreichen: Um eine beliebige Matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

in eine Hessenberg-Form zu überführen, müssen wir eine unitäre Ähnlichkeitstransformation finden, die die Elemente $a_{31}, \ldots, a_{n1}, a_{42}, \ldots, a_{n2}, \ldots$ eliminiert. Falls wir, wie zuvor

beschrieben, das Element a_{31} durch Kombination der ersten und dritten Zeile eliminieren würden, müssten wir, da wir jetzt eine Ähnlichkeitstransformation benötigen, auch die erste und dritte Spalte kombinieren, und dadurch könnte der Eintrag a_{31} wieder einen anderen Wert als null erhalten.

Also kombinieren wir die *zweite* mit der dritten Zeile, um a_{31} zu eliminieren. Die korrespondieren Transformation der Spalten beeinflusst dann nur die zweite und dritte Spalte, so dass die in a_{31} eingeführte Null erhalten bleibt:

$$\mathbf{G}_{31} = \begin{pmatrix} 1 & & & & & \\ & \bar{c} & \bar{s} & & & \\ & -s & c & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}, \qquad c = \frac{a_{21}}{r}, \qquad s = \frac{a_{31}}{r}, \qquad r = \sqrt{|a_{21}|^2 + |a_{31}|^2}.$$

Entsprechend lassen sich auch die weiteren Einträge eliminieren, bis die Hessenberg-Gestalt erreicht ist.

Algorithmus 6.4 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Der folgende Algorithmus überschreibt $\mathbf{A}$ mit einer Matrix $\mathbf{A}^{(0)}$ in Hessenberg-Form, die die Gleichung $\mathbf{A}^{(0)} = (\mathbf{Q}^{(0)})^* \mathbf{A} \mathbf{Q}^{(0)}$ erfüllt.*

$\mathbf{Q}^{(0)} \leftarrow \mathbf{I}$
for $j = 1$ *to* $n-2$ *do*
  for $i = j+1$ *to* $n-1$ *do begin*
    $\gamma \leftarrow \sqrt{|a_{ij}|^2 + |a_{i+1,j}|^2}$; $c \leftarrow a_{ij}/\gamma$; $s \leftarrow a_{i+1,j}/\gamma$
    for $k = j$ *to* n *do begin*
      $h \leftarrow a_{ik}$; $a_{ik} \leftarrow \bar{c}h + \bar{s}a_{i+1,k}$; $a_{i+1,k} \leftarrow -sh + ca_{i+1,k}$
    end
    for $k = 1$ *to* n *do begin*
      $h \leftarrow a_{ki}$; $a_{ki} \leftarrow \bar{c}h + \bar{s}a_{k,i+1}$; $a_{k,i+1} \leftarrow -sh + ca_{k,i+1}$
      $h \leftarrow q^{(0)}_{ki}$; $q^{(0)}_{ki} \leftarrow \bar{c}h + \bar{s}q^{(0)}_{k,i+1}$; $q^{(0)}_{k,i+1} \leftarrow -sh + cq^{(0)}_{k,i+1}$
    end
  end

Mit diesem Algorithmus können wir eine beliebige Matrix in $\sim n^3$ Operationen in Hessenberg-Form überführen, so dass sich dann die weiteren Schritte der QR-Iteration effizient durchführen lassen.

Falls zusätzliche Informationen über die Struktur von $\mathbf{A}$ vorliegen, können wir sie ausnutzen, um den Algorithmus effizienter zu gestalten: Bei Bandmatrizen der Bandbreite k etwa genügen $\sim n^2 k$ Operationen für die Transformation. Besonders günstig ist die Transformation auf Hessenberg-Form, falls $\mathbf{A}$ selbstadjungiert ist:

Bemerkung 6.5 *Falls $\mathbf{A}$ selbstadjungiert ist, muss auch $\mathbf{A}^{(0)}$ selbstadjungiert sein. Damit folgt aus $a^{(0)}_{ij} = 0$ auch $a^{(0)}_{ji} = 0$ für alle $1 < j+1 < i < n$, die Matrix $\mathbf{A}^{(0)}$ ist also tridiagonal. Man kann sich einfach überlegen, dass der beschriebene QR-Schritt für eine*

tridiagonale Matrix sogar nur $\sim n$ Operationen erfordert, die Iteration wird also im Fall selbstadjungierter Matrizen noch wesentlich effizienter als im allgemeinen Fall sein.

Da eine Hessenberg-Matrix schon „fast“ die obere Dreiecksgestalt besitzt, können wir besonders einfach feststellen, wann eine Teilmatrix unterhalb der Diagonalen konvergiert ist und wir die Dimension reduzieren können:

Bemerkung 6.6 *Sobald $|a_{i+1,i}|$ für ein $i \in \{1, \dots, n-1\}$ hinreichend klein geworden ist, können wir per Deflation zu einer kleineren Teilmatrix übergehen. In der Praxis verwendet man skalierungsinvariante Kriterien der Form*

$$|a_{i+1,i}| \leq \epsilon(|a_{ii}| + |a_{i+1,i+1}|)$$

mit einer Fehlerschranke $\epsilon \in \mathbb{R}_{>0}$, um zu erkennen, wann eine Teilmatrix unterhalb der Diagonalen konvergiert ist.

Die Hessenberg-Form bietet uns auch die Möglichkeit, den Shift-Parameter geschickter als bisher zu wählen: Die Idee des *Wilkinson-Shifts* besteht darin, nicht nur den rechten unteren Diagonaleintrag, also das Gegenstück des Rayleigh-Quotienten, zu verwenden, sondern stattdessen die Eigenwerte des rechten unteren 2×2-Diagonalblocks zu benutzen.

Im symmetrischen Fall besitzt $\mathbf{A}^{(m)}$ Tridiagonalgestalt, und der Wilkinson-Shift-Parameter ergibt sich aus der Untersuchung der Eigenwerte der 2×2-Matrix

$$\mathbf{S} := \begin{pmatrix} a_{n-1,n-1}^{(m)} & a_{n-1,n}^{(m)} \\ \bar{a}_{n-1,n}^{(m)} & a_{nn}^{(m)} \end{pmatrix}.$$

Sie sind Nullstellen des charakteristischen Polynoms

$$\begin{aligned} p_S(\lambda) = \det(\lambda\mathbf{I} - \mathbf{S}) &= \det \begin{pmatrix} \lambda - a_{n-1,n-1}^{(m)} & -a_{n-1,n}^{(m)} \\ -\bar{a}_{n-1,n}^{(m)} & \lambda - a_{nn}^{(m)} \end{pmatrix} \\ &= (\lambda - a_{n-1,n-1}^{(m)})(\lambda - a_{nn}^{(m)}) - |a_{n-1,n}^{(m)}|^2. \end{aligned}$$

Indem wir Mittelwert und halbe Differenz der Diagonalelemente mit

$$m := \frac{a_{n-1,n-1}^{(m)} + a_{nn}^{(m)}}{2}, \qquad d := \frac{a_{n-1,n-1}^{(m)} - a_{nn}^{(m)}}{2}$$

bezeichnen, erhalten wir die Darstellung

$$p_S(\lambda) = (\lambda - m - d)(\lambda - m + d) - |a_{n-1,n}^{(m)}|^2 = (\lambda - m)^2 - d^2 - |a_{n-1,n}^{(m)}|^2,$$

so dass die Nullstellen, und damit die Eigenwerte, durch

$$(\lambda - m)^2 = d^2 + |a_{n-1,n}^{(m)}|^2, \qquad \lambda = m \pm \sqrt{d^2 + |a_{n-1,n}^{(m)}|^2}$$

gegeben sind. Da wir darauf hoffen, dass das letzte Diagonalelement $a_{nn}^{(m)}$ gegen einen Eigenwert konvergiert, bietet es sich an, diejenige Nullstelle von p_S als Shift-Parameter μ zu wählen, die dem derzeitigen Wert von $a_{nn}^{(m)}$ am nächsten liegt. Für $d < 0$ ist $a_{nn}^{(m)}$ größer als der Mittelwert m, anderenfalls kleiner oder gleich, so dass wir die Formel

$$\mu = \begin{cases} m - \sqrt{d^2 + |a_{n-1,n}^{(m)}|^2} & \text{falls } d > 0, \\ m + \sqrt{d^2 + |a_{n-1,n}^{(m)}|^2} & \text{ansonsten} \end{cases} \tag{6.11}$$

erhalten. Indem wir dieses μ als Shift-Parameter in der QR-Iteration verwenden, ergibt sich die *QR-Iteration mit Wilkinson-Shift*, bei der wir auf bessere Konvergenz als bei der Verwendung des einfachen Rayleigh-Quotienten $\mu = a_{nn}^{(m)}$ hoffen dürfen.

Auch im Falle nicht selbstadjungierter Matrizen lassen sich verfeinerte Shift-Strategien entwickeln, allerdings muss hier berücksichtigt werden, dass komplexe Shift-Parameter auftreten können. Insbesondere bei der Behandlung einer nichtsymmetrischen reellen Matrix wäre man daran interessiert, mit einem komplexen Shift zu rechnen, ohne komplexe Zahlen verwenden zu müssen. Indem man geschickt einen QR-Schritt mit einem komplexen Shift μ und einen QR-Schritt mit dem komplex konjugierten Shift $\bar{\mu}$ kombiniert, lässt sich dieses Ziel erreichen.

Bemerkung 6.7 (Eigenvektoren) *Wir können den Rechenaufwand der QR-Iteration deutlich reduzieren, indem wir auf die Akkumulation der Transformationsmatrizen $\mathbf{Q}^{(m)}$ verzichten: Für Hessenberg-Matrizen reduziert sich so der Aufwand eines QR-Schritts von $\mathcal{O}(n^3)$ auf $\mathcal{O}(n^2)$, für tridiagonale Matrizen sogar auf $\mathcal{O}(n)$.*

Um die Eigenvektoren zu rekonstruieren bietet sich die Verwendung einer inversen Iteration an: Da alle Eigenwerte bekannt sind, können wir sehr gute Shift-Parameter wählen und also auf sehr schnelle Konvergenz hoffen.

Dabei ist es eine gute Idee, die inverse Iteration für die Hessenberg-Matrix $\mathbf{A}^{(0)}$ statt für $\mathbf{A}$ durchzuführen, denn wir haben bereits gesehen, dass sich eine QR-Zerlegung einer Hessenberg-Matrix mit geringem Aufwand konstruieren lässt, so dass auch das für die inverse Iteration erforderliche Lösen des linearen Gleichungssystems

$$(\mathbf{A}^{(0)} - \mu\mathbf{I})\mathbf{w}^{(m+1)} = \mathbf{z}^{(m)}$$

effizient möglich ist.

6.4 Implizite Verfahren

Bisher haben wir einen QR-Schritt explizit durchgeführt: Erst wird die QR-Zerlegung $\widehat{\mathbf{Q}}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{A}^{(m)}$ berechnet, dann wird daraus $\mathbf{A}^{(m+1)} = \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)}$ konstruiert. Bei dieser Vorgehensweise müssen wir sämtliche Givens-Rotationen während der Zerlegung speichern, um sie anschließend bei der Multiplikation verwenden zu können.

Wesentlich eleganter wäre es, wenn wir $\mathbf{A}^{(m+1)}$ direkt aus $\mathbf{A}^{(m)}$ berechnen könnten, wenn wir also eine Möglichkeit hätten, die Zerlegung und Multiplikation *implizit* durchzuführen. Das Speichern der Givens-Rotationen wäre dann überflüssig.

Diese Möglichkeit besteht, wenn wir mit Hessenberg-Matrizen arbeiten: Im QR-Schritt verwenden wir unitäre Ähnlichkeitstransformationen, die Hessenberg-Matrizen wieder in Hessenberg-Matrizen überführen. Derartige Transformationen unterliegen Gesetzmäßigkeiten, die sich praktisch ausnutzen lassen.

Bemerkung 6.8 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ in Hessenberg-Form. Die Matrix ist genau dann irreduzibel (siehe Definition 3.58), falls*

$$a_{i+1,i} \neq 0 \qquad \text{für alle } i \in \{1,\dots,n-1\}$$

gilt.

Mit Hilfe der bereits beschriebenen Deflation können wir dafür sorgen, dass wir die QR-Schritte nur für irreduzible Hessenberg-Matrizen durchführen müssen. Für derartige Matrizen gilt die Aussage:

Satz 6.9 (Transformation von Hessenberg-Matrizen) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine irreduzible Hessenberg-Matrix, sei $\mathbf{B} \in \mathbb{K}^{n\times n}$ eine zweite Hessenberg-Matrizen, und sei $\mathbf{Q} \in \mathbb{K}^{n\times n}$ eine unitäre Matrix, die*

$$\mathbf{A} = \mathbf{Q}^*\mathbf{B}\mathbf{Q}, \qquad |q_{11}| = 1 \tag{6.12}$$

erfüllt. Dann ist $\mathbf{Q}$ eine Diagonalmatrix.

Beweis. Durch Induktion über $n \in \mathbb{N}$. Für $n = 1$ ist die Behauptung trivial.

Sei nun $n \in \mathbb{N}$ so gewählt, dass die Behauptung gilt. Sei $\mathbf{A} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine irreduzible Hessenberg-Matrix, sei $\mathbf{B} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine zweite Hessenberg-Matrix, und sei $\mathbf{Q} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine unitäre Matrix, die (6.12) erfüllt.

Wir bezeichnen wieder mit $\mathbf{e}^i \in \mathbb{K}^{n+1}$ den i-ten kanonischen Einheitsvektor, und wir setzen $\mathbf{q}^{(i)} := \mathbf{Q}\mathbf{e}^{(i)} \in \mathbb{K}^{n+1}$. Wegen

$$1 = \|\mathbf{q}^{(1)}\|_2^2 = \sum_{j=1}^{n+1} |q_j^{(1)}|^2 = \sum_{j=1}^{n+1} |q_{j1}|^2 = 1 + \sum_{j=2}^{n+1} |q_{j1}|^2$$

muss $\mathbf{q}^{(1)} = \alpha\mathbf{e}^{(1)}$ mit $\alpha \in \mathbb{K}$ und $|\alpha| = 1$ gelten.

Für alle $i \in \{2,\dots,n+1\}$ folgt daraus bereits

$$q_{1i} = \langle \mathbf{e}^{(1)}, \mathbf{q}^{(i)} \rangle_2 = \alpha \langle \mathbf{q}^{(1)}, \mathbf{q}^{(i)} \rangle_2 = 0,$$

da die Matrix $\mathbf{Q}$ unitär ist und ihre Spalten deshalb eine Orthogonalbasis bilden.

Also besitzt $\mathbf{Q}$ die Gestalt

$$\mathbf{Q} = \begin{pmatrix} \alpha & \\ & \widehat{\mathbf{Q}} \end{pmatrix}, \qquad \widehat{\mathbf{Q}} := \begin{pmatrix} q_{22} & \cdots & q_{2,n+1} \\ \vdots & \ddots & \vdots \\ q_{n+1,2} & \cdots & q_{n+1,n+1} \end{pmatrix}.$$

Damit folgt aus $\mathbf{A} = \mathbf{Q}^*\mathbf{B}\mathbf{Q}$ die Gleichung

$$\widehat{\mathbf{A}} = \widehat{\mathbf{Q}}^*\widehat{\mathbf{B}}\widehat{\mathbf{Q}}, \quad \widehat{\mathbf{A}} := \begin{pmatrix} a_{22} & \dots & a_{2,n+1} \\ \vdots & \ddots & \vdots \\ a_{n+1,2} & \dots & a_{n+1,n+1} \end{pmatrix}, \quad \widehat{\mathbf{B}} := \begin{pmatrix} b_{22} & \dots & b_{2,n+1} \\ \vdots & \ddots & \vdots \\ b_{n+1,2} & \dots & b_{n+1,n+1} \end{pmatrix}$$

mit der irreduziblen Hessenberg-Matrix $\widehat{\mathbf{A}} \in \mathbb{K}^{n\times n}$, der Hessenberg-Matrix $\widehat{\mathbf{B}} \in \mathbb{K}^{n\times n}$ und der unitären Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{n\times n}$. Um die Induktionsvoraussetzung anwenden zu können, müssen wir nur noch nachweisen, dass $|\hat{q}_{11}| = |q_{22}| = 1$ gilt.

Nach Voraussetzung gilt

$$\mathbf{Q}\mathbf{A} = \mathbf{B}\mathbf{Q},$$

und durch Einsetzen von $\mathbf{e}^{(1)}$ folgt

$$\begin{aligned} a_{11}\mathbf{q}^{(1)} + a_{21}\mathbf{q}^{(2)} &= \mathbf{Q}(a_{11}\mathbf{e}^{(1)} + a_{21}\mathbf{e}^{(2)}) = \mathbf{Q}\mathbf{A}\mathbf{e}^{(1)} \\ &= \mathbf{B}\mathbf{Q}\mathbf{e}^{(1)} = \mathbf{B}\mathbf{q}^{(1)} = \alpha\mathbf{B}\mathbf{e}^{(1)} = \alpha b_{11}\mathbf{e}^{(1)} + \alpha b_{21}\mathbf{e}^{(2)}. \end{aligned}$$

Durch Subtraktion von $a_{11}\mathbf{q}^{(1)} = \alpha a_{11}\mathbf{e}^{(1)}$ auf beiden Seiten erhalten wir

$$a_{21}\mathbf{q}^{(2)} = \alpha(b_{11} - a_{11})\mathbf{e}^{(1)} + \alpha b_{21}\mathbf{e}^{(2)}.$$

Wir haben bereits gesehen, dass $q_1^{(2)} = q_{12} = 0$ gilt, also bleibt nur

$$a_{21}\mathbf{q}^{(2)} = \alpha b_{21}\mathbf{e}^{(2)},$$

übrig, und da $a_{21} \neq 0$ nach Voraussetzung gilt, haben wir

$$\mathbf{q}^{(2)} = \alpha\frac{b_{21}}{a_{21}}\mathbf{e}^{(2)}$$

bewiesen. Da $\mathbf{Q}$ unitär ist, muss $\mathbf{q}^{(2)}$ normiert sein, also gilt

$$\begin{aligned} |q_{22}|^2 = |\langle \mathbf{e}^{(2)}, \mathbf{q}^{(2)}\rangle_2|^2 &= \left|\alpha\frac{b_{21}}{a_{21}}\langle \mathbf{e}^{(2)}, \mathbf{e}^{(2)}\rangle_2\right|^2 = \left|\alpha\frac{b_{21}}{a_{21}}\right|^2 \\ &= \left\langle \alpha\frac{b_{21}}{a_{21}}\mathbf{e}^{(2)}, \alpha\frac{b_{21}}{a_{21}}\mathbf{e}^{(2)} \right\rangle = \langle \mathbf{q}^{(2)}, \mathbf{q}^{(2)}\rangle_2 = \|\mathbf{q}^{(2)}\|_2^2 = 1. \end{aligned}$$

Damit folgt aus der Induktionsvoraussetzung, dass die Matrix $\widehat{\mathbf{Q}}$ eine Diagonalmatrix ist, und damit auch die Matrix $\mathbf{Q}$. ∎

Tatsächlich lässt sich der Beweis auch auf den nicht vollständig irreduziblen Fall anwenden, um ein für unsere Zwecke völlig ausreichendes Teilresultat zu erzielen:

Bemerkung 6.10 (Nicht-irreduzibler Fall) *Falls* $\mathbf{A}$ *in Satz 6.9 nicht irreduzibel ist, sondern lediglich*

$$a_{i+1,i} \neq 0 \qquad \text{für alle } i \in \{1, \dots, k\}$$

mit einem $k < n$ erfüllt, können wir die Induktion nach k Schritten abbrechen und erhalten immer noch

$$\mathbf{Q} = \begin{pmatrix} \mathbf{D} & \\ & \widehat{\mathbf{Q}} \end{pmatrix}$$

mit einer unitären Diagonalmatrix $\mathbf{D} \in \mathbb{K}^{k\times k}$ und einer nicht weiter zu untersuchenden unitären Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{(n-k)\times(n-k)}$.

Der bisherige QR-Schritt beruhte darauf, die QR-Zerlegung $\widehat{\mathbf{Q}}^{(m+1)}\mathbf{R}^{(m+1)} = \mathbf{A}^{(m)}$ zu berechnen und dann

$$\mathbf{A}^{(m+1)} = \mathbf{R}^{(m+1)}\widehat{\mathbf{Q}}^{(m+1)} = (\widehat{\mathbf{Q}}^{(m+1)})^*\mathbf{A}^{(m)}\widehat{\mathbf{Q}}^{(m+1)}$$

zu konstruieren. Praktisch haben wir $\widehat{\mathbf{Q}}^{(m+1)}$ als Folge von $n-1$ Givens-Rotationen dargestellt (siehe (6.10)), also als

$$\widehat{\mathbf{Q}}^{(m+1)} = \mathbf{G}_1^*\mathbf{G}_2^*\dots\mathbf{G}_{n-1}^*,$$

bei denen $\mathbf{G}_i^*$ jeweils nur auf die i-te und die $(i+1)$-te Komponente eines Vektors wirkten. Man kann sich leicht überlegen, dass daraus folgt, dass die erste Spalte der Transformation $\widehat{\mathbf{Q}}^{(m+1)}$ ausschließlich von $\mathbf{G}_1$ abhängt, aber von keiner der anderen Transformationen.

Falls wir also eine zweite unitäre Transformation

$$\widetilde{\mathbf{Q}}^{(m+1)} = \mathbf{G}_1^*\widetilde{\mathbf{G}}_2^*\dots\widetilde{\mathbf{G}}_{n-1}^*$$

aus Givens-Rotationen $\widetilde{\mathbf{G}}_i^*$ konstruieren, die auch jeweils nur die i-te und die $(i+1)$-te Zeile beeinflussen, müssen beide dieselbe erste Spalte besitzen.

Wir setzen $\mathbf{P} := (\widehat{\mathbf{Q}}^{(m+1)})^*\widetilde{\mathbf{Q}}^{(m+1)}$. Als Produkt unitärer Matrizen ist $\mathbf{P}$ ebenfalls unitär, und infolge der Identität der ersten Spalten folgt $p_{11} = 1$.

Falls $\mathbf{A}^{(m+1)}$ irreduzibel ist und falls wir $\widetilde{\mathbf{G}}_2^*, \dots, \widetilde{\mathbf{G}}_{n-1}^*$ so gewählt haben, dass

$$\widetilde{\mathbf{A}}^{(m+1)} := (\widetilde{\mathbf{Q}}^{(m+1)})^*\mathbf{A}^{(m)}\widetilde{\mathbf{Q}}^{(m+1)}$$

wieder eine Hessenberg-Matrix ist, erhalten wir

$$\begin{aligned}\mathbf{A}^{(m+1)} &= (\widehat{\mathbf{Q}}^{(m+1)})^*\mathbf{A}^{(m)}\widehat{\mathbf{Q}}^{(m+1)} \\ &= (\widehat{\mathbf{Q}}^{(m+1)})^*\widetilde{\mathbf{Q}}^{(m+1)}\mathbf{A}^{(m+1)}(\widetilde{\mathbf{Q}}^{(m+1)})^*\widehat{\mathbf{Q}}^{(m+1)} = \mathbf{P}\widetilde{\mathbf{A}}^{(m+1)}\mathbf{P}^*,\end{aligned}$$

also können wir Satz 6.9 anwenden, um zu folgern, dass sich $\mathbf{A}^{(m+1)}$ und $\widetilde{\mathbf{A}}^{(m+1)}$ nur durch eine orthogonale Diagonalskalierung unterscheiden. Derartige Skalierungen sind für die Konvergenz der QR-Iteration ohne Belang, wir können also $\widetilde{\mathbf{A}}^{(m+1)}$ anstelle von $\mathbf{A}^{(m+1)}$ verwenden. Falls $\mathbf{A}^{(m+1)}$ nicht irreduzibel sein sollte, folgt aus Bemerkung 6.10, dass wir auf einen Teil der Matrix eine nicht näher festgelegte unitäre Transformation angewendet haben. Da in diesem Fall aber ohnehin als nächstes eine Deflation durchgeführt und diese Teilmatrix separat weiter behandelt würde, schadet diese überflüssige Transformation nicht.

Der Vorteil dieses Zugangs besteht darin, dass wir $\widetilde{\mathbf{A}}^{(m+1)}$ direkt aus $\mathbf{A}^{(m)}$ konstruieren können: Zunächst wählen wir die Givens-Rotation $\mathbf{G}_1$ so, dass das Nebendiagonalelement $a_{21}^{(m)}$ der geshifteten Matrix $\mathbf{A}^{(m)} - \mu\mathbf{I}$ eliminiert wird:

$$\mathbf{G}_1 = \begin{pmatrix} \bar{c} & \bar{s} & & & \\ -s & c & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}, \quad c = \frac{a_{11}^{(m)} - \mu}{r}, \quad s = \frac{a_{21}^{(m)}}{r}, \quad r = \sqrt{|a_{11}^{(m)} - \mu|^2 + |a_{21}^{(m)}|^2}.$$

Die durch $\mathbf{G}_1$ definierte unitäre Ähnlichkeitstransformation wenden wir auf $\mathbf{A}^{(m)}$ an, um die Matrix

$$\mathbf{B} := \mathbf{G}_1\mathbf{A}^{(m)}\mathbf{G}_1^* = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1,n-1} & b_{1n} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2,n-1} & b_{2n} \\ \mathbf{b_{31}} & b_{32} & b_{33} & \dots & b_{3,n-1} & b_{3n} \\ & & b_{43} & \dots & b_{4,n-1} & b_{4n} \\ & & & \ddots & \vdots & \vdots \\ & & & & b_{n,n-1} & b_{nn} \end{pmatrix}$$

zu erhalten. Da bei der Spaltentransformation (also der Multiplikation mit $\mathbf{G}_1^*$ von rechts) die erste und zweite Spalte kombiniert werden, wird im Allgemeinen $b_{31} \neq 0$ gelten, die Matrix $\mathbf{B}$ ist also keine Hessenberg-Matrix mehr.

Da wir Satz 6.9 nur auf Hessenberg-Matrizen anwenden können, müssen wir also nun die Matrix $\mathbf{B}$ unitär so transformieren, dass sie wieder die Hessenberg-Gestalt annimmt. Also eliminieren wir b_{31} mit Hilfe einer Givens-Rotation, die die zweite und dritte Zeile kombiniert:

$$\widetilde{\mathbf{G}}_2 = \begin{pmatrix} 1 & & & & & \\ & \bar{c} & \bar{s} & & & \\ & -s & c & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}, \qquad c = \frac{b_{21}}{r}, \qquad s = \frac{b_{31}}{r}, \qquad r = \sqrt{|b_{21}|^2 + |b_{31}|^2}.$$

Die Ähnlichkeitstransformation mit $\widetilde{\mathbf{G}}_2$ führt zwar dazu, dass der Eintrag b_{31} eliminiert wird, allerdings sorgt auch hier die Transformation der Spalten dafür, dass ein neuer Eintrag außerhalb der Hessenberg-Gestalt entsteht, denn da b_{43} in der Regel von null verschieden ist, entsteht so ein Eintrag in der vierten Zeile der zweiten Spalte, und die

resultierende Matrix hat die Form

$$\mathbf{C} := \widetilde{\mathbf{G}}_2 \mathbf{B} \widetilde{\mathbf{G}}_2^* = \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & \dots & c_{1,n-1} & c_{1n} \\ c_{21} & c_{22} & c_{23} & c_{24} & \dots & c_{2,n-1} & c_{2n} \\ & c_{32} & c_{33} & c_{34} & \dots & c_{3,n-1} & c_{3n} \\ & \mathbf{c_{42}} & c_{43} & c_{44} & \dots & c_{4,n-1} & c_{4n} \\ & & & c_{54} & \dots & c_{5,n-1} & c_{5n} \\ & & & & \ddots & \vdots & \vdots \\ & & & & & c_{n,n-1} & c_{nn} \end{pmatrix}.$$

Das aus der Hessenberg-Form herausfallende Element c_{42} können wir wiederum mit einer Givens-Rotation $\widetilde{\mathbf{G}}_3$ eliminieren, die die dritte und vierte Zeile kombiniert und zu einem neuen Element in der fünften Zeile der dritten Spalte führt. Mit jeder Givens-Rotation verschiebt sich also das störende Element weiter nach rechts unten, bis es mit einer letzten Rotation $\widetilde{\mathbf{G}}_{n-1}$ endgültig eliminiert werden kann und die Matrix

$$\widetilde{\mathbf{A}}^{(m+1)} = \widetilde{\mathbf{G}}_{n-1} \dots \widetilde{\mathbf{G}}_2 \mathbf{G}_1 \mathbf{A}^{(m)} \mathbf{G}_1^* \widetilde{\mathbf{G}}_2^* \dots \widetilde{\mathbf{G}}_{n-1}^*$$

wieder Hessenberg-Form hat. Gemäß Satz 6.9 unterscheidet sich diese Matrix von $\mathbf{A}^{(m+1)}$ nur durch eine unitäre Diagonalskalierung, also können wir fortfahren, als wäre sie die neue Iterierte.

Algorithmus 6.11 *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine irreduzible Hessenberg-Matrix. Der folgende Algorithmus überschreibt $\mathbf{A}$ mit einer Matrix, die im Wesentlichen (siehe Bemerkung 6.10) mit der nächsten Iterierten der QR-Iteration mit Shift übereinstimmt.*

$r \leftarrow \sqrt{|a_{11} - \mu|^2 + |a_{21}|^2}$; $c \leftarrow (a_{11} - \mu)/r$; $s \leftarrow a_{21}/r$
$a_{11} \leftarrow r$; $a_{21} \leftarrow 0$
for $j = 2$ **to** n **do begin**
  $h \leftarrow a_{1j}$; $a_{1j} \leftarrow \bar{c}h + \bar{s}a_{2j}$; $a_{2j} \leftarrow -sh + ca_{2j}$
end
$h \leftarrow a_{11}$; $a_{11} \leftarrow \bar{c}h + \bar{s}a_{12}$; $a_{12} \leftarrow -sh + ca_{12}$
$a_{21} \leftarrow \bar{s}a_{22}$; $a_{22} \leftarrow ca_{22}$
for $i = 2$ **to** $n - 1$ **do begin**
  $\gamma \leftarrow \bar{s}a_{i+1,i}$; $a_{i+1,i} \leftarrow ca_{i+1,i}$
  if $|\gamma| \le \epsilon(|a_{i,i-1}| + |a_{i+1,i}|)$ **then break**
  $r \leftarrow \sqrt{|a_{i,i-1}|^2 + |\gamma|^2}$; $c \leftarrow a_{i,i-1}/r$; $s \leftarrow \gamma/r$
  $a_{i,i-1} \leftarrow r$
  for $j = i$ **to** n **do begin**
    $h \leftarrow a_{ij}$; $a_{ij} \leftarrow \bar{c}h + \bar{s}a_{i+1,j}$; $a_{i+1,j} \leftarrow -sh + ca_{i+1,j}$
  end
  for $k = 1$ **to** i **do begin**
    $h \leftarrow a_{ki}$; $a_{ki} \leftarrow \bar{c}h + \bar{s}a_{k,i+1}$; $a_{k,i+1} \leftarrow -sh + ca_{k,i+1}$
  end
end

6.5 Singulärwertzerlegung*

Offensichtlich können nur quadratische Matrizen über Eigenwerte verfügen. Im Fall rechteckiger Matrizen kann gelegentlich die *Singulärwertzerlegung* an die Stelle der Schur-Zerlegung treten.

Satz 6.12 (Singulärwertzerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times m}$, sei $k := \min\{n, m\}$.*

Dann existieren zwei isometrische Matrizen $\mathbf{U} \in \mathbb{K}^{n\times k}$ und $\mathbf{V} \in \mathbb{K}^{m\times k}$ und reelle Zahlen $\sigma_1 \geq \sigma_2 \geq \sigma_k \geq 0$ mit

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix}. \tag{6.13}$$

Eine solche Faktorisierung nennt man eine Singulärwertzerlegung *der Matrix $\mathbf{A}$. Die Zahlen $\sigma_1, \ldots, \sigma_k$ werden die* Singulärwerte *der Matrix $\mathbf{A}$ genannt, die Spalten der Matrizen $\mathbf{U}$ und $\mathbf{V}$* linke und rechte Singulärvektoren*.*

Beweis. Wir führen den Beweis per Induktion über $k \in \mathbb{N}$.

Induktionsanfang: Gelte $k = 1$. Falls $\mathbf{A} = \mathbf{0}$ gilt, können wir $\sigma_1 = 0$ verwenden und beliebige isometrische Matrizen $\mathbf{U}$ und $\mathbf{V}$ wählen.

Anderenfalls setzen wir $\sigma_1 = \|\mathbf{A}\|_2 > 0$.

Falls $n = 1$ gilt, setzen wir $u_{11} = 1$ und $\mathbf{V} = \mathbf{A}^*/\sigma_1$.

Anderenfalls gilt $m = 1$ und wir setzen $v_{11} = 1$ und $\mathbf{U} = \mathbf{A}/\sigma_1$.

Induktionsvoraussetzung: Sei $k \in \mathbb{N}$ so gegeben, dass jede Matrix $\mathbf{A} \in \mathbb{K}^{n\times m}$ mit $k = \min\{n, m\}$ eine Singulärwertzerlegung der Form (6.13) besitzt.

Induktionsschritt: Sei $\mathbf{A} \in \mathbb{K}^{n\times m}$ mit $k+1 = \min\{n, m\}$ gegeben. Wir bezeichnen mit

$$\mathcal{S}_n := \{\mathbf{x} \in \mathbb{K}^n \ : \ \|\mathbf{x}\| = 1\}, \qquad \mathcal{S}_m := \{\mathbf{y} \in \mathbb{K}^m \ : \ \|\mathbf{y}\| = 1\}$$

die n- und die m-dimensionale Einheitssphäre und betrachten die Funktion

$$f\colon \mathcal{S}_n \times \mathcal{S}_m \to \mathbb{R}_{\geq 0}, \qquad (\mathbf{x}, \mathbf{y}) \mapsto |\langle \mathbf{x}, \mathbf{A}\mathbf{y}\rangle|.$$

Diese Funktion ist stetig, nimmt also auf der kompakten Menge $\mathcal{S}_n \times \mathcal{S}_m$ ein Maximum σ_1 an, für das wir $\mathbf{u} \in \mathcal{S}_n$ und $\mathbf{v} \in \mathcal{S}_m$ mit $\sigma_1 = f(\mathbf{u}, \mathbf{v}) = |\langle \mathbf{u}, \mathbf{A}\mathbf{v}\rangle|$ finden.

