Kapitel 4

Ausgleichsrechnung

Wir betrachten die ***Methode der kleinsten Quadrate*,** die bereits um 1800 von Gauß und unabhängig davon von Adrien Marie Legendre (1752-1833) entwickelt wurde.

Problem:

Gegeben seien Messdaten (t_i, y_i), $i = 1, ..., m$, diese müssten aus physikalischen oder ökonomischen Gründen alle auf einer Geraden liegen, welche durch

$$y = \alpha + \beta \cdot t \qquad (t \in \mathbb{R})$$

gegeben sei. Gesucht ist eine $''$vernünftige$''$ Näherung für α und β.

Beispiel: (Gewicht eines Neugeborenen)

Wir betrachten die Messdaten (t_n, g_n) der folgenden Tabelle:

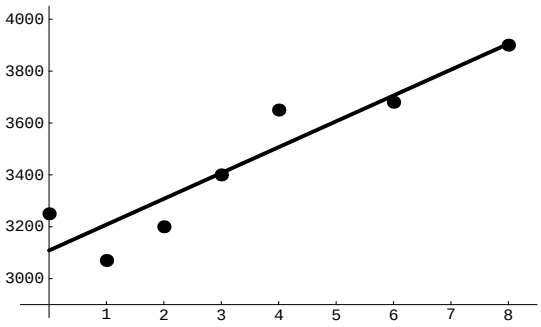
Zeit t (in Wochen)	0	1	2	3	4	6	8
Gewicht (in Gramm)	3250	3070	3200	3400	3650	3680	3900

Gesucht ist nun eine lineare Zunahme des Gewichtes. Durch Aufstellen einer Geradengleichung zu den jeweiligen Messzeitpunkten erhält man das folgende ***überbestimmte*** LGS:

$$\begin{aligned}
t = t_1 = 0: &\quad \alpha + \beta \cdot 0 &= 3250 = g_1 \\
t = t_2 = 1: &\quad \alpha + \beta \cdot 1 &= 3070 = g_2 \\
t = t_3 = 2: &\quad \alpha + \beta \cdot 2 &= 3200 = g_3 \\
t = t_4 = 3: &\quad \alpha + \beta \cdot 3 &= 3400 = g_4 \\
t = t_5 = 4: &\quad \alpha + \beta \cdot 4 &= 3650 = g_5 \\
t = t_6 = 6: &\quad \alpha + \beta \cdot 6 &= 3680 = g_6 \\
t = t_7 = 8: &\quad \alpha + \beta \cdot 8 &= 3900 = g_7
\end{aligned}$$

Es ergeben sich somit 7 Gleichungen für 2 Unbekannte, ein solches LGS ist im allgemeinen nicht lösbar. Gesucht ist nun eine $''$Ersatzlösung$''$, die das System $''$möglichst gut$''$ erfüllt: Wir betrachten die Residuen

$$\begin{aligned}
r_1 &= 3250 - \alpha - \beta \cdot 0 \\
r_2 &= 3070 - \alpha - \beta \cdot 1 \\
r_3 &= 3200 - \alpha - \beta \cdot 2 \\
r_4 &= 3400 - \alpha - \beta \cdot 3 \\
r_5 &= 3650 - \alpha - \beta \cdot 4 \\
r_6 &= 3680 - \alpha - \beta \cdot 6 \\
r_7 &= 3900 - \alpha - \beta \cdot 8
\end{aligned}$$



Ausgleichsgerade (siehe Beispiel)

Die ***Kleinste-Quadrate-Methode*** verfolgt folgende Idee: Man wähle α und β so, dass

$$Q(\alpha, \beta) \;=\; \sum_{n=1}^{7} r_n^2$$

minimal wird. Notwendiges Kriterium dafür ist, dass

$$\frac{\partial Q}{\partial \alpha} \;=\; \frac{\partial Q}{\partial \beta} \;=\; 0$$

gilt, also

$$\frac{\partial Q}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(\sum_{n=1}^{7}(g_n - \alpha - \beta \cdot t_n)^2\right) \;=\; 2 \cdot \sum_{n=1}^{7}(g_n - \alpha - \beta \cdot t_n) \cdot (-1) = 0$$

und

$$\frac{\partial Q}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\sum_{n=1}^{7}(g_n - \alpha - \beta \cdot t_n)^2\right) \;=\; 2 \cdot \sum_{n=1}^{7}(g_n - \alpha - \beta \cdot t_n) \cdot (-t_n) = 0.$$

Man erhält so die ***Gauß'schen Normalengleichungen***

$$\begin{aligned} \alpha \cdot \sum_{n=1}^{7} 1 + \beta \cdot \sum_{n=1}^{7} t_n &= \sum_{n=1}^{7} g_n, \\ \alpha \cdot \sum_{n=1}^{7} t_n + \beta \cdot \sum_{n=1}^{7} t_n^2 &= \sum_{n=1}^{7} t_n \cdot g_n. \end{aligned}$$

In diesem Beispiel ergibt sich explizit das LGS

$$\begin{bmatrix} 7 & 24 \\ 24 & 130 \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 24150 \\ 87550 \end{bmatrix},$$

das die Lösungen

$$\alpha \;=\; \frac{1038300}{334} \;\approx\; 3108,86 \qquad \beta \;=\; \frac{16625}{167} \;\approx 99,55$$

liefert. Der Wert für β entspricht dabei der durchschnittlichen wöchentlichen Gewichtszunahme des Babys (in Gramm).

Beachte:

Sogenannte ″Ausreißer″ (das heißt Messwerte, die im Vergleich mit den übrigen Messwerten deutlich von dem in etwa zu erwartenden Wert abweichen) haben großen Einfluss und verfälschen das Ergebnis. Deshalb werden diese bei der Berechnung der Ausgleichsgeraden oft einfach weggelassen. Berücksichtigt man dieses bei der erneuten Betrachtung des vorherigen Beispiels und vernachlässigt dort das Geburtsgewicht, also den ersten Messwert, so ergibt sich das LGS

$$\begin{bmatrix} 6 & 24 \\ 24 & 130 \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 20900 \\ 87550 \end{bmatrix},$$

das die Lösungen

$$\alpha = 3018,63 \qquad \beta = 116,18$$

liefert (und der Realität wesentlich näher kommt als die oben berechneten Werte!).

Allgemeine Form:

Bei gegebenen Messdaten (t_i, y_i), $i = 1, ..., m$ sind Werte für α und β gesucht, die eine ″möglichst gute″ Ausgleichsgerade

$$y = \alpha + \beta \cdot t \qquad (t \in \mathbb{R})$$

charakterisieren. Dazu ist das LGS

$$\begin{bmatrix} m & \sum_{i=1}^{m} t_i \\ \sum_{i=1}^{m} t_i & \sum_{i=1}^{m} t_i^2 \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} y_i \cdot t_i \end{bmatrix}$$

zu lösen. Definiert man

$$\underline{\mathbf{A}} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \in \mathbb{R}^{m \times 2}, \qquad \underline{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m,$$

so erhält man die Koeffizientenmatrix in der Form

$$\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} = \begin{bmatrix} m & \sum_{i=1}^{m} t_i \\ \sum_{i=1}^{m} t_i & \sum_{i=1}^{m} t_i^2 \end{bmatrix}$$

und die rechte Seite

$$\underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}} = \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} y_i \cdot t_i \end{bmatrix}.$$

Also ist das LGS

$$\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} \cdot \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}}$$

zu lösen; diese Gleichungen werden ***Gauß'sche Normalengleichungen*** genannt.

Verallgemeinerung:

Statt eines linearen Ansatzes

$$y = \alpha + \beta \cdot t$$

kann man zum Beispiel einen polynomialen Ansatz der Form

$$p_n(t) = \sum_{\nu=0}^{n} \alpha_\nu \cdot t^\nu$$

betrachten und die Koeffizienten α_ν so bestimmen, dass p_n "möglichst gut" durch die Punkte (t_i, y_i), $i = 1, ..., m$, verläuft. Sei dazu $m > n+1$, das heißt, man betrachtet ein ***überbestimmtes*** LGS der Form

$$\sum_{\nu=0}^{n} \alpha_\nu \cdot t_i^\nu = y_i \qquad (i = 1, ..., m).$$

Wir betrachten dann den Residuenvektor $\underline{\mathbf{r}} = (r_1, r_2, ..., r_m)^T$ mit

$$r_i = y_i - \sum_{\nu=0}^{n} \alpha_\nu \cdot t_i^\nu \qquad (i = 1, ..., m).$$

Ziel ist jetzt, die α_ν, $\nu = 1, ..., n$, so zu bestimmen, dass

$$Q(\alpha_0, ..., \alpha_n) = \sum_{i=1}^{m} r_i^2 = \sum_{i=1}^{m} \left(y_i - \sum_{j=0}^{n} \alpha_j \, t_i^j \right)^2$$

minimal wird. Notwendig dafür ist

$$\frac{\partial Q}{\partial \alpha_\nu} = 0 \qquad (\nu = 0, ..., n),$$

und man erhält analog zum linearen Fall das LGS

$$\begin{bmatrix} m & \sum\limits_{i=1}^{m} t_i & \cdots & \sum\limits_{i=1}^{m} t_i^n \\ \sum\limits_{i=1}^{m} t_i & \sum\limits_{i=1}^{m} t_i^2 & \cdots & \sum\limits_{i=1}^{m} t_i^{n+1} \\ \vdots & & & \vdots \\ \sum\limits_{i=1}^{m} t_i^n & \cdots & \cdots & \sum\limits_{i=1}^{m} t_i^{2n} \end{bmatrix} \cdot \begin{bmatrix} \alpha_0 \\ \vdots \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{m} y_i \\ \sum\limits_{i=1}^{m} t_i \cdot y_i \\ \vdots \\ \sum\limits_{i=1}^{m} t_i^n \cdot y_i \end{bmatrix}.$$

Mit

$$\underline{\mathbf{A}} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ \vdots & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \qquad \underline{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_m \end{bmatrix}, \qquad \underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \vdots \\ \vdots \\ \alpha_n \end{bmatrix}$$

ist das LGS äquivalent mit

$$\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\alpha} = \underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}},$$

das heißt, auch hier ergeben sich die Gauß'schen Normalengleichungen. Dabei war $\underline{\mathbf{A}} \cdot \underline{\alpha} = \underline{\mathbf{y}}$ das ursprüngliche (überbestimmte) LGS und $\underline{\mathbf{r}} = \underline{\mathbf{A}} \cdot \underline{\alpha} - \underline{\mathbf{y}}$ der Residuenvektor. Es ergibt sich das lineare Ausgleichsproblem

$$\sum_{i=1}^{m} r_i^2 \longrightarrow \min \qquad \Longleftrightarrow \qquad \|\underline{\mathbf{r}}\|_2^2 = \|\underline{\mathbf{A}} \cdot \underline{\alpha} - \underline{\mathbf{y}}\|_2^2 \longrightarrow \min . \qquad (*)$$

