Kapitel 2

Das symmetrische Eigenwertproblem

Mit diesem Kapitel wenden wir uns der *numerischen linearen Algebra* zu. Konkret geht es um die Lösung des algebraischen Eigenwertproblems

$$Ax = \lambda x,$$

wobei symmetrische (oder hermitesche) Matrizen $A \in \mathbb{R}^{n\times n}$ bzw. $A \in \mathbb{C}^{n\times n}$ im Vordergrund stehen. Diese Fragestellung ist in den Anwendungen von zentraler Bedeutung, da Eigenwerte eine prägnante Charakterisierung des dynamischen Systemverhaltens ermöglichen.

Zur numerischen Lösung des Eigenwertproblems benötigt man nicht nur Wissen und Methoden der linearen Algebra. Da die Eigenwerte durch die Nullstellen des charakteristischen Polynoms gegeben sind, hat man ein spezielles Nullstellenproblem zu lösen. Wie sich zeigen wird, ist dieser naheliegende Berechnungsweg jedoch extrem schlecht konditioniert, während alternative Ansätze, die auf unitären Ähnlichkeitstransformationen beruhen, auf ausgezeichnete Algorithmen führen.

Literaturhinweise:

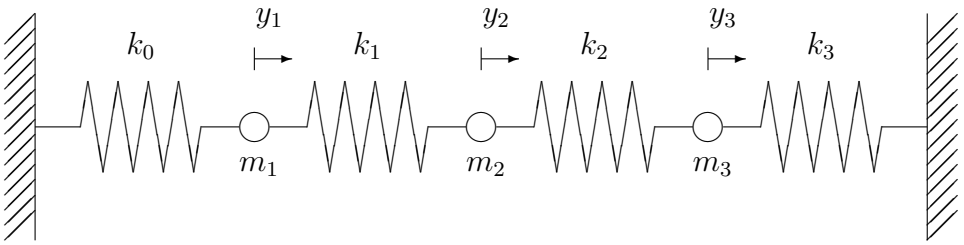
- G. H. Golub, Ch. F. van Loan: Matrix Computation. John Hopkins, 1996
- J. H. Wilkinson, C. Reinsch: Linear Algebra. Springer, 1971

2.1 Eigenwerte und Differentialgleichungen

Als Beispiel für ein Eigenwertproblem betrachten wir Schwingungen in der Punktmechanik:

Beispiel 2.1: *Feder-Masse-Schwinger*

Das folgende System besteht aus drei Massen m_1, m_2 und m_3, die durch Federn mit Federkonstanten k_0, k_1, k_2 und k_3 verbunden sind:



Mit dem *Newtonschen Kraftgesetz* $F = m\ddot{x}$ und dem *Hookeschen Federgesetz* $F = kx$ lassen sich für die drei Körper die Kräftebilanzen aufstellen. Dabei bezeichne y_i die Auslenkung des i-ten Körpers von seiner Ruhelage:

$$\begin{array}{rcrcrcl} m_1\ddot{y}_1 & + & k_1(y_1 - y_2) & + & k_0\, y_1 & = & 0 \\ m_2\ddot{y}_2 & + & k_2(y_2 - y_3) & + & k_1(y_2 - y_1) & = & 0 \\ m_3\ddot{y}_3 & + & k_3\, y_3 & + & k_2(y_3 - y_2) & = & 0 \end{array} \tag{2.1}$$

Für $y_1 = y_2 = y_3 = 0$ erhält man die (uninteressante) Gleichgewichtslösung. Gesucht ist die (nichttriviale) Lösung $y_j(t)$, $j = 1, 2, 3$, von (2.1), einer *linearen Differentialgleichung 2. Ordnung mit konstanten Koeffizienten.*

Aus der Analysis ist bekannt: Das Anfangswertproblem (2.1) zu gegebenen Anfangswerten $y_j(t_0)$, $\dot{y}_j(t_0)$ hat eine eindeutige Lösung.

Lösungsansatz: $y_j(t) = c_j\, e^{i\omega t}$ wobei c_j die Amplitude bezeichnet und ω die Frequenz der Schwingung ($e^{ix} = \cos x + i \sin x$).

Setzt man diesen Ansatz in (2.1) ein, so erhält man

$$\begin{array}{rcrcrcl} -m_1\omega^2 c_1 & + & k_0\, c_1 & + & k_1(c_1 - c_2) & = & 0 \\ -m_2\omega^2 c_2 & + & k_1(c_2 - c_1) & + & k_2(c_2 - c_3) & = & 0 \\ -m_3\omega^2 c_3 & + & k_2(c_3 - c_2) & + & k_3\, c_3 & = & 0 \end{array} \tag{2.2}$$

Mit $c := (c_1, c_2, c_3)^T$, Massenmatrix $M := \operatorname{diag}(m_1, m_2, m_3)$ (positiv definit) und der symmetrischen Steifigkeitsmatrix

$$K := \begin{pmatrix} k_0 + k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 & -k_2 \\ 0 & -k2 & k_2 + k_3 \end{pmatrix}$$

ist (2.2) äquivalent zu dem *verallgemeinerten Eigenwertproblem*

$$(K - \omega^2 M)c = 0 \quad \overset{c \neq 0}{\Longleftrightarrow} \quad \det(K - \lambda^2 M) = 0. \tag{2.3}$$

Überführung in ein gewöhnliches Eigenwertproblem mit dem Ansatz

$$x := M^{1/2}c = \operatorname{diag}(\sqrt{m_1}, \sqrt{m_2}, \sqrt{m_3})c, \quad A := M^{-1/2}KM^{-1/2}, \quad \lambda := \omega^2.$$

Dann schreibt sich das Eigenwertproblem in der Standardform

$$Ax = \lambda x \quad \overset{x \neq 0}{\Longleftrightarrow} \quad \det(A - \lambda I) = 0 \tag{2.4}$$

mit reeller, symmetrischer Matrix A.

Nicht nur Systeme von Punktmassen, sondern auch viele andere Anwendungen werden durch Differentialgleichungen zweiter Ordnung

$$M\ddot{y}(t) + Ky(t) = b(t)$$

mit Massenmatrix M, Steifigkeitsmatrix K und Anregung b beschrieben. Der Ansatz $y = c \exp i\omega t$ mit Frequenz ω liefert das System von Fundamentallösungen im homogenen Fall, und über die Variation der Konstanten ist dann auch der inhomogene Fall behandelbar. Wie in Beispiel 2.1 erhält man ein symmetrisches Eigenwertproblem der üblichen Form $Ax = \lambda x$ über die Transformation

$$x := L^T c \quad A := L^{-1}KL^{-T}, \quad \lambda := \omega^2$$

wobei nun die Cholesky-Zerlegung $M = LL^T$ verwendet wird (M ist positiv definit aber nicht immer diagonal).

Falls schon ein System erster Ordnung

$$\dot{y}(t) = Ay(t) + b(t)$$

vorliegt, ergibt sich mit $y = x \exp \lambda t$ direkt das (möglicherweise unsymmetrische) EWP $Ax = \lambda x$. Wichtige Unterscheidung hierbei:

a) A ist normale Matrix, d.h. es existiert ein orthogonales System von Eigenvektoren. Dann charakterisieren die Eigenwerte das Lösungsverhalten vollständig.

b) Andernfalls gelten Aussagen über die Stabilität und das Wachstum von Lösungen nur asymptotisch, d.h. für $t \to \infty$, denn dann kann $\exp At$ trotz negativer EW zunächst wachsen!

Skizze:

2.2 Theoretische Grundlagen

In diesem Abschnitt werden die wichtigsten Begriffe wiederholt und um die für die Numerik zentralen Aussagen erweitert.

Gegeben: $A \in \mathbb{C}^{n \times n}$ (oder $\mathbb{R}^{n \times n}$)
Gesucht : die „Fixpunkte“ bzw. die „Fixrichtungen“ von A:

$$x \neq 0 \quad \text{mit} \quad Ax = \lambda x, \quad \lambda \text{ skalar}$$

- λ heißt Eigenwert (EW)
- x heißt Eigenvektor (EV)

} der Matrix A

Es gibt auch sogenannte „Linkseigenvektoren“:

$$y \neq 0 \quad \text{mit} \quad y^H A = \lambda y^H, \quad \lambda \text{ skalar} \qquad (\Leftrightarrow \quad A^H y = \overline{\lambda} y)$$

Umformuliert: $x \neq 0$ mit $(A - \lambda I)x = 0$

$$\Leftrightarrow \qquad \phi(\lambda) := \det(A - \lambda I) = 0$$

$\phi \in \mathbb{P}_n$ heißt *charakteristisches Polynom* von A. Es existieren genau n EW (Nullstellen von ϕ).

Beachte: „Alle EW“ heißt n EW
„Alle EV“ heißt $\leq n$ EV (natürlich linear unabhängig).