Aus der Cauchy-Schwarz-Ungleichung (3.5) folgt

$$|\langle \mathbf{u}, \mathbf{A}\mathbf{v}\rangle| \leq \|\mathbf{u}\| \, \|\mathbf{A}\mathbf{v}\|,$$

und Gleichheit gilt genau dann, wenn $\mathbf{u}$ und $\mathbf{A}\mathbf{v}$ linear abhängig sind. Da σ_1 das Maximum dieses Ausdrucks ist und wir $\mathbf{u}$ unabhängig von $\mathbf{v}$ wählen dürfen, folgt, dass $\mathbf{u}$ und $\mathbf{A}\mathbf{v}$ linear abhängig sind, dass also ein $\alpha \in \mathbb{K}$ mit

$$\begin{aligned} &\mathbf{A}\mathbf{v} = \alpha\mathbf{u}, \\ &|\alpha| = |\alpha| \, |\langle \mathbf{u}, \mathbf{u}\rangle| = |\langle \mathbf{u}, \alpha\mathbf{u}\rangle| = |\langle \mathbf{u}, \mathbf{A}\mathbf{v}\rangle| = \sigma_1 \end{aligned}$$

existiert. Indem wir das Vorzeichen von $\mathbf{u}$ anpassen können wir

$$\mathbf{A}\mathbf{v} = \sigma_1 \mathbf{u}$$

sicher stellen. Analog folgt aus Lemma 3.16 und der Cauchy-Schwarz-Ungleichung auch

$$|\langle \mathbf{u}, \mathbf{A}\mathbf{v}\rangle| = |\langle \mathbf{A}^*\mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{A}^*\mathbf{u}\|\,\|\mathbf{v}\|,$$

und wir können wie zuvor folgern, dass $\mathbf{A}^*\mathbf{u}$ und $\mathbf{v}$ linear abhängig sind, dass also ein $\beta \in \mathbb{K}$ mit

$$\mathbf{A}^*\mathbf{u} = \beta\mathbf{v}$$

existiert. Mit unsere modifizierten Wahl des Vektors $\mathbf{u}$ folgt

$$\beta = \beta\langle \mathbf{v}, \mathbf{v}\rangle = \langle \mathbf{v}, \beta\mathbf{v}\rangle = \langle \mathbf{v}, \mathbf{A}^*\mathbf{u}\rangle = \langle \mathbf{A}\mathbf{v}, \mathbf{u}\rangle = \langle \sigma\mathbf{u}, \mathbf{u}\rangle = \sigma,$$

also haben wir insgesamt

$$\mathbf{A}\mathbf{v} = \sigma_1\mathbf{u}, \qquad \mathbf{A}^*\mathbf{u} = \sigma_1\mathbf{v}$$

bewiesen. Wir wählen Householder-Spiegelungen $\mathbf{U}_1 \in \mathbb{K}^{n\times n}$ und $\mathbf{V}_1 \in \mathbb{K}^{m\times m}$ derart, dass

$$\mathbf{u} = \mathbf{U}_1\delta^{(1)}, \qquad \mathbf{v} = \mathbf{V}_1\delta^{(1)}$$

gelten. Dann folgen

$$\begin{aligned}\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1\delta^{(1)} &= \mathbf{U}_1^*\mathbf{A}\mathbf{v} = \sigma_1\mathbf{U}_1^*\mathbf{u} = \sigma_1\delta^{(1)},\\ (\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1)^*\delta^{(1)} &= \mathbf{V}_1^*\mathbf{A}^*\mathbf{U}_1\delta^{(1)} = \mathbf{V}_1^*\mathbf{A}^*\mathbf{u} = \sigma_1\mathbf{V}_1^*\mathbf{v} = \sigma_1\delta^{(1)},\end{aligned}$$

so dass die transformierte Matrix die Gestalt

$$\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1 = \begin{pmatrix}\sigma_1 & \\ & \widehat{\mathbf{A}}\end{pmatrix}, \qquad \widehat{\mathbf{A}} \in \mathbb{K}^{(n-1)\times(m-1)}$$

aufweist. Nach der Induktionsvoraussetzung finden wir eine Singulärwertzerlegung der Matrix $\widehat{\mathbf{A}}$, wir finden also isometrische Matrizen $\widehat{\mathbf{U}} \in \mathbb{K}^{(n-1)\times k}$ und $\widehat{\mathbf{V}} \in \mathbb{K}^{(m-1)\times k}$ sowie $\sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_{k+1}$ mit

$$\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^*, \qquad \widehat{\mathbf{\Sigma}} = \begin{pmatrix}\sigma_2 & & \\ & \ddots & \\ & & \sigma_{k+1}\end{pmatrix}.$$

Zusammengesetzt erhalten wir

$$\mathbf{A} = \mathbf{U}_1\begin{pmatrix}\sigma_1 & \\ & \widehat{\mathbf{A}}\end{pmatrix}\mathbf{V}_1^* = \mathbf{U}_1\begin{pmatrix}\sigma_1 & \\ & \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^*\end{pmatrix}\mathbf{V}_1^* = \mathbf{U}_1\begin{pmatrix}1 & \\ & \widehat{\mathbf{U}}\end{pmatrix}\begin{pmatrix}\sigma_1 & \\ & \widehat{\mathbf{\Sigma}}\end{pmatrix}\begin{pmatrix}1 & \\ & \widehat{\mathbf{V}}\end{pmatrix}^*\mathbf{V}_1^*,$$

so dass wir mit

$$\mathbf{U} := \mathbf{U}_1 \begin{pmatrix} 1 & \\ & \widehat{\mathbf{U}} \end{pmatrix}, \qquad \mathbf{V} := \mathbf{V}_1 \begin{pmatrix} 1 & \\ & \widehat{\mathbf{V}} \end{pmatrix}, \qquad \mathbf{\Sigma} := \begin{pmatrix} \sigma_1 & \\ & \widehat{\mathbf{\Sigma}} \end{pmatrix}$$

die gewünschte Zerlegung gefunden haben. Mit $\mathbf{x} := \mathbf{U}\delta^{(2)}$ und $\mathbf{y} := \mathbf{V}\delta^2$ erhalten wir

$$\sigma_2 = |\sigma_2| = |\langle \delta^{(2)}, \mathbf{\Sigma}\delta^{(2)} \rangle| = |\langle \mathbf{U}^*\mathbf{x}, \mathbf{\Sigma}\mathbf{V}^*\mathbf{y} \rangle| = |\langle \mathbf{x}, \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{y} \rangle| = |\langle \mathbf{x}, \mathbf{A}\mathbf{y} \rangle| \leq \sigma_1,$$

so dass die Singulärwerte auch bereits die gewünschte Reihenfolge aufweisen. ∎

Eine solche Singulärwertzerlegung ist beispielsweise sehr nützlich, um lineare Ausgleichsprobleme zu lösen oder approximative Faktorisierungen der Matrix $\mathbf{A}$ zu konstruieren.

Die Singulärwertzerlegung ist eng verwandt mit der Schur-Zerlegung: Falls

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$$

mit isometrischen Matrizen $\mathbf{U}$ und $\mathbf{V}$ gilt, haben wir

$$\mathbf{A}\mathbf{A}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}\mathbf{U}^* = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^*, \qquad \mathbf{A}^*\mathbf{A} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^*,$$

die linken und rechten Singulärvektoren sind also Eigenvektoren der positiv semidefiniten und selbstadjungierten Matrizen $\mathbf{A}\mathbf{A}^*$ und $\mathbf{A}^*\mathbf{A}$.

Aus dieser Beobachtung lässt sich ein Algorithmus für die iterative Berechnung einer Singulärwertzerlegung gewinnen: Wir führen die implizite QR-Iteration für die Matrix $\mathbf{G} := \mathbf{A}\mathbf{A}^*$ durch und stellen fest, dass jedes dabei auftretende Zwischenergebnis

$$\mathbf{G}^{(m)} = (\mathbf{Q}^{(m)})^*\mathbf{G}\mathbf{Q}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

wieder in faktorisierter Form vorliegt: Es gilt

$$\mathbf{G}^{(m)} = (\mathbf{Q}^{(m)})^*\mathbf{G}\mathbf{Q}^{(m)} = (\mathbf{Q}^{(m)})^*\mathbf{A}((\mathbf{Q}^{(m)})^*\mathbf{A})^* = \mathbf{A}^{(m)}(\mathbf{A}^{(m)})^* \quad \text{für alle } m \in \mathbb{N}_0$$

mit $\mathbf{A}^{(m)} = (\mathbf{Q}^{(m)})^*\mathbf{A}$. Um einen Schritt der QR-Iteration für die Matrix $\mathbf{G}$ auszuführen, müssen wir also lediglich die korrespondierenden Transformationen auf die Spalten der Matrix $\mathbf{A}$ anwenden.

An der QR-Iteration für $\mathbf{G}$ ändert sich nichts, falls wir $\mathbf{A}^{(m)}$ durch eine LQ-Zerlegung ersetzen, also durch eine Faktorisierung

$$\mathbf{A}^{(m)} = \mathbf{L}^{(m)}\mathbf{P}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0,$$

mit unteren Dreiecksmatrizen $\mathbf{L}^{(m)} \in \mathbb{K}^{n\times m}$ und unitären Matrizen $\mathbf{P}^{(m)} \in \mathbb{K}^{m\times m}$, denn es gilt

$$\mathbf{G}^{(m)} = \mathbf{A}^{(m)}(\mathbf{A}^{(m)})^* = \mathbf{L}^{(m)}\mathbf{P}^{(m)}(\mathbf{P}^{(m)})^*(\mathbf{L}^{(m)})^* = \mathbf{L}^{(m)}(\mathbf{L}^{(m)})^* \quad \text{für alle } m \in \mathbb{N}_0.$$

Die Matrizen $\mathbf{L}^{(m)}$ können wir praktisch konstruieren, indem wir die zu der QR-Iteration gehörende Transformation auf die Zeilen einer Matrix $\mathbf{L}^{(m)}$ anwenden und anschließend mit Householder-Spiegelungen die LQ-Zerlegung des Ergebnisses berechnen.

Falls nun die QR-Iteration konvergiert, werden die Matrizen $\mathbf{G}^{(m)}$ gegen Diagonalform streben. Das bedeutet, dass die Zeilen der Matrizen $\mathbf{L}^{(m)}$ näherungsweise orthogonal aufeinander stehen werden. Infolge der Dreiecksstruktur kann das nur geschehen, wenn $\mathbf{L}^{(m)}$ auch gegen eine Diagonalmatrix konvergiert.

In dieser allgemeinen Form wäre die Berechnung allerdings zu aufwendig. Analog zu der für die Effizienz der QR-Iteration entscheidenden Hessenberg-Transformation empfiehlt es sich deshalb, in einem ersten Schritt die Matrix $\mathbf{A}$ zu *bidiagonalisieren*, sie also mit unitären Transformationen $\mathbf{U}^{(0)} \in \mathbb{K}^{n\times k}$ und $\mathbf{V}^{(0)} \in \mathbb{K}^{m\times k}$ in die Form

$$(\mathbf{U}^{(0)})^* \mathbf{A}\mathbf{V}^{(0)} = \begin{pmatrix} \alpha_1 & & & \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_{k-1} & \alpha_k \end{pmatrix}$$

zu bringen. Das gelingt dem *Golub-Kahan-Bidiagonalisierungsalgorithmus* [1], mit Hilfe von Householder-Spiegelungen: Zunächst wenden wir eine Spiegelung auf die Spalten der Matrix an, um die erste Zeile in die gewünschte Form zu bringen:

$$\mathbf{A}\mathbf{H}_1 = \begin{pmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ \times & \times & \times & \dots & \times \\ \times & \times & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \times \\ \times & \times & \times & \dots & \times \end{pmatrix}.$$

In einem zweiten Schritt wenden wir eine Spiegelung auf die Zeilen der Matrix an, die die erste Zeile unverändert lässt und Nullen im Rest der ersten Spalte einführt:

$$\mathbf{H}_2^* \mathbf{A}\mathbf{H}_1 = \begin{pmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ \beta_1 & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \dots & \times \end{pmatrix}$$

Im nächsten Schritt wird wieder eine Spiegelung auf die Spalten angewendet, wobei die

[1] G. Golub und W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, J. SIAM Num. Anal. B 2(2), 205–224 (1965)

erste Spalte unverändert gelassen wird:

$$\mathbf{H}_2^*\mathbf{A}\mathbf{H}_1\mathbf{H}_3 = \begin{pmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & 0 & \dots & 0 \\ 0 & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \dots & \times \end{pmatrix}.$$

Nun ist es wieder Zeit für eine Zeilentransformation, bei der die ersten *beiden* Zeilen nicht angefasst werden:

$$\mathbf{H}_4^*\mathbf{H}_2^*\mathbf{A}\mathbf{H}_1\mathbf{H}_3 = \begin{pmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & 0 & \dots & 0 \\ 0 & \beta_2 & \times & \dots & \times \\ 0 & 0 & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \times & \dots & \times \end{pmatrix}.$$

In dieser Weise können wir fortfahren, bis die gewünschte Bidiagonalform erreicht ist.

Wenn die Matrix $\mathbf{A}^{(0)} := (\mathbf{U}^{(0)})^*\mathbf{A}\mathbf{V}^{(0)}$ in Bidiagonalgestalt gegeben ist, ist das korrespondierende Produkt $\mathbf{G}^{(0)} = \mathbf{A}^{(0)}(\mathbf{A}^{(0)})^*$ eine Tridiagonalmatrix der Form

$$\mathbf{G}^{(0)} = \begin{pmatrix} |\alpha_1|^2 & \alpha_1\bar{\beta}_1 & & & & \\ \bar{\alpha}_1\beta_1 & |\alpha_2|^2 + |\beta_1|^2 & \alpha_2\bar{\beta}_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \bar{\alpha}_{k-2}\beta_{k-2} & |\alpha_{k-1}|^2 + |\beta_{k-2}|^2 & \alpha_{k-1}\bar{\beta}_{k-1} & \\ & & & \bar{\alpha}_{k-1}\beta_{k-1} & |\alpha_k|^2 + |\beta_{k-1}|^2 \end{pmatrix},$$

so dass sich ein Schritt der QR-Iteration besonders effizient durchführen lässt. Sehr elegant wird der Algorithmus, wenn wir eine *implizite* QR-Iteration verwenden: Wir bestimmen einen geeigneten Shift-Wert und wenden die erste Givens-Rotation auf die ersten beiden Zeilen der Matrix $\mathbf{A}^{(m)}$ an. Dadurch wird die Bidiagonalstruktur verletzt, in der ersten Zeile entsteht ein Eintrag oberhalb der Diagonalen. Diesen Eintrag können wir mit einer auf die Spalten angewandten Givens-Rotation eliminieren, erhalten dabei aber einen unerwünschten Eintrag in der dritten Zeile. Eine Givens-Rotation der zweiten und dritten Zeile beseitigt ihn, erzeugt aber einen Eintrag oberhalb der Diagonalen in der zweiten Zeile. Ihn beseitigen wir mit einer Givens-Rotation der zweiten und dritten Spalte, erzeugen aber einen störenden Eintrag in der vierten Zeile. In dieser Weise können wir die problematischen von null verschiedenen Einträge „aus der Matrix heraus schieben“, wie wir es schon im Fall des impliziten QR-Verfahrens getan haben.

Falls die QR-Iteration konvergiert, werden die Matrizen $\mathbf{G}^{(m)}$ gegen Diagonalform streben. Dank der Bidiagonalstruktur ist das äquivalent dazu, dass die Produkte $\bar{\alpha}_i\beta_i$ für $i \in [1 : k-1]$ gegen null konvergieren. Falls diese Konvergenz dadurch eintritt, dass

β_i gegen null geht, dürfen wir uns glücklich schätzen, denn dann verschwindet einer der Nebendiagonaleinträge der Matrix $\mathbf{A}^{(m)}$.

Falls dagegen α_i gegen null geht, können wir mit Hilfe einiger zusätzlicher Transformationen auch den Eintrag β_i eliminieren. Angenommen, α_2 ist gleich null. Die Matrix $\mathbf{A}^{(m)}$ sieht dann wie folgt aus:

$$\mathbf{A}^{(m)} = \begin{pmatrix} \alpha_1 & & & & & & \\ \beta_1 & 0 & & & & & \\ & \beta_2 & \alpha_3 & & & & \\ & & \beta_3 & \alpha_4 & & & \\ & & & \beta_4 & \alpha_5 & & \\ & & & & \ddots & \ddots & \\ & & & & & \beta_{k-1} & \alpha_k \end{pmatrix}$$

Wir wenden eine Givens-Rotation $\mathbf{R}_{2,3}$ auf die zweite und dritte Spalte an, um β_2 zu eliminieren. Das Ergebnis ist eine Matrix der folgenden Gestalt:

$$\mathbf{A}^{(m)}\mathbf{R}_{2,3} = \begin{pmatrix} \alpha_1 & & & & & & \\ \beta_1 & 0 & & & & & \\ & 0 & \alpha_3' & & & & \\ & \gamma_2 & \beta_3' & \alpha_4 & & & \\ & & & \beta_4 & \alpha_5 & & \\ & & & & \ddots & \ddots & \\ & & & & & \beta_{k-1} & \alpha_k \end{pmatrix},$$

bei der der Eintrag γ_2 die Bidiagonalgestalt stört. Wir eliminieren ihn mit Hilfe einer Givens-Rotation, die auf die zweite und vierte Spalte wirkt:

$$\mathbf{A}^{(m)}\mathbf{R}_{2,3}\mathbf{R}_{2,4} = \begin{pmatrix} \alpha_1 & & & & & & \\ \beta_1 & 0 & & & & & \\ & 0 & \alpha_3' & & & & \\ & 0 & \beta_3' & \alpha_4' & & & \\ & \gamma_3 & & \beta_4' & \alpha_5 & & \\ & & & & \ddots & \ddots & \\ & & & & & \beta_{k-1} & \alpha_k \end{pmatrix},$$

In dieser Weise können wir fortfahren, bis wir den Eintrag in der letzten Zeile der Matrix eliminiert haben und die Bidiagonalgestalt wiederhergestellt ist. Aus einer Null auf der Diagonalen können wir als eine auf der Nebendiagonalen gewinnen und so die Matrix analog zu der bei der QR-Iteration eingesetzten Deflation in unabhängige Diagonalblöcke zerlegen.

7 Verfahren für Tridiagonalmatrizen

Im vorigen Kapitel haben wir gesehen, dass sich eine beliebige selbstadjungierte Matrix mit Hilfe von Householder-Spiegelungen in eine Tridiagonalmatrix überführen lässt. Für Tridiagonalmatrizen lassen sich nicht nur die verschiedenen Varianten der QR-Iteration effizient durchführen, es existieren auch eine Reihe weiterer Verfahren, die die Tridiagonalgestalt ausnutzen können.

Ein Beispiel sind Bisektionsverfahren, mit deren Hilfe sich nach Nullstellen des charakteristischen Polynoms p_A suchen lässt. Nach Lemma 3.5 sind diese Nullstellen gerade die Eigenwerte der Matrix $\mathbf{A}$. Ein besonderer Vorteil dieser Verfahren besteht darin, dass wir mit ihrer Hilfe nicht nur nach dem kleinsten oder größten Eigenwert suchen können, sondern auch nach dem k-ten.

7.1 Auswertung des charakteristischen Polynoms

Wir fixieren eine selbstadjungierte Tridiagonalmatrix

$$\mathbf{A} = \begin{pmatrix} \alpha_1 & \bar{\beta}_1 & & & \\ \beta_1 & \alpha_2 & \bar{\beta}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n-2} & \alpha_{n-1} & \bar{\beta}_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{pmatrix}. \tag{7.1}$$

Für ein $i \in [1:n]$ und ein $x \in \mathbb{R}$ bezeichnen wir die Determinante der i-ten Hauptuntermatrix von $x\mathbf{I} - \mathbf{A}$ mit

$$p_i(x) := \det \begin{pmatrix} x-\alpha_1 & -\bar{\beta}_1 & & & \\ -\beta_1 & x-\alpha_2 & -\bar{\beta}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -\beta_{i-2} & x-\alpha_{i-1} & -\bar{\beta}_{i-1} \\ & & & -\beta_{i-1} & x-\alpha_i \end{pmatrix}.$$

Offenbar ist p_n gerade das charakteristische Polynom p_A. Für $i > 2$ erhalten wir durch Entwicklung nach der letzten Spalte und anschließend der letzten Zeile den Zusammen-

hang

$$
\begin{aligned}
p_i(x) := \det &\left(\begin{array}{ccccccc|c}
x-\alpha_1 & -\bar\beta_1 & & & & & & \\
-\beta_1 & x-\alpha_2 & -\bar\beta_2 & & & & & \\
& \ddots & \ddots & \ddots & & & & \\
& & -\beta_{i-3} & x-\alpha_{i-2} & -\bar\beta_{i-2} & & & \\
& & & -\beta_{i-2} & x-\alpha_{i-1} & -\bar\beta_{i-1} \\
& & & & -\beta_{i-1} & x-\alpha_i
\end{array}\right) \\
&= (x-\alpha_i)\det\begin{pmatrix}
x-\alpha_1 & -\bar\beta_1 & & & & \\
-\bar\beta_1 & x-\alpha_2 & -\bar\beta_2 & & & \\
& \ddots & \ddots & \ddots & & \\
& & -\beta_{i-3} & x-\alpha_{i-2} & -\bar\beta_{i-2} & \\
& & & -\beta_{i-2} & x-\alpha_{i-1}
\end{pmatrix} \\
&\quad + \bar\beta_{i-1}\det\left(\begin{array}{ccccc}
x-\alpha_1 & -\bar\beta_1 & & & \\
-\beta_1 & x-\alpha_2 & -\bar\beta_2 & & \\
& \ddots & \ddots & \ddots & \\
& & -\beta_{i-3} & x-\alpha_{i-2} & -\bar\beta_{i-2} \\
\hline
& & & & -\beta_{i-1}
\end{array}\right) \\
&= (x-\alpha_i)\det\begin{pmatrix}
x-\alpha_1 & -\bar\beta_1 & & & & \\
-\beta_1 & x-\alpha_2 & -\bar\beta_2 & & & \\
& \ddots & \ddots & \ddots & & \\
& & -\beta_{i-2} & x-\alpha_{i-2} & -\bar\beta_{i-2} & \\
& & & -\beta_{i-2} & x-\alpha_{i-1}
\end{pmatrix} \\
&\quad - |\beta_{i-1}|^2\det\begin{pmatrix}
x-\alpha_1 & -\bar\beta_1 & & & & \\
-\beta_1 & x-\alpha_2 & -\bar\beta_2 & & & \\
& \ddots & \ddots & \ddots & & \\
& & -\beta_{i-3} & x-\alpha_{i-3} & -\bar\beta_{i-3} & \\
& & & -\beta_{i-3} & x-\alpha_{i-2}
\end{pmatrix} \\
&= (x-\alpha_i)p_{i-1}(x) - |\beta_{i-1}|^2 p_{i-2}(x).
\end{aligned}
$$

Wenn wir zur Vereinfachung $p_0 = 1$ einführen, erhalten wir die Rekursionsformel

$$
p_i(x) = \begin{cases}
1 & \text{falls } i = 0, \\
x-\alpha_1 & \text{falls } i = 1, \\
(x-\alpha_i)p_{i-1}(x) - |\beta_{i-1}|^2 p_{i-2}(x) & \text{ansonsten}
\end{cases} \quad \text{für alle } i \in [0:n], x \in \mathbb{R}, \tag{7.2}
$$

aus der sich der folgende Algorithmus zur Auswertung des Tupels $(p_n(x), \ldots, p_0(x))$ ergibt:

Algorithmus 7.1 (Auswertung von p_i) *Der folgende Algorithmus berechnet das Tupel* $\mathbf{p} = (p_0(x), \ldots, p_n(x))$ *für ein beliebiges* $x \in \mathbb{R}$:

$p_0 \leftarrow 1; \quad p_1 \leftarrow x - \alpha_1$
for $i = 2$ *to* n *do*
  $p_i \leftarrow (x - \alpha_i)p_{i-1} - |\beta_{i-1}|^2 p_{i-2}$

Insbesondere lässt sich mit Hilfe von Algorithmus 7.1 das charakteristische Polynom $p_A = p_n$ in $\mathcal{O}(n)$ Operationen auswerten, und nebenbei werden die charakteristischen Polynome aller Hauptuntermatrizen berechnet, die sich für bestimmte Algorithmen als sehr nützlich erweisen können.

Da Eigenwerte von $\mathbf{A}$ Nullstellen des charakteristischen Polynoms p_A sind, das wir mit Algorithmus 7.1 elegant und effizient auswerten können, bietet es sich an, Standardverfahren zur Nullstellenberechnung auf p_A anzuwenden.

Eines der einfachsten und trotzdem zuverlässigsten Verfahren ist die Bisektion, bei der eine Folge von Intervallen berechnet wird, die jeweils mindestens eine Nullstelle enthalten: Wir beginnen mit einem Intervall $[a, b]$ und fordern, dass $p_A(a)$ und $p_A(b)$ unterschiedliche Vorzeichen besitzen. Dann muss nach Mittelwertsatz in $[a, b]$ eine Nullstelle von p_A enthalten sein. Wir unterteilen $[a, b]$ in zwei Teilintervall $[a, c]$ und $[c, b]$ mit $c = (b + a)/2$ und fahren mit demjenigen Intervall fort, für das die Vorzeichenbedingung immer noch erfüllt ist.

Algorithmus 7.2 (Bisektion) *Seien $a, b \in \mathbb{R}$ mit $a < b$ und $p_A(a)p_A(b) < 0$ gegeben. Dann approximiert der folgende Algorithmus einen Eigenwert von $\mathbf{A}$ im Intervall $[a, b]$:*

$p_a \leftarrow p_A(a); \quad p_b \leftarrow p_A(b)$
while $|b - a| > \epsilon$ *do begin*
  $c \leftarrow (b + a)/2$
  $p_c \leftarrow p_A(c)$
  if $p_a p_c < 0$ *then begin*
    $b \leftarrow c; \quad p_b \leftarrow p_c$
  end else begin
    $a \leftarrow c; \quad p_a \leftarrow p_c$
  end
end

Dieser Algorithmus benötigt pro Iterationsschritt nur eine Auswertung des Polynoms p_A, die sich, wie wir bereits gesehen haben, in $\mathcal{O}(n)$ Operationen durchführen lässt. Da sich mit jedem Schritt das Intervall halbiert, können wir in $\mathcal{O}(\log_2(1/\epsilon))$ Iterationen eine Genauigkeit von $\epsilon \in \mathbb{R}_{>0}$ erreichen.

Ein Vorteil dieses Algorithmus besteht darin, dass wir das zu untersuchende Intervall explizit vorgeben können und dass er sehr stabil arbeitet, falls die Auswertung von p_A stabil erfolgt. Ein Nachteil besteht darin, dass nicht klar ist, wie man ein Startintervall $[a, b]$ finden kann, das die benötigte Vorzeichenbedingung $p_A(a)p_A(b) < 0$ erfüllt.

Durch Differenzieren der Rekursionsformel für p_A können wir die Rekursionsformel

$$p_i'(x) = \begin{cases} 0 & \text{falls } i = 0, \\ 1 & \text{falls } i = 1, \\ p_{i-1}(x) + (x - \alpha_i)p_{i-1}'(x) & \\ \qquad - |\beta_{i-1}|^2 p_{i-2}'(x) & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in \mathbb{N}_0, x \in \mathbb{R} \tag{7.3}$$

gewinnen, mit der sich die erste Ableitung von p_A ebenfalls effizient berechnen lässt, so dass wir statt der Bisektion auch das Newton-Verfahren verwenden können. Zwar konvergiert das Newton-Verfahren unter Umständen wesentlich schneller als das Bisektionsverfahren, aber das passiert in der Regel nur, wenn ein guter Startwert vorliegt. Auch bei diesem Zugang stellt sich also die Frage nach geeigneten Startwerten.

7.2 Sturmsche Ketten

Wir sind daran interessiert, das einfache Bisektionsverfahren so zu modifizieren, dass wir nicht mehr auf eine Vorzeichenbedingung angewiesen sind, sondern eine Voraussetzung finden, die einfacher zu erfüllen ist.

Dazu gehen wir zunächst davon aus, dass alle Nullstellen von p_A einfach sind und wir sie deshalb in die Reihenfolge

$$\lambda_1 < \lambda_2 < \ldots < \lambda_n$$

bringen können. Nach dem Satz von Rolle liegt zwischen zwei dieser Nullstellen jeweils mindestens eine Nullstelle der Ableitung p_A' des charakteristischen Polynoms, wir können also für jedes $i \in [1 : n-1]$ ein $\lambda_i^{(1)} \in (\lambda_i, \lambda_{i+1})$ mit $p_A'(\lambda_i^{(1)}) = 0$ finden. Wir erhalten

$$\lambda_1 < \lambda_1^{(1)} < \lambda_2 < \lambda_2^{(1)} < \lambda_3 < \ldots < \lambda_{n-1} < \lambda_{n-1}^{(1)} < \lambda_n.$$

Da p_A' nur noch ein Polynom der Ordnung $n-1$ ist, kann es keine weiteren Nullstellen außer $\lambda_1^{(1)}, \ldots, \lambda_{n-1}^{(1)}$ besitzen.

In ähnlicher Weise können wir beweisen, dass die Nullstellen der $(i+1)$-ten Ableitung von p_A gerade die der i-ten Ableitung trennen, solange $i < n$ gilt.

Falls ξ eine einfache Nullstelle von p_A ist, können wir ein $\epsilon > 0$ so finden, dass p_A' in $[\xi - \epsilon, \xi + \epsilon]$ keine Nullstelle aufweist. Mit dem Mittelwertsatz der Differentialrechnung finden wir $\eta_+ \in [\xi, \xi + \epsilon]$ und $\eta_- \in [\xi - \epsilon, \xi]$ so, dass

$$\begin{aligned} \frac{p_A(\xi + \epsilon) - p_A(\xi)}{\epsilon} &= p_A'(\eta_+), & \frac{p_A(\xi) - p_A(\xi - \epsilon)}{\epsilon} &= p_A'(\eta_-), \\ p_A(\xi + \epsilon) &= p_A(\xi) + \epsilon p_A'(\eta_+), & p_A(\xi - \epsilon) &= p_A(\xi) - \epsilon p_A'(\eta_-) \end{aligned}$$

gilt, und es folgt

$$\begin{aligned} p_A(\xi + \epsilon)p_A(\xi - \epsilon) &= p_A(\xi)^2 + \epsilon p_A(\xi)(p_A'(\eta_+) - p_A'(\eta_-)) - \epsilon^2 p_A'(\eta_+)p_A'(\eta_-) \\ &= -\epsilon^2 p_A'(\eta_+)p_A'(\eta_-) < 0, \end{aligned}$$

weil ξ eine Nullstelle des Polynoms p_A ist und $p'_A(\eta_+)$ und $p'_A(\eta_-)$ dasselbe Vorzeichen besitzen. An jeder einfachen Nullstelle wechselt p_A also das Vorzeichen. Wenn wir Nullstellen zählen wollen, bietet es sich demnach an, nach Vorzeichenwechseln zu suchen.

Es lässt sich leicht nachprüfen, dass der führende Koeffizient des charakteristischen Polynoms gerade 1 ist, so dass

$$\lim_{x\to\infty} p_A^{(m)}(x) = \infty \qquad \text{für alle } m \in [0, n-1]$$

gilt. Wir haben bereits gesehen, dass die Nullstellen der Ableitung $p_A^{(m+1)}$ zwischen denen der Funktion $p_A^{(m)}$ liegen, also muss insbesondere

$$p_A^{(m)}(x) > 0 \qquad \text{für alle } x > \lambda_n,\ m \in [0, n]$$

gelten. Für hinreichend großes x sind also die Vorzeichen aller Ableitungen des Polynoms p_A identisch.

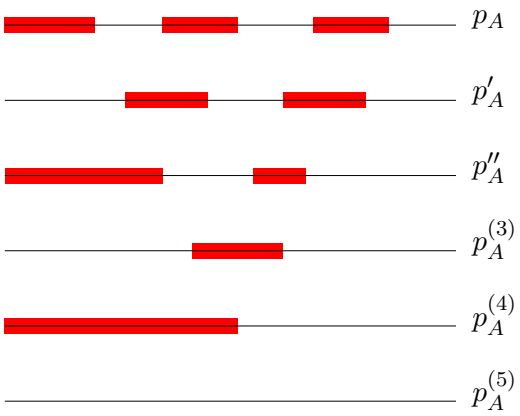


Abbildung 7.1: Vorzeichen eines charakteristischen Polynoms und seiner Ableitungen im Fall $n = 5$. Rot markiert Bereiche, in denen das Vorzeichen negativ ist.

In Abbildung 7.1 sind die Vorzeichen eines charakteristischen Polynoms und seiner Ableitungen für den Fall $n = 5$ dargestellt. Die Bereiche, in denen die einzelnen Funktionen negativ sind, sind rot markiert. Für unsere Zwecke von Bedeutung ist die Beobachtung, dass in jedem Punkt x die Anzahl der Vorzeichenwechsel *zwischen den Ableitungen* $p_A(x), p'_A(x), p''_A(x), \ldots, p_A^{(n)}(x)$ gerade die Anzahl der Nullstellen größer als x angibt.

Diese Eigenschaft verdanken wir der Tatsache, dass zwischen zwei einfachen Nullstellen unserer Polynome jeweils genau eine einfache Nullstelle seiner Ableitung liegt, so dass „rechts“ von einer Nullstelle von $p_A^{(m)}$ immer die Vorzeichen von $p_A^{(m)}$ und $p_A^{(m+1)}$ übereinstimmen. Falls $m > 1$ gilt, sind die Vorzeichen von $p_A^{(m)}$ und $p_A^{(m-1)}$ hingegen ungleich. Für $m > 1$ wird also ein Vorzeichenwechsel zwischen $p_A^{(m)}$ und $p_A^{(m-1)}$ durch

einen zwischen $p_A^{(m)}$ und $p_A^{(m+1)}$ ersetzt, lediglich für $m = 0$ reduziert sich die Gesamtzahl der Vorzeichenwechsel. Falls eine exakte Null auftreten sollte, können wir festlegen, ob sie als positiv oder negativ gelten soll.

