Kapitel 3

Iterative Verfahren für LGS

3.1 Der Banach'sche Fixpunktsatz

Sei $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ invertierbar und das LGS $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$ gegeben. Ein ***iteratives Verfahren*** besteht aus einer Berechnungsvorschrift

$$\underline{\mathbf{x}}^{(j+1)} = F\big(\underline{\mathbf{x}}^{(j)}\big)$$

mit der Hoffnung, dass für $j \longrightarrow \infty$ die iterativ gewonnenen Werte $\underline{\mathbf{x}}^{(j)}$ gegen die exakte Lösung $\underline{\mathbf{x}}$ streben. Wir berechnen also keine $''$exakte$''$ Lösung (soweit dies aufgrund der durch die Gleitkomma-Arithmetik auftretenden Rundungsfehler überhaupt möglich war), sondern versuchen, uns näher an die Lösung von $\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}}$ $''$heranzutasten$''$. Dazu verwenden wir folgende Idee: wir wählen eine (möglichst einfach zu berechnende) invertierbare Matrix $\underline{\mathbf{B}} \in \mathbb{R}^{n\times n}$ so, dass

$$\begin{aligned}
\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}} \quad &\Longleftrightarrow \quad \underline{\mathbf{B}} \cdot \underline{\mathbf{x}} + (\underline{\mathbf{A}} - \underline{\mathbf{B}}) \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}} \\
&\Longleftrightarrow \quad \underline{\mathbf{B}} \cdot \underline{\mathbf{x}} = \underline{\mathbf{b}} - (\underline{\mathbf{A}} - \underline{\mathbf{B}}) \cdot \underline{\mathbf{x}} \\
&\Longleftrightarrow \quad \underline{\mathbf{x}} = \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} - \underline{\mathbf{B}}^{-1} \cdot (\underline{\mathbf{A}} - \underline{\mathbf{B}}) \cdot \underline{\mathbf{x}} \\
&\Longleftrightarrow \quad \underline{\mathbf{x}} = \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} - \big(\underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}} - \underline{\mathbf{I}}\big) \cdot \underline{\mathbf{x}} \\
&\Longleftrightarrow \quad \underline{\mathbf{x}} = \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + \big(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\big) \cdot \underline{\mathbf{x}}.
\end{aligned}$$

Dann ist $\underline{\mathbf{x}}$ ist ein ***Fixpunkt*** der Funktion

$$F(\underline{\mathbf{x}}) = \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + \big(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\big) \cdot \underline{\mathbf{x}}.$$

Die Iteration ergibt sich aus der Vorschrift

$$\underline{\mathbf{x}}^{(j+1)} = \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + \big(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\big) \cdot \underline{\mathbf{x}}^{(j)}.$$

Definition 3.1:

Eine Abbildung $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ heißt ***Kontraktion*** bezüglich einer Vektornorm $\|\cdot\|$, wenn eine Konstante $\varrho < 1$ so existiert, dass

$$\|F(\underline{\mathbf{x}}) - F(\underline{\mathbf{y}})\| \leq \varrho \cdot \|\underline{\mathbf{x}} - \underline{\mathbf{y}}\| \qquad \forall \quad \underline{\mathbf{x}}, \underline{\mathbf{y}} \in \mathbb{R}^n$$

gilt.

Satz 3.2: (Banach'scher Fixpunktsatz für den $\mathbb{R}^n$)

Ist $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ eine Kontraktion bezüglich einer Vektornorm $\|\cdot\|$, dann besitzt F genau einen Fixpunkt $\underline{\mathbf{x}}^*$, d.h. $F(\underline{\mathbf{x}}^*) = \underline{\mathbf{x}}^*$, und für jedes $\underline{\mathbf{x}}^{(0)} \in \mathbb{R}^n$ konvergiert die Folge

$$\underline{\mathbf{x}}^{(j+1)} = F(\underline{\mathbf{x}}^{(j)}), \qquad j \in \mathbb{N}_0$$

gegen $\underline{\mathbf{x}}^*$.

Beweis:
Zur ***Eindeutigkeit*** des Fixpunktes: Angenommen, es existieren zwei verschiedene Fixpunkte, seien diese $\underline{\mathbf{x}}^*$ und $\underline{\mathbf{y}}^*$. Dann gilt

$$\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\| > 0$$

und damit

$$1 = \frac{\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|}{\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|}.$$

Nutzt man jetzt die Fixpunkteigenschaft von $\underline{\mathbf{x}}^*$ und $\underline{\mathbf{y}}^*$, so lässt sich die obige Gleichheit fortsetzen durch

$$\frac{\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|}{\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|} = \frac{\|F(\underline{\mathbf{x}}^*) - F(\underline{\mathbf{y}}^*)\|}{\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|} \leq \frac{\varrho \cdot \|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|}{\|\underline{\mathbf{x}}^* - \underline{\mathbf{y}}^*\|} = \varrho < 1,$$

wobei die erste Abschätzung durch die Kontraktionseigenschaft der Abbildung folgt und die letzte Abschätzung nach Definition der Kontraktion gilt. Das ist jedoch ein Widerspruch, damit muss die Annahme falsch gewesen sein, das heißt, der Fixpunkt ist eindeutig.

Zur ***Existenz*:** Es existiert ein $\underline{\mathbf{x}} \in \mathbb{R}^n$ mit

$$\|F(\underline{\mathbf{x}})\| < \infty$$

(ansonsten wäre die Definition der Abbildung wenig sinnvoll!) und damit folgt

$$\begin{aligned} \|F(\underline{\mathbf{x}}^{(0)})\| &\leq \|F(\underline{\mathbf{x}}^{(0)}) - F(\underline{\mathbf{x}})\| + \|F(\underline{\mathbf{x}})\| \\ &< \varrho \cdot \|\underline{\mathbf{x}}^{(0)} - \underline{\mathbf{x}}\| + \|F(\underline{\mathbf{x}})\| \\ &< \infty, \end{aligned}$$

denn nach Definition gilt $\varrho < 1$ und die Vektordifferenz $\underline{\mathbf{x}}^{(0)} - \underline{\mathbf{x}}$ ist beschränkt. Damit folgt, dass auch

$$\| \underbrace{\underline{\mathbf{x}}^{(1)}}_{F(\underline{\mathbf{x}}^{(0)})} - \underline{\mathbf{x}}^{(0)}\| \leq \|F(\underline{\mathbf{x}}^{(0)})\| + \|\underline{\mathbf{x}}^{(0)}\| < \infty.$$

Allgemein ist dann für $j \geq 1$

$$\begin{aligned} \|\underline{\mathbf{x}}^{(j+1)} - \underline{\mathbf{x}}^{(j)}\| &= \|F(\underline{\mathbf{x}}^{(j)}) - F(\underline{\mathbf{x}}^{(j-1)})\| \\ &\leq \varrho \cdot \|\underline{\mathbf{x}}^{(j)} - \underline{\mathbf{x}}^{(j-1)}\| \\ &\vdots \\ &\leq \varrho^j \cdot \|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\|. \end{aligned}$$

Also gilt für $N > 0$

$$\begin{aligned}\|\underline{\mathbf{x}}^{(j+N)} - \underline{\mathbf{x}}^{(j)}\| &\leq \sum_{k=1}^{N} \|\underline{\mathbf{x}}^{(j+k)} - \underline{\mathbf{x}}^{(j+k-1)}\| \leq \sum_{k=1}^{N} \varrho^{j+k-1} \cdot \|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\| \\ &= \sum_{k=0}^{N-1} \varrho^{j+k} \cdot \|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\| = \varrho^j \cdot \|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\| \cdot \sum_{k=0}^{N-1} \varrho^k \\ &\leq \varrho^j \cdot \|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\| \cdot \sum_{k=0}^{\infty} \varrho^k = \varrho^j \cdot \frac{1}{1-\varrho} \cdot \|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\|.\end{aligned}$$

Da $\|\underline{\mathbf{x}}^{(1)} - \underline{\mathbf{x}}^{(0)}\|$ konstant ist und $\varrho < 1$ nach Voraussetzung, strebt der Ausdruck für $j \longrightarrow \infty$ gegen Null. Damit ist die Folge der $\underline{\mathbf{x}}^{(j)}$ eine Cauchy-Folge (bzgl. der gewählten Vektornorm) und konvergiert gegen einen Grenzwert $\underline{\mathbf{x}}^*$. Dieser Grenzwert ist der gesuchte Fixpunkt, denn es gilt

$$\begin{aligned}\|F(\underline{\mathbf{x}}^*) - \underline{\mathbf{x}}^*\| &\leq \underbrace{\|F(\underline{\mathbf{x}}^*) - F(\underline{\mathbf{x}}^{(j)})\|}_{\leq \varrho \cdot \|\underline{\mathbf{x}}^* - \underline{\mathbf{x}}^{(j)}\|} + \|\underbrace{F(\underline{\mathbf{x}}^{(j)})}_{\underline{\mathbf{x}}^{(j+1)}} - \underline{\mathbf{x}}^*\| \\ &\leq \varrho \cdot \|\underline{\mathbf{x}}^* - \underline{\mathbf{x}}^{(j)}\| + \|\underline{\mathbf{x}}^{(j+1)} - \underline{\mathbf{x}}^*\|.\end{aligned}$$

Dabei ist $\underline{\mathbf{x}}^*$ aber der Grenzwert der Folge der $\underline{\mathbf{x}}^{(j)}$, so dass der letzte Ausdruck der Ungleichungskette für $j \longrightarrow \infty$ gegen Null strebt. Damit ist

$$F(\underline{\mathbf{x}}^*) = \underline{\mathbf{x}}^*,$$

$\underline{\mathbf{x}}^*$ also Fixpunkt der Abildung F. □

Für den obigen Ansatz

$$\underline{\mathbf{x}}^{(j+1)} = \underbrace{\underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}}^{(j)}}_{:= F(\underline{\mathbf{x}}^{(j)})}$$

muss also gelten, dass F eine Kontraktion ist, das heißt es muss

$$\begin{aligned}\|F(\underline{\mathbf{x}}) - F(\underline{\mathbf{y}})\| &= \|\underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} - (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{y}}\| \\ &= \|(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})(\underline{\mathbf{x}} - \underline{\mathbf{y}})\| \\ &\overset{!}{\leq} \varrho \cdot \|\underline{\mathbf{x}} - \underline{\mathbf{y}}\|\end{aligned}$$

gelten. F ist damit zum Beispiel Kontraktion, wenn

$$\|\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\| < 1$$

für eine verträgliche Matrixnorm ist.