Wichtige Begriffe:

- Die *algebraische Vielfachheit* eines EW ist die Vielfachheit von λ als Nullstelle von ϕ.
- Die *geometrische Vielfachheit* eines EW λ ist die Dimension des zugehörigen Nullraums, $\dim \mathcal{N}_{A-\lambda I}$. Es gilt:

$$\boxed{1 \leq \text{geom. Vielf.} \leq \text{alg. Vielf.}}$$

Gilt für alle Eigenwerte geom. Vielf. = alg. Vielf., dann gibt es ein *vollständiges System von EV*, d.h. n linear unabhängige EV, d.h. A ist *diagonalisierbar*.

Ähnlichkeitstransformationen (ÄT)

Zwei Matrizen A, B heißen ähnlich, falls eine reguläre Transformation T existiert mit $B = T^{-1}AT$:

$$A \rightsquigarrow B \quad :\Leftrightarrow \quad \exists T : \; B = T^{-1}AT$$

Es gilt für $A \rightsquigarrow B$:

λ EW von A	$\Leftrightarrow$	λ EW von B
x EV von A		$T^{-1}x$ EV von B

Beweis:

Falls $T^{-1} = T^H$: T unitär, und man spricht von einer unitären ÄT.

Welche Vorteile haben unitäre ÄT?

1. $A^H = A \Rightarrow B^H = B$ *strukturerhaltend*

   denn: $B^H = (T^{-1}AT)^H = T^H A^H (T^{-1})^H = T^{-1}AT = B$ wegen $T^H = T^{-1}$ und $A^H = A$.

2. $\|B\|_2 = \|A\|_2$ *normerhaltend*

   denn: $\|B\|_2 \overset{\text{def}}{=} \max\limits_{\|x\|_2=1} \|T^{-1}A\underbrace{Tx}_{y}\|_2 = \max\limits_{\|y\|_2=1} \|T^{-1}Ay\|_2$

   $= \max\limits_{\|y\|_2=1} \|Ay\|_2 = \|A\|_2$

3. $\kappa(A) = \kappa(B)$ *numerisch stabil*

   folgt aus 2. (beachte $\|A^{-1}\|_2 = \|T^{-1}A^{-1}T\|_2$)

Normalformen

Satz 2.1 *Satz von Jordan*

Jede quadratische Matrix läßt sich durch eine ÄT auf Jordansche Normalform bringen:

$$A \rightsquigarrow T^{-1}AT = \operatorname{diag}(J_1, \ldots, J_r) \quad \textit{mit}$$

$$J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_i \times n_i}.$$

$\det(J_i - \lambda I)$ *heißt Elementarteiler.*

Die Jordansche Normalform ist nicht für die Numerik geeignet. Selbst kleinste Rundungsfehler zerstören sofort die Struktur, da mehrfache Nullstellen zu Clustern von nahe benachbarten 1×1-Matrizen (=EW) entarten.

Schränkt man die ÄT auf unitäre Matrizen ein, so erhält man die Schursche Normalform:

Satz 2.2 *Schursche Normalform*

Zu jeder Matrix $A \in \mathbb{C}^{n\times n}$ und jeder Reihenfolge ihrer EW $\lambda_1, \ldots, \lambda_n$ existiert ein unitäres U, so daß

$$U^{-1}AU = U^H AU = \begin{pmatrix} \lambda_1 & * & * \\ & \ddots & * \\ 0 & & \lambda_n \end{pmatrix}.$$

Beweis: per Induktion (nicht konstruktiv, da λ_j als gegeben vorausgesetzt wird)

Induktionsanfang: $n = 1$ klar

$n - 1 \Rightarrow n$:

- Sei λ_1 beliebiger EW von A mit zugehörigem EV x_1, x_1 sei normiert:

$$\|x_1\|_2^2 = x_1^H x_1 = 1$$

$$Ax_1 = \lambda_1 x_1$$

- Konstruiere nun zu x_1 eine Orthonormalbasis $\{x_1, x_2, \ldots, x_n\}$ des $\mathbb{C}^n$
  $\Rightarrow$ Matrix $X = (x_1, x_2, \ldots, x_n)$ ist unitär: $X^H X = I$

- Führe Trafo mit X durch und bilde

$$X^H AXe_1 = X^H Ax_1 = \lambda_1 X^H x_1 = \lambda_1 e_1 \qquad \text{wegen } X^H = X^{-1}$$

$$\Rightarrow \qquad X^H AX = \begin{pmatrix} \lambda_1 & * \\ 0 & A_1 \end{pmatrix} \quad \text{mit } A_1 \in \mathbb{C}^{n-1\times n-1}$$

- Wende die Induktionsvoraussetzung auf A_1 an, d.h. es existiert unitäres U_1 mit

$$U_1^H A U_1 = \begin{pmatrix} \lambda_2 & * & \dots & * \\ & \ddots & \ddots & \vdots \\ & & \ddots & * \\ 0 & & & \lambda_n \end{pmatrix}$$

- Konstruiere U durch

$$U = X \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix}.$$

  Damit folgt

$$U^H A U = \begin{pmatrix} 1 & 0 \\ 0 & U_1^H \end{pmatrix} \underbrace{X^H A X}_{= \begin{pmatrix} \lambda_1 & * \\ 0 & A_1 \end{pmatrix}} \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & * \\ 0 & U_1^H A U_1 \end{pmatrix}$$

$$= \begin{pmatrix} \lambda_1 & * & \dots & * \\ & \lambda_2 & \ddots & \vdots \\ & & \ddots & * \\ 0 & & & \lambda_n \end{pmatrix}$$

□

Eine Folgerung aus der Schurschen Normalform charakterisiert die *normalen Matrizen*:

Satz 2.3 *Unitäre Diagonalisierbarkeit*

Zu jeder normalen Matrix A, d.h. $AA^H = A^H A$ bzw. $AA^T = A^T A$ existiert eine unitäre (orthogonale) Matrix U, so daß

$$U^{H/T} A U = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

mit den Eigenwerten $\lambda_1, \dots, \lambda_n$.

Beweis:

Zu den normalen Matrizen gehören:

• reell symmetrische:

• reell antisymmetrische:

• reell orthogonale:

• selbstadjungierte:

• unitäre:

Speziell folgt weiter

Satz 2.4 *Hauptachsentheorem*

Zu jeder Hermiteschen (symmetrischen) Matrix A ($= A^H$, A^T) existiert eine unitäre (orthogonale) Matrix U, so daß

$$U^{H/T} A U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

mit $\lambda_j \in \mathbb{R}$, $j = 1, \ldots, n$.

In Kurzform: „Symmetrische Matrizen sind diagonalisierbar und besitzen reelle Eigenwerte.“

Zusammenfassung:

A durch unitäre ÄT diagonalisierbar

$\Leftrightarrow$ A normal, $AA^H = A^H A$

$\Leftrightarrow$ A besitzt vollständiges System von orthonormierten Eigenvektoren (geom. Vielf. = alg. Vielf.)

2.3 Der Satz von Gershgorin und die Kondition einfacher Eigenwerte

Bevor wir an die Konstruktion numerischer Verfahren für das EWP gehen, müssen wir noch untersuchen, ob das EWP überhaupt eine sinnvolle Aufgabe, d.h. gut konditioniert ist.

Der Satz von Gershgorin (1931) erlaubt Aussagen über die Verteilung der Eigenwerte. Zu seiner Formulierung wird jeder Zeile einer Matrix A eine abgeschlossene Kreisscheibe in der komplexen Zahlenebene zugeordnet, mit dem Diagonalelement als Mittelpunkt und der Betragssumme der übrigen Elemente dieser Zeile als Radius:

$$K_j := \left\{ z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{k=1, k\neq j}^{n} |a_{jk}| \right\}$$

Dann gilt:

Satz 2.5 *Satz von Gershgorin*

Jeder Eigenwert der Matrix $A \in \mathbb{C}^{n\times n}$ liegt in mindestens einer Kreisscheibe K_j. Alle Eigenwerte liegen in der Vereinigung aller Kreisscheiben. Kein Eigenwert liegt außerhalb dieser Vereinigung.

Beweis: Nur die erste Aussage muss bewiesen werden. Sei λ ein Eigenwert und x ein zugehöriger Eigenvektor mit betragsgrößter Komponente x_j (legt j fest). Aus

$$\sum_{k=1}^{n} a_{jk} x_k = \lambda x_j$$

folgt

$$\lambda - a_{jj} = \sum_{k=1, k\neq j}^{n} a_{jk} x_k / x_j$$

und

$$|\lambda - a_{jj}| \leq \sum_{k=1, k\neq j}^{n} |a_{jk}||x_k|/|x_j| \leq \sum_{k=1, k\neq j}^{n} |a_{jk}|,$$

also $\lambda \in K_j$. □

Mann kann den Satz noch verschärfen und zeigen:
Sind m Kreisscheiben disjunkt zu den übrigen, dann enthalten sie zusammen genau m Eigenwerte.