Die Ableitungen des charakteristischen Polynoms können wir zwar im Prinzip berechnen, sehr viel eleganter ist es allerdings, festzustellen, dass die bei seiner Auswertung berechneten Hilfspolynoms p_m die Rolle der Ableitungen $p_A^{(n-m)}$ übernehmen können. Dadurch können wir die Vorzeichenwechsel mit sehr geringem zusätzlichen Aufwand zählen.

Die für uns interessanten Eigenschaften eines Tupels von Funktionen fasst die folgende Definition zusammen:

Definition 7.3 (Sturmsche Kette) *Ein Tupel $(p_0, p_1, \ldots, p_n)$ von reellen Polynomen heißt* Sturmsche Kette*, wenn*

1. $\overline{p_{n-1}(\xi)} p_n'(\xi) > 0$ *für alle Nullstellen $\xi \in \mathbb{K}$ von p_n erfüllt ist,*
2. $\overline{p_{i+1}(\xi)} p_{i-1}(\xi) < 0$ *für alle $i \in [1 : n-1]$ und Nullstellen $\xi \in \mathbb{K}$ von p_i gilt, sowie*
3. *p_0 keine Nullstelle besitzt.*

Bedingung 1 stellt sicher, dass in den Nullstellen des Polynom p_n die Vorzeichen der Ableitung p_n' und des Polynoms p_{n-1} übereinstimmen. Insbesondere müssen diese Nullstellen einfach sein, da p_n' nicht gleich null sein kann.

Bedingung 2 sorgt dafür, dass bei einer Nullstelle eines Polynoms p_i die „benachbarten“ Polynome p_{i-1} und p_{i+1} entgegengesetzte Vorzeichen aufweisen, so dass die Gesamtzahl der Vorzeichenwechsel unverändert bleibt.

Bedingung 3 schließlich benötigen wir, um ein „Referenzvorzeichen“ festzulegen, an dem wir die anderen Vorzeichen messen.

Nun stehen wir vor der Aufgabe, nachzuweisen, dass die durch die Rekursionsformel (7.2) definierten Polynome eine Sturmsche Kette bilden.

Lemma 7.4 (Sturmsche Kette) *Falls $\mathbf{A}$ irreduzibel ist, falls also $\beta_i \neq 0$ für alle $i \in [1 : n-1]$ gilt, bilden die durch (7.2) definierten Polynome $p_0, \ldots, p_n$ eine Sturmsche Kette.*

Beweis. Der Beweis beruht darauf, Eigenvektoren zu den Eigenwerten der Matrix $\mathbf{A}$ zu konstruieren. Für ein $x \in \mathbb{K}$ suchen wir einen Vektor $\mathbf{e}(x) \in \mathbb{K}^n$, der „möglichst nahe“ am Kern von $x\mathbf{I} - \mathbf{A}$ liegt, für den nämlich die ersten $n-1$ Komponenten des Vektors $(x\mathbf{I} - \mathbf{A})\mathbf{e}(x)$ verschwinden.

Damit $\mathbf{e}(x)$ nicht der Nullvektor wird, setzen wir $e_1(x) = 1$. Durch Einsetzen in die erste Zeile der Matrix erhalten wir nun

$$(x - \alpha_1)e_1(x) - \bar{\beta}_1 e_2(x) = 0,$$
$$e_2(x) = \frac{x - \alpha_1}{\bar{\beta}_1} e_1(x) = \frac{p_1(x)}{\bar{\beta}_1}$$

während Einsetzen in die i-te Zeile für $i \in [2 : n-1]$ die Gleichung

$$(-\beta_{i-1}e_{i-1}(x) + (x-\alpha_i)e_i(x) - \bar{\beta}_i e_{i+1}(x)) = 0,$$
$$e_{i+1}(x) = \frac{(x-\alpha_i)e_i(x) - \beta_{i-1}e_{i-1}(x)}{\bar{\beta}_i}$$

für $i \in [1 : n-1]$ ergibt. Im Zähler dieses Terms erkennen wir Teile der Rekursionsformel (7.2) wieder und wählen den Ansatz

$$e_i(x) = \frac{p_{i-1}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{i-1}} \qquad \text{für } i \in [2 : n].$$

Mit ihm nimmt unsere Gleichung die Form

$$\begin{aligned} e_{i+1}(x) &= \frac{(x-\alpha_i)p_{i-1}(x)/(\bar{\beta}_1 \dots \bar{\beta}_{i-1}) - \beta_{i-1}p_{i-2}(x)/(\bar{\beta}_1 \dots \bar{\beta}_{i-2})}{\bar{\beta}_i} \\ &= \frac{(x-\alpha_i)p_{i-1}(x)/(\bar{\beta}_1 \dots \bar{\beta}_{i-1}) - |\beta_{i-1}|^2 p_{i-2}(x)/(\bar{\beta}_1 \dots \bar{\beta}_{i-1})}{\bar{\beta}_i} \\ &= \frac{(x-\alpha_i)p_{i-1}(x) - |\beta_{i-1}|^2 p_{i-2}(x)}{\bar{\beta}_1 \dots \bar{\beta}_i} = \frac{p_i(x)}{\bar{\beta}_1 \dots \bar{\beta}_i} \end{aligned}$$

annimmt. Mit einer einfachen Induktion folgt, dass der durch

$$e_i(x) = \begin{cases} 1 & \text{falls } i = 1, \\ \frac{p_{i-1}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{i-1}} & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in [1 : n]$$

definierte Vektor die gewünschten Gleichungen erfüllt.

Für die n-te Zeile schließlich erhalten wir

$$\begin{aligned} -\beta_{n-1}e_{n-1}(x) + (x-\alpha_n)e_n(x) &= -\beta_{n-1}\frac{p_{n-2}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{n-2}} + (x-\alpha_n)\frac{p_{n-1}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{n-1}} \\ &= -|\beta_{n-1}|^2\frac{p_{n-2}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{n-1}} + (x-\alpha_n)\frac{p_{n-1}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{n-1}} \\ &= \frac{(x-\alpha_n)p_{n-1}(x) - |\beta_{n-1}|^2 p_{n-2}(x)}{\bar{\beta}_1 \dots \bar{\beta}_{n-1}} \\ &= \frac{p_n(x)}{\bar{\beta}_1 \dots \bar{\beta}_{n-1}} =: \gamma(x). \end{aligned}$$

Falls $p_n(x) = 0$ gilt, haben wir auch $\gamma(x) = 0$, also folgt die Gleichung $(x\mathbf{I} - \mathbf{A})\mathbf{e}(x) = \mathbf{0}$ und $\mathbf{e}(x)$ ist ein Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert x. Im allgemeinen Fall haben wir

$$(x\mathbf{I} - \mathbf{A})\mathbf{e}(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \gamma(x) \end{pmatrix} \qquad \text{für alle } x \in \mathbb{K} \tag{7.4}$$

bewiesen. Indem wir diese Funktion nach x differenzieren, erhalten wir

$$\mathbf{e}(x) + (x\mathbf{I} - \mathbf{A})\mathbf{e}'(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \gamma'(x) \end{pmatrix} \qquad \text{für alle } x \in \mathbb{K}. \tag{7.5}$$

Sei nun $\xi \in \mathbb{K}$ eine Nullstelle des Polynoms p_n, also auch des Polynoms γ. Indem wir (7.5) mit $\mathbf{e}(\xi)$ multiplizieren erhalten wir wegen $(x\mathbf{I} - \mathbf{A})\mathbf{e}(x) = \mathbf{0}$ und da $\mathbf{A}$ selbstadjungiert ist die Gleichung

$$\begin{aligned} 0 < \|\mathbf{e}(\xi)\|^2 &= \|\mathbf{e}(\xi)\|^2 + \langle(\xi\mathbf{I} - \mathbf{A})\mathbf{e}(\xi), \mathbf{e}'(\xi)\rangle = \|\mathbf{e}(\xi)\|^2 + \langle\mathbf{e}(\xi), (\xi\mathbf{I} - \mathbf{A})\mathbf{e}'(\xi)\rangle \\ &= \langle\mathbf{e}(\xi), \mathbf{e}(\xi) + (\xi\mathbf{I} - \mathbf{A})\mathbf{e}'(\xi)\rangle = \overline{e_n(\xi)}\gamma'(\xi) = \frac{\overline{p_{n-1}(\xi)}}{\beta_1 \dots \beta_{n-1}} \frac{p_n'(\xi)}{\bar{\beta}_1 \dots \bar{\beta}_{n-1}} \\ &= \frac{\overline{p_{n-1}(\xi)} p_n'(\xi)}{|\beta_1|^2 \dots |\beta_{n-1}|^2}, \end{aligned}$$

also insbesondere Bedingung 1 der Definition 7.3.

Zum Nachweis der Bedingung 2 dieser Definition setzen wir die Rekursionsformel (7.2) ein. Sei $i \in [1 : n-1]$, und sei $\xi \in \mathbb{K}$ eine Nullstelle von p_i. Dann folgt aus (7.2) die Gleichung

$$p_{i+1}(\xi) = -|\beta_i|^2 p_{i-1}(\xi),$$

also

$$\overline{p_{i+1}(\xi)} p_{i-1}(\xi) = -|\beta_i|^2 \overline{p_{i-1}(\xi)} p_{i-1}(\xi) = -|\beta_i|^2 |p_{i-1}(\xi)|^2 \leq 0.$$

Falls nun $p_{i+1}(\xi) = 0$ gelten würde, hätten wir wegen $\beta_i \neq 0$ auch $p_{i-1}(\xi) = 0$ und könnten mit der Rekurrenzform (7.2) induktiv fortfahren, um zu dem Widerspruch $0 = p_{i+1}(\xi) = p_i(\xi) = \ldots = p_0(\xi) = 1$ zu gelangen. Also müssen $p_{i+1}(\xi), p_{i-1}(\xi) \neq 0$ gelten, und damit $\overline{p_{i+1}(\xi)} p_{i-1}(\xi) < 0$. ∎

Aus diesem Lemma lassen sich bereits erste nützliche Aussagen über selbstadjungierte Tridiagonalmatrizen gewinnen.

Folgerung 7.5 (Einfache Eigenwerte) *Eine irreduzible selbstadjungierte Tridiagonalmatrix besitzt nur einfache Eigenwerte.*

Beweis. Mit Lemma 7.4 folgt die Aussage unmittelbar aus der ersten Bedingung in Definition 7.3. ∎

Nach Satz 3.43 wissen wir bereits, dass selbstadjungierte Matrizen reell diagonalisierbar sind, dass also charakteristischen Polynome in reelle Linearfaktoren zerfallen. Bei einer irreduziblen selbstadjungierten Tridiagonalmatrix sind diese Linearfaktoren alle unterschiedlich, so dass wir reelle Eigenwerte $\lambda_1 < \lambda_2 < \ldots < \lambda_n$ finden.

Folgerung 7.6 (Trennungseigenschaft) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine irreduzible selbstadjungierte Tridiagonalmatrix. Seien $\lambda_1 < \lambda_2 < \ldots < \lambda_n$ ihre Eigenwerte und $\lambda'_1 < \lambda'_2 < \ldots < \lambda'_{n-1}$ die Eigenwerte ihrer $(n-1)$-ten Hauptuntermatrix $\mathbf{A}_{n-1}$. Dann gilt*

$$\lambda_1 < \lambda'_1 < \lambda_2 < \ldots < \lambda_{n-1} < \lambda'_{n-1} < \lambda_n.$$

Beweis. Sei $i \in [1 : n-1]$. Nach der Bedingung 1 der Definition 7.3 können $p'_A(\lambda_i)$ und $p'_A(\lambda_{i+1})$ nicht gleich null sein, λ_i und λ_{i+1} sind also einfache Nullstellen des Polynoms p_A. Nach unserer Vorbetrachung besitzt p'_A dann genau eine einfache Nullstelle in $(\lambda_i, \lambda_{i+1})$, also müssen sich die Vorzeichen von $p'_A(\lambda_i)$ und $p'_A(\lambda_{i+1})$ unterscheiden.

Nach der Bedingung 1 der Definition 7.3 weist p_{n-1} in λ_i und λ_{i+1} dieselben Vorzeichen wie $p'_A = p'_n$ auf, also muss auch p_{n-1} mindestens eine Nullstelle λ'_i in $(\lambda_i, \lambda_{i+1})$ besitzen.

Damit haben wir $n-1$ Nullstellen von p_{n-1} gefunden, und da p_{n-1} höchstens den Grad $n-1$ aufweist und nicht das Nullpolynom ist, können nach dem Identitätssatz für Polynome keine weiteren Nullstellen existieren. ∎

Die für uns entscheidende Eigenschaft der Sturmschen Kette besteht darin, dass ihre Vorzeichenwechsel eine Beziehung zu der Anzahl der Nullstellen besitzen, so dass wir Aussagen darüber treffen können, wieviele Nullstellen in einem Intervall liegen, ohne sie explizit berechnen zu müssen.

Satz 7.7 (Nullstellenzähler) *Sei $(p_0, p_1, \ldots, p_n)$ eine Sturmsche Kette. Wir definieren für alle $x \in \mathbb{R}$*

$$W_x := \{i \in [0 : n-1] \ : \ p_i(x)p_{i+1}(x) < 0 \text{ oder } p_i(x) = 0\}$$

und die Funktion

$$w : \mathbb{R} \to \mathbb{N}_0, \quad x \mapsto |W_x|,$$

die x die Mächtigkeit von W_x zuordnet. Seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben. Dann besitzt p_n genau $w(a) - w(b)$ Nullstellen im Intervall $(a, b]$.

Beweis. Offensichtlich ändert sich w nur, wenn eines der Polynome $(p_i)_{i=0}^n$ sein Vorzeichen ändert, also eine Nullstelle passiert. Wir definieren für alle $x \in \mathbb{R}$ die Menge

$$N_x := \{i \in [0 : n] \ : \ p_i(x) = 0\}.$$

Infolge der Bedingung 3 in Definition 7.3 gilt $0 \notin N_x$ für alle $x \in \mathbb{R}$. Wir werden nun nachweisen, dass die Mächtigkeit von W_x genau dann um eins sinkt, wenn x eine Nullstelle von p_n passiert.

Sei $\xi \in \mathbb{R}$ mit $N_\xi \neq \emptyset$. Da nicht-konstante Polynome nur endlich viele Nullstellen besitzen können, können sich die Nullstellen nicht häufen, also muss ein $\epsilon \in \mathbb{R}_{>0}$ so existieren, dass $N_x = \emptyset$ für alle $x \in [\xi - \epsilon, \xi + \epsilon] \setminus \{\xi\}$ gilt.

Sei $i \in N_\xi$. Falls $i < n$ gilt, folgt aus Bedingung 2 in Definition 7.3 die Ungleichung $p_{i-1}(\xi)p_{i+1}(\xi) < 0$, also insbesondere $i - 1 \notin N_\xi$ und $i + 1 \notin N_\xi$. Nach Wahl von ϵ besitzen dann p_{i-1} und p_{i+1} keine Nullstellen in $[\xi - \epsilon, \xi + \epsilon]$, so dass wir

$$p_{i-1}(x)p_{i+1}(x) < 0 \qquad \text{für alle } x \in [\xi - \epsilon, \xi + \epsilon]$$

erhalten. Sei $x \in [\xi - \epsilon, \xi + \epsilon]$.

Falls $p_{i-1}(x)p_i(x) < 0$ gilt, folgt $p_{i+1}(x)p_i(x) > 0$, also $|W_x \cap \{i-1, i\}| = |\{i-1\}| = 1$. Falls $p_{i-1}(x)p_i(x) > 0$ gilt, folgt $p_{i+1}(x)p_i(x) < 0$, also $|W_x \cap \{i-1, i\}| = |\{i\}| = 1$. Falls schließlich $p_{i-1}(x)p_i(x) = 0$ gilt, folgt $p_i(x) = 0$, also $|W_x \cap \{i-1, i\}| = |\{i\}| = 1$. Also ändern Nullstellen von p_i für $i > 0$ nichts an der Mächtigkeit der Menge W_x.

Falls $i = n$ gilt, folgt aus der Bedingung 1 in Definition 7.3 die Ungleichung $p_{n-1}(\xi)p'_n(\xi) > 0$, also insbesondere $p_{n-1}(\xi) \neq 0$ und damit $n-1 \notin N_\xi$. Da auch $p'_n(\xi) \neq 0$ gilt, können wir ein $\delta \in (0, \epsilon]$ so wählen, dass p'_n in $[\xi - \delta, \xi + \delta]$ keine Nullstellen besitzt. Dann folgt aus der Stetigkeit der Ableitung

$$p_{n-1}(\xi)p'_n(y) > 0 \qquad \text{für alle } y \in [\xi - \delta, \xi + \delta].$$

Sei $x \in (\xi, \xi + \delta]$. Mit dem Mittelwertsatz der Differentialrechnung finden wir ein $\eta_+ \in [\xi, \xi + \delta]$ mit

$$p_{n-1}(\xi)p_n(x) = p_{n-1}(\xi)(p_n(x) - p_n(\xi)) = p_{n-1}(\xi)\,(x - \xi)p'_n(\eta_+) > 0.$$

Für $x \in [\xi - \delta, \xi)$ finden wir entsprechend ein $\eta_- \in [\xi - \delta, \xi]$ mit

$$p_{n-1}(\xi)p_n(x) = p_{n-1}(\xi)(p_n(x) - p_n(\xi)) = p_{n-1}(\xi)\,(x - \xi)p'_n(\eta_-) < 0.$$

Da p_{n-1} nach Definition keine Nullstellen in $[\xi-\delta, \xi+\delta]$ besitzen kann, folgt per Stetigkeit schließlich

$$\begin{aligned}
&p_{n-1}(x)p_n(x) < 0, \text{ also } W_x \cap \{n-1, n\} = \{n-1\} && \text{für alle } x \in [\xi - \delta, \xi),\\
&p_{n-1}(x)p_n(x) > 0, \text{ also } W_x \cap \{n-1, n\} = \emptyset && \text{für alle } x \in (\xi, \xi + \delta],\\
&p_{n-1}(\xi)p_n(\xi) = 0, \text{ also } W_x \cap \{n-1, n\} = \emptyset.
\end{aligned}$$

Also reduzieren Nullstellen des Polynoms p_n die Mächtigkeit der Menge W_x um eins. ∎

Der Satz 7.7 enthält lediglich eine Aussage über die Differenzen $w(b) - w(a)$ der Funktion w. Das folgende Lemma bestimmt w näher:

Lemma 7.8 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine irreduzible selbstadjungierte Tridiagonalmatrix, sei λ_1 ihr kleinster Eigenwert und λ_n ihr größter. Für alle $x \in \mathbb{R}_{<\lambda_1}$ gilt $w(x) = n$, und für alle $x \in \mathbb{R}_{\geq\lambda_n}$ gilt $w(x) = 0$.*

Beweis. Nach Konstruktion des Polynoms p_i hat sein führender Koeffizient ein positives Vorzeichen. Für $x \to \infty$ muss also $p_i(x) \to \infty$ gelten, somit existiert ein $x_0 \in \mathbb{R}$ mit

$$p_i(x) > 0 \qquad \text{für alle } i \in [0:n],\ x \in \mathbb{R}_{\geq x_0}.$$

Es folgt $w(x_0) = 0$. Da w seinen Wert nur bei Nullstellen von p_n ändert, folgt die zweite Aussage. Da p_n genau n einfache Nullstellen besitzt, folgt die erste Aussage direkt aus Satz 7.7. ∎

Auf dieser Grundlage können wir einen Algorithmus konstruieren, der einen beliebigen Eigenwert bestimmt:

Sei $k \in [1:n]$, und seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben.

Falls $w(b) \leq n - k$ gilt, enthält das unendliche Intervall $(-\infty, b]$ gerade $n - w(b) \geq k$ Eigenwerte, also insbesondere auch den k-ten Eigenwert, den wir suchen.

Falls $w(a) > n - k$ gilt, enthält das unendliche Intervall $(-\infty, a]$ gerade $n - w(a) < k$ Eigenwerte, also gerade *nicht* den k-ten Eigenwert, den wir suchen.

Also muss der k-te Eigenwert in dem Intervall $(a, b] = (-\infty, b] \setminus (-\infty, a]$ liegen.

Algorithmus 7.9 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine irreduzible selbstadjungierte Tridiagonalmatrix, seien $\lambda_1 < \ldots < \lambda_n$ ihre Eigenwerte. Sei $k \in [1, n]$, und seien Intervallgrenzen $a, b \in \mathbb{R}$ mit $a < b$, $w(b) \leq n - k < w(a)$ gegeben. Dann bestimmt der folgende Algorithmus eine Folge von $(a, b]$, die λ_k enthalten:*

```
while |b − a| > ϵ do begin
  c ← (b + a)/2
  if w(c) ≤ n − k then
    b ← c
  else
    a ← c
end
```

Auch dieser Bisektionsalgorithmus halbiert den Fehler in jedem Schritt und benötigt dank Algorithmus 7.1 nur $\mathcal{O}(n)$ Operationen dafür. Allerdings ist auch er auf geeignete Startwerte angewiesen und nur auf selbstadjungierte irreduzible Tridiagonalmatrizen anwendbar.

Wir haben bereits gesehen, dass wir jede beliebige selbstadjungierte Matrix mit unitären Ähnlichkeitstransformationen auf Tridiagonalgestalt bringen können. Falls die Tridiagonalmatrix nicht irreduzibel ist, können wir sie in eine Blockdiagonalmatrix mit irreduziblen Diagonalblöcken zerlegen.

Damit bleibt nur noch zu klären, wie wir ein geeignetes Anfangsintervall finden. Einen einfachen Zugang bieten *Gerschgorin-Kreise*, die uns eine Möglichkeit bieten, das Spektrum einer beliebigen Matrix einzugrenzen.

Satz 7.10 (Gerschgorin-Kreise) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Die abgeschlossenen Kreisscheiben*

$$\mathcal{D}_i := \{z \in \mathbb{C} \ : \ |z - a_{ii}| \leq r_i\}, \qquad r_i := \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|,$$

bezeichnen wir als die Gerschgorin-Kreise *zu der Matrix $\mathbf{A}$. Es gilt*

$$\sigma(\mathbf{A}) \subseteq \bigcup_{i=1}^{n} \mathcal{D}_i,$$

jeder Eigenwert ist also in mindestens einer der Kreisscheiben enthalten.

Beweis. Sei $\lambda \in \sigma(\mathbf{A})$ ein Eigenwert der Matrix $\mathbf{A}$. Sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein korrespondierender Eigenwert. Wir bezeichnen den Index des betragsgrößten Koeffizienten mit $i \in [1:n]$, es gilt also

$$|x_j| \leq |x_i| \qquad \text{für alle } j \in [1:n].$$

Da $\mathbf{x}$ ein Eigenvektor ist, folgt mit der Dreiecksungleichung

$$\begin{aligned}
\lambda x_i = (\lambda \mathbf{x})_i = (\mathbf{A}\mathbf{x})_i &= \sum_{j=1}^{n} a_{ij} x_j, \\
(\lambda - a_{ii}) x_i &= \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j, \\
|\lambda - a_{ii}|\,|x_i| &\leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\,|x_j| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\,|x_i| = r_i |x_i|, \\
|\lambda - a_{ii}| &\leq r_i,
\end{aligned}$$

also bereits $\lambda \in \mathcal{D}_i$. ∎

Im von uns untersuchten Fall selbstadjungierter Tridiagonalmatrizen $\mathbf{A}$ sind die *Gerschgorin-Kreise* besonders einfach zu bestimmen: Wenn wir $\beta_0 = \beta_n = 0$ setzen, erhalten wir

$$\mathcal{D}_i = \{z \in \mathbb{C} \ : \ |z - \alpha_i| \leq |\beta_{i-1}| + |\beta_i|\}.$$

Da $\mathbf{A}$ selbstadjungiert ist, ist das Spektrum reell, es gilt also

$$\sigma(\mathbf{A}) \subseteq \bigcup_{i=1}^{n} [\alpha_i - (|\beta_{i-1}| + |\beta_i|), \alpha_i + (|\beta_{i-1}| + |\beta_i|)],$$

so dass wir folgern können, dass das Spektrum in dem Intervall $[a, b]$ mit

$$\begin{aligned}
a &:= \min\{\alpha_i - |\beta_{i-1}| - |\beta_i| \ : \ i \in \{1, \ldots, n\}\}, \\
b &:= \max\{\alpha_i + |\beta_{i-1}| + |\beta_i| \ : \ i \in \{1, \ldots, n\}\}
\end{aligned}$$

enthalten ist Dieses Intervall erfüllt also die Voraussetzungen von Algorithmus 7.9 für jedes $k \in [1:n]$, so dass wir gezielt jeden beliebigen Eigenwert berechnen können.

7.3 Trägheitssatz und Dreieckszerlegungen

Der Einsatz der Sturmschen Ketten kann zu Schwierigkeiten führen, falls Rundungsfehler die Vorzeichen der einzelnen Polynome verfälschen. Es gibt allerdings alternative Möglichkeiten, um festzustellen, wie viele Eigenwerte kleiner oder größer als eine gegebene Zahl sind: Wir untersuchen, wieviele negative und positive Eigenwerte die „spektral

verschobene“ Matrix $\mathbf{A} - \mu\mathbf{I}$ besitzt. Die Anzahl der Eigenwerte, die echt kleiner als μ sind, ist gerade die Anzahl der negativen Eigenwerte der Matrix $\mathbf{A} - \mu\mathbf{I}$.

Die Anzahl der negativen Eigenwerte lässt sich bestimmen, ohne sie explizit zu berechnen, indem wir geeignete Transformationen auf die Matrix anwenden, die wesentlich einfacher handzuhaben sind als die bisher verwendeten Ähnlichkeitstransformationen.

Definition 7.11 (Kongruenztransformation) *Sei $\mathbf{T} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix. Die Abbildung*

$$\mathbb{K}^{n\times n} \to \mathbb{K}^{n\times n}, \qquad \mathbf{A} \mapsto \mathbf{T}^*\mathbf{A}\mathbf{T},$$

nennen wir die zu $\mathbf{T}$ gehörende Kongruenztransformation.

Für unitäre Matrizen $\mathbf{T}$ ist die Kongruenztransformation auch eine Ähnlichkeitstransformation, im allgemeinen Fall gilt das allerdings nicht. Trotzdem können Kongruenztransformationen sehr nützlich sein: Der *Trägheitssatz von Sylvester* besagt, dass die Anzahl der positiven und negativen Eigenwerte unter Kongruenztransformationen unverändert bleibt.

Satz 7.12 (Trägheitssatz) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix. Sei $\mathbf{T} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix und*

$$\widehat{\mathbf{A}} := \mathbf{T}^*\mathbf{A}\mathbf{T}.$$

Dann besitzen $\mathbf{A}$ und $\widehat{\mathbf{A}}$ jeweils gleich viele echt positive und echt negative Eigenwerte.

Beweis. Nach Satz 3.43 existieren unitäre Matrizen $\mathbf{Q}, \widehat{\mathbf{Q}} \in \mathbb{K}^{n\times n}$ und reelle Diagonalmatrizen $\mathbf{D}, \widehat{\mathbf{D}} \in \mathbb{R}^{n\times n}$ mit

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*, \qquad \mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n,$$

$$\widehat{\mathbf{A}} = \widehat{\mathbf{Q}}\widehat{\mathbf{D}}\widehat{\mathbf{Q}}^*, \qquad \widehat{\mathbf{D}} = \begin{pmatrix} \hat{\lambda}_1 & & \\ & \ddots & \\ & & \hat{\lambda}_n \end{pmatrix}, \qquad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_n.$$

Wir bezeichnen mit $p, \hat{p} \in [0:n]$ die Anzahl der echt positiven Eigenwerte der Matrizen $\mathbf{A}$ und $\widehat{\mathbf{A}}$, es sollen also

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0 \geq \lambda_{p+1}, \qquad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{\hat{p}} > 0 \geq \hat{\lambda}_{\hat{p}+1}$$

gelten. Unsere Aufgabe ist es, nachzuweisen, dass $p = \hat{p}$ gilt.

Dazu definieren wir die Spaltenvektoren $\mathbf{q}^{(j)}, \widehat{\mathbf{q}}^{(j)} \in \mathbb{K}^n$ der Matrizen $\mathbf{Q}$ und $\widehat{\mathbf{Q}}$ durch

$$q_i^{(j)} := q_{ij}, \qquad \hat{q}_i^{(j)} := \hat{q}_{ij} \qquad \text{für alle } i, j \in [1:n].$$

Die von den Eigenvektoren zu echt positiven Eigenwerten aufgespannten Vektorräume bezeichnen wir mit

$$\mathcal{P} := \operatorname{span}\{\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(p)}\}, \qquad \widehat{\mathcal{P}} := \operatorname{span}\{\widehat{\mathbf{q}}^{(1)}, \dots, \widehat{\mathbf{q}}^{(\hat{p})}\}.$$

Wenn wir zeigen können, dass beide Räume dieselbe Dimension besitzen, sind wir fertig.

Als Hilfsmittel führen wir die Räume

$$\mathcal{N} := \operatorname{span}\{\mathbf{q}^{(p+1)}, \dots, \mathbf{q}^{(n)}\}, \qquad \widehat{\mathcal{N}} := \operatorname{span}\{\widehat{\mathbf{q}}^{(\hat{p}+1)}, \dots, \widehat{\mathbf{q}}^{(n)}\}$$

zu den nicht-positiven Eigenwerten ein und untersuchen die Abbildungen

$$\begin{aligned} f &: \mathbb{K}^n \to \mathbb{R}, & \mathbf{x} &\mapsto \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle, \\ \hat{f} &: \mathbb{K}^n \to \mathbb{R}, & \widehat{\mathbf{x}} &\mapsto \langle \widehat{\mathbf{x}}, \widehat{\mathbf{A}}\widehat{\mathbf{x}}\rangle. \end{aligned}$$

Für einen Vektor $\mathbf{x} \in \mathcal{P} \setminus \{\mathbf{0}\}$ finden wir Koeffizienten $\alpha_1, \dots, \alpha_j \in \mathbb{K}$ mit

$$\mathbf{x} = \sum_{j=1}^{p} \alpha_j \mathbf{q}^{(j)},$$

die wegen $\mathbf{x} \neq \mathbf{0}$ nicht alle gleich null sein können, und erhalten

$$f(\mathbf{x}) = \sum_{i=1}^{p}\sum_{j=1}^{p} \bar{\alpha}_i \alpha_j \langle \mathbf{q}^{(i)}, \mathbf{A}\mathbf{q}^{(j)}\rangle = \sum_{i=1}^{p} |\alpha_i|^2 \langle \mathbf{q}^{(i)}, \lambda_i \mathbf{q}^{(i)}\rangle = \sum_{i=1}^{p} |\alpha_i|^2 \lambda_i > 0.$$

Für $\hat{f}$ und $\widehat{\mathcal{N}}$ erhalten wir ein entsprechendes Resultat und dürfen

$$f(\mathbf{x}) > 0, \qquad \hat{f}(\widehat{\mathbf{x}}) \leq 0 \qquad \text{für alle } \mathbf{x} \in \mathcal{P} \setminus \{\mathbf{0}\},\ \widehat{\mathbf{x}} \in \widehat{\mathcal{N}}$$

festhalten. Sei nun $\widehat{\mathbf{x}} \in \widehat{\mathcal{N}}$ gegeben. Wir stellen fest, dass

$$0 \geq \hat{f}(\widehat{\mathbf{x}}) = \langle \widehat{\mathbf{x}}, \widehat{\mathbf{A}}\widehat{\mathbf{x}}\rangle = \langle \widehat{\mathbf{x}}, \mathbf{T}^*\mathbf{A}\mathbf{T}\widehat{\mathbf{x}}\rangle = \langle \mathbf{T}\widehat{\mathbf{x}}, \mathbf{A}\mathbf{T}\widehat{\mathbf{x}}\rangle = f(\mathbf{T}\widehat{\mathbf{x}})$$

gilt. Also folgt

$$f(\mathbf{x}) \leq 0 \qquad \text{für alle } \mathbf{x} \in \mathbf{T}\widehat{\mathcal{N}},$$

also kann der Schnitt der Teilräume $\mathcal{P}$ und $\mathbf{T}\widehat{\mathcal{N}}$ nur den Nullvektor enthalten.

Da $\mathbf{T}$ invertierbar ist, folgt daraus

$$n \geq \dim(\mathcal{P}) + \dim(\mathbf{T}\widehat{\mathcal{N}}) = \dim(\mathcal{P}) + \dim(\widehat{\mathcal{N}}) = p + n - \hat{p},$$

also $0 \geq p - \hat{p}$ und damit $\hat{p} \geq p$.

Indem wir entsprechend mit den Räumen $\widehat{\mathcal{P}}$ und $\mathcal{N}$ verfahren, folgt auch $p \geq \hat{p}$, so dass $p = \hat{p}$ bewiesen ist.

Wir können dieselbe Argumentation auf die Matrizen $-\mathbf{A}$ und $-\widehat{\mathbf{A}}$ anwenden, um zu zeigen, dass auch die Anzahl der echt negativen Eigenwerte identisch ist. ∎

Unser Ziel ist es, die Anzahl der negativen Eigenwerte der Matrix $\mathbf{A}-\mu\mathbf{I}$ zu berechnen. Dank des Trägheitssatzes 7.12 ändert sich diese Anzahl nicht, wenn wir Kongruenztransformationen auf die Matrix anwenden, also bietet es sich an, nach Kongruenztransformationen zu suchen, die die Matrix in eine Form bringen, an der wir die Anzahl der negativen Eigenwerte unmittelbar ablesen können. Ideal geeignet wäre eine Diagonalmatrix, wir suchen also eine reguläre Matrix $\mathbf{T} \in \mathbb{K}^{n\times n}$ derart, dass $\mathbf{T}^*(\mathbf{A}-\mu\mathbf{I})\mathbf{T}$ eine Diagonalmatrix ist, denn dann stehen die Eigenwerte auf der Diagonalen.