Satz 4.1:

Das lineare Ausgleichsproblem $(*)$ besitzt stets eine Lösung. Falls $\underline{\mathbf{A}}$ maximalen Rang besitzt, das heißt, falls $\mathrm{rg}(\underline{\mathbf{A}}) = n+1$, ist die Lösung eindeutig bestimmt und es gilt

$$\underline{\alpha} = (\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}})^{-1} \cdot \underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}}$$

Beweis:

1) Wir beweisen zunächst die ***Existenz*** einer Lösung. Sei dazu U der Untervektorraum des $\mathbb{R}^m$, der von den Spaltenvektoren von $\underline{\mathbf{A}}$ aufgespannt wird, es sei also

$$U = \mathrm{span}\left\{ \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix}, ..., \begin{bmatrix} t_1^n \\ \vdots \\ t_m^n \end{bmatrix} \right\}.$$

Dann ist

$$\underline{\mathbf{r}} = \underline{\mathbf{A}} \cdot \underline{\alpha} - \underline{\mathbf{y}} = \underbrace{\alpha_0 \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \alpha_1 \cdot \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix} + ... + \alpha_n \cdot \begin{bmatrix} t_1^n \\ \vdots \\ t_m^n \end{bmatrix}}_{:= \underline{\mathbf{a}} \in U} - \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

und das Minimierungsproblem $(*)$ ist äquivalent damit, ein $\underline{\mathbf{a}}^* \in U$ zu finden, so dass

$$\|\underline{\mathbf{a}}^* - \underline{\mathbf{y}}\|_2^2 = \min_{\underline{\mathbf{a}} \in U} \|\underline{\mathbf{a}} - \underline{\mathbf{y}}\|_2^2$$

gilt. Aus der Linearen Algebra ist bekannt, dass das Lot $\underline{\mathbf{n}}$ von $\underline{\mathbf{y}} \in \mathbb{R}^m$ auf $U \subseteq R^m$ existiert und eindeutig bestimmt ist und den Abstand $\|\underline{\mathbf{a}} - \underline{\mathbf{y}}\|_2^2$ für $\underline{\mathbf{a}} \in U$ minimiert. Sei $\underline{\mathbf{a}}^*$ ist die orthogonale Projektion von $\underline{\mathbf{y}}$ auf U, so dass

$$\underline{\mathbf{n}} = \underline{\mathbf{y}} - \underline{\mathbf{a}}^* \perp U$$

gilt (es ist also ein LGS zur Berechnung von $\underline{\mathbf{a}}^*$ zu lösen!).

2) Zur ***Eindeutigkeit*** der Lösung: Besitzt $\underline{\mathbf{A}}$ maximalen Rang, so sind die Vektoren

$$\underline{\mathbf{a}}_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, ..., \quad \underline{\mathbf{a}}_n = \begin{bmatrix} t_1^n \\ \vdots \\ t_m^n \end{bmatrix}$$

linear unabhängig. Für das Lot $\underline{\mathbf{n}} = \underline{\mathbf{y}} - \underline{\mathbf{a}}^*$ gilt

$$(\underline{\mathbf{y}} - \underline{\mathbf{a}}^*)^T \cdot \underline{\mathbf{a}} = 0 \qquad \forall \quad \underline{\mathbf{a}} \in U,$$

das heißt, es gilt $\underline{\mathbf{y}} - \underline{\mathbf{a}}^* \perp U$. Mit der Darstellung

$$\underline{\mathbf{a}}^* = \sum_{i=0}^{n} \alpha_i \cdot \underline{\mathbf{a}}_i$$

für $\underline{\mathbf{a}}^*$ folgt

$$\begin{aligned}
& \left(\sum_{i=0}^{n} \alpha_i \cdot \underline{\mathbf{a}}_i - \underline{\mathbf{y}} \right)^T \cdot \underline{\mathbf{a}}_j = 0 && (j = 0, ..., n) \\
\iff \quad & (\underline{\mathbf{A}} \cdot \underline{\alpha} - \underline{\mathbf{y}})^T \cdot \underline{\mathbf{a}}_j = 0 && (j = 0, ..., n) \\
\iff \quad & (\underline{\mathbf{A}} \cdot \underline{\alpha} - \underline{\mathbf{y}})^T \cdot \underbrace{[\underline{\mathbf{a}}_0 \ \ \underline{\mathbf{a}}_1 \ \ \cdots \ \ \underline{\mathbf{a}}_n]}_{\underline{\mathbf{A}}} = \underbrace{(0, ..., 0)}_{\underline{\mathbf{0}}^T} \\
\iff \quad & \underline{\alpha}^T \cdot \underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} - \underline{\mathbf{y}}^T \cdot \underline{\mathbf{A}} = \underline{\mathbf{0}}^T \\
\iff \quad & \underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\alpha} - \underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}} = \underline{\mathbf{0}} \\
\iff \quad & \underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\alpha} = \underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}}.
\end{aligned}$$

Dabei ist $\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}}$ invertierbar, denn: Sei

$$\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{0}},$$

und

$$\underline{\mathbf{u}} := \underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = x_0 \cdot \underline{\mathbf{a}}_0 + ... + x_n \cdot \underline{\mathbf{a}}_n \qquad \underline{\mathbf{x}} = (x_0, ..., x_n)^T.$$

Andererseits gilt

$$\underline{\mathbf{A}}^T \cdot (\underline{\mathbf{A}} \cdot \underline{\mathbf{x}}) = \underline{\mathbf{0}} \iff \underline{\mathbf{A}}^T \cdot \underline{\mathbf{u}} = \underline{\mathbf{0}} \iff \underline{\mathbf{a}}_i^T \cdot \underline{\mathbf{u}} = \underline{\mathbf{0}} \, (i = 0, ..., n) \iff \underline{\mathbf{u}} \perp U.$$

Damit muss $\underline{\mathbf{u}} = \underline{\mathbf{0}}$ gelten, denn $\underline{\mathbf{u}} \in U$ und $\underline{\mathbf{u}} \perp U$ lässt nur diese eine Möglichkeit zu. Da die Vektoren $\underline{\mathbf{a}}_0, ..., \underline{\mathbf{a}}_n$ linear unabhängig waren, muss damit die Linearkombination

$$\underline{\mathbf{u}} = x_0 \cdot \underline{\mathbf{a}}_0 + ... + x_n \cdot \underline{\mathbf{a}}_n$$

trivial sein, das heißt es muss

$$x_0 = x_1 = \cdots = x_n = 0$$

gelten. Damit ist aber $\underline{\mathbf{x}} = \underline{\mathbf{0}}$ und $(\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}})$ invertierbar. Die Lösung $\underline{\alpha}$ berechnet sich dann durch

$$\underline{\alpha} = (\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}})^{-1} \cdot \underline{\mathbf{A}}^T \cdot \underline{\mathbf{y}}$$

und ist eindeutig.

Beachte: Besitzt $\underline{\mathbf{A}}$ ***nicht*** maximalen Rang, so ist zwar das Lot auf U eindeutig bestimmt, aber die Koeffizienten α_i nicht, denn die Spalten von $\underline{\mathbf{A}}$ bilden keine Basis von U sondern nur ein Erzeugendensystem, das heißt, es sind unter Umständen verschiedene Darstellungen des Lotes möglich. □

Bemerkung:

Besitzt $\underline{\mathbf{A}}$ maximalen Rang, so ist $\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}}$ positiv definit und symmetrisch, also ist das cg-Verfahren anwendbar, auch das Cholesky-Verfahren ist unter Umständen sinnvoll.

Achtung:

Die Matrix $\underline{\mathbf{A}}^T \cdot \underline{\mathbf{A}}$ ist oft ***sehr schlecht konditioniert***! Deshalb gibt es spezielle orthogonale Verfahren zur Lösung des LGS, siehe Kapitel 5.3.

Kapitel 5

Matrizeneigenwertprobleme

Problem:

Bestimmung aller Eigenwerte und zugehörigen Eigenvektoren einer gegebenen Matrix $\underline{\mathbf{A}}$.

Theorie:

Die Eigenwerte sind die Nullstellen des charakteristischen Polynoms

$$p(\lambda) = \det\left(\underline{\mathbf{A}} - \lambda \cdot \underline{\mathbf{I}}\right) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} - \lambda \end{bmatrix},$$

das heißt, die Bestimmung der Eigenwerte macht zwei Berechnungsschritte nötig:

- Berechnung der Polynomkoeffizienten des charakteristischen Polynoms (Berechnung der obigen Determinante)
- Berechnung der Nullstellen des entstandenen Polynoms

Die durchzuführenden Berechnungen sind jedoch numerisch äußerst instabil! Wir wollen iterative Methoden zur Berechnung der Eigenwerte und Eigenvektoren herleiten, die die Probleme der Determinantenberechnung und der Nullstellenberechnung umgehen.

5.1 Vektoriteration

Die hier zunächst behandelte ***direkte Vektoriteration*** geht auf ***von Mises*** zurück.

Idee:

Betrachte die Folge $\{\underline{\mathbf{x}}_k\}$ mit

$$\underline{\mathbf{x}}_{k+1} = \underline{\mathbf{A}} \cdot \underline{\mathbf{x}}_k, \qquad \underline{\mathbf{x}}_0 \neq \underline{\mathbf{0}}, \qquad k = 0, 1, 2, ...$$

Ist λ ein einfacher Eigenwert von $\underline{\mathbf{A}}$, der betragsmäßig größer als alle anderen Eigenwerte von $\underline{\mathbf{A}}$ ist, so $''$setzt sich der zugehörige Eigenvektor durch$''$, das heißt, die Folge $\{\underline{\mathbf{x}}_k\}$ konvergiert gegen diesen Eigenvektor.