Ziel:

Man wähle eine günstige (und leicht zu berechnende) Matrix $\underline{\mathbf{B}}$, so dass die Abbildung

$$\tilde{F}(\underline{\mathbf{x}}) = (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})\, \underline{\mathbf{x}}$$

eine Kontraktion ist (ideal wäre natürlich $\underline{\mathbf{B}} = \underline{\mathbf{A}}$, jedoch macht die Berechnung der Inversen von $\underline{\mathbf{A}}$ mehr Aufwand als die Lösung des LGS nach herkömmlichen direkten Verfahren, somit ist diese Wahl für $\underline{\mathbf{B}}$ nicht empfehlenswert. Man wird versuchen, eine leicht zu berechnende Matrix zu finden, die möglichst viel mit $\underline{\mathbf{A}}$ ″zu tun″ hat).

3.2 Spektralradius und Konvergenz

Definition 3.3:

Das ***Spektrum*** $S(\underline{\mathbf{A}})$ einer Matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ ist die Menge aller Eigenwerte von $\underline{\mathbf{A}}$, das heißt es ist

$$S(\underline{\mathbf{A}}) \;=\; \big\{\, \lambda \in \mathbb{C} \;\big|\; \det(\underline{\mathbf{A}} - \lambda \underline{\mathbf{I}}) \;=\; 0 \big\}.$$

Der ***Spektralradius*** ist definiert als

$$\rho\,(\underline{\mathbf{A}}) \;=\; \max_{\lambda \in S(\underline{\mathbf{A}})} |\lambda|.$$

Satz 3.4:

Sei $\|\cdot\|_\infty$ die Maximum-Vektornorm und $\|\cdot\|_\infty$ bezeichne außerdem die zugehörige Matrixnorm, das heißt, hier die Zeilensummennorm. Dann gilt

$$\rho\,(\underline{\mathbf{A}}) \;=\; \limsup_{j\to\infty} \|\underline{\mathbf{A}}^j\|_\infty^{\frac{1}{j}}.$$

Insbesondere ist damit

$$\rho\,(\underline{\mathbf{A}}) \;\le\; \|\underline{\mathbf{A}}\|_\infty.$$

Bemerkung:

Da Vektornormen im $\mathbb{R}^n$ äquivalent sind, das heißt

$$c_1 \cdot \|\underline{\mathbf{x}}\|_\infty \;\le\; \|\underline{\mathbf{x}}\|_v \;\le\; c_2 \cdot \|\underline{\mathbf{x}}\|_\infty$$

für beliebige Vektornormen $\|\cdot\|_v$ mit passend gewählten Konstanten c_1, c_2 gilt, gilt die Aussage von Satz 3.4 für beliebige Vektornormen und damit verträgliche und konsistente Matrixnormen.

Beweis von Satz 3.4:

Wir betrachten die Fortsetzung von $\|\cdot\|_\infty$ auf $\mathbb{C}^n$ (das heißt, die auftretenden Beträge sind gegebenenfalls im komplexen Sinne zu verstehen).

1) Sei λ ein Eigenwert von $\underline{\mathbf{A}}$, also $\lambda \in S(\underline{\mathbf{A}})$, und $\underline{\mathbf{x}}$ sei ein zugehöriger Eigenvektor mit

$$\|\underline{\mathbf{x}}\|_\infty \;=\; 1.$$

Dann ist

$$\underline{\mathbf{A}} \cdot \underline{\mathbf{x}} \;=\; \lambda \cdot \underline{\mathbf{x}}$$

und es folgt

$$\|\underline{\mathbf{A}}^j\|_\infty \cdot \overbrace{\|\underline{\mathbf{x}}\|_\infty}^{=1} \ge \|\underline{\mathbf{A}}^j \cdot \underline{\mathbf{x}}\|_\infty = \|\lambda^j \cdot \underline{\mathbf{x}}\|_\infty = |\lambda^j| \cdot \underbrace{\|\underline{\mathbf{x}}\|_\infty}_{=1} = |\lambda^j| = |\lambda|^j,$$

also

$$\|\underline{\mathbf{A}}^j\|_\infty \;\ge\; |\lambda|^j$$

für beliebiges $\lambda \in S(\underline{\mathbf{A}})$. Damit ist

$$\|\underline{\mathbf{A}}^j\|_\infty^{\frac{1}{j}} \geq |\lambda| \qquad \forall \quad \lambda \in S(\underline{\mathbf{A}}),$$

das heißt,

$$\|\underline{\mathbf{A}}^j\|_\infty^{\frac{1}{j}} \geq \rho(\underline{\mathbf{A}})$$

für alle $j = 1, 2, \ldots$ und es folgt

$$\limsup_{j\to\infty} \|\underline{\mathbf{A}}^j\|_\infty^{\frac{1}{j}} \geq \rho(\underline{\mathbf{A}}).$$

2) Es existiert eine invertierbare Matrix $\underline{\mathbf{S}} \in \mathbb{C}^{n\times n}$ mit

$$\underline{\mathbf{S}}^{-1} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{S}} = \begin{bmatrix} \lambda_1 & * & & \\ & \ddots & \ddots & \\ & & \ddots & * \\ & & & \lambda_n \end{bmatrix} = \underline{\mathbf{J}}.$$

$\underline{\mathbf{J}}$ ist eine ***Jordan'sche Normalform*** von $\underline{\mathbf{A}}$ mit $* \in \{0, 1\}$ und o.B.d.A. sei $|\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_n|$. Sei ferner $\varepsilon > 0$ und $\underline{\mathbf{D}}_\varepsilon$ eine Diagonalmatrix der Form

$$\underline{\mathbf{D}}_\varepsilon = \begin{bmatrix} 1 & & & & \\ & \varepsilon & & & \\ & & \varepsilon^2 & & \\ & & & \ddots & \\ & & & & \varepsilon^{n-1} \end{bmatrix}.$$

Dann ist

$$\begin{aligned}
&\underline{\mathbf{D}}_\varepsilon^{-1} \cdot \underline{\mathbf{J}} \cdot \underline{\mathbf{D}}_\varepsilon \\
&= \begin{bmatrix} 1 & & & & \\ & \varepsilon^{-1} & & & \\ & & \varepsilon^{-2} & & \\ & & & \ddots & \\ & & & & \varepsilon^{-n+1} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & * & & \\ & \ddots & \ddots & \\ & & \ddots & * \\ & & & \lambda_n \end{bmatrix} \cdot \begin{bmatrix} 1 & & & & \\ & \varepsilon & & & \\ & & \varepsilon^2 & & \\ & & & \ddots & \\ & & & & \varepsilon^{n-1} \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 & * & & & \\ & \varepsilon^{-1}\lambda_2 & \varepsilon^{-1}* & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon^{-n+2}* \\ & & & & \varepsilon^{-n+1}\lambda_n \end{bmatrix} \cdot \begin{bmatrix} 1 & & & & \\ & \varepsilon & & & \\ & & \varepsilon^2 & & \\ & & & \ddots & \\ & & & & \varepsilon^{n-1} \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 & \varepsilon* & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon* \\ & & & \lambda_n \end{bmatrix} =: \underline{\mathbf{J}}_\varepsilon,
\end{aligned}$$

das heißt, es gilt

$$\underbrace{\|\underline{\mathbf{D}}_\varepsilon^{-1} \cdot \underline{\mathbf{J}} \cdot \underline{\mathbf{D}}_\varepsilon\|_\infty}_{\|\underline{\mathbf{J}}_\varepsilon\|_\infty} \leq \underbrace{|\lambda_n|}_{\rho(\underline{\mathbf{A}})} + \varepsilon = \rho(\underline{\mathbf{A}}) + \varepsilon.$$