Diese Aufgabe lässt sich lösen, indem wir $\mathbf{T}$ als Dreiecksmatrix wählen und eine verallgemeinerte Form der *Cholesky-Zerlegung* berechnen. Zur Abkürzung setzen wir $\mathbf{B} := \mathbf{A}-\mu\mathbf{I}$ und suchen nach einer normierten unteren Dreiecksmatrix $\mathbf{L} \in \mathbb{K}^{n\times n}$ und einer reellen Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ mit

$$\mathbf{B} = \mathbf{L}\mathbf{D}\mathbf{L}^* \qquad \Longleftrightarrow \qquad \mathbf{L}^{-1}(\mathbf{A}-\mu\mathbf{I})\mathbf{L}^{-*} = \mathbf{D}.$$

Wir zerlegen die auftretenden Matrizen in Teilmatrizen:

$$\begin{aligned}
\mathbf{B} &= \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ \mathbf{B}_{*1} & \mathbf{B}_{**} \end{pmatrix} && \text{mit } \mathbf{B}_{1*} \in \mathbb{K}^{1\times(n-1)},\ \mathbf{B}_{**} \in \mathbb{K}^{(n-1)\times(n-1)},\\
\mathbf{L} &= \begin{pmatrix} 1 & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} && \text{mit } \mathbf{L}_{*1} \in \mathbb{K}^{1\times(n-1)},\ \mathbf{L}_{**} \in \mathbb{K}^{(n-1)\times(n-1)},\\
\mathbf{D} &= \begin{pmatrix} d_1 & \\ & \mathbf{D}_{**} \end{pmatrix} && \text{mit}\mathbf{D}_{**} \in \mathbb{R}^{(n-1)\times(n-1)}.
\end{aligned}$$

Durch Einsetzen in die definierende Gleichung erhalten wir

$$\begin{aligned}
\begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ \mathbf{B}_{*1} & \mathbf{B}_{**} \end{pmatrix} = \mathbf{B} = \mathbf{L}\mathbf{D}\mathbf{L}^* &= \begin{pmatrix} 1 & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} \begin{pmatrix} d_1 & \\ & \mathbf{D}_{**} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{L}_{*1}^* \\ & \mathbf{L}_{**}^* \end{pmatrix} \\
&= \begin{pmatrix} d_1 & \\ \mathbf{L}_{*1}d_1 & \mathbf{L}_{**}\mathbf{D}_{**} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{L}_{*1}^* \\ & \mathbf{L}_{**}^* \end{pmatrix} = \begin{pmatrix} d_1 & d_1\mathbf{L}_{*1}^* \\ \mathbf{L}_{*1}d_1 & \mathbf{L}_{*1}d_1\mathbf{L}_{*1}^* + \mathbf{L}_{**}\mathbf{D}_{**}\mathbf{L}_{**}^* \end{pmatrix}.
\end{aligned}$$

Aus $\mathbf{B}^* = \mathbf{B}$ folgt $\mathbf{B}_{1*} = \mathbf{B}_{*1}^*$, so dass lediglich die folgenden drei Gleichungen zu erfüllen sind:

$$\begin{aligned}
b_{11} &= d_1,\\
\mathbf{B}_{*1} &= \mathbf{L}_{*1}d_1,\\
\mathbf{B}_{**} &= \mathbf{L}_{*1}d_1\mathbf{L}_{*1}^* + \mathbf{L}_{**}\mathbf{D}_{**}\mathbf{L}_{**}^*.
\end{aligned}$$

Falls $b_{11} \neq 0$ gilt, können wir sie umformen und finden

$$\begin{aligned}
d_1 &= b_{11},\\
\mathbf{L}_{*1} &= \mathbf{B}_{*1}/d_1,\\
\mathbf{L}_{**}\mathbf{D}_{**}\mathbf{L}_{**}^* &= \widehat{\mathbf{B}} := \mathbf{B}_{**} - \mathbf{L}_{*1}d_1\mathbf{L}_{*1}^*.
\end{aligned}$$

Die letzte Gleichung ist von der Form des ursprünglichen Problems, so dass wir sie per Induktion behandeln können.

In unserem Fall ist $\mathbf{B}$ eine Tridiagonalmatrix der Form

$$\mathbf{B} = \begin{pmatrix} \alpha_1 - \mu & \bar{\beta}_1 & & \\ \beta_1 & \alpha_2 - \mu & \ddots & \\ & \ddots & \ddots & \bar{\beta}_{n-1} \\ & & \beta_{n-1} & \alpha_n - \mu \end{pmatrix},$$

so dass die Gleichungen

$$\mathbf{L}_{*1} = \begin{pmatrix} \beta_1/(\alpha_1 - \mu) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \widehat{\mathbf{B}} = \begin{pmatrix} \alpha_2 - \mu - \frac{|\beta_1|^2}{(\alpha_1-\mu)^2} & \bar{\beta}_2 & & \\ \beta_2 & \alpha_3 - \mu & \ddots & \\ & \ddots & \ddots & \bar{\beta}_{n-1} \\ & & \beta_{n-1} & \alpha_n - \mu \end{pmatrix}$$

gelten und sich demnach der Induktionsschritt mit wenigen Rechenoperationen vollziehen lässt.

8 Lanczos-Verfahren für schwachbesetzte Matrizen

In der Praxis treten sehr häufig Matrizen auf, die besondere Eigenschaften aufweisen, die zur Beschleunigung bestimmter Operationen ausgenutzt werden können: Die Matrix-Vektor-Multiplikation mit einer Tridiagonalmatrix der Dimension n oder Bestimmung einer QR-Zerlegung lassen sich mit einem Aufwand von $\mathcal{O}(n)$ Operationen durchführen. Ein Gleichungssystem mit einer Hessenberg-Matrix lässt sich mit einem Aufwand von $\mathcal{O}(n^2)$ Operationen lösen, während bei einem allgemeinen System $\mathcal{O}(n^3)$ Operationen erforderlich sind.

Wir werden uns in diesem Kapitel auf eine relativ allgemeine Klasse von Matrizen konzentrieren, nämlich auf solche, die nur eine von n unabhängige Anzahl von nicht verschwindenden Elementen pro Zeile und Spalte aufweisen. Derartige Matrizen treten typischerweise bei der Diskretisierung partieller Differentialgleichungen auf.

8.1 Zweidimensionales Modellproblem

Wir untersuchen als Modellproblem die numerische Approximation des Poisson-Eigenwertproblems

$$-\Delta u(x,y) = \lambda u(x,y)$$

mit $(x,y) \in]0,1[^2$, der Randbedingung

$$u(0,\cdot) = u(1,\cdot) = u(\cdot,0) = u(\cdot,1) = 0$$

und dem Laplace-Operator

$$\Delta u(x,y) := \frac{\partial^2 u}{\partial x}(x,y) + \frac{\partial^2 u}{\partial y}(x,y). \tag{8.1}$$

Zur Diskretisierung wählen wir ein $N \in \mathbb{N}$ und setzen

$$h := \frac{1}{N+1}, \qquad \mathcal{I} := [1:N]^2, \qquad \Omega_h := \{(hi,hj) \ : \ (i,j) \in \mathcal{I}\}.$$

Analog zum Beispiel in Abschnitt 2.1 approximieren wir die Differentialquotienten in (8.1) durch Differenzenquotienten:

$$-\Delta u(x,y) \approx h^{-2}\left(4u(x,y) - u(x-h,y) - u(x+h,y) - u(x,y-h) - u(x,y+h)\right).$$

Wie im eindimensionalen Fall schränken wir den Definitionsbereich auf Ω_h ein und erhalten ein lineares Gleichungssystem im Raum $\mathbb{R}^{\mathcal{I}}$: Wir ersetzen den Differentialoperator

$-\Delta$ durch eine Matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ und die Funktion u durch einen Vektor $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$, die durch

$$a_{ij} := \begin{cases} 4h^{-2} & \text{falls } i_x = j_x \text{ und } i_y = j_y \\ -h^{-2} & \text{falls } |i_x - i_y| + |j_x - j_y| = 1 \\ 0 & \text{ansonsten} \end{cases},$$
$$x_j := u(hj_x, hj_y) \qquad \text{für alle } i = (i_x, i_y) \in \mathcal{I}, j = (j_x, j_y) \in \mathcal{I}$$

definiert sind und erhalten das Gleichungssystem

$$\mathbf{A}\mathbf{x} \approx \lambda\mathbf{x}.$$

Wie im eindimensionalen Fall ist die Matrix $\mathbf{A}$ symmetrisch und positiv definit, und auch in diesem Fall sind wir an ihren kleinsten Eigenwerten interessiert. Unsere Aufgabe besteht also darin, ein n-dimensionales Eigenwertproblem zu lösen, wobei $n = N^2 \approx h^{-2}$ gilt und sich beweisen lässt, dass die Eigenwerte und Eigenvektoren des diskreten Systems mit einem Fehler proportional zu h^2 gegen die des ursprünglichen kontinuierlichen Problems konvergieren. Um eine brauchbare Genauigkeit zu erreichen, müssen wir also darauf vorbereitet sein, sehr große Eigenwertprobleme zu behandeln.

Während allerdings im eindimensionalen Fall eine Tridiagonalmatrix entstand, lässt sich im zweidimensionalen Fall keine Anordnung der Indexmenge $\mathcal{I}$ finden, die aus $\mathbf{A}$ eine Bandmatrix mit von N unabhängiger Bandbreite macht: Man kann beweisen, dass die Bandbreite immer mindestens $N/2$ betragen muss. In der Praxis verwendet man sehr häufig eine Anordnung von $\mathcal{I}$, bei der die Matrix $\mathbf{A}$ die Block-Tridiagonaldarstellung

$$\mathbf{A} = h^{-2}\begin{pmatrix} \mathbf{T} & -\mathbf{I} & & \\ -\mathbf{I} & \ddots & \ddots & \\ & \ddots & \ddots & -\mathbf{I} \\ & & -\mathbf{I} & \mathbf{T} \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad \mathbf{T} = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N\times N},$$

besitzt, die eine Bandbreite von N aufweist.

Eine Besonderheit von Tridiagonalmatrizen weist allerdings auch die Matrix $\mathbf{A}$ unabhängig von der Anordnung der Indizes auf: Da pro Zeile dieser Matrizen lediglich höchsten fünf von null verschiedene Einträge auftreten, lässt sich das Matrix-Vektor-Produkt $\mathbf{y} := \mathbf{A}\mathbf{x}$ für $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ in $\mathcal{O}(n)$ Operationen auswerten. Matrizen mit dieser Eigenschaft werden als *schwachbesetzt* bezeichnet.

Theoretisch können wir alle bereits diskutierten Verfahren auf die Matrix $\mathbf{A}$ anwenden, praktisch stoßen wir dabei allerdings auf Schwierigkeiten: Sobald wir Linearkombinationen von Zeilen oder Spalten berechnen, werden dadurch in der Regel neue von null verschiedene Einträge in der Matrix verursacht. Das bedeutet, dass der Speicherbedarf und der Rechenaufwand sehr stark zunehmen, für hohe Problemdimensionen ist dieser Ansatz nicht mehr sinnvoll durchführbar.

Günstig dagegen wäre die Vektoriteration, da sie lediglich Matrix-Vektor-Multiplikationen erfordert, die sich, wie gesagt, sehr effizient durchführen lassen. Dieser Ansatz lässt sich ausbauen, um mehrere der größten und kleinsten Eigenwerte mit den entsprechenden Eigenvektoren zu berechnen.

8.2 Krylow-Räume

Bei allgemeinen Matrizen können wir den Rechenaufwand wesentlich reduzieren, indem wir sie mit Hilfe unitärer Transformationen auf Hessenberg-Gestalt bringen. Falls es uns gelingt, dasselbe für schwachbesetzte Matrizen zu bewerkstelligen, könnten wir im symmetrischen Fall einfach Standardverfahren wie die QR-Iteration oder die Bisektion mit einer Sturmschen Kette auf die resultierende Tridiagonalmatrix anwenden und so ein effizienteres Verfahren erhalten.

Bei einer vollständigen Tridiagonalisierung würde der Aufwand trotzdem mindestens $\mathcal{O}(n^2)$ betragen, weil die unitären Matrizen $\mathcal{O}(n^2)$ Einträge aufweisen. Deshalb bestimmen wir ein Tupel von orthonormalen Matrizen $(\mathbf{Q}^{(k)})_{k=1}^n$ mit $\mathbf{Q}^{(k)} \in \mathbb{R}^{n\times k}$ derart, dass die für unsere Anwendungen interessanten größten und kleinsten Eigenwerte der transformierten Matrizen $(\mathbf{Q}^{(k)})^*\mathbf{A}\mathbf{Q}^{(k)}$ möglichst schnell gegen die größten und kleinsten Eigenwerte von $\mathbf{A}$ konvergieren.

Sei im folgenden $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix mit den Eigenwerten $\lambda_1 \leq \ldots \leq \lambda_n$. Dank des Satzes 3.42 von Courant und Fischer wissen wir, dass der Rayleigh-Quotient

$$\Lambda_A \colon \mathbb{K}^n \setminus \{\mathbf{0}\} \to \mathbb{R}, \qquad \mathbf{x} \mapsto \frac{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2},$$

die Gleichungen

$$\begin{aligned}\lambda_1 &= \min\{\Lambda_A(\mathbf{x}) \ : \ \mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\},\\ \lambda_n &= \max\{\Lambda_A(\mathbf{x}) \ : \ \mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\}\end{aligned}$$

erfüllt. In ähnlicher Weise lassen sich auch andere Eigenwerte mit Hilfe geeigneter lokaler Minima und Maxima des Rayleigh-Quotienten charakterisieren.

Eine gute Strategie zur Approximation der Eigenwerte könnte also darin bestehen, den Rayleigh-Quotienten zu minimieren oder zu maximieren. Für derartige Aufgaben ist das *Gradientenverfahren* geeignet: Der *Gradient* einer Funktion $f \colon \mathbb{R}^n \to \mathbb{R}$ in einem Punkt $\mathbf{x} \in \mathbb{R}^n$ ist der Vektor $\nabla f(\mathbf{x}) \in \mathbb{R}^n$, für den

$$Df(\mathbf{x}) \cdot \mathbf{y} = \langle \nabla f(\mathbf{x}), \mathbf{y}\rangle_2 \qquad \text{für alle } \mathbf{y} \in \mathbb{R}^n$$

gilt, mit dem sich also alle Richtungsableitungen von f durch ein Skalarprodukt beschreiben lassen. Es lässt sich nachweisen, dass die Steigung der Funktion f im Punkt $\mathbf{x}$ in Richtung des Gradienten am stärksten ist.

Die Idee des Gradientenverfahrens besteht nun darin, eine Folge von Iterierten so zu bestimmen, dass sich die neue Iterierte $\mathbf{x}'$ ergibt, indem zu der alten Iterierten $\mathbf{x}$ ein geeignetes Vielfaches des Gradienten $\nabla f(\mathbf{x})$ hinzuaddiert wird, denn diese Richtung ist zumindest lokal am vielversprechendsten.

Wir werden nun in ähnlicher Weise versuchen, den Rayleigh-Quotienten zu maximieren beziehungsweise zu minimieren. Dazu benötigen wir seinen Gradienten:

Lemma 8.1 (Gradient) *Es gilt*

$$\nabla\Lambda_A(\mathbf{x}) = \frac{2}{\langle \mathbf{x}, \mathbf{x}\rangle_2}(\mathbf{A}\mathbf{x} - \Lambda_A(\mathbf{x})\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

Insbesondere ist der Gradient $\nabla\Lambda_A(\mathbf{x})$ genau dann gleich null, wenn $\mathbf{x}$ ein Eigenvektor von $\mathbf{A}$ ist.

Beweis. Wir führen die Hilfsfunktionen

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2, \qquad g(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}\rangle_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n$$

ein und erhalten mit Hilfe der Produktregel die Gleichungen

$$\begin{aligned} Df(\mathbf{x}) \cdot \mathbf{y} &= \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle_2 + \langle \mathbf{x}, \mathbf{A}\mathbf{y}\rangle_2 = \langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle_2 + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle_2 = 2\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle_2, \\ Dg(\mathbf{x}) \cdot \mathbf{y} &= 2\langle \mathbf{x}, \mathbf{y}\rangle_2 \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n. \end{aligned}$$

Aus $\Lambda_A(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$ und der Quotientenregel folgt

$$\begin{aligned} D\Lambda_A(\mathbf{x}) \cdot \mathbf{y} &= \frac{g(\mathbf{x})Df(\mathbf{x}) \cdot \mathbf{y} - f(\mathbf{x})Dg(\mathbf{x}) \cdot \mathbf{y}}{g^2(\mathbf{x})} = \frac{2\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2} - \frac{f(\mathbf{x})}{g(\mathbf{x})}\frac{2\langle \mathbf{x}, \mathbf{y}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2} \\ &= \frac{2}{\langle \mathbf{x}, \mathbf{x}\rangle_2}\langle \mathbf{A}\mathbf{x} - \Lambda_A(\mathbf{x})\mathbf{x}, \mathbf{y}\rangle_2 \qquad \text{für alle } \mathbf{y} \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \end{aligned}$$

und damit die gesuchte Gleichung. ∎

Wir beginnen mit der Suche nach einem Minimum mit einem Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^n$. Um die nächste Iterierte zu berechnen, bestimmen wir den Gradienten des Rayleigh-Quotienten Λ_A in $\mathbf{x}^{(0)}$ und setzen

$$\begin{aligned} \mathbf{x}^{(m+1)} &:= \mathbf{x}^{(m)} + \alpha_m \nabla\Lambda_A(\mathbf{x}^{(m)}) = \mathbf{x}^{(m)} - \frac{2\alpha_m\Lambda_A(\mathbf{x}^{(m)})}{\langle \mathbf{x}^{(m)}, \mathbf{x}^{(m)}\rangle_2}\mathbf{x}^{(m)} + \frac{2\alpha_m}{\langle \mathbf{x}^{(m)}, \mathbf{x}^{(m)}\rangle_2}\mathbf{A}\mathbf{x}^{(m)} \\ &\in \operatorname{span}\{\mathbf{x}^{(m)}, \mathbf{A}\mathbf{x}^{(m)}\} \qquad \text{für alle } m \in \mathbb{N}_0 \end{aligned}$$

mit geeigneten Skalierungsfaktoren α_m. Die nächste Iterierte liegt also im Aufspann von $\mathbf{x}^{(m)}$ und $\mathbf{A}\mathbf{x}^{(m)}$. Mit einer einfachen Induktion folgt

$$\mathbf{x}^{(m)} \in \operatorname{span}\{\mathbf{x}^{(0)}, \mathbf{A}\mathbf{x}^{(0)}, \ldots, \mathbf{A}^m\mathbf{x}^{(0)}\} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Die derart durch das wiederholte Multiplizieren eines Startvektors mit der Matrix $\mathbf{A}$ definierten Räume sind für uns von besonderem Interesse.

Definition 8.2 (Krylow-Raum) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$, $\mathbf{q}^{(0)} \in \mathbb{K}^n$ und $m \in \mathbb{N}_0$. Der Raum*

$$\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) := \operatorname{span}\{\mathbf{A}^i\mathbf{q}^{(0)} \ : \ i \in [0:m]\}$$

wird als m-ter Krylow-Raum *zu der Matrix $\mathbf{A}$ und dem Startvektor $\mathbf{q}^{(0)}$ bezeichnet.*

Offenbar gilt immer $\dim \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) \leq m+1$. Die Dimension kann allerdings deutlich kleiner sein: Falls beispielsweise $\mathbf{q}^{(0)}$ ein Eigenvektor von $\mathbf{A}$ ist, ist $\mathbf{A}\mathbf{q}^{(0)}$ ein Vielfaches von $\mathbf{q}^{(0)}$, also gilt $\dim \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) = 1$ für alle $m \in \mathbb{N}_0$.

Diese Beobachtung lässt sich etwas verallgemeinern: Falls $\mathbf{q}^{(0)}$ in einem ℓ-dimensionalen invarianten Unterraum der Matrix $\mathbf{A}$ enthalten ist, muss offenbar auch jeder Krylow-Raum $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m)$ in diesem Unterraum enthalten sein, also erfüllt die Dimension die Abschätzung $\dim \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) \leq \ell$ für alle $m \in \mathbb{N}_0$.

8.3 Arnoldi-Basis

Der Vorteil der Krylow-Räume besteht darin, dass sie sich relativ einfach mit Hilfe von Matrix-Vektor-Multiplikationen konstruieren lassen. Außerdem eignen sie sich, wie bereits gesehen, um den Rayleigh-Quotienten zu minimieren beziehungsweise zu maximieren und so Approximationen des kleinsten beziehungsweise größten Eigenwertes zu erhalten.

Allerdings ist dafür die kanonische Basis $\mathbf{q}^{(0)}, \mathbf{A}\mathbf{q}^{(0)}, \ldots, \mathbf{A}^m\mathbf{q}^{(0)}$ in der Regel nicht gut geeignet: Für große Werte von m ist zu befürchten, dass die Basisvektoren, wie schon bei der Vektoriteration gesehen, gegen einen Eigenraum konvergieren und damit „numerisch linear abhängig“ werden können. Die Lösung besteht, wie schon bei der orthogonalen Iteration, darin, zu einer orthonormalen Basis zu wechseln.

Da sich die kanonische Basis des Krylow-Raums $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m+1)$ von der des Raums $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m)$ nur durch den Vektor $\mathbf{A}^{m+1}\mathbf{q}^{(0)}$ unterscheidet, bietet es sich an, auch orthonormale Basen zu verwenden, die sich lediglich durch einen Vektor unterscheiden. Wir suchen also eine orthonormale Vektoren $\mathbf{q}^{(0)}, \mathbf{q}^{(1)}, \ldots$ derart, dass

$$\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) = \operatorname{span}\{\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}\} \qquad \text{für alle } m \in [0 : m_0] \tag{8.2}$$

gilt, wobei

$$m_0 := \min\{m \in \mathbb{N} \ : \ \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) \subseteq \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m-1)\}$$

die maximale Dimension ist, die ein mit dem Startvektor $\mathbf{q}^{(0)}$ konstruierter Krylow-Raum erreichen kann. Wie wir bereits gesehen haben, entspricht m_0 gerade der Dimension des kleinsten invarianten Unterraums, der $\mathbf{q}^{(0)}$ enthält.

Aus der Gleichung (8.2) können wir direkt eine induktive Konstruktion für die Vektoren $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m_0-1)}$ gewinnen: Wir gehen davon aus, dass ein $m \in [0 : m_0 - 1]$ so gegeben ist, dass $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}$ eine Basis des Krylow-Raums $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m)$ ist. Damit ist

$$\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}, \mathbf{A}^{m+1}\mathbf{q}^{(0)}$$

eine Basis des Raums $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m+1)$. Da wir damit rechnen müssen, dass die Vektoren $\mathbf{A}^{m+1}\mathbf{q}^{(0)}$ gegen einen Eigenraum konvergieren und damit die Basis „numerisch linear abhängig“ wird, soll nun $\mathbf{A}^{m+1}\mathbf{q}^{(0)}$ durch einen „numerisch stabileren“ Vektor ersetzt werden: Wegen

$$\mathbf{A}^m\mathbf{q}^{(0)} \in \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) = \operatorname{span}\{\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}\}$$

existieren $\alpha_0, \ldots, \alpha_m \in \mathbb{K}$ mit

$$\mathbf{A}^m \mathbf{q}^{(0)} = \alpha_0 \mathbf{q}^{(0)} + \ldots + \alpha_m \mathbf{q}^{(m)},$$

also folgt

$$\mathbf{A}^{m+1} \mathbf{q}^{(0)} = \alpha_0 \mathbf{A}\mathbf{q}^{(0)} + \ldots + \alpha_m \mathbf{A}\mathbf{q}^{(m)}.$$

Nach Definition haben wir

$$\alpha_0 \mathbf{A}\mathbf{q}^{(0)} + \ldots + \alpha_{m-1} \mathbf{A}\mathbf{q}^{(m-1)} \in \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m),$$

also ist auch $\mathbf{A}\mathbf{q}^{(m)}$ ein Vektor, der in $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m+1)$, aber nicht in $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m)$ liegt. Dieser Vektor lässt sich numerisch stabiler konstruieren, und

$$\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}, \mathbf{A}\mathbf{q}^{(m)}$$

ist eine Basis des Raums $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m+1)$. Per Gram-Schmidt-Orthonormalisierung konstruieren wir nun aus $\mathbf{A}\mathbf{q}^{(m)}$ den gewünschten Vektor $\mathbf{q}^{(m+1)}$.

Definition 8.3 (Arnoldi-Basis) *Sei $\mathbf{q}^{(0)} \in \mathbb{K}^n$ mit $\|\mathbf{q}^{(0)}\|_2 = 1$ gegeben. Die* Arnoldi-Basis $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m_0-1)}$ *ist durch*

$$\mathbf{p}^{(m+1)} := \mathbf{A}\mathbf{q}^{(m)} - \sum_{i=0}^{m} \langle \mathbf{q}^{(i)}, \mathbf{A}\mathbf{q}^{(m)} \rangle_2 \, \mathbf{q}^{(i)} \qquad \text{für alle } m \in [0 : m_0 - 1],$$

$$\mathbf{q}^{(m+1)} := \frac{\mathbf{p}^{(m+1)}}{\|\mathbf{p}^{(m+1)}\|_2} \qquad \text{für alle } m \in [0 : m_0 - 2]$$

definiert. Nach Definition von m_0 gelten $\mathbf{p}^{(m)} \neq \mathbf{0}$ für alle $m \in [1 : m_0 - 1]$ sowie $\mathbf{p}^{(m_0)} = \mathbf{0}$, also ist die Arnoldi-Basis wohldefiniert.

Nach Konstruktion ist die Basis auch orthonormal und besitzt die Eigenschaft (8.2).

Wir sind daran interessiert, die kleinsten und größten Eigenwerte zu approximieren. Der kleinste Eigenwert λ_1 ist gerade das Minimum des Rayleigh-Quotienten Λ_A auf dem Raum $\mathbb{K}^n \setminus \{\mathbf{0}\}$, und indem wir diesen Raum durch die Teilräume

$$\mathcal{Q}^{(m)} := \operatorname{span}\{\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}\} \qquad \text{für alle } m \in [0 : m_0 - 1]$$

ersetzen, können wir Approximationen

$$\lambda_1^{(m)} := \min\{\Lambda_A(\mathbf{x}) \;:\; \mathbf{x} \in \mathcal{Q}^{(m)} \setminus \{\mathbf{0}\}\} \qquad \text{für alle } m \in [0 : m_0 - 1]$$

konstruieren. Das Minimum können wir berechnen, indem wir die Basisvektoren zu Matrizen

$$\mathbf{Q}^{(m)} := \begin{pmatrix} \mathbf{q}^{(0)} & \ldots & \mathbf{q}^{(m)} \end{pmatrix} \in \mathbb{K}^{n \times [0:m]} \qquad \text{für alle } m \in [0 : m_0 - 1]$$

zusammenfassen (die Notation $\mathbb{K}^{n\times[0:m]}$ soll betonen, dass die Spalten der Matrix von 0 bis m statt von 1 bis $m+1$ numeriert sind) und feststellen, dass

$$\begin{aligned}
\lambda_1^{(m)} &= \min\{\Lambda_A(\mathbf{x}) \ : \ \mathbf{x} \in \mathcal{Q}^{(m)} \setminus \{\mathbf{0}\}\} \\
&= \min\{\Lambda_A(\mathbf{Q}^{(m)}\widehat{\mathbf{x}}) \ : \ \widehat{\mathbf{x}} \in \mathbb{K}^{[0:m]} \setminus \{\mathbf{0}\}\} \\
&= \min\left\{ \frac{\langle \mathbf{Q}^{(m)}\widehat{\mathbf{x}}, \mathbf{A}\mathbf{Q}^{(m)}\widehat{\mathbf{x}}\rangle_2}{\langle \mathbf{Q}^{(m)}\widehat{\mathbf{x}}, \mathbf{Q}^{(m)}\widehat{\mathbf{x}}\rangle_2} \ : \ \widehat{\mathbf{x}} \in \mathbb{K}^{[0:m]} \setminus \{\mathbf{0}\} \right\} \\
&= \min\left\{ \frac{\langle \widehat{\mathbf{x}}, (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)}\widehat{\mathbf{x}}\rangle_2}{\langle (\mathbf{Q}^{(m)})^*\mathbf{Q}^{(m)}\widehat{\mathbf{x}}, \widehat{\mathbf{x}}\rangle_2} \ : \ \widehat{\mathbf{x}} \in \mathbb{K}^{[0:m]} \setminus \{\mathbf{0}\} \right\} \\
&= \min\{\Lambda_{(Q^{(m)})^*AQ^{(m)}}(\widehat{\mathbf{x}}) \ : \ \widehat{\mathbf{x}} \in \mathbb{K}^{[0:m]} \setminus \{\mathbf{0}\}\}
\end{aligned}$$

gilt, nach dem Satz 3.42 von Courant und Fischer sind also die Werte $\lambda_1^{(m)}$ gerade die kleinsten Eigenwerte der Matrizen

$$\widehat{\mathbf{A}}^{(m)} := (\mathbf{Q}^{(m)})^*\mathbf{A}\mathbf{Q}^{(m)} \in \mathbb{K}^{[0:m]\times[0:m]} \qquad \text{für alle } m \in [0 : m_0 - 1], \tag{8.3}$$

können also durch Lösen eines $(m+1)$-dimensionalen Eigenwertproblems berechnet werden. Für $m \ll n$ lässt sich diese Berechnung wesentlich effizienter durchführen als für die ursprüngliche Matrix $\mathbf{A}$. Durch die folgende Beobachtung wird die Berechnung sogar noch weiter vereinfacht:

Lemma 8.4 (Hessenberg-Form) *Für jedes $m \in [0 : m_0 - 1]$ besitzt die Matrix $\widehat{\mathbf{A}}^{(m)}$ Hessenberg-Gestalt.*

Beweis. Sei $m \in [0 : m_0 - 1]$ und $j \in [0 : m]$. Nach Konstruktion gilt

$$\mathbf{q}^{(j)} \in \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, j), \qquad \mathbf{A}\mathbf{q}^{(j)} \in \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, j+1) = \operatorname{span}\{\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(j+1)}\}.$$

Da $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m)}$ eine orthonormale Basis ist, muss also

$$\hat{a}_{ij}^{(m)} = \langle \mathbf{q}^{(i)}, \mathbf{A}\mathbf{q}^{(j)}\rangle_2 = 0 \qquad \text{für alle } i \in [j+2 : m]$$

gelten. Somit ist $\widehat{\mathbf{A}}^{(m)}$ eine obere Hessenberg-Matrix. ∎

Die Berechnung der unteren Nebendiagonalelemente gestaltet sich besonders einfach, weil nach Definition

$$\begin{aligned}
\hat{a}_{i+1,i}^{(m)} &= \langle \mathbf{q}^{(i+1)}, \mathbf{A}\mathbf{q}^{(i)}\rangle = \langle \mathbf{q}^{(i+1)}, \mathbf{p}^{(i+1)} + \sum_{\ell=1}^{i} \hat{a}_{\ell i}^{(m)} q^{(\ell)}\rangle \\
&= \langle \mathbf{q}^{(i+1)}, \mathbf{p}^{(i+1)}\rangle = \|\mathbf{p}^{(i+1)}\| \qquad \text{für alle } i \in [0 : m-1],\ m \in [0 : m_0 - 1]
\end{aligned}$$

gilt, wir also mit der für die Orthonormalisierung ohnehin benötigten Norm auch noch einen Eintrag der Matrix „nebenbei“ berechnen.

Für eine praktische Konstruktion der Arnoldi-Basis wäre es von Vorteil, wenn wir auch den Parameter m_0 berechnen könnten. Nach Definition 8.3 gilt $\mathbf{p}^{(m_0)} = \mathbf{0}$, also können wir diesen Vektor verwenden, um m_0 zu ermitteln. In der numerischen Praxis dürfen wir nicht auf einen exakten Nullvektor hoffen, stattdessen müssen wir prüfen, ob eine vorgegebene Fehlerschranke unterschritten wurde. Hier bietet sich ein Kriterium der Form

$$\|\mathbf{p}^{(m+1)}\|_2 \leq \epsilon_{\mathrm{ir}} \|\mathbf{A}\mathbf{q}^{(m)}\|_2$$

mit einem $\epsilon_{\mathrm{ir}} \in \mathbb{R}_{>0}$ an, denn diese Bedingung wird genau dann erfüllt, wenn der „größte Teil“ des Vektors $\mathbf{A}\mathbf{q}^{(m)}$ bereits in $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m)$ enthalten ist und deshalb bei der Orthonormalisierung entfernt wurde.