Beispiel:

Betrachte die Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} 4 & -1 \\ -5 & 0 \end{bmatrix},$$

diese besitzt das charakteristische Polynom

$$p(\lambda) = \det \begin{bmatrix} 4-\lambda & -1 \\ -5 & -\lambda \end{bmatrix} = \lambda^2 - 4\lambda - 5$$

und damit die Eigenwerte

$$\lambda_1 = 5, \qquad \lambda_2 = -1.$$

Für den betragsmäßig größten Eigenwert λ_1 ergibt sich der Eigenvektor

$$\underline{\mathbf{x}} = c \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad c \in \mathbb{R}.$$

Wählt man speziell $c = 1$ oder $c = -1$, so folgt $\|\underline{\mathbf{x}}\|_\infty = 1$ (Normierung). Setzt man beispielsweise

$$\underline{\mathbf{x}}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

so liefert obige Iterationsvorschrift folgende Tabelle (die Vektoren mit ″Schlange″ bezeichnen die bereits normierten Iterationsergebnisse):

$\underline{\mathbf{x}}_0$	$\underline{\mathbf{x}}_1$	$\underline{\tilde{\mathbf{x}}}_1$	$\underline{\mathbf{x}}_2$	$\underline{\tilde{\mathbf{x}}}_2$	$\underline{\tilde{\mathbf{x}}}_3$
$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 3 \\ -5 \end{bmatrix}$	$\begin{bmatrix} 0,6 \\ -1 \end{bmatrix}$	$\begin{bmatrix} 3,4 \\ -3 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -0,8824 \end{bmatrix}$	$\begin{bmatrix} -0,97647 \\ 1 \end{bmatrix}$

$\underline{\tilde{\mathbf{x}}}_4$	$\underline{\tilde{\mathbf{x}}}_5$	$\underline{\tilde{\mathbf{x}}}_6$
$\begin{bmatrix} 1 \\ -0,995 \end{bmatrix}$	$\begin{bmatrix} -0,99904 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -0,9998 \end{bmatrix}$

Satz 5.1:

Es sei $\underline{\mathbf{A}}$ diagonalisierbar, das heißt, $\underline{\mathbf{A}}$ besitze n linear unabhängige Eigenvektoren, und es gelte für die Eigenwerte $\lambda_1, ..., \lambda_n$

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|,$$

sei also λ_1 der betragsmäßig größte Eigenwert von $\underline{\mathbf{A}}$. Sei weiter $\underline{\mathbf{x}}_0 \in \mathbb{R}^n \setminus \{\underline{\mathbf{0}}\}$ und $\underline{\mathbf{x}}_0$ sei ***nicht*** in dem Untervektorraum enthalten, der von den Eigenvektoren zu den Eigenwerten $\lambda_2, ..., \lambda_n$ aufgespannt wird. Dann konvergiert die Folge

$$\underline{\tilde{\mathbf{x}}}_k = \frac{\underline{\mathbf{x}}_k}{\|\underline{\mathbf{x}}_k\|}$$

mit der Iterationsvorschrift

$$\underline{\mathbf{x}}_{k+1} = \underline{\mathbf{A}} \cdot \underline{\tilde{\mathbf{x}}}_k$$

gegen einen normierten Eigenvektor von $\underline{\mathbf{A}}$ zum Eigenwert λ_1.

Beweis:

Sei $\{\underline{\mathbf{y}}_1, ..., \underline{\mathbf{y}}_n\}$ eine Basis von Eigenvektoren von $\underline{\mathbf{A}}$ mit

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{y}}_i = \lambda_i \cdot \underline{\mathbf{y}}_i \qquad (i = 1, ..., n). \tag{$*$}$$

Für $\underline{\mathbf{x}}_0$ gibt es dann nach Voraussetzung eine Darstellung

$$\underline{\mathbf{x}}_0 = c_1 \cdot \underline{\mathbf{y}}_1 + c_2 \cdot \underline{\mathbf{y}}_2 + ... + c_n \cdot \underline{\mathbf{y}}_n$$

mit $c_1 \neq 0$. Damit folgt

$$\begin{aligned} \underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0 &= c_1 \cdot \underline{\mathbf{A}}^k \cdot \underline{\mathbf{y}}_1 + c_2 \cdot \underline{\mathbf{A}}^k \cdot \underline{\mathbf{y}}_2 + ... + c_n \cdot \underline{\mathbf{A}}^k \cdot \underline{\mathbf{y}}_n \\ &\overset{(*)}{=} c_1 \cdot \lambda_1^k \cdot \underline{\mathbf{y}}_1 + c_2 \cdot \lambda_2^k \cdot \underline{\mathbf{y}}_2 + ... + c_n \cdot \lambda_n^k \cdot \underline{\mathbf{y}}_n \end{aligned}$$

und wir erhalten

$$\frac{\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0}{c_1 \cdot \lambda_1^k} = \underline{\mathbf{y}}_1 + \frac{c_2}{c_1} \cdot \left(\frac{\lambda_2}{\lambda_1}\right)^k \cdot \underline{\mathbf{y}}_2 + ... + \frac{c_n}{c_1} \cdot \left(\frac{\lambda_n}{\lambda_1}\right)^k \cdot \underline{\mathbf{y}}_n.$$

Für den Grenzübergang $k \to \infty$ ergibt sich

$$\lim_{k\to\infty} \frac{\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0}{c_1 \cdot \lambda_1^k} = \underline{\mathbf{y}}_1 \qquad (\text{denn } \frac{|\lambda_i|}{|\lambda_1|} < 1 \text{ für } i \geq 2).$$

Wegen

$$\frac{\|\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0\|}{|c_1 \cdot \lambda_1^k|} = \underbrace{\left\| \pm \left[\underline{\mathbf{y}}_1 \pm \frac{c_2}{|c_1|} \cdot \left(\frac{\lambda_2}{|\lambda_1|}\right)^k \cdot \underline{\mathbf{y}}_2 \pm ... \pm \frac{c_n}{|c_1|} \cdot \left(\frac{\lambda_n}{|\lambda_1|}\right)^k \cdot \underline{\mathbf{y}}_n \right] \right\|}_{\overset{k\to\infty}{\longrightarrow} \|\underline{\mathbf{y}}_1\|}$$

folgt

$$\lim_{k\to\infty} \underbrace{\frac{\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0}{\|\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0\|}}_{\underline{\mathbf{A}}^k \cdot \tilde{\underline{\mathbf{x}}}} = \lim_{k\to\infty} \pm \frac{\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0}{c_1 \cdot \lambda_1^k} \cdot \frac{|c_1 \cdot \lambda_1^k|}{\|\underline{\mathbf{A}}^k \cdot \underline{\mathbf{x}}_0\|} = \pm \frac{\underline{\mathbf{y}}_1}{\|\underline{\mathbf{y}}_1\|}.$$

□

Bemerkung:

Problematisch ist unter Umständen die Wahl des Startvektors, da man vorher nicht weiß, ob der gewählte Vektor im Untervektorraum liegt, der von den Eigenvektoren zu den restlichen Eigenwerten aufgespannt wird. Nur dann liefert das Verfahren aber tatsächlich den gesuchten Eigenvektor. An dieser Stelle kommen uns jedoch die durch die Fließkomma-Arithmetik entstehenden numerischen Fehler zu Hilfe, so dass das Verfahren praktisch unabhängig vom Startvektor $\underline{\mathbf{x}}^{(0)}$ funktioniert.

Satz 5.2:

Die Matrix $\underline{\mathbf{A}}$ sei symmetrisch, besitze also ein vollständiges Orthonormalsystem von Eigenvektoren, und es bezeichne $(\lambda_i, \underline{\mathbf{y}}_i)$ ein Eigenwert-Eigenvektor-Paar. Ist $\underline{\mathbf{x}}$ eine Eigenvektor-Näherung der Form

$$\underline{\mathbf{x}} = \underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{r}}$$

mit $\|\underline{\mathbf{y}}_i\|_2 = 1$, $\|\underline{\mathbf{r}}\|_2 = 1$ und außerdem $\underline{\mathbf{r}} \perp \underline{\mathbf{y}}_i$, so liefert

$$\rho := \frac{\underline{\mathbf{x}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{x}}}{\underline{\mathbf{x}}^T \cdot \underline{\mathbf{x}}} = \lambda_i + \varepsilon^2 \cdot c$$

eine Näherung von λ_i (c ist eine Konstante). Der Ausdruck

$$\frac{\underline{\mathbf{x}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{x}}}{\underline{\mathbf{x}}^T \cdot \underline{\mathbf{x}}}$$

heißt ***Rayleigh-Quotient.***

Beweis:

Der Vektor $\underline{\mathbf{r}} \perp \underline{\mathbf{y}}_i$ besitze die Basisdarstellung

$$\underline{\mathbf{r}} = \sum_{\substack{j=1 \\ j \neq i}}^{n} c_j \cdot \underline{\mathbf{y}}_j.$$

Dann ergibt sich

$$\begin{aligned}
\rho &= \frac{(\underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{r}})^T \cdot \underline{\mathbf{A}} \cdot (\underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{r}})}{(\underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{r}})^T \cdot (\underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{r}})} \\
&= \frac{\underline{\mathbf{y}}_i^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{y}}_i + \varepsilon \cdot \underline{\mathbf{y}}_i^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{r}} + \varepsilon^2 \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{r}}}{\underline{\mathbf{y}}_i^T \cdot \underline{\mathbf{y}}_i + \underbrace{\varepsilon \cdot \underline{\mathbf{y}}_i^T \cdot \underline{\mathbf{r}}}_{=0} + \underbrace{\varepsilon \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{y}}_i}_{=0} + \varepsilon^2 \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{r}}} \\
&= \frac{\lambda_i \cdot \underline{\mathbf{y}}_i^T \cdot \underline{\mathbf{y}}_i + \overbrace{\varepsilon \cdot \lambda_i \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{y}}_i}^{=0} + \overbrace{\varepsilon \cdot \lambda_i \cdot \underline{\mathbf{y}}_i^T \cdot \underline{\mathbf{r}}}^{=0} + \varepsilon^2 \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{r}}}{\underbrace{\|\underline{\mathbf{y}}_i\|_2^2}_{=1} + \varepsilon^2 \cdot \underbrace{\|\underline{\mathbf{r}}\|_2}_{=1}} \\
&= \frac{\lambda_i \cdot \overbrace{\|\underline{\mathbf{y}}_i\|_2^2}^{=1} + \varepsilon^2 \cdot \underline{\mathbf{r}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{r}}}{1 + \varepsilon^2} \\
&= \frac{\lambda_i \cdot (1 + \varepsilon^2) + \varepsilon^2 \cdot (\underline{\mathbf{r}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{r}} - \lambda_i)}{1 + \varepsilon^2} \\
&= \lambda_i + \varepsilon^2 \cdot \underbrace{\frac{\underline{\mathbf{r}}^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{r}} - \lambda_i}{1 + \varepsilon^2}}_{:= c}.
\end{aligned}$$

Damit folgt die Behauptung. □

Probleme:

- Man erhält nur ***einen*** Eigenvektor, die übrigen bleiben weiter unbekannt.
- Die Konvergenzgeschwindigkeit ist unter Umständen sehr schlecht, sie ist abhängig vom Quotienten $\left|\frac{\lambda_2}{\lambda_1}\right|$ (vgl. oben).

