Kapitel 6

Numerische Berechnung von Eigenwerten und Eigenvektoren

6.1 Einleitung

Im folgenden sei $A \in \mathbb{R}^{n,n}$ eine reelle quadratische Matrix. Das allgemeine Eigenwertproblem ist folgendes: Bestimme sämtliche komplexe Eigenwerte von A, und zu jedem Eigenwert λ eine Menge von Vektoren, welche eine Basis des Eigenraumes von λ bilden. Das Problem in dieser Allgemeinheit ist jedoch numerisch nur mit recht großem Aufwand zu lösen. Probleme treten z.B. bei nicht-reellen oder auch bei mehrfachen Eigenwerten auf:

Beispiel: Das Problem, einen mehrfachen Eigenwert einer Matrix zu finden, ist schlecht konditioniert: Die Matrix $A(\epsilon) = \left(\begin{smallmatrix} 0 & \epsilon \\ 1 & 0 \end{smallmatrix}\right)$, $\epsilon > 0$, hat die Eigenwerte $\lambda_\pm(\epsilon) = \pm\sqrt{\epsilon}$. $A(0)$ hat also den doppelten Eigenwert 0. Wenn man $A(\epsilon)$ als Störung von $A(0)$ betrachtet, so gilt für den Fehler von A, d.h. für den Fehler in den Daten

$$\|A(\epsilon) - A(0)\|_\infty = \epsilon,$$

aber für den Fehler in den Resultaten

$$|\lambda_\pm(\epsilon) - \lambda_\pm(0)| = \sqrt{\epsilon}.$$

Ein Fehler in den Daten von ϵ bewirkt also einen Fehler in den Resultaten von $\sqrt{\epsilon}$; für kleine ϵ ist das Verhältnis $\sqrt{\epsilon}/\epsilon = 1/\sqrt{\epsilon}$ sehr groß!

Wir werden in diesem Kapitel nur folgende Teilprobleme betrachten:

1. Bestimmung des betragsgrößten Eigenwertes einer Matrix und eines zugehörigen Eigenvektors, und

2. Bestimmung aller Eigenwerte einer symmetrischen Matrix.

6.2 Die Vektoriteration

Die Potenzmethode oder Vektoriteration ist ein sehr einfach durchzuführendes Verfahren zur Bestimmung des größten Eigenwertes einer Matrix. Sie hat den Vorteil, daß sie keine Manipulationen an der Matrix selber erfordern.

Sei $A \in \mathbb{R}^{n,n}$ mit n paarweise verschiedenen reellen Eigenwerten, die dem Betrage nach geordnet seien sollen: $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. In diesem Fall gibt es eine Basis $\{v_1, \ldots, v_n\}$ aus Eigenvektoren von A mit $Av_i = \lambda_i v_i$, $i = 1, \ldots, n$. Wir wählen nun ein $x_0 \in \mathbb{R}^n$, und stellen es bezüglich der Basis dar:

$$x_0 = \sum_{i=1}^{n} c_i v_i.$$

x_0 sei dabei so gewählt, daß

$$\|x_0\| = 1 \text{ und } c_1 \neq 0. \tag{6.1}$$

Für die Iteration $x_{k+1} = Ax_k$ gilt dann:

$$x_1 = \sum_{i=1}^{n} c_i \lambda_i v_i, \quad x_2 = \sum_{i=1}^{n} c_i \lambda_i^2 v_i, \quad \text{induktiv: } x_k = \sum_{i=1}^{n} c_i \lambda_i^k v_i.$$

Hieraus folgt

$$\frac{x_k}{\lambda_1^k} = c_1 v_1 + \sum_{i=2}^{n} c_i \frac{\lambda_i^k}{\lambda_1^k} v_i \to c_1 v_1 \tag{6.2}$$

für $k \to \infty$, denn $|\lambda_i/\lambda_1| < 1$ für $i \geq 2$. Diese Gleichung normieren wir wie folgt:

$$\frac{x_k}{\lambda_1^k} \Big/ \left\| \frac{x_k}{\lambda_1^k} \right\| = \frac{x_k}{\|x_k\|} \cdot \frac{|\lambda_1^k|}{\lambda_1^k} := \bar{x}_k \cdot \sigma_k \to \frac{c_1 v_1}{\|c_1 v_1\|} := \bar{v}_1. \tag{6.3}$$

Hierbei haben wir ausgenutzt, daß die Normierung $x \to x/\|x\|$ für $x \in \mathbb{R}^n \setminus \{0\}$ eine stetige Abbildung ist und man somit Limes und Normierung vertauschen kann. Für die Folge $\{\bar{x}_k\}$ gilt damit folgendes:

$$\bar{x}_k \cdot \sigma_k \to \bar{v}_1, \quad \text{also } A\bar{x}_k \cdot \sigma_k \to A\bar{v}_1 = \lambda_1 \bar{v}_1.$$

Durch Bilden des Skalarproduktes dieser beiden Gleichungen erhalten wir

$$\langle A\bar{x}_k \cdot \sigma_k, \bar{x}_k \cdot \sigma_k \rangle = \sigma_k^2 \langle A\bar{x}_k, \bar{x}_k \rangle = \langle A\bar{x}_k, \bar{x}_k \rangle \to \langle A\bar{v}_1, \bar{v}_1 \rangle = \lambda_1 \langle \bar{v}_1, \bar{v}_1 \rangle = \lambda_1 \|\bar{v}_1\|^2 = \lambda_1.$$

Wir fassen die bisherigen Ergebnisse in einem Satz zusammen:

Satz 6.1 (Vektoriteration) *Die Folgen $\{\bar{y}_k\} \subset \mathbb{R}^n$ und $\{\mu_k\} \subset \mathbb{R}$, definiert durch*

$$\bar{y}_0 = x_0 \textit{ aus } (6.1), \quad y_k = A\bar{y}_k, \quad \mu_k = \langle y_k, \bar{y}_k \rangle, \quad \bar{y}_{k+1} = \frac{y_k}{\|y_k\|} \tag{6.4}$$

haben folgende Konvergenzeigenschaften: $\|\bar{y}_k\| = 1$ für alle k, $\bar{y}_{2k} \to \bar{v}_1$, und $\mu_k \to \lambda_1$ für $k \to \infty$.