Algorithmus 8.5 (Arnoldi-Basis) *Seien $\mathbf{A} \in \mathbb{K}^{n\times n}$ und $\mathbf{q}^{(0)} \in \mathbb{K}^n$ mit $\|\mathbf{q}^{(0)}\|_2 = 1$ gegeben. Der folgende Algorithmus berechnet die Arnoldi-Basis $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m_0-1)}$ und die gemäß (8.3) definierten Matrizen $\widehat{\mathbf{A}}^{(m)}$. Der Algorithmus endet mit $m = m_0$.*

$\mathbf{p} \leftarrow \mathbf{A}\mathbf{q}^{(0)}$; $\quad \gamma \leftarrow \|\mathbf{p}\|_2$
$\hat{a}_{00} \leftarrow \langle \mathbf{q}^{(0)}, \mathbf{p}\rangle_2$; $\quad \mathbf{p} \leftarrow \mathbf{p} - \hat{a}_{00}\mathbf{q}^{(0)}$
$\hat{a}_{10} \leftarrow \|\mathbf{p}\|_2$
$m \leftarrow 1$
while $\hat{a}_{m,m-1} \geq \epsilon_{ir}\gamma$ **do begin**
  $\mathbf{q}^{(m)} \leftarrow \mathbf{p}/\hat{a}_{m,m-1}$
  $\mathbf{p} \leftarrow \mathbf{A}\mathbf{q}^{(m)}$; $\quad \gamma \leftarrow \|\mathbf{p}\|_2$
  for $i \in [0:m]$ **do begin**
    $\hat{a}_{im} \leftarrow \langle \mathbf{q}^{(i)}, \mathbf{p}\rangle_2$; $\quad \mathbf{p} \leftarrow \mathbf{p} - \hat{a}_{im}\mathbf{q}^{(i)}$
  end
  $\hat{a}_{m+1,m} \leftarrow \|\mathbf{p}\|_2$
  $m \leftarrow m + 1$
end

Falls $\mathbf{A}$ selbstadjungiert ist, gilt dasselbe nach Konstruktion auch für $\widehat{\mathbf{A}}^{(m)}$, und da diese Matrix auch eine Hessenberg-Matrix ist, muss sie dann sogar tridiagonal sein, so dass sich der Rechenaufwand für die Orthonormalisierung und die Bestimmung von $\lambda_1^{(m)}$ noch weiter reduzieren lässt: Es gilt

$$\hat{a}_{im} = \overline{\hat{a}_{mi}} = 0 \qquad \text{für alle } i \in [0:m-2],$$

so dass wir die meisten der für die Orthogonalisierung erforderlichen Skalarprodukte einsparen können. Wenn wir die Tridiagonalmatrix $\widehat{\mathbf{A}}$ in der bereits aus Kapitel 7 bekannten Form

$$\widehat{\mathbf{A}} = \begin{pmatrix} \alpha_1 & \bar{\beta}_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \bar{\beta}_{m-1} \\ & & & \beta_{m-1} & \alpha_m \end{pmatrix}$$

darstellen, ergibt sich

$$\hat{a}_{m-1,m} = \overline{\hat{a}_{m,m-1}} = \bar{\beta}_{m-1},$$

so dass ein weiteres Skalarprodukt entfällt. Der resultierende sehr effiziente *Lanczos-Algorithmus* nimmt damit die folgende Form an:

Algorithmus 8.6 (Lanczos-Algorithmus) *Seien* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *selbstadjungiert und* $\mathbf{q}^{(0)} \in \mathbb{K}^n$ *mit* $\|\mathbf{q}^{(0)}\|_2 = 1$ *gegeben. Der folgende Algorithmus berechnet die Arnoldi-Basis* $\mathbf{q}^{(0)}, \ldots, \mathbf{q}^{(m_0-1)}$ *und die Matrizen* $\widehat{\mathbf{A}}^{(m)}$*. Der Algorithmus endet mit* $m = m_0$*.*

$\mathbf{p} \leftarrow \mathbf{A}\mathbf{q}^{(0)}$; $\quad \gamma \leftarrow \|\mathbf{p}\|_2$;
$\alpha_1 \leftarrow \langle \mathbf{q}^{(0)}, \mathbf{p}\rangle_2$; $\quad \mathbf{p} \leftarrow \mathbf{p} - \alpha_1 \mathbf{q}^{(0)}$;
$\beta_1 \leftarrow \|\mathbf{p}\|_2$;
$m \leftarrow 1$;
while $\beta_m \geq \epsilon_{ir}\gamma$ *do begin*
  $\mathbf{q}^{(m)} \leftarrow \mathbf{p}/\beta_m$;
  $\mathbf{p} \leftarrow \mathbf{A}\mathbf{q}^{(m)}$; $\quad \gamma \leftarrow \|\mathbf{p}\|_2$;
  $\alpha_{m+1} \leftarrow \langle \mathbf{q}^{(m)}, \mathbf{p}\rangle_2$; $\quad \mathbf{p} \leftarrow \mathbf{p} - \bar{\beta}_m \mathbf{q}^{(m-1)} - \alpha_{m+1}\mathbf{q}^{(m)}$;
  $\beta_{m+1} \leftarrow \|\mathbf{p}\|_2$;
  $m \leftarrow m + 1$
end

Da der Algorithmus abbricht, sobald $\beta_m = 0$ gilt, berechnet er nicht nur eine selbstadjungierte Tridiagonalmatrix, sondern sogar eine *irreduzible* selbstadjungierte Tridiagonalmatrix. Deshalb lässt sich die Theorie aus Kapitel 7 anwenden, um beispielsweise die Eigenwerte mit Hilfe Sturmscher Ketten zu berechnen.

Es muss leider darauf hingewiesen werden, dass in der Praxis Rundungsfehler häufig zu einem Verlust der Orthogonalität der mit dem Lanczos-Verfahren berechneten Basis führen. Dadurch kann es beispielsweise passieren, dass $\widehat{\mathbf{A}}^{(m)}$ bestimmte Eigenwerte bis auf Rundungsfehler mehrfach aufweist, die in $\mathbf{A}$ nur einfach auftreten.

8.4 Konvergenz

Offensichtlich gilt

$$\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m) = \{p(\mathbf{A})\mathbf{q}^{(0)} \ : \ p \in \Pi_m\},$$

wobei Π_m den Raum der Polynome vom Grad $\leq m$ bezeichnet.

Bei der folgenden Untersuchung beschränken wir uns auf den Fall selbstadjungierter Matrizen, setzen also $\mathbf{A}^* = \mathbf{A}$ voraus. Mit Folgerung 3.41 finden wir eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine reelle Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

derart, dass

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$$

gilt. Daraus folgt

$$p(\mathbf{A}) = \mathbf{Q}p(\mathbf{D})\mathbf{Q}^* \qquad \text{für alle } p \in \Pi_m,\ m \in \mathbb{N}_0, \tag{8.4}$$

und mit Hilfe dieser Gleichung können wir die Elemente des Krylow-Raums analysieren. Falls beispielsweise alle Eigenwerte verschieden sind, können wir für $j \in [1:n]$ das *j-te Lagrange-Polynom*

$$\ell_j(x) := \prod_{\substack{i=1\\ i\neq j}}^{n} \frac{x-\lambda_i}{\lambda_j-\lambda_i}$$

untersuchen. Mit $\widehat{\mathbf{q}}^{(0)} := \mathbf{Q}^*\mathbf{q}^{(0)}$ und $\mathbf{x}^{(i)} := \mathbf{Q}\delta^{(i)}$ erhalten wir

$$\begin{aligned}\ell_j(\mathbf{A})\mathbf{q}^{(0)} &= \mathbf{Q}\ell_j(\mathbf{D})\mathbf{Q}^*\mathbf{q}^{(0)} = \mathbf{Q}\ell_j(\mathbf{D})\widehat{\mathbf{q}}^{(0)} = \mathbf{Q}\ell_j(\mathbf{D})\sum_{i=1}^{n}\delta^{(i)}\hat{q}_i^{(0)}\\ &= \mathbf{Q}\sum_{i=1}^{n}\delta^{(i)}\ell_j(\lambda_i)\hat{q}_i^{(0)} = \sum_{i=1}^{n}\mathbf{x}^{(i)}\ell_j(\lambda_i)\hat{q}_i^{(0)} = \mathbf{x}^{(j)}\hat{q}_j^{(0)},\end{aligned}$$

haben also im Fall $\hat{q}_j^{(0)} \neq 0$ einen Eigenvektor zu dem Eigenwert λ_j gefunden.

Allerdings gilt $\ell_j \in \Pi_n$, so dass wir den Krylow-Raum

$$\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m_0 - 1) \supseteq \mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, n)$$

maximaler Ordnung verwenden müssten, um in dieser Weise Eigenvektoren zu konstruieren. Der dazu erforderliche Aufwand ist in der Regel zu hoch.

Deshalb interessieren wir uns eher für Polynome, die die Lagrange-Polynome in geeigneter Weise *approximieren*, um aus ihnen Approximationen der Eigenvektoren zu konstruieren.

Lemma 8.7 (Eigenvektor-Approximation) *Sei $\mathbf{q}^{(0)} \in \mathbb{K}^n$ mit $\|\mathbf{q}^{(0)}\| = 1$ gegeben und sei $\mathbf{e}^{(1)} := \mathbf{Q}\delta^{(1)}$. Sei $p \in \Pi_m$ ein Polynom mit*

$$p(\lambda_1) = 1, \qquad |p(x)| \leq \epsilon \qquad \text{für alle } x \in \{\lambda_2, \ldots, \lambda_n\}.$$

Dann gilt

$$\tan\angle(p(\mathbf{A})\mathbf{q}^{(0)}, \mathbf{e}^{(1)}) \leq \epsilon \tan\angle(\mathbf{q}^{(0)}, \mathbf{e}^{(1)}).$$

Beweis. Mit (8.4) und $\widehat{\mathbf{q}}^{(0)} := \mathbf{Q}^*\mathbf{q}^{(0)}$ folgt

$$\begin{aligned}\tan^2\angle(p(\mathbf{A})\mathbf{q}^{(0)}, \mathbf{e}^{(1)}) &= \tan^2\angle(\mathbf{Q}p(\mathbf{D})\mathbf{Q}^*\mathbf{q}^{(0)}, \mathbf{Q}\delta^{(1)})\\ &= \tan^2\angle(p(\mathbf{D})\widehat{\mathbf{q}}^{(0)}, \delta^{(1)}) = \frac{\sum_{i=2}^{n}|p(\lambda_i)|^2|\hat{q}_i^{(0)}|^2}{|p(\lambda_1)|^2|\hat{q}_1^{(0)}|^2}\end{aligned}$$

$$\leq \frac{\epsilon^2 \sum_{i=2}^{n} |\hat{q}_i^{(0)}|^2}{|\hat{q}_1^{(0)}|^2} = \epsilon^2 \tan^2 \angle(\widehat{\mathbf{q}}^{(0)}, \delta^{(1)})$$
$$= \epsilon^2 \tan^2 \angle(\mathbf{Q}\widehat{\mathbf{q}}^{(0)}, \mathbf{Q}\delta^{(1)}) = \epsilon^2 \tan^2 \angle(\mathbf{q}^{(0)}, \mathbf{x}^{(1)}).$$

Das ist die gewünschte Abschätzung. ∎

Um ein effizientes Verfahren zu erhalten, sollten wir darauf achten, dass die Ordnung von p nicht zu hoch wird. Zur Konstruktion derartiger Polynome sind die *Tschebyscheff-Polynome* besonders gut geeignet:

Definition 8.8 (Tschebyscheff-Polynome) *Wir definieren durch*

$$\begin{aligned} c_0 &: \mathbb{K} \to \mathbb{K}, \quad x \mapsto 1, \\ c_1 &: \mathbb{K} \to \mathbb{K}, \quad x \mapsto x, \\ c_{m+2} &: \mathbb{K} \to \mathbb{K}, \quad x \mapsto 2xc_{m+1}(x) - c_m(x), \end{aligned}$$

für alle $m \in \mathbb{N}_0$ die Folge der Tschebyscheff-Polynome.

Aus dieser Darstellung lässt sich direkt ableiten, dass $c_m \in \Pi_m$ gilt, und die Rekurrenzrelation kann sich bei der Implementierung als nützlich erweisen. Für die theoretische Untersuchung sind alternative Darstellungen der Tschebyscheff-Polynome nützlicher:

Lemma 8.9 (Alternative Darstellungen) *Für alle $x \in [-1, 1]$ und alle $m \in \mathbb{N}_0$ gilt*

$$c_m(x) = \cos(m \arccos x).$$

Für alle $x \in \mathbb{R}$ und alle $m \in \mathbb{N}_0$ gilt

$$c_m(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2 - 1}\right)^m + \left(x + \sqrt{x^2 - 1}\right)^{-m}\right),$$

wobei im Falle $x \in (-1, 1)$ die komplexe Wurzelfunktion so gewählt werden muss, dass $\sqrt{-1} = i$ gilt.

Beweis. Die erste Aussage erhalten wir, indem wir mit Hilfe des Additionstheorems

$$\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$$

die Gleichung

$$2x\cos((k+1)\arccos x) = \cos(k \arccos x) + \cos((k+2)\arccos x)$$

nachweisen und daraus per Induktion die gewünschte Gleichung folgern.

Zum Nachweis der zweiten Aussage nutzen wir aus, dass es wegen des Identitätssatzes für holomorphe Funktionen ausreicht, die Gleichung für $x \in [-1, 1]$ nachzuweisen. Diesen Nachweis führen wir, indem wir ausgehend von der ersten Gleichung die Darstellung

$$\cos(k\xi) = \mathrm{Re}(e^{ik\xi}) = \frac{1}{2}(e^{i\xi k} + e^{-i\xi k}) = \frac{1}{2}\left((e^{i\xi})^k + (e^{i\xi})^{-k}\right)$$

mit $\xi = \arccos x$ verwenden und anschließend

$$e^{i\xi} = \cos(\xi) + i\sin(\xi) = \cos(\xi) + i\sqrt{1-\cos^2(\xi)} = \cos(\xi) + \sqrt{\cos^2(\xi) - 1}$$

einsetzen. ∎

Lemma 8.10 (Optimalität) *Sei $x_0 \in \mathbb{R} \setminus [-1,1]$. Unter allen Polynomen $p \in \Pi_m$ mit $p(x_0) = c_m(x_0)$ besitzt c_m die kleinste Maximumnorm auf dem Intervall $[-1,1]$.*

Beweis. Die Maximumnorm von c_m auf dem Intervall $[-1,1]$ beträgt wegen $c_m(x) = \cos(m \arccos x)$ gerade eins. Falls also ein Polynom p existiert, dessen Maximumnorm nicht größer als eins ist, muss $|p(x_\nu)| \leq 1 = |c_m(x_\nu)|$ in den Punkten

$$x_\nu = \cos(\pi\nu/m) \qquad \text{für alle } \nu \in \{0,\ldots,m\}$$

gelten. Da c_m zwischen zwei dieser Punkte sein Vorzeichen wechselt, muss auch $c_m - p$ sein Vorzeichen wechseln. Daraus folgert man mit Hilfe des Mittelwertsatzes, dass $c_m - p$ auf jedem Intervall $[x_\nu, x_{\nu+1}]$ mindestens eine Nullstelle besitzen muss, also insgesamt mindestens m auf dem Intervall $[-1,1]$. Da x_0 außerhalb dieses Intervalls liegt und $c_m(x_0) - p(x_0) = 0$ gilt, besitzt $c_m - p$ mindestens $m+1$ Nullstellen. Aus dem Identitätssatz für Polynome folgt $c_m = p$. ∎

Um eine Approximation des kleinsten Eigenwerts λ_1 zu erhalten, benötigen wir ein Polynom niedriger Ordnung, das auf dem Intervall $[\lambda_2, \lambda_n]$ möglichst kleine Werte annimmt. Die Tschebyscheff-Polynome sind auf $[-1,1]$ in der oben präzisierten Weise minimal, also bietet es sich an, sie so zu transformieren, dass diese Minimalität sich auf $[\lambda_2, \lambda_n]$ überträgt.

Die Abbildung

$$\Phi : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \frac{\lambda_n - x}{\lambda_n - \lambda_2} + \frac{x - \lambda_2}{\lambda_n - \lambda_2}(-1) = \frac{\lambda_n + \lambda_2 - 2x}{\lambda_n - \lambda_2}$$

erfüllt $\Phi(\lambda_2) = 1$ und $\Phi(\lambda_n) = -1$ sowie

$$x_0 := \Phi(\lambda_1) = \frac{\lambda_n + \lambda_2 - 2\lambda_1}{\lambda_n - \lambda_2} = \frac{\lambda_n - \lambda_2 + 2(\lambda_2 - \lambda_1)}{\lambda_n - \lambda_2} = 1 + 2\frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_2} > 1. \tag{8.5}$$

Mit Hilfe dieser Transformation können wir nun das für die Berechnung der Eigenvektoren benötigte p_ϵ optimal wählen: Zu einer gegebenen Dimension $m \in \mathbb{N}$ setzen wir

$$p_m(x) := \frac{1}{c_m(x_0)} c_m(\Phi(x)) \tag{8.6}$$

und erhalten

$$p_m(\lambda_1) = \frac{1}{c_m(x_0)} c_m(\Phi(\lambda_1)) = 1, \qquad |p_m(x)| \leq \frac{1}{c_m(x_0)} \qquad \text{für alle } x \in [\lambda_2, \lambda_n].$$

Um konkret angeben zu können, wie gut die durch p_m induzierte Approximation des Eigenvektors ist, benötigen wir eine Abschätzung für $1/c_m(x_0)$:

Lemma 8.11 (Konvergenzrate) *Wir definieren*

$$\kappa := \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} > 1, \qquad \varrho := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \tag{8.7}$$

und erhalten

$$\frac{1}{|c_m(x_0)|} = \frac{2\varrho^m}{1 + \varrho^{2m}} \leq 2\varrho^m \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweis. Da wir die Tschebyscheff-Polynome in der zweiten Darstellung aus Lemma 8.9 verwenden wollen, müssen wir $x_0 + \sqrt{x_0^2 - 1}$ geeignet darstellen. Wir beginnen mit

$$\begin{aligned} x_0^2 - 1 &= \left(1 + 2\frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_2}\right)^2 - 1 = 1 + 4\frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_2} + 4\frac{(\lambda_2 - \lambda_1)^2}{(\lambda_n - \lambda_2)^2} - 1 \\ &= 4\frac{(\lambda_2 - \lambda_1)(\lambda_n - \lambda_2) + (\lambda_2 - \lambda_1)^2}{(\lambda_n - \lambda_2)^2} = 4\frac{(\lambda_2 - \lambda_1)(\lambda_n - \lambda_1)}{(\lambda_n - \lambda_2)^2}. \end{aligned}$$

Aus dieser Gleichung und (8.5) ergibt sich

$$\begin{aligned} x_0 + \sqrt{x_0^2 - 1} &= \frac{(\lambda_n - \lambda_2) + 2(\lambda_2 - \lambda_1) + 2\sqrt{\lambda_2 - \lambda_1}\sqrt{\lambda_n - \lambda_1}}{\lambda_n - \lambda_2} \\ &= \frac{(\lambda_2 - \lambda_1) + (\lambda_n - \lambda_1) + 2\sqrt{\lambda_2 - \lambda_1}\sqrt{\lambda_n - \lambda_1}}{\lambda_n - \lambda_2} \\ &= \frac{(\sqrt{\lambda_2 - \lambda_1} + \sqrt{\lambda_n - \lambda_1})^2}{\lambda_n - \lambda_2} = \frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_2}\left(1 + \sqrt{\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}}\right)^2 \\ &= \frac{\lambda_2 - \lambda_1}{(\lambda_n - \lambda_1) - (\lambda_2 - \lambda_1)}(1 + \sqrt{\kappa})^2 = \left(\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} - 1\right)^{-1}(1 - \sqrt{\kappa})^2 \\ &= \frac{1}{\kappa - 1}(1 + \sqrt{\kappa})^2 = \frac{(\sqrt{\kappa} + 1)^2}{(\sqrt{\kappa} + 1)(\sqrt{\kappa} - 1)} = \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} = 1/\varrho. \end{aligned}$$

Einsetzen in die zweite Darstellung in Lemma 8.9 ergibt

$$c_m(x_0) = \frac{1}{2}((1/\varrho)^m + (1/\varrho)^{-m}) = \frac{1}{2}(\varrho^{-m} + \varrho^m) = \frac{1 + \varrho^{2m}}{2\varrho^m}$$

und damit

$$\frac{1}{c_m(x_0)} = \frac{2\varrho^m}{\varrho^{2m} + 1}.$$

■

Die für uns wichtigsten Eigenschaften des durch (8.6) gegebenen transformierten Tschebyscheff-Polynoms fassen wir in folgendem Lemma zusammen:

Lemma 8.12 (Transformiertes Polynom) *Sei $m \in \mathbb{N}_0$, und seien $\lambda_1 < \lambda_2 \leq \lambda_n$ gegeben. Dann existiert ein Polynom $p_m \in \Pi_m$ mit*

$$p_m(\lambda_1) = 1, \qquad \max\{|p_m(x)| \ : \ x \in [\lambda_2, \lambda_n]\} \leq \frac{2\varrho^m}{1+\varrho^{2m}}$$

mit ϱ aus (8.7).

Beweis. Wir definieren p_m durch (8.6) und wenden Lemma 8.11 an. Da c_m auf $[-1,1]$ nur Werte zwischen -1 und 1 annehmen kann, kann p_m auf $[\lambda_2, \lambda_n]$ nur Werte zwischen $-1/c_m(x_0)$ und $1/c_m(x_0)$ annehmen. ∎

Aus diesem Lemma folgt unmittelbar die folgende Konvergenzaussage für den Eigenvektor $\mathbf{x}^{(1)}$:

Folgerung 8.13 (Eigenvektor-Approximation) *Es gilt*

$$\tan \angle(p_m(\mathbf{A})\mathbf{q}^{(0)}, \mathbf{e}^{(1)}) \leq \frac{2\varrho^m}{1+\varrho^{2m}} \tan \angle(\mathbf{q}^{(0)}, \mathbf{e}^{(1)}) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Wir kombinieren Lemma 8.7 mit Lemma 8.12. ∎

Die Größe κ beschreibt, wie groß der Abstand zwischen λ_1 und λ_2 im Vergleich zum „Durchmesser“ $\lambda_n - \lambda_1$ des gesamten Spektrums ist. Je kleiner dieser Abstand wird, desto größer wird κ und desto näher rückt die „Konvergenzrate“ ϱ an eins heran. Wie schon bei der Vektoriteration ist es also auch hier von Vorteil, wenn die Eigenwerte, die wir berechnen wollen, einen möglichst großen Abstand zum Rest des Spektrums aufweisen.

Man beachte, dass Folgerung 8.13 nur eine Existenzaussage bietet: Wir wissen, dass mit $p_m(\mathbf{A})\mathbf{q}^{(0)}$ ein Vektor in dem Krylow-Raum $\mathcal{K}(\mathbf{A}, \mathbf{q}^{(0)}, m)$ existiert, der $\mathbf{e}^{(1)}$ approximiert, aber wir können ihn nicht berechnen, weil wir die exakten Werte von λ_1, λ_2 und λ_n nicht kennen, die für die Konstruktion von p_m erforderlich wären.

Bei Eigenwerten dagegen können wir einen direkten Bezug zwischen den berechenbaren Näherungswerten $\lambda_1^{(m)}$ und dem Eigenwert λ_1 herstellen:

Satz 8.14 *Sei $\mathbf{A} = \mathbf{A}^* \in \mathbb{K}^{n\times n}$. Seien $\lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$ die Eigenwerte von $\mathbf{A}$. Dann ist $\mathbf{e}^{(1)} = \mathbf{Q}\delta^{(1)}$ ein Eigenvektor zu dem Eigenwert λ_1. Sei $m \in [0 : m_0 - 1]$ und sei $\lambda_1^{(m)}$ der kleinste Eigenwert von $\widehat{\mathbf{A}}^{(m)}$. Dann gilt*

$$\lambda_1 \leq \lambda_1^{(m)} \leq \lambda_1 + (\lambda_n - \lambda_1)\left(\frac{2\varrho^m}{1+\varrho^{2m}}\right)^2 \tan^2 \angle(\mathbf{q}^{(0)}, \mathbf{e}^{(1)}).$$

für die Konstante ϱ aus (8.7).

Beweis. Die linke Abschätzung folgt direkt aus

$$\lambda_1^{(m)} = \min\{\Lambda_{\widehat{A}^{(m)}}(\widehat{\mathbf{x}}) \ : \ \widehat{\mathbf{x}} \in \mathbb{R}^k \setminus \{\mathbf{0}\}\} = \min\{\Lambda_A(\mathbf{x}) \ : \ \mathbf{x} \in \operatorname{Bild} \mathbf{Q}^{(m)} \setminus \{\mathbf{0}\}\}$$

$$\geq \min\{\Lambda_A(\mathbf{x}) \ : \ \mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\} = \lambda_1.$$

Für die rechte Abschätzung kombinieren wir Folgerung 8.13 mit Lemma 5.11. Wir setzen $\mathbf{y} := p_m(\mathbf{A})\mathbf{q}^{(0)}$ und erhalten

$$\begin{aligned}\lambda_1^{(m)} - \lambda_1 \leq \Lambda_A(\mathbf{y}) - \lambda_1 = |\Lambda_A(\mathbf{y}) - \lambda_1| &\leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \sin^2(\mathbf{y}, \mathbf{e}^{(1)}) \\ &\leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \tan^2(p_m(\mathbf{A})\mathbf{q}^{(0)}, \mathbf{e}^{(1)}) \\ &\leq \|\mathbf{A} - \lambda_1\mathbf{I}\| \left(\frac{2\varrho^m}{1+\varrho^{2m}}\right)^2 \tan^2 \angle(\mathbf{q}^{(0)}, \mathbf{e}^{(1)}).\end{aligned}$$

Es bleibt nur noch die Norm der Matrix $\mathbf{A} - \lambda_1\mathbf{I}$ abzuschätzen:

$$\|\mathbf{A} - \lambda_1\mathbf{I}\| = \|\mathbf{Q}\mathbf{D}\mathbf{Q}^* - \lambda_1\mathbf{Q}\mathbf{Q}^*\| = \|\mathbf{Q}(\mathbf{D} - \lambda_1\mathbf{I})\mathbf{Q}^*\| = \|\mathbf{D} - \lambda_1\mathbf{I}\|.$$

Mit der Abschätzung

$$\|(\mathbf{D} - \lambda_1\mathbf{I})\mathbf{z}\|^2 = \sum_{i=1}^n (\lambda_i - \lambda_1)^2 |z_i|^2 \leq \sum_{i=1}^n (\lambda_n - \lambda_1)^2 |z_i|^2 = (\lambda_n - \lambda_1)^2 \|z\|^2$$

folgt $\|\mathbf{A} - \lambda_1\mathbf{I}\| = \|\mathbf{D} - \lambda_1\mathbf{I}\| \leq \lambda_n - \lambda_1$, also die Behauptung. ∎

Bemerkung 8.15 (Mehrfache Eigenwerte) *Die Voraussetzung $\lambda_1 < \lambda_2$ ist in diesem Fall nicht entscheidend, sie dient lediglich der Vereinfachung des Beweises. Falls $\lambda_1 = \ldots = \lambda_k < \lambda_{k+1}$ gelten sollte, können wir den Beweis wie gehabt durchführen und müssen lediglich das Polynom p_m so wählen, dass es auf $[\lambda_{k+1}, \lambda_n]$ kleine Werte annimmt. Außerdem müssen wir den Tangens des Winkels zwischen $\mathbf{q}^{(1)}$ und $\mathbf{x}^{(1)}$ durch den Tangens des Winkels zwischen $\mathbf{q}^{(1)}$ und dem von $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ aufgespannten Eigenraum zum Eigenwert λ_1 ersetzen.*

Die Fehlerabschätzung gilt dann mit λ_{k+1} anstelle von λ_2, entscheidend ist also der Abstand zwischen $\lambda_1 = \ldots = \lambda_k$ und dem Rest des Spektrums.

Bemerkung 8.16 (Modellproblem) *Im Fall des zweidimensionalen Modellproblems sind wir vor allem daran interessiert, mit kleinen Gitterschrittweiten $h \in \mathbb{R}_{>0}$ zu arbeiten. Für kleines h stellt man fest, dass $\lambda_n \approx ch^{-2}$ für eine Konstante $c \in \mathbb{R}_{>0}$ gilt, der größte Eigenwert wird also sehr groß werden. Damit folgt auch $\kappa \approx ch^{-2}$, und wir erhalten*

$$\varrho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \frac{2}{\sqrt{\kappa}+1} \approx 1 - \frac{2}{\sqrt{ch^{-2}}+1} \approx 1 - \frac{2}{\sqrt{c}}h,$$

die Konvergenzgeschwindigkeit unseres Verfahrens wird also für zunehmend feine Gitter zunehmend langsamer werden.

9 Eigenwertverfahren für sehr große Matrizen

Die Konvergenz des Lanczos-Verfahrens hängt von dem Verhältnis zwischen dem größten und dem kleinsten Eigenwert der Matrix $\mathbf{A}$ ab, also von ihrer Konditionszahl. Gerade bei Eigenwertproblemen, die im Kontext partieller Differentialgleichungen auftreten, wird diese Konditionszahl sehr hoch sein, so dass wir mit einer relativ langsamen Konvergenz rechnen müssen. Deshalb interessiert man sich für Verfahren, die weniger empfindlich auf die Konditionszahl reagieren. Die Idee besteht darin, einen *Vorkonditionierer* zu verwenden, also eine Matrix $\mathbf{B}$, die so gewählt, ist, dass einerseits die Konditionszahl der Matrix $\mathbf{BA}$ günstiger als die der ursprünglichen Matrix ist und sich andererseits Matrix-Vektor-Multiplikationen mit der Matrix $\mathbf{B}$ effizient durchführen lassen. Mit Hilfe der so entstehenden *vorkonditionierten Eigenwertverfahren* lassen sich auch sehr große Eigenwertprobleme effizient behandeln.

9.1 Richardson-Iteration

Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ wieder eine selbstadjungierte Matrix, es gelte also $\mathbf{A} = \mathbf{A}^*$. Nach Folgerung 3.47 finden wir eine orthogonale Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},$$

mit

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{D},$$

und da $\mathbf{A}$ selbstadjungiert ist, sind alle Eigenwerte reell. Offenbar lässt sich leicht sicherstellen, dass sie aufsteigend sortiert sind, dass also

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

gilt. Wir interessieren uns für den kleinsten Eigenwert λ_1 der Matrix.

Eines der einfachsten iterativen Lösungsverfahren für lineare Gleichungssysteme der Form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{9.1}$$

mit vorgegebener rechter Seite $\mathbf{b}$ ist die *Richardson-Iteration*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \theta(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}) \qquad \text{für alle } m \in \mathbb{N}_0,$$

die ausgehend von einem Startvektor $\mathbf{x}^{(0)}$ eine Folge von Näherungslösungen berechnet. Dabei ist die korrekte Wahl des *Dämpfungsparameters* $\theta \in \mathbb{R}_{>0}$ entscheidend für das Konvergenzverhalten.

Wir sind allerdings nicht daran interessiert, lineare Gleichungssysteme zu lösen, sondern daran, den kleinsten Eigenwert und einen passenden Eigenvektor zu berechnen. Zu diesem Zweck ersetzen wir (9.1) durch die Eigenwertgleichung

$$\mathbf{A}\mathbf{x} = \lambda_1 \mathbf{x},$$

die sich in die Form eines linearen Gleichungssystems

$$(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x} = \mathbf{0}$$

mit der rechten Seite $\mathbf{b} = \mathbf{0}$ bringen lässt. Wenn wir annehmen, dass uns eine Näherung $\mu \in \mathbb{R}$ des Eigenwerts λ_1 zur Verfügung steht, können wir das Richardson-Verfahren anwenden und erhalten die Iterationsvorschrift

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \theta(\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0. \tag{9.2}$$

Für die Analyse führen wir, wie schon häufiger, die transformierten Vektoren

$$\widehat{\mathbf{x}}^{(m)} := \mathbf{Q}^*\mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

ein und erhalten mit $\mathbf{x}^{(m)} = \mathbf{Q}\widehat{\mathbf{x}}^{(m)}$ aus (9.2) die Gleichung

$$\begin{aligned}
\widehat{\mathbf{x}}^{(m+1)} &= \mathbf{Q}^*\mathbf{x}^{(m+1)} = \mathbf{Q}^*\mathbf{x}^{(m)} - \theta(\mathbf{Q}^*\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{Q}^*\mathbf{x}^{(m)}) \\
&= \widehat{\mathbf{x}}^{(m)} - \theta(\mathbf{Q}^*\mathbf{A}\mathbf{Q}\widehat{\mathbf{x}}^{(m)} - \mu\widehat{\mathbf{x}}^{(m)}) \\
&= \widehat{\mathbf{x}}^{(m)} - \theta(\mathbf{D}\widehat{\mathbf{x}}^{(m)} - \mu\widehat{\mathbf{x}}^{(m)}) \\
&= (\mathbf{I} - \theta(\mathbf{D} - \mu\mathbf{I}))\widehat{\mathbf{x}}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0.
\end{aligned}$$

Für die einzelnen Komponenten des Vektors $\widehat{\mathbf{x}}^{(m)}$ folgt daraus

$$\widehat{x}_i^{(m)} = (1 - \theta(\lambda_i - \mu))^m \widehat{x}_i^{(0)} \qquad \text{für alle } m \in \mathbb{N}_0,\ i \in [1:n]. \tag{9.3}$$

Bei der Untersuchung dieses Ausdrucks sind zwei Fälle zu unterscheiden: Für $i \in [1:n]$ mit $\lambda_i \leq \mu$ erhalten wir

$$1 - \theta(\lambda_i - \mu) = 1 + \theta(\mu - \lambda_i) \geq 1, \tag{9.4}$$

diese Komponenten des Iterationsvektors werden also wachsen. Am stärksten wächst dabei die erste Komponente, da λ_1 der kleinste Eigenwert ist.

Für $i \in [1:n]$ mit $\lambda_i > \mu$ dagegen ergibt sich

$$1 + \theta\mu - \theta\lambda_i = 1 - \theta(\lambda_i - \mu) < 1.$$

Falls wir

$$\mu < \lambda_n, \qquad \theta < \frac{2}{\lambda_n - \mu} \tag{9.5}$$

sicherstellen können, folgt

$$-1 < 1 - \theta(\lambda_n - \mu) \leq 1 - \theta(\lambda_i - \mu) < 1,$$

also wird diese Komponente des Iterationsvektors fallen.

Falls wir die beiden Bedingungen (9.5) sicherstellen können, werden demnach alle Komponenten zu großen Eigenwerten gegen null konvergieren, die Richardson-Iteration wird also einen Vektor aus dem zu den kleinen Eigenwerten gehörenden invarianten Unterraum berechnen. Die erste der beiden Bedingungen lässt sich einfach erfüllen, indem wir den Rayleigh-Quotienten (vgl. Definition 5.9) verwenden: Für jeden beliebigen Vektor $\mathbf{x} \neq \mathbf{0}$ erfüllt

$$\mu := \Lambda_A(\mathbf{x}) = \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle_2}{\langle \mathbf{x}, \mathbf{x} \rangle_2}$$

nach dem Satz 3.42 von Courant und Fischer die Ungleichungen

$$\lambda_1 \leq \mu \leq \lambda_n, \tag{9.6}$$

und falls $x_i \neq 0$ für ein beliebiges $i \in [1 : n]$ mit $\lambda_i < \lambda_n$ gilt, folgt auch $\mu < \lambda_n$.

Offenbar kann die Rayleigh-Iteration nur gegen einen gewünschten Eigenvektor konvergieren, falls der Startvektor $\mathbf{x}^{(0)}$ Anteile dieses Eigenvektors enthält, und in diesem Fall muss $\Lambda_A(\mathbf{x}^{(0)}) < \lambda_n$ gelten.

Falls wir μ über den Rayleigh-Quotienten berechnen, folgt aus (9.6), dass wir die zweite der Bedingungen (9.5) erfüllen können, indem wir grob durch

$$\frac{2}{\lambda_n - \mu} \geq \frac{2}{\lambda_n - \lambda_1}$$

abschätzen und

$$\theta < \frac{2}{\lambda_n - \lambda_1}$$

fordern. Da häufig wenigstens grobe Schranken für das Spektrum einer zu untersuchenden Matrix berechnet werden können, beispielsweise mit Hilfe der Gerschgorin-Kreise (vgl. Satz 7.10), ist es möglich, in dieser Weise die Konvergenz der Richardson-Iteration zu garantieren.