MAPLE-Prozedur:

Die Vektoriteration lässt sich mit Hilfe der folgenden MAPLE-Prozedur implementieren:

```
> restart;
> with(LinearAlgebra);   printlevel:=0;
> vectoriteration:=proc(A, x, eps)
  local x0, x1, x2, nor, ew;
```

```
x0:=VectorScalarMultiply(x,1/Norm(x, infitity));        # Normierung von x
x1:= MatrixVectorMultiply(A, x0);
x1:=VectorScalarMultiply(x1,1/Norm(x1, infitity));      # Normierung von x1
while (Norm(VectorAdd(x0,x1,1,1), infinity) > eps) and
                        (Norm(VectorAdd(x0,x1,1,-1), infinity) > eps) do
   x2:= MatrixVectorMultiply(A, x1);
   x2:=VectorScalarMultiply(x2,1/Norm(x2, infitity));
   x0:= x1;
   x1:= x2;
end do;
print(x1);
# Eigenwertberechnung
nor:=DotProduct(x1, x1);    # quadrierte Euklidische Norm von x1
x2:=MatrixVectorMultiply(A, x1);
ew:=DotProduct(x1, x2)/nor;
print('Eigenwert', ew);
end proc;
```

Die Prozedur kann folgendermaßen aufgerufen werden:

```
A:= Matrix([[4., −1.], [−5., 0.]]);
x0:= Vector([1., 1.]);
eps:=0.001;
vectoriteration(A, x0, eps);
```

5.2 Inverse Vektoriteration

Die ***inverse Vektoriteration*** geht auf ***Wieland (1945)*** zurück.

Idee:

Sei $\overline{\lambda}$ eine Schätzung für den ***einfachen*** Eigenwert λ_i der Matrix $\underline{\mathbf{A}}$ und es gelte

$$|\overline{\lambda} - \lambda_i| \; < \; |\overline{\lambda} - \lambda_j| \qquad\qquad j \neq i.$$

(An dieser Stelle geht ein, dass λ_i einfacher Eigenwert sein muss, ansonsten funktioniert die Idee nicht!) Dann ist $(\lambda_i - \overline{\lambda})^{-1}$ der betragsmäßig größte Eigenwert von $(\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}})^{-1}$, denn es gilt

$$\det\left[(\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}}) - \mu \cdot \underline{\mathbf{I}}\right] \; = \; \det\left[(\underline{\mathbf{A}} - \underbrace{(\overline{\lambda} + \mu)}_{\lambda} \cdot \underline{\mathbf{I}}\right],$$

das heißt, ist λ_i Eigenwert von $\underline{\mathbf{A}}$, so ist $\mu_i \; = \; \lambda_i - \overline{\lambda}$ Eigenwert von $(\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}})$ und daher $\dfrac{1}{\mu_i} \; = \; \dfrac{1}{\lambda_i - \overline{\lambda}}$ Eigenwert von $(\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}})^{-1}$. Dabei ergibt sich die letzte Äquivalenz aus

$$(\underline{\mathbf{A}} - \overline{\lambda}\underline{\mathbf{I}}) \cdot \underline{\mathbf{x}} \; = \; \mu_i \cdot \underline{\mathbf{x}} \quad \Longleftrightarrow \quad \underline{\mathbf{x}} \; = \; \mu_i \cdot (\underline{\mathbf{A}} - \overline{\lambda}\underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{x}} \quad \Longleftrightarrow \quad \frac{1}{\mu_i} \cdot \underline{\mathbf{x}} \; = \; (\underline{\mathbf{A}} - \overline{\lambda}\underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{x}}.$$

Man wende jetzt die direkte Vektoriteration auf die Matrix $(\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}})^{-1}$ an, das heißt, man wählt einen Startvektor $\underline{\mathbf{x}}_0$ und berechnet

$$\underline{\mathbf{x}}_{k+1} \; = \; (\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{x}}_k$$

beziehungsweise

$$(\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}}) \cdot \underline{\mathbf{x}}_{k+1} \; = \; \underline{\mathbf{x}}_k,$$

In jedem Iterationsschritt muss also ein LGS gelöst werden, dies jedoch immer nur für verschiedene rechte Seiten und mit identischer Koeffizientenmatrix, so dass nach Berechnung

einer $\underline{\mathbf{L}}\cdot\underline{\mathbf{U}}$-Zerlegung der Matrix $\underline{\mathbf{A}}-\overline{\lambda}\cdot\underline{\mathbf{I}}$ dann pro Iterationsschritt jeweils nur eine Vorwärts- und eine Rückwärtselimination notwendig ist.

Beispiel:

Betrachte die Matrix

$$\underline{\mathbf{A}} = \begin{bmatrix} -1 & 3 \\ -2 & 4 \end{bmatrix}.$$

Diese besitzt die Eigenwerte $\lambda_1 = 2,\ \lambda_2 = 1$. Wir betrachten jetzt $\overline{\lambda} = 1.1$ als Schätzung von λ_1. Dann ergibt sich

$$\underline{\mathbf{A}} - \overline{\lambda}\cdot\underline{\mathbf{I}} = \begin{bmatrix} -1 & 3 \\ -2 & 4 \end{bmatrix} - \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix} = \begin{bmatrix} -2.1 & 3 \\ -2 & 2.9 \end{bmatrix}$$

und

$$\det\left(\underline{\mathbf{A}} - \overline{\lambda}\cdot\underline{\mathbf{I}}\right) = -0.09.$$

Damit lautet die Inverse zu $(\underline{\mathbf{A}} - \overline{\lambda}\cdot\underline{\mathbf{I}})$:

$$\left(\underline{\mathbf{A}} - \overline{\lambda}\cdot\underline{\mathbf{I}}\right)^{-1} = -\frac{1}{0.09}\cdot\begin{bmatrix} 2.9 & -3 \\ 2 & -2.1 \end{bmatrix}.$$

Mit dem Startvektor

$$\underline{\mathbf{y}}_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

erhält man im ersten Iterationsschritt

$$\underline{\mathbf{y}}_1 = -\frac{1}{0.09}\cdot\begin{bmatrix} 2.9 & -3 \\ 2 & -2.1 \end{bmatrix}\cdot\begin{bmatrix} 2 \\ 1 \end{bmatrix} = -\frac{1}{0.09}\cdot\begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix}.$$

Normierung liefert

$$\tilde{\underline{\mathbf{y}}}_1 = \frac{-\frac{1}{0.09}\cdot\begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix}}{\left\| -\frac{1}{0.09}\cdot\begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix} \right\|} = \frac{-\frac{1}{0.09}\cdot\begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix}}{\left| -\frac{1}{0.09} \right| \cdot \left\| \begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix} \right\|} = \frac{\begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix}}{\left\| \begin{bmatrix} 2.8 \\ 1.9 \end{bmatrix} \right\|}.$$

Durch Normierung wird also der Faktor

$$\frac{1}{\det\left(\underline{\mathbf{A}} - \overline{\lambda}\cdot\underline{\mathbf{I}}\right)}$$

herausgekürzt. Dadurch wird die Berechnung stabil. Wir erhalten so

$\tilde{\underline{\mathbf{y}}}_1$	$\tilde{\underline{\mathbf{y}}}_2$	$\tilde{\underline{\mathbf{y}}}_3$	$\tilde{\underline{\mathbf{y}}}_4$	$\tilde{\underline{\mathbf{y}}}_5$
$-\begin{bmatrix} 1 \\ 0.67857 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.665289 \end{bmatrix}$	$-\begin{bmatrix} 1 \\ 0.666819 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.6666497 \end{bmatrix}$	$-\begin{bmatrix} 1 \\ 0.6666685 \end{bmatrix}$

Nun ist $\tilde{\underline{\mathbf{y}}}_k$ eine Näherung für den Eigenvektor zum Eigenwert $(\lambda_i - \overline{\lambda})^{-1}$ der Matrix $(\underline{\mathbf{A}} - \overline{\lambda}\cdot\underline{\mathbf{I}})^{-1}$ und ***gleichzeitig*** Näherung für den Eigenvektor zum Eigenwert λ_i der Matrix $\underline{\mathbf{A}}$,

denn es gilt

$$\begin{aligned} & \underline{\mathbf{A}} \cdot \underline{\mathbf{y}} = \lambda_i \cdot \underline{\mathbf{y}} \\ \iff \quad & (\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}}) \cdot \underline{\mathbf{y}} = \underbrace{(\lambda_i - \overline{\lambda})}_{\mu_i} \cdot \underline{\mathbf{y}} \\ \iff \quad & \frac{1}{\lambda_i - \overline{\lambda}} \cdot \underline{\mathbf{y}} = (\underline{\mathbf{A}} - \overline{\lambda} \cdot \underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{y}}. \end{aligned}$$

Den zugehörigen Eigenwert kann man jetzt mit Hilfe der Matrix $\underline{\mathbf{A}}$ berechnen, es ist im Beispiel

$$\frac{(\tilde{\underline{\mathbf{y}}}_5)^T \cdot \underline{\mathbf{A}} \cdot \tilde{\underline{\mathbf{y}}}_5}{(\tilde{\underline{\mathbf{y}}}_5)^T \cdot \tilde{\underline{\mathbf{y}}}_5} = 0.9999992745$$

eine Näherung für λ_2.

Bemerkung:

Verfeinerte Methoden zur Eigenwertberechnung sind

- QR-Algorithmus
- Singulärwertzerlegung (siehe z.B. Hämmerlin/Hoffmann: Numerische Mathematik, Kapitel 2 und 3).

5.3 QR-Verfahren

Zur Berechnung der Eigenwerte einer Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$ lässt sich auch das QR-Verfahren verwenden, das auf einer Faktorisierung der Matrix $\underline{\mathbf{A}}$ in ein Produkt aus einer orthogonalen Matrix $\underline{\mathbf{Q}}$ und einer oberen Dreiecksmatrix $\underline{\mathbf{R}}$ beruht, d.h. $\underline{\mathbf{A}} = \underline{\mathbf{Q}}\,\underline{\mathbf{R}}$.
Eine Matrix $\underline{\mathbf{Q}}$ heißt dabei orthogonal, falls $\underline{\mathbf{Q}}^{-1} = \underline{\mathbf{Q}}^T$ gilt. Die n Spalten (Zeilen) einer orthogonalen Matrix $\underline{\mathbf{Q}}$ bilden ein ONS im $\mathbb{R}^n$, d.h. für die Spalten von $\underline{\mathbf{Q}} = (\underline{\mathbf{q}}_1, \ldots, \underline{\mathbf{q}}_n)$ gilt $\underline{\mathbf{q}}_i^T\, \underline{\mathbf{q}}_j = \sigma_{ij},\ i, j = 1, \ldots, n$. Eine QR-Faktorisierung von Matrizen ist aus folgendem Grund interessant.