Beispiel 2.2: $A = \begin{pmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{pmatrix}$

Eigenwertverteilung:

Im folgenden geht es uns um die Änderung der EW bei (infinitesimal kleinen) Änderungen der Matrixelemente. Wie wirken sich diese Störungen der Daten aus? Ist das Eigenwertproblem gut konditioniert?

Bezeichnungen: $A \in \mathbb{C}^{n\times n}$ fest, „ungestörte Matrix“
$\Delta A \in \mathbb{C}^{n\times n}$ beliebig, $\|\Delta A\| \leq \varepsilon$ mit ε Störparameter
$\lambda_i(A + \Delta A) =$ i-ter EW von $A + \Delta A$
$\Delta A = \varepsilon \cdot B$ mit $\|B\| \leq 1$

Definition 2.1: *Konditionszahl des i-ten Eigenwertes*

$$\kappa_i := \lim_{\varepsilon\to 0} \sup_{\|\Delta A\|\leq\varepsilon} \frac{|\lambda_i(A + \Delta A) - \lambda_i(A)|}{\varepsilon} \quad \in [0, \infty]$$

Falls $\kappa_i = \infty \Leftrightarrow$ „λ_i ist ∞-schlecht konditioniert“

Die Konditionszahl κ_i liefert den bestmöglichen Parameter κ für die Abschätzung

$$|\lambda_i(A + \Delta A) - \lambda_i(A)| \leq \kappa\|\Delta A\| + o(\|\Delta A\|)$$

Falls λ_i im „Punkt“ A differenzierbar ist, so gilt $\kappa_i = \|\lambda'(A)\|$ (ist der Fall, falls λ_i *algebraisch einfacher EW* von A ist).

Beachte: Die Definition der Konditionszahl κ_i ist invariant unter unitären Ähnlichkeitstransformationen. Sie setzt voraus, daß sich die i-ten EW von A und $A + \Delta A$ eindeutig zuordnen lassen. Dies folgt aus der Stetigkeit der EW von $A + \varepsilon B$ (stetige Abhängigkeit von ε).

Beweis mit Satz von Rouché: f, g analytisch in $\mathcal{L} \subseteq \mathbb{C}$ und $|f(z)| > |g(z)| \; \forall z \in \partial\mathcal{L} \quad \Rightarrow \quad f$ und $f \pm g$ haben in $\mathcal{L}$ gleich viele Nullstellen. Anwendung auf $\det(A + \varepsilon B - \lambda I)$ als $f(\lambda) \pm g(\lambda)$ und $\det(A - \lambda I)$ als $f(\lambda)$ liefert das Gewünschte.

Für den Fall algebraisch einfacher EW läßt sich Folgendes zeigen:

Satz 2.6 *Kondition einfacher EW*

Sei λ ein algebraisch einfacher EW von A. Dann gilt

$$\kappa = \frac{1}{|y^H x|} = \frac{1}{\cos \angle(x, y)},$$

wobei $\begin{cases} x & \text{eindeutiger Rechts-EV von } A: \quad Ax = \lambda x, \qquad \|x\|_2 = 1 \\ y & \text{eindeutiger Links-EV von } A: \quad y^H A = \lambda y^H, \quad \|y\|_2 = 1. \end{cases}$

Ist A normal, d.h. $A^H A = A A^H$, so gilt sogar $\kappa = 1$.

Bemerkungen:

- Das symmetrische EWP ist für algebraisch einfache EW gut konditioniert: $\kappa = 1$
- Falls λ mehrfacher EW zu lauter linearen Elementarteilern ist, so ist κ beschränkt: $\kappa \leq const$ (ohne Beweis).
- Für die EV gelten diese Aussagen nicht!

Beweis zu Satz 2.6:

Da λ algebraisch einfach ist, gilt: $\phi(\lambda) = 0, \; \phi'(\lambda) \neq 0$ für $\phi(z) := \det(A - zI)$. Betrachte nun das gestörte EWP für die Matrix $A + \varepsilon B$:

$$(A + \varepsilon B)x(\varepsilon) = \lambda(\varepsilon)x(\varepsilon) \quad \text{und} \quad \det(A + \varepsilon B - \lambda(\varepsilon)I) = 0.$$

Mit dem Satz über implizite Funktionen

$$\left.\begin{array}{c} f(x,y) \text{ analytisch mit} \\ f(x_0,y_0)=0,\ \left(\frac{\partial f}{\partial y}\right)_{x_0,y_0} \neq 0 \end{array}\right\} \Rightarrow \quad \exists_1 y(x) \text{ analytisch mit } f(x,y(x))=0$$

ergibt sich mit der Zuordnung $x \mapsto \varepsilon,\ y \mapsto \lambda,\ x_0 \mapsto 0,\ y_0 \mapsto \lambda$ und

$$f(x,y) \equiv \det(A+\varepsilon B-\lambda I), \quad f(x_0,y_0)=\det(A-\lambda I):$$

$\lambda(\varepsilon)$ ist eine analytische Funktion, d.h.

$$\lambda(\varepsilon)=\alpha_0+\alpha_1\varepsilon+\frac{\alpha_2}{2}\varepsilon^2+\ldots \qquad \text{mit } \alpha_i \in \mathbb{C} \text{ geeignet.}$$

Anwendung der Cramerschen Regel, homogene Fassung, liefert

$$x(\varepsilon)=\xi_o+\xi_1\varepsilon+\frac{\xi_2}{2}\varepsilon^2+\ldots, \qquad \xi_i \in \mathbb{C}^n \text{ geeignet.}$$

Einsetzen und Koeffizientenvergleich:

$$\begin{aligned}&(A+\varepsilon B)\left(\xi_0+\xi_1\varepsilon+\frac{\xi_2}{2}\varepsilon^2+\ldots\right)\\ &= \left(\alpha_0+\alpha_1\varepsilon+\frac{\alpha_2}{2}\varepsilon^2+\ldots\right)\left(\xi_0+\xi_1\varepsilon+\frac{\xi_2}{2}\varepsilon^2+\ldots\right)\end{aligned}$$

(1) ε^0 : $A\xi_0=\alpha_0\xi_0$ $\Rightarrow\ \alpha_0=\lambda,\ \xi_0=x$ (Fall $\varepsilon=0$)
(2) ε^1 : $A\xi_1+B\xi_0=\alpha_0\xi_1+\alpha_1\xi_0$ usw.

Setze (1) in (2) ein, $A\xi_1+Bx=\lambda\xi_1+\alpha_1 x$, und multipliziere mit y^H von links:

$$\underbrace{y^HA}_{\lambda y^H}\xi_1+y^HBx=\lambda y^H\xi_1+\alpha_1 y^Hx$$

$$\Rightarrow \qquad \lambda(A+\varepsilon B)=\lambda+\varepsilon\frac{y^HBx}{y^Hx}+\mathcal{O}(\varepsilon^2)$$

Damit:

$$\kappa=\lim_{\varepsilon\to 0}\sup_{\|B\|\le 1}\left|\frac{y^HBx}{y^Hx}+\mathcal{O}(\varepsilon)\right|=\sup_{\|B\|\le 1}\left|\frac{y^HBx}{y^Hx}\right|=\frac{1}{|y^Hx|}.$$

Die letzte Gleichung gilt, denn $|y^HBx| \le \|y\|_2\,\|Bx\|_2 \le \|y\|_2\,\|B\|_2\,\|x\|_2 = 1\cdot\|B\|_2\cdot 1 \le 1$, und für $B=yx^H$ mit $\|B\|=1$ gilt Gleichheit. Somit

$$\kappa=\frac{1}{|y^Hx|}=\frac{1}{\cos\angle(x,y)},$$

da x und y normiert sind. Für normale Matrizen gilt $y^H x = x^H x = 1$, denn

$$A = U \begin{pmatrix} \lambda & & & \\ & * & & \\ & & \ddots & \\ & & & * \end{pmatrix} U^H \quad \text{mit } U^H = U^{-1}$$

$\Rightarrow$ $x = Ue_1$ ist Rechtseigenvektor, $y^H = x^H$ ist Linkseigenvektor, $x^H x = \|x\|_2^2 = \|Ue_1\|_2^2 = \|e_1\|_2^2 = 1 \quad \Rightarrow\ y^H x = x^H x = 1.$ □

Eigenwerte zu nicht-linearen Elementarteilern dagegen sind ∞-schlecht konditioniert.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \Delta A = \varepsilon B, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Die Eigenwerte von A sind $\lambda_1 = \lambda_2 = 0$. Für die Eigenwerte $\tilde{\lambda}$ von

$$A + \varepsilon B = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}$$

ergibt sich $\tilde{\lambda}^2 - \varepsilon = 0 \quad \Rightarrow \quad \tilde{\lambda}_{1/2} = \pm\sqrt{\varepsilon}$

$$\Rightarrow \qquad \frac{\tilde{\lambda}_{1/2} - \lambda_{1/2}}{\varepsilon} = \pm\varepsilon^{-\frac{1}{2}} \ \xrightarrow{\varepsilon \to 0}\ \infty,$$

damit: $\kappa_i = \infty$.