Wegen

$$\begin{aligned}
\|\underline{\mathbf{A}}^j\|_\infty &= \|\underline{\mathbf{S}} \cdot \underbrace{\overbrace{(\underline{\mathbf{S}}^{-1} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{S}})}^{\underline{\mathbf{J}}} \cdot \overbrace{(\underline{\mathbf{S}}^{-1} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{S}})}^{\underline{\mathbf{J}}} \cdots \overbrace{(\underline{\mathbf{S}}^{-1} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{S}})}^{\underline{\mathbf{J}}}}_{j-mal} \cdot \underline{\mathbf{S}}^{-1}\|_\infty \\
&= \|\underline{\mathbf{S}} \cdot \underline{\mathbf{J}}^j \cdot \underline{\mathbf{S}}^{-1}\|_\infty \\
&\leq \|\underline{\mathbf{S}}\|_\infty \cdot \|\underline{\mathbf{J}}^j\|_\infty \cdot \|\underline{\mathbf{S}}^{-1}\|_\infty \\
&= \kappa(\underline{\mathbf{S}}) \cdot \|\underline{\mathbf{J}}^j\|_\infty \\
&= \kappa(\underline{\mathbf{S}}) \cdot \|\underline{\mathbf{D}}_\varepsilon \cdot \underbrace{\overbrace{(\underline{\mathbf{D}}_\varepsilon^{-1} \cdot \underline{\mathbf{J}} \cdot \underline{\mathbf{D}}_\varepsilon)}^{\underline{\mathbf{J}}_\varepsilon} \cdot \overbrace{(\underline{\mathbf{D}}_\varepsilon^{-1} \cdot \underline{\mathbf{J}} \cdot \underline{\mathbf{D}}_\varepsilon)}^{\underline{\mathbf{J}}_\varepsilon} \cdots \overbrace{(\underline{\mathbf{D}}_\varepsilon^{-1} \cdot \underline{\mathbf{J}} \cdot \underline{\mathbf{D}}_\varepsilon)}^{\underline{\mathbf{J}}_\varepsilon}}_{j-mal} \cdot \underline{\mathbf{D}}_\varepsilon^{-1}\|_\infty \\
&\leq \kappa(\underline{\mathbf{S}}) \cdot \|\underline{\mathbf{D}}_\varepsilon\|_\infty \cdot \|\underline{\mathbf{J}}_\varepsilon^j\|_\infty \cdot \|\underline{\mathbf{D}}_\varepsilon^{-1}\|_\infty \\
&= \kappa(\underline{\mathbf{S}}) \cdot \kappa(\underline{\mathbf{D}}_\varepsilon) \cdot \|\underline{\mathbf{J}}_\varepsilon^j\|_\infty \\
&\leq \kappa(\underline{\mathbf{S}}) \cdot \kappa(\underline{\mathbf{D}}_\varepsilon) \cdot \underbrace{\|\underline{\mathbf{J}}_\varepsilon\|_\infty \cdot \|\underline{\mathbf{J}}_\varepsilon\|_\infty \cdots \|\underline{\mathbf{J}}_\varepsilon\|_\infty}_{j-mal} \\
&\leq \kappa(\underline{\mathbf{S}}) \cdot \kappa(\underline{\mathbf{D}}_\varepsilon) \cdot (\rho(\underline{\mathbf{A}}) + \varepsilon)^j
\end{aligned}$$

(Ausnutzung der Konsistenz !) gilt

$$\|\underline{\mathbf{A}}^j\|_\infty \leq \kappa(\underline{\mathbf{S}}) \cdot \kappa(\underline{\mathbf{D}}_\varepsilon) \cdot (\rho(\underline{\mathbf{A}}) + \varepsilon)^j \qquad \forall \quad j \in \mathbb{N}.$$

Dann folgt

$$\begin{aligned}
\limsup_{j\to\infty} \|\underline{\mathbf{A}}^j\|_\infty^{\frac{1}{j}} &\leq \limsup_{j\to\infty} \left[(\kappa(\underline{\mathbf{S}}) \cdot \kappa(\underline{\mathbf{D}}_\varepsilon))^{\frac{1}{j}} \cdot (\rho(\underline{\mathbf{A}}) + \varepsilon)\right] \\
&= (\rho(\underline{\mathbf{A}}) + \varepsilon) \cdot \limsup_{j\to\infty} (\kappa(\underline{\mathbf{S}}) \cdot \kappa(\underline{\mathbf{D}}_\varepsilon))^{\frac{1}{j}} = \rho(\underline{\mathbf{A}}) + \varepsilon.
\end{aligned}$$

Strebt jetzt ε gegen Null, so liefert dies

$$\limsup_{j\to\infty} \|\underline{\mathbf{A}}^j\|_\infty^{\frac{1}{j}} \leq \rho(\underline{\mathbf{A}}),$$

und zusammen mit dem ersten Teil des Beweises folgt die Behauptung. □

Bemerkung:

Aus Satz 3.4 folgt: Für beliebiges $\varepsilon > 0$ existiert ein $j_0 \in \mathbb{N}$, so dass für alle $j \geq j_0$

$$\|\underline{\mathbf{A}}^j\|^{\frac{1}{j}} \leq \rho(\underline{\mathbf{A}}) + \varepsilon$$

gilt.

Wir betrachten jetzt das Iterationsverfahren:

Satz 3.5:

Das Iterationsverfahren

$$\underline{\mathbf{x}}^{(j+1)} = \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}}^{(j)}$$

konvergiert genau dann für einen beliebigen Startvektor $\underline{\mathbf{x}}^{(0)}$, wenn

$$\rho(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) < 1$$

gilt. Hinreichend für die Konvergenz ist nach Satz 3.4 auch

$$\|\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\| \; < \; 1$$

oder

$$\|(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^j\|^{\frac{1}{j}} \; < \; 1$$

für eine beliebige Matrixnorm, die konsistent und zu einer Vektornorm (Operatornorm) verträglich ist.

Beweis:

1) Sei

$$\rho\,(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \; < \; 1.$$

Dann existiert nach Satz 3.4 ein $m \in \mathbb{N}$ mit

$$\|(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^m\|_\infty \; < \; 1.$$

Sei weiter

$$F(\underline{\mathbf{x}}) \; = \; (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}} + \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}},$$

das heißt, es gilt

$$F(\underline{\mathbf{x}}^{(j)}) \; = \; \underline{\mathbf{x}}^{(j+1)}$$

und deshalb z.B.

$$\begin{aligned} F^2(\underline{\mathbf{x}}) &= (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \left[(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}} + \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}}\right] + \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} \\ &= (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^2 \cdot \underline{\mathbf{x}} + (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}} + \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}}. \end{aligned}$$

Allgemein erhält man folgende Berechnungsvorschrift:

$$F^m(\underline{\mathbf{x}}) \; = \; (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^m \cdot \underline{\mathbf{x}} + \underbrace{\sum_{k=0}^{m-1} (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^k \cdot \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}}}_{:=\underline{\mathbf{c}}_m}.$$

Dann ist F^m eine Kontraktion, denn es ist

$$\begin{aligned} \|F^m(\underline{\mathbf{x}}) - F^m(\underline{\mathbf{y}})\|_\infty &= \|(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^m (\underline{\mathbf{x}} - \underline{\mathbf{y}})\|_\infty \\ &\leq \underbrace{\|(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^m\|_\infty}_{<1} \cdot \|\underline{\mathbf{x}} - \underline{\mathbf{y}}\|_\infty. \end{aligned}$$

Damit besitzt F^m nach Satz 3.2 einen eindeutigen Fixpunkt $\underline{\mathbf{x}}^*$ für einen beliebigen Startvektor. Für $k = 0, ..., m-1$ ist dann

$$F^{jm}(\underline{\mathbf{x}}^{(k)}) \; = \; \underline{\mathbf{x}}^{(jm+k)},$$

und jede Teilfolge $(\underline{\mathbf{x}}^{(jm+k)})_{j=0}^{\infty}$ strebt für $j \to \infty$ gegen $\underline{\mathbf{x}}^*$. Also konvergiert die gesamte Folge $(\underline{\mathbf{x}}^{(j)})_{j=0}^{\infty}$ gegen $\underline{\mathbf{x}}^*$.

2) Das Iterationsverfahren konvergiere mit Grenzwert $\underline{\mathbf{x}}^*$, das heißt, es gelte

$$\|\underline{\mathbf{x}}^{(j)} - \underline{\mathbf{x}}^*\|_\infty \quad \stackrel{j \to \infty}{\longrightarrow} \quad 0.$$

Sei $\underline{\mathbf{y}} \in \mathbb{C}^n$ Eigenvektor des betragsmäßig größten Eigenwertes λ von $(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})$ und sei der Startvektor der Iteration in der Weise

$$\underline{\mathbf{x}}^{(0)} \; = \; \underline{\mathbf{x}}^* + \underline{\mathbf{y}}$$

gewählt. Dann ist

$$\begin{aligned}
\underline{\mathbf{x}}^{(j)} - \underline{\mathbf{x}}^* &= \left[(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}}^{(j-1)} + \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}}\right] - \overbrace{\left[(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}}^* + \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{b}}\right]}^{\underline{\mathbf{x}}^* \, ist \, Fixpunkt} \\
&= (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})(\underline{\mathbf{x}}^{(j-1)} - \underline{\mathbf{x}}^*) \\
&= \ldots \\
&= (\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})^j \underbrace{(\underline{\mathbf{x}}^{(0)} - \underline{\mathbf{x}}^*)}_{=\underline{\mathbf{y}}} \\
&= \lambda^j \cdot \underline{\mathbf{y}}.
\end{aligned}$$

Da

$$\|\underline{\mathbf{x}}^{(j)} - \underline{\mathbf{x}}^*\|_\infty \quad \stackrel{j\to\infty}{\longrightarrow} \quad 0$$

gilt, folgt

$$\lim_{j\to\infty} \|\lambda^j \cdot \underline{\mathbf{y}}\|_\infty = 0,$$

das aber ist aufgrund der Normaxiome gleichbedeutend mit

$$\lim_{j\to\infty} |\lambda^j| \cdot \|\underline{\mathbf{y}}\|_\infty = 0.$$

Das liefert schließlich

$$|\lambda| < 1.$$

Da λ nach Voraussetzung der betragsmäßig größte Eigenwert von $(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})$ war, also

$$|\lambda| = \rho\,(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})$$

gilt, folgt die Behauptung. □

Bemerkung:

Die Konvergenz des Iterationsverfahrens ist umso schneller, je kleiner der Spektralradius von $(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})$ ist.