Satz 5.3:

Die Kondition einer Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$ bzgl. der Spektralnorm und der Frobenius-Norm ändert sich nicht bei Multiplikation mit einer orthogonalen Matrix.

Beweis:

Ist $\underline{\mathbf{Q}} \in \mathbb{R}^{n \times n}$ orthogonal, so folgt für jedes $\underline{\mathbf{x}} \in \mathbb{R}^n$

$$\|\underline{\mathbf{Q}}\,\underline{\mathbf{x}}\|_2^2 = (\underline{\mathbf{Q}}\,\underline{\mathbf{x}})^T\, \underline{\mathbf{Q}}\,\underline{\mathbf{x}} = \underline{\mathbf{x}}^T\, \underline{\mathbf{Q}}^T\, \underline{\mathbf{Q}}\,\underline{\mathbf{x}} = \underline{\mathbf{x}}^T\, \underline{\mathbf{x}} = \|\underline{\mathbf{x}}\|_2^2.$$

Also folgt auch $\|\underline{\mathbf{Q}}\,\underline{\mathbf{A}}\,\underline{\mathbf{x}}\|_2 = \|\underline{\mathbf{A}}\ \underline{\mathbf{x}}\|_2$ und damit

$$\|\underline{\mathbf{Q}}\,\underline{\mathbf{A}}\|_2 = \max_{\underline{\mathbf{x}} \neq 0} \frac{\|\underline{\mathbf{Q}}\underline{\mathbf{A}}\underline{\mathbf{x}}\|_2}{\|\underline{\mathbf{x}}\|_2} = \max_{\underline{\mathbf{x}} \neq 0} \frac{\|\underline{\mathbf{A}}\underline{\mathbf{x}}\|_2}{\|\underline{\mathbf{x}}\|_2} = \|\underline{\mathbf{A}}\|_2.$$

Da mit $\underline{\mathbf{Q}}$ auch $\underline{\mathbf{Q}}^{-1} = \underline{\mathbf{Q}}^T$ orthogonal ist, folgt analog $\|(\underline{\mathbf{Q}}\underline{\mathbf{A}})^{-1}\|_2 = \|\underline{\mathbf{A}}^{-1}\underline{\mathbf{Q}}^{-1}\|_2 = \|\underline{\mathbf{A}}^{-1}\|_2$ und damit die Behauptung für die Spektralnorm. Analog lässt sich die Behauptung für die Frobenius-Norm zeigen. □

Zur Berechnung der QR-Zerlegung wenden wir das Householder-Verfahren an.

Definition 5.4:

Eine Matrix $\underline{\mathbf{H}} \in \mathbb{R}^{n\times n}$ der Form $\underline{\mathbf{H}} = \underline{\mathbf{I}} - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T$ mit der Einheitsmatix $\underline{\mathbf{I}} \in \mathbb{R}^{n\times n}$ und einem Vektor $\underline{\mathbf{u}} \in \mathbb{R}^n$ mit $\|\underline{\mathbf{u}}\|_2 = 1$ heißt **Householder Matrix**.

Wegen $\underline{\mathbf{H}}^T = (\underline{\mathbf{I}} - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T)^T = \underline{\mathbf{I}} - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T = \underline{\mathbf{H}}$ und

$$\begin{aligned}\underline{\mathbf{H}}^T\underline{\mathbf{H}} &= \underline{\mathbf{H}}\cdot\underline{\mathbf{H}} = (\underline{\mathbf{I}} - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T)(\underline{\mathbf{I}} - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T)\\ &= \underline{\mathbf{I}} - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T - 2\underline{\mathbf{u}}\,\underline{\mathbf{u}}^T + 4\underline{\mathbf{u}}\underbrace{\underline{\mathbf{u}}^T\underline{\mathbf{u}}}_{=1}\underline{\mathbf{u}}^T = \underline{\mathbf{I}}\end{aligned}$$

ist die Householder-Matrix symmetrisch und orthogonal. Wir betrachten nun zunächst folgendes Problem. Für einen gegebenen Vektor $\underline{\mathbf{a}} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ ist eine Householder-Matrix $\underline{\mathbf{H}}$ so zu bestimmen, so dass

$$\underline{\mathbf{H}}\,\underline{\mathbf{a}} = c\cdot\underline{\mathbf{e}}_1 = \begin{pmatrix} c\\ 0\\ \vdots\\ 0\end{pmatrix}$$

gilt. Wegen $\|\underline{\mathbf{H}}\underline{\mathbf{a}}\|_2^2 = \underline{\mathbf{a}}^T\underline{\mathbf{H}}^T\underline{\mathbf{H}}\,\underline{\mathbf{a}} = \underline{\mathbf{a}}^T\underline{\mathbf{a}} = \|\underline{\mathbf{a}}\|_2^2 = c^2\cdot\underbrace{\|\underline{\mathbf{e}}_1\|_2^2}_{=1}$ folgt $c = \pm\|\underline{\mathbf{a}}\|_2$.

Wir wählen nun

$$\underline{\mathbf{v}} := \underline{\mathbf{a}} + \underbrace{\operatorname{sign}(a_1)}_{:=\begin{cases}1 & a_1\geq 0\\ -1 & a_1<0\end{cases}}\cdot\|\underline{\mathbf{a}}\|_2\cdot\underline{\mathbf{e}}_1 \qquad \text{und} \qquad \underline{\mathbf{u}} := \frac{\underline{\mathbf{v}}}{\|\underline{\mathbf{v}}\|_2}.$$

Dann folgt

$$\underline{\mathbf{H}} = \underline{\mathbf{I}} - 2\cdot\frac{\underline{\mathbf{v}}\,\underline{\mathbf{v}}^T}{\|\underline{\mathbf{v}}\|_2^2} = \underline{\mathbf{I}} - \tilde{c}\cdot\underline{\mathbf{v}}\,\underline{\mathbf{v}}^T$$

mit

$$\begin{aligned}\tilde{c} &= \frac{2}{\|\underline{\mathbf{v}}\|_2^2} = 2\cdot\|\underline{\mathbf{a}} + \operatorname{sign}(a_1)\cdot\|\underline{\mathbf{a}}\|_2\cdot\underline{\mathbf{e}}_1\|_2^{-2} = 2\cdot\left\|\begin{pmatrix} a_1 + \operatorname{sign}(a_1)\|a\|_2\\ a_2\\ \vdots\\ a_n\end{pmatrix}\right\|_2^{-2}\\ &= \frac{2}{\|\underline{\mathbf{a}}\|_2^2 + \|\underline{\mathbf{a}}\|_2^2 + 2a_1\operatorname{sign}(a_n)\|\underline{\mathbf{a}}\|_2} = (\|\underline{\mathbf{a}}\|_2^2 + |a_1|\|\underline{\mathbf{a}}\|_2)^{-1}.\end{aligned}$$

Tatsächlich erhalten wir in diesem Fall

$$\begin{aligned}\underline{\mathbf{H}}\mathbf{a} &= (\underline{\mathbf{I}} - \tilde{c}\,\underline{\mathbf{v}}\,\underline{\mathbf{v}}^T)\underline{\mathbf{a}} = \underline{\mathbf{a}} - \tilde{c}\,\underline{\mathbf{v}}\underbrace{(\underline{\mathbf{a}} + \operatorname{sign}(a_1)\|\underline{\mathbf{a}}\|_2\cdot\underline{\mathbf{e}}_1)^T\underline{\mathbf{a}}}_{(\|\underline{\mathbf{a}}\|_2^2 + \operatorname{sign}(a_1)\cdot\|\underline{\mathbf{a}}\|_2\cdot a_1) = \tilde{c}^{-1}}\\ &= \underline{\mathbf{a}} - \underline{\mathbf{v}} = \operatorname{sign}(a_1)\|\underline{\mathbf{a}}\|_2\cdot\underline{\mathbf{e}}_1.\end{aligned}$$

Wir können nun zeigen:

Satz 5.5:

Jede Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ lässt sich als Produkt $\underline{\mathbf{A}} = \underline{\mathbf{Q}}\,\underline{\mathbf{R}}$ mit einer Orthogonalmatrix $\underline{\mathbf{Q}}$ und

einer oberen Dreiecksmatrix $\underline{\mathbf{R}}$ schreiben.

Beweis:

Wir führen den Beweis mittels vollständiger Induktion über n. Sei $\underline{\mathbf{A}} = (\underline{\mathbf{a}}_1, \ldots, \underline{\mathbf{a}}_n)$.
Für $n = 1$ folgt $a_{11} = 1 \cdot a_{11}$.
Für $n > 1$ wählen wir eine Householder-Matrix $\underline{\mathbf{H}} \in \mathbb{R}^{n \times n}$ mit $\underline{\mathbf{H}}\,\underline{\mathbf{a}}_1 = \pm\|\underline{\mathbf{a}}_1\|_2\,\underline{\mathbf{e}}_1$.
Dann folgt

$$\underline{\mathbf{H}}\,\underline{\mathbf{A}} = \begin{pmatrix} \pm\|a_1\|_2 & * & \ldots & * \\ 0 & & & \\ \vdots & & \underline{\tilde{\mathbf{A}}} & \\ 0 & & & \end{pmatrix}$$

mit $\underline{\tilde{\mathbf{A}}} \in \mathbb{R}^{(n-1)\times(n-1)}$. Die Induktionsvoraussetzung liefert eine QR-Zerlegung der Form $\underline{\tilde{\mathbf{A}}} = \underline{\tilde{\mathbf{Q}}}\,\underline{\tilde{\mathbf{R}}}$, d.h. $\underline{\tilde{\mathbf{Q}}}^T\,\underline{\tilde{\mathbf{A}}} = \underline{\tilde{\mathbf{R}}}$.
Setzen wir nun

$$\underline{\mathbf{H}}_1 = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & \underline{\tilde{\mathbf{Q}}}^T & \\ 0 & & & \end{pmatrix}$$

so folgt, dass $\underline{\mathbf{H}}_1\,\underline{\mathbf{H}}\,\underline{\mathbf{A}} = \underline{\mathbf{R}}$ eine obere Dreiecksmatix ist. □

Im QR-Algorithmus werden also sukzessive Householder-Matrizen immer kleinerer Dimensionen berechnet.