Aus der Gleichung (9.3) können wir eine einfache Konvergenzaussage gewinnen:

Satz 9.1 (Konvergenz) *Seien $k \in [1 : n-1]$ und $\theta \in \mathbb{R}_{>0}$ so gewählt, dass*

$$\lambda_1 \leq \ldots \leq \lambda_k \leq \mu < \lambda_{k+1} \leq \ldots \leq \lambda_n$$

gilt. Sei mit

$$\mathbf{P} := \mathbf{Q}\widehat{\mathbf{P}}\mathbf{Q}^*, \qquad \widehat{\mathbf{P}} := \begin{pmatrix} \mathbf{I}_k & \\ & \mathbf{0} \end{pmatrix} \in \mathbb{K}^{n \times n}$$

die orthogonale Projektion auf den von den ersten k Eigenvektoren aufgespannten invarianten Teilraum bezeichnet (vgl. Lemma 5.37). Dann gelten

$$\|(\mathbf{I} - \mathbf{P})\mathbf{x}^{(m+1)}\|_2 \leq \varrho \|(\mathbf{I} - \mathbf{P})\mathbf{x}^{(m)}\|_2,$$

$$\|\mathbf{P}\mathbf{x}^{(m+1)}\|_2 \geq \|\mathbf{P}\mathbf{x}^{(m)}\|_2 \qquad \textit{für alle } m \in \mathbb{N}_0$$

mit der durch

$$\varrho := \max\{1-\theta(\lambda_{k+1}-\mu), \theta(\lambda_n-\mu)-1\}$$

gegebenen Konvergenzrate. Die optimale Konvergenzrate

$$\varrho_{opt} = \frac{(\lambda_n-\mu)-(\lambda_{k+1}-\mu)}{(\lambda_n-\mu)+(\lambda_{k+1}-\mu)} < 1$$

wird für den Parameter

$$\theta_{opt} = \frac{2}{(\lambda_n-\mu)+(\lambda_{k+1}-\mu)}$$

angenommen, liegt aber für alle $\theta \in (0, \theta_{opt}]$ unter eins.

Beweis. Infolge der Anordnung der Eigenwerte erhalten wir

$$1-\theta(\lambda_n-\mu) \leq 1-\theta(\lambda_i-\mu) \leq 1-\theta(\lambda_{k+1}-\mu) \qquad \text{für alle } i \in [k+1:n].$$

Damit kann insbesondere

$$|1-\theta(\lambda_n-\mu)| \geq |1-\theta(\lambda_{k+1}-\mu)|$$

nur dann gelten, wenn $1-\theta(\lambda_n-\mu)$ negativ ist und

$$\theta(\lambda_n-\mu)-1 \geq 1-\theta(\lambda_{k+1}-\mu)$$

erfüllt. Also folgt

$$|1-\theta(\lambda_i-\mu)| \leq \max\{1-\theta(\lambda_{k+1}-\mu), \theta(\lambda_n-\mu)-1\} = \varrho \quad \text{für alle } i \in [k+1:n].$$

Mit (9.4) erhalten wir

$$|1-\theta(\lambda_i\mu)| \geq 1 \qquad \text{für alle } i \in [1:k].$$

Dank (9.3) folgen daraus

$$\begin{aligned}
\|(\mathbf{I}-\widehat{\mathbf{P}})\widehat{\mathbf{x}}^{(m+1)}\|_2^2 &= \sum_{i=k+1}^{n} |\widehat{x}_i^{(m+1)}|^2 = \sum_{i=k+1}^{n} |1-\theta(\lambda_i-\mu)|^2\, |\widehat{x}_i^{(m)}|^2 \\
&\leq \sum_{i=k+1}^{n} \varrho^2 |\widehat{x}_i^{(m)}|^2 = \varrho^2 \|(\mathbf{I}-\widehat{\mathbf{P}})\widehat{\mathbf{x}}^{(m+1)}\|_2^2, \\
\|\widehat{\mathbf{P}}\widehat{\mathbf{x}}^{(m+1)}\|_2^2 &= \sum_{i=1}^{k} |\widehat{x}_i^{(m+1)}|^2 = \sum_{i=1}^{k} |1-\theta(\lambda_i-\mu)|^2 |\widehat{x}_i^{(m)}|^2 \\
&\geq \sum_{i=1}^{k} |\widehat{x}_i^{(m)}|^2 = \|\widehat{\mathbf{P}}\widehat{\mathbf{x}}^{(m)}\|_2^2 \qquad \text{für alle } m \in \mathbb{N}_0.
\end{aligned}$$

Durch Rücktransformation ergibt sich das gewünschte Ergebnis

$$\|(\mathbf{I}-\mathbf{P})\mathbf{x}^{(m+1)}\|_2 = \|(\mathbf{I}-\widehat{\mathbf{P}})\widehat{\mathbf{x}}^{(m+1)}\|_2 \le \varrho\|(\mathbf{I}-\widehat{\mathbf{P}})\widehat{\mathbf{x}}^{(m)}\|_2 = \varrho\|(\mathbf{I}-\mathbf{P})\mathbf{x}^{(m)}\|_2,$$
$$\|\mathbf{P}\mathbf{x}^{(m+1)}\|_2 = \|\widehat{\mathbf{P}}\widehat{\mathbf{x}}^{(m+1)}\|_2 \ge \|\widehat{\mathbf{P}}\widehat{\mathbf{x}}^{(m)}\|_2 = \|\mathbf{P}\mathbf{x}^{(m)}\|_2 \qquad \text{für alle } m \in \mathbb{N}_0.$$

Für die Bestimmung des optimalen Dämpfungsparameters θ gehen wir davon aus, dass ϱ als Maximum einer monoton fallenden und einer monoton wachsenden Funktion nur dann minimal sein kann, wenn

$$1-\theta(\lambda_{k+1}-\mu) = \theta(\lambda_n-\mu)-1$$

gilt, und diese Gleichung ist äquivalent mit

$$2 = \theta((\lambda_{k+1}-\mu)+(\lambda_n-\mu)).$$

Daraus ergibt sich die Formel für θ_{opt}. Durch Einsetzen in die Definition der Konvergenzrate ϱ erhalten wir

$$\begin{aligned}\varrho_{\text{opt}} &= 1-\theta_{\text{opt}}(\lambda_{k+1}-\mu) = 1-\frac{2(\lambda_{k+1}-\mu)}{(\lambda_n-\mu)+(\lambda_{k+1}-\mu)}\\ &= \frac{(\lambda_n-\mu)+(\lambda_{k+1}-\mu)-2(\lambda_{k+1}-\mu)}{(\lambda_n-\mu)+(\lambda_{k+1}-\mu)} = \frac{(\lambda_n-\mu)-(\lambda_{k+1}-\mu)}{(\lambda_n-\mu)+(\lambda_{k+1}-\mu)},\end{aligned}$$

und wegen $\mu < \lambda_{k+1}$ folgt $\varrho_{\text{opt}} < 1$. Für ein beliebiges $\theta \in (0, \theta_{\text{opt}}]$ erhalten wir

$$1 > 1-\theta(\lambda_{k+1}-\mu) \ge 1-\theta_{\text{opt}}(\lambda_{k+1}-\mu) = \theta_{\text{opt}}(\lambda_n-\mu)-1 \ge \theta(\lambda_n-\mu)-1,$$

also insbesondere $\varrho < 1$. ∎

Die Konvergenzaussage lässt sich kürzer fassen, indem wir den Winkel zwischen den Vektoren und dem invarianten Unterraum mit Hilfe der Gleichungen (5.16) ausdrücken.

Folgerung 9.2 (Konvergenz) *Seien k, θ und ϱ wie in Satz 9.1 gegeben. Sei $\mathcal{E}_k$ der von den Eigenvektoren zu den Eigenwerten $\lambda_1, \dots, \lambda_k$ aufgespannte Teilraum.*

Falls $\tan\angle(\mathbf{x}^{(0)}, \mathcal{E}_k) < \infty$ gilt, falls also der Startvektor Komponenten in $\mathcal{E}_k$ aufweist, folgt

$$\tan\angle(\mathbf{x}^{(m)}, \mathcal{E}_k) \le \varrho^m \tan\angle(\mathbf{x}^{(0)}, \mathcal{E}_k) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Sei $m \in \mathbb{N}_0$. Mit (5.16) gilt

$$\tan\angle(\mathbf{x}^{(m)}, \mathcal{E}_k) = \frac{\|(\mathbf{I}-\mathbf{P})\mathbf{x}^{(m)}\|_2}{\|\mathbf{P}\mathbf{x}^{(m)}\|_2}.$$

Aus Satz 9.1 folgt

$$\begin{aligned}\tan\angle(\mathbf{x}^{(m+1)}, \mathcal{E}_k) &= \frac{\|(\mathbf{I}-\mathbf{P})\mathbf{x}^{(m+1)}\|_2}{\|\mathbf{P}\mathbf{x}^{(m+1)}\|_2}\\ &\le \varrho\frac{\|(\mathbf{I}-\mathbf{P})\mathbf{x}^{(m)}\|_2}{\|\mathbf{P}\mathbf{x}^{(m)}\|_2} = \varrho\tan\angle(\mathbf{x}^{(m)}, \mathcal{E}_k) \qquad \text{für alle } m \in \mathbb{N}_0,\end{aligned}$$

und mit einer einfachen Induktion folgt daraus die Behauptung. ∎

9.2 Optimale Dämpfung

Es wäre von Vorteil, wenn wir den Dämpfungsparameter θ der Richardson-Iteration in optimaler Weise durch einen geeigneten Algorithms wählen lassen könnten.

Gemäß (9.2) ist die neue Iterierte $\mathbf{x}^{(m+1)}$ eine Linearkombination zwischen der alten Iterierten $\mathbf{x}^{(m)}$ und dem *Residuum*

$$\mathbf{r}^{(m)} := \mu\mathbf{x}^{(m)} - \mathbf{A}\mathbf{x}^{(m)},$$

also stellt sich die Frage, wie wir diese Linearkombination wählen sollen, um dem gesuchten Eigenvektor möglichst nahe zu kommen.

Bei der Beantwortung dieser Frage können wir uns wieder von Satz 3.42 leiten lassen: Der gewünschte Eigenwert λ_1 ist das Minimum des Rayleigh-Quotienten, also bietet es sich an,

$$\mathbf{x}^{(m+1)} \in \operatorname{span}\{\mathbf{x}^{(m)}, \mathbf{r}^{(m)}\}$$

so zu wählen, dass

$$\Lambda_A(\mathbf{x}^{(m+1)}) \leq \Lambda_A(\mathbf{z}) \qquad \text{für alle } \mathbf{z} \in \operatorname{span}\{\mathbf{x}^{(m)}, \mathbf{r}^{(m)}\} \tag{9.7}$$

gilt. Indem wir $\mathbf{x}^{(m)}$ und $\mathbf{r}^{(m)}$ als Spalten der Matrix

$$\mathbf{V}^{(m)} := \begin{pmatrix}\mathbf{x}^{(m)} & \mathbf{r}^{(m)}\end{pmatrix}$$

verwenden und die nächste Iterierte durch

$$\mathbf{x}^{(m+1)} = \mathbf{V}^{(m)}\mathbf{y}^{(m+1)}$$

ausdrücken, folgt

$$\lambda_A(\mathbf{V}^{(m)}\mathbf{y}^{(m+1)}) = \frac{\langle \mathbf{A}\mathbf{V}^{(m)}\mathbf{y}^{(m+1)}, \mathbf{V}^{(m)}\mathbf{y}^{(m+1)}\rangle_2}{\langle \mathbf{V}^{(m)}\mathbf{y}^{(m+1)}, \mathbf{V}^{(m)}\mathbf{y}^{(m+1)}\rangle_2} = \frac{\langle (\mathbf{V}^{(m)})^*\mathbf{A}\mathbf{V}^{(m)}\mathbf{y}^{(m+1)}, \mathbf{y}^{(m+1)}\rangle_2}{\langle (\mathbf{V}^{(m)})^*\mathbf{V}^{(m)}\mathbf{y}^{(m+1)}, \mathbf{y}^{(m+1)}\rangle_2},$$

wir können also die Minimierung des Rayleigh-Quotienten des ursprünglichen Problems auf die Minimierung eines sehr ähnlichen Quotienten

$$\frac{\langle \mathbf{B}^{(m)}\mathbf{y}^{(m+1)}, \mathbf{y}^{(m+1)}\rangle_2}{\langle \mathbf{G}^{(m)}\mathbf{y}^{(m+1)}, \mathbf{y}^{(m+1)}\rangle_2}$$

mit zweidimensionalen Matrizen

$$\mathbf{B}^{(m)} := (\mathbf{V}^{(m)})^*\mathbf{A}\mathbf{V}^{(m)}, \qquad \mathbf{G}^{(m)} := (\mathbf{V}^{(m)})^*\mathbf{V}^{(m)}$$

zurückführen. Stünde im Nenner lediglich das Skalarprodukt des Vektors $\mathbf{y}^{(m+1)}$ mit sich selbst, wüssten wir nach Satz 3.42, dass das Minimum des Rayleigh-Quotienten gerade der kleinste Eigenwert der Matrix $\mathbf{B}^{(m)}$ wäre, wir müssten also lediglich ein zweidimensionales Eigenwertproblem lösen.

Offenbar tritt diese vorteilhafte Situation nach Lemma 3.32 genau dann ein, wenn $(\mathbf{V}^{(m)})^*\mathbf{V}^{(m)} = \mathbf{I}$ gilt, wenn also $\mathbf{V}^{(m)}$ orthogonal ist. Da wir lediglich an dem Bild dieser Matrix interessiert sind, bietet es sich an, sie zu orthogonalisieren. Im einfachsten Fall könnten wir beispielsweise die *Gramsche Matrix* $\mathbf{G}^{(m)}$ berechnen und ihre Cholesky-Zerlegung

$$\mathbf{L}\mathbf{L}^* = \mathbf{G}^{(m)}$$

verwenden, um mit

$$\widehat{\mathbf{V}}^{(m)} := \mathbf{V}^{(m)}(\mathbf{L}^*)^{-1}$$

eine orthogonale Basis zu erhalten:

$$(\widehat{\mathbf{V}}^{(m)})^*\widehat{\mathbf{V}}^{(m)} = \mathbf{L}^{-1}(\mathbf{V}^{(m)})^*\mathbf{V}^{(m)}(\mathbf{L}^*)^{-1} = \mathbf{L}^{-1}\mathbf{G}^{(m)}(\mathbf{L}^*)^{-1} = \mathbf{I}.$$

Allerdings ist damit zu rechnen, dass im Laufe der Iteration $\mathbf{x}^{(m)}$ dem gesuchten Eigenvektor relativ nahe kommen wird, so dass $\mathbf{r}^{(m)}$ im Verhältnis sehr klein zu werden droht. Damit wäre die Gramsche Matrix $\mathbf{G}^{(m)}$ schlecht konditioniert und die Berechnung der Cholesky-Zerlegung deshalb potentiell ungenau.

Eine bessere Alternative besteht darin, wie schon bei der orthogonalen Iteration die numerisch sehr stabilen *Householder-Spiegelungen* zu verwenden, um die Matrix $\mathbf{V}^{(m)}$ auf obere Dreiecksgestalt zu bringen:

$$\mathbf{Q}_2\mathbf{Q}_1\mathbf{V}^{(m)} = \mathbf{R}.$$

Dabei eliminiert $\mathbf{Q}_1$ in der ersten Spalte der Matrix alle Einträge unterhalb der Diagonalen, $\mathbf{Q}_2$ bewirkt dasselbe für die zweite Spalte des Ergebnisses. Da $\mathbf{R}$ eine obere Dreiecksmatrix mit lediglich zwei Spalten ist, existiert eine Matrix $\widehat{\mathbf{R}} \in \mathbb{K}^{2\times 2}$ mit

$$\mathbf{R} = \begin{pmatrix}\widehat{\mathbf{R}}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{I}_2\\ \mathbf{0}\end{pmatrix}\widehat{\mathbf{R}},$$

und wir erhalten

$$\mathbf{Q}_2\mathbf{Q}_1\mathbf{V}^{(m)} = \begin{pmatrix}\mathbf{I}_2\\ \mathbf{0}\end{pmatrix}\widehat{\mathbf{R}}.$$

Da die Householder-Spiegelungen unitär sind, folgt

$$\mathbf{V}^{(m)} = \mathbf{Q}_1^*\mathbf{Q}_2^*\begin{pmatrix}\mathbf{I}_2\\ \mathbf{0}\end{pmatrix}\widehat{\mathbf{R}}.$$

Damit haben wir eine QR-Zerlegung der Matrix $\mathbf{V}^{(m)}$ gefunden:

$$\mathbf{V}^{(m)} = \mathbf{Q}^{(m)}\widehat{\mathbf{R}}, \qquad \mathbf{Q}^{(m)} = \mathbf{Q}_1^*\mathbf{Q}_2^*\begin{pmatrix}\mathbf{I}_2\\ \mathbf{0}\end{pmatrix}.$$

Im Unterschied zu konventionellen QR-Zerlegungen ist $\mathbf{Q}^{(m)}$ nicht quadratisch, sondern hat die Dimensionen der ursprünglichen Matrix $\mathbf{V}^{(m)}$. Nach Konstruktion enthält das Bild der orthogonalen Matrix $\mathbf{Q}^{(m)}$ das der Matrix $\mathbf{V}^{(m)}$, für injektives $\mathbf{V}^{(m)}$ stimmen beide sogar überein.

Indem wir $\mathbf{V}^{(m)}$ in dieser Weise orthogonalisieren und das zweidimensionale Eigenwertproblem für die Berechnung der nächsten Iterierten verwenden, erhalten wir den folgenden Algorithmus:

Algorithmus 9.3 (Gradientenverfahren) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert. Der folgende Algorithmus berechnet eine Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$, die unter geeigneten Bedingungen gegen einen Eigenvektor zu dem kleinsten Eigenwert von $\mathbf{A}$ konvergiert.*

$\mathbf{x} \leftarrow \mathbf{x}/\|\mathbf{x}\|_2$;
$\lambda \leftarrow \langle \mathbf{Ax}, \mathbf{x}\rangle_2$;
$\mathbf{r} \leftarrow \mathbf{Ax} - \lambda\mathbf{x}$;
while $\|\mathbf{r}\|_2$ *zu groß* *do begin*
  $\mathbf{V} \leftarrow \begin{pmatrix}\mathbf{x} & \mathbf{r}\end{pmatrix} \in \mathbb{K}^{n\times 2}$;
  Berechne eine QR-Zerlegung $\mathbf{QR} = \mathbf{V}$ *mit* $\mathbf{Q} \in \mathbb{K}^{n\times 2}$;
  $\mathbf{C} \leftarrow \mathbf{AQ}$;
  $\mathbf{B} \leftarrow \mathbf{Q}^*\mathbf{C}$; $\quad \{ = \mathbf{Q}^*\mathbf{AQ} \}$
  Finde einen normierten Eigenvektor $\mathbf{y}$ *für den minimalen Eigenwert* λ *der Matrix* $\mathbf{B}$;
  $\mathbf{x} \leftarrow \mathbf{Qy}$;
  $\mathbf{r} \leftarrow \mathbf{Cy} - \lambda\mathbf{x}$ $\quad \{ = \mathbf{Ax} - \lambda\mathbf{x} \}$
end

Die Bezeichnung „Gradientenverfahren“ ist dadurch motiviert, dass wegen Lemma 8.1

$$\operatorname{span}\{\mathbf{x}^{(m)}, \mathbf{r}^{(m)}\} = \operatorname{span}\{\mathbf{x}^{(m)}, \nabla\Lambda_A(\mathbf{x}^{(m)})\}$$

gilt, wir verbessern unsere Lösung also gerade in Richtung des Gradienten des Rayleigh-Quotienten. Da unser Algorithmus dabei die bestmögliche Schrittweite verwendet, entspricht er dem konventionellen Gradientenverfahren für die Minimierung des Rayleigh-Quotienten, allerdings mit der Besonderheit, dass die Iterationsvektoren in jedem Schritt normiert werden. Da der Rayleigh-Quotient invariant unter Skalierung des Arguments ist, ändert sich dadurch das Konvergenzverhalten nicht.

Anders als im Fall der Richardson-Iteration aktualisieren wir im Gradientenverfahren den Shift-Parameter μ in jedem Schritt. Da wir Konvergenz gegen den kleinsten Eigenwert λ_1 erwarten, bietet es sich an, eher die im Fall der Vektoriteration (vgl. Kapitel 5) verwendeten Argumente zu verwenden: Unser Ziel ist es, den Tangens der Winkels zwischen den Iterationsvektoren $\mathbf{x}^{(m)}$ und dem Eigenvektor $\mathbf{e}^{(1)}$ zu reduzieren.

Satz 9.4 (Konvergenz) *Gelte $\lambda_1 < \lambda_2$. Mit*

$$\varrho := \frac{(\lambda_n - \lambda_1) - (\lambda_2 - \lambda_1)}{(\lambda_n - \lambda_1) + (\lambda_2 - \lambda_1)} < 1$$

erhalten wir die Abschätzung

$$\tan\angle(\mathbf{x}^{(m+1)}, \mathbf{e}^{(1)}) \leq \varrho \tan\angle(\mathbf{x}^{(m)}, \mathbf{e}^{(1)}) \qquad \text{für alle } m \in \mathbb{N}_0,$$

das Gradientenverfahren konvergiert also gegen einen Eigenvektor zu dem kleinsten Eigenwert λ_1.

Beweis. Da das Gradientenverfahren nach Konstruktion den Rayleigh-Quotienten in dem Krylow-Raum $\mathcal{K}(A, \mathbf{x}^{(m)}, 1)$ minimiert, können wir seine Analyse auf die eines Lanczos-Verfahrens mit einem Schritt zurückführen.

Mit Lemma 8.11 erhalten wir eine Rate von

$$\varrho = \frac{2\hat{\varrho}}{1+\hat{\varrho}^2}$$

mit

$$\hat{\varrho} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \kappa = \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1},$$

so dass sich

$$\begin{aligned}
\varrho &= 2\left(\frac{1+\hat{\varrho}^2}{\hat{\varrho}}\right)^{-1} = 2\left(\frac{1}{\hat{\varrho}} + \hat{\varrho}\right)^{-1} = 2\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{-1} \\
&= 2\left(\frac{(\sqrt{\kappa}+1)^2 + (\sqrt{\kappa}-1)^2}{(\sqrt{\kappa}+1)(\sqrt{\kappa}-1)}\right)^{-1} = 2\left(\frac{\kappa + 2\sqrt{\kappa} + 1 + \kappa - 2\sqrt{\kappa} + 1}{\kappa - 1}\right)^{-1} \\
&= 2\left(\frac{2\kappa+2}{\kappa-1}\right)^{-1} = \frac{\kappa-1}{\kappa+1} = \frac{\frac{\lambda_n-\lambda_1}{\lambda_2-\lambda_1} - 1}{\frac{\lambda_n-\lambda_1}{\lambda_2-\lambda_1} + 1} = \frac{(\lambda_n-\lambda_1)-(\lambda_2-\lambda_1)}{(\lambda_n-\lambda_1)+(\lambda_2-\lambda_1)}
\end{aligned}$$

ergibt. Nun können wir fortfahren wie in Folgerung 8.13. ∎

Bemerkung 9.5 (Mehrfacher Eigenwert) *Falls $\lambda_1 = \lambda_2 = \ldots = \lambda_k < \lambda_{k+1}$ gilt, erhalten wir ein leicht modifiziertes Konvergenzresultat: Aus der Darstellung*

$$\mathbf{x}^{(0)} = \sum_{\ell=1}^{n} \alpha_\ell \mathbf{e}^{(\ell)}$$

des Startvektors können wir einen Vektor

$$\mathbf{e} := \sum_{\ell=1}^{k} \alpha_\ell \mathbf{e}^{(\ell)}$$

konstruieren. Falls $\mathbf{e} \neq \mathbf{0}$ gilt, ist $\mathbf{e}$ ein Eigenvektor zu dem mehrfachen Eigenwert $\lambda_1 = \ldots = \lambda_k$, und wir können wir in Satz 9.4 vorgehen, um

$$\tan \angle(\mathbf{x}^{(m+1)}, \mathbf{e}) \leq \varrho \tan \angle(\mathbf{x}^{(m)}, \mathbf{e}) \qquad \text{für alle } m \in \mathbb{N}_0$$

mit der Konvergenzrate

$$\varrho := \frac{\lambda_n - \lambda_{k+1}}{\lambda_n - \lambda_1} = 1 - \frac{\lambda_{k+1} - \lambda_1}{\lambda_n - \lambda_1} < 1$$

zu beweisen. Auch in diesem Fall konvergieren die Iterierten also gegen einen Eigenvektor zu dem kleinsten Eigenwert.

In vielen Anwendungen ist λ_n sehr viel größer als λ_1, so dass die Konvergenzrate ϱ des Gradientenverfahrens relativ nahe bei eins liegen wird und deshalb sehr viele Iterationsschritte durchgeführt werden müssen, um eine hinreichend gute Näherung eines Eigenvektors zu erhalten.

9.3 Vorkonditionierer

Unser Ziel besteht nun darin, die Geschwindigkeit des Gradientenverfahrens zu verbessern. Insbesondere stellt die Abhängigkeit der Konvergenz von dem größten Eigenwert λ_n ein Problem dar, da bei vielen Aufgabenstellungen, beispielsweise auch bei unserem Modellproblem, dieser Eigenwert sehr groß werden kann.

Die inverse Iteration kennt dieses Problem nicht, für sie konnten wir eine Konvergenzrate von $|\lambda_2|/|\lambda_1|$ nachweisen. Leider ist es unrealistisch, bei sehr großen Matrizen $\mathbf{A}$ die Inverse exakt zu berechnen, und die Approximation der Eigenwerte einer genäherten Inversen ist etwas unbefriedigend.

Wir können allerdings die inverse Iteration so umschreiben, dass wir mit Näherungen arbeiten können, aber weiterhin die „richtigen“ Eigenwerte bestimmen: Da wir lediglich an der Richtung der Vektoren interessiert sind, können wir die Iterierten der inversen Iteration beliebig skalieren. Wir wählen die Skalierung mit dem Rayleigh-Quotienten und erhalten

$$\begin{aligned}\mathbf{x}^{(m+1)} &= \Lambda_A(\mathbf{x}^{(m)})\mathbf{A}^{-1}\mathbf{x}^{(m)} = \mathbf{x}^{(m)} - \mathbf{A}^{-1}\mathbf{A}\mathbf{x}^{(m)} + \Lambda_A(\mathbf{x}^{(m)})\mathbf{A}^{-1}\mathbf{x}^{(m)} \\ &= \mathbf{x}^{(m)} - \mathbf{A}^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}).\end{aligned}$$

Falls $\mathbf{x}^{(m)}$ ein Eigenvektor der Matrix $\mathbf{A}$ ist, gilt $\mathbf{A}\mathbf{x}^{(m)} = \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}$, also $\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)}$, exakte Eigenvektoren sind also Fixpunkte dieser Iteration.

Diese Eigenschaft bleibt erhalten, wenn wir die exakte Inverse durch eine Näherung $\mathbf{N}$ ersetzen, um zu der *vorkonditionierten Richardson-Iteration*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{N}(\mathbf{A}\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0$$

zu gelangen. Für *jede* invertierbare *Vorkonditionierungsmatrix* $\mathbf{N}$ sind die Fixpunkte dieser Iteration genau die Eigenvektoren der Matrix $\mathbf{A}$.

Bemerkung 9.6 (Vorkonditionierung) *Der Begriff* Vorkonditionierung *stammt dabei aus der Welt der linearen Gleichungssysteme: Falls wir*

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

lösen wollen, aber die Konditionszahl der Matrix $\mathbf{A}$ groß ist, werden viele iterative Lösungsverfahren nur sehr langsam konvergieren. Deshalb ersetzt man das System durch das äquivalente System

$$\mathbf{N}\mathbf{A}\mathbf{x} = \mathbf{N}\mathbf{b},$$

in dem durch eine geeignete Vorkonditionierungsmatrix $\mathbf{N}$ die Konditionszahl reduziert und so das Konvergenzverhalten verbessert werden kann.

In unserem Fall wird, wie schon bei der Richardson-Iteration, die rechte Seite des linearen Gleichungssystems durch $\Lambda_A(\mathbf{x})\mathbf{x}$ ersetzt, aber die Idee der Vorkonditionierung bleibt unverändert.

Die vorkonditionierte Richardson-Iteration ähnelt sehr der uns bereits bekannten nicht vorkonditionierten Version, wir haben lediglich das

$$\mathbf{r}^{(m)} = \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)} - \mathbf{A}\mathbf{x}^{(m)}$$

durch ein *vorkonditioniertes Residuum*

$$\mathbf{q}^{(m)} = \mathbf{N}\mathbf{r}^{(m)} = \mathbf{N}(\Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)} - \mathbf{A}\mathbf{x}^{(m)})$$

ersetzt. Indem wir wieder einen optimalen Dämpfungsparameter wählen, gelangen wir zu dem *vorkonditionierte Gradientenverfahren* (auch bekannt als PINVIT, *preconditioned inverse iteration*).

Algorithmus 9.7 (Vorkonditioniertes Gradientenverfahren) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix, und sei $\mathbf{N} \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit. Der folgende Algorithmus berechnet eine Folge $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$, die unter geeigneten Bedingungen gegen einen Eigenvektor konvergiert.*

$\mathbf{x} \leftarrow \mathbf{x}/\|\mathbf{x}\|_2$;
$\lambda \leftarrow \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle_2$;
$\mathbf{r} \leftarrow \mathbf{A}\mathbf{x} - \lambda\mathbf{x}$;
$\mathbf{p} \leftarrow \mathbf{N}\mathbf{r}$;
while $\|\mathbf{r}\|_2$ *zu groß* **do begin**
  $\mathbf{V} \leftarrow \begin{pmatrix}\mathbf{x} & \mathbf{p}\end{pmatrix} \in \mathbb{K}^{n\times 2}$;
  Berechne eine QR-Zerlegung $\mathbf{QR} = \mathbf{V}$ *mit* $\mathbf{Q} \in \mathbb{K}^{n\times 2}$;
  $\mathbf{C} \leftarrow \mathbf{A}\mathbf{Q}$;
  $\mathbf{B} \leftarrow \mathbf{Q}^*\mathbf{C}$; $\quad \{ = \mathbf{Q}^*\mathbf{A}\mathbf{Q} \}$
  Finde einen normierten Eigenvektor $\mathbf{y}$ *für den minimalen Eigenwert* λ
    der Matrix $\mathbf{B}$;
  $\mathbf{x} \leftarrow \mathbf{Q}\mathbf{y}$;
  $\mathbf{r} \leftarrow \mathbf{C}\mathbf{y} - \lambda\mathbf{x}$ $\quad \{ = \mathbf{A}\mathbf{x} - \lambda\mathbf{x} \}$
  $\mathbf{p} \leftarrow \mathbf{N}\mathbf{r}$;
end

Auch dieses Verfahren hat den großen Vorteil, dass in jedem Schritt nur zwei Multiplikationen mit der Matrix $\mathbf{A}$ und sogar nur eine mit der Matrix $\mathbf{N}$ benötigt werden, so dass man es auch in Situationen anwenden kann, in denen diese Matrizen nicht explizit im Speicher vorliegen. Beispielsweise wird bei manchen Diskretisierungen die Matrix $\mathbf{A}$ nicht gespeichert, weil sich Speicherplatz sparen lässt, indem man ihre Koeffizienten nach Bedarf aus der Beschreibung der zugrundeliegenden Geometrie rekonstruiert.

In der Praxis wird $\mathbf{N}$ häufig so gewählt, dass es eine Näherung der Inversen $\mathbf{A}^{-1}$ ist, die sich effizient berechnen lässt.

Zur Motivation untersuchen wir das vorkonditionierte Richardson-Verfahren

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \theta\mathbf{N}(\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0,$$

zunächst mit $\mathbf{N} = \mathbf{A}^{-1}$, $\theta = 1$ und $\mu > 0$. Wir zerlegen $\mathbf{x}^{(m)}$ wieder in Eigenvektoranteile

$$\mathbf{x}^{(m)} = \sum_{\ell=1}^{n} \alpha_\ell \mathbf{e}^{(\ell)}$$

und berechnen

$$\begin{aligned}\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{A}^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{x}^{(m)}) &= \sum_{\ell=1}^{n}(1 - (\lambda_\ell - \mu)/\lambda_\ell)\alpha_\ell \mathbf{e}^{(\ell)} \\ = \sum_{\ell=1}^{n}(1 - (1 - \mu/\lambda_\ell))\alpha_\ell \mathbf{e}^{(\ell)} &= \sum_{\ell=1}^{n}(\mu/\lambda_\ell)\alpha_\ell \mathbf{e}^{(\ell)}.\end{aligned}$$

Für die Konvergenz des Verfahrens gibt den Ausschlag, wie sich der Faktor für $\ell = 1$ zu denen für $\ell > 1$ verhält. Da

$$\mu/\lambda_2 \geq \mu/\lambda_3 \geq \ldots \geq \mu/\lambda_n$$

gilt, genügt es, die obere Schranke

$$\varrho := \frac{\mu/\lambda_2}{\mu/\lambda_1} = \frac{\lambda_1}{\lambda_2} < 1$$

der Konvergenzrate zu untersuchen. Sie entspricht der Schranke, die wir für die inverse Iteration erhalten haben und ist insbesondere von λ_n vollständig unabhängig. Damit haben wir zwar unser Ziel einer deutlichen Verbesserung der Konvergenzrate erreicht, allerdings dafür den hohen Preis bezahlt, in jedem Schritt des Verfahrens ein Gleichungssystem *exakt* lösen zu müssen.

Wesentlich attraktiver wäre es, $\mathbf{N}$ lediglich als eine Näherung der Inversen wählen zu können, die sich effizient berechnen lässt.

Leider würden die zur Zeit bekannten Beweise für die zentralen Konvergenzaussagen den Rahmen dieser Vorlesung sprengen, deshalb wird hier nur ein Ergebnis von K. Neymeyr[1] ohne Beweis angegeben.