3.3 Gauß-Seidel-Verfahren und Jacobi-Verfahren

Problem:

Es ist die Matrix $\underline{\mathbf{B}}$ so zu wählen, dass

- $\underline{\mathbf{B}}^{-1}$ leicht zu berechnen ist
- $\rho\,(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) < 1$ gilt (je kleiner der Spektralradius ist, desto schneller konvergiert das Verfahren)

1. Idee:

Man wähle die Matrix $\underline{\mathbf{B}}$ als Diagonalmatrix der Form

$$\underline{\mathbf{B}} = \begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix}$$

falls $a_{jj} \neq 0$, $j = 1, ..., n$, das heißt, $\underline{\mathbf{B}}$ enthält nur die Einträge der Hauptdiagonalen von $\underline{\mathbf{A}}$. Das Verfahren wird dann ***Jacobi-Verfahren (Gesamtschritt-Verfahren)*** genannt. Man erhält als Iterationsvorschrift

$$\underline{\mathbf{B}} \cdot \underline{\mathbf{x}}^{(j+1)} = \underline{\mathbf{b}} + (\underline{\mathbf{B}} - \underline{\mathbf{A}}) \cdot \underline{\mathbf{x}}^{(j)}$$

bzw. explizit

$$\begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(j+1)} \\ \vdots \\ x_n^{(j+1)} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} - \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \begin{bmatrix} x_1^{(j)} \\ \vdots \\ x_n^{(j)} \end{bmatrix}.$$

Komponentenweise ergibt sich somit die folgende Berechnungsvorschrift:

$$x_k^{(j+1)} = \frac{b_k - \sum\limits_{\substack{r=1 \\ r \neq k}}^{n} a_{kr} \cdot x_r^{(j)}}{a_{kk}}, \qquad k = 1, ..., n.$$

Bezeichnet l die Anzahl der Nicht-Null-Elemente pro Zeile von $\underline{\mathbf{A}}$, so benötigen wir $l - 1$ Multiplikationen, $l - 1$ Additionen und eine Division zur Berechnung einer Komponente x_k^{j+1}, also $(2l - 1)n$ flops pro Iterationsschritt.

2. Idee:

Man wähle die Matrix $\underline{\mathbf{B}}$ in der Form

$$\underline{\mathbf{B}} = \begin{bmatrix} a_{11} & & \\ \vdots & \ddots & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix},$$

das heißt, $\underline{\mathbf{B}}$ ist eine untere Dreiecksmatrix (nämlich das untere Dreieck der Ausgangsmatrix $\underline{\mathbf{A}}$). Es sei $a_{jj} \neq 0$, $j = 1, ..., n$ vorausgesetzt (sonst ist $\underline{\mathbf{B}}$ nicht invertierbar!). Das sich ergebende Verfahren heißt ***Gauß-Seidel-Verfahren (Einzelschritt-Verfahren)***. Zu lösen ist die folgende Gleichung:

$$\begin{bmatrix} a_{11} & & & \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} x_1^{(j+1)} \\ \vdots \\ x_n^{(j+1)} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} - \underbrace{\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}}_{\underline{\mathbf{A}} - \underline{\mathbf{B}}} \cdot \begin{bmatrix} x_1^{(j)} \\ \vdots \\ x_n^{(j)} \end{bmatrix}.$$

Komponentenweise ergibt sich die Berechnungsvorschrift (k-te Zeile)

$$\sum_{r=1}^{k} a_{kr} \cdot x_r^{(j+1)} = b_k - \sum_{r=k+1}^{n} a_{kr} \cdot x_r^{(j)}$$

$$\iff x_k^{(j+1)} = \frac{b_k - \sum\limits_{r=1}^{k-1} a_{kr} \cdot x_r^{(j+1)} - \sum\limits_{r=k+1}^{n} a_{kr} \cdot x_r^{(j)}}{a_{kk}}.$$

Der Rechenaufwand pro Iterationsschritt hängt wieder linear von der Anzahl der Nicht-Null-Elemente in einer Zeile von $\underline{\mathbf{A}}$ ab (beide Verfahren sind somit günstig für große, dünn besetzte Matrizen). Besitzt $\underline{\mathbf{A}}$ nur l Nichtnulleinträge pro Zeile, so ergibt sich zur Berechnung der

k-ten Komponente $x_k^{(j+1)}$ wieder ein Aufwand von $(l-1)$ Multiplikationen, $l-1$ Additionen und einer Division , also wiederum $(2l-1)n$ flops pro Iterationsschritt.

Bemerkung:

Während beim Jacobi-Verfahren alle Komponenten des neuen Vektors $\underline{\mathbf{x}}^{(j+1)}$ aus dem alten Vektor $\underline{\mathbf{x}}^{(j)}$ gleichzeitig berechnet werden können, sind die Komponenten von $\underline{\mathbf{x}}^{(j+1)}$ beim Gauß-Seidel-Verfahren auch von den schon berechneten Werten des neuen Vektors abhängig. Daher spricht man von Gesamtschritt- bzw. Einzelschrittverfahren.

Für welche Matrizen konvergieren die Verfahren ?

Satz 3.6:

Ist $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, dann konvergiert das Gauß-Seidel-Verfahren.

Beweis:

Wir betrachten eine Zerlegung von $\underline{\mathbf{A}}$ der Form

$$\underline{\mathbf{A}} = \underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{L}}^T$$
$$= \begin{bmatrix} 0 & & & \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} + \begin{bmatrix} a_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} + \begin{bmatrix} 0 & a_{21} & \cdots & a_{n1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n,n-1} \\ & & & 0 \end{bmatrix}$$

(beachte, dass wegen der Symmetrie von $\underline{\mathbf{A}}$ die Indizierung der dritten Matrix korrekt ist, es gilt $a_{jk} = a_{kj}$). Dann folgt mit $\underline{\mathbf{B}} = \underline{\mathbf{L}} + \underline{\mathbf{D}}$

$$\begin{aligned} \underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}} &= \left(\underline{\mathbf{I}} - (\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1}(\underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{L}}^T)\right) \\ &= \underline{\mathbf{I}} - \underline{\mathbf{I}} - (\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1} \cdot \underline{\mathbf{L}}^T = -(\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1} \cdot \underline{\mathbf{L}}^T. \end{aligned}$$

Zu zeigen ist also, dass

$$\rho\left(-(\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1} \cdot \underline{\mathbf{L}}^T\right) < 1$$

gilt. Da $\underline{\mathbf{A}}$ positiv definit ist, folgt $a_{jj} > 0$, dies ist wegen

$$a_{jj} = \underline{\mathbf{e}}_j^T \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{e}}_j > 0$$

leicht einzusehen. Setzen wir

$$\underline{\mathbf{D}}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{a_{11}} & & \\ & \ddots & \\ & & \sqrt{a_{nn}} \end{bmatrix},$$

so folgt

$$\begin{aligned}
& \underline{\mathbf{D}}^{\frac{1}{2}} \cdot \left(-(\underline{\mathbf{L}}+\underline{\mathbf{D}})^{-1} \cdot \underline{\mathbf{L}}^T\right) \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} \\
= & -\left(\underline{\mathbf{D}}^{-\frac{1}{2}}\right)^{-1} \cdot (\underline{\mathbf{L}}+\underline{\mathbf{D}})^{-1} \cdot \underline{\mathbf{L}}^T \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} \\
= & -\left[(\underline{\mathbf{L}}+\underline{\mathbf{D}})\left(\underline{\mathbf{D}}^{-\frac{1}{2}}\right)\right]^{-1} \cdot \underline{\mathbf{L}}^T \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} \\
= & -\big[\underline{\mathbf{L}} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} + \underbrace{\underline{\mathbf{D}} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}}}_{\underline{\mathbf{D}}^{\frac{1}{2}}}\big]^{-1} \cdot \underline{\mathbf{L}}^T \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} \\
= & -\left[\underline{\mathbf{L}} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} + \underline{\mathbf{D}}^{\frac{1}{2}}\right]^{-1} \cdot \underbrace{\left(\underline{\mathbf{D}}^{-\frac{1}{2}}\right)^{-1} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}}}_{\underline{\mathbf{I}}} \cdot \underline{\mathbf{L}}^T \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} \\
= & -\left[\left(\underline{\mathbf{D}}^{-\frac{1}{2}}\right)\left(\underline{\mathbf{L}} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} + \underline{\mathbf{D}}^{\frac{1}{2}}\right)\right]^{-1} \cdot \underbrace{\underline{\mathbf{D}}^{-\frac{1}{2}} \cdot \underline{\mathbf{L}}^T \cdot \underline{\mathbf{D}}^{-\frac{1}{2}}}_{:=\underline{\mathbf{L}}_1^T} \\
= & -\big[\underbrace{\underline{\mathbf{D}}^{-\frac{1}{2}} \cdot \underline{\mathbf{L}} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}}}_{\underline{\mathbf{L}}_1} + \underbrace{\underline{\mathbf{D}}^{-\frac{1}{2}} \cdot \underline{\mathbf{D}}^{\frac{1}{2}}}_{\underline{\mathbf{I}}}\big]^{-1} \cdot \underline{\mathbf{L}}_1^T \\
= & -(\underline{\mathbf{L}}_1+\underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{L}}_1^T.
\end{aligned}$$