Beweis: Wir zeigen mit Induktion über k, daß $\bar{y}_k = \bar{x}_k$ für alle $k \geq 0$.
Für $k = 0$ ist $\bar{y}_0 = x_0 = \bar{x}_0$, die Aussage stimmt also. Sei nun die Behauptung für ein k bewiesen. Dann gilt nach Induktionsvoraussetzung

$$\bar{y}_{k+1} = \frac{y_k}{||y_k||} = \frac{A\bar{y}_k}{||A\bar{y}_k||} = \frac{A\bar{x}_k}{||A\bar{x}_k||} = \frac{A\frac{x_k}{||\bar{x}_k||}}{||A\frac{x_k}{||\bar{x}_k||}||} = \frac{Ax_k}{||Ax_k||} = \frac{x_{k+1}}{||x_{k+1}||} = \bar{x}_{k+1},$$

also insbesondere $||\bar{y}_k|| = 1$. Außerdem folgt

$$\mu_k = \langle y_k, \bar{x}_k \rangle = \langle A\bar{x}_k, \bar{x}_k \rangle \to \lambda_1.$$

Schließlich gilt $\sigma_{2k} = 1$, so daß nach (6.3) $\bar{y}_{2k} \to \bar{v}_1$. □

Bemerkung:

1. Die Folge $\{\bar{y}_k\}$ konvergiert genau dann gegen $\bar{v}_1$, wenn $\sigma_k = 1$, also $\lambda_1 > 0$. Für $\sigma_k = (-1)^k$, also $\lambda_1 < 0$, hat die Folge $\{\bar{y}_k\}$ die beiden Häufungspunkte $\pm\bar{v}_1$.

2. Die Bedingung $c_1 \neq 0$ ist für die Praxis nicht relevant. Selbst bei $c_1 = 0$ sorgen Rundungsfehler im Laufe der Iteration $x_{k+1} = Ax_k$ dafür, daß von Null verschiedene Koeffizienten von v_1 in der Basisdarstellung von x_k entstehen.

Beispiel: Die Matrix

$$A = \begin{pmatrix} 5 & -5 & 1 \\ 8 & -8 & 1 \\ 8 & -10 & 3 \end{pmatrix} \tag{6.5}$$

hat die Eigenwerte $\lambda_1 = -3$, $\lambda_2 = 2$, $\lambda_3 = 1$. Ein Eigenvektor zu λ_1 mit $||x|| = 1$ ist $\bar{y}_0 = (1/3, 2/3, 2/3)^T$. Die Iteration aus Satz 6.1 mit $x_0 = (1, 0, 0)^T$ liefert hierzu bei 4-stelliger Rechnung folgende Ergebnisse:

k	μ_k	$\bar{y}_{k+1}^T$
0	5.0	$(0.4042, 0.6467, 0.6467)$
1	-1.901	$(-0.2959, -0.6756, -0.6756)$
2	-3.554	$(0.3435, 0.6640, 0.6640)$
3	-2.846	$(-0.3296, -0.6677, -0.6677)$
4	-3.057	$(0.3346, 0.6663, 0.6663)$
5	-2.980	$(-0.3332, -0.6668, -0.6668)$
6	-3.003	$(0.3334, 0.6665, 0.6669)$

Konvergenzbeschleunigung durch Shift-Strategie

Wir betrachten noch einmal die Gleichung (6.2), welche grundlegend für die Konvergenz der Vektoriteration war. Da die Eigenwerte betragsmäßig geordnet sind, gilt:

$$\sum_{i=2}^{n} c_i \frac{\lambda_i^k}{\lambda_1^k} = O\left(\frac{\lambda_2^k}{\lambda_1^k}\right),$$

d.h. die Summe konvergiert wie $(\lambda_2/\lambda_1)^k$ gegen Null. Die Idee bei der Shift-Strategie ist, die Matrix A derart zu transformieren, daß der Quotient $|\lambda_2/\lambda_1|$ möglichst klein ist. Dazu dienen folgende einfache Beobachtungen: Falls σ kein Eigenwert von A ist, gilt

$$\begin{aligned} & x \in \mathbb{R}^n \text{ ist Eigenvektor von } A \text{ zum Eigenwert } \lambda \\ \Leftrightarrow & x \in \mathbb{R}^n \text{ ist Eigenvektor von } A-\sigma I \text{ zum Eigenwert } \lambda-\sigma \\ \Leftrightarrow & x \in \mathbb{R}^n \text{ ist Eigenvektor von } (A-\sigma I)^{-1} \text{ zum Eigenwert } (\lambda-\sigma)^{-1} \end{aligned}$$

Diese Eigenschaft macht man sich bei der *Shiftstrategie* zunutze.