Es ist einleuchtend, daß in diesem schlecht konditionierten Fall eine Berechnung der EW über die Nullstellen des charakteristischen Polynoms auch nicht weiterhelfen kann.

Für das gut konditionierte reell-symmetrische EWP könnte man jedoch zunächst daran denken, das charakteristische Polynom aufzustellen und anschließend seine Nullstellen zu bestimmen. Leider „verschwindet“ die Information der Eigenwerte, falls man das charakteristische Polynom in Koeffizientendarstellung behandelt. Auch im Fall disjunkter Nullstellen ist das

Problem der Nullstellenbestimmung von Polynomen in der Regel schlecht konditioniert.

Hierzu ein warnendes Beispiel von Wilkinson:

$$P(\lambda) = (\lambda - 1)(\lambda - 2) \cdots (\lambda - 20) \in \mathbb{P}_{20}$$

Ausmultiplizieren der Wurzeldarstellung (Hornerschema) ergibt Koeffizienten in der Größenordnung zwischen 1 (Koeffizient von λ^{20}) und 10^{20} (Koeffizient von $\lambda_0 : 20!$). Für das leicht gestörte Polynom ($\varepsilon := 2^{-23} \approx 10^{-7}$)

$$\tilde{P}(\lambda) := P(\lambda) - \varepsilon\lambda^{19}$$

(Koeffizient von λ^{19} in $P : \mathcal{O}(10^3)$) ergeben sich die folgenden Nullstellen:

1.000000000	$10.095266145 \pm 0.643500904i$
2.000000000	$11.793633881 \pm 1.652329728i$
3.000000000	$13.992358137 \pm 2.518830070i$
4.000000000	$16.730737466 \pm 2.812624894i$
4.999999928	$19.502439400 \pm 1.940330347i$
6.000006944	20.846908101
6.999697234	
8.007267603	
8.917250249	

Man ist daher auf iterative Verfahren angewiesen.

2.4 Reduktionsalgorithmen

Der Satz von Schur sagt aus, daß für jede Matrix $A \in \mathbb{C}^{n\times n}$ eine unitäre Matrix U existiert, so dass

$$U^H A U = \begin{pmatrix} \lambda_1 & * & \dots & * \\ & \ddots & \ddots & \vdots \\ & & \ddots & * \\ 0 & & & \lambda_n \end{pmatrix}$$

mit den EW $\lambda_1, \ldots, \lambda_n$ von A.

Eine erste Idee, sämtliche EW von A gleichzeitig zu berechnen, wäre, U in endlich vielen Schritten zu bestimmen. Da die EW die Wurzeln des charakteristischen Polynoms sind, hätte man damit auch ein endliches Verfahren zur Bestimmung der Nullstellen von Polynomen beliebigen Grades gefunden. Das ist jedoch nicht möglich (Satz von Abel).

Eine zweite Idee führt hier zum Ziel: In einem ersten Schritt bringt man die Matrix durch ÄT (am besten unitär) auf möglichst einfache Gestalt. Man spricht hierbei von Reduktionsalgorithmen. In einem zweiten Schritt löst man iterativ das (nun viel billigere) EWP für die reduzierte Matrix.

Eine alternative Methode im Falle symmetrischer Matrizen stellt das **Jacobi-Verfahren** dar. Man wendet sukzessive Rotationsmatrizen als ÄT an, die Matrixelemente zu Null machen. Für dieses Verfahren läßt sich zeigen, daß die Matrix gegen $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ konvergiert.

Idee der Reduktionsalgorithmen

Vereinfache Ausgangsmatrix A durch unitäre ÄT:

$$A = A^{(0)} \;\to\; A^{(1)} \;\to\; A^{(2)} \;\to\; \ldots$$

$$\to \quad A^{(n-2)} = \begin{cases} \begin{pmatrix} * & \ldots & \ldots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{pmatrix} & \text{obere Hessenberg-Form} \\ & \text{im unsymmetrischen Fall} \\ \begin{pmatrix} * & * & & \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ & & * & * \end{pmatrix} & \text{Tridiagonalmatrix} \\ & \text{im symmetrischen Fall} \end{cases}$$

Als Werkzeug verwendet man *Householder-Reflexionen.*

Definition 2.2: *Die Matrix*

$$T := I - 2 \cdot v \cdot v^H$$

mit $v \in \mathbb{C}^m$ *und* $\|v\|_2 = 1$ *heisst Spiegelungsmatrix.*

Schema für T:

Einen beliebigen Vektor $x \in \mathbb{C}^m$ kann man zerlegen in $x = p + s$ mit $p := v(v^H x)$ parallel zu v und $s := x - p$ senkrecht zu v.

Die Matrix T spiegelt dann den parallelen Anteil p an s,

$$T \cdot (s + p) = T \cdot (s + v(v^H x)) = (I - 2vv^H) \cdot (s + v(v^H x)) = s - p\,.$$

Skizze:

Die Spiegelungsmatrix T hat besondere Eigenschaften:

(i) T ist hermitesch (symmetrisch im reellen Fall), $T^H = T$

(ii) T ist involutorisch, $T^{-1} = T$

(iii) T ist unitär (orthogonal)

Statt mittels eines normierten Vektors v kann man T auch durch

$$T := I - uu^H/\kappa\,, \quad \kappa := \frac{1}{2}u^H u$$

mit $u \in \mathbb{C}^m/\{0\}$ definieren.

Die Anwendung von T auf einen Vektor x vermöge

$$y = T \cdot x = x - uu^H x/\kappa$$

benötigt keine Matrix-Vektormultiplikation. Man formt zunächst das Skalarprodukt $\sigma := u^H x/\kappa$ und dann $y := x - \sigma u$. Analog geht man bei der Anwendung auf Matrizen vor:

$$T \cdot A = A - u(u^H A/\kappa)\,, \quad \text{jede Spalte wie zuvor.}$$

Die Matrix A muss dabei nicht quadratisch sein.

In den Anwendungen hat man es immer mit einem Spezialfall zu tun. T bzw. u ist so zu bestimmen, dass die Anwendung auf ein gegebenes x gerade ein Vielfaches des ersten Einheitsvektors e_1 ergibt,

$$T \cdot x = y = -\zeta e_1 = \begin{pmatrix} -\zeta \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Die folgende Vorschrift berechnet diese spezielle Spiegelung:

$$\begin{aligned} \zeta &:= \|x\|_2 \cdot x_1/|x_1| \\ u &:= x + \zeta e_1 = (x_1 + \zeta, x_2, \ldots, x_m)^T \\ \kappa &:= x^H x + \|x\|_2 \cdot |x_1| \\ T &:= I - uu^H/\kappa \end{aligned} \tag{2.5}$$

Man nennt die Spiegelung T dann eine *Householdertransformation*, und es gilt $Tx = -\zeta e_1 = (-\zeta, 0, \ldots, 0)^T$ (Beweis per Nachrechnen).

Anwendung bei der Reduktion des EWP: Die Ähnlichkeitstrafo $A \rightsquigarrow U^H AU$ würde die Dreiecksform zerstören, daher ist eine Modifikation notwendig. Ziel ist es, die Matrix A durch eine unitäre ÄT auf obere Hessenberg-Form zu bringen:

$$A \;\rightsquigarrow\; U^H AU = \begin{pmatrix} * & \dots & \dots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix},$$

wobei U das Produkt von $n-2$ Householder-Reflexionen ist.

1. Teilschritt:

• Partitioniere A in

$$A = \left(\begin{array}{c|ccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{array}\right)$$

und betrachte den Vektor $u = (0, u_2, \dots, u_n)^T$.