Beispiel:

1) Zu $\underline{\mathbf{a}} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$ ist eine Householder-Matrix $\underline{\mathbf{H}}$ gesucht mit

$$\underline{\mathbf{H}}\,\underline{\mathbf{a}} = \pm\|\underline{\mathbf{a}}\|_2 \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \pm 3 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Wähle

$$\underline{\mathbf{v}} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} + 1 \cdot 3 \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix} \qquad \text{und} \qquad \tilde{c} = \frac{2}{\|\underline{\mathbf{v}}\|_2^2} = \frac{2}{30} = \frac{1}{15}.$$

Wir erhalten

$$\underline{\mathbf{H}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{15} \begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix} (5, 2, 1) = \frac{1}{15} \begin{pmatrix} -10 & -10 & -5 \\ -10 & 11 & -2 \\ -5 & -2 & 14 \end{pmatrix}.$$

2) Berechne die QR-Zerlegung von

$$\underline{\mathbf{A}} = \begin{pmatrix} 3 & -9 & 7 \\ -4 & -13 & -1 \\ 0 & -20 & -35 \end{pmatrix}.$$

1. Schritt: Wir wählen

$$\underline{\mathbf{v}}_1 = \begin{pmatrix} 3 \\ -4 \\ 0 \end{pmatrix} + 1 \cdot 5 \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ -4 \\ 0 \end{pmatrix} \qquad \text{und} \qquad \tilde{c} = \frac{2}{80} = \frac{1}{40},$$

und erhalten die erste Householder-Matrix

$$\underline{\mathbf{H}}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{40} \begin{pmatrix} 64 & -32 & 0 \\ -32 & 16 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} -3 & 4 & 0 \\ 4 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}.$$

Multiplikation mit $\underline{\mathbf{A}}$ ergibt

$$\underline{\mathbf{H}}_1 \, \underline{\mathbf{A}} = \begin{pmatrix} -5 & -5 & -5 \\ 0 & -15 & 5 \\ 0 & -20 & -35 \end{pmatrix}.$$

2. Schritt: Wir wählen

$$\underline{\mathbf{v}}_2 = \begin{pmatrix} -15 \\ -20 \end{pmatrix} - 1 \cdot 25 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -40 \\ -20 \end{pmatrix} \qquad \text{und} \qquad \tilde{c} = \frac{2}{2000} = \frac{1}{1000},$$

und erhalten mit

$$\underline{\mathbf{v}}_2 \, \underline{\mathbf{v}}_2^T = \begin{pmatrix} 1600 & 800 \\ 800 & 400 \end{pmatrix}$$

die zweite Householder-Matrix

$$\underline{\mathbf{H}}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{1000} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1600 & 800 \\ 0 & 800 & 400 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.6 & -0.8 \\ 0 & -0.8 & 0.6 \end{pmatrix}.$$

Damit folgt

$$\underline{\mathbf{H}}_2 \, \underline{\mathbf{H}}_1 \, \underline{\mathbf{A}} = \begin{pmatrix} -5 & -5 & -5 \\ 0 & 25 & 25 \\ 0 & 0 & -25 \end{pmatrix} = \underline{\mathbf{R}}.$$

Also gilt $\underline{\mathbf{A}} = \underline{\mathbf{H}}_1^T \, \underline{\mathbf{H}}_2^T \, \underline{\mathbf{R}} = \underline{\mathbf{Q}} \, \underline{\mathbf{R}}$ mit

$$\underline{\mathbf{Q}} = \underline{\mathbf{H}}_1^T \underline{\mathbf{H}}_2^T = \begin{pmatrix} -0.6 & -0,48 & -0.64 \\ 0.8 & -0.36 & -0.48 \\ 0 & -0.8 & 0.6 \end{pmatrix}.$$

Algorithmus:

Die Matrix $\underline{\mathbf{Q}}$ wird durch die einzelnen Vektoren $\underline{\mathbf{v}}_1, \ldots, \underline{\mathbf{v}}_{n-1}$ der Householder-Matrizen festgelegt, wobei $\underline{\mathbf{v}}_j = (v_{jj}, \ldots, v_{nj})$ nur noch $n - j + 1$ Einträge hat. Bei der Speicherung der QR-Zerlegung wird daher statt $\underline{\mathbf{H}}_j = \underline{\mathbf{I}} - \tilde{c}_j \underline{\mathbf{v}}_j \, \underline{\mathbf{v}}_j^T$ nur der Vektor $\underline{\mathbf{v}}_j$ und der Wert $\tilde{c}_j$ abgespeichert. Man kann $\underline{\mathbf{v}}_j$ (bis auf v_{jj}) im unteren Teil der $\underline{\mathbf{A}}$-Matrix speichern, rechts oben speichern wir $\underline{\mathbf{R}}$. Dann müssen noch die Werte v_{jj} und die Werte $\tilde{c}_j$ in getrennten Vektoren $\underline{\mathbf{v}}$ und $\underline{\mathbf{c}}$ abgespeichert werden.

MAPLE-Prozedur:

Die QR-Zerlegung einer Matrix $\underline{\mathbf{A}}$ lässt sich mit Hilfe der folgenden MAPLE-Prozedur implementieren:

```
> restart;
> with(LinearAlgebra): printlevel:=0;
> qr_zerlegung:=proc(A)
  local n, j, nor, k, i, summe;
  global v, c;
  # option trace;
  n:=RowDimension(A);
  v:=Vector(n); c:=Vector(n);
  for j from 1 to n - 1 do
    nor:=0;
    for k from j to n do nor := nor + A[k, j] * A[k, j]; end do;
    nor:=sqrt(nor);
    if nor = 0 then v[j] := 0;
    else
      c[j] := 1.0/(nor * nor + nor * abs(A[j, j]));
      if A[j, j] < 0 then nor := -nor end if;
      v[j] := A[j, j] + nor;
      A[j, j] := -nor;
      for k from j + 1 to n do
        summe := v[j] * A[j, k];
        for i from j + 1 to n do summe := summe + A[i, j] * A[i, k]; end do;
        summe := summe * c[j];
        A[j, k] := A[j, k] - v[j] * summe;
        for i from j + 1 to n do A[i, k] := A[i, k] - A[i, j] * summe; end do;
      end do;
    end if;
  end do;
  print(A, v, c);
  end proc:
```

Anwendung:

```
> A:=Matrix([[3.,-9.,7.], [-4.,-13.,-1.],[0.,-20.,-35.]]);
> qr_zerlegung(A);
```

Wir erhalten das Resultat

$$\underline{\mathbf{A}} = \begin{pmatrix} -5 & -5 & -5 \\ -4 & 25 & 25 \\ 0 & -20 & -25 \end{pmatrix}, \qquad \underline{\mathbf{v}} = \begin{pmatrix} 8.00 \\ -40.00 \\ 0 \end{pmatrix}, \qquad \underline{\mathbf{c}} = \begin{pmatrix} 0.025 \\ 0.001 \\ 0 \end{pmatrix}.$$

Bemerkung:
Die QR-Zerlegung lässt sich auch auf Matrizen der Form $\underline{\mathbf{A}} \in \mathbb{R}^{m \times n}$ mit $m > n$ anwenden, und wir erhalten $\underline{\mathbf{A}} = \underline{\mathbf{Q}}\,\underline{\mathbf{R}}$, wobei

$$\underline{\mathbf{R}} = \begin{pmatrix} \underline{\mathbf{R}}_1 \\ \underline{\mathbf{0}} \end{pmatrix}$$

mit einer Dreiecksmatrix $\underline{\mathbf{R}}_1 \in \mathbb{R}^{n \times n}$ und einer Nullmatrix $\underline{\mathbf{0}} \in \mathbb{R}^{m-n \times n}$. Daher ist die QR-Zerlegung auch in der Ausgleichsrechnung anwendbar. Falls keine Pivotisierungsprobleme auftreten, folgt dann

$$\begin{aligned} \|\underline{\mathbf{A}}\,\underline{\mathbf{x}} - \underline{\mathbf{b}}\|_2 &= \|\underline{\mathbf{Q}}\,\underline{\mathbf{R}}\,\underline{\mathbf{x}} - \underline{\mathbf{b}}\|_2 = \|\underline{\mathbf{Q}}^T\,(\underline{\mathbf{Q}}\,\underline{\mathbf{R}}\,\underline{\mathbf{x}} - \underline{\mathbf{b}})\|_2 \\ &= \|\underline{\mathbf{R}}\,\underline{\mathbf{x}} - \underline{\mathbf{Q}}^T\,\underline{\mathbf{b}}\|_2 \to \min. \end{aligned}$$

Daher ist dann nur noch $\underline{\mathbf{R}}_1 \, \underline{\mathbf{x}} = (\underline{\mathbf{Q}}^T \, \underline{\mathbf{b}})_n$ zu berechnen, wobei $(\underline{\mathbf{Q}}^T \, \underline{\mathbf{b}})_n$ den Teilvektor von $\underline{\mathbf{Q}}^T \, \underline{\mathbf{b}}$ bezeichnet, der nur die ersten n Komponenten von $\underline{\mathbf{Q}}^T \, \underline{\mathbf{b}}$ enthält.

Anwendung des QR-Verfahrens zur Berechnung der Eigenwerte

Definition 5.6: (QR-Transformation)

Gegeben sei eine Matrix $\underline{\mathbf{A}}_0 := \underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$. Für $m = 0, 1, \ldots$ zerlege man $\underline{\mathbf{A}}_m$ in der Form $\underline{\mathbf{A}}_m = \underline{\mathbf{Q}}_m \, \underline{\mathbf{R}}_m$ mit einer orthogonalen Matrix $\underline{\mathbf{Q}}_m$ und einer oberen Dreiecksmatrix $\underline{\mathbf{R}}_m$, und bilde das vertauschte Produkt

$$\underline{\mathbf{A}}_{m+1} := \underline{\mathbf{R}}_m \, \underline{\mathbf{Q}}_m.$$

Wegen $\underline{\mathbf{Q}}_m^T \, \underline{\mathbf{A}}_m = \underline{\mathbf{R}}_m$ gilt dann

$$\underline{\mathbf{A}}_{m+1} = \underline{\mathbf{Q}}_m^T \, \underline{\mathbf{A}}_m \, \underline{\mathbf{Q}}_m,$$

d.h., alle $\underline{\mathbf{A}}_m$ haben dieselben Eigenwerte wie $\underline{\mathbf{A}}_0 = \underline{\mathbf{A}}$.