Die entscheidende Voraussetzung an den Vorkonditionierer hat die Form

$$(1-\gamma)\langle \mathbf{N}^{-1}\mathbf{x}, \mathbf{x}\rangle_2 \leq \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle_2 \leq (1+\gamma)\langle \mathbf{N}^{-1}\mathbf{x}, \mathbf{x}\rangle_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n \tag{9.8}$$

mit einer Konstanten $\gamma \in [0, 1)$. Gemäß Lemma 3.32 entspricht $\gamma = 0$ gerade dem Fall $\mathbf{N}^{-1} = \mathbf{A}$, den wir soeben untersucht haben. Falls $\lambda_k \leq \Lambda_A(\mathbf{x}^{(m)}) < \lambda_{k+1}$ für ein $k \in \{1, \ldots, n-1\}$ gilt, folgt entweder

$$\Lambda_A(\mathbf{x}^{(m+1)}) \leq \lambda_k$$

oder

$$\frac{\Lambda_A(\mathbf{x}^{(m+1)}) - \lambda_k}{\lambda_{k+1} - \Lambda_A(\mathbf{x}^{(m+1)})} \leq \varrho^2 \frac{\Lambda_A(\mathbf{x}^{(m)}) - \lambda_k}{\lambda_{k+1} - \Lambda_A(\mathbf{x}^{(m)})} \tag{9.9}$$

[1] K. Neymeyr, *A geometric convergence theory for the preconditioned steepest descent iteration*, Universität Rostock (2010)

mit dem durch

$$\varrho := \frac{\kappa + \gamma(2-\kappa)}{(2-\kappa)+\gamma\kappa} < 1, \qquad \kappa := \frac{\lambda_k(\lambda_n - \lambda_{k+1})}{\lambda_{k+1}(\lambda_n - \lambda_k)} < 1$$

gegebenen Konvergenzfaktor. Die Abschätzung (9.9) ist *scharf*, es gibt also einen Vektor $\mathbf{x}^{(m)}$, für den beide Seiten gleich sind.

Interessant an der Abschätzung ist ihr Verhalten für den Fall, dass λ_n sehr groß wird: Wir haben

$$\kappa = \frac{\lambda_k(\lambda_n - \lambda_{k+1})}{\lambda_{k+1}(\lambda_n - \lambda_k)} = \frac{\lambda_k}{\lambda_{k+1}} \frac{(\lambda_n - \lambda_k) - (\lambda_{k+1} - \lambda_k)}{\lambda_n - \lambda_k} = \frac{\lambda_k}{\lambda_{k+1}} \left(1 - \frac{\lambda_{k+1} - \lambda_k}{\lambda_n - \lambda_k}\right),$$

die Größe κ wird also für $\lambda_n \to \infty$ von unten gegen λ_k/λ_{k+1} konvergieren. Insbesondere lässt sie sich unabhängig von λ_n beschränken.

Für die Konvergenzrate haben wir

$$\varrho = \frac{\kappa + \gamma(2-\kappa)}{(2-\kappa)+\gamma\kappa} = \frac{(1-\gamma)\kappa + \gamma\kappa + \gamma(2-\kappa)}{(1-\gamma)(2-\kappa) + \gamma(2-\kappa) + \gamma\kappa} = \frac{(1-\gamma)\kappa + 2\gamma}{(1-\gamma)(2-\kappa) + 2\gamma},$$

so dass wir für $\gamma \to 1$ eine Rate von eins, also keine Konvergenz, und für $\gamma \to 0$ eine Rate von

$$\frac{\kappa}{2-\kappa} \leq \kappa < 1$$

erhalten.

Im Gegensatz zu den bisher behandelten Abschätzungen bezieht sich die Aussage nicht direkt auf die Konvergenz des Iterationsvektors, sondern lediglich auf die des Rayleigh-Quotienten.

9.4 Erweiterungen

Das Gradientenverfahren für lineare Gleichungssysteme

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

kann als Minimierungsverfahren für das quadratische Funktional

$$f : \mathbb{K}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto \frac{1}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle_2 - \mathrm{Re}\langle \mathbf{b}, \mathbf{x}\rangle_2,$$

interpretiert werden: Es gilt

$$f(\mathbf{x}) \text{ minimal} \qquad \Longleftrightarrow \qquad \mathbf{A}\mathbf{x} = \mathbf{b},$$

und für $\mathbb{K} = \mathbb{R}$ gilt gerade

$$\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b},$$

so dass das Residuum gerade dem negativen Gradienten entspricht. Aus der Form des Funktionals folgt, dass auch für Teilräume eine Minimalitätseigenschaft gilt: Für jeden Teilraum $\mathcal{V} \subseteq \mathbb{K}^n$ und jedes $\mathbf{x} \in \mathbb{K}^n$ gilt

$$f(\mathbf{x}) \leq f(\mathbf{x} + \mathbf{y}) \qquad \text{für alle } \mathbf{y} \in \mathcal{V}$$

genau dann, wenn

$$\langle \mathbf{b} - \mathbf{A}\mathbf{x}, \mathbf{y} \rangle_2 = 0 \qquad \text{für alle } \mathbf{y} \in \mathcal{V}$$

gilt, $\mathbf{x}$ lässt sich also durch Addition eines Vektors aus $\mathcal{V}$ genau dann nicht mehr verbessern, wenn das Residuum senkrecht auf $\mathcal{V}$ steht. Diese Beobachtung nutzt das *Verfahrens der konjugierten Gradienten*, kurz das *cg-Verfahren*: Statt zu $\mathbf{x}^{(m)}$ in Richtung des Residuums $\mathbf{r}^{(m)}$ zu verändern, um $\mathbf{x}^{(m+1)}$ zu erhalten, konstruieren wir aus dem Residuum (bzw. dem Gradienten) eine andere Richtung, die die Optimalität bezüglich der vorangehenden Richtungen erhält. Aus $\mathbf{A} = \mathbf{A}^*$ lässt sich mit etwas Aufwand schließen, dass die nächste Iterierte als Lösung des folgenden Minimierungsproblems berechnet werden kann:

Finde $\mathbf{x}^{(m+1)} \in \operatorname{span}\{\mathbf{x}^{(m-1)}, \mathbf{x}^{(m)}, \mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}\}$ mit

$$f(\mathbf{x}^{(m+1)}) \leq f(\mathbf{y}) \qquad \text{für alle } \mathbf{y} \in \operatorname{span}\{\mathbf{x}^{(m-1)}, \mathbf{x}^{(m)}, \mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}\}.$$

Dieser Ansatz lässt sich offensichtlich auf unsere Aufgabenstellung, nämlich die Minimierung des Rayleigh-Quotienten, übertragen: Wir berechnen die nächste Näherung des Eigenvektors als Lösung des folgenden Minimierungsproblems:

Finde $\mathbf{x}^{(m+1)} \in \operatorname{span}\{\mathbf{x}^{(m-1)}, \mathbf{x}^{(m)}, \mathbf{A}\mathbf{x}^{(m)} - \lambda\mathbf{x}^{(m)}\}$ mit

$$\Lambda_A(\mathbf{x}^{(m+1)}) \leq \Lambda_A(\mathbf{y}) \quad \text{für alle } \mathbf{y} \in \operatorname{span}\{\mathbf{x}^{(m-1)}, \mathbf{x}^{(m)}, \mathbf{A}\mathbf{x}^{(m)} - \lambda\mathbf{x}^{(m)}\}.$$

Da der Rayleigh-Quotient kein quadratisches Funktional ist, vererbt sich die Optimalität einer Näherung bezüglich eines Teilraums leider nicht auf im folgenden Schritt berechnete Näherungen. Trotzdem wird die Näherung im Vergleich zum Gradientenverfahren besser sein, da das Minimum in einem größeren Raum gesucht wird. Deshalb spricht man nur von einem *lokal* optimalen cg-Verfahren (LOPCG, *locally optimal preconditioned cg*).

Algorithmus 9.8 (Lokal optimales vorkonditioniertes cg-Verfahren) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine selbstadjungierte Matrix, und sei* $\mathbf{N} \in \mathbb{K}^{n\times n}$ *selbstadjungiert und positiv definit. Der folgende Algorithmus berechnet eine Folge* $(\mathbf{x}^{(m)})_{m\in\mathbb{N}_0}$*, die unter geeigneten Bedingungen gegen einen Eigenvektor konvergiert.*

$\mathbf{x} \leftarrow \mathbf{x}/\|\mathbf{x}\|_2$*;*
$\lambda \leftarrow \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle_2$*;*
$m \leftarrow 0$*;*
$\mathbf{r} \leftarrow \mathbf{A}\mathbf{x} - \lambda\mathbf{x}$*;*
$\mathbf{p} \leftarrow \mathbf{N}\mathbf{r}$*;*

```
while ‖r‖₂ zu groß do begin
  if m = 0 then
    V ← (x  p) ∈ K^{n×2}
  else
    V ← (z  x  p) ∈ K^{n×3};
  Berechne eine QR-Zerlegung QR = V;
  C ← AQ;
  B ← Q*C;        { = Q*AQ }
  Finde einen normierten Eigenvektor y für den minimalen Eigenwert λ
    der Matrix B;
  z ← x;
  x ← Qy;
  r ← Cy − λx        { = Ax − λx }
  p ← Nr;
  m ← m + 1
end
```

Ein *global* optimales Verfahren haben wir bereits kennen gelernt: Der Lanczos-Algorithmus 8.6 berechnet das Minimum jeweils im Raum *aller* bisher aufgetretenen Residuen, benötigt allerdings für die Darstellung der Arnoldi-Basis auch sehr viel mehr Speicher als das lokal optimale Verfahren.

Bemerkung 9.9 (Geschachtelte Iteration) *Im Vergleich zu den quadratisch oder kubisch konvergenten Rayleigh- und QR-Iterationen scheinen die in diesem Kapitel vorgestellten lediglich linear konvergenten Verfahren auf den ersten Blick wenig attraktiv zu sein. Auf den zweiten Blick zeigt sich allerdings, dass sie auch wesentlich besser für große Matrizen geeignet sind, da sie ohne die Berechnung der exakten Inversen (wie bei der Rayleigh-Iteration) oder vollständiger orthonormaler Basen (wie bei der QR-Iteration) auskommen.*

In der Praxis entstehen die Matrizen $\mathbf{A}$*, deren kleinsten Eigenwert wir berechnen wollen, durch die Diskretisierung einer Differentialgleichung. In diesem Fall lässt sich die relativ langsame Konvergenz ausgleichen, indem geeignete Startvektoren berechnet werden: Die Differentialgleichung wird mit unterschiedlichen Genauigkeiten diskretisiert, so dass eine Hierarchie von Matrizen entsteht. Die Matrizen niedriger Genauigkeit sind klein und lassen sich deshalb schnell bearbeiten, eventuell sogar mit einer QR-Iteration. Die so berechneten Näherungen der Eigenvektoren werden dann als Startvektoren für zunehmend genauere Diskretisierungen verwendet, bis die gewünschte Genauigkeit erreicht ist. In dieser Weise genügen für jede Diskretisierung wenige Iterationsschritte, und der Gesamtaufwand bleibt beherrschbar.*

9.5 Block-Verfahren

Bisher haben wir uns lediglich mit der Frage nach der Berechnung eines Eigenvektors für den kleinsten Eigenwert beschäftigt. Ähnlich wie im Fall der orthogonalen Iteration

(vgl. Abschnitt 5.5) können wir den einzelnen Iterationsvektor durch eine Basis solcher Vektoren ersetzen und in dieser Weise versuchen, eine Anzahl der kleinsten Eigenwerte zu ermitteln.

Wir führen wieder mehrere Iterationen simultan durch und fassen die dabei entstehenden Vektoren zu Matrizen $\mathbf{X}^{(m)} \in \mathbb{K}^{n \times k}$ zusammen. Statt zu verlangen, dass die Iterationsvektoren Einheitsvektoren sind, fordern wir wieder, dass $\mathbf{X}^{(m)}$ eine orthogonale Matrix ist. Im Fall $k = 1$ sind beide Eigenschaften äquivalent, im Fall $k > 1$ stellt die zweite Eigenschaft sicher, dass nicht alle Spalten der Matrix gegen denselben Eigenraum konvergieren.

Wir hoffen, dass die ℓ-ten Spalten der Matrizen $\mathbf{X}^{(m)}$ gegen den Eigenraum des ℓ-ten Eigenwerts konvergieren werden. Falls die Eigenwerte unterschiedlich sind, sollten wir bei der Berechnung des Residuums also unterschiedliche Eigenwerte einsetzen. Da wir in diesem Abschnitt häufig mit einzelnen Spalten bestimmter Matrizen arbeiten müssen, führen wir die Notation ein, dass $\mathbf{M}_\ell = \mathbf{M}\mathbf{e}^{(\ell)}$ gerade die ℓ-te Spalte einer Matrix $\mathbf{M}$ bezeichnet.

Die Rayleigh-Quotienten zu den Spalten der Matrix $\mathbf{X}^{(m)}$ sind dann durch

$$\lambda_\ell^{(m)} := \langle \mathbf{A}\mathbf{X}_\ell^{(m)}, \mathbf{X}_\ell^{(m)} \rangle_2 \qquad \text{für alle } \ell \in \{1, \ldots, k\}$$

gegeben, da die Spalten nach Voraussetzung normiert sind. Die korrespondierenden Residuen fassen wir zu einer Matrix $\mathbf{R}^{(m)} \in \mathbb{K}^{n \times k}$ zusammen, die durch

$$\mathbf{R}_\ell^{(m)} := \mathbf{A}\mathbf{X}_\ell^{(m)} - \lambda_\ell^{(m)} \mathbf{X}_\ell^{(m)} \qquad \text{für alle } \ell \in \{1, \ldots, k\}$$

definiert ist. Das vorkonditionierte Residuum ergibt sich dann als

$$\mathbf{P}_\ell^{(m)} := \mathbf{N}\mathbf{R}^{(m)}.$$

Wir konstruieren eine Block-Variante des vorkonditionierten Gradientenverfahrens (vgl. Algorithmus 9.7), bei der wir die verbesserten Näherungen der Eigenvektoren nicht nur in den von $\mathbf{X}_\ell^{(m)}$ und $\mathbf{P}_\ell^{(m)}$ aufgespannten zweidimensionalen Teilräumen suchen, sondern im gesamten Bild der Matrizen $\mathbf{X}^{(m)}$ und $\mathbf{P}^{(m)}$, also in einem Teilraum, der höchstens $2k$-dimensional sein kann. Da wir nur an der Berechnung weniger Eigenwerte interessiert sind, können wir davon ausgehen, dass das Eigenwertproblem in diesem Teilraum effizient lösbar ist. Damit erhalten wir den folgenden Algorithmus:

Algorithmus 9.10 (Block-Gradientenverfahren) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine selbstadjungierte Matrix, und sei $\mathbf{N} \in \mathbb{K}^{n \times n}$ selbstadjungiert und positiv definit. Der folgende Algorithmus berechnet eine Folge von Matrizen $(\mathbf{X}^{(m)})_{m \in \mathbb{N}_0}$, deren Spalten unter geeigneten Bedingungen gegen Eigenvektoren zu den kleinsten Eigenwerten konvergieren.*

for $\ell \in \{1, \ldots, k\}$ *do begin*
  $\lambda_\ell \leftarrow \langle \mathbf{A}\mathbf{X}_\ell, \mathbf{X}_\ell \rangle_2$;
  $\mathbf{R}_\ell \leftarrow \mathbf{A}\mathbf{X}_\ell - \lambda_\ell \mathbf{X}_\ell$;
end;

$\mathbf{P} \leftarrow \mathbf{NR}$;
while $\|\mathbf{R}\|_2$ *zu groß* `do begin`
  $\mathbf{V} \leftarrow \begin{pmatrix}\mathbf{X} & \mathbf{P}\end{pmatrix} \in \mathbb{K}^{n\times(2k)}$;
  Berechne eine QR-Zerlegung $\mathbf{QR} = \mathbf{V}$ *mit* $\mathbf{Q} \in \mathbb{K}^{n\times(2k)}$;
  $\mathbf{C} \leftarrow \mathbf{AQ}$;
  $\mathbf{B} \leftarrow \mathbf{Q}^*\mathbf{C}$; $\{ = \mathbf{Q}^*\mathbf{AQ} \}$
  Finde eine orthogonale Matrix $\mathbf{Y} \in \mathbb{K}^{(2k)\times k}$, *deren Spalten Eigenvektoren zu den* k *kleinsten Eigenwerten* $\lambda_1 \leq \ldots \leq \lambda_k$ *der Matrix* $\mathbf{B}$ *enthält;*
  $\mathbf{X} \leftarrow \mathbf{QY}$;
  `for` $\ell \in \{1, \ldots, k\}$ `do`
    $\mathbf{R}_\ell \leftarrow \mathbf{AX}_\ell - \lambda_\ell \mathbf{X}_\ell$;
  $\mathbf{P} \leftarrow \mathbf{NR}$;
end

Das Block-Gradientenverfahren bietet dieselben Vorteile wie die anderen bisher besprochenen Blockverfahren: Es lassen sich mehrere Eigenvektoren simultan berechnen, so dass sich beispielsweise zu mehrfachen Eigenwerten die vollständigen Eigenräume konstruieren lassen. Für das Konvergenzverhalten ist nicht mehr das Verhältnis des ersten und zweiten Eigenwerts relevant, sondern das des k-ten und des $(k+1)$-ten. Unter geeigneten Bedingungen lässt sich sogar zeigen, dass für die Konvergenz der ℓ-ten Spalte lediglich das Verhältnis des ℓ-ten und des $(k+1)$-ten Eigenwerts von Bedeutung ist, so dass das Blockverfahren auch dann von Vorteil sein kann, wenn lediglich eine Näherung des ersten Eigenwerts gesucht ist.

Bemerkung 9.11 (Residuum) *Im Vergleich mit der orthogonalen Iteration stellt sich die Frage, ob unsere Berechnung des Residuums angemessen ist: Wir berechnen*

$$\mathbf{R}^{(m)} = \mathbf{AX}^{(m)} - \mathbf{X}^{(m)}\mathbf{D}^{(m)}, \qquad \mathbf{D}^{(m)} = \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_k\end{pmatrix},$$

während wir bei der orthogonalen Iteration und der QR-Iteration statt der Diagonalmatrix $\mathbf{D}^{(m)}$ *die Matrix*

$$\mathbf{\Lambda}^{(m)} := (\mathbf{X}^{(m)})^*\mathbf{AX}^{(m)}$$

verwendet haben, die auch die Interaktion zwischen verschiedenen Spalten der Matrix $\mathbf{X}^{(m)}$ *erfasst. Mit Hilfe dieser Matrix ließe sich ein modifiziertes Residuum der Form*

$$\mathbf{R}^{(m)} = \mathbf{AX}^{(m)} - \mathbf{X}^{(m)} - \mathbf{\Lambda}^{(m)}$$

berechnen. Der Unterschied zwischen beiden Zugängen besteht darin, dass der zweite Ansatz schon dann kleine Residuen ermittelt, wenn die Spalten der Matrix $\mathbf{X}^{(m)}$ *lediglich einen invarianten Unterraum approximieren, während der ursprüngliche Ansatz nur dann zu kleinen Residuen fehlt, wenn die Spalten der Matrix Eigenvektoren sind.*

Selbstverständlich können wir wie im vorigen Abschnitt auch eine Blockvariante des lokal optimalen cg-Verfahrens konstruieren: Wir nehmen $\mathbf{X}^{(m-1)}$ hinzu und erhalten so das *lokal optimale vorkonditionierte Block-cg-Verfahren* (LOBPCG, *locally optimal block preconditioned cg*):

Algorithmus 9.12 (Block-cg-Verfahren) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix, und sei $\mathbf{N} \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit. Der folgende Algorithmus berechnet eine Folge von Matrizen $(\mathbf{X}^{(m)})_{m\in\mathbb{N}_0}$, deren Spalten unter geeigneten Bedingungen gegen Eigenvektoren zu den kleinsten Eigenwerten konvergieren.*

for $\ell \in \{1,\dots,k\}$ *do begin*
  $\lambda_\ell \leftarrow \langle \mathbf{A}\mathbf{X}_\ell, \mathbf{X}_\ell\rangle_2$;
  $\mathbf{R}_\ell \leftarrow \mathbf{A}\mathbf{X}_\ell - \lambda_\ell \mathbf{X}_\ell$;
end;
$m \leftarrow 0$;
$\mathbf{P} \leftarrow \mathbf{N}\mathbf{R}$;
while $\|\mathbf{R}\|_2$ *zu groß do begin*
  if $m = 0$ *then*
    $\mathbf{V} \leftarrow \begin{pmatrix}\mathbf{X} & \mathbf{P}\end{pmatrix} \in \mathbb{K}^{n\times(2k)}$
  else
    $\mathbf{V} \leftarrow \begin{pmatrix}\mathbf{Z} & \mathbf{X} & \mathbf{P}\end{pmatrix} \in \mathbb{K}^{n\times(3k)}$;
  Berechne eine QR-Zerlegung $\mathbf{Q}\mathbf{R} = \mathbf{V}$;
  $\mathbf{C} \leftarrow \mathbf{A}\mathbf{Q}$;
  $\mathbf{B} \leftarrow \mathbf{Q}^*\mathbf{C}$; $\{ = \mathbf{Q}^*\mathbf{A}\mathbf{Q} \}$
  Finde eine orthogonale Matrix $\mathbf{Y}$*, deren Spalten Eigenvektoren zu den* k *kleinsten Eigenwerten* $\lambda_1 \le \ldots \le \lambda_k$ *der Matrix* $\mathbf{B}$ *enthält;*
  $\mathbf{Z} \leftarrow \mathbf{X}$;
  $\mathbf{X} \leftarrow \mathbf{Q}\mathbf{Y}$;
  for $\ell \in \{1,\dots,k\}$ *do*
    $\mathbf{R}_\ell \leftarrow \mathbf{A}\mathbf{X}_\ell - \lambda_\ell \mathbf{X}_\ell$;
  $\mathbf{P} \leftarrow \mathbf{N}\mathbf{R}$;
  $m \leftarrow m + 1$
end

9.6 Eigenwert-Mehrgitterverfahren

Bei sehr großen Matrizen führt kein Weg daran vorbei, deren strukturelle Eigenschaften so gut wie möglich auszunutzen. Im Fall partieller Differentialgleichungen besteht eine solche Eigenschaft darin, dass wir sie in unterschiedlichen Auflösungen diskretisieren können, um unterschiedlich große Matrizen zu erhalten. Da diese Matrizen Näherungen desselben kontinuierlichen Problems darstellen, dürfen wir darauf hoffen, dass sie zueinander in Beziehung stehen und dass sich diese Beziehung für unsere Zwecke ausnutzen lässt.

Wir gehen davon aus, dass uns eine Familie $(\mathbf{A}_\ell)_{\ell=0}^L$ selbstadjungierter Matrizen

$$\mathbf{A}_\ell \in \mathbb{K}^{\mathcal{I}_\ell \times \mathcal{I}_\ell} \qquad \text{für alle } \ell \in [0:L]$$

gegeben ist, wobei $\mathbf{A}_{\ell+1}$ jeweils zu einer feineren Diskretisierung als $\mathbf{A}_\ell$ gehören soll.

Unsere Aufgabe besteht darin, einen der kleineren Eigenwerte der Matrix $\mathbf{A} = \mathbf{A}_L$ zu berechnen, die dann zu der am höchsten auflösenden Diskretisierung gehört.

Die Beziehung zwischen den einzelnen Diskretisierungen wird durch *Prolongationsmatrizen*

$$\mathbf{P}_\ell \in \mathbb{K}^{\mathcal{I}_\ell \times \mathcal{I}_{\ell-1}} \qquad \text{für alle } \ell \in [1:L]$$

dargestellt, die die Einbettung einer auf dem zu der Stufe $\ell - 1$ gehörenden Gitter gegebenen Funktion auf der Stufe ℓ beschreiben, beispielsweise durch Interpolation.

Häufig lassen sich diese Matrizen so wählen, dass die *Galerkin-Eigenschaft*

$$\mathbf{A}_{\ell-1} = \mathbf{P}_\ell^* \mathbf{A}_\ell \mathbf{P}_\ell \qquad \text{für alle } \ell \in [1:L] \tag{9.10}$$

gilt, dass also die Matrizen auf gröberen Gittern perfekt durch die auf feineren dargestellt werden können.

Seien nun $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ mit $n := \#\mathcal{I}_L$ die Eigenwerte der Matrix $\mathbf{A}$, und sei $(\mathbf{e}_i)_{i=1}^n$ eine orthonormale Basis zugehöriger Eigenvektoren. Wir untersuchen die einfache Richardson-Iteration

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \theta(\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0$$

für Linearkombinationen

$$\mathbf{x}^{(m)} = \sum_{i=1}^n \alpha_i \mathbf{e}_i, \qquad \alpha \in \mathbb{K}^n$$

von Eigenvektoren. Es gilt

$$\begin{aligned}
\mathbf{x}^{(m+1)} &= \sum_{i=1}^n \alpha_i \Big(\mathbf{e}_i - \theta(\mathbf{A}\mathbf{e}_i - \mu\mathbf{e}_i)\Big) \\
&= \sum_{i=1}^n \alpha_i \Big(1 - \theta(\lambda_i \mathbf{e}_i - \mu\mathbf{e}_i)\Big) \\
&= \sum_{i=1}^n \alpha_i \big(1 - \theta(\lambda_i - \mu)\big)\mathbf{e}_i.
\end{aligned}$$

Wir wissen bereits, dass für $\lambda_i < \mu$ der Anteil des Eigenvektors $\mathbf{e}_i$ zunehmen wird. Das stört uns nicht, da wir ja kleine Eigenwerte approximieren wollen.

Unser Ziel besteht darin, dafür zu sorgen, dass die besonders großen Eigenwerte möglichst stark reduziert werden. Dazu wählen wir den Dämpfungsparameter θ anders als zuvor, indem wir

$$\theta := \frac{1}{\lambda_n - \lambda_1}$$

setzen. Für $i \in [1:n]$ mit $\lambda_i - \mu \geq (\lambda_n - \lambda_1)/2$ folgt

$$1 - \theta(\lambda_i - \mu) \leq 1 - \theta(\lambda_n - \lambda_1)/2 = 1 - 1/2 = 1/2,$$

Anteile des Vektors $\mathbf{x}^{(m)}$, die zu großen Eigenwerten gehören, werden also mindestens um den Faktor $1/2$ reduziert.

Bei Matrizen, die aus der Diskretisierung einer partiellen Differentialgleichung entstehen, gehören große Eigenwerte zu stark oszillierenden Eigenvektoren, während kleine Eigenwerte zu glatten Eigenvektoren gehören. Also reduziert die Richardson-Iteration die Oszillationen und sorgt so dafür, dass der Vektor $\mathbf{x}^{(m+1)}$ „glatter“ als der Vektor $\mathbf{x}^{(m)}$ sein wird.

Iterationsverfahren mit dieser Eigenschaft werden deshalb als *Glättungsverfahren* bezeichnet.

Wir sind aber nicht an einem glatten Vektor interessiert, sondern an einem Eigenvektor. Theoretisch könnten wir einen Schritt der inversen Iteration mit Shift durchführen, also

$$\mathbf{x}^{(m+1)} = (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{x}^{(m)},$$

um dem Eigenvektor näher zu kommen, allerdings müssten wir dafür ein Gleichungssystem mit der Matrix $\mathbf{A} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{I}$ lösen, die dafür aber viel zu groß sein dürfte.

Stattdessen nutzen wir aus, dass nach einigen Schritten der Richardson-Iteration der Vektor $\mathbf{x}^{(m)}$ so glatt sein wird, dass wir ihn mit der gröberen Diskretisierung der Stufe $L-1$ gut approximieren können.

Da wir dabei die exakte Matrix auf der Stufe L durch eine Näherung auf der Stufe $L-1$ ersetzen werden, empfiehlt es sich, wie bei der vorkonditionierte inversen Iteration dafür zu sorgen, dass Eigenvektoren Fixpunkte sind, wir verwenden also die Umformulierung

$$\begin{aligned}
\mathbf{x}^{(m+1)} &= (\Lambda_A(\mathbf{x}^{(m)}) - \mu)(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{x}^{(m)} \\
&= \mathbf{x}^{(m)} - (\mathbf{A} - \mu\mathbf{I})^{-1}\Big((\mathbf{A} - \mu\mathbf{I})\mathbf{x}^{(m)} - (\Lambda_A(\mathbf{x}^{(m)}) - \mu)\mathbf{x}^{(m)}\Big) \\
&= \mathbf{x}^{(m)} - (\mathbf{A} - \mu\mathbf{I})^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}).
\end{aligned}$$

Um einen Schritt dieses Verfahrens durchzuführen, müssten wir das Gleichungssystem

$$(\mathbf{A} - \mu\mathbf{I})\mathbf{y} = \mathbf{A}\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}$$

lösen und dann $\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{y}$ setzen. Nach einigen Schritten des Glättungsverfahrens dürfen wir darauf hoffen, dass wir $\mathbf{y}$ auf dem gröberen Gitter der Stufe $L-1$ approximieren können, dass also ein Vektor $\mathbf{y}_{L-1}$ mit

$$\begin{aligned}
\mathbf{y} &\approx \mathbf{P}_L\mathbf{y}_{L-1}, \\
(\mathbf{A}_L - \mu\mathbf{I})\mathbf{y} &\approx (\mathbf{A}_L - \mu\mathbf{I})\mathbf{P}_L\mathbf{y}_{L-1}
\end{aligned}$$

existiert. Um aus dieser Approximationseigenschaft ein Gleichungssystem der Stufe $L-1$ zu machen, das sich praktisch lösen lässt, multiplizieren wir mit $\mathbf{P}_L^*$ und erhalten

$$\mathbf{P}_L^*(\mathbf{A}_L - \mu\mathbf{I})\mathbf{P}_L\mathbf{y}_{L-1} = \mathbf{P}_L^*(\mathbf{A}_L - \mu\mathbf{I})\mathbf{y} = \mathbf{P}_L^*(\mathbf{A}_L\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}).$$

Falls wir die Galerkin-Eigenschaft (9.10) voraussetzen und annehmen, dass $\mathbf{P}_L^*$ näherungsweise eine Linksinverse der Prolongation $\mathbf{P}_L$ ist, also $\mathbf{P}_L^*\mathbf{P}_L \approx \mathbf{I}$ gilt, gelangen wir zu

$$(\mathbf{A}_{L-1} - \mu\mathbf{I})\mathbf{y}_{L-1} = \mathbf{P}_L^*(\mathbf{A}_L\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}),$$

wir müssen also ein Gleichungssystem auf der Stufe $L-1$ lösen, um $\mathbf{y}_{L-1}$ zu berechnen.

Insgesamt erhalten wir die *Grobgitterkorrektur*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \mathbf{P}_L(\mathbf{A}_{L-1} - \mu\mathbf{I})^{-1}\mathbf{P}_L^*(\mathbf{A}_L\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}).$$

Falls $\mathbf{x}^{(m)}$ sich hinreichend gut auf dem gröberen Gitter darstellen lässt, dürfen wir darauf hoffen, dass sie ein ähnliches Verhalten wie die inverse Iteration mit Shift aufweist, also mit einer von der Gitterstufe unabhängigen Geschwindigkeit konvergiert.

Allerdings wird in vielen Anwendungsfällen die Matrix $\mathbf{A}_{L-1}$ auf der Gitterstufe $L-1$ immer noch viel zu groß sein, um direkt das Gleichungssystem

$$(\mathbf{A}_{L-1} - \mu\mathbf{I})\mathbf{y}_{L-1} = \mathbf{b}_{L-1}$$

zu lösen. Wir hoffen natürlich, dass μ in der Nähe eines Eigenwerts der Matrix $\mathbf{A}_L$ liegt, und da unser Grobgittersystem diese Matrix approximieren soll, müssen wir damit rechnen, dass μ auch in der Nähe eines Eigenwerts der Matrix $\mathbf{A}_{L-1}$ liegt.

Eine elegante Lösung besteht darin, dieses Problem nicht zu ignorieren, sondern auszunutzen: Wir ersetzen μ durch den nächstgelegenen *exakten* Eigenwert μ_{L-1} der Matrix $\mathbf{A}_{L-1}$, und nehmen an, dass uns der zugehörige Eigenraum bekannt ist. Mit $\mathbf{E}_{L-1} \in \mathbb{K}^{\mathcal{I}_{L-1}\times\mathcal{I}_{L-1}}$ bezeichnen wir die orthogonale Projektion auf diesen Eigenraum, so dass die Matrix $\mathbf{A}_{L-1} - \mu_{L-1}\mathbf{I}$ auf dem orthogonalen Komplement $\mathcal{W}_{L-1} = \operatorname{Bild}(\mathbf{I} - \mathbf{E}_{L-1})$ des Eigenraums invertierbar ist. Das Grobgittersystem ersetzen wir durch

$$(\mathbf{I} - \mathbf{E}_{L-1})(\mathbf{A}_{L-1} - \mu\mathbf{I})\mathbf{y}_{L-1} = (\mathbf{I} - \mathbf{E}_{L-1})\mathbf{b}_{L-1},$$

und halten fest, dass dieses System in $\mathcal{W}_{L-1}$ eindeutig lösbar ist.

Diese Lösung approximieren wir wieder mit einem Iterationsverfahren: Ausgehend von dem Startvektor $\mathbf{y}_{L-1}^{(0)}$ führen wir einige Schritte der Richardson-Iteration

$$\mathbf{y}_{L-1}^{(m+1)} = \mathbf{y}^{(m)} - \theta(\mathbf{A}_{L-1}\mathbf{y}^{(m)} - \mu\mathbf{y}^{(m)} - \mathbf{b}_{L-1}) \qquad \text{für alle } m \in \mathbb{N}_0$$

durch, um den Fehler $\mathbf{y}_{L-1} - \mathbf{y}_{L-1}^{(m)}$ zu glätten. Anschließend wenden wir die Projektion $\mathbf{I} - \mathbf{E}_{L-1}$ an, um den Fehler in den Raum $\mathcal{W}_{L-1}$ zu befördern.

Den glatten Fehler können wir auf dem nächstgröberen Gitter, also auf der Gitterstufe $L-2$, approximieren, indem wir dort die Gleichung

$$(\mathbf{A}_{L-2} - \mu_{L-1}\mathbf{I})\mathbf{y}_{L-2} = \mathbf{P}_{L-1}^*(\mathbf{A}_{L-1}\mathbf{y}^{(m)} - \mu\mathbf{y}^{(m)} - \mathbf{b}_{L-1})$$

lösen und anschließend die Korrektur $\mathbf{y}_{L-2}$ von $\mathbf{y}_{L-1}^{(m)}$ subtrahieren. Sicherheitshalber können wir an dieser Stelle auch noch einmal die Projektion $\mathbf{I} - \mathbf{E}_{L-1}$ anwenden.

Falls die Matrix $\mathbf{A}_{L-2}$ immer noch zu groß ist, können wir rekursiv fortfahren, bis wir auf dem gröbsten Gitter angekommen sind.