Diese Matrix ist ähnlich zu $-(\underline{\mathbf{L}}+\underline{\mathbf{D}})^{-1} \cdot \underline{\mathbf{L}}^T$ und besitzt daher dieselben Eigenwerte. Sei λ ein solcher Eigenwert von $-(\underline{\mathbf{L}}_1+\underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{L}}_1^T$ mit zugehörigem Eigenvektor $\underline{\mathbf{x}} \in \mathbb{C}^n$ und $\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{x}} = 1$ (das heißt, es wird ein bereits normierter Eigenvektor betrachtet). Dann ergibt sich

$$\begin{aligned}
& -(\underline{\mathbf{L}}_1+\underline{\mathbf{I}})^{-1} \cdot \underline{\mathbf{L}}_1^T \cdot \underline{\mathbf{x}} = \lambda \cdot \underline{\mathbf{x}} \\
\iff & -\underline{\mathbf{L}}_1^T \cdot \underline{\mathbf{x}} = \lambda \cdot (\underline{\mathbf{L}}_1+\underline{\mathbf{I}}) \cdot \underline{\mathbf{x}} \\
\implies & -\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1^T \cdot \underline{\mathbf{x}} = \lambda \cdot \overline{\underline{\mathbf{x}}}^T \cdot (\underline{\mathbf{L}}_1+\underline{\mathbf{I}}) \cdot \underline{\mathbf{x}} \\
\iff & -\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1^T \cdot \underline{\mathbf{x}} = \lambda \cdot \overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1 \cdot \underline{\mathbf{x}} + \lambda \cdot \underbrace{\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{x}}}_{=1}
\end{aligned}$$

Die Ausdrücke auf beiden Seiten der entstandenen Gleichung sind skalar. Durch Transponieren der linken Seite der Gleichung bleibt die Aussage also richtig. Außerdem ist die Matrix $\underline{\mathbf{L}}_1$ reell, so dass $\underline{\mathbf{L}}_1 = \overline{\underline{\mathbf{L}}}_1$ gilt. Damit wird die letzte Gleichung zu

$$-\underline{\mathbf{x}}^T \cdot \underline{\mathbf{L}}_1 \cdot \overline{\underline{\mathbf{x}}} = \lambda \cdot \overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1 \cdot \underline{\mathbf{x}} + \lambda \iff -\overline{(\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1 \cdot \underline{\mathbf{x}})} = \lambda \cdot \overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1 \cdot \underline{\mathbf{x}} + \lambda.$$

Setzt man jetzt

$$\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1 \cdot \underline{\mathbf{x}} = \alpha + i\beta,$$

so lässt sich die letzte Gleichung umschreiben in

$$\begin{aligned}
& -(\alpha - i\beta) = \lambda \cdot (\alpha + i\beta) + \lambda = \lambda \cdot (\alpha + i\beta + 1) \\
\implies & \alpha^2 + \beta^2 = |\lambda|^2 \left((\alpha+1)^2 + \beta^2\right) \\
\implies & |\lambda|^2 = \frac{\alpha^2+\beta^2}{1+2\alpha+\alpha^2+\beta^2}.
\end{aligned}$$

(bei der auftretenden Division ist zunächst $(\alpha+1)^2 + \beta^2 = 0$ auszuschliessen, dieser Fall kann aber nach den folgenden Rechnungen gar nicht auftreten). Kann man jetzt zeigen, dass

$$1 + 2\alpha > 0$$

gilt, so wäre die Aussage $|\lambda| < 1$ bewiesen. Da $\underline{\mathbf{A}}$ und damit auch

$$\begin{aligned}
\underline{\mathbf{D}}^{-\frac{1}{2}} \cdot \underline{\mathbf{A}} \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} &= \underline{\mathbf{D}}^{-\frac{1}{2}} \cdot (\underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{L}}^T) \cdot \underline{\mathbf{D}}^{-\frac{1}{2}} \\
&= \underline{\mathbf{L}}_1 + \underline{\mathbf{I}} + \underline{\mathbf{L}}_1^T
\end{aligned}$$

positiv definit ist, folgt wegen $\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1 \cdot \underline{\mathbf{x}} = \alpha + i\beta$ und $\overline{\underline{\mathbf{x}}}^T \cdot \underline{\mathbf{L}}_1^T \cdot \underline{\mathbf{x}} = \alpha - i\beta$

$$\begin{aligned} 0 &< \overline{\underline{\mathbf{x}}}^T \cdot (\underline{\mathbf{I}} + \underline{\mathbf{L}}_1 + \underline{\mathbf{L}}_1^T) \cdot \underline{\mathbf{x}} \\ &= 1 + (\alpha + i\beta) + (\alpha - i\beta) = 1 + 2\alpha \end{aligned}$$

(hierdurch wird der kritische Fall in der obigen Division ausgeschlossen). Also folgt

$$|\lambda| < 1$$

und damit die Behauptung. □

Definition 3.7:

Eine Matrix $\underline{\mathbf{A}} = (a_{jk})_{j,k=1}^n \in \mathbb{R}^{n\times n}$ heißt ***zeilenweise (strikt) diagonaldominant***, falls für $j = 1, \ldots, n$ gilt:

$$|a_{jj}| > \sum_{\substack{k=1 \\ k\neq j}}^{n} |a_{jk}|,$$

das heißt, in jeder Zeile ist das Diagonalelement betragsmäßig größer als die Summe der Beträge der weiteren Einträge in der Zeile.
Eine Matrix $\underline{\mathbf{A}} = (a_{jk})_{j,k=1}^n \in \mathbb{R}^{n\times n}$ heißt ***spaltenweise (strikt) diagonaldominant***, falls für $k = 1, \ldots, n$ gilt:

$$|a_{kk}| > \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{jk}|.$$

Satz 3.8:

1) Ist $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ zeilenweise (strikt) diagonaldominant, so konvergieren das Jacobi- und das Gauß-Seidel-Verfahren.

2) Ist $\underline{\mathbf{A}} \in \mathbb{R}^{n\times n}$ spaltenweise (strikt) diagonaldominant, so konvergieren das Jacobi-Verfahren und das Gauß-Seidel-Verfahren.

Beweis:

1) Betrachte zunächst das Jacobi-Verfahren. Hierbei wurde $\underline{\mathbf{B}}$ in der Form

$$\underline{\mathbf{B}} = \begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix}$$

gewählt, daraus erhält man

$$\begin{aligned} \|\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\|_\infty &= \left\| \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -\frac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\frac{a_{n1}}{a_{nn}} & \cdots & \cdots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{bmatrix} \right\|_\infty \\ &= \max_{j=1,\ldots,n} \sum_{\substack{k=1 \\ k\neq j}}^{n} \frac{|a_{jk}|}{|a_{jj}|}. \end{aligned}$$

Sei nun $\underline{\mathbf{A}}$ zeilenweise diagonaldominant, so dass

$$|a_{jj}| \;>\; \sum_{\substack{k=1\\k\neq j}}^{n} |a_{jk}| \qquad\qquad j=1,...,n$$

gilt. Daraus folgt

$$1 \;>\; \sum_{\substack{k=1\\k\neq j}}^{n} \frac{|a_{jk}|}{|a_{jj}|} \qquad\qquad j=1,...,n,$$

bzw.

$$1 \;>\; \max_{j=1,...,n} \sum_{\substack{k=1\\k\neq j}}^{n} \frac{|a_{jk}|}{|a_{jj}|} = \|\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1}\underline{\mathbf{A}}\|_\infty \,,$$

das heißt, für zeilenweise Diagonaldominanz konvergiert das Jacobi-Verfahren.

2) Es sei nun $\underline{\mathbf{A}}$ spaltenweise diagonaldominant, also

$$|a_{kk}| \;>\; \sum_{\substack{j=1\\j\neq k}}^{n} |a_{jk}| \qquad\qquad k=1,...,n.$$

Wir betrachten zunächst die Matrix $\underline{\mathbf{I}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{B}}^{-1}$. Hierfür folgt

$$\begin{aligned}
\|\underline{\mathbf{I}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{B}}^{-1}\|_1 \;&=\; \left\| \begin{bmatrix} 0 & -\frac{a_{12}}{a_{22}} & \cdots & \cdots & -\frac{a_{1n}}{a_{nn}} \\ -\frac{a_{21}}{a_{11}} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -\frac{a_{n-1,n}}{a_{nn}} \\ -\frac{a_{n1}}{a_{11}} & \cdots & \cdots & -\frac{a_{n,n-1}}{a_{n-1,n-1}} & 0 \end{bmatrix} \right\|_1 \\
&=\; \max_{k=1,...,n} \sum_{\substack{j=1\\j\neq k}}^{n} \frac{|a_{jk}|}{|a_{kk}|} = \max_{k=1,...,n} \underbrace{\frac{1}{|a_{kk}|} \cdot \sum_{\substack{j=1\\j\neq k}}^{n} |a_{jk}|}_{<1} < 1.
\end{aligned}$$

Also folgt

$$\rho\,(\underline{\mathbf{I}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{B}}^{-1}) \;<\; 1.$$

Betrachte nun die Matrix

$$\underline{\mathbf{B}}^{-1} \cdot (\underline{\mathbf{I}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{B}}^{-1}) \cdot \underline{\mathbf{B}} = \underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1}\underline{\mathbf{A}}.$$

Diese ist ähnlich zu $\underline{\mathbf{I}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{B}}^{-1}$, besitzt demzufolge dieselben Eigenwerte, so dass auch

$$\rho\,(\underline{\mathbf{B}}^{-1} \cdot (\underline{\mathbf{I}} - \underline{\mathbf{A}} \cdot \underline{\mathbf{B}}^{-1}) \cdot \underline{\mathbf{B}}) = \rho\,(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}) < 1.$$

gilt. Das aber liefert die Konvergenz des Jacobi-Verfahrens im Falle der spaltenweisen Diagonaldominanz von $\underline{\mathbf{A}}$.