Wähle die Komponenten von u so, daß der $(n-1)$-dimensionale Vektor $(a_{21}, a_{31}, \dots, a_{n1})^T$ gespiegelt wird:

$$\begin{pmatrix} a_{21} \\ \vdots \\ \vdots \\ a_{n1} \end{pmatrix} \;\rightsquigarrow\; T \cdot \begin{pmatrix} a_{21} \\ \vdots \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

mit $T = I - \kappa u u^H$, $\kappa = \frac{2}{u^H u}$, wobei

$$\left.\begin{array}{l} u_2 = a_{21} + \frac{a_{21}}{|a_{21}|}\sqrt{s}, \quad s = |a_{21}|^2 + \ldots + |a_{n1}|^2 \\ u_k = a_{k1},\ k = 3, \dots, n \end{array}\right\} \;\frac{1}{\kappa} = \frac{1}{2} u^H u = s + |a_{21}|\sqrt{s}$$

• Anwendung der unitären ÄT ergibt (nachrechnen!):

$$\underset{A^{(0)}=A}{\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & * & \dots & * \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{n1} & * & \dots & * \end{pmatrix}} \rightsquigarrow \underset{A'=TA}{\left(\begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline * & \star & \dots & \star \\ 0 & \vdots & \text{geändert} & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & \star & \dots & \star \end{array}\right)} \rightsquigarrow$$

$A^{(1)} = A'' = TAT$ (wegen $T^H = T$)

$$A^{(1)} = \left(\begin{array}{c|ccc} a_{11} & \times & \dots & \times \\ * & \vdots & & \vdots \\ 0 & \vdots & \text{geändert} & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & \times & \dots & \times \end{array}\right)$$

2. bis $(n-2)$. Teilschritt: Rekursive Berechnung analog zum 1. Teilschritt.

Ergebnis: $A^{(n-2)}$ hat obere Hessenberg-Form.

Falls $A \in \mathbb{C}^{n\times n}$ Hermitesch ist, erhält man durch dieses Vorgehen sogar Tridiagonalform:

$$A \rightsquigarrow U^H A U = \begin{pmatrix} * & * & & \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ & & * & * \end{pmatrix}$$

Damit kann tatsächlich wie gefordert eine Reduktion in $n-2$ Teilschritten durchgeführt werden.

Vorgehen beim EW- / EV-Problem:

1.) Transformiere A auf einfachere Form,
$A \rightsquigarrow$ Tridiagonal- bzw. obere Hessenberg-Matrix
mittels ÄT, i.A. $(n-2)$ Teilschritte

2.) Wende *iteratives Verfahren* auf die *vereinfachte Matrix* an!

- Inverse Vektoriteration zur Bestimmung eines EW / EV (Abschnitt 2.5)
- QR-Algorithmus mit Shift-Technik für alle EW / EV (Abschnitt 2.6)
- Prinzip der *Sturmschen Kette* mit *Bisektion* für spezielle EW (Abschnitt 2.7)

2.5 Vektoriteration (Potenzmethoden)

Die Konvergenzeigenschaften des QR-Verfahrens (vgl. Abschnitt 2.6) beruhen wesentlich auf der einfachsten direkten Möglichkeit, Eigenwerte und -vektoren einer *normalen* Matrix $A \in \mathbb{C}^{n\times n}$ zu bestimmen: der Vektoriteration.

a) Direkte Vektoriteration nach von Mises

Betrachte die *Krylov-Sequenz*

$$v^{(i+1)} := Av^{(i)} = A^{i+1}v^{(0)}, \quad i = 0, 1, 2, \ldots \tag{2.6}$$

zu einem Startwert $v^{(0)} \in \mathbb{C}^n \setminus \{0\}$.

Falls λ_1 dominanter EW ist, d.h. λ_1 einfach und $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$, so konvergieren die normierten Krylov-Vektoren $v^{(i)}/\|v^{(i)}\|_2$ und die quadratischen Formen (*Rayleigh-Quotienten*) $v^{(i)\,H}Av^{(i)}/\|v^{(i)}\|_2^2$ gegen den normierten EV x_1 zum dominanten EW λ_1:

$$\lim_{i\to\infty} \frac{v^{(i)}}{\|v^{(i)}\|_2} = x_1, \qquad \lim_{i\to\infty} \frac{v^{(i)\,H}Av^{(i)}}{\|v^{(i)}\|_2^2} = \lambda_1.$$

Denn stellt man $v^{(0)}$ in der Orthonormalbasis $\{x_1, x_2, \ldots, x_n\}$ von A dar,

$$v^{(0)} = \sum_{i=1}^{n} \alpha_i x_i,$$

folgt

$$v^{(k)} = A^k v^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k x_i = \alpha_1 \lambda_1^k \left(x_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i \right),$$

wobei $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$ (dabei wird $\alpha_1 \neq 0$ angenommen).

Die einfache Vektoriteration (2.6) hat jedoch gravierende Nachteile:

- Man erhält nur den EV zum betragsgrößten EW λ_1. Aber: Falls $|\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$, so ergibt der Ansatz mit $A - \lambda_1 x_1 x_1^H / x_1^H x_1$ statt A Konvergenz gegen x_2 bzw. λ_2.
- Falls $v^{(0)} \perp x_1$ (d.h. $\alpha_1 = 0$), ist der Startwert nicht hinreichend allgemein gewählt. Allerdings ergibt sich nur theoretisch keine Konvergenz gegen x_1, λ_1, praktisch doch aufgrund der Rundungsfehler (diese können also durchaus erwünscht sein).
- Die Konvergenz ist linear mit Konvergenzfaktor $|\lambda_2/\lambda_1|$. Liegen die EW λ_2 und λ_1 betragsmäßig dicht zusammen, so konvergiert die direkte Vektoriteration nur sehr langsam.

Bessere Ergebnisse liefert die

b) Inverse Vektoriteration nach Wielandt

Betrachte nun die Folge

$$(A - \lambda I) v^{(i+1)} = v^{(i)}, \quad i = 0, 1, 2, \ldots \tag{2.7}$$

zu einem Startwert $v^{(0)} \in \mathbb{C}^n \setminus \{0\}$ für eine Näherung λ zum EW λ_j, wobei angenommen wird, dass $0 \neq |\lambda_j - \lambda| \ll |\lambda_i - \lambda|$, $i \neq j$. Dann ergibt sich

$$v^{(k)} = (A - \lambda I)^{-k} v^{(0)} \overset{\text{ONB}}{=} (A - \lambda I)^{-k} \left(\sum_{i=1}^{n} \alpha_i x_i \right) = \sum_{i=1}^{n} \alpha_i \frac{1}{(\lambda_i - \lambda)^k} x_i$$

wegen $(A - \lambda I) x_i = (\lambda_i - \lambda) x_i$. Also

$$(\lambda_j - \lambda)^k v^{(k)} = \alpha_j x_j + \sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \left(\underbrace{\frac{\lambda_j - \lambda}{\lambda_i - \lambda}}_{|\cdot| \ll 1} \right)^k x_i$$

und somit folgt analog zur direkten Vektoriteration

$$\lim_{i \to \infty} \frac{v^{(i)}}{\|v^{(i)}\|_2} = x_j, \qquad \lim_{i \to \infty} \frac{v^{(i)\,H} A v^{(i)}}{\|v^{(i)}\|_2^2} = \lambda_j.$$

Da man bei der inversen Iteration die Eigenwerte von A mit der Näherung λ verschiebt auf die Eigenwerte von $A - \lambda I$, bezeichnet man λ als *Shift-Parameter*. Falls keine gute Näherung bekannt ist, nimmt man den Rayleigh-Quotienten $\lambda = v^{(0)\,H} A v^{(0)} / \|v^{(0)}\|_2^2$ als Shift. Bei einer Variante der inversen Iteration, der *Rayleigh-Quotienten-Iteration*, wird der Shift im übrigen in jedem Schritt über den Rayleigh-Quotienten aktualisiert.

Zwei Punkte sind bei der inversen Iteration noch von Bedeutung: Erstens liegt wieder lineare Konvergenz mit Konvergenzfaktor

$$\max_{i \neq j} \left| \frac{\lambda_j - \lambda}{\lambda_i - \lambda} \right| < 1$$

vor. Falls λ eine besonders gute Schätzung von λ_j ist, so gilt

$$\left| \frac{\lambda_j - \lambda}{\lambda_i - \lambda} \right| \ll 1 \qquad \forall i \neq j.$$

Dann konvergiert das Verfahren sehr schnell, i.A. reicht $k = 1, 2$ aus.

Zweitens muß im Gegensatz zur direkten Vektoriteration bei jedem Iterationsschritt das lineare Gleichungssystem (2.7) für verschiedene rechte Seiten $v^{(i)}$ gelöst werden. Daher bringt man A zuerst auf Tridiagonalform, wie im Abschnitt 9.4 beschrieben.

Im folgenden werden wir sehen, daß das QR-Verfahren, das den *state-of-the-art* für die EW- / EV-Berechnung darstellt, eng verwandt ist mit den beiden Versionen der Vektoriteration.

2.6 Der QR-Algorithmus

DAS Verfahren zur Berechnung aller Eigenwerte einer hermiteschen Tridiagonalmatrix. Die Idee dazu geht auf Francis und Kublanovskaja (1961) zurück:

Bilde eine Folge von Matrizen $\{A_k\}_{k\in\mathbb{N}}$ mit

$$A_0 := A$$
$$A_k = Q_k R_k \qquad \text{„QR-Zerlegung“} \tag{2.8}$$
$$A_{k+1} = R_k Q_k \tag{2.9}$$

wobei Q_k unitär ist und R_k eine obere Dreiecksmatrix.