Satz 5.7:

Die Eigenwerte der Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n \times n}$ seien paarweise dem Betrage nach und von Null verschieden. Ferner gestatte die Inverse $\underline{\mathbf{T}}^{-1}$ der Matrix der (linear unabhängigen) Eigenvektoren von $\underline{\mathbf{A}}$ eine LU-Zerlegung (ohne Pivotisierung). Dann gilt für die mittels der QR-Transformation erzeugte Folge von Matrizen $\underline{\mathbf{A}}_m$:

- Die Diagonalelemente streben gegen die Eigenwerte von $\underline{\mathbf{A}}$. Diese sind dem Betrage nach geordnet.
- Die Elemente unter der Diagonale gehen gegen Null.
- Die Folgen $\{\underline{\mathbf{A}}_{2m}\}$ und $\{\underline{\mathbf{A}}_{2m+1}\}$ streben jeweils gegen obere Dreiecksmatrizen.

Ferner konvergiert die Folge der $\underline{\mathbf{Q}}_m$ gegen eine Diagonalmatrix, die nur 1 oder -1 auf der Diagonale hat.

Beweis:
R. Schaback, H. Wendland: Numerische Mathematik, 5. Auflage, Springer, 2005. □

5.4 Eigenwerteinschließungen

Wir betrachten nun zwei Verfahren, mit denen das Spektrum einer Matrix $\mathbf{A} \in K^{n \times n}$ (mit $K = \mathbb{R}$ oder $K = \mathbb{C}$) grob eingegrenzt werden kann. Solche Abschätzungen können zur Wahl geeigneter Startwerte für Iterationsverfahren ausgenutzt werden.

Satz 5.8: (Satz von Gerschgorin)

Sei $\underline{\mathbf{A}} = (a_{jk})_{j,k=1}^n \in K^{n \times n}$ und λ ein beliebiger Eigenwert von $\underline{\mathbf{A}}$. Dann gilt

$$\lambda \in \bigcup_{i=1}^n K_i = \bigcup_{i=1}^n \left\{ z : \; |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \right\}.$$

Die Kreise K_i heißen ***Gerschgorin-Kreise***.

Beweis:
Sei $\underline{\mathbf{A}}\underline{\mathbf{x}} = \lambda\underline{\mathbf{x}}$ mit $\underline{\mathbf{x}} = (x_1, \ldots, x_n)^T \neq \mathbf{0}$. Dann existiert eine Komponente x_i mit $|x_j| \leq |x_i|$ für alle $j \neq i$. Sei $(\underline{\mathbf{A}}\underline{\mathbf{x}})_i$ die i-te Komponente von $\underline{\mathbf{A}}\underline{\mathbf{x}}$ dann ist

$$\lambda x_i = (\underline{\mathbf{A}}\underline{\mathbf{x}})_i = \sum_{j=1}^{n} a_{ij} x_j,$$

und somit folgt

$$|\lambda - a_{ii}| = |\sum_{\substack{j=1 \\ j\neq 1}}^{n} a_{ij}\frac{x_j}{x_i}| \leq \sum_{\substack{j=1 \\ j\neq 1}}^{n} |a_{ij}|.$$

Also ist $\lambda \in K_i \subset \bigcup_{j=1}^{n} K_j$. □

Bemerkung:
Ist $\lambda \in S(\underline{\mathbf{A}})$ (S Spektrum von $\underline{\mathbf{A}}$), dann ist $\overline{\lambda} \in S(\underline{\mathbf{A}}^*)$ wobei

$$\underline{\mathbf{A}}^* = \begin{cases} \underline{\mathbf{A}}^T & K = \mathbb{R}, \\ \overline{\underline{\mathbf{A}}}^T & K = \mathbb{C}. \end{cases}$$

Tatsächlich gilt

$$\det(\underline{\mathbf{A}} - \lambda I) = 0 \quad \Rightarrow \quad \det(\overline{\underline{\mathbf{A}} - \lambda\underline{\mathbf{I}}})^T = 0 \quad \Rightarrow \quad \det(\overline{\underline{\mathbf{A}}}^T - \overline{\lambda}\underline{\mathbf{I}}) = 0.$$

Für $\overline{\lambda} \in S(\underline{\mathbf{A}}^*)$ gilt der Satz von Gerschgorin nun entsprechend, d.h.,

$$\overline{\lambda} \in \bigcup_{i=1}^{n} \left\{ z : |z - \overline{a_{ii}}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}| \right\}$$

bzw.

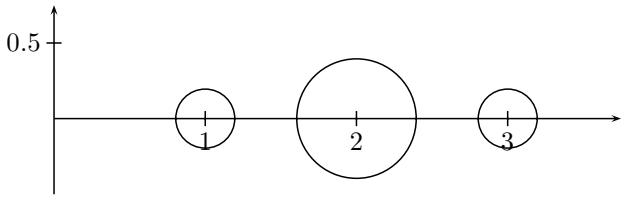
$$\lambda \in \bigcup_{i=1}^{n} \left\{ z : |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq 1}}^{n} |a_{ji}| \right\} = \bigcup_{i=1}^{n} K_i^*.$$

Beispiel:
Wir wollen die Eigenwerte der Matrix

$$\underline{\mathbf{A}} = \begin{pmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{pmatrix}$$

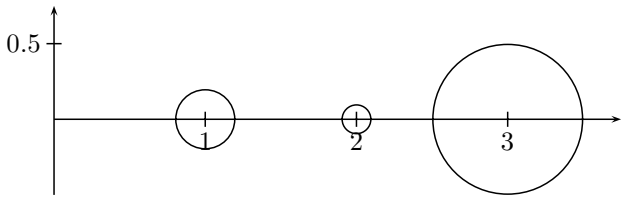
mit Hilfe des Satzes von Gerschgorin abschätzen. Wir erhalten die Gerschgorin-Kreise

$$\begin{aligned} K_1 &= \{\mu : |\mu - 1| \leq 0.2\}, \\ K_2 &= \{\mu : |\mu - 2| \leq 0.4\}, \\ K_3 &= \{\mu : |\mu - 3| \leq 0.2\}. \end{aligned}$$



Da $S(\underline{\mathbf{A}}) = S(\underline{\mathbf{A}}^T)$ wenden wir den Satz auch auf $\underline{\mathbf{A}}^T$ an und erhalten die Gerschgorin-Kreise

$$\begin{aligned} K_1^* &= \{\mu : |\mu - 1| \leq 0.2\}, \\ K_2^* &= \{\mu : |\mu - 2| \leq 0.1\}, \\ K_3^* &= \{\mu : |\mu - 3| \leq 0.5\}. \end{aligned}$$



Auf diese Weise läßt sich die erste Abschätzung des Eigenwerts in der Nähe von 2 noch verbessern!

Bemerkung:

Die Aussage des Satzes von Gerschgorin beinhaltet, dass strikt diagonaldominante Matrizen nicht singulär sind.

Weitere Einschließungsresultate für $S(\underline{\mathbf{A}})$ beruhen auf dem Konzept des Wertebereichs einer Matrix.

Definition 5.9:

Unter dem Wertebereich einer Matrix $\underline{\mathbf{A}} \in \mathbb{C}^{n\times n}$ versteht man die Menge aller Rayleigh-Quotienten $\frac{\overline{\underline{\mathbf{x}}}^T \underline{\mathbf{A}}\,\underline{\mathbf{x}}}{\overline{\underline{\mathbf{x}}}^T \underline{\mathbf{x}}}$ mit $\underline{\mathbf{x}} \in \mathbb{C}^n \setminus \{0\}$.

$$\begin{aligned} W(\underline{\mathbf{A}}) &:= \left\{ z = \frac{\overline{\underline{\mathbf{x}}}^T \underline{\mathbf{A}}\,\underline{\mathbf{x}}}{\overline{\underline{\mathbf{x}}}^T \underline{\mathbf{x}}} : \underline{\mathbf{x}} \in \mathbb{C}^n \setminus \{0\} \right\}, \\ &= \{ z = \overline{\underline{\mathbf{x}}}^T \underline{\mathbf{A}}\,\underline{\mathbf{x}} : \underline{\mathbf{x}} \in \mathbb{C}^n \setminus \{0\}, \|\underline{\mathbf{x}}\|_2 = 1 \}. \end{aligned}$$

Lemma 5.10:

(a) $W(\underline{\mathbf{A}})$ ist zusammenhängend.

(b) Ist $\underline{\mathbf{A}} \in \mathbb{C}^{n\times n}$ hermitesch, dann ist $W(\underline{\mathbf{A}})$ das reelle Intervall $[\lambda_n, \lambda_1]$, wobei λ_1 den größten und λ_n den kleinsten Eigenwert von $\underline{\mathbf{A}}$ bezeichnet.

(c) Ist $\underline{\mathbf{A}}$ schiefhermitesch, d.h. $\overline{\underline{\mathbf{A}}}^T = -\underline{\mathbf{A}}$, dann ist $W(\underline{\mathbf{A}})$ ein rein imaginäres Intervall, nämlich die konvexe Hülle aller Eigenwerte von $\underline{\mathbf{A}}$.

Beweis:

a) Liegen z_0 und $z_1 \neq z_0$ in $W(\underline{\mathbf{A}})$, dann existieren $\underline{\mathbf{x}}_0$, $\underline{\mathbf{x}}_1 \in \mathbb{C}^n \setminus \{0\}$ mit

$$z_0 = \frac{\overline{\underline{\mathbf{x}}}_0^T \underline{\mathbf{A}}\,\underline{\mathbf{x}}_0}{\overline{\underline{\mathbf{x}}}_0^T \underline{\mathbf{x}}_0}, \quad z_1 = \frac{\overline{\underline{\mathbf{x}}}_1^T \underline{\mathbf{A}}\,\underline{\mathbf{x}}_1}{\overline{\underline{\mathbf{x}}}_1^T \underline{\mathbf{x}}_1}.$$

Wegen $z_0 \neq z_1$ sind $\underline{\mathbf{x}}_0$ und $\underline{\mathbf{x}}_1$ linear unabhängig, so dass die Verbindungsstrecke

$$[\underline{\mathbf{x}}_0, \underline{\mathbf{x}}_1] := \{\underline{\mathbf{x}}_t = \underline{\mathbf{x}}_0 + t(\underline{\mathbf{x}}_1 - \underline{\mathbf{x}}_0) : \ t \in [0,1]\}$$

nicht den Nullpunkt enthält. Also ist

$$z_t := \frac{\overline{\underline{\mathbf{x}}}_t^T \underline{\mathbf{A}}\ \underline{\mathbf{x}}_t}{\overline{\underline{\mathbf{x}}}_t^T\ \underline{\mathbf{x}}_t}, \qquad 0 \leq t \leq 1,$$

eine stetige Kurve in $W(\underline{\mathbf{A}})$ die z_0 und z_1 verbindet.