3) Wir betrachten jetzt das Gauß-Seidel-Verfahren. Es sei $\underline{\mathbf{A}}$ zeilenweise diagonaldominant, das heißt, es gelte

$$|a_{jj}| \;>\; \sum_{\substack{k=1\\k\neq j}}^{n} |a_{jk}|, \qquad\qquad (j=1,...,n). \qquad\qquad (*)$$

Setzt man jetzt

$$\begin{aligned}\underline{\mathbf{A}} &= \underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{U}} \\ &= \begin{bmatrix} 0 & & & \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} + \begin{bmatrix} a_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} + \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{bmatrix},\end{aligned}$$

so folgt

$$\underline{\mathbf{B}} \;=\; \underline{\mathbf{L}} + \underline{\mathbf{D}}.$$

Das liefert mit $\underline{\mathbf{L}}_1 := -\underline{\mathbf{D}}^{-1} \cdot \underline{\mathbf{L}}$ und $\underline{\mathbf{U}}_1 := -\underline{\mathbf{D}}^{-1} \cdot \underline{\mathbf{U}}$

$$\begin{aligned}\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}} &= \underline{\mathbf{I}} - (\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1}(\underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{U}}) \\ &= \underline{\mathbf{I}} - \underline{\mathbf{I}} - (\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1}\underline{\mathbf{U}} \\ &= -(-\underline{\mathbf{D}}\,\underline{\mathbf{L}}_1 + \underline{\mathbf{D}})^{-1}\,(-\underline{\mathbf{D}}\,\underline{\mathbf{U}}_1) \\ &= (-\underline{\mathbf{L}}_1 + \underline{\mathbf{I}})^{-1}\,\underline{\mathbf{D}}^{-1}\,\underline{\mathbf{D}}\,\underline{\mathbf{U}}_1 = (\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1}\,\underline{\mathbf{U}}_1.\end{aligned}$$

Wir zeigen nun, dass

$$\|(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1} \cdot \underline{\mathbf{U}}_1\|_\infty \;<\; 1$$

gilt. Wir nutzen hier folgende Bezeichnungen. Für eine Matrix $\underline{\mathbf{G}} = (g_{jk})_{j=1,k=1}^{m,n} \in \mathbb{R}^{m\times n}$ ist $|\underline{\mathbf{G}}| := (|g_{jk}|)_{j=1,k=1}^{m,n}$ die sogenannte Betragsmatrix. Weiter sei für zwei Matrizen $\underline{\mathbf{G}}, \underline{\mathbf{H}} \in \mathbb{R}^{m\times n}$ die Beziehung $\underline{\mathbf{G}} < \underline{\mathbf{H}}$ richtig falls für alle Elemente von $\underline{\mathbf{G}}$ und $\underline{\mathbf{H}}$ gilt $g_{jk} < h_{jk}$, $j = 1, \ldots, m$, $k = 1, \ldots, n$.
Dann finden wir folgende Abschätzung.

$$\begin{aligned}|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1| \cdot \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} &= \begin{bmatrix} 0 & \left|\frac{a_{12}}{a_{11}}\right| & \cdots & \left|\frac{a_{1n}}{a_{11}}\right| \\ \left|\frac{a_{21}}{a_{22}}\right| & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \left|\frac{a_{n-1,n}}{a_{n-1,n-1}}\right| \\ \left|\frac{a_{n1}}{a_{nn}}\right| & \cdots & \left|\frac{a_{n,n-1}}{a_{nn}}\right| & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \\ &\leq \underbrace{\|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_\infty}_{<1 \; wegen \; (*)} \cdot \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} < \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}.\end{aligned}$$

Da insbesondere

$$|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1| \;=\; |\underline{\mathbf{L}}_1| + |\underline{\mathbf{U}}_1|$$

gilt (beide Matrizen sind Dreiecksmatrizen ($\underline{\mathbf{L}}_1$ eine untere, $\underline{\mathbf{U}}_1$ eine obere) mit Nullen auf der Hauptdiagonalen), folgt

$$\begin{aligned}|\underline{\mathbf{U}}_1| \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} &= (|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1| - |\underline{\mathbf{L}}_1|) \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\ &\leq \|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_\infty \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - |\underline{\mathbf{L}}_1| \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.\end{aligned}$$

Weiter sind sowohl $\underline{\mathbf{L}}_1$ als auch $|\underline{\mathbf{L}}_1|$ ***nilpotent***, das heißt, es gilt

$$\underline{\mathbf{L}}_1^n \;=\; |\underline{\mathbf{L}}_1|^n \;=\; \underline{\mathbf{0}}$$

und man erhält folgende nützlichen Zusammenhänge,

$$(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1} \;=\; \underline{\mathbf{I}} + \underline{\mathbf{L}}_1 + \underline{\mathbf{L}}_1^2 + \cdots + \underline{\mathbf{L}}_1^{n-1}$$

und

$$(\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1|)^{-1} \;=\; \underline{\mathbf{I}} + |\underline{\mathbf{L}}_1| + |\underline{\mathbf{L}}_1|^2 + \cdots + |\underline{\mathbf{L}}_1|^{n-1}.$$

Damit folgt

$$\begin{aligned} |(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1}| &= |\underline{\mathbf{I}} + \underline{\mathbf{L}}_1 + \cdots + \underline{\mathbf{L}}_1^{n-1}| \\ &\leq \underline{\mathbf{I}} + |\underline{\mathbf{L}}_1| + \cdots + |\underline{\mathbf{L}}_1|^{n-1} \\ &= (\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1|)^{-1}. \end{aligned}$$

Also erhalten wir

$$\begin{aligned} & |(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1} \cdot \underline{\mathbf{U}}_1| \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\ \leq\; & |(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1}| \cdot |\underline{\mathbf{U}}_1| \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\ \leq\; & (\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1|)^{-1} \cdot \left(\|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_\infty \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - |\underline{\mathbf{L}}_1| \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right) \\ =\; & (\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1|)^{-1} \cdot \big((\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1|) \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \underbrace{\|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_\infty \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \underline{\mathbf{I}} \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{(\|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_\infty - 1) \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}} \big) \end{aligned}$$

$$\begin{aligned} =\; & \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \underbrace{(\overbrace{\|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_\infty}^{<1} - 1)}_{<0} \cdot \underbrace{(\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1|)^{-1}}_{>\underline{\mathbf{I}}} \cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\ <\; & \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \end{aligned}$$

Das liefert aber

$$\|(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1} \cdot \underline{\mathbf{U}}_1\|_\infty \;<\; 1,$$

also Konvergenz des Verfahrens.

4) Sei nun $\underline{\mathbf{A}}$ spaltenweise diagonaldominant. Der Beweis der Konvergenz des Gauß-Seidel-Verfahrens lässt sich in ähnlicher Weise führen wie für zeilenweise Diagonaldominanz. Seien $\underline{\mathbf{L}}_1 := -\underline{\mathbf{L}}\underline{\mathbf{D}}^{-1}$ und $\underline{\mathbf{U}}_1 := -\underline{\mathbf{U}}\underline{\mathbf{D}}^{-1}$. Wir erhalten dann

$$\underline{\mathbf{I}} - \underline{\mathbf{A}}\underline{\mathbf{B}}^{-1} = \underline{\mathbf{I}} - (\underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{U}})(\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1} = -\underline{\mathbf{U}}(\underline{\mathbf{L}} + \underline{\mathbf{D}})^{-1} = \underline{\mathbf{U}}_1(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1}.$$

Weiter folgt

$$|\underline{\mathbf{L}}_1^T + \underline{\mathbf{U}}_1^T| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \leq \|\underline{\mathbf{L}}_1^T + \underline{\mathbf{U}}_1^T\|_\infty \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_1 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

und mit $|\underline{\mathbf{U}}_1^T| = |\underline{\mathbf{L}}_1^T + \underline{\mathbf{U}}_1^T| - |\underline{\mathbf{L}}_1^T|$ wieder

$$|\underline{\mathbf{U}}_1^T| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \leq \|\underline{\mathbf{L}}_1^T + \underline{\mathbf{U}}_1^T\|_1 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - |\underline{\mathbf{L}}_1^T| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Außerdem gilt

$$(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1^T)^{-1} = \sum_{k=0}^{n-1} (\underline{\mathbf{L}}_1^T)^k \leq \sum_{k=0}^{n-1} |\underline{\mathbf{L}}_1^T|^k = (\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1^T|)^{-1}$$

und wir erhalten wie vorher

$$|(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1^T)^{-1}\, \underline{\mathbf{U}}_1^T| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \leq \underbrace{(\|\underline{\mathbf{L}}_1 + \underline{\mathbf{U}}_1\|_1 - 1)}_{<0} (\underline{\mathbf{I}} - |\underline{\mathbf{L}}_1^T|)^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} < \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Also ergibt sich

$$\|(\underline{\mathbf{U}}_1\,(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1})^T\|_\infty = \|\underline{\mathbf{U}}_1(\underline{\mathbf{I}} - \underline{\mathbf{L}}_1)^{-1}\|_1 < 1$$

und damit $\|\underline{\mathbf{I}} - \underline{\mathbf{A}}\underline{\mathbf{B}}^{-1}\|_1 < 1$. Analog wie schon im Beweisschritt 2) ist daher

$$\rho(\underline{\mathbf{I}} - \underline{\mathbf{A}}\underline{\mathbf{B}}^{-1}) = \rho(\underline{\mathbf{B}}^{-1}(\underline{\mathbf{I}} - \underline{\mathbf{A}}\underline{\mathbf{B}}^{-1}))\underline{\mathbf{B}}) = \rho(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1}\underline{\mathbf{A}}) < 1$$

und die Behauptung folgt. □

MAPLE-Prozeduren:

Wir betrachten die Maple-Prozeduren für das Jacobi- und das Gauß-Seidel-Verfahren. Ein Iterationsschritt des Jacobi-Verfahrens lässt sich folgendermaßen programmieren:

```
> restart;
> with(LinearAlgebra):
> printlevel:=0:
> JacobiIt:=proc(A, x, b)        #Jacobi-Iterationsschritt
  local j, k, n, y;
  # option trace;
  n:=Dimension(b); y:=Vector(n);
  for j from 1 to n do
      y[j] := b[j];
      for k from 1 to j − 1 do y[j] := y[j] − A[j, k] ∗ x[k] end do;
      for k from j + 1 to n do y[j] := y[j] − A[j, k] ∗ x[k] end do;
      y[j] := y[j]/A[j, j]
  end do;
  for j from 1 to n do x[j] := y[j] end do;
  end proc:
```

Dabei ist A die Koeffizientenmatrix des LGS, b die rechte Seite und x ein beliebiger Vektor der Länge n. Für $x = x^{(j)}$ berechnet die Prozedur $x^{(j+1)}$. Für das gesamte Jacobi-Verfahren erhalten wir:

```
> #Jacobi-Verfahren, Maximumnorm des Residuums soll kleiner tol sein
  Jacobi:=proc(A, b, tol)
  global res, x;
  local n, k, JacobiIt, vectornorm, residuum, vnorm;
  JacobiIt:=proc(A, x, b) ... end proc;
  vectornorm:=proc(x)        #Maximumnorm des Vektors x
  local j, n, vnorm;
  vnorm := 0; n :=Dimension(x);
  for j from 1 to n do
     if abs(x[j]) > vnorm then vnorm :=abs(x[j]) end if
  end do;
  eval(vnorm)
  end proc;
  residuum:=proc(A, x, b)        #Residuum Ax − b
  global res;
  local j, k, n;
  n:=Dimension(b); res:=Vector(n);
  for j from 1 to n do
     for k from 1 to n do res[j] := res[j] + A[j, k] ∗ x[k] end do;
     res[j] := res[j] − b[j]
  end do;
  end proc;
  n:=Dimension(b); x:=Vector(n); res:=Vector(n);
  residuum(A, x, b);
  vnorm:= vectornorm(res); k := 0;
  while (vnorm > tol) do
     JacobiIt(A, x, b);
     residuum(A, x, b);
     vnorm:=vectornorm(res);
     k := k + 1;
     print(k, x);        # druckt nach jedem Iterationsschritt Näherung x
  end do;
  end proc:
```

Der Aufruf erfolgt nun beispielsweise folgendermaßen:

```
> A:=Matrix([[10., −1., 0., 2.], [1., 12., −1., 2.], [−2., 1., 15., 0.], [1., −2., 0., 20.]]):
  b:=Vector([11, 14, 14, 19]): tol := 0.001:
  Jacobi(A, b, tol);
```

mit dem Resultat

$$
\begin{aligned}
&1, \quad [1.100000000, 1.166666667, .9333333333, .9500000000],\\
&2, \quad [1.026666667, .9944444442, 1.002222222, 1.011666667],\\
&3, \quad [.9971111106, .9960185183, 1.003925926, .9981111110],\\
&4, \quad [.9999796298, 1.000882717, .9998802467, .9997462965],\\
&5, \quad [1.000139013, 1.000034003, .9999384360, 1.000089290],\\
&6, \quad [.9999855420, .9999684042, 1.000016269, .9999964500].
\end{aligned}
$$

Nach sechs Iterationsschritten bricht das Programm ab, da die Fehlerschranke *tol* erreicht ist. Analog erhalten wir für das Gauß-Seidel-Verfahren den Iterationsschritt:

```
> GaussSeidelIt:=proc(A, x, b)        #Gauss-Seidel-Iterationsschritt
  local j, n, k, z, y;
  n:=Dimension(b); y:=Vector(n);
  for j from 1 to n do
      z := b[j];
      for k from 1 to j − 1 do z := z − A[j, k] ∗ y[k] end do;
      for k from j + 1 to n do z := z − A[j, k] ∗ x[k] end do;
      y[j] := z/A[j, j]
  end do;
  for j from 1 to n do x[j] := y[j] end do;
  end proc:
```

und das gesamte Verfahren lautet analog zum Jacobi-Verfahren:

```
>  GaussSeidel:=proc(A, b, tol)        # Gauss-Seidel-Verfahren
   global res, x;
   local n, k, vectornorm, GaussSeidelIt, residuum, vnorm;
   GaussSeidelIt:=proc(A, x, b) ... end proc;
   vectornorm:=proc(x)        # Maximumnorm
   local j, n, vnorm;
   vnorm := 0; n:=Dimension(x);
   for j from 1 to n do if abs(x[j]) > vnorm then vnorm :=abs(x[j]) end if; end do;
   eval(vnorm)
   end;
   residuum:=proc(A, x, b)        #Residuum Ax − b
   global res;
   local j, k, n;
   n:=Dimension(b);
   res:=Vector(n);
   for j from 1 to n do
       for k from 1 to n do res[j] := res[j] + A[j, k] ∗ x[k] end do;
       res[j] := res[j] − b[j]
   end do;
   end proc;
   n:=Dimension(b); x:=Vector(n);
   res:=Vector(n);
   residuum(A, x, b);
   vnorm:=vectornorm(res);
   k := 0;
   while (vnorm > tol) do
       GaussSeidelIt(A, x, b);
       residuum(A, x, b);
       vnorm:= vectornorm(res);
       k := k + 1;
       print(k, x);
   end do;
   end proc:
```

Der Aufruf erfolgt (mit dem selben Beispiel für die Matrix $\underline{\mathbf{A}}$ und den Vektor $\underline{\mathbf{b}}$ wie oben durch

```
GaussSeidel(A, b, 0.001);
```

mit dem Resultat

$$\begin{array}{ll} 1 & [1.100000000, 1.075000000, 1.008333333, 1.002500000], \\ 2 & [1.007000000, .9996944442, 1.000953704, .9996194445], \\ 3 & [1.000045555, 1.000139105, .9999968007, 1.000011633], \\ 4 & [1.000011584, .9999968292, 1.000001756, .9999991040]. \end{array}$$

Schon nach vier Iterationsschritten ist der Lösungsvektor mit der vorgegebenen Toleranz angenähert.

3.4 Relaxationsverfahren

Die Konvergenzgeschwindigkeit der Iterationsverfahren hängt wesentlich von $\rho\left(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}}\right)$ ab. Unter Umständen lässt sich der Spektralradius $\rho(\underline{\mathbf{I}} - \underline{\mathbf{B}}^{-1} \cdot \underline{\mathbf{A}})$ durch eine leichte Änderung der Matrix $\underline{\mathbf{B}}$ noch verkleinern. Das hier betrachtete ***Relaxationsverfahren*** ist eine spezielle ***Variante des Gauß-Seidel-Verfahrens*** mit

$$\underline{\mathbf{B}}(\omega) = \frac{1}{\omega} \cdot (\underline{\mathbf{D}} + \omega \cdot \underline{\mathbf{L}}) = \frac{1}{\omega} \cdot \underline{\mathbf{D}} + \underline{\mathbf{L}},$$

wobei

$$\underline{\mathbf{A}} \;=\; \underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{U}}.$$

Speziell ergibt sich für $\omega = 1$ das bereits bekannte Gauß-Seidel-Verfahren, für $\omega < 1$ nennt man das entstehende Verfahren ***Unterrelaxation,*** im Falle $\omega > 1$ wird es als ***Überrelaxation*** bezeichnet. Insbesondere erhält man das ***SOR-Verfahren (successive overrelaxation)*** (auf dieses wird im Folgenden noch genauer eingegangen). Formal lautet die Iterationsvorschrift

$$\underline{\mathbf{B}}(\omega) \cdot \underline{\mathbf{x}}^{(j+1)} \;=\; \underline{\mathbf{b}} - (\underline{\mathbf{A}} - \underline{\mathbf{B}}(\omega)) \cdot \underline{\mathbf{x}}^{(j)},$$

die äquivalent ist mit

$$\begin{aligned} \frac{1}{\omega} \cdot (\underline{\mathbf{D}} + \omega \cdot \underline{\mathbf{L}}) \cdot \underline{\mathbf{x}}^{(j+1)} &= \underline{\mathbf{b}} - \left(\underline{\mathbf{U}} + \left(1 - \frac{1}{\omega}\right) \cdot \underline{\mathbf{D}}\right) \cdot \underline{\mathbf{x}}^{(j)} \\ &= \underline{\mathbf{b}} - \frac{1}{\omega} \cdot \left((\omega - 1) \cdot \underline{\mathbf{D}} + \omega \cdot \underline{\mathbf{U}}\right) \cdot \underline{\mathbf{x}}^{(j)} \\ &= \underline{\mathbf{b}} + \frac{1}{\omega} \cdot \left((1 - \omega) \cdot \underline{\mathbf{D}} - \omega \cdot \underline{\mathbf{U}}\right) \cdot \underline{\mathbf{x}}^{(j)}. \end{aligned}$$