Shift ohne Invertierung

Wir bestimmen ein $\sigma \in \mathbb{R}$, so daß

$$r(\sigma) := \max_{2 \le i \le n} \left| \frac{\lambda_i - \sigma}{\lambda_1 - \sigma} \right|$$

möglichst klein (jedenfalls $< r(0) < 1$) ist. Dann ist $\lambda_1 - \sigma$ der betragsgrößte Eigenwert von $A - \sigma I$. Wenn wir die Vektoriteration nun auf $A - \sigma I$ anwenden, können wir erwarten, daß die Iteration schneller (entsprechend $r(\sigma)$) gegen $\lambda_1 - \sigma$ konvergiert. Zur Berechnung von λ_1 anstatt $\lambda_1 - \sigma$ müssen wir einfach σ zur Folge $\langle y_k, \bar{x}_k \rangle$ hinzuaddieren. Man überlegt sich, daß die Folgen

$$\bar{y}_0 = x_0 \text{ aus (6.1)}, \quad y_k = (A - \sigma I)\bar{y}_k, \quad \mu_k = \langle y_k, \bar{y}_k \rangle + \sigma, \quad \bar{y}_{k+1} = \frac{y_k}{\|y_k\|} \tag{6.6}$$

dieselben Konvergenzeigenschaften wie (6.4) haben.

Beispiel: Für (6.5) erhalten wir $r(0) = 2/3$ und $r(1.5) = 1/9$. Wir können demnach erwarten, daß die Vektoriteration mit Shift $\sigma = 1.5$ um einiges schneller konvergiert als ohne Shift (d.h. mit $\sigma = 0$). Die folgende Tabelle bestätigt unsere Vermutung:

k	μ_k	$\bar{y}_{k+1}^T$
0	5.0	$(0.2956, 0.6757, 0.6757)$
1	-3.559	$(-0.3296, -0.6675, -0.6675)$
2	-3.054	$(0.3331, 0.6669, 0.6669)$
3	-3.006	$(-0.3333, -0.6666, -0.6666)$

Leider ist es schwierig, ohne weiteres einen geeigneten Shift-Parameter σ zu erhalten. Geeigneter ist die *inverse Shift-Strategie*:

Shift mit Invertierung

Wir nehmen an, daß wir schon eine gute Approximation σ von λ_1 haben (z.B. durch einige Schritte der Vektoriteration ohne Shift), jedenfalls so gut, daß $\lambda_1 - \sigma$ der *betragskleinste* Eigenwert von $A - \sigma I$ ist. Falls σ selber kein Eigenwert von A ist, dann ist $(\lambda_1 - \sigma)^{-1}$ der betragsgrößte Eigenwert von $(A - \sigma I)^{-1}$. Bezeichnen wir mit $\bar{\lambda}_i$ die ihrem Betrage nach geordneten Eigenwerte von $(A-\sigma I)^{-1}$, ist die Konvergenzgeschwindigkeit zur Berechnung von $\bar{\lambda}_1$ dann gegeben durch

$$\frac{|\bar{\lambda}_2|}{|\bar{\lambda}_1|} = \frac{|\bar{\lambda}_2|}{|(\lambda_1-\sigma)^{-1}|} = |\bar{\lambda}_2(\lambda_1-\sigma)|.$$

Für $\sigma \simeq \lambda_1$ können wir also erwarten, daß die inverse Iteration sehr schnell konvergiert. Zur Berechnung von λ_1 anstatt $(\lambda_1-\sigma)^{-1}$ müssen wir dann lediglich die Folge $\langle y_k, \bar{y}_k\rangle$ invertieren und σ hinzuaddieren. Die Definition der inversen Iteration ist also

$$\bar{y}_0 = x_0 \text{ aus (6.1)}, \quad y_k = (A-\sigma I)^{-1}\bar{y}_k, \quad \mu_k = \langle y_k, \bar{y}_k\rangle^{-1} + \sigma, \quad \bar{y}_{k+1} = \frac{y_k}{\|y_k\|} \tag{6.7}$$

Beispiel: Um dies zu demonstrieren, nehmen wir $x_0 = (-0.3296, -0.6677, -0.6677)^T$ als Startvektor und $\sigma = -2.846$ als Shift (das sind die Folgenglieder μ_k und $\bar{y}_{k+1}$ für $k = 3$ im Beispiel ohne Shift). Bei Weiterrechnung mit 8 Stellen mit inverser Iteration erhalten wir

k	μ_k	$\bar{y}_{k+1}^T$
0	-2.9977240	$(0.33348240, 0.66662942, 0.66662936)$
1	-3.0000897	$(-0.33332736, -0.66666817, -0.66666811)$

Bemerkung: Es ist bei der inversen Iteration nicht nötig, $A - \sigma I$ zu invertieren. Falls A symmetrisch ist, kann man z.B. die Cholesky-Zerlegung von $A - \sigma I$ in $O(n^3)$ Operationen berechnen. Jede Lösung des Gleichungssystems $(A - \sigma I)y_k = \bar{y}_k$ nach y_k erfordert dann nur noch $O(n^2)$ Operationen, also nicht mehr als bei einer Matrix-Vektor-Multiplikation. Für nichtsymmetrische Matrizen existieren ähnliche Zerlegungen, basierend auf dem Standard-Gauß-Algorithmus in Kapitel 1.

6.3 Das Jacobiverfahren für symmetrische Matrizen

Wir nehmen für diesen Abschnitt an, daß $A \in \mathbb{R}^{n,n}$ symmetrisch ist. Insbesondere sind dann alle Eigenwerte reell, und es existiert eine Transformation $D = UAU^T$, wobei D eine Diagonalmatrix ist mit den Eigenwerten von A auf der Diagonalen und U eine orthogonale Matrix mit einer Orthogonalbasis von Eigenvektoren von A als Spalten. Offensichtlich haben wir das Problem, alle Eigenwerte von A zu berechnen, gelöst, wenn wir ein solches D (was eindeutig bis auf Permutationen der Diagonalelemente ist) kennen. Das Jacobiverfahren liefert eine Folge $A = A_1, A_2, \ldots$ von Matrizen, die gegen ein solches D konvergiert.