Erste wesentliche Beobachtung: Für das obige *QR-Verfahren* gilt:

(i) Die Matrizen A_k sind ähnlich zu A.

(ii) Ist A hermitesch, so auch A_k.

denn: $A_k = Q_k R_k$ und $A_{k+1} = R_k Q_k \quad \Rightarrow \quad R_k = Q_k^H A_k$. Also liegt eine ÄT $A_{k+1} = Q_k^H A_k Q_k$ vor, und die EW bleiben erhalten. Die Eigenschaft (ii) zeigt man induktiv mittels $(A_{k+1})^H = Q_k^H A_k^H Q_k^{HH} \overset{A_k^H = A_k}{=} Q_k^H A_k Q_k = A_{k+1}$.

Bemerkungen:

- Für reguläre Matrizen A ist die Zerlegung bis auf Phasenfaktoren eindeutig, d.h.
$$Q_k R_k = \tilde{Q}_k \tilde{R}_k \quad \Rightarrow \quad \tilde{Q}_k = Q_k S,\ \tilde{R}_k = S^H R_k$$
mit $S = \text{diag}\,(e^{i\varphi_1}, e^{i\varphi_2}, \ldots, e^{i\varphi_n})$.
- Jeder Schritt ist eine unitäre ÄT $\Rightarrow$ EW bleiben erhalten.
- Das QR-Verfahren kann so implementiert werden, daß gilt: ist A hermitesch und tridiagonal, so auch A_k (s.u.).
- Wie unten gezeigt wird, gilt mit der Zerlegung $A_k = L_k + D_k + L_k^H$:
$$\lim_{k\to\infty} L_k = 0$$
$$\lim_{k\to\infty} D_k = \Lambda = \text{diag}\,(\text{EW von } A),$$
falls die EW von A paarweise verschieden sind.

Bevor wir das Konvergenzverhalten des QR-Verfahrens genauer betrachten, wollen wir zuerst den engen Zusammenhang des QR-Verfahrens mit den Vektoriterationen des letzten Abschnitts aufzeigen (für reelles A):

Wegen der ÄT $A_{k+1} = Q_k^T A_k Q_k$ ist

$$A_k = P_k^T A P_k, \quad P_k := Q_1 \cdot \ldots \cdot Q_{k-1}.$$

Daraus schliesst man auf

$$AP_k = P_k A_k = P_k Q_k R_k = P_{k+1} R_k \tag{2.10}$$
$$P_k^T A^{-1} = R_k^{-1} P_{k+1}^T. \tag{2.11}$$

Wir beobachten: Die jeweils erste Spalte der Transformationsmatrix P_k folgt der gewöhnlichen Vektoriteration mit *Normierung*:

$$(2.10) \quad \Rightarrow \qquad AP_k e_1 = r_{11}^{(k)} P_{k+1} e_1$$

Darüberhinaus folgt die jeweils letzte Spalte dieser Transformationsmatrix bzw. die letzte Zeile der linksseitigen Transformationsmatrix P_k^T der inversen Vektoriteration (für einen Links-EV) nach Wielandt mit Normierung, aber *ohne Shift* :

$$(2.11) \quad \Rightarrow \qquad e_n^T P_k^T A^{-1} = e_n^T P_{k+1}^T / r_{nn}^{(k)}$$

Damit ist das asymptotische Verhalten der ersten Spalte und letzten Zeile von P_k klar, und wir können auf die Resultate aus Abschnitt 2.5 zurückgreifen:

Seien $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$ die EW von A. Dann konvergiert

$$P_k e_1 \overset{k\to\infty}{\longrightarrow} x_1 \qquad \text{(normierter EV zu } \lambda_1\text{)},$$
$$P_k e_1 = x_1 + f_k \qquad \text{mit Fehler } f_k = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right).$$

Für die euklidische Länge der ersten Spalte a_1 (ohne a_{11}) der Matrix A_k

ergibt sich

$$\begin{aligned} v_k &:= \underbrace{\left\| \begin{pmatrix} a_{21}^{(k)} & a_{31}^{(k)} & \dots & a_{n1}^{(k)} \end{pmatrix} \right\|_2}_{= \|a_{21}^{(k)}\|_2 \text{ für Tridi.}} = \|(A_k - a_{11} I)\, e_1\|_2 \\ &= \min_\lambda \|(A_k - \lambda I)\, e_1\|_2 = \min_\lambda \left\| P_k^T\, (A - \lambda I)\, P_k e_1 \right\|_2 \\ &= \min_\lambda \|(A - \lambda I)\, P_k e_1\|_2 = \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right). \end{aligned} \tag{2.12}$$

Der Skalar v_k ist also das kleinstmögliche Residuum zu $P_k e_1$ als Näherung für (irgendeinen) EV. Mit jedem QR-Schritt wird v_k um den Faktor $|\lambda_2/\lambda_1|$ verkleinert.

Für die Betrachtung der letzten Spalte sei nun auch $|\lambda_{n-1}| > |\lambda_n|$. Somit konvergiert

$$\begin{aligned} &e_n^T P_k^T \overset{k\to\infty}{\longrightarrow} x_n^T \qquad \text{(normierter EV zu } \lambda_n\text{)}, \\ &e_n^T P_k^T = x_n^T + \bar{f}_k^T \qquad \text{mit Fehler } \bar{f}_k = \mathcal{O}\left(\left| \frac{\lambda_n}{\lambda_{n-1}} \right|^k \right). \end{aligned}$$

Für die euklidische Länge der n-ten Zeile von A ohne a_{nn} (entspricht der n-ten Spalte von A für symmetrische A!) findet man analog

$$\begin{aligned} \tilde{v}_k &:= \left\| \begin{pmatrix} a_{n1}^{(k)} & a_{n2}^{(k)} & \dots & a_{n,n-1}^{(k)} \end{pmatrix} \right\|_2 = \dots \\ &= \min_\lambda \left\| e_n^T P_k^T (A - \lambda I) \right\|_2 = \mathcal{O}\left(\left| \frac{\lambda_n}{\lambda_{n-1}} \right|^k \right). \end{aligned} \tag{2.13}$$

Dabei ist $\tilde{v}_k$ das kleinstmögliche Residuum zu $e_n^T P_k^T$ als Näherung für (irgendeinen) EV. Mit jedem QR-Schritt wird $\tilde{v}_k$ um den Faktor $|\lambda_n/\lambda_{n-1}|$ verkleinert.

Diese nahe Verwandtschaft des QR-Verfahrens mit beiden Vektoriterationsalgorithmen motiviert den folgenden Satz, der die Konvergenz des QR-Verfahrens sicherstellt, falls alle EW von A paarweise verschieden sind.

Allerdings sieht man auch, daß die Konvergenzgeschwindigkeit vom Verhältnis $|\lambda_2/\lambda_1|$ bzw. $|\lambda_n/\lambda_{n-1}|$ der EW abhängt, was für nahe benachbarte EW zu beliebig langsamer Konvergenz führt. Die Shiftstrategie der inversen Iteration kann die Konvergenz jedoch so weit verbessern, dass man für symmetrische Matrizen A sogar kubische Konvergenz erhält. Näheres dazu später.

Satz 2.7 *Konvergenz des QR-Verfahrens*

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit den paarweise verschiedenen EW $\lambda_1, \ldots, \lambda_n$, und $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0$. Dann gilt mit den Bezeichnungen aus (2.8) - (2.9) und $A_k = \left(a_{ij}^{(k)}\right)$:

(i) $\lim_{k\to\infty} Q_k = I$

(ii) $\lim_{k\to\infty} R_k = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: \Lambda$

(iii) $\lim_{k\to\infty} a_{ij}^{(k)} = \mathcal{O}\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right)$ für $i > j$

Beweis: Idee: bringe Zerlegung von A_k in Zusammenhang mit A^k-Zerlegung und zeige die Aussagen für A^k (einfacher!).

- $A^k = \underbrace{Q_0 \cdot \ldots \cdot Q_{k-1}}_{P_k} \cdot \underbrace{R_{k-1} \cdot \ldots \cdot R_0}_{U_k}$

Beweis per vollständiger vollständige Induktion:

$k = 0:$ $\quad A^1 = A_0 = Q_0 R_0$

$k \to k+1:$

$$\begin{aligned} A_k &= Q_k R_k \qquad \text{nach QR-Zerlegung} \\ &= \underbrace{Q_{k-1}^T \cdot \ldots \cdot Q_0^T}_{P_k^T} \cdot A \cdot \underbrace{Q_0 \cdot \ldots \cdot Q_{k-1}}_{P_k} \\ \Rightarrow AP_k &= P_k Q_k R_k \end{aligned}$$

Damit ist

$$\begin{aligned} A^{k+1} &= A \cdot A^k = AP_kU_k \quad \text{nach Induktionsvoraussetzung} \\ &= \underbrace{P_kQ_k}_{P_{k+1}} \cdot \underbrace{R_kU_k}_{U_{k+1}} \end{aligned}$$

Ergebnis: die Zerlegung von A^k ist mittels der A_k-Zerlegung bestimmbar!