Ein Problem bleibt allerdings: Woher kennen wir die exakten Eigenwerte und Eigenvektoren auf den gröberen Gittern? Eine elegante Lösung bietet wieder die geschachtelte Iteration: Zunächst berechnen wir einen Eigenwert und den zugehörigen Eigenraum auf dem gröbsten Gitter. Dann verwenden wir die Prolongation $\mathbf{P}_1$, um die Basis dieses Eigenraums auf die nächstfeinere Gitterstufe zu transportieren. Sie dient uns als Startwert für die Iteration auf dieser Stufe, die wir durchführen können, weil uns auf Stufe 0 alles Nötige bekannt ist. Sobald die Eigenwerte und Eigenvektoren auf Stufe 1 hinreichend genau bestimmt sind, können wir die Prozedur wiederholen, um Näherungen auf Stufe 2 zu berechnen. In dieser Weise fahren wir fort, bis wir auf der gewünschten Stufe L angekommen sind.

10 Verwandte Fragestellungen

Bisher haben wir uns ausschließlich mit der Frage nach der Berechnung einzelner, mehrerer oder aller Eigenwerte und eventuell der zugehörigen Eigenvektoren befasst. In diesem Kapitel beschäftigen wir uns mit Problemen, die eng mit Eigenwertproblemen verwandt sind, aber modifizierte Lösungsverfahren erfordern.

10.1 Verallgemeinerte Eigenwertprobleme

Für zwei Matrizen $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{n\times n}$ können wir uns die Frage stellen, ob Zahlen $\lambda \in \mathbb{K}$ und Vektoren $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ existieren, die die Gleichung

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \tag{10.1}$$

erfüllen. Diese Aufgabe bezeichnet man als *verallgemeinertes Eigenwertproblem*. Falls $\mathbf{B}$ invertierbar ist, können wir durch Multiplikation mit $\mathbf{B}^{-1}$ zu dem gewöhnlichen Eigenwertproblem

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{10.2}$$

übergehen und die bereits diskutierten Verfahren einsetzen.

Von Interesse sind deshalb zwei Typen verallgemeinerter Eigenwertprobleme: Einerseits diejenigen, bei denen $\mathbf{B}$ nicht invertierbar ist, und andererseits diejenigen, bei denen $\mathbf{A}$ selbstadjungiert ist und diese nützliche Eigenschaft erhalten bleiben sollte. Diesem zweiten Fall widmen wir uns in einem separaten Abschnitt.

Wir untersuchen zunächst den Fall, dass $\mathbf{B}$ nicht invertierbar oder sehr schlecht konditioniert ist. In diesem Fall ist man daran interessiert, eine *verallgemeinerte Schur-Zerlegung* zu berechnen, nämlich unitäre Matrizen $\mathbf{Q}, \mathbf{P} \in \mathbb{K}^{n\times n}$, für die die Matrizen

$$\widehat{\mathbf{A}} := \mathbf{Q}\mathbf{A}\mathbf{P}^*, \qquad \widehat{\mathbf{B}} := \mathbf{Q}\mathbf{B}\mathbf{P}^*$$

obere Dreiecksmatrizen sind. An den Diagonalelementen der beiden transformierten Matrizen lassen sich dann die verallgemeinerten Eigenwerte direkt ablesen, durch Rückwärtseinsetzen in $\widehat{\mathbf{A}} - \lambda\widehat{\mathbf{B}}$ können wir auch Eigenvektoren bestimmen.

Bei der Berechnung der verallgemeinerten Schur-Zerlegung können wir uns an der Vorgehensweise für die gewöhnliche Schur-Zerlegung orientieren: Mit Hilfe geeigneter unitärer Matrizen $\mathbf{Q}$ und $\mathbf{P}$ können wir das Eigenwertproblem auf die Form

$$\widehat{\mathbf{A}}\mathbf{x} = \lambda\widehat{\mathbf{B}}\mathbf{x}, \qquad \widehat{\mathbf{A}} = \mathbf{Q}\mathbf{A}\mathbf{P}^*, \qquad \widehat{\mathbf{B}} = \mathbf{Q}\mathbf{B}\mathbf{P}^*$$

bringen, bei der $\widehat{\mathbf{B}}$ bereits eine rechte obere Dreiecksmatrix ist, $\widehat{\mathbf{A}}$ allerdings nur eine Hessenberg-Matrix.

Implizite Iteration für reguläres B

Die Dreiecksstruktur der Matrix $\widehat{\mathbf{B}}$ lässt sich ausnutzen, um ähnlich wie im Fall der Berechnung der Singulärwertzerlegung vorzugehen: Falls $\widehat{\mathbf{B}}$ regulär ist, könnte man *implizit* das Gegenstück der Formulierung (10.2) behandeln, also die Schur-Zerlegung der Matrix $\mathbf{H}^{(0)} = \widehat{\mathbf{B}}^{-1}\widehat{\mathbf{A}}$ berechnen, indem man die in der QR-Iteration auftretenden Transformationen $\mathbf{Q}^{(0)}, \mathbf{Q}^{(1)}, \ldots$ anwendet:

$$\mathbf{H}^{(0)} := \widehat{\mathbf{B}}^{-1}\widehat{\mathbf{A}}, \qquad \mathbf{H}^{(m+1)} := (\mathbf{Q}^{(m)})^* \mathbf{H}^{(m)} \mathbf{Q}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Da $\widehat{\mathbf{B}}$ schlecht konditioniert sein kann, wollen wir die Multiplikation mit $\widehat{\mathbf{B}}^{-1}$ vermeiden. Wie bei der Singulärwertzerlegung definieren wir

$$\begin{aligned} \widehat{\mathbf{A}}^{(0)} &:= \widehat{\mathbf{A}}, & \widehat{\mathbf{A}}^{(m+1)} &:= (\mathbf{P}^{(m)})^* \widehat{\mathbf{A}}^{(m)} \mathbf{Q}^{(m)}, \\ \widehat{\mathbf{B}}^{(0)} &:= \widehat{\mathbf{B}}, & \widehat{\mathbf{B}}^{(m+1)} &:= (\mathbf{P}^{(m)})^* \widehat{\mathbf{B}}^{(m)} \mathbf{Q}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0 \end{aligned}$$

mit später noch genauer zu spezifizierenden unitären Matrizen $\mathbf{P}^{(m)}$ und erhalten wegen

$$\mathbf{H}^{(0)} = (\widehat{\mathbf{B}}^{(0)})^{-1} \widehat{\mathbf{A}}^{(0)}$$

und

$$\begin{aligned} \mathbf{H}^{(m+1)} &= (\mathbf{Q}^{(m)})^* \mathbf{H}^{(m)} \mathbf{Q}^{(m)} = (\mathbf{Q}^{(m)})^* (\widehat{\mathbf{B}}^{(m)})^{-1} \widehat{\mathbf{A}}^{(m)} \mathbf{Q}^{(m)} \\ &= (\mathbf{Q}^{(m)})^* (\widehat{\mathbf{B}}^{(m)})^{-1} \mathbf{P}^{(m)} (\mathbf{P}^{(m)})^* \widehat{\mathbf{A}}^{(m)} \mathbf{Q}^{(m)} \\ &= ((\mathbf{P}^{(m)})^* \widehat{\mathbf{B}} \mathbf{Q}^{(m)})^{-1} ((\mathbf{P}^{(m)})^* \widehat{\mathbf{A}} \mathbf{Q}^{(m)}) = (\widehat{\mathbf{B}}^{(m+1)})^{-1} \widehat{\mathbf{A}}^{(m+1)} \qquad \text{für alle } m \in \mathbb{N}_0 \end{aligned}$$

mit einer einfachen Induktion die faktorisierte Darstellung

$$\mathbf{H}^{(m)} = (\widehat{\mathbf{B}}^{(m)})^{-1} \widehat{\mathbf{A}}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0. \tag{10.3}$$

Wir können also die Matrizen $\mathbf{H}^{(m)}$ während der gesamten Iteration in Produktform darstellen und müssen insbesondere niemals die vollständige Inverse der Matrizen $\widehat{\mathbf{B}}^{(m)}$ berechnen. Allerdings müssen wir darauf achten, dass während der Iteration weder die Dreiecksform der Matrizen $\widehat{\mathbf{B}}^{(m)}$ noch die Hessenberg-Form der Matrizen $\widehat{\mathbf{A}}^{(m)}$ verloren geht. Dazu greifen wir wieder auf Satz 6.9 zurück: Wir führen die erste Givens-Rotation der QR-Zerlegung der Matrix $\mathbf{H}^{(m)} - \mu\mathbf{I}$ mit einem geeigneten Shift-Parameter μ durch und sorgen anschließend mit weiteren Givens-Rotationen dafür, dass die Matrizen wieder die vorgesehene Form haben. Mit Satz 6.9 folgt dann, dass sich das Resultat allenfalls im Vorzeichen von dem der ursprünglichen QR-Iteration unterscheidet. Wir stellen die ursprünglichen Matrizen in der Form

$$\widehat{\mathbf{A}}^{(m)} = \begin{pmatrix} a_{11} & a_{12} & \ldots & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ & a_{32} & a_{33} & \ldots & a_{3n} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \widehat{\mathbf{B}}^{(m)} = \begin{pmatrix} b_{11} & b_{12} & \ldots & \ldots & b_{1n} \\ & b_{22} & b_{23} & \ldots & b_{2n} \\ & & b_{33} & \ldots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & b_{nn} \end{pmatrix}$$

dar. Die erste Givens-Rotation bei der Berechnung der QR-Zerlegung von $\mathbf{H}^{(m)}$ wirkt auf die ersten beiden Spalten der Matrizen, die die Form

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & \dots & a_{1n} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23} & \dots & a_{2n} \\ \gamma_{\mathbf{1}} & a_{32}^{(1)} & a_{33} & \dots & a_{3n} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \begin{pmatrix} b_{11}^{(1)} & b_{12}^{(1)} & \dots & \dots & b_{1n} \\ \delta_{\mathbf{1}} & b_{22}^{(1)} & b_{23} & \dots & b_{2n} \\ & & b_{33} & \dots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & b_{nn} \end{pmatrix}$$

annehmen. Indem wir eine passende Givens-Rotation auf die beiden ersten Zeilen anwenden, können wir δ_1 eliminieren:

$$\begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & \dots & a_{1n}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} \\ \gamma_{\mathbf{1}} & a_{32}^{(1)} & a_{33} & \dots & a_{3n} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \begin{pmatrix} b_{11}^{(2)} & b_{12}^{(2)} & \dots & \dots & b_{1n}^{(2)} \\ & b_{22}^{(2)} & b_{23}^{(2)} & \dots & b_{2n}^{(2)} \\ & & b_{33} & \dots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & b_{nn} \end{pmatrix}.$$

Den Eintrag γ_1 eliminieren wir mit einer Givens-Rotation, die auf die zweite und dritte Zeile wirkt, so dass sich

$$\begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & \dots & a_{1n}^{(2)} \\ a_{21}^{(3)} & a_{22}^{(3)} & a_{23}^{(3)} & \dots & a_{2n}^{(3)} \\ & a_{32}^{(3)} & a_{33}^{(3)} & \dots & a_{3n}^{(3)} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \begin{pmatrix} b_{11}^{(2)} & b_{12}^{(2)} & \dots & \dots & b_{1n}^{(2)} \\ & b_{22}^{(3)} & b_{23}^{(3)} & \dots & b_{2n}^{(3)} \\ & \delta_{\mathbf{2}} & b_{33}^{(3)} & \dots & b_{3n}^{(3)} \\ & & & \ddots & \vdots \\ & & & & b_{nn} \end{pmatrix}$$

ergibt. Wir beseitigen δ_2 mit einer Givens-Rotation, die auf die zweite und dritte Spalte wirkt und zu einem Eintrag γ_2 in der zweiten Spalte der vierten Zeile der linken Matrix führt. Diesen Eintrag eliminieren wir mit einer Givens-Rotation, die auf die dritte und vierte Zeile wirkt und einen Eintrag δ_3 in die dritten Spalte der vierten Zeile der rechten Matrix erzeugt. Indem wir entsprechend fortfahren, können wir die störenden Einträge wieder „nach rechte unten aus den Matrizen heraus schieben“ und so die ursprüngliche Struktur der Matrizen wiederherstellen. Damit sind $\widehat{\mathbf{A}}^{(m+1)}$ und $\widehat{\mathbf{B}}^{(m+1)}$ gefunden. Da auf die erste Spalte der Matrizen lediglich die Givens-Rotation des ersten Schritts der QR-Zerlegung wirkt, lässt sich Satz 6.9 anwenden und folgern, dass die so berechneten Matrizen denen entsprechen, die im Rahmen der konventionellen QR-Iteration entstehen würden. Demzufolge ist zu erwarten, dass die Matrizen $\mathbf{H}^{(m)}$ gegen obere Dreiecksform konvergieren werden, dass also die unteren Nebendiagonaleinträge gegen Null streben werden. Falls $h_{k+1,k}^{(m)} = 0$ gilt, folgt aus (10.3)

$$0 = h_{k+1,k}^{(m)} = a_{k+1,k}^{(m)} / b_{k+1,k+1}^{(m)},$$

also muss $a_{k+1,k}^{(m)} = 0$ gelten, wir hätten in diesem Fall also auch einen Nebendiagonaleintrag der Matrix $\widehat{\mathbf{A}}^{(m)}$ eliminiert und damit einen Schritt in Richtung der gewünschten Dreiecksmatrix vollzogen.

Deflation

Falls $\hat{a}_{k+1,k}^{(m)} = 0$ für ein $k \in \{1, \dots, n-1\}$ gelten sollte, können wir wie üblich eine Deflation durchführen und die Iteration mit kleineren Teilmatrizen fortführen.

Die Deflation kann uns aber auch dabei helfen, den Fall einer nicht invertierbaren Matrix $\mathbf{B}$ zu behandeln: Falls $\mathbf{B}$ nicht invertierbar ist, kann die Dreiecksmatrix $\widehat{\mathbf{B}}$ ebenfalls nicht invertierbar sein. Bei einer Dreiecksmatrix ist das genau dann der Fall, wenn ein Diagonaleintrag gleich null ist, wir können diese Situation also sehr einfach erkennen.

Falls beispielsweise $\hat{b}_{11}^{(m)} = 0$ gilt, sind wir in der Situation

$$\widehat{\mathbf{A}}^{(m)} = \begin{pmatrix} a_{11} & a_{12} & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ & a_{32} & a_{33} & \dots & a_{3n} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \widehat{\mathbf{B}}^{(m)} = \begin{pmatrix} \mathbf{0} & b_{12} & \dots & \dots & b_{1n} \\ & b_{22} & b_{23} & \dots & b_{2n} \\ & & b_{33} & \dots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & b_{nn} \end{pmatrix}$$

und können eine Givens-Rotation auf die ersten beiden Zeilen anwenden, um den Eintrag a_{21} zu eliminieren. Wir erhalten

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & \dots & a_{1n} \\ & a_{22}^{(1)} & a_{23} & \dots & a_{2n} \\ & a_{32}^{(1)} & a_{33} & \dots & a_{3n} \\ & & \ddots & \ddots & \vdots \\ & & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \begin{pmatrix} \mathbf{0} & b_{12}^{(1)} & \dots & \dots & b_{1n} \\ & b_{22}^{(1)} & b_{23} & \dots & b_{2n} \\ & & b_{33} & \dots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & b_{nn} \end{pmatrix}$$

Nun ist ein Nebendiagonalelement in der linken Matrix gleich null, so dass wir per Deflation zu den Teilmatrizen

$$\begin{pmatrix} a_{22}^{(1)} & a_{23} & \dots & a_{2n} \\ a_{32}^{(1)} & a_{33} & \dots & a_{3n} \\ & \ddots & \ddots & \vdots \\ & & a_{n,n-1} & a_{nn} \end{pmatrix}, \qquad \begin{pmatrix} b_{22}^{(1)} & b_{23} & \dots & b_{2n} \\ & b_{33} & \dots & b_{3n} \\ & & \ddots & \vdots \\ & & & b_{nn} \end{pmatrix}$$

übergehen können. Nicht invertierbare Matrizen können wir also nicht nur einfach bei unseren Berechnungen berücksichtigen, sie führen sogar dazu, dass wir besonders früh mit einer Deflation die Problemgröße reduzieren können.

Der aus den impliziten QR-Schritten und der Deflation entstehende Algorithmus ist unter dem Namen *QZ-Iteration* bekannt. Da ausschließlich unitäre Transformationen zum Einsatz kommen, ist er in der Praxis relativ unanfällig für Rundungsfehler und somit auch für eher schlecht konditionierte Eigenwertprobleme anwendbar.

10.2 Selbstadjungierte positiv definite verallgemeinerte Eigenwertprobleme

Wir wenden uns einem wichtigen Spezialfall unter den verallgemeinerten Eigenwertproblemen zu: Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert, und sei $\mathbf{B} \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit, es gelte also

$$0 < \langle \mathbf{Bx}, \mathbf{x}\rangle_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}.$$

Wir suchen nach Lösungen λ, $\mathbf{x} \neq \mathbf{0}$ des verallgemeinerten Eigenwertproblems

$$\mathbf{Ax} = \lambda \mathbf{Bx}.$$

Unser Ziel besteht darin, dieses verallgemeinerte Eigenwertproblem auf ein gewöhnliches Eigenwertproblem zurückzuführen, ohne dabei die Selbstadjungiertheit der Matrix $\mathbf{A}$ zu verlieren. Das Hilfsmittel der Wahl ist die *Cholesky-Zerlegung*

$$\mathbf{B} = \mathbf{LL}^*$$

der Matrix $\mathbf{B}$, die für selbstadjungiert positive Matrizen immer existiert und sich in der Praxis auch häufig stabil berechnen lässt. Wir erhalten

$$\begin{aligned} \mathbf{Ax} &= \lambda \mathbf{Bx}, \\ \mathbf{Ax} &= \lambda \mathbf{LL}^*\mathbf{x}, \\ \mathbf{LL}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{L}^*\mathbf{x} &= \lambda \mathbf{LL}^*\mathbf{x}, \end{aligned}$$

und durch Multiplikation mit $\mathbf{L}^{-1}$ von links folgt

$$\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}(\mathbf{L}^*\mathbf{x}) = \lambda \mathbf{L}^*\mathbf{x}.$$

Wir führen die Hilfsgrößen

$$\widehat{\mathbf{x}} := \mathbf{L}^*\mathbf{x}, \qquad \widehat{\mathbf{A}} := \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}$$

ein und erhalten das gewöhnliche Eigenwertproblem

$$\widehat{\mathbf{A}}\widehat{\mathbf{x}} = \lambda \widehat{\mathbf{x}}. \tag{10.4}$$

Offenbar ist die Matrix $\widehat{\mathbf{A}}$ selbstadjungiert, es ist uns also gelungen, diese wichtige Eigenschaft zu erhalten.

Damit ist das Problem (10.4) unseren sämtlichen bisher untersuchten Verfahren zugänglich. Beispielsweise folgt aus Folgerung 3.47, dass eine unitäre Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{n\times n}$ und eine reelle Diagonalmatrix $\widehat{\mathbf{D}} \in \mathbb{R}^{n\times n}$ mit

$$\widehat{\mathbf{Q}}\widehat{\mathbf{D}}\widehat{\mathbf{Q}}^* = \widehat{\mathbf{A}}$$

existieren. Indem wir die Transformation rückgängig machen, erhalten wir

$$\mathbf{Q} := (\mathbf{L}^*)^{-1}\widehat{\mathbf{Q}}$$

und stellen fest, dass

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \widehat{\mathbf{Q}}^*\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\widehat{\mathbf{Q}} = \widehat{\mathbf{Q}}^*\widehat{\mathbf{A}}\widehat{\mathbf{Q}} = \mathbf{D},$$
$$\mathbf{Q}^*\mathbf{B}\mathbf{Q} = \widehat{\mathbf{Q}}^*\mathbf{L}^{-1}\mathbf{L}\mathbf{L}^*(\mathbf{L}^*)^{-1}\widehat{\mathbf{Q}} = \widehat{\mathbf{Q}}^*\widehat{\mathbf{Q}} = \mathbf{I}$$

gelten, dass also die Matrix $\mathbf{Q}$ sowohl $\mathbf{A}$ als auch $\mathbf{B}$ auf Diagonalform transformiert.

Die Eigenwerte $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ des transformierten Problems (und damit auch der verallgemeinerten Eigenwertproblems) lassen sich mit Hilfe des Rayleigh-Quotienten charakterisieren, der sich in der Form

$$\Lambda_{\widehat{A}}(\widehat{\mathbf{x}}) = \frac{\langle\widehat{\mathbf{A}}\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2}{\langle\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2} = \frac{\langle\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{L}^*\mathbf{x},\mathbf{L}^*\mathbf{x}\rangle_2}{\langle\mathbf{L}^*\mathbf{x},\mathbf{L}^*\mathbf{x}\rangle_2} = \frac{\langle\mathbf{A}\mathbf{x},(\mathbf{L}^*)^{-1}\mathbf{L}^*\mathbf{x}\rangle_2}{\langle\mathbf{L}\mathbf{L}^*\mathbf{x},\mathbf{x}\rangle_2} = \frac{\langle\mathbf{A}\mathbf{x},\mathbf{x}\rangle_2}{\langle\mathbf{B}\mathbf{x},\mathbf{x}\rangle_2}$$

darstellen lässt. Diese Eigenschaft ist vor allem für große verallgemeinerte Eigenwertproblem wichtig, da es mit ihrer Hilfe häufig möglich ist, auf die zeitaufwendige Berechnung der transformierten Matrix $\widehat{\mathbf{A}}$ zu verzichten und direkt mit $\mathbf{A}$ und $\mathbf{B}$ zu arbeiten.

Mit kleinen Modifikationen lassen sich auch andere Verfahren durchführen, ohne explizit mit $\widehat{\mathbf{A}}$ arbeiten zu müssen. Ein Beispiel ist der Lanczos-Algorithmus 8.6, der sich elegant formulieren lässt, wenn man das euklidische Skalarprodukt durch das zu der Matrix $\mathbf{B}$ gehörende *Energie-Skalarprodukt* ersetzt, das durch

$$\langle\mathbf{x},\mathbf{y}\rangle_B := \langle\mathbf{B}\mathbf{x},\mathbf{y}\rangle_2 \qquad \text{für alle } \mathbf{x},\mathbf{y} \in \mathbb{K}^n$$

definiert ist. Es induziert die *Energienorm*

$$\|\mathbf{x}\|_B := \sqrt{\langle\mathbf{x},\mathbf{x}\rangle_B} \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n.$$

Wir bezeichnen mit

$$\widehat{\mathbf{p}}^{(m)} := \mathbf{L}^*\mathbf{p}^{(m)}, \qquad \widehat{\mathbf{q}}^{(m)} := \mathbf{L}^*\mathbf{q}^{(m)} \qquad \text{für alle } m \in \mathbb{N}$$

die transformierten Vektoren und untersuchen die im Rahmen des Algorithmus 8.6 auftretenden Gleichungen. Die Berechnung von

$$\gamma = \|\widehat{\mathbf{p}}^{(m)}\|_2 = \sqrt{\langle\widehat{\mathbf{p}}^{(m)},\widehat{\mathbf{p}}^{(m)}\rangle_2} = \sqrt{\langle\mathbf{L}^*\mathbf{p}^{(m)},\mathbf{L}^*\mathbf{p}^{(m)}\rangle_2}$$
$$= \sqrt{\langle\mathbf{L}\mathbf{L}^*\mathbf{p}^{(m)},\mathbf{p}^{(m)}\rangle_2} = \sqrt{\langle\mathbf{B}\mathbf{p}^{(m)},\mathbf{p}^{(m)}\rangle_2} = \|\mathbf{p}^{(m)}\|_B$$

lässt sich auf die Energienorm zurückführen, und ebenso die von

$$\beta_m = \langle\widehat{\mathbf{q}}^{(m)},\widehat{\mathbf{p}}^{(m)}\rangle_2 = \langle\mathbf{L}^*\mathbf{q}^{(m)},\mathbf{L}^*\mathbf{p}^{(m)}\rangle_2 = \langle\mathbf{L}\mathbf{L}^*\mathbf{q}^{(m)},\mathbf{p}^{(m)}\rangle_2 = \langle\mathbf{q}^{(m)},\mathbf{p}^{(m)}\rangle_B$$

auf das Energie-Skalarprodukt. Für die Berechnung von $\widehat{\mathbf{p}}^{(m)}$ erhalten wir

$$\widehat{\mathbf{p}}^{(m)} = \widehat{\mathbf{A}}\widehat{\mathbf{q}}^{(m)} = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{L}^*\mathbf{q}^{(m)} = \mathbf{L}^{-1}\mathbf{A}\mathbf{q}^{(m)},$$

so dass sich

$$\mathbf{p}^{(m)} = (\mathbf{L}^*)^{-1}\widehat{\mathbf{p}}^{(m)} = (\mathbf{L}^*)^{-1}\mathbf{L}^{-1}\mathbf{A}\mathbf{q}^{(m)} = (\mathbf{L}\mathbf{L}^*)^{-1}\mathbf{A}\mathbf{q}^{(m)} = \mathbf{B}^{-1}\mathbf{A}\mathbf{q}^{(m)}$$

ergibt. Anders als im Fall des ursprünglichen Lanczos-Algorithmus müssen wir also für das verallgemeinerte Eigenwertproblem dazu in der Lage sein, Gleichungssysteme der Form

$$\mathbf{B}\mathbf{p}^{(m)} = \mathbf{A}\mathbf{q}^{(m)}$$

effizient zu lösen. Der resultierende verallgemeinerte Lanczos-Algorithmus nimmt mit diesen Modifikationen die folgende Form an:

Algorithmus 10.1 (Verallgemeinerter Lanczos-Algorithmus) *Seien $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert, sei $\mathbf{B} \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit, und sei ein Startvektor $\mathbf{q}^{(1)} \in \mathbb{K}^n$ mit $\|\mathbf{q}^{(1)}\|_B = 1$ gegeben. Der folgende Algorithmus berechnet die bezüglich des zu $\mathbf{B}$ gehörenden Energieskalarprodukts orthonormale Arnoldi-Basis $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(m_0)}$ und die Matrizen $\widehat{\mathbf{A}}^{(m)}$. Der Algorithmus endet mit $m = m_0$.*

$m \leftarrow 1$;
$\mathbf{a} \leftarrow \mathbf{A}\mathbf{q}^{(m)}$;
Löse $\mathbf{B}\mathbf{p} = \mathbf{a}$;
$\gamma \leftarrow \sqrt{\langle \mathbf{a}, \mathbf{p}\rangle_2}$;
$\alpha_1 \leftarrow \langle \mathbf{q}^{(1)}, \mathbf{a}\rangle_2$; $\quad \mathbf{p} \leftarrow \mathbf{p} - \alpha_1 \mathbf{q}^{(1)}$;
$\beta_1 \leftarrow \sqrt{\langle \mathbf{a}, \mathbf{p}\rangle_2}$;
while $\beta_m \geq \epsilon_{ir}\gamma$ **do begin**
  $\mathbf{q}^{(m+1)} \leftarrow \mathbf{p}/\beta_m$;
  $m \leftarrow m + 1$;
  $\mathbf{a} \leftarrow \mathbf{A}\mathbf{q}^{(m)}$;
  Löse $\mathbf{B}\mathbf{p} = \mathbf{a}$;
  $\gamma \leftarrow \sqrt{\langle \mathbf{a}, \mathbf{p}\rangle_2}$;
  $\alpha_m \leftarrow \langle \mathbf{q}^{(m)}, \mathbf{a}\rangle_2$; $\quad \mathbf{p} \leftarrow \mathbf{p} - \bar{\beta}_{m-1}\mathbf{q}^{(m-1)} - \alpha_m \mathbf{q}^{(m)}$;
  $\beta_m \leftarrow \sqrt{\langle \mathbf{a}, \mathbf{p}\rangle_2}$
end

Im Algorithmus wird bei der Berechnung von β_m ein kleiner Trick verwendet: Streng genommen müssten wir

$$\langle \mathbf{p}, \mathbf{p}\rangle_B = \langle \mathbf{B}\mathbf{p}, \mathbf{p}\rangle_2$$

berechnen und würden deshalb $\mathbf{B}\mathbf{p}$ benötigen, also eine Multiplikation mit der Matrix $\mathbf{B}$. Glücklicherweise steht $\mathbf{p}$ an dieser Stelle des Algorithmus nach Konstruktion senkrecht (bezüglich des Energie-Skalarprodukts) auf $\mathbf{q}^{(m-1)}$ und $\mathbf{q}^{(m)}$, erfüllt also

$$\langle \mathbf{B}\mathbf{q}^{(m-1)}, \mathbf{p}\rangle_2 = \langle \mathbf{q}^{(m-1)}, \mathbf{p}\rangle_B = 0, \qquad \langle \mathbf{B}\mathbf{q}^{(m)}, \mathbf{p}\rangle_2 = \langle \mathbf{q}^{(m)}, \mathbf{p}\rangle_B = 0,$$

so dass wir

$$\begin{aligned}\langle \mathbf{a}, \mathbf{p}\rangle_2 &= \langle \mathbf{A}\mathbf{q}^{(m)}, \mathbf{p}\rangle_2 = \langle \mathbf{B}\mathbf{B}^{-1}\mathbf{A}\mathbf{q}^{(m)}, \mathbf{p}\rangle_2 \\ &= \langle \mathbf{B}(\mathbf{B}^{-1}\mathbf{A}\mathbf{q}^{(m)} - \bar{\beta}_{m-1}\mathbf{q}^{(m-1)} - \alpha_m \mathbf{q}^{(m)}), \mathbf{p}\rangle_2 = \langle \mathbf{B}\mathbf{p}, \mathbf{p}\rangle_2\end{aligned}$$

erhalten und so die zusätzliche Multiplikation vermeiden können. Der verallgemeinerte Lanczos-Algorithmus berechnet eine Tridiagonalmatrix, deren Eigenwerte die der Matrix $\widehat{\mathbf{A}}$ approximieren, und damit auch die des verallgemeinerten Eigenwertproblems.

Im Fall der vorkonditionierten Eigenwertverfahren ist die Situation etwas besser, da sich die Inverse der Matrix $\mathbf{B}$ in der Matrix des Vorkonditionierers unterbringen lässt: Wenn $\widehat{\mathbf{N}}$ ein Vorkonditionierer für die Matrix $\widehat{\mathbf{A}}$ ist, nimmt die Richardson-Iteration die Form

$$\widehat{\mathbf{x}}^{(m+1)} = \widehat{\mathbf{x}}^{(m)} - \theta\widehat{\mathbf{N}}(\widehat{\mathbf{A}}\widehat{\mathbf{x}}^{(m)} - \mu\widehat{\mathbf{x}}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0$$

an. Indem wir wie zuvor

$$\widehat{\mathbf{x}}^{(m)} = \mathbf{L}^*\mathbf{x}^{(m)}, \qquad \mathbf{x}^{(m)} = (\mathbf{L}^*)^{-1}\widehat{\mathbf{x}}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

einsetzen, erhalten wir

$$\begin{aligned}\mathbf{x}^{(m+1)} &= (\mathbf{L}^*)^{-1}\widehat{\mathbf{x}}^{(m)} - \theta(\mathbf{L}^*)^{-1}\widehat{\mathbf{N}}(\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\widehat{\mathbf{x}}^{(m)} - \mu\widehat{\mathbf{x}}^{(m)})\\ &= (\mathbf{L}^*)^{-1}\widehat{\mathbf{x}}^{(m)} - \theta(\mathbf{L}^*)^{-1}\widehat{\mathbf{N}}\mathbf{L}^{-1}(\mathbf{A}(\mathbf{L}^*)^{-1}\widehat{\mathbf{x}}^{(m)} - \mu\mathbf{L}\widehat{\mathbf{x}}^{(m)})\\ &= \mathbf{x}^{(m)} - \theta(\mathbf{L}^*)^{-1}\widehat{\mathbf{N}}\mathbf{L}^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{B}\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0.\end{aligned}$$

Wir definieren

$$\mathbf{N} := (\mathbf{L}^*)^{-1}\widehat{\mathbf{N}}\mathbf{L}^{-1}, \qquad \widehat{\mathbf{N}} = \mathbf{L}^*\mathbf{N}\mathbf{L}$$

und erhalten

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \theta\mathbf{N}(\mathbf{A}\mathbf{x}^{(m)} - \mu\mathbf{B}\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Für die Untersuchung der Konvergenz ist wichtig, dass

$$\widehat{\mathbf{N}}^{-1} = \mathbf{L}^{-1}\mathbf{N}^{-1}(\mathbf{L}^*)^{-1}, \qquad \widehat{\mathbf{A}} = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}$$

gelten, so dass wir die Gleichungen

$$\begin{aligned}\langle\widehat{\mathbf{N}}^{-1}\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2 &= \langle\mathbf{L}^{-1}\mathbf{N}^{-1}(\mathbf{L}^*)^{-1}\mathbf{L}^*\mathbf{x},\mathbf{L}^*\mathbf{x}\rangle_2 = \langle\mathbf{N}^{-1}\mathbf{x},\mathbf{x}\rangle_2,\\ \langle\widehat{\mathbf{A}}\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2 &= \langle\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{L}^*\mathbf{x},\mathbf{L}^*\mathbf{x}\rangle_2 = \langle\mathbf{A}\mathbf{x},\mathbf{x}\rangle_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n, \widehat{\mathbf{x}} = \mathbf{L}^*\mathbf{x}\end{aligned}$$

erhalten. Damit gilt die für die Konvergenz des vorkonditionierten Verfahrens wichtige Bedingung (9.8) genau dann, wenn

$$(1-\gamma)\langle\widehat{\mathbf{N}}^{-1}\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2 \leq \langle\widehat{\mathbf{A}}\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2 \leq (1+\gamma)\langle\widehat{\mathbf{N}}^{-1}\widehat{\mathbf{x}},\widehat{\mathbf{x}}\rangle_2 \qquad \text{für alle } \widehat{\mathbf{x}} \in \mathbb{R}^n$$

erfüllt ist. Falls $\widehat{\mathbf{N}}$ also ein guter Vorkonditionierer für $\widehat{\mathbf{A}}$ ist, ist $\mathbf{N}$ auch ein guter Vorkonditionierer für $\mathbf{A}$.

Index

Literaturverzeichnis

[1] G. Frobenius. Ueber Matrizen aus nicht negativen Elementen. *Sitzungsber. Königl. Preuss. Akad. Wiss.*, pages 456–477, 1921.

[2] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, London, 1996.

[3] O. Perron. Zur Theorie der Matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[4] H. Wielandt. Unzerlegbare, nicht negative Matrizen. *Mathematische Zeitschrift*, 52:642–648, 1950.