Der Fall $n = 2$

Zur Herleitung des Jacobiverfahrens betrachten wir zunächst den Fall $n = 2$. In diesem Fall ist A von der Gestalt

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

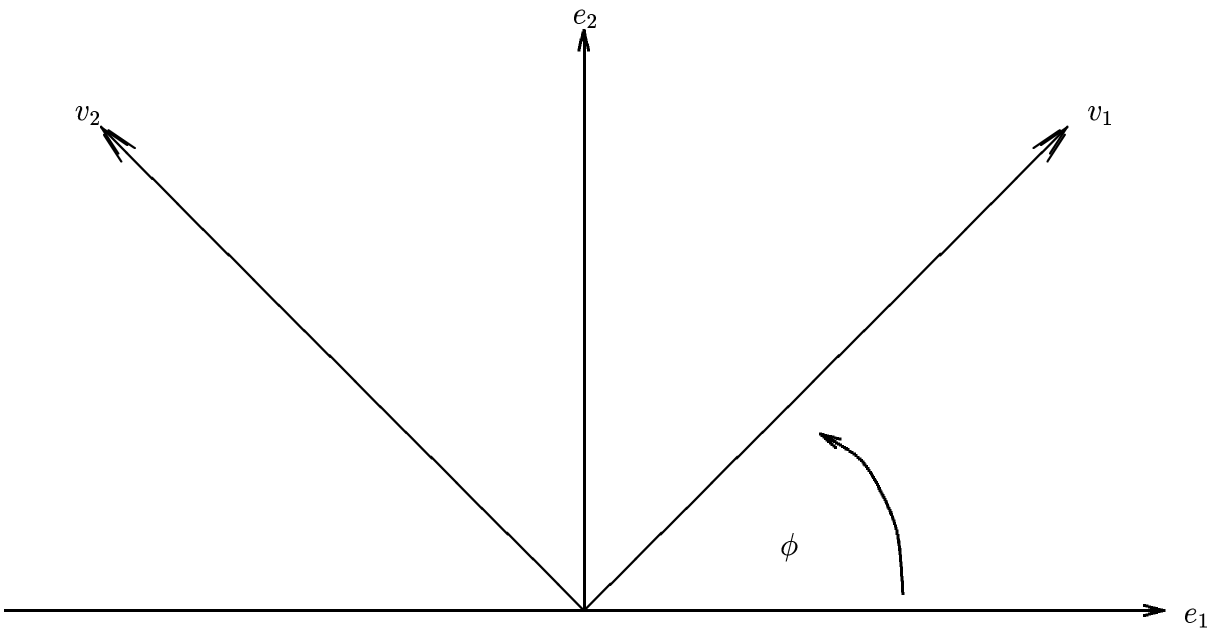

Abbildung 6.1: Schematische Darstellung der Eigenvektoren einer symmetrischen 2×2-Matrix

und die Lage der Eigenvektoren $\{v_1, v_2\}$ ist schematisch in Abb. 6.1 dargestellt. Die Idee zur Berechnung von D ist folgende: Wir drehen die kanonische Basis $\{e_1, e_2\}$ um einen geeigneten Winkel ϕ, sodaß sie mit den Eigenvektoren $\{v_1, v_2\}$ in Übereinstimmung gelangen. Dann wenden wir A an und drehen zurück. Bezeichnen wir die Matrix, die bei Verknüpfung dieser drei Operationen entstehen mit A', so läßt sich A' schreiben als

$$A' = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a & b \\ b & d \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \tag{6.8}$$

mit $c = \cos(\phi)$, $s = \sin(\phi)$. Sind λ_1, λ_2 die Eigenwerte von A zu den Eigenvektoren v_1, v_2, so gilt nach Konstruktion $A'e_1 = \lambda_1 e_1$, $A'e_2 = \lambda_2 e_2$, A' ist also eine Diagonalmatrix. Dieses Ergebnis können wir in eine Bedingung für ϕ umwandeln: Zunächst ist A' ebenfalls symmetrisch, da A' von der Form $A' = UAU^T$ mit A symmetrisch ist. Damit A' eine Diagonalmatrix ist, muß offenbar $a'_{1,2} = 0$ gelten. $a'_{1,2}$ kann man aber direkt aus (6.8) mit Matrixmultiplikation bestimmen, es gilt

$$a'_{1,2} = a'_{2,1} = sc(d-a) + (c^2 - s^2)b = 0,$$

was wir mit einigen trigonometrischen Identitäten umformen zu

$$\sin(2\phi)(a-d) = 2\cos(2\phi)b$$

und schließlich

$$\phi = \frac{1}{2}\arctan\left(\frac{2b}{a-d}\right) \quad \text{für } a \neq d.$$

Dabei setzen wir $\phi = +\pi/4$ oder $\phi = -\pi/4$ für $a = d$. Hat man ϕ auf diese Weise bestimmt, so erhalten wir den Rest von A' und damit die gesuchten Eigenwerte aus (6.8) durch

$$\begin{aligned} a'_{1,1} &= \lambda_1 = c^2 a + 2csb + s^2 d, \\ a'_{2,2} &= \lambda_2 = c^2 a - 2csb + s^2 d. \end{aligned}$$