Untersuche jetzt *(i) - (iii)* für A^k.

• Für die EW von A^k gilt:

$$\begin{aligned} & \Lambda = Q^T A Q \qquad \text{(Hauptachsentrafo)} \\ \Rightarrow \quad & \Lambda^k = Q^T A^k Q = \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k) \end{aligned}$$

• Trick von Wilkinson: Zerlege zur weiteren Untersuchung die orthogonale Matrix Q^T in

$$Q^T = L \cdot R.$$

Eine solche Zerlegung existiert stets nach einer geeigneten Permutation (auch unitäre ÄT) von A und damit Q (A ist regulär!). Damit gilt: $A^k = Q\Lambda^k Q^T = Q\Lambda^k LR = Q\big(\Lambda^k L\Lambda^{-k}\big)\big(\Lambda^k R\big)$. Da L normierte untere Dreiecksmatrix ist, folgt speziell

$$\big(\Lambda^k L\Lambda^{-k}\big)_{ij} = l_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k \quad \text{mit} \quad \left|\frac{\lambda_i}{\lambda_j}\right| < 1, \quad i > j,$$

$\Lambda^k L\Lambda^{-k}$ ist untere Dreiecksmatrix und

$$\Lambda^k L\Lambda^{-k} = I + E_k \text{ mit } E_k = \begin{pmatrix} 0 & & & \\ * & \ddots & & \\ \vdots & \ddots & \ddots & \\ * & \ldots & * & 0 \end{pmatrix} \overset{k\to\infty}{\longrightarrow} 0$$

Angewandt auf die A^k-Darstellung ergibt sich

$$A^k = Q\big(\Lambda^k L\Lambda^{-k}\big)\big(\Lambda^k R\big) = Q\big(I + E_k\big)\big(\Lambda^k R\big)$$

Nächster Trick: Zerlege $I + E_k$ formal nach QR: $I + E_k = \hat{Q}_k \hat{R}_k$ (mit positiven Diagonalelementen von $\hat{R}_k$ $\Rightarrow$ eindeutig!). Wegen $E_k \overset{k\to\infty}{\longrightarrow} 0$ folgt dann $\hat{Q}_k, \hat{R}_k \overset{k\to\infty}{\longrightarrow} I$.
• Ergebnis: Wir haben eine weitere QR-Zerlegung von A^k berechnet:

$$A^k = \big(Q\hat{Q}_k\big)\big(\hat{R}^k \Lambda^k R\big)$$

Bis auf Vorzeichen in der Hauptdiagonale gilt

$$A^k = P_k U_k, \qquad P_k = Q\hat{Q}_k, \qquad U_k = \hat{R}_k \Lambda^k R$$

mit den Grenzwerteigenschaften

$$\begin{aligned} P_{k+1} = P_k Q_k \quad \Rightarrow \quad Q_k &= P_k^T P_{k+1} = \big(Q\hat{Q}_k\big)^T \big(Q\hat{Q}_{k+1}\big) \\ &= \hat{Q}_k^T \hat{Q}_{k+1} \overset{k\to\infty}{\longrightarrow} I \\ U_{k+1} = R_k U_k \quad \Rightarrow \quad R_k &= U_{k+1} U_k^{-1} = \big(\hat{R}_{k+1}\Lambda^{k+1} R\big)\big(\hat{R}_k \Lambda^k R\big)^{-1} \\ &= \hat{R}_{k+1} \Lambda \hat{R}_k^{-1} \overset{k\to\infty}{\longrightarrow} \Lambda \end{aligned}$$

Mit den beiden letzten Aussagen folgt zu guter letzt $\lim_{k\to\infty} A_k = \Lambda$. □

Analog zur direkten Vektoriteration ist die Konvergenz des QR-Verfahrens sehr langsam, falls $|\lambda_i/\lambda_j| \approx 1$ ist. Wie bei der inversen Vektoriteration nach Wielandt kann hier eine Shiftstrategie abhelfen:

Sei $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}| > |\lambda_n|$ und $|\lambda_{n-1}| \approx |\lambda_n|$.

Wähle Shift σ so, dass

$$|\lambda_1 - \sigma| \geq |\lambda_2 - \sigma| \geq \ldots \geq |\lambda_{n-1} - \sigma| \gg |\lambda_n - \sigma|.$$

Dann konvergiert gemäß (2.13) $a_{n,n-1}^{(k)}$ wie

$$\tilde{v}_k = \|a_{n,n-1}^{(k)}\|_2 = \mathcal{O}\left(\underbrace{\left|\frac{\lambda_n - \sigma}{\lambda_{n-1} - \sigma}\right|}_{\ll 1}^{k}\right).$$

Wählt man als Shift $\sigma_{k-1} = a_{nn}^{(k-1)}$, so hat man im Normalfall quadratische, für hermitesche Tridiagonalmatrizen sogar kubische Konvergenz (im Gegensatz zur linearen Konvergenz des einfachen QR-Verfahrens!).

Begründung: (nur einfache EW) Sei $\tilde{v}_k \leq \varepsilon$. Dann folgt nach Gerschgorin $|\lambda_n - a_{nn}^{(k)}| = \mathcal{O}(\varepsilon)$ und

$$\tilde{v}_{k+1} = \underbrace{\tilde{v}_k}_{\leq \varepsilon} \cdot \underbrace{\left|\frac{\lambda_n - a_{nn}^{(k)}}{\lambda_{n-1} - a_{nn}^{(k)}}\right|}_{\mathcal{O}(\varepsilon)} = \mathcal{O}(\varepsilon^2).$$

Sei $\tilde{v}_k = \|a_{n,n-1}^{(k)}\|_2 = \|a_{n-1,n}^{(k)}\|_2 \leq \varepsilon$ im Spezialfall einer symmetrischen Tridiagonalmatrix. Man kann zeigen, dass

$$|\lambda_n - a_{nn}^{(k)}| \leq \varepsilon^2$$

und somit

$$\tilde{v}_{k+1} = \underbrace{\tilde{v}_k}_{\leq \varepsilon} \cdot \underbrace{\left|\frac{\lambda_n - a_{nn}^{(k)}}{\lambda_{n-1} - a_{nn}^{(k)}}\right|}_{\mathcal{O}(\varepsilon^2)} = \mathcal{O}(\varepsilon^3).$$

In der Praxis ist auch bei mehrfachen EW kein Beispiel bekannt, bei dem sich eine geringere Konvergenzordnung zeigen würde.

Statt $\sigma_k = a_{nn}^{(k)}$ wird nach einem Vorschlag von Wilkinson der EW des unteren 2×2-Blocks gewählt, welcher näher an $a_{nn}^{(k)}$ ist. Das gibt raschere Konvergenz, ohne allerdings die Konvergenzordnung zu verbessern. Um im Falle reeller nichtsymmetrischer Matrizen A_k eine komplexe Matrix A_{k+1} für einen komplexen Shift σ_k zu vermeiden, wird ein Doppelschritt mit σ_k und $\overline{\sigma_k}$ nach Francis verwandt (vgl. Stoer/Bulirsch).

Somit ergibt sich

Algorithmus 2.1 *QR-Verfahren mit Shift*

$A_0 := A$

Für $k = 0, 1, 2, \ldots$

- Wähle Shift σ_k
- Zerlegung: $A_k - \sigma_k I =: Q_k R_k$ (2.14)
- Rekombination: $R_k Q_k + \sigma_k I =: A_{k+1}$ (2.15)

Wie im Fall ohne Shift zeigt man sofort, daß alle Matrizen A_k ähnlich zu A sind und die Symmetrie erhalten bleibt.

Technische Durchführung

Implementiert wird die QR-Zerlegung mittels ebener Rotationen in der sogenannten *impliziten Shift-Strategie*. Diese hat folgende Vorteile:

- Tridiagonalform bleibt erhalten.

- Falls $|\sigma_k|$ groß ist, so kann die Addition und Subtraktion von $\sigma_k I$ in (2.14) und (2.15) zu Genauigkeitsverlusten führen. Die implizite Shift-Strategie hingegen berechnet A_{k+1} ohne explizite Addition und Subtraktion von $\sigma_k I$.
- Ist $|a_{n,n-1}^{(k)}| \leq \varepsilon$ (Schwellenwert), so ist nach Gerschgorin der n-te EW λ_n bis auf $\mathcal{O}(\varepsilon^2)$ im symmetrischen Fall genau. Dieser kann dann abgelesen werden, und mit der Restmatrix der Dimension $n-1$ kann weitergerechnet werden (sogenannte *Deflation*).