b) Sei $\underline{\mathbf{A}}$ hermitesch. Dann ist $W(\underline{\mathbf{A}})$ reell, denn $\overline{\underline{\mathbf{x}}}^T\underline{\mathbf{x}} = \|\underline{\mathbf{x}}\|^2 > 0$ ist reell und

$$\overline{\overline{\underline{\mathbf{x}}}^T\,\underline{\mathbf{A}}\,\underline{\mathbf{x}}} = \underline{\mathbf{x}}^T\overline{\underline{\mathbf{A}}}\,\overline{\underline{\mathbf{x}}} = (\underline{\mathbf{x}}^T\overline{\underline{\mathbf{A}}}\,\overline{\underline{\mathbf{x}}})^T = \overline{\underline{\mathbf{x}}}^T\underbrace{\overline{\underline{\mathbf{A}}}^T}_{=\underline{\mathbf{A}}}\underline{\mathbf{x}} = \overline{\underline{\mathbf{x}}}^T\underline{\mathbf{A}}\,\underline{\mathbf{x}}.$$

Wegen a) ist $W(\underline{\mathbf{A}})$ also ein abgeschlossenes reelles Intervall. Wir berechnen den maximalen Wert in $W(\underline{\mathbf{A}})$:
Wähle dazu ein $\alpha > 0$ so, dass $\underline{\mathbf{A}} + \alpha\,\underline{\mathbf{I}}$ positiv definit ist. Dann existiert eine Cholesky-Zerlegung

$$\underline{\mathbf{A}} + \alpha\,\underline{\mathbf{I}} = \underline{\mathbf{G}}\cdot\overline{\underline{\mathbf{G}}}^T,$$

und für $\|\underline{\mathbf{x}}\|_2 = 1$ gilt

$$\overline{\underline{\mathbf{x}}}^T\underline{\mathbf{A}}\;\underline{\mathbf{x}} = \overline{\underline{\mathbf{x}}}^T(\underline{\mathbf{A}} + \alpha\underline{\mathbf{I}})\underline{\mathbf{x}} - \underbrace{\overline{\underline{\mathbf{x}}}^T\alpha\underline{\mathbf{I}}\,\underline{\mathbf{x}}}_{=\alpha} = \overline{\underline{\mathbf{x}}}^T\underline{\mathbf{G}}\,\overline{\underline{\mathbf{G}}}^T\underline{\mathbf{x}} - \alpha = \|\overline{\underline{\mathbf{G}}}^T\underline{\mathbf{x}}\|_2^2 - \alpha.$$

Also folgt

$$\begin{aligned}\max W(\underline{\mathbf{A}}) &= \max_{\|\underline{\mathbf{x}}\|_2=1}\overline{\underline{\mathbf{x}}}^T\underline{\mathbf{A}}\,\underline{\mathbf{x}} = \max_{\|\underline{\mathbf{x}}\|=1}\|\overline{\underline{\mathbf{G}}}^T\underline{\mathbf{x}}\|_2^2 - \alpha\\ &= \rho(\underline{\mathbf{G}}\cdot\overline{\underline{\mathbf{G}}}^T) - \alpha = \rho(\underline{\mathbf{A}} + \alpha\underline{\mathbf{I}}) - \alpha = \lambda_1,\end{aligned}$$

denn λ ist ein Eigenwert von $\underline{\mathbf{A}}$ genau dann, wenn $\lambda + \alpha$ ein Eigenwert von $\underline{\mathbf{A}} + \alpha\underline{\mathbf{I}}$ ist:

$$\begin{aligned}\underline{\mathbf{A}}\,\underline{\mathbf{x}} = \lambda\,\underline{\mathbf{x}} \quad &\Leftrightarrow \quad \underline{\mathbf{A}}\,\underline{\mathbf{x}} + \alpha\underline{\mathbf{x}} = (\lambda+\alpha)\underline{\mathbf{x}}\\ &\Leftrightarrow \quad (\underline{\mathbf{A}} + \alpha\underline{\mathbf{I}})\,\underline{\mathbf{x}} = (\lambda+\alpha)\,\underline{\mathbf{x}}.\end{aligned}$$

Durch Übergang von $W(\underline{\mathbf{A}})$ zu $W(-\underline{\mathbf{A}})$ beweist man entsprechend, dass der linke Endpunkt von $W(\underline{\mathbf{A}})$ der kleinste Eigenwert λ_n von $\underline{\mathbf{A}}$ ist.

c) Wegen $\overline{\underline{\mathbf{A}}}^T = -\underline{\mathbf{A}}$ ist $i\underline{\mathbf{A}}$ hermitesch, denn $\overline{(i\,\underline{\mathbf{A}})}^T = \overline{i}\,\overline{\underline{\mathbf{A}}}^T = -i\,\overline{\underline{\mathbf{A}}}^T = i\,\underline{\mathbf{A}}$. Außerdem gilt $W(i\underline{\mathbf{A}}) = iW(\underline{\mathbf{A}})$ und $S(i\underline{\mathbf{A}}) = iS(\underline{\mathbf{A}})$. Die Behauptung folgt nun aus Teil b). □

Bemerkung:
Jede Matrix $\underline{\mathbf{A}} \in \mathbb{C}^{n\times n}$ lässt sich in eine hermitesche Matrix $\frac{1}{2}(\underline{\mathbf{A}} + \overline{\underline{\mathbf{A}}}^T)$ und eine schiefhermitesche Matrix $\frac{1}{2}(\underline{\mathbf{A}} - \overline{\underline{\mathbf{A}}}^T)$ zerlegen,

$$\underline{\mathbf{A}} = \frac{\underline{\mathbf{A}} + \overline{\underline{\mathbf{A}}}^T}{2} + \frac{\underline{\mathbf{A}} - \overline{\underline{\mathbf{A}}}^T}{2}.$$

Satz 5.11: (Satz von Bendixson)

Das Spektrum von $\underline{\mathbf{A}} \in \mathbb{C}^{n\times n}$ ist in dem Rechteck

$$R = W\left(\frac{\underline{\mathbf{A}} + \underline{\mathbf{A}}^*}{2}\right) + W\left(\frac{\underline{\mathbf{A}} - \underline{\mathbf{A}}^*}{2}\right)$$

enthalten. Dabei bezeichnet für Mengen $W_1, W_2 \subset \mathbb{C}^n$ die Summe $W_1 + W_2$ die Menge $\{a + b : a \in W_1, b \in W_2\}$.

Beweis:

Wir zeigen, dass $W(\underline{\mathbf{A}}) \subset R$. Für $\underline{\mathbf{x}} \in \mathbb{C}^n$ mit $\|\underline{\mathbf{x}}\|_2 = 1$ gilt

$$\begin{aligned}\overline{\underline{\mathbf{x}}}^T \underline{\mathbf{A}}\,\underline{\mathbf{x}} &= \overline{\underline{\mathbf{x}}}^T \left(\frac{\underline{\mathbf{A}} + \overline{\underline{\mathbf{A}}}^T}{2} + \frac{\underline{\mathbf{A}} - \overline{\underline{\mathbf{A}}}^T}{2} \right) \underline{\mathbf{x}} \\ &= \overline{\underline{\mathbf{x}}} \left(\frac{\underline{\mathbf{A}} + \overline{\underline{\mathbf{A}}}^T}{2} \right) \underline{\mathbf{x}} + \overline{\underline{\mathbf{x}}}^T \left(\frac{\underline{\mathbf{A}} + \overline{\underline{\mathbf{A}}}^T}{2} \right) \underline{\mathbf{x}} \in W\left(\frac{\underline{\mathbf{A}} + \underline{\mathbf{A}}^*}{2} \right) + W\left(\frac{\underline{\mathbf{A}} - \underline{\mathbf{A}}^*}{2} \right).\end{aligned}$$

Aus Lemma 5.10 folgt, dass diese Einschlussmenge ein Rechteck ist. □

Beispiel:

Wir betrachten die Matrix

$$A = \begin{pmatrix} 4 & 0 & -3 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{pmatrix}$$

und wollen mit Hilfe der Sätze von Gerschgorin und Bendixson deren Eigenwerte abschätzen. Für die Gerschgorinkreise von $\underline{\mathbf{A}}$ bzw. $\underline{\mathbf{A}}^T$ erhalten wir

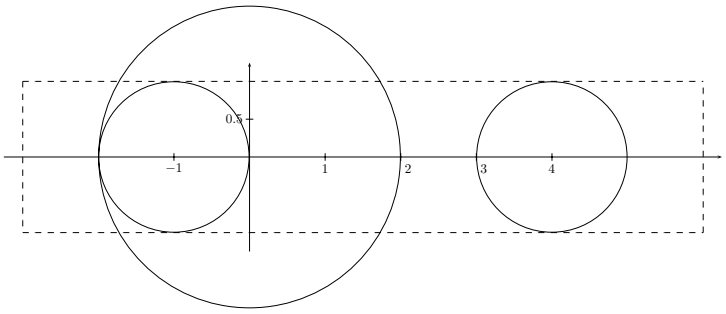
$$\begin{aligned} K_1 &= \{z : |z-4| \le 3\}, & K_1^* &= \{z : |z-4| \le 1\}, \\ K_2 &= \{z : |z+1| \le 1\}, & K_2^* &= \{z : |z+1| \le 1\}, \\ K_3 &= \{z : |z| \le 2\}, & K_3^* &= \{z : |z| \le 4\}. \end{aligned}$$

Anwendung des Satzes von Bendixson ergibt

$$\underline{\mathbf{H}} = \frac{\underline{\mathbf{A}} + \overline{\underline{\mathbf{A}}}^T}{2} = \begin{pmatrix} 4 & 0 & -2 \\ 0 & -1 & 1 \\ -2 & 1 & 0 \end{pmatrix}, \qquad \underline{\mathbf{S}} = \frac{\underline{\mathbf{A}} - \overline{\underline{\mathbf{A}}}^T}{2} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Wir schätzen die Eigenwerte von $\underline{\mathbf{H}}$ nun wieder mit Hilfe von Satz 5.8 ab und erhalten $[4-2, 4+2] \cup [-1-1, -1+1] \cup [0-3, 0+3] = [-3, 6]$.

Die Eigenwerte von $\underline{\mathbf{S}}$ sind in $[0-i, 0+i] = [-i, i]$ enthalten. Nach dem Satz von Bendixson liegen die Eigenwerte von $\underline{\mathbf{A}}$ also im Rechteck $R = [-3, 6] + [-i, i]$.



Die tatsächlichen Eigenwerte von $\underline{\mathbf{A}}$ sind

$$S(\underline{\mathbf{A}}) = \{-1.7838\ldots, 0.1198\ldots, 4.6679\ldots\}.$$