Der Fall $n > 2$

Für allgemeines n wählen wir zwei Indizes $1 \leq p < q \leq n$ und betrachten die 2×2-Matrix

$$\tilde{A} = \begin{pmatrix} a_{p,p} & a_{p,q} \\ a_{q,p} & a_{q,q} \end{pmatrix}.$$

Hierauf können wir den Fall $n = 2$ anwenden, d.h. eine ebene Rotation bestimmen, sodaß $\tilde{A}$ auf eine Diagonalmatrix transformiert wird. Bei der vollen Matrix A entspricht dies einer Rotation in der $p-q$-Ebene. Natürlich können wir nicht erwarten, daß mit dieser Rotation *alle* Nichtdiagonaleinträge von A verschwinden, sondern nur $a'_{p,q} = a'_{q,p}$. Wenn wir mit $U_{p,q}(\phi)$ die entsprechende Rotationsmatrix bezeichnen, dann gilt

$$A' = U_{p,q}(\phi) A U_{p,q}(\phi)^T \tag{6.9}$$

mit

$$U_{p,q}(\phi) = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \quad \text{also} \quad (U_{p,q}(\phi))_{i,j} = \begin{cases} c: & i = j \wedge (i = p \vee i = q) \\ s: & i = p \wedge j = q \\ -s: & i = q \wedge j = p \\ 1: & i = j \wedge i \neq p \wedge i \neq q \\ 0: & \text{sonst} \end{cases}$$

und $c = \cos(\phi)$, $s = \sin(\phi)$. Wir rechnen jetzt die Komponenten von A' aus und leiten wie im Fall $n = 2$ aus der Bedingung $a'_{p,q} = a'_{q,p} = 0$ eine Bedingung für ϕ her. Zunächst überlegt man sich, daß allgemein für die Komponenten eines Matrixproduktes ABC aus drei $n \times n$-Matrizen gilt:

$$(ABC)_{i,j} = \sum_{k,l=1}^{n} a_{i,k} b_{k,l} c_{l,j}.$$

Für die Komponenten von A' gilt dann, wenn $u_{i,j}$ die Komponenten von $U_{p,q}(\phi)$ sind:

$$a'_{i,j} = \sum_{k,l=1}^{n} u_{i,k} u_{j,l} a_{k,l}.$$

Also gilt:

$$\begin{aligned}
a'_{p,p} &= \sum_{k,l=1}^{n} u_{p,k}u_{p,l}a_{k,l} = u_{p,p}^2 a_{p,p} + u_{p,p}u_{p,q}a_{p,q} + u_{p,q}u_{p,p}a_{q,p} + u_{p,q}^2 a_{q,q} \\
&= c^2 a_{p,p} + 2cs a_{p,q} + s^2 a_{q,q}, \qquad (6.10)\\
a'_{q,q} &= \sum_{k,l=1}^{n} u_{q,k}u_{q,l}a_{k,l} = u_{q,q}^2 a_{q,q} + u_{q,q}u_{q,p}a_{q,p} + u_{q,p}u_{q,q}a_{q,q} + u_{q,p}^2 a_{q,q} \\
&= s^2 a_{p,p} - 2cs a_{p,q} + c^2 a_{q,q}, \qquad (6.11)\\
a'_{p,q} = a'_{q,p} &= \sum_{k,l=1}^{n} u_{p,k}u_{q,l}a_{k,l} = u_{p,p}u_{q,p}a_{p,p} + u_{p,p}u_{q,q}a_{p,q} + u_{p,q}u_{q,p}a_{q,p} + u_{p,q}u_{q,q}a_{q,q} \\
&= (c^2 - s^2)a_{p,q} + cs(a_{q,q} - a_{p,p}), \qquad (6.12)\\
a'_{p,j} = a'_{j,p} &= \sum_{k,l=1}^{n} u_{p,k}u_{j,l}a_{k,l} = \sum_{k=1}^{n} u_{p,k}a_{k,j} = ca_{p,j} + sa_{q,j} \text{ für } j \neq p, q, \qquad (6.13)\\
a'_{q,j} = a'_{j,q} &= \sum_{k,l=1}^{n} u_{q,k}u_{j,l}a_{k,l} = \sum_{k=1}^{n} u_{q,k}a_{k,j} = -sa_{p,j} + ca_{q,j} \text{ für } j \neq p, q, \qquad (6.14)\\
a'_{i,j} &= a_{i,j} \text{ für } i, j \neq p, q. \qquad (6.15)
\end{aligned}$$

A' ist also außer in der p-ten und q-ten Zeile und Spalte mit A identisch.

Damit $a'_{p,q} = a'_{q,p} = 0$ gilt, muß also gelten

$$(c^2 - s^2)a_{p,q} + cs(a_{q,q} - a_{p,p}) = 0 \qquad (6.16)$$

woraus wir analog zum Fall $n = 2$ für $a_{p,p} \neq a_{q,q}$ erhalten

$$\phi = \frac{1}{2} \arctan\left(\frac{2a_{p,q}}{a_{p,p} - a_{q,q}}\right). \qquad (6.17)$$

Für $a_{p,p} = a_{q,q}$ setzen wir $\phi = \pi/4$. Umgekehrt sieht man, daß diese Wahl von ϕ $a'_{p,q} = a'_{q,p} = 0$ impliziert.