Nach Vorabreduktion auf Tridiagonalgestalt mittels Householdertransformationen (Aufwand $4/3n^3$ flops) kann das eigentliche QR-Verfahren auf diese Art und Weise in $O(n^2)$ flops realisiert werden.

Details: *Einsatz ebener Rotationen*

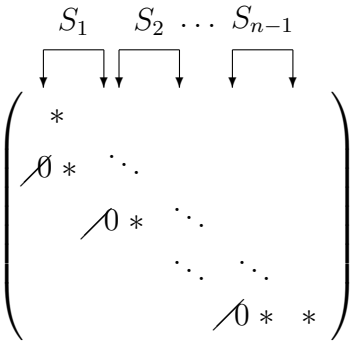
$$S_j = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos & \sin & & & \leftarrow & j \\ & & & -\sin & \cos & & & \leftarrow & j+1 \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \end{pmatrix}$$

a)

$$\begin{matrix} S_1^H \\ S_2^H \\ \vdots \\ S_{n-1}^H \end{matrix} \begin{pmatrix} a_{11}-\sigma_k & & & & \\ \not{*}\,0 & \ddots & & & \\ & \not{*}\,0 & \ddots & & \\ & & \ddots & \ddots & \\ & & & \not{*}\,0 & a_{nn}-\sigma_k \end{pmatrix} = R_k$$

$$A_k - \sigma_k I \;\rightsquigarrow\; S_{n-1}^H \cdot \ldots \cdot S_1^H \,(A_k - \sigma_k I) = R_k$$

b)



Wiederauffüllen der unteren Nebendiagonale:

$$R_k \;\rightsquigarrow\; R_k \underbrace{S_1 \cdot \ldots \cdot S_{n-1}}_{Q_k} = R_k Q_k$$

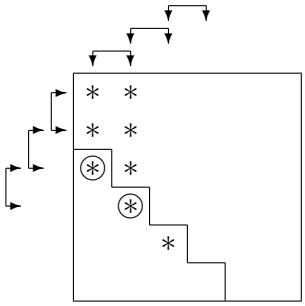
Wegen der Assoziativität der Matrixmultiplikation gilt:

$$R_k Q_k = \underbrace{\Big(S_{n-1}^H \ldots \big(\ldots \underbrace{\big(S_1^H (A_k - \sigma_k I) S_1 \big)}_{} \ldots \big) \ldots S_{n-1} \Big)}_{}.$$

Daher ergibt sich das sogenannte **Chasing:**

• 1 Extra-Element außerhalb der Subdiagonalen nach 1 Schritt von links nach rechts:

$$A_k - \sigma_k I \;\rightsquigarrow\; S_1^H (A_k - \sigma_k I) S_1$$



• Für den zweiten Schritt von links: (3,1)-Element kann als Bestimmung für den Drehwinkel genommen werden, so daß (3,1)-Element $\to 0$ wird.

Beachte: wegen $A_{k+1} - \sigma_k I = R_k Q_k$ folgt

$$A_{k+1} = S_{n-1}^H \cdot \ldots \cdot S_1^H A_k S_1 \cdot \ldots \cdot S_{n-1}.$$

Dies ermöglicht folgendes implizite Vorgehen:

- $B := S_1^H A_k S_1$ (mit S_1 wie oben definiert).
- Jage nun die beiden Elemente außerhalb der Subdiagonalen die Diagonale entlang:

  $$B \;\rightsquigarrow\; S_2^H B S_2 \;\rightsquigarrow\; \ldots \;\rightsquigarrow\; A_{k+1} = S_{n-1}^H \cdot \ldots \cdot S_2^H B S_2 \cdot \ldots \cdot S_{n-1}.$$

  Keine explizite Berechnung von $\pm\sigma_k I$ ist mehr nötig!

2.7 Bisektion zur Ermittlung einzelner Eigenwerte

In diesem Abschnitt geht es um die Berechnung einzelner Eigenwerte einer reellen, symmetrischen Tridiagonalmatrix

$$T = \begin{pmatrix} a_1 & b_2 & & 0 \\ b_2 & a_2 & . & \\ & . & . & b_n \\ 0 & & b_n & a_n \end{pmatrix}.$$

Natürlich kann man dazu auch das im vorigen Abschnitt diskutierte QR-Verfahren einsetzen. Wenn es aber nur um einige wenige EW geht, ist die im folgenden beschriebene Bisektionsmethode eine sehr gute Alternative.

Die Methode beruht auf dem Trägheitssatz von Sylvester:

Zwei symmetrische Matrizen A und B haben dann und nur dann gleichviele positive/negative/verschwindende Eigenwerte, wenn

$$A = S^T B S \quad \text{mit nichtsingulärem } S\,.$$

Die Trägheit ändert sich also nicht bei *Kongruenztransformationen*.

Damit liegt eine einfache Methode vor, um die Anzahl ν der EW $\lambda_1 \leq \ldots \leq \lambda_n$ der Tridiagonalmatrix T zu bestimmen, die kleiner sind als ein beliebiges $\lambda \in \mathbb{R}$. Man zerlegt

$$T - \lambda I = LR = LDL^T$$

mittels Gauß-Elimination in die Diagonalmatrix $D = \text{diag}\,(d_1, \ldots, d_n)$ und die normierte untere Dreiecks- bzw. Bidiagonalmatrix L. Wenn $T - \lambda I$ regulär ist, dann auch L, und der Satz von Sylvester ist anwendbar:

Anzahl $\nu(\lambda)$ der negativen Diagonalelemente (negative Pivots) d_i
= Anzahl der negativen Eigenwerte von $T - \lambda I$
= Anzahl der Eigenwerte von T kleiner als λ

Algorithmus 2.2 *Berechnung der negativen Pivots*
= Anzahl $\nu(\lambda)$ der EW von T kleiner als λ

$\nu := 0$; $d_1 := a_1 - \lambda$;
if $d_1 < 0 : \ \nu := 1$;
for $i := 2 : n$,
$\quad d_i := a_i - b_i^2/d_{i-1} - \lambda$;
$\quad$ *if* $d_i < 0 : \ \nu := \nu + 1$;

Folgende Anwendungen hat dieser Algorithmus:

a) Berechnung des k-ten Eigenwertes einer Matrix auf eine vorgegebene Genauigkeit. Man startet mit einem Intervall $[a, b]$, welches λ_k mit Sicherheit enthält ($\rightarrow$ Gerschgorin). Dann führt man aus

Algorithmus 2.3 *Bisektion zur Berechnung λ_k*

while $b - a > 2$ TOL:
$\quad \lambda := (a + b)/2$;
$\quad$ *if* $k > \nu(\lambda) : \ a := \lambda$
$\quad$ *else* $b := \lambda$
$\lambda := (a + b)/2$;

Mit einer Rückwärtsanalyse kann man zeigen, dass dieser Algorithmus auch unter dem Einfluss von Rundungsfehlern ausgezeichnete Ergebnisse liefert.

b) Berechnung der Häufigkeitsverteilung $m(i)$ der Eigenwerte einer sehr großen Matrix im Bereich $[a, b]$, unterteilt in die Intervalle $[a+(i-1)h, a+ih]$:

$h := (b - a)/N$; $k := \nu(a)$;
for $i := 1 : N$,
$\quad j := k$; $k := \nu(a + ih)$; $m(i) := k - j$;

2.8 Singulärwertzerlegung

Als Abschluss des Kapitels über Eigenwerte soll noch die Singulärwertzerlegung erwähnt werden. Sie stellt ein sehr nützliches Mittel zur Analyse von Matrizen (auch nichtquadratischer) dar und findet z.B. in der linearen Ausgleichsrechnung und bei der numerischen Berechnung der Kondition einer Matrix Anwendung.

Satz 2.8 *Singulärwertzerlegung*

Sei $A \in \mathbb{R}^{m\times n}$ eine beliebige relle Matrix. Dann gibt es orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$, so dass

$$U^T A V = \Sigma = diag\ (\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$$

wobei $p = min\ (m, n)$ und $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

Die σ_i sind die *Singulärwerte* der Matrix A. Ihre Quadrate sind gerade die Eigenwerte der Matrix $A^T A$.

Die Singulärwerte geben Auskunft über den Rang der Matrix A. Sind die ersten k Werte ungleich 0 und die restlichen gleich 0, so hat A den Rang k. Weiterhin ist $\sigma_1 = \|A\|_2$.

Bei der numerischen Berechnung der Singulärwertzerlegung wird das Produkt $A^T A$ jedoch nicht explizit geformt. Zunächst wird A mittels eines Reduktionsalgorithmus (Householder-Trafos, ähnlich oben) auf Bidiagonalgestalt gebracht. Danach wird der QR-Algorithmus in modifizierter Gestalt angewandt, so dass nur auf der Bidiagonalmatrix operiert wird. Bezüglich der Details wird auf die Literatur verwiesen.