Explizit erhalten wir in Matrixschreibweise

$$\begin{bmatrix} \frac{a_{11}}{\omega} & & & \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & \frac{a_{nn}}{\omega} \end{bmatrix} \begin{bmatrix} x_1^{(j+1)} \\ \vdots \\ \vdots \\ x_n^{(j+1)} \end{bmatrix}$$

$$= \begin{bmatrix} b_1 \\ \vdots \\ \vdots \\ b_n \end{bmatrix} - \begin{bmatrix} \frac{-1+\omega}{\omega} a_{11} & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & \frac{-1+\omega}{\omega} a_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(j)} \\ \vdots \\ \vdots \\ x_n^{(j)} \end{bmatrix}.$$

Komponentenweise ergibt sich der Zusammenhang (k-te Zeile)

$$\sum_{r=1}^{k-1} a_{kr} \cdot x_r^{(j+1)} + \frac{a_{kk}}{\omega} \cdot x_k^{(j+1)} \;=\; b_k - \frac{\omega - 1}{\omega} \cdot a_{kk} \cdot x_k^{(j)} - \sum_{r=k+1}^{n} a_{kr} \cdot x_r^{(j)},$$

also

$$x_k^{(j+1)} \;=\; \frac{\omega}{a_{kk}} \cdot \left(b_k - \sum_{r=1}^{k-1} a_{kr} \cdot x_r^{(j+1)} - \sum_{r=k+1}^{n} a_{kr} \cdot x_r^{(j)} \right) + (1 - \omega) \cdot x_k^{(j)}.$$

Satz 3.9:

Bezeichnet man

$$\underline{\mathbf{H}}(\omega) \;:=\; \underline{\mathbf{I}} - \underline{\mathbf{B}}(\omega)^{-1} \cdot \underline{\mathbf{A}},$$

so gilt

$$\rho\left(\underline{\mathbf{H}}(\omega)\right) \;\geq\; |\omega - 1|.$$

Insbesondere folgt, dass nur der Relaxationsparameter $\omega \in (0, 2)$ interessant sind, da sonst $|\omega - 1| \geq 1$ und somit keine Konvergenz des Verfahrens vorliegt.

Beweis:

Es sei $\underline{\mathbf{L}}_1 := -\underline{\mathbf{D}}^{-1}\,\underline{\mathbf{L}}$ und $\underline{\mathbf{U}}_1 := -\underline{\mathbf{D}}^{-1}\,\underline{\mathbf{U}}$. Dann gilt wegen

$$\underline{\mathbf{B}}(\omega) = \frac{1}{\omega}\underline{\mathbf{D}} + \underline{\mathbf{L}} = \frac{1}{\omega}\underline{\mathbf{D}}(\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1)$$

der Zusammenhang

$$\begin{aligned}
\underline{\mathbf{H}}(\omega) &= \underline{\mathbf{I}} - \underline{\mathbf{B}}(\omega)^{-1}\,\underline{\mathbf{A}} = \underline{\mathbf{I}} - (\frac{1}{\omega}\underline{\mathbf{D}} + \underline{\mathbf{L}})^{-1}\,(\underline{\mathbf{L}} + \underline{\mathbf{D}} + \underline{\mathbf{U}}) \\
&= \underline{\mathbf{I}} - (\frac{1}{\omega}\underline{\mathbf{D}} + \underline{\mathbf{L}})^{-1} \left((\frac{1}{\omega}\underline{\mathbf{D}} + \underline{\mathbf{L}}) + (1 - \frac{1}{\omega})\underline{\mathbf{D}} + \underline{\mathbf{U}} \right) \\
&= \underline{\mathbf{I}} - \underline{\mathbf{I}} - \left(\frac{1}{\omega}\underline{\mathbf{D}}(\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1) \right)^{-1} \left((1 - \frac{1}{\omega})\underline{\mathbf{D}} + \underline{\mathbf{U}} \right) \\
&= (\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1)^{-1} \left(-(\omega - 1)\underline{\mathbf{I}} - \omega\underline{\mathbf{D}}^{-1}\underline{\mathbf{U}} \right) \\
&= (\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1)^{-1}\,((1 - \omega)\underline{\mathbf{I}} + \omega\underline{\mathbf{U}}_1).
\end{aligned}$$

Nun ist $\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1$ eine untere Dreiecksmatrix mit Einsen auf der Hauptdiagonale, also ergibt sich $\det\,(\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1) = 1$ für alle $\omega \in \mathbb{R}$. Für das charakteristische Polynom von $\underline{\mathbf{H}}(\omega)$ folgt

$$\begin{aligned}
p(\lambda) &= \det\,(\underline{\mathbf{H}}(\omega) - \lambda\underline{\mathbf{I}}) = \det\,(\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1) \cdot \det\,(\underline{\mathbf{H}}(\omega) - \lambda\underline{\mathbf{I}}) \\
&= \det\,((1 - \omega)\underline{\mathbf{I}} + \omega\underline{\mathbf{U}}_1 - \lambda(\underline{\mathbf{I}} - \omega\underline{\mathbf{L}}_1)) \\
&= \det\,((1 - \omega - \lambda)\underline{\mathbf{I}} + \omega\underline{\mathbf{U}}_1 + \lambda\,\omega\underline{\mathbf{L}}_1).
\end{aligned}$$

Der konstante Term dieses Polynoms $(-1)^n p(0)$ ist das Produkt der Eigenwerte λ_i von $\underline{\mathbf{H}}(\omega)$,

$$(-1)^n p(0) = \prod_{i=1}^{n} \lambda_i = (-1)^n \det\,((1 - \omega)\underline{\mathbf{I}} + \omega\underline{\mathbf{U}}_1) = (\omega - 1)^n.$$

Damit folgt aber

$$\rho(\underline{\mathbf{H}}(\omega)) = \max_{i=1,\ldots,n} |\lambda_i| \geq |\omega - 1|. \qquad \square$$

Bemerkung:

Für bestimmte Matrizen (positiv definite!), für die das Gauß-Seidel-Verfahren konvergiert, existiert ein $\overline{\omega} \in (1,2)$, so dass $\rho(\underline{\mathbf{H}}(\omega))$ für $\omega \in [1,\overline{\omega}]$ monoton fällt (dabei bezeichnet $\overline{\omega}$ das $''$optimale$''$ ω). Für leichte Überrelaxation erhält man also verhältnismässig schnell konvergierende Verfahren, sukzessives Vergrößern des Relaxationsparameters liefert so das SOR-Verfahren.

MAPLE-Prozedur:

Man kann den Iterationsschritt mit Hilfe der folgenden MAPLE-Prozedur implementieren:

```
> restart;
> with(LinearAlgebra);  printlevel:=0;
> # Relaxation-Iterationsschritt
> RelaxIt:=proc(A, x, b, r)        # r Relaxationsparameter
  local j, k, n, z, y;
  n:=Dimension(b);
  y:=x;
  for j from 1 to n do
      z := b[j];
      for k from 1 to j − 1 do z := z − A[j, k] ∗ y[k] end do;
      for k from j + 1 to n do z := z − A[j, k] ∗ y[k] end do;
      y[j] := r ∗ z/A[j, j] + (1 − r) ∗ y[j]
  end do;
  x:=y;
  end proc;
```

Dabei bezeichnet r den Relaxationsparameter, x die Lösung des letzten Iterationsschrittes und b die rechte Seite des LGS. Die fertige Lösungsprozedur mit Verwendung eines Relaxationsverfahrens lautet dann wie folgt:

```
> # Relaxationsverfahren
> relax:=proc(A, b, tol, r)
  local n, k, vectornorm, residuum, RelaxIt;
  global res, vnorm, x;
  RelaxIt:= proc(A, x, b, r) ... end proc;
  vectornorm:=proc(x)
      local j, n, vnorm;
      vnorm := 0;
      n:=Dimension(x);
      for j to n do if vnorm < abs(x[j]) then vnorm :=abs(x[j]) end if end do;
      evalf(vnorm)
  end proc;
  residuum;=proc(A, x, b)
      global res;
      local j, k, n;
      n:=Dimension(b);
      res:=Vector(n);
      for j to n do
          for k to n do res[j] := res[j] + A[j, k] ∗ x[k] end do;
          res[j] := res[j] − b[j]
      end do;
  end proc;
  n:=Dimension(b);
```

```
k := 0; x:=Vector(n);
residuum(A, x, b);
vnorm:=vectornorm(res);
while (vnorm > tol) do
    RelaxIt(A, x, b, r);
    residuum(A, x, b);
    vnorm:=vectornorm(res);
    k := k + 1;
    print(k, x);
end do
end proc;
```

Bemerkung:

Für symmetrische positiv definite Matrizen **A** ist das Relaxationsverfahren konvergent für $0 < \omega < 2$.

Beweis:

Stoer/Burlisch: *Numerische Mathematik 2*, Seite 267-269.