Die Jacobi-Iteration

Mit der berechneten Jacobi-Transformation ist man natürlich noch nicht am Ziel. Man hat nur erreicht, daß $a'_{p,q} = a'_{q,p} = 0$, die übrigen Einträge sind aber noch von Null verschieden, A' ist wahrscheinlich noch weit von der gewünschten Diagonalgestalt entfernt. Es ist nun naheliegend, ein weiteres Indexpaar auszuwählen und A' erneut zu transformieren: Man wählt geeignete neue p', q', berechnet ein neues ϕ' wie oben und $U_{p',q'}(\phi')$ und erhält als neue Transformierte $A'' = U_{p',q'}(\phi')A'U_{p',q'}(\phi')^T$, usw. So fortfahrend, erhält man eine Folge $\{p_k, q_k\}$ von Indizes, eine Folge $\{\phi_k\}$ von Rotationswinkeln, eine Folge $U_k = U_{p_k,q_k}(\phi_k)$ von orthogonalen Matrizen sowie eine Folge A_k von symmetrischen Matrizen mit

$$A_{k+1} = U_k A_k U_k^T.$$

Man hofft, daß auf diese Weise immer mehr Nicht-Diagonaleinträge der Folge $\{A_k\}$ verschwinden und man so schließlich zu einer Diagonalmatrix D gelangt. Da alle A_k zueinander ähnlich sind (die Jacobitransformation ist eine Ähnlichkeitstransformation und ändert somit die Eigenwerte nicht), stehen die Eigenwerte von $A = A_0$ am Schluß auf der Diagonalen von D.

Diese Hoffnung ist in der Tat zumindest teilweise berechtigt. Allerdings kommt man nicht mit einer endlichen Anzahl von Iterationen zur Diagonalgestalt, da im allgemeinen ein neuer Jacobischritt einen durch einen vorherigen Jacobischritt erzeugten Null-Eintrag wieder zunichte macht. Man kann aber zeigen, daß bei geeigneter Auswahl der p_k, q_k die Nichtdiagonaleinträge für $k \to \infty$ gegen Null konvergieren, A_k also immer mehr Diagonalgestalt annimmt. Wenn die Nichtdiagonaleinträge für ein k hinreichend klein sind, bricht man die Iteration ab und schätzt die Eigenwerte mit dem Satz von Gerschgorin (im nächsten Abschnitt) ab.

Bei der Auswahl des p_k, q_k gibt es mehrere Strategien. Beim *klassischen Jacobiverfahren* wählt man p_k, q_k derart, daß

$$|a^{(k)}_{p_k,q_k}| = \max_{1 \le p < q \le n} |a^{(k)}_{p,q}|, \tag{6.18}$$

d.h. man wählt das Indexpaar mit dem betragsgrößten Matrixeintrag. (Die Einträge von A_k werden mit $a^{(k)}_{i,j}$ bezeichnet.) Die klassische Jacobi-Iteration $A_k \to A_{k+1}$ ist also wie folgt definiert:

1. Bestimme $p_k < q_k$, so daß (6.18) gilt

2. Berechne

$$\phi_k = \begin{cases} \frac{1}{2} \arctan\left(\frac{2a^{(k)}_{p_k,q_k}}{a_{p_k,p_k} - a_{q_k,q_k}}\right) : & a_{p_k,p_k} \neq a_{q_k,q_k} \\ \frac{\pi}{4} : & \text{sonst} \end{cases} \tag{6.19}$$

3. Berechne A_{k+1} als die Jacobi-Transformation von A_k mit ϕ_k

Konvergenz der Jacobi-Iteration

Wir zeigen jetzt, daß das klassische Jacobiverfahren in folgendem Sinne konvergiert:

Satz 6.2 $\lim_{k\to\infty} a^{(k)}_{i,j} = 0$ *für* $i \neq j$.

Beweis: Es seien

$$s := \sum_{i \neq j} a^2_{i,j}, \quad s' := \sum_{i \neq j} (a'_{i,j})^2$$

die Summen der Quadrate der Nichtdiagonaleinträge von A bzw. A'. Nach (6.13)–(6.15) gilt für $j \neq p, q$:

$$(a'_{p,j})^2 + (a'_{q,j})^2 = (ca_{p,j} + sa_{q,j})^2 + (ca_{q,j} - sa_{p,j})^2 = (a_{p,j})^2 + (a_{q,j})^2.$$

Wenn wir mit

$$I := \{(i,j) \ : \ i \neq j, \ i \neq p, q, \ j \neq p, q\} \subset \{1, \ldots, n\} \times \{1, \ldots, n\}$$

die Menge der Indexpaare zu Matrixeinträgen außerhalb der Diagonalen sowie außerhalb der p-ten und q-ten Zeile und Spalte bezeichnen, dann gilt also

$$\begin{aligned} s' = \sum_{i \neq j} (a'_{i,j})^2 &= \sum_{(i,j) \in I} (a'_{i,j})^2 + 2 \sum_{j \neq p,q} (a'_{p,j})^2 + (a'_{q,j})^2 \\ &= \sum_{(i,j) \in I} a_{i,j}^2 + 2 \sum_{j \neq p,q} a_{p,j}^2 + a_{q,j}^2 \\ &= \sum_{i \neq j} a_{i,j}^2 - 2a_{p,q}^2. \end{aligned}$$

Für $s_k := \sum_{i \neq j} (a_{i,j}^{(k)})^2$ haben wir also gezeigt:

$$s_{k+1} = s_k - 2(a_{p_k,q_k}^{(k)})^2.$$

Wegen (6.18) gilt

$$s_k \leq (n^2 - n)(a_{p_k,q_k}^{(k)})^2,$$

also

$$s_{k+1} \leq s_k - 2\frac{s_k}{n^2 - n} = \left(1 - \frac{2}{n^2 - n}\right) s_k := q s_k$$

mit $0 \leq q < 1$. Induktiv folgt daraus: $s_k \leq q^k s_0$, also $\lim_{k \to \infty} s_k = 0$ und damit die Behauptung. □

Effiziente und stabile Durchführung der Jacobi-Rotation

Die Berechnung der Rotation in dieser Form ist rundungsfehleranfällig. Durch einige geschickte Umformungen kann dies aber verbessert werden. Dazu benutzen wir die folgenden trigonometrischen Identitäten:

$$\tan^2(\phi) + 2\tan(\phi)\cot(2\phi) - 1 = 0, \quad \cot^2(\phi) - 2\cot(\phi)\cot(2\phi) - 1 = 0, \tag{6.20}$$

$$\tan\left(\phi - \frac{\pi}{2}\right) = -\cot(\phi), \quad \cos(\phi) = \frac{1}{\sqrt{1 + \tan^2(\phi)}}, \quad \sin(\phi) = \tan(\phi)\cos(\phi). \tag{6.21}$$

Aus (6.16) erhalten wir hiermit

$$\sin(2\phi)(a_{p,p} - a_{q,q}) = 2\cos(2\phi)a_{p,q},$$

also

$$\cot(2\phi) = \frac{a_{p,p} - a_{q,q}}{2a_{p,q}} := \theta. \tag{6.22}$$

Nach (6.20) sind nun die beiden Nullstellen des Polynoms zweiten Grades $p(t) := t^2 + 2t\theta - 1$ gegeben durch

$$t_1 = \tan(\phi), \quad t_2 = -\cot(\phi) = \tan\left(\phi - \frac{\pi}{2}\right).$$

Wir wählen von diesen beiden die betragsmäßig kleinste Nullstelle und bezeichnen sie mit t_0; es gilt

$$t_0 = \frac{\operatorname{sign}(\theta)}{|\theta| + \sqrt{\theta^2 + 1}}. \tag{6.23}$$

(man vergewissere sich anhand von Abbildung 6.1, daß es egal ist, ob man um ϕ oder $\frac{\pi}{2} - \phi$ dreht.) Aus t_0 können wir nun c und s bestimmen:

$$c = \frac{1}{\sqrt{1 + t_0^2}}, \quad s = t_0 c. \tag{6.24}$$

Setzen wir noch

$$\tau := s/(1 + c), \tag{6.25}$$

so erhalten wir aus (6.10)-(6.15) und (6.22)–(6.25):

$$a'_{p,p} = a_{p,p} + t_0 a_{p,q}, \tag{6.26}$$
$$a'_{q,q} = a_{q,q} - t_0 a_{p,q}, \tag{6.27}$$
$$a'_{p,q} = a'_{q,p} = 0, \tag{6.28}$$
$$a'_{p,j} = a'_{j,p} = a_{p,j} + s(a_{q,j} - \tau a_{p,j}) \text{ für } j \neq p, q, \tag{6.29}$$
$$a'_{q,j} = a'_{j,q} = a_{q,j} - s(a_{p,j} + \tau a_{p,j}) \text{ für } j \neq p, q, \tag{6.30}$$
$$a'_{i,j} = a_{i,j} \text{ für } i, j \neq p, q. \tag{6.31}$$

Beispiel: Wir führen den ersten Jacobi-Schritt für die Matrix

$$A = A_0 = \begin{pmatrix} 8 & 5 & 0 & -2 \\ 5 & 7 & -2 & -1 \\ 0 & -2 & 2 & 2 \\ -2 & -1 & 2 & 6 \end{pmatrix}$$

durch. Das betragsgrößte Nichtdiagonal-Element ist $a_{1,2}$, es gilt also $p_0 = 1$, $q_0 = 2$. Daraus erhalten wir mit (6.22)–(6.25) bei 4-stelliger Rechnung:

$$\theta = 0.1, \quad t_0 = 0.905, \quad c = 0.7413, \quad s = 0.6709, \quad \tau = 0.3854$$

woraus sich mit (6.26)–(6.30)

$$A_1 = \begin{pmatrix} 12.53 & 0 & -1.342 & -2.154 \\ 0 & 2.475 & -1.483 & 0.6 \\ -1.342 & -1.483 & 2 & 2 \\ -2.154 & 0.6 & 2 & 6 \end{pmatrix}$$

ergibt. Bei Weiteriteration erhält man schließlich

$$A_{10} = \begin{pmatrix} 13.49 & 0 & -0.15\text{E-}5 & 0.2937\text{E-}3 \\ 0 & 3.545 & -0.3703\text{E-}3 & -0.1743\text{E-}5 \\ -0.15\text{E-}5 & -0.3703\text{E-}3 & 0.08509 & 0 \\ 0.2937\text{E-}3 & -0.1743\text{E-}5 & 0 & 5.89 \end{pmatrix}.$$

Der nun folgende Satz von Gerschgorin ermöglicht es uns, anhand von A_{10} die Eigenwerte von A recht genau abzuschätzen.

6.4 Der Satz von Gerschgorin zur Abschätzung von Eigenwerten

Satz 6.3 (Gerschgorin) *Es sei $A \in \mathbb{C}^{n,n}$ eine beliebige quadratische Matrix. Dann gilt:*

1. *Die Vereinigung $\bigcup_{i=1}^n B_i$ aller abgeschlossenen Kreise*
$$B_i := \left\{ \mu \in \mathbb{C} \ : \ |\mu - a_{i,i}| \le \sum_{j \ne i} |a_{i,j}| \right\}$$
enthält alle Eigenwerte von A.

2. *Ist $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ eine Indexmenge und*
$$M_1 := \bigcup_{i \in I} B_i, \quad M_2 := \bigcup_{i \in \{1,\ldots,n\} \setminus I} B_i, \quad M_1 \cap M_2 = \emptyset,$$
dann enthält M_1 genau k Eigenwerte und M_2 genau $n-k$ Eigenwerte von A, wenn man die Eigenwerte entsprechend ihrer algebraischen Vielfachheit im charakteristischen Polynom zählt.

Beweis:

1. Sei $\lambda \in \mathbb{C}$, $x \in \mathbb{C}^n$, $x \ne 0$ mit $Ax = \lambda x$. Wähle ein $i \in \{1, \ldots, n\}$ mit $\|x\|_\infty = \max_j \|x_j\| = |x_i|$. Hieraus erhalten wir
$$\begin{aligned} (Ax)_i &= \sum_{j=1}^n a_{i,j} x_j = a_{i,i} x_i + \sum_{j \ne i} a_{i,j} x_j \\ \Rightarrow & (\lambda - a_{i,i}) \frac{x_i}{\|x\|_\infty} = \sum_{j \ne i} a_{i,j} \frac{x_j}{\|x\|_\infty} \\ \Rightarrow & |\lambda - a_{i,i}| \le \sum_{j \ne i} |a_{i,j}| \end{aligned}$$

also $\lambda \in B_i$.

2. Wir betrachten für $0 \leq s \leq 1$ $A(s) := D + s(A - D)$, wobei D der Diagonalteil von A ist. $A(0)$ ist also die Diagonalmatrix mit der Diagonalen von A, und $A(1) = A$. Außerdem definieren wir

$$\begin{aligned} B_i(s) &:= \left\{ \mu \in \mathbb{C} \ : \ |\mu - a_{i,i}| \leq s \sum_{j \neq i} |a_{i,j}| \right\}, \\ M_1(s) &:= \bigcup_{i \in I} B_i(s), \quad M_2(s) := \bigcup_{i \in \{1,\ldots,n\} \setminus I} B_i(s), \\ \kappa_1(s) &:= \text{ Anzahl der algebraischen Eigenwerte in } M_1(s), \\ \kappa_2(s) &:= \text{ Anzahl der algebraischen Eigenwerte in } M_2(s). \end{aligned}$$

Die Idee ist, die Aussage für $A(0)$ (d.h. $\kappa_1(0) = k$ und $\kappa_2(0) = n - k$) zu zeigen und dann mit einem Stetigkeitsargument zu beweisen, daß für $0 < s \leq 1$ „nichts passieren“ kann. Dazu benutzen wir ohne vollständigen Beweis die folgende Tatsache: Für $i = 1, \ldots, n$ gibt es Kurven $\lambda_i(s) \subset \mathbb{C}$ mit $\lambda_i(0) = a_{i,i}$, die stetig von $s \in [0,1]$ abhängen, sodaß $\{\lambda_1(s), \ldots, \lambda_n(s)\}$ die Menge aller Eigenwerte von $A(s)$ ist. Zum vollständigen Beweis dieser Tatsache zeigt man zunächst mit dem Satz über implizite Funktionen, daß die Nullstellen eines Polynoms stetig von seinen Koeffizienten abhängen (wobei man allerdings eventuelle mehrfache Nullstellen gesondert behandeln muß). Die Tatsache folgt nun daraus, daß die Eigenwerte von $A(s)$ Nullstellen eines Polynoms sind, dessen Koeffizienten stetig von s abhängen, da $A(s)$ stetig von s abhängt.

Wir betrachten nun zunächst den Fall $s = 0$. Hierfür ist

$$B_i(0) = \{a_{i,i}\}, \quad M_1(0) = \bigcup_{i \in I} \{a_{i,i}\}, \quad M_2(0) = \bigcup_{i \in \{1,\ldots,n\} \setminus I} \{a_{i,i}\}, \quad \lambda_i(0) = a_{i,i}.$$

Offensichtlich gilt $\kappa_1(0) = k$ und $\kappa_2(0) = n - k$. Nach Voraussetzung gilt nun weiter: $M_1(1) \cap M_2(1) = \emptyset$, woraus auch $M_1(s) \cap M_2(s) = \emptyset$ für $0 \leq s \leq 1$ folgt, denn $M_1(s) \subset M_1(1)$ und $M_2(s) \subset M_2(1)$. Insbesondere gibt es eine Gerade $g \subset \mathbb{C}$, welche $M_1(s)$ und $M_2(s)$ für $0 \leq s \leq 1$ trennt. Wäre etwa $\kappa_1(1) \neq k$, dann müßte mindestens ein Eigenwert von M_1 nach M_2 wandern (oder von M_2 nach M_1), wobei er wegen der Stetigkeit g überqueren müßte. Anders ausgedrückt gäbe es ein $i \in \{1, \ldots, n\}$ und ein $s_0 \in [0,1]$ mit $\lambda_i(s_0) \in g$, also $\lambda_i(s_0) \notin M_1(s_0)$ und $\lambda_i(s_0) \notin M_2(s_0)$. Insbesondere gilt also: $\lambda_i(s_0) \notin B_j(s_0)$, $j = 1, \ldots, n$, woraus nach dem ersten Teil des Satzes folgt, daß $\lambda_i(s_0)$ kein Eigenwert von $A(s_0)$ sein kann, Widerspruch.

$\square$

Mit dem Satz von Gerschgorin erhalten wir folgende Abschätzung für die Eigenwerte von

A in obigem Beispiel:

$$\lambda_1 = 13.49 \pm 2.952\texttt{E-4}, \quad \lambda_2 = 5.89 \pm 2.954\texttt{E-4},$$
$$\lambda_3 = 3.545 \pm 3.720\texttt{E-4}, \quad \lambda_4 = 0.08509 \pm 3.718\texttt{E-4}.$$